Lecture Notes in Physics

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

The Editorial Policy for Edited Volumes

The series *Lecture Notes in Physics* (LNP), founded in 1969, reports new developments in physics research and teaching - quickly, informally but with a high degree of quality. Manuscripts to be considered for publication are topical volumes consisting of a limited number of contributions, carefully edited and closely related to each other. Each contribution should contain at least partly original and previously unpublished material, be written in a clear, pedagogical style and aimed at a broader readership, especially graduate students and nonspecialist researchers wishing to familiarize themselves with the topic concerned. For this reason, traditional proceedings cannot be considered for this series though volumes to appear in this series are often based on material presented at conferences, workshops and schools.

Acceptance

A project can only be accepted tentatively for publication, by both the editorial board and the publisher, following thorough examination of the material submitted. The book proposal sent to the publisher should consist at least of a preliminary table of contents outlining the structure of the book together with abstracts of all contributions to be included. Final acceptance is issued by the series editor in charge, in consultation with the publisher, only after receiving the complete manuscript. Final acceptance, possibly requiring minor corrections, usually follows the tentative acceptance unless the final manuscript differs significantly from expectations (project outline). In particular, the series editors are entitled to reject individual contributions if they do not meet the high quality standards of this series. The final manuscript must be ready to print, and should include both an informative introduction and a sufficiently detailed subject index.

Contractual Aspects

Publication in LNP is free of charge. There is no formal contract, no royalties are paid, and no bulk orders are required, although special discounts are offered in this case. The volume editors receive jointly 30 free copies for their personal use and are entitled, as are the contributing authors, to purchase Springer books at a reduced rate. The publisher secures the copyright for each volume. As a rule, no reprints of individual contributions can be supplied.

Manuscript Submission

The manuscript in its final and approved version must be submitted in ready to print form. The corresponding electronic source files are also required for the production process, in particular the online version. Technical assistance in compiling the final manuscript can be provided by the publisher‘s production editor(s), especially with regard to the publisher’s own LaTeX macro package which has been specially designed for this series.

LNP Homepage (springerlink.com)

On the LNP homepage you will find:
–The LNP online archive. It contains the full texts (PDF) of all volumes published since 2000. Abstracts, table of contents and prefaces are accessible free of charge to everyone. Information about the availability of printed volumes can be obtained.
–The subscription information. The online archive is free of charge to all subscribers of the printed volumes.
–The editorial contacts, with respect to both scientific and technical matters.
–The author’s / editor’s instructions.

R. Livi A. Vulpiani (Eds.)

The Kolmogorov Legacy in Physics

Springer

Editors

Roberto Livi
Universitá di Firenze
Dipartimento di Fisica
Via Sansone 1
50019 Sesto Fiorentino, Italy

Angelo Vulpiani
Universitá di Roma "La Sapienza"
Dipartimento di Fisica
Piazzale A. Moro 2
00185 Roma, Italy

Translation from the French language edition of *"L'Héritage de Kolmogorov en Physique"* edited by Roberto Livi and Angelo Vulpiani
© 2003 Editions Éditions Belin, ISBN 2-7011-3558-3, France

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at http://dnb.ddb.de
ISSN 0075-8450
ISBN 3-540-20307-9 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by the authors/editor
Data conversion: PTP-Berlin Protago-TeX-Production GmbH
Cover design: *design & production*, Heidelberg

Printed on acid-free paper
54/3141/du - 5 4 3 2 1 0

Lecture Notes in Physics

For information about Vols. 1–589
please contact your bookseller or Springer-Verlag
LNP Online archive: springerlink.com

Vol.590: D. Benest, C. Froeschlé (Eds.), Singularities in Gravitational Systems. Applications to Chaotic Transport in the Solar System.

Vol.591: M. Beyer (Ed.), CP Violation in Particle, Nuclear and Astrophysics.

Vol.592: S. Cotsakis, L. Papantonopoulos (Eds.), Cosmological Crossroads. An Advanced Course in Mathematical, Physical and String Cosmology.

Vol.593: D. Shi, B. Aktaş, L. Pust, F. Mikhailov (Eds.), Nanostructured Magnetic Materials and Their Applications.

Vol.594: S. Odenbach (Ed.),Ferrofluids. Magnetical Controllable Fluids and Their Applications.

Vol.595: C. Berthier, L. P. Lévy, G. Martinez (Eds.), High Magnetic Fields. Applications in Condensed Matter Physics and Spectroscopy.

Vol.596: F. Scheck, H. Upmeier, W. Werner (Eds.), Noncommutative Geometry and the Standard Model of Elememtary Particle Physics.

Vol.597: P. Garbaczewski, R. Olkiewicz (Eds.), Dynamics of Dissipation.

Vol.598: K. Weiler (Ed.), Supernovae and Gamma-Ray Bursters.

Vol.599: J.P. Rozelot (Ed.), The Sun's Surface and Subsurface. Investigating Shape and Irradiance.

Vol.600: K. Mecke, D. Stoyan (Eds.), Morphology of Condensed Matter. Physcis and Geometry of Spatial Complex Systems.

Vol.601: F. Mezei, C. Pappas, T. Gutberlet (Eds.), Neutron Spin Echo Spectroscopy. Basics, Trends and Applications.

Vol.602: T. Dauxois, S. Ruffo, E. Arimondo (Eds.), Dynamics and Thermodynamics of Systems with Long Range Interactions.

Vol.603: C. Noce, A. Vecchione, M. Cuoco, A. Romano (Eds.), Ruthenate and Rutheno-Cuprate Materials. Superconductivity, Magnetism and Quantum Phase.

Vol.604: J. Frauendiener, H. Friedrich (Eds.), The Conformal Structure of Space-Time: Geometry, Analysis, Numerics.

Vol.605: G. Ciccotti, M. Mareschal, P. Nielaba (Eds.), Bridging Time Scales: Molecular Simulations for the Next Decade.

Vol.606: J.-U. Sommer, G. Reiter (Eds.), Polymer Crystallization. Obervations, Concepts and Interpretations.

Vol.607: R. Guzzi (Ed.), Exploring the Atmosphere by Remote Sensing Techniques.

Vol.608: F. Courbin, D. Minniti (Eds.), Gravitational Lensing:An Astrophysical Tool.

Vol.609: T. Henning (Ed.), Astromineralogy.

Vol.610: M. Ristig, K. Gernoth (Eds.), Particle Scattering, X-Ray Diffraction, and Microstructure of Solids and Liquids.

Vol.611: A. Buchleitner, K. Hornberger (Eds.), Coherent Evolution in Noisy Environments.

Vol.612: L. Klein, (Ed.), Energy Conversion and Particle Acceleration in the Solar Corona.

Vol.613: K. Porsezian, V.C. Kuriakose, (Eds.), Optical Solitons. Theoretical and Experimental Challenges.

Vol.614: E. Falgarone, T. Passot (Eds.), Turbulence and Magnetic Fields in Astrophysics.

Vol.615: J. Büchner, C.T. Dum, M. Scholer (Eds.), Space Plasma Simulation.

Vol.616: J. Trampetic, J. Wess (Eds.), Particle Physics in the New Millenium.

Vol.617: L. Fernández-Jambrina, L. M. González-Romero (Eds.), Current Trends in Relativistic Astrophysics, Theoretical, Numerical, Observational

Vol.618: M.D. Esposti, S. Graffi (Eds.), The Mathematical Aspects of Quantum Maps

Vol.619: H.M. Antia, A. Bhatnagar, P. Ulmschneider (Eds.), Lectures on Solar Physics

Vol.620: C. Fiolhais, F. Nogueira, M. Marques (Eds.), A Primer in Density Functional Theory

Vol.621: G. Rangarajan, M. Ding (Eds.), Processes with Long-Range Correlations

Vol.622: F. Benatti, R. Floreanini (Eds.), Irreversible Quantum Dynamics

Vol.623: M. Falcke, D. Malchow (Eds.), Understanding Calcium Dynamics, Experiments and Theory

Vol.624: T. Pöschel (Ed.), Granular Gas Dynamics

Vol.625: R. Pastor-Satorras, M. Rubi, A. Diaz-Guilera (Eds.), Statistical Mechanics of Complex Networks

Vol.626: G. Contopoulos, N. Voglis (Eds.), Galaxies and Chaos

Vol.627: S.G. Karshenboim, V.B. Smirnov (Eds.), Precision Physics of Simple Atomic Systems

Vol.628: R. Narayanan, D. Schwabe (Eds.), Interfacial Fluid Dynamics and Transport Processes

Vol.630: T. Brandes, S. Kettemann (Eds.), Anderson Localization and Its Ramifications

Vol.631: D. J. W. Giulini, C. Kiefer, C. Lämmerzahl (Eds.), Quantum Gravity, From Theory to Experimental Search

Vol.632: A. M. Greco (Ed.), Direct and Inverse Methods in Nonlinear Evolution Equations

Vol.633: H.-T. Elze (Ed.), Decoherence and Entropy in Complex Systems, Based on Selected Lectures from DICE 2002

Vol.634: R. Haberlandt, D. Michel, R. Stannarius (Eds.), Direct and Inverse Methods in Nonlinear Evolution Equations

Vol.635: D. Alloin, W. Gieren (Eds.), Stellar Candles for the Extragalactic Distance Scale

Vol.636: R. Livi, A. Vulpiani (Eds.), The Kolmogorov Legacy in Physics, A Century of Turbulence and Complexity

Preface

I was delighted to learn that R. Livi and A. Vulpiani will edit the book dedicated to the legacy of Kolmogorov in physics. Also, I was very much honored when they invited me to write an introduction for this book. Certainly, it is a very difficult task. Andrei N. Kolmogorov (1903-1987) was a great scientist of the 20th Century, mostly known as a great mathematician. He also had classical results in some parts of physics. Physicists encounter his name at conferences, meetings, and workshops dedicated to turbulence. He wrote his famous papers on this subject in the early Forties. Soon after the results became known worldwide they completely changed the way of thinking of researchers working in hydrodynamics, atmospheric sciences, oceanography, etc. An excellent book by U. Frisch *Turbulence, the Legacy of A.N. Kolmogorov*, published by the Cambridge University Press in 1995 gives a very detailed exposition of Kolmogorov's theory. Sometimes it is stressed that the powerful renormalization group method in statistical physics and quantum field theory that is based upon the idea of scale invariance has as one of its roots the Kolmogorov theory of turbulence. I had heard several times Kolmogorov talking about turbulence and had always been given the impression that these were talks by a pure physicist. One could easily forget that Kolmogorov was a great mathematician. He could discuss concrete equations of state of real gases and liquids, the latest data of experiments, etc. When Kolmogorov was close to eighty I asked him about the history of his discoveries of the scaling laws. He gave me a very astonishing answer by saying that for half a year he studied the results of concrete measurements. In the late Sixties Kolmogorov undertook a trip on board a scientific ship participating in the experiments on oceanic turbulence. Kolmogorov was never seriously interested in the problem of existence and uniqueness of solutions of the Navier-Stokes system. He also considered his theory of turbulence as purely phenomenological and never believed that it would eventually have a mathematical framework.

Kolmogorov laid the foundation for a big mathematical direction, now called the theory of deterministic chaos. In problems of dynamics he always stressed the importance of dynamical systems generated by differential equations and he considered this to be the most important part of the theory. Two great discoveries in non-linear dynamics are connected with the name of Kolmogorov: KAM-theory where the letter K stands for Kolmogorov and

Kolmogorov entropy and Kolmogorov systems, which opened new fields in the analysis of non-linear dynamical systems.

The histories of both discoveries are sufficiently well known. A friend of mine, who was a physicist once told me that KAM-theory is so natural that it is strange that it was not invented by physicists. The role of Kolmogorov's work on entropy in physics is not less than in mathematics. It is not so well known that there was a time when Kolmogorov believed in the importance of dynamical systems with zero entropy and had unpublished notes where he constructed an invariant of dynamical system expressed in terms of the growth of entropies of partitions over big intervals of time. Later, Kolmogorov changed his point of view and formulated a conjecture according to which the phase space of a typical dynamical system consists up to a negligible subset of measure zero of invariant tori and mixing components with positive entropy. To date we have no tools to prove or disprove this conjecture. Also, Kolmogorov's ideas on complexity grew up from his wowhen Kolmogorov believed in the importance of dynamical systems with zero entropy and had unpublished notes where he constructed an invariant of dynamical system expressed in terms of the growth of entropies of partitions over big intervals of time. Later, Kolmogorov changed his point of view and formulated a conjecture according to which the phase space of a typical dynamical system consists up to a negligible subset of measure zero of invariant tori and mixing components with positive entropy. To date we have no tools to prove or disprove this conjecture. Also, Kolmogorov's ideas on complexity grew up from his work on entropy. Physical intuition can be seen in Kolmogorov works on diffusion processes. One of his classmates at the University was M. A. Leontovich who later became a leading physicist working on problems of thermo-nuclear fusion. In 1933 Kolmogorov and Leontovich wrote a joint paper on what was later called *Wiener Sausage.* Many years later Kolmogorov used his intuition to propose the answer to the problem of chasing Brownian particle, which was studied by E. Mishenko and L. Pontrijagin. The joint paper of three authors gave its complete solution.

Kolmogorov made important contributions to biology and linguistics. His knowledge of various parts of human culture was really enormous. He loved music and knew very well poetry and literature. His public lectures like the one delivered on the occasion of his 60th birthday and another one under the title, *Can a Computer Think?* were great social events. For those who ever met or knew Kolmogorov personally, memories about this great man stay forever.

Princeton, *Yakov G. Sinai*
April 2003

Introduction

The centenary of A.N. Kolmogorov, one of the greatest scientists of the 20th century, falls this year, 2003. He was born in Russia on the 25th of April 1903.[1]This is typically the occasion for apologetic portraits or hagiographic surveys about such an intense human and scientific biography.

Various meetings and publications will be devoted to celebrate the work and the character of the great mathematician. So one could wonder why pubblishing a book which simply aims at popularizing his major achievements in fields out of pure mathematics? We are deeply convinced that Kolmogorov's contributions are the cornerstone over which many modern research fields, from physics to computer science and biology, are based and still keep growing. His ideas have been transmitted also by his pupils to generations of scientists. The aim of this book is to extend such knowledge to a wider audience, including cultivated readers, students in scientific disciplines and active researchers.

Unfortunately, we never had the opportunity for sharing, with those who met him, the privilege of discussing and interacting with such a personality. Our only credentials for writing about Kolmogorov come from our scientific activity, which has been and still now is mainly based on some of his fundamental contributions.

In this book we do not try to present the great amount, in number and quality, of refined technical work and intuitions that Kolmogorov devoted to research in pure mathematics, ranging from the theory of probability to stochastic processes, theory of automata and analysis. For this purpose we address the reader to a collection of his papers,[2] which contains also illuminating comments by his pupils and collaborators. Here we want to pursue the goal of accounting for the influence of Kolmogorov's seminal work on several

[1] A short biography of Kolmogorov can be found in P.M.B. Vitanyi, CWI Quarterly **1**, page+3 (1988), (http://www.cwi.nl/∼paulv/KOLMOGOROV.BIOGRAPHY.html); a detailed presentation of the many facets of his scientific activities is contained in *Kolmogorov in Perspective* (History of Mathematics, Vol. 20, American Mathematical Society, 2000).

[2] V.M. Tikhomirov and A.N. Shiryayev (editors): "Selected works of A.N. Kolmogorov",, Vol.1, 2 and 3, Kluwer Academic Publishers, Dordrecht, Boston London (1991)

modern research fields in science, namely chaos, complexity, turbulence, mathematical description of biological and chemical phenomena (e.g. reaction diffusion processes and ecological communities).

This book is subdivided into four parts: chaos and dynamical systems (Part I), algorithmic complexity and information theory (Part II), turbulence (Part III) and applications of probability theory (Part IV). A major effort has been devoted to point out the importance of Kolmogorov's contribution in a modern perspective. The use of mathematical formulae is unavoidable for illustrating crucial aspects. At least part of them should be accessible also to readers without a specific mathematical background.

The issues discussed in the first part concern quasi–integrability and chaotic behaviour in Hamiltonian systems. Kolmogorov's work, together with the important contributions by V.I. Arnol'd and J. Moser, yielded the celebrated *KAM theorem*. These pioneering papers have inspired many analytical and computational studies applied to the foundations of statistical mechanics, celestial mechanics and plasma physics. An original and fruitful aspect of his approach to deterministic chaos came from the appreciation of the theoretical relevance of Shannon's information theory. This led to the introduction of what is nowadays called "Kolmogorov–Sinai entropy". This quantity measures the amount of information generated by chaotic dynamics.

Moreover, Kolmogorov's complexity theory, which is at the basis of modern *algorithmic information theory*, introduces a conceptually clear and well defined notion of randomness, dealing with the amount of information contained in individual objects. These fundamental achievements crucially contributed to the understanding of the deep relations among the basic concepts at the heart of chaos, information theory and "complexity". Nonetheless, it is also worth mentioning the astonishingly wide range of applications, from linguistic to biology, of Kolmogorov's complexity. These issues are discussed in the second part.

The third part is devoted to turbulence and reaction-diffusion systems. With great physical intuition, in two short papers of 1941 Kolmogorov determined the scaling laws of turbulent fluids at small scale. His theory (usually called K41) was able to provide a solid basis to some ideas of L.F. Richardson and G.I. Taylor that had never been brought before to a proper mathematical formalization. We can say that still K41 stays among the most important contributions in the longstanding history of the theory of turbulence. The second crucial contribution to turbulence by Kolmogorov (known as *K62 theory*) originated with experimental findings at the Moscow Institute of Atmospheric Physics, created by Kolmogorov and Obukhov. K62 was the starting point of many studies on the small scale structure of fully developed turbulence, i.e. fractal and multifractal models. Other fascinating problems from different branches of science, like "birth and death" processes and genetics, raised Kolmogorov's curiosity. With N.S. Piscounov and I.V. Petrovsky, he proposed a mathematical model for describing the spreading of an advantageous gene – a problem that was also considered independently by R.A. Fisher. Most of the

modern studies ranging from spreading of epidemics to chemical reactions in stirred media and combustion processes can be traced back to his work.

In the last part of this book some recent developments and applications of the theory of probability are presented. One issue inspired by K62 is the application of "wild" stochastic processes (characterized by "fat tails" and intermittent behaviour), to the study of the statistical properties of financial time series. In fact, in most cases the classical central limit theorem cannot be applied and one must consider stable distributions. The very existence of such processes opens questions of primary importance for renormalization group theory, phase transitions and, more generally, for scale invariant phenomena, like in K41.

We are indebted with the authors, from France, Germany, Italy, Spain, and Russia, who contributed to this book, that was commissioned with a very tight deadline. We were sincerely impressed by their prompt response, and effective cooperation.

We warmly thank Prof. Ya.G. Sinai, who agreed to outline in the Preface the character of A.N. Kolmogorov.

A particular acknowledgement goes to Dr. Patrizia Castiglione (staff of Belin Editions): this book has been made possible thanks to her enthusiastic interest and professionality.

Florence and Rome, Spring 2003

Roberto Livi and Angelo Vulpiani

Contents

Part IV. Applications of Probability Theory

List of Contributors

Franco Bagnoli
Dipartimento di Energetica
"S. Stecco", Università Firenze,
Via S. Marta, 3
Firenze, Italy, 50139.
franco.bagnoli@unifi.it

Luca Biferale
Dept. of Physics and INFM,
University of Tor Vergata,
Via della Ricerca Scientifica 1,
Rome, Italy, 00133
Luca.Biferale@roma2.infn.it

Franco Bignone
Istituto Nazionale
per la Ricerca sul Cancro, IST
Lr.go Rosanna Benzi 10,
Genova, Italy, 16132
abignone@unige.it

Guido Boffetta
Dept. of General Physics and INFM,
University of Torino,
Via P.Giuria 1,
Torino, Italy, 10125
boffetta@to.infn.it

Jean-Philippe Bouchaud
Service de Physique de l'État Condensé,
Centre d'études de Saclay,
Orme des Merisiers,
Gif-sur-Yvette Cedex, France, 91191
and
Science & Finance, Capital Fund Management,
rue Victor-Hugo 109-111,
Levallois, France, 92532
bouchau@drecam.saclay.cea.fr

Bernard Castaing
Ecole Normale Superieure de Lyon,
46 Allee d'Italie,
Lyon, France, 69364 Lyon Cedex 07
bcastain@ens-lyon.fr

Fabio Cecconi
Università degli Studi di Roma "La Sapienza",
INFM Center for Statistical Mechanics and Complexity,
P.le Aldo Moro 2,
Rome, Italy, 00185
Fabio.Cecconi@roma1.infn.it

Antonio Celani
CNRS, INLN,
1361 Route des Lucioles,
06560 Valbonne, France
celani@inln.cnrs.fr

Alessandra Celletti
Dipartimento di Matematica,
Università di Roma "Tor Vergata",
Via della Ricerca Scientifica,
Roma, Italy, 00133
celletti@mat.uniroma2.it

Massimo Cencini
INFM Center for Statistical Mechanics and Complexity, Dipartimento di Fisica di Roma "La Sapienza",
P.zzle Aldo Moro, 2
Roma, Italy, 00185
Massimo.Cencini@roma1.infn.it

Massimo Falcioni
University of Rome "La Sapienza", Physics Department
and
INFM Center for Statistical Mechanics and Complexity,
P.zzle Aldo Moro, 2
Rome, Italy, 00185
massimo.falcioni@phys.uniroma1.it

Claude Froeschlé
Observatoire de Nice,
BP 229, France, 06304 Nice Cedex 4
Claude.Froeschle@obs-nice.fr

Giovanni Jona-Lasinio
Dipartimento di Fisica, Universitá "La Sapienza" and INFN,
Piazza A. Moro 2,
Roma, Italy, 00185
Gianni.Jona@roma1.infn.it

Elena Lega
Observatoire de Nice,
BP 229, France, 06304 Nice Cedex 4
Elena.Lega@obs-nice.fr

Roberto Livi
Dipartimento di Fisica
via G. Sansone 1
Sesto Fiorentino, Italy, 50019
Roberto.Livi@fi.infn.it

Cristobal Lopez
Instituto Mediterraneo de Estudios Avanzados (IMEDEA) CSIC-UIB,
Campus Universitat Illes Balears
Palma de Mallorca, Spain, 07122
clopez@imedea.uib.es

Vittorio Loreto
University of Rome "La Sapienza", Physics Department and
INFM Center for Statistical Mechanics and Complexity,
P.zzle Aldo Moro, 2
Rome, Italy, 00185
loreto@pil.phys.uniroma1.it

Andrea Mazzino
ISAC-CNR, Lecce Section,
Lecce, Italy, 73100
and
Department of Physics,
Genova Unviersity,
Genova, Italy, 16146
mazzino@fisica.unige.it

Jean-François Muzy
Laboratoire SPE, CNRS UMR 6134,
Université de Corse,
Corte, France, 20250
muzy@univ-corse.fr

Giorgio Parisi
Dipartimento di Fisica, Sezione INFN, SMC and UdRm1 of INFM,
Università di Roma "La Sapienza",
Piazzale Aldo Moro 2,
Rome, Italy, 00185
Giorgio.Parisi@roma1.infn.it

Arkady Pikovsky
Potsdam University,
Potsdam, Germany, 14469
pikovsky@stat.physik.uni-potsdam.de

Antonio Politi
Istituto Nazionale di Ottica Applicata, INOA,
Lr.go Enrico Fermi 6,
Firenze, Italy, 50125
politi@ino.it

Alain Pumir
CNRS, INLN,
1361 Route des Lucioles,
06560 Valbonne, France
Alain.Pumir@inln.cnrs.fr

Stefano Ruffo
Dipartimento di Energetica,
via S. Marta, 3,
Florence, Italy, 50139
ruffo@avanzi.de.unifi.it

Dimitrij Shepelyansky
Lab. de Phys. Quantique, UMR du CNRS 5626, Univ. P.Sabatier,
Toulouse Cedex 4, France, 31062
dima@irsamc.ups-tlse.fr

Davide Vergni
Istituto Applicazioni del Calcolo, IAC-CNR
V.le del Policlinico, 137
Rome, Italy, 00161
Davide.Vergni@roma1.infn.it

Angelo Vulpiani
University of Rome "La Sapienza",
Physics Department and
INFM Center for Statistical Mechanics and Complexity,
P.zzle Aldo Moro, 2
Rome, Italy, 00185
angelo.vulpiani@roma1.infn.it

Michael Zaks
Humboldt University of Berlin,
Berlin, Germany, 12489
zaks@physik.hu-berlin.de

Kolmogorov Pathways from Integrability to Chaos and Beyond

Roberto Livi[1], Stefano Ruffo[2], and Dima Shepelyansky[3]

[1] Dipartimento di Fisica via G. Sansone 1, Sesto Fiorentino, Italy, 50019
Roberto.Livi@fi.infn.it
[2] Dipartimento di Energetica, via S. Marta, 3, Florence, Italy, 50139
ruffo@avanzi.de.unifi.it
[3] Lab. de Phys. Quantique, UMR du CNRS 5626, Univ. P. Sabatier, Toulouse Cedex 4, France, 31062 dima@irsamc.ups-tlse.fr

Abstract. Two limits of Newtonian mechanics were worked out by Kolmogorov. On one side it was shown that in a generic integrable Hamiltonian system, regular quasi-periodic motion persists when a small perturbation is applied. This result, known as Kolmogorov-Arnold-Moser (KAM) theorem, gives mathematical bounds for integrability and perturbations. On the other side it was proven that almost all numbers on the interval between zero and one are uncomputable, have positive Kolmogorov complexity and, therefore, can be considered as random. In the case of nonlinear dynamics with exponential (i.e. Lyapunov) instability this randomnesss, hidden in the initial conditions, rapidly explodes with time, leading to unpredictable chaotic dynamics in a perfectly deterministic system. Fundamental mathematical theorems were obtained in these two limits, but the generic situation corresponds to the intermediate regime between them. This intermediate regime, which still lacks a rigorous description, has been mainly investigated by physicists with the help of theoretical estimates and numerical simulations. In this contribution we outline the main achievements in this area with reference to specific examples of both low-dimensional and high-dimensional dynamical systems. We shall also discuss the successes and limitations of numerical methods and the modern trends in physical applications, including quantum computations.

1 A General Perspective

At the end of the 19$^{\text{th}}$ century H. Poincaré rigorously showed that a generic Hamiltonian system with few degrees of freedom described by Newton's equations is not integrable [1]. It was the first indication that dynamical motion can be much more complicated than simple regular quasi–periodic behavior. This result puzzled the scientific community, because it is difficult to reconcile it with Laplace determinism, which guarantees that the solution of dynamical equations is uniquely determined by the initial conditions. The main developments in this direction came from mathematicians; they were worked out only in the middle of 20$^{\text{th}}$ century by A.N. Kolmogorov and his school. In the limiting case of regular integrable motion they showed that a generic nonlinear pertubation does not destroy integrability. This result is nowadays

R. Livi, S, Ruffo, and D. Shepelyansky, Kolmogorov Pathways from Integrability to Chaos and Beyond, Lect. Notes Phys. **636**, 3–32 (2003)
http://www.springerlink.com/

formulated in the well–known Kolmogorov–Arnold–Moser (KAM) theorem [2]. This theorem states that invariant surfaces in phase space, called tori, are only slightly deformed by the perturbation and the regular nature of the motion is preserved. The rigorous formulation and proof of this outstanding theorem contain technical difficulties that would require the introduction of refined mathematical tools. We cannot enter in such details here. In the next we shall provide the reader a sketch of this subject by a simple physical illustration. More or less at the same time, Kolmogorov analyzed another highly nontrivial limit, in which the dynamics becomes unpredictable, irregular or, as we say nowadays, *chaotic* [3]. This was a conceptual breakthrough, which showed how unexpectedly complicated the solution of simple deterministic equations can be. The origin of chaotic dynamics is actually hidden in the initial conditions. Indeed, according to Kolmogorov and Martin-Löf [3,4], almost all numbers in the interval $[0, 1]$ are uncomputable. This means that the length of the best possible numerical code aiming at computing n digits of such a number increases proportionally to n, so that the number of code lines becomes infinite in the limit of arbitrary precision. For a given n, we can define the number of lines l of the program that is able to generate the bit string. If the limit of the ratio l/n as $n \to \infty$ is positive, then the bit string has positive Kolmogorov complexity. In fact, in real (computer) life we work only with computable numbers, which have zero Kolmogorov complexity and zero–measure on the [0,1] interval. On the other hand, Kolmogorov numbers contain infinite information and their digits have been shown to satisfy all tests on randomness. However, if the motion is stable and regular, then this randomness remains confined in the tails of less significant digits and it has no practical effect on the dynamics. Conversely, there are systems where the dynamics is unstable, so that close trajectories separate exponentially fast in time. In this case the randomness contained in the far digits of the initial conditions becomes relevant, since it extends to the more significant digits, thus determining a chaotic and unpredictable dynamics. Such chaotic motion is robust with respect to generic smooth perturbations [5]. A well known example of such a chaotic dynamics is given by the Arnold "cat" map

$$\begin{aligned} x_{t+1} &= x_t + y_t \mod 1 \\ y_{t+1} &= x_t + 2y_t \mod 1 \ , \end{aligned} \tag{1}$$

where x and y are real numbers in the $[0, 1]$ interval, and the subscript $t = 0, 1, \ldots$ indicates discrete time. The transformation of the cat's image after six iterations is shown in Fig. 1. It clearly shows that the cat is chopped in small pieces, that become more and more homogeneously distributed on the unit square. Rigorous mathematical results for this map ensure that the dynamics is ergodic and mixing [6,7]. Moreover, it belongs to the class of K-systems, which exhibit the K-property, i.e. they have positive Kolmogorov-Sinai entropy [8–10]. The origin of chaotic behavior in this map is related to the exponential instability of the motion, due to which the distance $\delta r(t)$

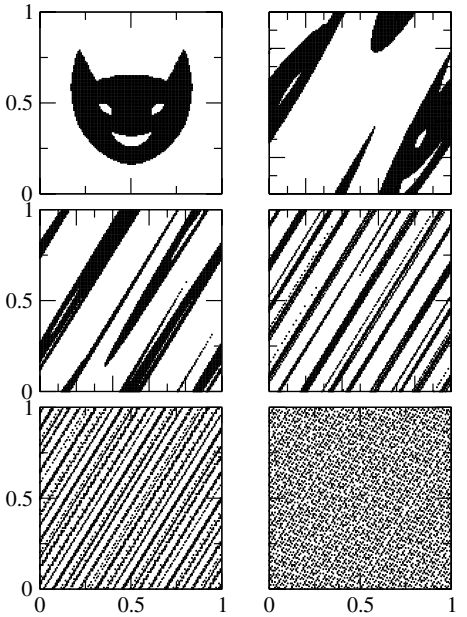

Fig. 1. Arnold "cat" map: six iterations of map (1) from left to right and from top to bottom

between two initially close trajectories grows exponentially with the number of iterations t as

$$\delta r(t) \sim \exp(ht)\ \delta r(0). \tag{2}$$

Here, h is the Kolmogorov-Sinai (KS) entropy (the extension of these concepts to dynamical systems with many degrees of freedom will be discussed in Sect. 5). For map (1) one proves that $h = \ln[(3+\sqrt{5})/2] \approx 0.96$ so that for $\delta r(0) \sim O(10^{-16})$, approximately at $t = 40$, $\delta r(40) \sim O(1)$. Hence, an orbit iterated on a Pentium IV computer in double precision will be completely different from the ideal orbit generated by an infinite string of digits defining the initial conditions with infinite precision. This implies that different computers will simulate different chaotic trajectories even if the initial conditions are the same. The notion of sensitive dependence on initial conditions, expressed in (2), is due to Poincaré [11] and was first emphasized in numerical experiments in the seminal papers by Lorenz [12], Zaslavsky and Chirikov [13] and Henon-Heiles [14]. However, the statistical, i.e. average, properties associated with such a dynamics are robust with respect to small perturbations [5]. It is worth stressing that this rigorous result does not apply to non–analytic perturbations in computer simulations due to round–off errors. Nonetheless, all experiences in numerical simulations of dynamical chaos confirm the stability of statistical properties in this case as well, even if no mathematical rigorous proof exists. Physically, the appearance of statistical properties is related to

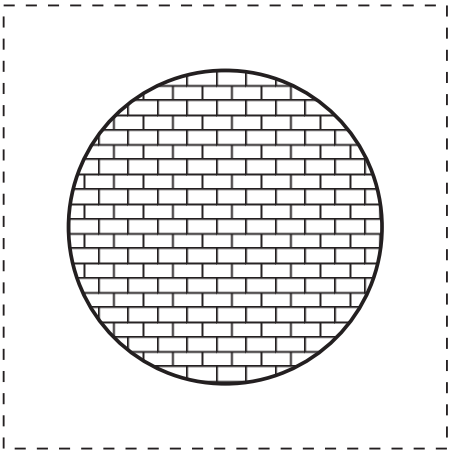

Fig. 2. Sinai billiard: the disc is an elastic scatterer for a point mass particle which freely moves between collisions with the disc. The dashed contour lines indicate periodic boundary conditions: a particle that crosses them on the right (*top*) reappears with the same velocity on the left (*bottom*) (the motion develops topologically into a torus)

the decay in time of correlation functions of the dynamical variables, which for map (1) is exponential.

These results are the cornerstones of the origin of statistical behavior in deterministic motion, even for low–dimensional dynamical systems. However, a K-system (like Arnold cat map (1)) is not generic. Significant progress towards the description of generic physical systems was made by Sinai [15], who proved the K-property for the billiard shown in Fig. 2. It was also proved by Bunimovich [16] that the K-property persists also for "focusing" billiards, like the stadium (see Fig. 3). However, physics happens to be much richer than basic mathematical models. As we will discuss in the following sections, the phase space of generic dynamical systems (including those with many degrees of freedom) contains intricately interlaced chaotic and regular components. The lack of rigorous mathematical results in this regime left a broad possibility for physical approaches, involving analytical estimates and numerical simulations.

2 Two Degrees of Freedom: Chirikov's Standard Map

A generic example of such a chaotic Hamiltonian system with divided phase-space is given by the Chirikov standard map [17,18]:

$$I_{t+1} = I_t + K \sin(\theta_t) \; ; \;\; \theta_{t+1} = \theta_t + I_{t+1} \pmod{2\pi} \; . \tag{3}$$

In this area-preserving map the conjugated variables (I, θ) represent the action I and the phase θ. The subscript t indicates time and takes non-negative

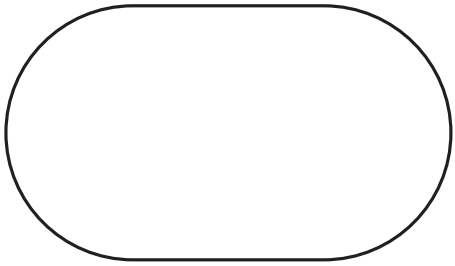

Fig. 3. Bunimovich or "stadium" billiard: the boundary acts as an elastic wall for colliding point mass particles, which otherwise move freely

integer values $t = 0, 1, 2, \ldots$. This mapping can be derived from the motion of a mechanical system made of a planar rotor of inertia M and length l that is periodically kicked (with period τ) with an instantaneous force of strength K/l. Angular momentum I will then vary only at the kick, the variation being given by $\Delta I = (K/l)l \sin\theta$, where θ is the in-plane angle formed by the rotor with a fixed direction when the kick is given. Solving the equations of motion, one obtains map (3) by relating the motion after the kick to the one before (having put $\tau/M = 1$). Since this is a forced system, its energy could increase with time, but this typically happens only if the perturbation parameter K is big enough. Map (3) displays all the standard behaviors of the motion of both one-degree-of-freedom Hamiltonians perturbed by an explicit time-dependence (so-called 1.5 degree of freedom systems) and two-degree-of-freedom Hamiltonians. The extended phase-space has dimension three in the former case and four in the latter. The phase-space of map (3) is topologically the surface of a cylinder, whose axial direction is along I and extends to infinity, and whose orthogonal direction, running along circumferences of unit radius, displays the angle θ. For $K = 0$ the motion is integrable, meaning that all trajectories are explicitly calculable and given by $I_t = I_0, \theta_t = \theta_0 + tI_0 (mod\ \ 2\pi)$. If $I_0/2\pi$ is the rational p/q (with p and q integers), every initial point closes onto itself at the q-th iteration of the map, i.e. it generates a periodic orbit of period q. A special case is $I_0 = 0$, which is a line made of an infinity of fixed points, a very degenerate situation indeed. All irrationals $I_0/(2\pi)$, which densely fill the I axis, generate quasi-periodic orbits: As the map is iterated, the points progressively fill the line $I = const$. Hence, at $K = 0$ the motion is periodic or quasi-periodic. What happens if a small perturbation is switched on, i.e. $K \neq 0$, but small? This is described by two important results: the Poincaré-Birkhoff fixed point theorem (see Chap. 3.2b of [19]) and the Kolmogorov-Arnold-Moser (KAM) theorem [2](see also the contribution by A. Celletti et al. in this volume).

The Poincaré-Birkhoff theorem states that the infinity of periodic orbits issuing from rational $I_0/(2\pi)$ values collapse onto two orbits of period q, one stable (elliptic) and the other unstable (hyperbolic). Around the stable orbits,

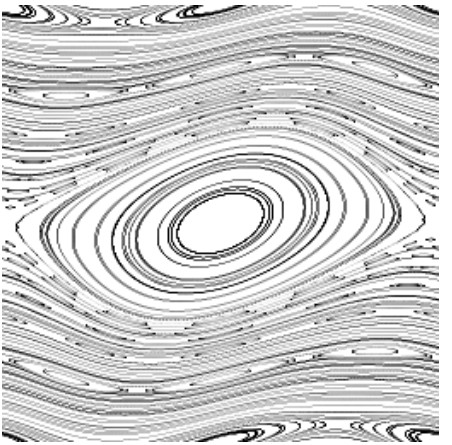

Fig. 4. Phase-space of the Chirikov standard map (3) in the square $(2\pi \times 2\pi)$ for $K = 0.5$

"islands" of stability form, where the motion is quasi-periodic. The biggest of such islands is clearly visible in Fig. 4 and has at the center the elliptic fixed point $(I = 0, \theta = \pi)$ which originates from the degenerate line of fixed points $I = 0$ as soon as $K \neq 0$.

The KAM theorem states that most of the irrational $I_0/2\pi$ initial values generate, at small K, slightly deformed quasi-periodic orbits called KAM-tori. Traces of the integrability of the motion survive the finite perturbations. Since irrationals are dense on a line, this is the most generic situation when K is small. This result has been transformed into a sort of paradigm: slight perturbations of an integrable generic Hamiltonian do not destroy the main features of integrability, which are represented by periodic or quasi-periodic motion. This is also why the KAM result was useful to Chirikov and coworkers to interpret the outcome of the numerical experiment by Fermi, Pasta and Ulam, as we discuss in Sects. 3 and 4.

There is still the complement to the periodic and quasi-periodic KAM motion to be considered! Even at very small K, a tiny but non vanishing fraction of initial conditions performs neither a periodic nor a quasi-periodic motion. This is the motion that has been called "chaotic", because, although deterministic, it has the feature of being sensible to the smallest perturbations of the initial condition [11–14,18].

Let us summarize all of these features by discussing the phase-space structure of map (3), as shown for three different values of K: $K = 0.5$ (Fig. 4), $K = K_g = 0.971635\ldots$ (Fig. 5) and $K = 2.0$ (Fig. 6).

For $K = 0.5$, successive iterates of an initial point θ_0, I_0 trace lines on the plane. The invariant curves $I = \text{const}$, that fill the phase-space when $K = 0$, are only slightly deformed, in agreement with the KAM theorem. A region foliated by quasi-periodic orbits rotating around the fixed point

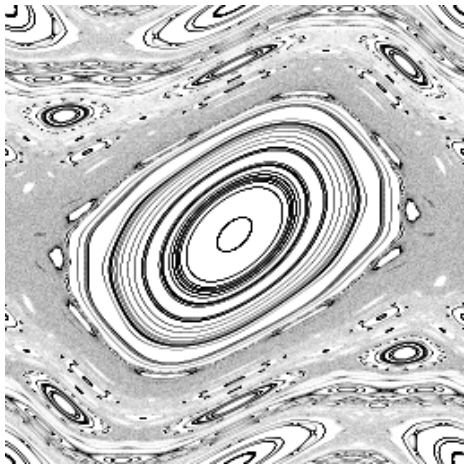

Fig. 5. Same as Fig. 4 for $K = K_g = 0.971635...$

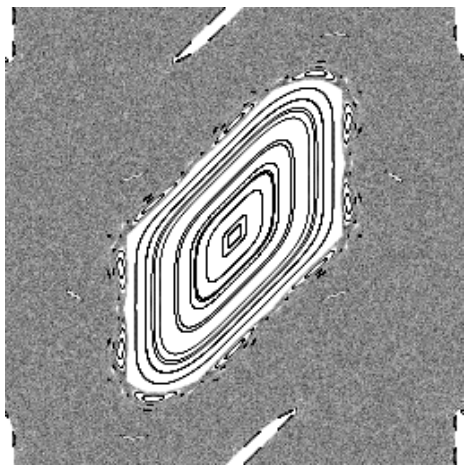

Fig. 6. Same as Fig. 4 for $K = 2$

$(I = 0, \theta = \pi)$ appears; it is called "resonance". Resonances of higher order appear around periodic orbits of longer periods. Their size in phase-space is smaller, but increases with K. Chaos is bounded in very tiny layers. Due to the presence of so many invariant curves, the dynamics in I remains bounded. Physically, it means that although work is done on the rotor, its energy does not increase. A distinctive quantity characterizing a KAM torus is its rotation number, defined as

$$r = \lim_{t\to\infty} \frac{\theta_t - \theta_0}{2\pi t} \ . \tag{4}$$

One can readily see that it equals the time averaged action $< I_t/(2\pi) >_t$ of the orbit, and its number theoretic properties, namely its "irrationality", are central to the dynamical behavior of the orbit. Numerical simulations indicate that for model (3) the most robust KAM torus corresponds to the "golden mean" irrational rotation number $r = r_g = (\sqrt{5}-1)/2$. Let us recall some number theoretic properties. Let a_i be positive integers and denote by

$$\frac{1}{a_1 + \dfrac{1}{a_2 + \cdots}} \equiv [a_1, a_2, \dots] \tag{5}$$

the continued fraction representation of any real number smaller than one. It turns out that r_g contains the minimal positive integers in the continued fraction, $r_g = [1, 1, 1, \dots]$. Indeed, this continued fraction can be resummed by solving the algebraic equation $r_g^{-1} = 1 + r_g$, which clearly has two solutions that correspond to two maximally robust KAM tori. The "golden mean" rotation number r_g corresponds to the "most irrational" number; in some nontrivial sense, it is located as far as possible from rationals. Rational winding numbers correspond to "resonances", and are the major source of perturbation of KAM curves. It is possible to study numerically the stability of periodic orbits with the Fibonacci approximation to the golden mean value $r_n = p_n/q_n \to r_g$ with $q_n = 1, 2, 3, 5, 8, 13 \dots$ and $p_n = q_{n-1}$. This approach has been used by Greene and MacKay and it has allowed them to determine the critical value of the perturbation parameter $K_g = 0.971635...$ at which the last invariant golden curve is destroyed [20,21]. The phase-space of map (3) at $K = K_g$ is shown in Fig. 5. It is characterized by a hierarchical structure of islands of regular quasi-periodic motion centered around periodic orbits with Fibonacci winding number surrounded by a chaotic sea. Such a hierarchy has been fully characterized by MacKay [21] for the Chirikov standard map using renormalization group ideas. A similar study had been conducted by Escande and Doveil [22] for a "paradigm" 1.5-degrees of freedom Hamiltonian describing the motion of a charged particle in two longitudinal waves. Recently, these results have been made rigorous[23], by implementing methods very close to the Wilson renormalization group [24].

For $K > K_g$ the last KAM curve is destroyed and unbounded diffusion in I takes place. With the increase of K, the size of stable islands decreases (see Fig. 6) and for $K \gg 1$, the measure of integrable components becomes very small. In this regime of strong chaos the values of the phases between different map iterations become uncorrelated and the distribution function $f(I)$ of trajectories in I can be approximately described by a Fokker-Planck equation

$$\frac{\partial f}{\partial t} = \frac{D}{2}\frac{\partial^2 f}{\partial I^2} \ , \tag{6}$$

where $D =< (I_{t+1} - I_t)^2 >_t$ is the diffusion constant. For $K \gg 1$, $D \approx K^2/2$ (so-called quasi-linear theory). Thus, due to chaos, deterministic motion can

be described by a statistical diffusive process. As a result, the average square action grows linearly with the number of iterations $< I_t^2 >= I_0^2 + Dt$ for large t.

From the analytical viewpoint the onset of chaos described above has been first obtained by Chirikov on the basis of the resonance-overlap criterion [25]. Let us come back to the representation of the Chirikov standard map in terms of the equations of motion of the Hamiltonian of the kicked rotor

$$H(I,\theta,t) = I^2/2 + K\cos\theta \sum_m \delta(t-m) = I^2/2 + K\sum_m \cos(\theta - 2\pi m t) \ , \tag{7}$$

where $\delta(t)$ is the Dirac δ-function and the sum represents the action of the periodic kicks. The expansion of the periodic δ-function in Fourier series leads to the second expression for the Hamiltonian (7), where the sum runs over all positive/negative integers m. This second form of the Hamiltonian clearly shows the importance of resonances, where the derivative of the phase θ is equal to the external driving frequency $\dot{\theta} = I_m = 2\pi m$. Assuming that the perturbation is weak ($K \ll 1$), we obtain that, in the vicinity of the resonant value of the action, the dynamics is approximately described by the Hamiltonian of a pendulum $H_p = (I - I_m)^2/2 + K\cos\phi$ where $\phi = \theta - 2\pi m t$ is the resonant phase (with respect to the usual pendulum, this one has gravity pointing upward). Indeed, in the first approximation, all non-resonant terms can be averaged out so that the slow motion in the vicinity of I_m becomes similar to the dynamics of a pendulum, given by the term with $m = 0$. The pendulum has two qualitatively different types of motion: phase rotations for an energy $H_p > K$ and phase oscillations for an energy $H_p < K$. In the phase-space (I, θ) these two motions are separated from each other by the separatrix curve $I - I_m = \pm 2\sqrt{K}\sin(\phi/2)$ which at $H_p = K$ starts from the unstable equilibrium point at $\phi = 0$. Thus, the size of the separatrix is, $\Delta\omega_r = \Delta I = 4\sqrt{K}$, while the distance between the resonances $\dot{\phi} = \Omega_m = 2\pi m$ is $\Omega_d = \Omega_{m+1} - \Omega_m = 2\pi$. Two close unperturbed nonlinear resonances overlap when the size of the resonance becomes larger than the distance between them, $\Delta\omega_r > \Omega_d$. Above this resonance-overlap border, a trajectory can move from one resonance to another and the motion becomes chaotic on large scale (as we have commented above, chaos is present even for the smaller K values, but it is restricted to thin layers). In the case of the map (3) this simple criterion gives the critical parameter $K_c = \pi^2/4 \approx 2.5$,larger than the real value $K_g = 0.971635...$ determined by the Greene method. In fact, this simple criterion does not take into account the effects of secondary order resonances and of the finite size of chaotic layers appearing around the separatrix. Considering both effects reduces the border approximately by a factor 2.5 [18]. Thus, in the final form, the Chirikov resonance-overlap criterion can be written as

$$K_c \approx 2.5(\Delta\omega_r/\Omega_d)^2 > 1 \ . \tag{8}$$

Invented by Chirikov in 1959, this physical criterion remains the main analytical tool for determining the chaos border in deterministic Hamiltonian systems. When Chirikov presented his criterion to Kolmogorov, the latter said: "one should be a very brave young man to claim such things!". Indeed, a mathematical proof of the criterion is still lacking and there are even known counterexamples of nonlinear systems with a hidden symmetry, such as the Toda lattice (see Chap. 1.3c of [19]), where the dynamics remains integrable for $K \gg K_c$. However, such systems with a hidden symmetry are quite rare and specific, while for generic Hamiltonian systems the criterion works nicely and determines very well the border for the onset of chaos. An extension and a deep understanding of Chirikov criterion in the renormalization group approach has allowed an improvement and its extensive application to systems with many degrees of freedom [26]. Chirikov resonance overlap criterion finds also applications in such diverse physical systems as particles in magnetic traps [25,18,27], accelerator physics [28], highly excited hydrogen atoms in a microwave field [29], mesoscopic resonance tunneling diodes in a tilted magnetic field [30].

In fact, the Chirikov standard map gives a local description of interacting resonances, assuming that resonance amplitudes slowly change with action I. This is the main reason why this map finds such diverse applications. For example, a modest modification of the kick function $f(\theta) = \sin\theta$ and the dispersion relation $\theta_{t+1} = \theta_t + {I_t}^{-3/2}$ in (3) is sufficient to give a description of the dynamics of the Halley's comet in the solar system [31].

For small perturbations, chaos initially appears in a chaotic layer around the separatrix of a nonlinear resonance. Some basic questions about the effects of nonlinear perturbations in the vicinity of the separatrix were first addressed by Poincaré [1], who estimated the angle of separatrix splitting. The width of the chaotic layer was determined by Chirikov on the basis of the overlap criterion (8) in [17,18]. In fact, for small perturbations, *e.g.* K in map(3), the external frequency ω is much larger than the resonance oscillation frequency ω_0. In such a case, the relative energy w of a trajectory randomly fluctuates inside the chaotic separatrix layer whose width is exponentially small, *e.g.* for the map (3) $|w| < w_s \approx 8\pi\lambda^3 \exp(-\pi\lambda/2)$, where $\lambda = \omega/\omega_0 = 2\pi/\sqrt{K} \gg 1$. Even for $K = 0.5$ the width of the layer is very small and it is hardly visible in Fig. 4 ($w_s \approx 0.015$). It is interesting to note that the dynamics inside the chaotic layer is described by a simple separatrix map, which is similar to the map (3): $y_{t+1} = y_t + \sin x_t, x_{t+1} = x_t - \lambda \ln|y_{t+1}|$ where $y = \lambda w/w_s$ and x is the phase of the rotation [18]. The width of the separatrix layer increases with K as well as the size of primary and secondary resonances. At some critical value K_c the last invariant curve becomes critical. For map (3) $K_c = K_g = 0.971635....$ For $K > K_g$ the golden invariant curve is destroyed and it is replaced by an invariant Cantor set ("cantorus") which allows trajectories to propagate diffusively in action I. Rigorous mathematical results prove the existence of the cantori [32–34]. However, in spite of fundamental advances

in ergodic theory [6,7], a rigorous proof of the existence of a finite measure set of chaotic orbits for map (3) is still missing, even for specific values of K.

The absence of diffusion for small perturbations is typical of 1.5 and 2 degrees of freedom systems. For three or more degrees of freedom, resonances are no longer separated by invariant KAM curves and form a connected web that is dense in action space. Hence, chaotic motion along resonances can carry the orbit arbitrarily close to any region of the phase space compatible with energy conservation. This mechanism is called Arnold diffusion, since Arnold [35] first described its existence. Arnold diffusion is present also for negligible perturbations, but its rate becomes vanishingly small. A theoretical calculation of this rate was first performed by Chirikov[18] and later refined by several authors (see chapter 6 of [19] for a review). Beautiful illustrations of the Arnold web have been obtained by Laskar through the use of frequency analysis [36].

While the local structure of divided phase space is now well understood, the statistical properties of the dynamics remain unclear, in spite of the simplicity of these systems. Among the most important statistical characteristics is the decay of the time correlation function $C(\tau)$ in time and the statistics of Poincaré recurrences $P(\tau)$. The latter is defined as $P(\tau) = N_\tau/N$, where N_τ is the number of recurrences in a given region with recurrence time $t > \tau$ and N is the total number of recurrences. According to the Poincaré theorem (for an easy illustration see Chap. 7.1.3 of [37]), an orbit of a Hamiltonian system always returns sufficiently close to its initial position. However, the statistics of these recurrences depends on the dynamics and is different for integrable and chaotic motion. In the case of strong chaos without any stability islands (e.g. the Arnold cat map (1)), the probability $P(\tau)$ decays exponentially with τ. This case is similar to the coin flipping, where the probability to stay head for more than τ flips decays exponentially. The situation turns out to be different for the more general case of the dynamics inside the chaotic component of an area-preserving map with divided phase space. Studies of $P(\tau)$ for such a case showed that, at a large times, recurrences decay with a power law $P(\tau) \propto 1/\tau^p$ with an exponent $p \approx 1.5$ (see [38] and Fig. 7). Investigations of different maps also indicated approximately the same value of p, even if it was remarked that p can vary from map to map, and that the decay of $P(\tau)$ can even oscillate with $\ln \tau$. This result is of general importance. It can also be shown that it determines the correlation function decay $C(\tau)$ via the relation $C(\tau) \propto \tau P(\tau)$. The statistics of $P(\tau)$ is also well suited for numerical simulations, due to the natural property $P(\tau) > 0$ and to its statistical stability. Such a slow decay of Poincaré recurrences is related to the sticking of a trajectory near a critical KAM curve, which restricts the chaotic motion in phase space [38]. Indeed, when approaching the critical curve with the border rotation number r_g , the local diffusion rate D_n goes to zero as $D_n \sim |r_g - r_n|^{\alpha/2} \sim 1/q_n^\alpha$ with $\alpha = 5$, where $r_n = p_n/q_n$ are the rational convergents for r_g as determined by the continued fraction expansion. The theoretical value $\alpha = 5$ follows from a resonant theory of critical

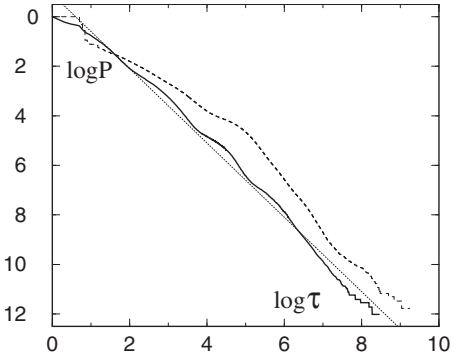

Fig. 7. Poincaré recurrences $P(\tau)$ in the Chirikov standard map (3) at $K = K_g$ (*dashed curve*) and in the separatrix map (see text) with the critical golden boundary curve at $\lambda = 3.1819316$ (*full curve*). The return line is $I = y = 0$. The *dotted straight line* shows the power-law decay $P(\tau) \propto 1/\tau^p$ with $p = 1.5$. [From [38]]

invariant curves [21,38] and is confirmed by numerical measurements of the local diffusion rate in the vicinity of the critical golden curve in the Chirikov standard map [39]. Such a decrease of the diffusion rate near the chaos border would give the exponent $p = 3$, if everything was determined by the local properties of principal resonances p_n/q_n. However, the value $p = 3$ is significantly different from the numerically found value $p \approx 1.5$ (see [38,40] and Fig. 7). At the same time, the similarity of the decay of $P(\tau)$ in two very different maps with critical golden curves is in favor of the universal decay of Poincaré recurrences; it is possible that the expected value $p = 3$ will be reached at very large τ.

3 Many Degrees of Freedom: The Numerical Experiment of Fermi, Pasta, and Ulam

At the beginning of the 50's one of the first digital computers, MANIAC 1, was available at Los Alamos National Laboratories in the US. It had been designed by the mathematician J. von Neumann for supporting investigations in several research fields, where difficult mathematical problems could not be tackled by rigorous proofs[1]. Very soon, Enrico Fermi realized the great potential of this revolutionary computational tool for approaching some basic physical questions, that had remained open for decades. In particular, MANIAC 1 appeared to be suitable for analyzing the many aspects of nonlinear problems, that could not be accessible to standard perturbative methods. Thanks to his deep physical intuition, Fermi pointed out a crucial problem,

[1] It should be mentioned that MANIAC 1 was mainly designed for supporting research in nuclear physics, which yielded the production of the first atomic bomb.

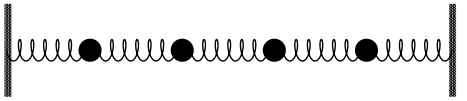

Fig. 8. The FPU chain of oscillators coupled by nonlinear springs

that had been raised already in 1914 by the dutch physicist P. Debye. He had suggested that the finiteness of thermal conductivity in crystals should be due to the nonlinearities inherent in the interaction forces acting among the constituent atoms. Although experimental results seemed to support such a conjecture, a convincing explanation based on a microscopic theory was still lacking fourty years later[2]. In collaboration with the mathematician S. Ulam and the physicist J. Pasta, Fermi proposed to integrate, on the MANIAC 1 the dynamical equations of the simplest mathematical model of an anharmonic crystal: a chain of harmonic oscillators coupled by nonlinear forces (see Fig. 8). In practice, this is described by a classical Hamiltonian of the form

$$H = \sum_{i=1}^{N} \frac{p_i^2}{2m} + \frac{\omega^2}{2}(q_{i+1} - q_i)^2 + \frac{\nu}{n}(q_{i+1} - q_i)^n, \tag{9}$$

where the integer space index i labels the oscillators, whose displacements with respect to equilibrium positions and momenta are q_i and p_i, respectively. The integer exponent $n > 2$ identifies the nonlinear potential, whose strength is determined by the coupling parameter ν. For the sake of simplicity, Fermi, Pasta and Ulam considered the cases $n = 3, 4$, with ν denoted as α and β, respectively (from which the names "α" and "β" models).

The complex interactions among the constituent atoms or molecules of a real solid are reduced to harmonic and nonlinear springs, acting between nearest-neighbor equal–mass particles. Nonlinear springs apply restoring forces proportional to the cubic or quartic power of the elongation of particles from their equilibrium positions[3]. Despite such simplifications, the basic ingredients that one can reasonably conjecture to be responsible for the main physical effect (i.e. the finiteness of thermal conductivity) had been taken into account in the model.

In this form the problem was translated into a program containing an integration algorithm that MANIAC 1 could efficiently compute. It should be stressed that further basic conceptual implications of this numerical experiment were known from the very beginning to Fermi and his collaborators.

[2] Only recently further progress has been made in the understanding of the role of nonlinearity and disorder, together with spatial constraints, in determining transport properties in models of solids and fluids; for a review see [41].

[3] These simplifications can be easily justified by considering that any interaction between atoms in a crystal can be well approximated by such terms, for amplitudes of atomic oscillations much smaller than the interatomic distance: this is the typical situation for real solids at room temperature and pressure.

In fact, they also expected to verify a common belief that had never been amened to a rigorous mathematical proof: In an isolated mechanical system with many degrees of freedom (i.e. made of a large number of atoms or molecules), a generic nonlinear interaction among them should eventually yield equilibrium through "thermalization" of the energy. On the basis of physical intuition, nobody would object to this expectation if the mechanical system starts its evolution from an initial state very close to thermodynamic equilibrium. Nonetheless, the same should also be observed for an initial state where the energy is supplied to a small subset of oscillatory modes of the crystal; nonlinearities should make the energy flow towards all oscillatory modes, until thermal equilibrium is eventually reached. Thermalization corresponds to energy equipartition among all the modes[4]. In physical terms, this can be considered as a formulation of the "ergodic problem". This was introduced by the austrian physicist L. Boltzmann at the end of the 19^{th} century to provide a theoretical explanation of the apparently paradoxical fact, namely that

the time–reversible microscopic dynamics of a gas of hard spheres should naturally evolve on a macroscopic scale towards thermodynamic equilibrium, thus yielding the "irreversible" evolution compatible with the second principle of thermodynamics.

In this perspective, the FPU[5] numerical experiment was intended to test also if and how equilibrium is approached by a relatively large number of nonlinearly coupled oscillators, obeying the classical laws of Newtonian mechanics. Furthermore, the measurement of the time interval needed for approaching the equilibrium state, i.e. the "relaxation time" of the chain of oscillators, would have provided an indirect determination of thermal conductivity[6].

In their numerical experiment FPU considered relatively short chains, up to 64 oscillators[7], with fixed boundary conditions.[8] The energy was initially stored in one of the low, i.e. long–wavelength, oscillatory modes.

---

[4] The "statistical" quality of this statement should be stressed. The concept of energy equipartition implies that the time average of the energy contained in each mode is constant. In fact, fluctuations prevent the possibility that this might exactly occur at any instant of time.

[5] In the following we shall use the usual acronym for Fermi-Pasta-Ulam.

[6] More precisely, according to Boltzmann's kinetic theory, the relaxation time τ_r represents an estimate of the time scale of energy exchanges inside the crystal: Debye's argument predicts that thermal conductivity κ is proportional to the specific heat at constant volume of the crystal, C_v, and inversely proportional to τ_r, in formulae $\kappa \propto C_v/\tau_r$.

[7] Such sizes were already at the limit of computational performances of MANIAC 1, whose execution speed was much smaller than a modern home pc.

[8] The particles at the chain boundaries are constrained to interact with infinite mass walls, see Fig. 8.

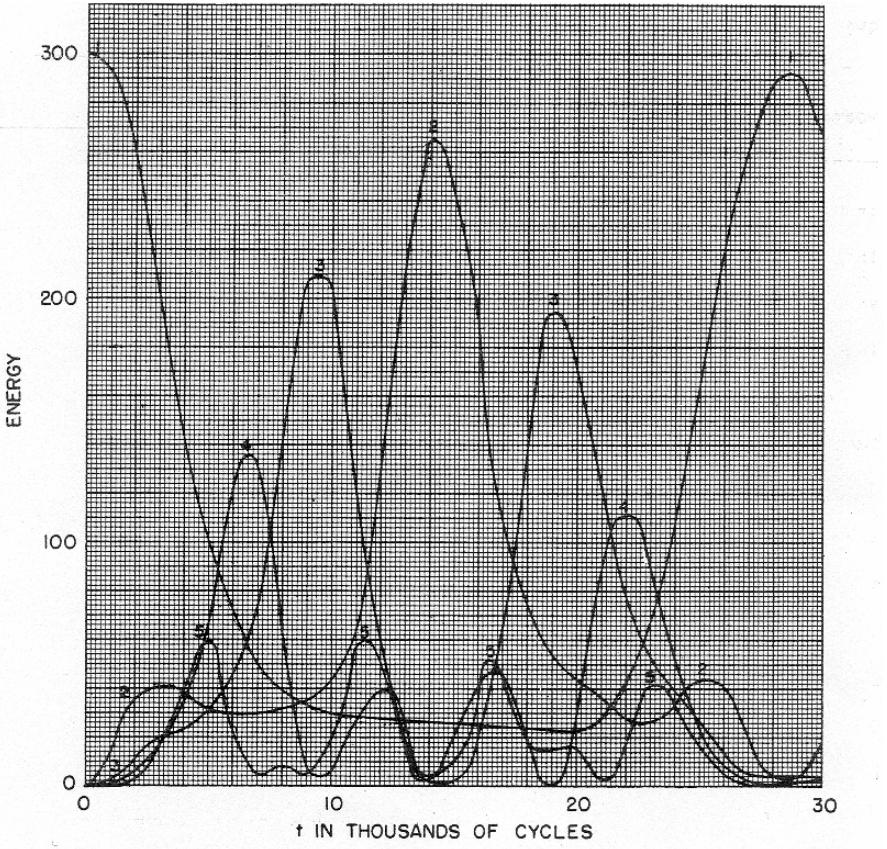

Fig. 1. The quantity plotted is the energy (kinetic plus potential in each of the first five modes). The units for energy are arbitrary. N = 32; $\alpha = 1/4$; $\delta t^2 = 1/8$. The initial form of the string was a single sine wave. The higher modes never exceeded in energy 20 of our units. About 30,000 computation cycles were calculated.

Fig. 9. Energy recurrence in the first 5 Fourier modes in the FPU α model. The figure is taken from [44]

Surprisingly enough, the expected scenario did not appear. Contrary to any intuition the energy did not flow to the higher modes, but was exchanged only among a small number of low modes, before flowing back almost exactly to the initial state, yielding the recurrent behavior shown in Fig. 9.

Even though nonlinearities were at work neither a tendency towards thermalization, nor a mixing rate of the energy could be identified. The dynamics exhibited regular features very close to those of an integrable system.

Almost at the same time as this numerical experiment, A.N. Kolmogorov outlined the first formulation of the KAM theorem (see Sect. 2). FPU

certainly were not aware of his achievement, that indicated that all regular features of the dynamics are kept by integrable hamiltonian systems subject to a small enough perturbation. This could have guided the authors to realize that the nonlinear effects were too small a perturbation of the integrable harmonic chain to prevent regular motion. A deeper understanding of the implications of the FPU experiment on ergodicity and KAM theorem had to wait for more than one decade, for the numerical experiment of Izrailev and Chirikov [42] and Chirikov's overlap criterion [43] (see also Sect. 5).

It should be mentioned that Fermi was quite disappointed by the difficulties in finding a convincing explanation, thus deciding not to publish the results. They were finally published in 1965, one decade after his death, in a volume containing his Collected Papers [44]. The FPU report is probably the most striking example of a crucial achievement which never appeared as a regular paper in a scientific journal, but which, nonetheless, has been a major source of inspiration for future developments in science. Actually, while the understanding of the mechanisms of relaxation to equilibrium and ergodicity mainly concerned the later efforts of european scientists, some american researchers concentrated their attention in trying to interpret the regular motion of the FPU chain in a different way. The first contribution came from a seminal paper by the M.D. Kruskal, a physicist at Princeton, and N.J. Zabusky, a mathematician at Bell Laboratories, in 1965 [45]. This was the starting point for the large physical literature on nonlinear lattice vibrations, that are nowadays called "solitons". In fact, Kruskal and Zabusky were interested in studying the continuum limit of the FPU chain. In particular, Zabusky later conjectured that the dynamical conditions investigated by FPU in their numerical experiment could be explained by an appropriate equation in the continuum limit [46]. This idea is quite natural, since the FPU experiment showed that when a long–wavelength, i.e. low–frequency, mode was initially excited, the energy did not flow towards the small–wavelength, i.e. high–frequency, modes. Since discreteness effects are associated with the latter modes, one can reduce the set of ordinary differential equations describing the chain to an effective partial differential equation that should provide a confident description of long–wavelength excitations. Actually, the continuum limit of the FPU chain was found to correspond to a Korteweg-deVries like equation[9]

$$u_t + \epsilon u^{n-2} u_x + \mu u_{xxx} = 0 \ , \tag{10}$$

where u is the spatial derivative of the displacement field once the right-going wave is selected, and n is the order of the nonlinearity in 9. Exact solutions of such equations can be explicitly found in the form of propagating nonlinear waves. The reader should take into account that the coefficients ϵ

[9] It should be mentioned that performing continuum limits of lattice equations is quite a delicate mathematical problem, as discussed in [47] and also, more recently, in [48]

and μ depend on crucial parameters of the model: the energy of the initial excitation, or, equivalently, the strength of the nonlinear force. For large strength or high energy, the "dispersive" term μu_{xxx} becomes negligible with respect to the nonlinear term $\epsilon u^{n-2} u_x$ and (10) reduces to the first two terms on the left hand side. This reduced partial differential equation has running wave solutions that become unstable after a specific time scale, so-called "shocks". This time scale can be estimated on the basis of the parameters appearing in the equation. Without entering into mathematical details, one can say that the reduced equation describes excitations similar to sea waves, which break their shape because the top of the wave propagates more rapidly than the bottom[10]. This analysis provides a convincing explantion for the FPU experiment. In fact, one can easily conclude that FPU performed their numerical simulations in conditions where the chain was well represented by (10), with a sufficiently large dispersion coefficient μ. Accordingly, the typical instabilities due to discreteness effects might have become manifest only after exceedingly long times, eventually yielding destruction of the regular motion. Moreover, this analysis is consistent with the (almost) contemporary findings of the numerical experiment by Izrailev and Chirikov [42], which show that at high energies or high nonlinearities, the regular motion is rapidly lost.

4 Energy Thresholds

An alternative explanation for the localization of the energy in a small portion of long–wavelength Fourier modes in the FPU chain can be obtained using the resonance–overlap criterion discussed in Sect. 2. It is worth pointing out that the same criterion provides a quantitative estimate of the value of the energy density above which regular motion is definitely lost.

In order to illustrate this interesting issue, we have to introduce some simple mathematical tools. Let us first recall that the Hamiltonian of the Fermi-Pasta-Ulam model (9) can be rewritten in linear normal Fourier coordinates (Q_k, P_k) (phonons)

$$H = \frac{1}{2} \sum_k \left(P_k^2 + \omega_k^2 Q_k^2 \right) + \beta V(\mathbf{Q}) \ , \tag{11}$$

where the nonlinear potential $V(\mathbf{Q})$, whose strength is determined by the coupling constant β[11], controls the energy exchange among the normal modes and ω_k is the the k-th phonon frequency (e.g. $\omega_k = 2\sin(\pi k/N)$ for periodic boundary conditions). The harmonic energy of the k-th normal mode is defined as $E_k = (P_k^2 + \omega_k^2 Q_k^2)/2$. If the energy H is small enough the time–averaged phonon energies $\bar{E}_k(T) = T^{-1} \int_0^T E_k(t) dt$ show an extremely

[10] A clear survey on this class of partial differential equations can be found in [50], Sects. 7 and 8. See also [49]

[11] We restrict ourselves to the quartic nonlinearity $n = 4$ in (9), hence $\nu \equiv \beta$

slow relaxation towards the equipartition state (defined by $E_k = const$) as T increases. On the contrary, at higher energies, the equipartition state is reached in a relatively short time. The presence of these qualitatively different behaviors when the energy is varied was in fact predicted by Chirikov and Izrailev [42] using the "resonance overlap" criterion. Let us give here just a brief sketch of the application of this criterion to the FPU β model. The corresponding Hamiltonian can be written in action-angle variables and, as an approximation, one can consider just one Fourier mode. In fact, this is justified at the beginning of the evolution, when most of the energy is still kept by the initially excited mode.

$$H = H_0 + \beta H_1 \approx \omega_k J_k + \frac{\beta}{2N} \left(\omega_k J_k\right)^2 \ , \tag{12}$$

where $J_k = \omega_k Q_k^2$ is the action variable. In practice, only the nonlinear self-energy of a mode is considered in this approximation. H_0 and H_1 are the unperturbed (integrable) Hamiltonian and the perturbation, respectively. Indeed $\omega_k J_k \approx H_0 \approx E$ if the energy is initially put in mode k. It is then easy to compute the nonlinear correction to the linear frequency ω_k, giving the renormalized frequency ω_k^r

$$\omega_k^r = \frac{\partial H}{\partial J_k} = \omega_k + \frac{\beta}{N} \omega_k^2 J_k = \omega_k + \Omega_k. \tag{13}$$

When $N \gg k$, then

$$\Omega_k \approx \frac{\beta H_0 k}{N^2}. \tag{14}$$

The "resonance overlap" criterion consists of verifying whether the frequency shift is on the order of the distance between two resonances:

$$\Delta\omega_k = \omega_{k+1} - \omega_k \approx N^{-1} \ , \tag{15}$$

(the last approximation being again valid only when $N \gg k$), i.e.

$$\Omega_k \approx \Delta\omega_k \ . \tag{16}$$

One obtains from this equation an estimate of ϵ_c, the "critical" energy density multiplied by β, above which sizeable chaotic regions develop and a fast diffusion takes place in phase space while favouring relaxation to equipartition. the form of ϵ_c is

$$\epsilon_c = \left(\frac{\beta H_0}{N}\right)_c \approx k^{-1} \ , \tag{17}$$

with $k = O(1) \ll N$. Summarizing, primary resonances are weakly coupled below ϵ_c and this in turn induces a slow relaxation process to equipartition.

Conversely, above ϵ_c, fast relaxation to equipartition is present, due to "primary resonance" overlap.

The presence of an energy threshold in the FPU–model separating different dynamical regimes was first identified numerically by Bocchieri et al. [51]. A numerical confirmation of the predictions of the resonance overlap criterion was obtained by Chirikov and coworkers [52]. Further confirmations came for more refined numerical experiments [53,54], showing that, for sufficiently high energies, regular behaviors disappear, while equipartition among the Fourier modes sets in rapidly. Later on [55], the presence of the energy threshold was characterized in full detail by introducing an appropriate Shannon entropy, which counts the number of effective Fourier modes involved in the dynamics (at equipartition this entropy is maximal). Around ϵ_c, the scaling with energy of the maximal Lyapunov exponent (see Sect. 5) also changes, revealing what has been called the "strong stochasticity threshold" [56]. Below ϵ_c, although primary resonances do not overlap, higher order resonances may, yielding a slower evolution towards equipartition [57,58]. The time scale for such an evolution has been found to be inversely proportional to a power of the energy density [59].

After having illustrated the main developments along the lines suggested by the resonance–overlap criterion, it is worth adding some further comments about the existence of an energy threshold, which separates the regular dynamics observed by FPU at low energies from the highly chaotic dynamical phase observed at higher energies.

In their pioneering contribution, Bocchieri and coworkers [51] were mainly concerned by the implications for ergodic theory of the presence of an energy threshold. In fact, the dynamics at low energies seems to violate ergodicity, although the FPU system is known to be chaotic. This is quite a delicate and widely debated issue for its statistical implications. Actually, one expects that a chaotic dynamical system made of a large number of degrees of freedom should naturally evolve towards equilibrium. We briefly summarize here the state of the art on this problem. The approach to equipartition below and above the energy threshold is just a matter of time scales, that actually turn out to be very different from each other. An analytical estimate of the maximum Lyapunov exponent λ (see Sect. 5) of the FPU problem [60] has pointed out that there is a threshold value, ϵ_T, of the energy density, $\epsilon = \beta H/N$, at which the scaling of λ with ϵ changes drastically:

$$\lambda(\epsilon) \sim \begin{cases} \epsilon^{1/4} & \text{if } \epsilon > \epsilon_T; \\ \epsilon^2 & \text{if } \epsilon < \epsilon_T. \end{cases} \tag{18}$$

This implies that the typical *relaxation time*, i.e. the inverse of λ, may become exceedingly large for very small values of ϵ below ϵ_T. It is worth stressing that this result holds in the thermodynamic limit, indicating that the different relaxation regimes represent a statistically relevant effect. To a high degree of confidence, it is found that ϵ_T in (18) coincides with ϵ_c in (17). A more controversial scenario has been obtained by thoroughly investigating

the relaxation dynamics for specific classes of initial conditions. When a few long–wavelength modes are initially excited, regular motion may persist over times much longer than $1/\lambda$ [57]. On the other hand, numerical simulations and analytic estimates indicate that any threshold effect should vanish in the thermodynamic limit [58,59,61]. An even more complex scenario is obtained when a few short-wavelength modes are excited: solitary wave dynamics is observed, followed by slow relaxation to equipartition [62]. It is worth mentioning that some regular features of the dynamics have been found to persist even at high energies (e.g., see [63]), irrespectively of the initial conditions. While such regularities can still play a crucial role in determining energy transport mechanisms [41], they do not significantly affect the robustness of the statistical properties of the FPU model in equilibrium at high energies. In this regime, the model exhibits highly chaotic dynamics, which can be quantified by the spectrum of characteristic Lyapunov exponents. A general description of these chaoticity indicators and their relation with the concept of "metric entropy", introduced by Kolmogorov, is the subject of the following section.

5 Lyapunov Spectra and Characterization of Chaotic Dynamics

The possibility that unpredictable evolution may emerge from deterministic equations of motion is a relatively recent discovery in science. In fact, a Laplacian view of the laws of mechanics had not taken into account such a possibility: the universality of these laws guaranteed that cosmic order should extend its influence down to human scale. The metaphore of divinity as a "clockmaker" was suggested by the regularity of planetary orbits and by the periodic appearance of celestial phenomena, described by the elegant mathematical language of analytical mechanics. Only at the end of the 19^{th} century did the french mathematician H. Poincaré realize that unpredictability is in order as a manifestation of the dynamical instability typical of mechanical systems described by a sufficiently large number of variables[12]. His studies on the stability of the three–body problem with gravitational interaction led him to introduce the concept of "sensitivity with respect to the initial conditions" (see also the contribution by A. Celletti et al. in this volume). He meant that two trajectories, whose initial conditions were separated by an infinitesimal difference, could yield completely different evolution after a suitable lapse of time. This finding is at the basis of what we nowadays call "deterministic chaos", which has been identified as a generic feature of a host of dynamical models of major interest in science and its applications. Here we do not aim at providing the reader a full account of the fascinating history of deterministic chaos. Many interesting books and articles for specialists and newcomers

[12] In fact, such a number is not that large: three independent dynamical variables are enough to allow for unpredictable evolution.

in science are available (for instance, an introductory survey to the subject can be found in [50,37]). We rather want to focus our attention on the crucial contribution of A.N. Kolmogorov in this field.

In order to fully appreciate Kolmogorov's achievements it is useful to discuss certain concepts, introduced for quantifying deterministic chaos. In a chaotic dynamical system two infinitesimally close trajectories, say at distance $\delta(0)$ at time $t = 0$, evolve in time by amplifying exponentially their distance, i.e. $\delta(t) \sim \delta(0) \exp \lambda t$. The exponential rate of divergence $\lambda > 0$ measures the degree of chaoticity of the dynamics. In an isolated dynamical system described by a finite number of variables, such an exponential increase cannot last forever, due to the finiteness of the available phase space. Nonetheless, Oseledec's multiplicative theorem [64] guarantees that, under quite general conditions, the following limit exists

$$\lambda = \lim_{t\to\infty} \lim_{\delta r(0)\to 0} \frac{1}{t} \ln \frac{\delta r(t)}{\delta r(0)} \ . \tag{19}$$

Accordingly λ can be interpreted as the "average" exponential rate of divergence of nearby trajectories, where the average is made over the portion of phase space accessible to the trajectory (see also (2)). It is worth stressing that this quantity is independent of the choice of the initial conditions, provided they belong to the same chaotic component of the phase space. More generally, in a deterministic system described by N dynamical variables or, as one should say, "degrees–of–freedom", it is possible to define a *spectrum of Lyapunov exponents*, λ_i with $i = 1, \cdots, N$, i.e. one for each degree–of–freedom. Conventionally, the integer i labels the exponents from the highest to the smallest one. The stability of a generic trajectory in a multi–dimensional space is, in principle, subject to the contribution of as many components as there are degrees of freedom. This is quite a difficult concept that requires a rigorous mathematical treatment, to be fully appreciated[13]. Intuitively, one can say that the sum $S_n = \sum_{i=1}^{n} \lambda_i$ measures the average exponential rates of expansion, or contraction, of a volume of geometric dimension n in phase space. Accordingly, $S_1 = \lambda_1 \equiv \lambda$ is equivalent to the definition (19), since a "1–dimensional volume" is a generic trajectory in phase space; $S_2 = \lambda_1 + \lambda_2$ gives the divergence rate of a surface; $S_N = \sum_{i=1}^{N} \lambda_i$ is the average divergence rate of the whole phase space. In dissipative dynamical systems, S_N is negative, so that the phase space volume is subject to a global contraction. Nonetheless, the presence of at least one positive Lyapunov exponent, $\lambda_1 > 0$, is enough for making the evolution chaotic: in this case, the trajectory approaches a *chaotic (strange) attractor*. For Hamiltonian systems, according to Liouville's theorem, any volume in phase space is conserved and $S_N = 0$; moreover, for each $\lambda_i > 0$ there exists $\lambda_{N-i} = -\lambda_i$[14]. In summary,

[13] For this purpose we refer the reader to [65].

[14] For each conserved quantity like energy, momentum etc., there is a pair of conjugated exponents that are zero. Stated differently, each conservation law amounts

chaotic evolution implies that a small region in phase space (for instance, the volume identifying the uncertainity region around an initial condition) is expanded and contracted with exponential rates along different directions in phase space. After a time on the order of $1/\lambda$ the distance between two infinitesimally close initial conditions will take the size of the accessible phase space: accordingly, we have no means of predicting where the image of an initial point will be in phase space, by simply knowing the image of an initially closeby point. An infinite precision in the determination of the initial conditions would be required in order to cope with this task. From a mathematical point of view, the determinism of the equations of motion remains unaffected by a chaotic evolution; from a physical point of view, determinism is lost, since the possibility of "predicting" is guaranteed only in the presence of a stable deterministic evolution. In fact, in contrast with mathematics, physics has to deal with precision and errors: in a chaotic dynamics we cannot control the propagation of an initial, arbitrarily small uncertainty.

At this point the very meaning of physics as a predictive science can become questionable, since chaotic dynamics seems to be present in the great majority of natural phenomena. On the other hand, the impossibility of an exact determination of the trajectories does not exclude the possibility of having statistical knowlodge about a chaotic system. The theory of Statistical Mechanics by Boltzmann is the first example where deterministic dynamical rules were replaced by statistical concepts. Actually, the practical impossibility of following the evolution equations of a large number of particles in a diluted gas interacting by elastic collisions led Boltzmann to encompass the problem by introducing an evolution equation for a distribution function $f(\mathbf{r}, \mathbf{v}, t)$. This function tells us about the probability of finding, at time t, a particle of the gas in a given position $\mathbf{r}$ and with velocity $\mathbf{v}$. This probably depends on some global properties of the gas, like the temperature and the occupied volume, rather than on the fine details of the collision dynamics. Boltzmann showed that the evolution equation for $f(\mathbf{r}, \mathbf{v}, t)$ is irreversible and consistent with the second principle of thermodynamics: entropy tends naturally to increase while approaching the equilibrium state, which corresponds to maximal entropy. The great intuition of A.N. Kolmogorov was that a similar, thermodynamic like, description could be adapted to chaotic dynamics. It is important to point out also the main conceptual difference of Kolmogorov's approach with respect to Boltzmann. There is no need for replacing chaotic equations with something else. The crucial observation is that unpredictable dynamical systems can depend on some global feature, i.e. an internal time, like $1/\lambda$, and on the geometric structure of the phase space

to a geometrical constraint that limits the access of the trajectory to a submanifold of phase space. Integrability can be a consequence of all λ_i's being zero, i.e. there can be as many conservation laws as the number of degrees of freedom. However, it can happen that the system is not necessarily integrable and the rate of divergence is weaker than exponential.

(possibly including different kinds of attractors). As a substitute for thermodynamic entropy, Kolmogorov introduced the concept of *metric entropy*. The conceptual breakthrough is that a mechanical description is replaced by a statistical description in terms of a measure: more precisely, we study the evolution of regions of the phase space rather than single trajectories. On this basis, one can easily notice that the concept of "metric entropy" was taken by Kolmogorov directly from information theory. Let us sketch his approach: some mathematics is necessary even if we shall not enter into the technical details[15]. Consider a set of n possible events, that in an experiment can be observed with probabilities $p_1, p_2, \cdots, p_n$, respectively ($\sum_i p_i = 1$). Information theory attributes the information content $-\ln p_j$ to the observation of the j-th event. Accordingly, the average information content associated with an experiment with n possible outcomes is $H = -\sum_{j=1}^{n} p_j \ln p_j$. As a first step towards extending this definition to chaotic dynamics, Kolmogorov introduced a partition of the phase space A into n disjoint subsets $A_1, A_2, \cdots, A_n$, with $A_i \cap A_j = 0$ if $i \neq j$: finding, at some instant of time, the trajectory in one of these subsets is the "event" for chaotic dynamics. By identifying the probability p_j with the measure $\mu(A_j)$ of the subset A_j, one can define the "entropy" associated with the partition A as

$$H(A) = -\sum_{j=1}^{n} \mu(A_j) \ln \mu(A_j). \tag{20}$$

Let us indicate with the symbol ϕ^{-t} the backward in time evolution operator (or "flux") over a time span $-t$, so that $\phi^{-t}A$ represents the partition generated by ϕ^{-t} from A, by taking the intersection of all the back iterates of each initial subset A_i. After n iterations, the application ϕ^{-t} generates a partition

$$A^{(n)} = A \cap (\phi^{-t}A) \cap (\phi^{-2t}A) \cap \cdots \cap (\phi^{-nt}A) \ , \tag{21}$$

where the symbol $\cap$ also denotes the intersection of two partitions. One can say that the proliferation with n of the elements of the partition (21) provides us with a measure of how fast the dynamics divides the original partition A, making it finer and finer. The main idea of Kolmogorov is to obtain a quantitative measure of the degree of chaoticity, or *mixing*, by the average information produced between two iterations

$$H(A, \phi^{-t}) = \lim_{n\to\infty} [H(A^{(n+1)}) - H(A^{(n)})] \tag{22}$$

Finally, since one aims to obtain an upper estimate of the information produced by the dynamics, the definition of *metric Kolmogorov-Sinai entropy* amounts to

$$h(\phi^{-t}) = \sup_{A} H(A, \phi^{-t}). \tag{23}$$

[15] We refer the reader aiming at a rigorous approach to [66]

This quantity is a dynamical indicator, which depends only on the nature of the dynamics. The *internal time* of the system is then given by $1/h$. Three different situations may then appear: $h = 0$ for regular motion (e.g., periodic dynamics), $h = \infty$ for a fully non–deterministic evolution (e.g., a dynamics subject to the influence of external noise), and $0 < h < \infty$ for a deterministic chaotic system. The russian mathematician Ya. B. Pesin [67] proved a remarkable relation between Kolmogorov's metric entropy and the positive component of the Lyapunov spectrum

$$h = \sum_{j=1}^{m} \lambda_j \quad , \quad \lambda_m > 0 > \lambda_{m+1}. \tag{24}$$

It is now evident that for systems with one degree of freedom, $h = \lambda$. The russian mathematician Ya.G. Sinai was the first to propose a simple dynamical model exhibiting mixing properties [15]. He considered a billiard with convex reflecting walls (see Fig. 2) and he proved that the flux associated with the dynamics of a bouncing ball has positive metric entopy. Later, another russian mathematician L.A. Bunimovich showed that the same result is obtained for the stadium billiard [16], where there is no convexity (see Fig. 3), thus indicating that the presence of mixing requires weaker conditions. These contributions also shed some light on the possibility that metric entropy could be at the basis of a statistical description of more physical models, like a gas of hard spheres (the mathematical model of a diluted gas as introduced by Boltzmann) or the FPU chain discussed in Sect. 3. Nonetheless, we should at least point out that the relation between mixing and statistical measure necessarily has to deal with the introduction of the so-called thermodynamic limit, i.e. the limit in which the number of degrees of freedom goes to infinity. In general, this limit does not commute with the limit $t \to \infty$ in (19) and (22). In other words, the results of the measurement of λ and h may depend on the order according to which these limits are performed. Stimulated by a discussion with D. Ruelle at IHES in Paris in 1984, two of the authors and their colleague A. Politi numerically investigated this problem for the FPU chain and other similar dynamical models. They obtained evidence for the existence of a limit curve for the spectrum of Lyapunov exponents in the thermodynamic limit [68] (see Fig. 10). Further numerical indications for the existence of such a limit for a variety of physical systems have been obtained afterwards, but a rigorous mathematical proof is still lacking, although some attempts in this direction have been made [69–71]. The existence of a Lyapunov spectrum in the thermodynamic limit is also used as an hypothesis in the proof of the Gallavotti-Cohen fluctuation-dissipation relation for forced reversible systems [72].

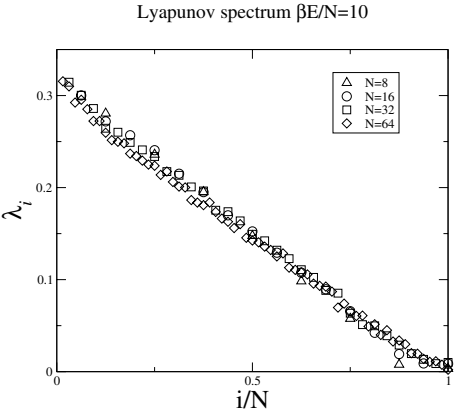

Fig. 10. The spectrum of positive Lyapunov exponents of the FPU *beta* model for different chain lengths, from 8 up to 64 oscillators

6 Quantum Computers and Quantum Chaos

In spite of the fundamental contributions by mathematicians and physicists in the understanding of chaotic dynamics, the role on numerical simulations of chaos can hardly be overestimated. Indeed, computer simulations made possible the investigation of the richness of chaos in all of its details and made the image of chaos familiar to the public. At present, the new technological developments related to quantum information and computation open new horizons to the simulations of chaos.

Indeed, a great deal of attention has been devoted in the last years to the possibility of performing numerical simulations on a quantum computer. As it was already stressed long time ago by Feynman [73], the massive parallelism allowed by quantum mechanics enables us to operate on an exponential number of states using a single quantum transformation. The recent development of quantum information processing has shown that computers designed on the basis of the laws of quantum mechanics can perform some tasks exponentially faster than any known classical computational algorithm (see *e.g.* [74]). The best known example of such a task is the integer factorization algorithm proposed by Shor. The quantum computer can be viewed as a system of qubits (two-level systems), on which "one-qubit" rotations and "two-qubit" transformations allow one to realize any unitary transformation in the exponentially large Hilbert space [74]. At present simple algorithms with up to seven qubits have been realized with nuclear spins in a molecule (NMR) and cold trapped ions.

Quantum computation sheds a new light on chaotic dynamics. Indeed, due to quantum parallelism, a quantum computer can iterate the Liouville density distribution for $O(2^{2n_q})$ classical trajectories in the Arnold cat map (1) in $O(n_q)$ quantum operations (e.g. "one-qubit" rotations and control-NOT

"two-qubit" gates), while a classical simulation requires $O(2^{2n_q})$ operations [75]. For these simulations the phase-space of (1) is discretized in N^2 cells with coordinates (x_i, y_j), $x_i = i/N$ and $y_j = j/N$, $i, j = 0, \ldots, N-1$, $N = 2^{n_q}$. The quantum algorithm simulates this discretized dynamics with the help of 3 quantum registers. The first two registers describe the position x_i and the momentum y_j of N^2 points in the discretized phase-space, where each register contains n_q qubits. The remaining $n_q - 1$ qubits in the third register are used as a work space. An initial classical Liouville distribution can then be represented by a quantum state proportional to $\sum_{i,j} a_{ij}|x_i > |y_j > |0 >$ where the coefficients $a_{i,j}$ are 0 or 1, corresponding to the classical density. The classical dynamics of map (1) is performed with the help of modular additions on the basis of the quantum algorithm described in [76]. The third register holds the curries of the addition and the result is taken modulo N by eliminating the last curry. One map iteration is done in two additions performed in parallel for all classical trajectories in $O(n_q)$ quantum gates.

An example of classical dynamics on a 128×128 lattice is shown in Fig. 11 (left). After $t = 10$ iterations the cat image becomes completely chaotic. Even if the exact dynamics is time reversible the minimal random errors in the last bit (round-off errors) make it effectively irreversible due to dynamical chaos and exponential growth of errors.

Hence, the initial image is not recovered after 10 (or 200) iterations forward/backward (see Fig. 11), even if these minimal errors are done only once at the moment of time inversion. On the contrary the quantum computation remains stable to 1% random errors in the phase of unitary rotation performed by each quantum gate: accordingly, the time reversibility of motion is preserved (see Fig. 11 (right)). In fact the precision of quantum computation remains sufficiently good during a time scale $t_f \propto 1/(\epsilon^2 n_q)$ where ϵ is the error amplitude in quantum gate rotations [75]. The physical origin of this result is related to the fact that the imperfection at each gate rotation transfers probability of the order ϵ^2 from the exact state to all the other states. At the same time the classical error propagates exponentially, due to chaotic deterministic dynamics. This result demonstrates a qualitative difference in the nature of classical and quantum errors. Indeed, quantum perturbation theory guarantees that small quantum errors weakly perturb the evolution. Conversely, from the viewpoint of quantum mechanics (spin flip) a classical error is large even in the last bit and this is the reason why it propagates exponentially in the case of simulations of chaotic dynamics. Thus, despite the common lore that quantum computers are very vulnerable to noise, the study of the Arnold cat map dynamics shows that classical unstable motion, for which classical computers display exponential sensibility to errors, can be simulated accurately with exponential efficiency by a realistic quantum computer [75].

There are also other quantum algorithms which allow the simulations of complex quantum dynamics in a polynomial number of gates for an exponentially large Hilbert space. For example, the quantum dynamics of map

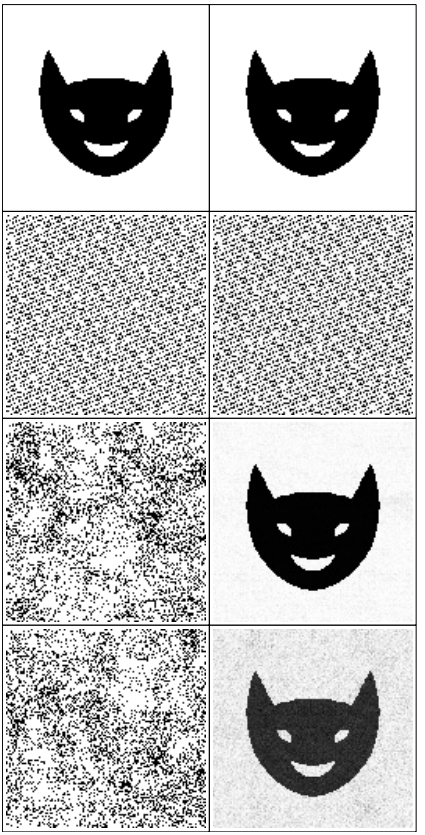

Fig. 11. Dynamics of the Arnold cat map (obtained by interchanging x and y in (1)) simulated on a classical computer (left) and quantum computer (right), on a 128×128 lattice. *Upper row:* initial distribution; *second row:* distributions after $t = 10$ iterations; *third row:* $t = 20$, with time inversion made after 10 iterations; *bottom row:* distributions at $t = 400$, with time inversion made at $t = 200$. Left: the classical error of one cell size ($\epsilon = 1/128$) is done only at the moment of time inversion; right: all quantum gates operate with quantum errors of amplitude $\epsilon = 0.01$; grayness is proportional to the probability $|a_{i,j}|^2$, $n_q = 7$, 20 qubits in total. [From [75]]

(3) can be simulated in $O(n_q^3)$ gates for the vector size $N = 2^{n_q}$ [77]. This opens new links between quantum computation and the field of quantum chaos, which investigates the properties of quantum systems with a chaotic classical limit. The field of quantum chaos has become an independent area of research, closely related to mesoscopic physics, Anderson localization in disordered potential, random matrix theory and periodic orbit quantization in the regime of chaos. due to space constraints, we cannot describe in any

detail this novel and fertile research field. We can only address the reader to reviews and books that can provide her/him with an exhaustive overview on this subject [29,78–81].

References

1. H. Poincaré, *Les methodes nouvelles de la mecanique celeste*, Gauthier-Villars, Paris (1892)
2. A.N. Kolmogorov, Dokl. Akad. Nauk. USSR **98**, 525 (1954); V.I. Arnol'd, Russ. Math. Surv. **18**, 85 (1963); J. Moser, Nachr. Akad. Wiss. Goettingen Math. Phys. **K1**, 1 (1962)
3. A.N. Kolmogorov, Prob. Inform. Trans. **1**, 3 (1965)
4. P. Martin-Löf, Information and Control, **9**, 602 (1966)
5. D.V. Anosov, Soviet Mathematics-Doklady **3**, 1068 (1962)
6. V.I. Arnol'd and A. Avez, *Problemes Ergodiques de la Mecanique Classique*, (Gauthier-Villars, Paris 1967)
7. I.P. Kornfeld, S.V. Fomin, and Y.G. Sinai, *Ergodic theory* (Springer 1982)
8. N.S. Krylov, in *Works on the Foundations of Statistical Physics* (Princeton University Press, Princeton 1979)
9. A.N. Kolmogorov, Dokl. Akad. Nauk USSR **119**, 861 (1958); *ibid.* **124**, 754 (1959)
10. Y.G. Sinai, Dokl. Akad. Sci. USSR **124**, 768 (1959)
11. H. Poincaré, *Science et Méthode* (Flammarion, Paris 1908)
12. E. Lorenz, J. Atmos. Sci, **20**, 130 (1963)
13. G.M. Zaslavsky and B.V. Chirikov: Dokl. Akad. Nauk USSR, **159**, 3081 (1964)
14. M. Henon and C. Heiles, Astron. J., **69**, 73 (1964)
15. Y.G. Sinai, Dokl. Akad. Sci. USSR **153**, 1261 (1963); Russ. Math. Surv. **25**, 137 (1970)
16. L.A. Bunimovich, Comm. Math. Phys. **65**, 295 (1979)
17. B.V. Chirikov, *Research concerning the theory of nonlinear resonance and stochasticity*, Preprint N 267, Institute of Nuclear Physics, Novosibirsk (1969) [Engl. Trans., CERN Trans. 71-40 (1971)]
18. B.V. Chirikov, Phys. Rep., **52**, 263 (1979)
19. A.J. Lichtenberg and M.A. Lieberman, *Regular and chaotic dynamics* (Springer 1992)
20. J.M. Greene, J. Math. Phys. **20**, 1183 (1979)
21. R.S. MacKay, Physica D **7**, 283 (1983)
22. D.F. Escande and F. Doveil, J. Stat. Phys, **26**, 257 (1981)
23. C. Chandre and H.R. Jauslin, Phys. Rep. **365**, 1 (2002)
24. K.G. Wilson, Rev. Mod. Phys. **47**, 773 (1975)
25. B.V. Chirikov, At. Energ. **6**, 630 (1959) [Engl. Transl., J. Nucl. Energy Part C: Plasma Phys. **1** 253 (1960)]
26. Y. Elskens and D.F. Escande, *Microscopic dynamics of plasmas and chaos* (IOP, Bristol and Philadelphia 2003)
27. B.V. Chirikov, Proc. R. Soc. London, Ser. A **413**, 145 (1987)
28. F.M. Izrailev, Physica D **1**, 243 (1980)
29. G. Casati, I. Guarneri, and D.L. Shepelyansky, IEEE Jour. of Quant. Elect. **24**, 1420 (1988)

30. D.L. Shepelyansky and A.D. Stone, Phys. Rev. Lett. **74**, 2098 (1995)
31. B.V. Chirikov and V.V. Vecheslavov, Astron. Astrophys. **221**, 146 (1989)
32. S. Aubry and P.Y. LeDaeron, Physica D, **8**, 381 (1983)
33. J.N. Mather, Ergod. Th. Dynam. Sys., **4**, 301 (1984)
34. R.S. MacKay and I.C. Percival, Comm. Math. Phys., **98**, 469 (1985)
35. V.I. Arnold, Russ. Math. Surveys **18**, 85 (1964)
36. J. Laskar, Physica D **67**, 257 (1993)
37. E. Ott, "Chaos in Dynamical Systems" (Cambridge University Press, Cambridge 1993)
38. B.V. Chirikov and D.L. Shepelyansky, Phys. Rev. Lett. **61**, 1039 (1988); *ibid.* **82**, 528 (1999); *ibid.* **89**, 239402 (2002)
39. S. Ruffo and D.L. Shepelyansky, Phys. Rev. Lett. **76**, 3300 (1996)
40. M. Weiss, L. Hufnagel and R. Ketzmerick, Phys. Rev. Lett. **89**, 239401 (2002)
41. S. Lepri, R. Livi, and A. Politi, Physics Reports **377**, 1 (2003)
42. F.M. Izrailev and B.V. Chirikov, Sov. Phys. Dokl. **11**, 30 (1966)
43. B.V. Chirikov: Comp. Phys. Rev. A **2**, 2013 (1970)
44. E. Fermi, J.R. Pasta and S. Ulam: *Studies of Nonlinear Problems*, Los Alamos Rept. LA-1940 (1955); also appeared in *Collected works of Enrico Fermi*, **2**, 978, University of Chicago Press, Chicago 1965
45. N.J. Zabusky and M.D. Kruskal, Phys. Rev. Lett. **15**, 240 (1965)
46. N.J. Zabusky, in Proceedings of the Symposium on *Nonlinear Partial Differential Equations*, p. 223, W. Ames ed. (Academic Press, NY 1967)
47. C. Cercignani, Rivista del Nuovo Cimento **7**, 429 (1977)
48. P. Rosenau, Phys. Lett. A **311**, 39 (2003)
49. P. Poggi, S. Ruffo and H. Kantz, Phys. Rev. E **52**, 307 (1995)
50. M. Tabor, *Chaos and Integrability in Nonlinear Dynamics* (John Wiley & Sons. Inc., New York 1989)
51. P. Bocchieri, A. Scotti, B. Bearzi and A. Loinger, Phys. Rev. A **2**, 2013 (1970)
52. B. Chirikov, F. Izrailev, and V. Tayurskij, Comp. Phys. Comm. **5**, 11 (1973)
53. M. Casartelli, G. Casati, E. Diana, L. Galgani, and A. Scotti, Theor. Math. Phys. **29**, 205 (1976)
54. G. Benettin and A. Tenenbaum, Phys. Rev. A **28**, 3020 (1983)
55. R. Livi, M. Pettini, S. Ruffo, M. Sparpaglione, and A. Vulpiani, Phys. Rev. A **31** (1985) 1039; R. Livi, M. Pettini, S. Ruffo, and A. Vulpiani, Phys. Rev. A **31** (1985) 2740
56. M. Pettini and M. Landolfi, Phys. Rev. A **41**, 768 (1990)
57. J. De Luca, A.J. Lichtenberg, and M.A. Lieberman, CHAOS **5**, 283 (1995)
58. D. Shepelyansky, Nonlinearity **10**, 1331 (1997)
59. J. De Luca, A.J. Lichtenberg, and S. Ruffo, Phys. Rev. E **60**, 3781 (1999)
60. L. Casetti, R. Livi, and M. Pettini, Phys. Rev. Lett. **74**, 375 (1995)
61. H. Kantz, R. Livi, and S. Ruffo, J. Stat. Phys. **76**, 627 (1994)
62. T. Cretegny, T. Dauxois, S. Ruffo, and A. Torcini, Physica D **121**, 106 (1998); Y. Kosevich and S. Lepri, Phys. Rev. B **61**, 299 (2000); V.V. Mirnov, A.J Lichtenberg, and H Guclu, Physica D **157**, 251 (2001)
63. C. Alabiso, M. Casartelli, and P. Marenzoni, J. Stat. Phys. **79**, 451 (1995)
64. V.I. Oseledec, Trans. Moscow Math. Soc. **19**, 197 (1968)
65. G. Benettin, L. Galgani, A. Giorgilli, and J.M. Strelcyn, Meccanica **15**, 9 and 21 (1980)
66. J.P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985)

67. Y.B. Pesin, Russ. Math. Surveys **32**, 55 (1977)
68. R. Livi, A. Politi, and S. Ruffo, J. Phys. A **19**, 2033, (1986)
69. J.P. Eckmann and E. Wayne, J. Stat. Phys. **50**, 853 (1988)
70. Y.G. Sinai, Int. J. Bifur. Chaos, **6**, 1137 (1996)
71. S. Tanase-Nicola and J. Kurchan, cond-mat/0302448
72. G. Gallavotti and E.G.D. Cohen, J. Stat. Phys. **80**, 931 (1995)
73. R.P. Feynman, Found. Phys. **16**, 507 (1986)
74. M.A. Nielsen and I.L. Chuang, *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge 2000)
75. B. Georgeot and D.L. Shepelyansky, Phys. Rev. Lett. **86**, 5393 (2001); **88**, 219802 (2002)
76. V. Vedral, A. Barenco, and A. Ekert, Phys. Rev. A **54**, 147 (1996)
77. B. Georgeot and D.L. Shepelyansky, Phys. Rev. Lett. **86**, 2890 (2001)
78. M.C. Gutzwiller, *Chaos in Classical and Quantum mechanics* (Springer, New York, 1990)
79. F.M. Izrailev, Phys. Rep. **196**, 299 (1990)
80. M.J. Giannoni, A. Voros, and J. Zinn-Justin, eds., *Chaos and Quantum Physics* (Elsevier Science, Amsterdam, 1990)
81. T. Guhr, A. Müller-Groeling, and H. Weidenmüller, Phys. Rep. **299**, 189 (1998)

From Regular to Chaotic Motions through the Work of Kolmogorov

Alessandra Celletti[1], Claude Froeschlé[2], and Elena Lega[3]

[1] Dipartimento di Matematica, Università di Roma "Tor Vergata", Via della Ricerca Scientifica, 00133 Roma, Italy, celletti@mat.uniroma2.it
[2] Observatoire de Nice, BP 229, 06304 Nice Cedex 4, France, claude@obs-nice.fr
[3] Observatoire de Nice, BP 229, 06304 Nice Cedex 4, France, elena@obs-nice.fr

Abstract. Since ancient times the problem of the stability of the solar system has been investigated by a wide range of scientists, from astronomers to mathematicians, to physicists. The early studies of P.S. Laplace, U. Leverrier, C.E. Delaunay and J.L. Lagrange were based on perturbation theory. Later H. Poincaré proved the non integrability of the three-body problem. It was only in the '50s that A. Kolmogorov provided a theorem which can be used in a constructive way to prove the stability of motions in nearly-integrable systems. A few years later, the pioneer work of Kolmogorov was extended by V.I. Arnold and J. Moser, providing the so-called KAM theory. Though the original estimates of the KAM theorem do not provide rigorous results in agreement with the astronomical predictions, the recent implementations of computer- assisted proofs show that KAM theory can be efficiently used in concrete applications. In particular, the stability of some asteroids (in the context of a simplified three-body problem) has been proved for the realistic values of the parameters, like the Jupiter-Sun mass ratio, the eccentricity and the semimajor axis.

KAM theory was the starting point for a broad investigation of the stability of nearly-integrable Hamiltonian systems. In particular, we review the theorems developed by N.N. Nekhoroshev, which allows to prove the stability on an open set of initial data for exponentially long times.

The numerical simulations performed by M. Henon and C. Heiles filled the gap between theory and experiments, opening a bridge toward the understanding of periodic, quasiperiodic and chaotic motions. In particular, the concept of chaos makes its appearance in a wide range of physical problems. The extent of chaotic motions is provided by the computation of Lyapunov's exponents, which allow to measure the divergence of nearby trajectories. This concept was recently refined to investigate the behaviour of weakly chaotic motions, through the implemention of frequency analysis and Fast Lyapunov Indicators (FLI). We review the applications of such methods to investigate the topology of the phase space. Moreover, the computation of the FLI's allowed to study the transition from Chirikov to Nekhoroshev regime and therefore it provides results about diffusion in Hamiltonian systems.

A. Celletti, C. Froeschlé, and E. Lega, From Regular to Chaotic Motions through the Work of Kolmogorov, Lect. Notes Phys. **636**, 33–59 (2003)
http://www.springerlink.com/

1 Introduction

We know that the solar system has survived catastrophic events since 5 billion years, which correspond to the estimated age of our star. Although the Sun will become a red giant in about 5 billion more years thus destroying the regularity of the planetary motions, the question of the stability of the solar system within this time interval remains open.

This problem has been open since antiquity and it has motivated the development of many studies in mathematical physics. In this context, the so-called *perturbation theory* was developed by Laplace, Leverrier, Delaunay and Lagrange, with the aim of investigating the dynamics of planetary motion. To be more precise, let us start with the simplest model of planetary evolution, the two-body problem, consisting of the motion of one planet under the gravitational influence of the Sun. As is well known, the solution is provided by Kepler's laws which guarantee that (in this approximation) the planet orbits the Sun on an elliptic orbit, the Sun being at one of the foci. The two-body problem is the typical example of an *integrable* system, since its solution can be explicitly provided. In order to get a more realistic model, we should include the gravitational attraction of the other bodies of the solar system. Therefore, let us start by considering the *three-body problem,* where the effect of a third body is included. Despite the many efforts of scientists in the 19^{th} century, a mathematical solution to this problem cannot be found: indeed, H. Poincaré proved the non-integrability of the three-body problem [50]. However, if we consider that the action of the third body (whatever it is: planet, satellite or comet) is small compared to the force exerted by the Sun, we can include the three-body problem in the class of the *nearly-integrable* systems. In our example, the integrable approximation is provided by the two-body problem and the strength of the nonlinearity is relatively small (being proportional to the mass ratio of the primaries, i.e. the third body and the Sun). In this framework, perturbation theory provides useful tools to give (in particular cases) an approximate solution to the equations of motion. However, such theory does not allow definite conclusions to be drawn about the stability of the dynamics, since the mathematical series involved in the construction of the approximate solution are generally divergent.

It was only in the '50s that A.N. Kolmogorov made a major breakthrough in the study of nearly-integrable systems. He proved [34] that for slightly perturbed systems, under quite general assumptions, some regions of the phase space remain regular. We stress that Kolmogorov's theorem can be used in a constructive way to prove the stability of motions. A few years later, the pioneering work of Kolmogorov was extended by V.I. Arnold [1] and J. Moser [46], giving rise to the so-called KAM theory. Though the original estimates of the KAM theorem do not provide rigorous results in agreement with the physical predictions, recent implementations of computer-assisted proofs show that KAM theory can be efficiently used in model problems, yielding results in agreement with the experimental observations. KAM theory was

the starting point for a wider investigation of the stability of nearly-integrable systems, which is briefly reviewed in Sect. 2. Moreover, it opened a new field of research, related to the development of numerical experiments.

In this framework, the simulations performed by M. Hénon and C. Heiles [32] filled the gap between theory and experiments, opening a bridge toward the global study of the dynamical properties of the whole phase space. In particular, the concept of chaos makes its appearance in a wide range of physical problems. The presence of chaotic motions is provided by the computation of the Lyapunov exponents, which allow the divergence of nearby trajectories to be measured. This concept was recently refined to investigate the behaviour of weakly chaotic motions, through the implementation of frequency analysis and Fast Lyapunov Indicators (hereafter, FLI). In Sect. 3 we review some applications of such methods to investigate the topology of the phase space. Moreover, the computation of the FLIs allowed the study of the transition from Chirikov's to Nekhoroshev's regime. It therefore provides results about diffusion in Hamiltonian systems.

2 Stable Motions

2.1 Integrable and Non-integrable Systems

When dealing with real physical problems, it is rather exceptional to find a mathematical solution which describes the dynamics of the system. If the solution cannot be found, the typical approach is to diminish the complexity of the problem by introducing a reasonable set of assumptions. If a solution of the most simplified model can be found, then one can proceed to reintroduce the ingredients which make the model closer to reality. However, such procedure can fail even at the basic step, when the simplest model is considered.

A familiar example of this approach is provided by the investigation of the stability of the solar system. The reality consists of 9 planets orbiting around the Sun, a lot of satellites moving about the planets, thousands of small objects like comets and asteroids. Moreover, one should take into account the oblateness of the solar system bodies, the existence of rings around the major planets, tides, solar wind effects and so on. Without proceeding in the discouraging list of events which contribute to the description of the real world, we start our investigation by reducing the problem to the study of the simplest model arising in Celestial Mechanics: the two-body problem. Consider, for example, the motion of an asteroid about the Sun, neglecting all additional effects like the gravitational influence of the other planets. The solution of this simplified problem dates back to J. Kepler, according to whom the asteroid revolves around the Sun on an elliptic orbit and the motion can be described by elementary formulae. The two-body problem is the archetype of an *integrable* system, for which an explicit solution can be found.

According to our strategy we increase the difficulty of the problem, just adding the effect of a third body, for example Jupiter, the largest planet of the solar system. The three-body problem has been investigated since the 18^{th} century by several astronomers and mathematicians. However, an explicit solution cannot be found. Indeed, H. Poincaré [50] proved that the three-body problem is *non-integrable,* since there are not sufficiently many independent integrals (namely, quantities which are constant during the evolution of the dynamics), which would allow the derivation of mathematical expressions to describe the motion of an asteroid under the Newtonian attraction of the Sun and Jupiter. Since the mass of Jupiter is much smaller than that of the Sun (the mass ratio amounts to about 10^{-3}), the asteroid–Sun–Jupiter problem is close to the asteroid–Sun integrable system whose solution is provided by Kepler's laws. Therefore the effect of Jupiter can be considered as a small perturbation and the three-body problem enters the class of the so-called *nearly-integrable* systems. In this case, even if the complete answer cannot be given, one can proceed by trying to find an *approximate* solution to the equations of motion, which would describe the dynamics with *sufficient* accuracy. The development of this kind of approach is known as *perturbation theory,* which we shall describe in the next section.

In order to make these concepts quantitative, let us consider a mechanical system with n degrees of freedom, described by the Hamiltonian function $H(\underline{p}, \underline{q})$, where $\underline{p} \in \mathbf{R}^n$, $\underline{q} \in \mathbf{R}^n$. If the system is integrable, there exists a canonical change of variables $\mathcal{C} : (\underline{I}, \underline{\varphi}) \in \mathbf{R}^n \times \mathbf{T}^n \to (\underline{p}, \underline{q}) \in \mathbf{R}^{2n}$ (with $\mathbf{T} \equiv \mathbf{R}/2\pi\mathbf{Z}$), such that the transformed Hamiltonian becomes

$$H \circ \mathcal{C}(\underline{I}, \underline{\varphi}) = h(\underline{I}) \ ,$$

having the property that it depends only on the variables $\underline{I}$. The coordinates $(\underline{I}, \underline{\varphi})$ are known in the literature as *action–angle* variables [4]. The integrability of the system can be immediately observed by writing down Hamilton's equations. In fact, denoting by

$$\underline{\omega} = \underline{\omega}(\underline{I}) = \frac{\partial h(\underline{I})}{\partial \underline{I}}$$

the *frequency* or *rotation number* of the system, one has

$$\begin{aligned} \dot{\underline{I}} &= -\frac{\partial h(\underline{I})}{\partial \underline{\varphi}} = \underline{0} \\ \dot{\underline{\varphi}} &= \frac{\partial h(\underline{I})}{\partial \underline{I}} = \underline{\omega}(\underline{I}) \ . \end{aligned}$$

Therefore, the vector $\underline{I}$ is constant during the motion, say $\underline{I} = \underline{I}_0$, while from the second equation we have $\underline{\varphi} = \underline{\omega}(\underline{I}_0)t + \underline{\varphi}_0$, where $(\underline{I}_0, \underline{\varphi}_0)$ denote the initial conditions.

A nearly-integrable system can be conveniently described in terms of a Hamiltonian of the form

$$H(\underline{I},\underline{\varphi}) \;=\; h(\underline{I}) + \varepsilon f(\underline{I},\underline{\varphi}) \;, \tag{1}$$

where $h(\underline{I})$ represents the integrable part, while $\varepsilon f(\underline{I},\underline{\varphi})$ is the perturbation whose size, measured by the parameter ε, is supposed to be small.

The dynamics associated with the three-body problem can be described in terms of a nearly-integrable Hamiltonian system of the previous form. Indeed, referring to the example of the motion of the asteroid under the attraction of the Sun and Jupiter, the integrable part represents the two-body asteroid–Sun interaction (whose solution is given by Kepler's laws), while the perturbation is due to the gravitational influence of Jupiter; in particular, the perturbing parameter ε represents the Jupiter-Sun mass ratio. Though an explicit solution of this problem cannot be found, perturbation theory can provide useful tools to predict the motion with good accuracy.

2.2 Perturbation Theory

Celestial Mechanics stimulated the development of perturbation theory. To give an example, due to unexplained perturbations in the motion of Uranus, the discovery of the planet Neptune was anticipated theoretically by J. Adams and U. Leverrier; their mathematical computations, performed using perturbation theory, predicted the position of Neptune with astonishing accuracy.

In order to explore in more detail the fundamentals of perturbation theory [4,5,26,52], let us start by considering a Hamiltonian function of the form (1). Neglecting the perturbation, we can solve explicitly the equations corresponding to the Hamiltonian $h(\underline{I})$; this provides an ε-approximation of the true solution up to a time of the order of $1/\varepsilon$. The goal of perturbation theory is to introduce a suitable change of coordinates such that we can integrate the equations of motion for a longer time. In particular, we describe a simple algorithm that allows the perturbation to be removed to order ε^2. More precisely, we define a convenient change of variables, say $\mathcal{C} : (\underline{I}',\underline{\varphi}') \to (\underline{I},\underline{\varphi})$ with the property that the transformed Hamiltonian $H'(\underline{I}',\underline{\varphi}')$ has the form

$$H'(\underline{I}',\underline{\varphi}') \;=\; H \circ \mathcal{C}(\underline{I}',\underline{\varphi}') \;\equiv\; h'(\underline{I}') + \varepsilon^2 f'(\underline{I}',\underline{\varphi}') \;, \tag{2}$$

with suitable functions h' and f'. We remark that the perturbing function has now been removed to orders ε^2 and therefore Hamilton's equations associated with (2) can be integrated up to a time of order $1/\varepsilon^2$. Since ε is a small quantity ($0 < \varepsilon << 1$) we get the solution of the motion for a longer time. This idea was developed in the 18^{th} century to compute the dynamics of the bodies of the solar system with more precision than Kepler's laws. The ingredients needed to implement perturbation theory are very elementary;

they are based on the construction of a coordinate transformation, a Taylor series development as a function of the perturbing parameter ε, and a Fourier series expansion to determine explicitly the change of coordinates.

It is rather instructive to derive a first order perturbation theory. We ask the reader to make a little effort in trying to follow the mathematical formulae, so that she/he can control the basics of the computations that allow refined predictions of the solar system dynamics. Let us start by introducing the canonical change of variables $(\underline{I}, \underline{\varphi}) \to (\underline{I}', \underline{\varphi}')$ by means of the system of equations

$$\begin{aligned} \underline{I} &= \underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} \\ \underline{\varphi}' &= \underline{\varphi} + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{I}'} \ , \end{aligned}$$

where the *unknown* function $\Phi(\underline{I}', \underline{\varphi})$ is usually referred to as the *generating function.* We remark that Φ depends on the old angle variables $\underline{\varphi}$ and on the new action variables $\underline{I}'$, as usually defined in classical mechanics textbooks (see, e.g., [4]). In order to give explicit formulae for the transformed Hamiltonian, let us start by splitting the perturbing function $f(\underline{I}, \underline{\varphi})$ into a part $f_0(\underline{I}) = \frac{1}{(2\pi)^n} \int_{\mathbf{T}^n} f(\underline{I}, \underline{\varphi}) d\underline{\varphi}$, depending only on the variables $\underline{I}$ and a remainder function $\tilde{f}(\underline{I}, \underline{\varphi})$ defined by the expression $\tilde{f}(\underline{I}, \underline{\varphi}) \equiv f(\underline{I}, \underline{\varphi}) - f_0(\underline{I})$. Next, we insert the change of variables in the Hamiltonian (1) and we expand in a Taylor series up to second order around $\varepsilon = 0$; more precisely, we obtain

$$\begin{aligned} & h(\underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}}) + \varepsilon f(\underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}}, \varphi) \\ & = h(\underline{I}') + \underline{\omega}(\underline{I}') \cdot \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} + \varepsilon f_0(\underline{I}') + \varepsilon \tilde{f}(\underline{I}', \underline{\varphi}) + O(\varepsilon^2) \ , \end{aligned}$$

where we recall that $\underline{\omega} \equiv \frac{\partial h}{\partial \underline{I}}$. Since we want that the new Hamiltonian is integrable up the second power of ε, we need to kill the first order term in ε by requiring that the function Φ satisfies the equation

$$\underline{\omega}(\underline{I}') \cdot \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} + \tilde{f}(\underline{I}', \underline{\varphi}) = \underline{0} \ . \tag{3}$$

The above equation is well-posed, since its average over the angle variables is zero. If we succeed in finding a solution to (3), then we immediately recognize that the new integrable Hamiltonian can be written as

$$h'(\underline{I}') \ = \ h(\underline{I}') + \varepsilon f_0(\underline{I}') \ ,$$

whose associated Hamilton's equations provide the solution of the motion up to $O(\varepsilon^2)$. Equation (3) can be used to derive an explicit expression for the generating function. More precisely, let us expand the functions Φ and $\tilde{f}$ in Fourier series as

$$\Phi(\underline{I}', \underline{\varphi}) = \sum_{\underline{m} \in \mathbf{Z}^n \setminus \{\underline{0}\}} \hat{\Phi}_{\underline{m}}(\underline{I}') \; e^{i \underline{m} \cdot \underline{\varphi}} \; ,$$
$$\tilde{f}(\underline{I}', \underline{\varphi}) = \sum_{\underline{m} \in \mathbf{Z}^n \setminus \{\underline{0}\}} \hat{f}_{\underline{m}}(\underline{I}') \; e^{i \underline{m} \cdot \underline{\varphi}} \; .$$

Inserting the Fourier expansion inside (3) we obtain

$$i \sum_{\underline{m} \in \mathbf{Z}^n \setminus \{\underline{0}\}} \underline{\omega}(\underline{I}') \cdot \underline{m} \; \hat{\Phi}_{\underline{m}}(\underline{I}') \; e^{i \underline{m} \cdot \underline{\varphi}} = - \sum_{\underline{m} \in \mathbf{Z}^n \setminus \{\underline{0}\}} \hat{f}_{\underline{m}}(\underline{I}') \; e^{i \underline{m} \cdot \underline{\varphi}} \; ,$$

from which we derive

$$\hat{\Phi}_{\underline{m}}(\underline{I}') = - \frac{\hat{f}_{\underline{m}}(\underline{I}')}{i \; \underline{\omega}(\underline{I}') \cdot \underline{m}} \; .$$

Casting together the above formulae, we obtain that the generating function is given by the expression

$$\Phi(\underline{I}', \underline{\varphi}) = i \sum_{\underline{m} \in \mathbf{Z}^n \setminus \{\underline{0}\}} \frac{\hat{f}_{\underline{m}}(\underline{I}')}{\underline{\omega}(\underline{I}') \cdot \underline{m}} \; e^{i \underline{m} \cdot \underline{\varphi}} \; .$$

We strongly remark that the algorithm presented above is constructive in the sense that it provides explicit formulae for the computation of the transformed Hamiltonian, as well as of the canonical change of variables. However, a problem might arise in the above implementation whenever the terms $\underline{\omega}(\underline{I}') \cdot \underline{m}$ are zero, meaning that the frequency vector is rationally dependent. In such a case, the generating function is not defined and the algorithm fails. Indeed, if the vector $\underline{\omega}$ is rationally independent, the divisors are not identically zero; they can however become arbitrarily small and they can lead to the divergence of the series defining the generating function. The *small divisor* problem is the obstacle that prevents the iteration of the algorithm described above to higher orders in ε. If the divisors are very small, the Fourier coefficients defining the generating function become very big and the series may not converge. However, the theory founded by A.N. Kolmogorov in 1954 allows infinitely many iterations of the above procedure and, under suitable assumptions, leads to the proof of the existence of *stable* motions.

2.3 The Kolmogorov–Arnold–Moser Theorem

The success of perturbation theory crucially depends on the character of the frequency vector $\underline{\omega}$. Let us focus on the case $\underline{\omega} = (\omega_1, ..., \omega_n)$ rationally independent, meaning that there does not exist a vanishing linear combination

with rational coefficients (unless all the coefficients are zero). A *quasi-periodic* motion is a solution that can be expressed (as the time t varies) as a function of the form $F = F(\omega_1 t, ..., \omega_n t)$, where $F(\varphi_1, ..., \varphi_n)$ is a multiperiodic function, 2π-periodic in the angles φ_i and $\underline{\omega}$ is rationally independent. In order to illustrate the difference between periodic and quasi-periodic motions, let us consider the dynamics on a two-dimensional torus (i.e., fix $n = 2$), which is graphically represented by a life-belt. A point on the torus is located through two angles φ_1, φ_2 (with $0 \le \varphi_1 \le 2\pi$, $0 \le \varphi_2 \le 2\pi$) measured on the parallel and on the meridian of the torus. If we assume that the evolution is given by $\varphi_1(t) = \omega_1 t + \varphi_1(0)$, $\varphi_2(t) = \omega_2 t + \varphi_2(0)$. The difference between periodic and quasi-periodic motions is given by the rational or irrational character of the quantity $\frac{\omega_1}{\omega_2}$. In particular, if $\frac{\omega_1}{\omega_2}$ is a rational number, say $\frac{\omega_1}{\omega_2} = \frac{p}{q}$ with $p, q \in \mathbf{Z}_+$, then the motion is periodic and it retraces the same steps. On the contrary, if $\frac{\omega_1}{\omega_2}$ is irrational, the evolution never returns to the initial position and it can be shown that the trajectory is everywhere dense on the torus, defining a quasi-periodic motion [2].

As described in Sect. 2.2, perturbation theory concerns the fate of the solutions of an integrable system under a small perturbation that preserves the Hamiltonian structure. Standard theories approached this problem by fixing the initial conditions and investigating the consequent evolution. Instead of looking for the stability of the solution with preassigned initial conditions, Kolmogorov changed the point of view by investigating the stability of the dynamics with a fixed frequency $\underline{\omega}$. At the International Congress of Mathematicians held in Amsterdam in 1954, Kolmogorov announced a result which marked a milestone in the development of the stability analysis of nearly-integrable Hamiltonian systems [34].

Kolmogorov's theorem can be stated as follows: consider a real-analytic Hamiltonian function of the form (1) and fix a rationally independent frequency vector $\underline{\omega}$; under suitable assumptions on the unperturbed Hamiltonian h and on the frequency $\underline{\omega}$, if the strength of the perturbation (namely the parameter ε) is sufficiently small, there exists an invariant torus on which a quasi-periodic motion with frequency $\underline{\omega}$ takes place. Moreover, Kolmogorov's theorem states that the collection of such invariant tori form a set of positive measure in the phase space. In [34] Kolmogorov also gave an outline of the proof: his scheme turned out to be particularly rich and useful. In 1963 V.I. Arnold [1,2] published a detailed alternative proof of Kolmogorov's theorem, while in 1962 J. Moser [46] developed a proof about the existence of invariant curves for smooth enough area-preserving mappings in the plane. Since then, the overall theory is known with the acronym of KAM theory.

In order to illustrate KAM theory in more detail, let us consider a Hamiltonian function of the form (1). As far as the integrable system is concerned (namely setting $\varepsilon = 0$), we have seen that Hamilton's equations lead to quasi-periodic motions where the actions are constant, say $\underline{I} = \underline{I}_0$ with $\underline{\omega}(\underline{I}_0) \equiv \underline{\omega}_0$, while the angles vary linearly with time ($\underline{\varphi}(t) = \underline{\omega}_0 t + \underline{\varphi}_0$). The assumptions

under which KAM theory can be applied are very general. The first hypothesis concerns the unperturbed Hamiltonian $h(\underline{I})$ which we assume to be *non-degenerate,* in the sense that the determinant of the Hessian matrix is different from zero [1]:

$$\det \frac{\partial^2 h(\underline{I})}{\partial \underline{I}^2} \neq 0 \ , \qquad \forall \ \underline{I} \in \mathbf{R}^n \ . \tag{4}$$

The second assumption concerns the frequency vector: in order to avoid the small-divisor problem, one requires that $\underline{\omega}_0$ satisfies a strong irrationality condition, namely that $\underline{\omega}_0$ fulfills the *diophantine* inequalities:

$$|\underline{\omega}_0 \cdot \underline{m}|^{-1} \ \leq \ C|\underline{m}|^{\tau} \ , \qquad \forall \ \underline{m} \in \mathbf{Z}^n \backslash \{\underline{0}\} \ , \tag{5}$$

where C, τ are positive constants. We remark that the above assumption is more stringent than requiring a *non-resonance* condition of the form $|\underline{\omega}_0 \cdot \underline{m}| \neq 0$ for any $\underline{m} \in \mathbf{Z}^n \backslash \{\underline{0}\}$.

Since the diophantine condition (5) plays a key role in the statement of the KAM theorem, we want to make it clearer through a concrete example. Take $n = 2$ and assume that the frequency vector has the form $\underline{\omega}_0 \equiv (\gamma, 1)$ with γ being a real number. Let us introduce the *continued fraction* representation of the number γ as the sequence of positive integers $\{a_0, a_1, a_2, a_3 ...\}$, such that

$$\gamma = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots}}} \ .$$

Using the standard notation, we write $\gamma = [a_0; a_1, a_2, ...]$. A simple computation shows that if γ is a rational number, then its continued fraction is composed by a finite number of terms: there exists a positive integer N such that $\gamma = [a_0; a_1, a_2, ..., a_N]$. On the contrary, if γ is an irrational number, then the continued fraction representation is infinite. Number theory guarantees that the diophantine condition (5) with $\tau = 1$ is satisfied, for example, by the so-called *noble* numbers, i.e. irrational numbers whose continued fraction is definitely 1: there exists a positive integer M such that $\gamma = [a_0; a_1, a_2, ..., a_M, 1, 1, 1, ...]$. Obviously, the most "noble" irrational number is represented by a continued fraction composed by all 1's; this number, known since antiquity, is the *golden ratio* $\frac{\sqrt{5}+1}{2} = [1; 1, 1, 1, ...]$. The golden ratio satisfies condition (5) with the smallest diophantine constant C, being $C = \frac{3+\sqrt{5}}{2}$. Finally, let us mention that each irrational number can be approximated by a sequence of rational numbers $\{\frac{p_k}{q_k}\}_{k \in \mathbf{Z}}$ provided by the successive truncations of the continued fraction, i.e.

$$\frac{p_1}{q_1} \ = \ a_0 + \frac{1}{a_1} \ , \qquad \frac{p_2}{q_2} \ = \ a_0 + \frac{1}{a_1 + \frac{1}{a_2}} \ , \qquad \frac{p_3}{q_3} \ = \ a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3}}} \ ...$$

[1] Notice that the non-degeneracy condition implies that the frequencies satisfy the relation: $\det \frac{\partial \underline{\omega}(\underline{I})}{\partial \underline{I}} \neq 0$.

For the golden ratio, the rational approximants are given by the ratio of the so-called Fibonacci's numbers and the sequence $\{1, \frac{2}{1}, \frac{3}{2}, \frac{5}{3}, \frac{8}{5}, \frac{13}{8}, \frac{21}{13}, ...\}$ converges to $\frac{\sqrt{5}+1}{2}$.

Coming back to the description of the content of KAM theory, let us consider a Hamiltonian system described by (1), assuming that the conditions (4) and (5) are satisfied. We ask ourselves whether there still exists, for the perturbed problem (namely taking $\varepsilon \neq 0$), an invariant torus on which a quasi-periodic motion with frequency $\underline{\omega}_0$ takes place. KAM theory yields a positive answer to this question, provided that the size of the perturbing parameter ε is sufficiently small, say $\varepsilon \leq \varepsilon_{KAM}(\underline{\omega}_0)$. The smallness condition is required in order to guarantee the convergence of the series expansions involved in the proof. We remark that for low-dimensional iso-energetic non-degenerate Hamiltonian systems, the existence of a KAM torus provides a strong stability property. More precisely, if $n = 2$, the phase space has dimension 4 and the constant energy surfaces have dimension 3. Therefore, the 2-dimensional KAM tori separate the constant energy levels providing the stability of the motion, due to the uniqueness of the solutions of ordinary differential equations. This property ceases to be valid as $n \geq 3$; indeed, already for $n = 3$, the dimensions of the phase space and of the constant energy surfaces are, respectively, 6 and 5. Thus, the 3-dimensional KAM tori do not provide a separation of the phase space into invariant regions as it is the case for the 2-dimensional Hamiltonian systems.

Before sketching the proof of the KAM theorem, let us briefly describe the qualitative behaviour of the KAM torus as the perturbing parameter is varied. If ε is very small, the invariant torus lives close to the location of the unperturbed torus, but it becomes more and more displaced and deformed as the perturbing parameter increases. At a certain value of ε, say $\varepsilon = \varepsilon_c(\underline{\omega}_0)$, the invariant torus ceases to be regular and it breaks down for $\varepsilon > \varepsilon_c(\underline{\omega}_0)$. In order to estimate the critical break-down threshold of an invariant torus with frequency $\underline{\omega}_0$, we can use the analytical approach given by the KAM theorem to get a lower bound $\varepsilon_{KAM}(\underline{\omega}_0)$. Alternatively, we can implement a numerical investigation to determine the break-down value $\varepsilon_c(\underline{\omega}_0)$. For this approach, the most used numerical method is due to J. Greene [28], who provided an efficient algorithm to compute the critical threshold. The original work [28] was developed for a simple discrete system known as the *standard mapping,* characterized by a single frequency $\omega_0 \in \mathbf{R}$. Greene's method relies on the conjecture that the break-down of an invariant surface with frequency ω_0 is strictly related to the transition from stability to instability of the periodic orbits with frequency given by the rational approximants $\{\frac{p_k}{q_k}\}_{k \in \mathbf{Z}}$ converging to ω_0. The stability of the periodic orbits can be easily determined by computing the Lyapunov exponents; therefore, Greene's method can be easily adapted to numerical computations and it has been extensively applied to a

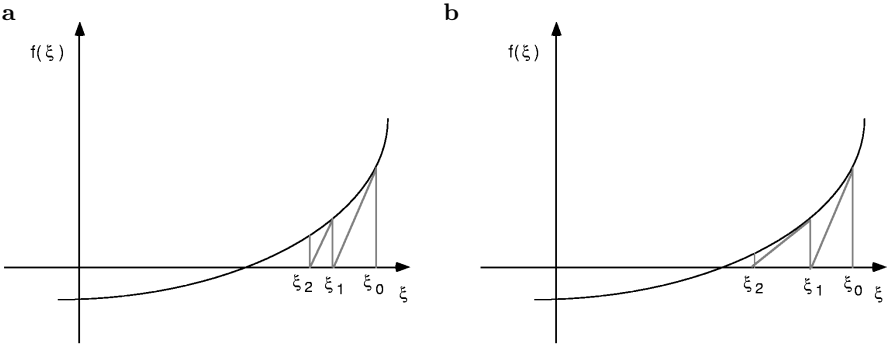

Fig. 1. Newton's method for finding the root of the equation $f(\xi) = 0$. **a** Linear convergence; **b** quadratic convergence

wide variety of continuous and discrete systems. Partial proofs of Greene's technique have been presented in [18] and [44].

Concerning the analytical estimate, the idea behind Kolmogorov's theorem is to overcome the small divisor problem by producing a *superconvergent* sequence of canonical transformations. We have reviewed the standard perturbation theory which allows the transformation of the initial Hamiltonian, say $H_1 = h_1 + \varepsilon f_1$, to a Hamiltonian denoted as $H_2 = h_2 + \varepsilon^2 f_2$, where the perturbation is of order ε^2. One can try to iterate the algorithm to obtain the Hamiltonian $H_3 = h_3 + \varepsilon^3 f_3$ and, at the j-th step, $H_j = h_j + \varepsilon^j f_j$. However, the hindrance due to the convergence of the method cannot in general be overcome. Kolmogorov's idea consists in the development of a *quadratic* method according to which the initial Hamiltonian is transformed to $H_2 = h_2 + \varepsilon^2 f_2$. Iterating, one has $H_3 = h_3 + \varepsilon^4 f_3$ and, at the j-th step, $H_j = h_j + \varepsilon^{2^{(j-1)}} f_j$. The fact that the perturbation decreases with ε faster than linearly allows to control the small divisors appearing in the sequence of transformations.

Newton's method for finding the real root of an equation $f(\xi) = 0$ offers a convenient opportunity to explain the difference between the linear (standard perturbation theory) and quadratic (KAM theory) methods (see Fig. 1). Let us start with an initial approximation ξ_0 and let $e_0 \equiv |\xi - \xi_0|$ be an estimate on the initial error. If we proceed to find the solution by a *linear* convergence, the successive approximation ξ_1 is determined as the intersection of the tangent to the curve at the point $(\xi_0, f(\xi_0))$ with the ξ-axis. Let the derivative of f at ξ_0 be $\eta_0 \equiv f'(\xi_0)$. The next approximation ξ_2 is computed as the abscissa of the line through $(\xi_1, f(\xi_1))$ with slope η_0. Iterating this algorithm, the error at step j is $e_j = |\xi - \xi_j| = O(e_0^{j+1})$, in analogy to standard perturbation theory.

In the quadratic procedure, the successive approximations are computed as the intersection between the ξ-axis and the tangent to the function

computed at the previous step, i.e.

$$\xi_{j+1} = \xi_j - \frac{f(\xi_j)}{f'(\xi_j)} \qquad j = 0, 1, 2...$$

Let us expand $f(\xi)$ around ξ_j; the second order expansion is given by

$$0 = f(\xi) = f(\xi_j) + f'(\xi_j)(\xi - \xi_j) + \frac{f''(\xi_j)}{2!}(\xi - \xi_j)^2 + O(|\xi - \xi_j|^3) \ ,$$

which yields

$$\xi_{j+1} - \xi = \frac{1}{2!} \ \frac{f''(\xi_j)}{f'(\xi_j)}(\xi - \xi_j)^2 \ .$$

Therefore, the error at the step j goes quadratically as

$$e_{j+1} = O(e_j^2) = O(e_0^{2^j}) \ .$$

The proof of the KAM theorem requires a long set of estimates in order to ensure the convergence of the method. We refer the reader to [34,1,46,5,26,52] for the detailed proof.

2.4 The Stability of a Model Associated to the Three-Body Problem

The study of the stability of the three-body problem in Celestial Mechanics has a long tradition. A mathematical approach based on first-order perturbation theory was introduced by Lagrange and Laplace. In particular, they focused on the effect of small perturbations on the dynamics of the planets in order to see whether, after a sufficiently long time, collisions or escape to infinity might take place. In [2], the planetary motion in the framework of the three-body and many-body problems was investigated. Quoting directly V.I. Arnold ([2], p. 125, see also [1]), his result is the following: *"for the majority of the initial conditions under which the instantaneous orbits of the planets are close to circles lying in a single plane, perturbation of the planets on one another produces, in the course of an infinite interval of time, little change on these orbits provided the masses of the planets are sufficiently small"*. To give concrete estimates, M. Hénon [31] applied the original version of Arnold's theorem to the three-body problem, allowing him to prove the existence of invariant tori for values of the perturbing parameter (namely the Jupiter–Sun mass ratio) less or equal than 10^{-333}. This estimate was improved to 10^{-50} by applying Moser's theorem. However, astronomical observations show that the mass of Jupiter compared to that of the Sun is about 10^{-3}. The discrepancy between the KAM estimate and the physical value led to the idea that KAM theory is unable to produce efficient results. Nevertheless, at least for some model problems, the refinement of the KAM estimates and *computer-assisted* series expansions provide results which are

in agreement with the astronomical observations and numerical experiments [9–13]). Let us describe a recent result obtained in [13], concerning the stability of some asteroids in a 3-body framework (see also [12]). Recall that asteroids are thousands minor bodies of the solar system which populate a belt between the orbits of Mars and Jupiter.

Let us introduce a mathematical model, which describes the motion of an asteroid moving under the gravitational influence of the Sun and Jupiter. We assume that the three bodies move on the same plane and that the mass of the asteroid is so small that it does not influence the motion of the primaries. Moreover, we assume that Jupiter orbits the Sun on a circular (Keplerian) orbit. This model is known as the *planar, circular, restricted three-body problem.* It can be conveniently described in terms of suitable action–angle coordinates, known as *Delaunay's variables* [16,53] and the corresponding Hamiltonian function has two degrees of freedom. The unperturbed frequencies (namely the frequencies obtained considering the two-body problems provided by the pairs Sun–asteroid or Sun–Jupiter) are inversely proportional to the periods of revolution about the Sun. We remark that a resonance condition occurs whenever the ratio of the periods of revolution of the asteroid and of Jupiter about the Sun is a rational number; there are many examples in the asteroidal belt which meet this condition, usually denoted as *orbit–orbit resonance.*

With reference to Sect. 2.3, since the Hamiltonian system is two-dimensional, we fix a level of energy (corresponding to the Keplerian motion of the asteroid) in order to have some control on the dynamics. The stability of the minor body can be obtained by proving the existence of invariant surfaces which confine the motion of the asteroid on the given energy level. Refined KAM estimates for the isoenergetic case have been developed in [13]. The proof involves heavy computations of the series expansion of the solution. To this end, a suitable implementation on the computer has been performed. We remark that all numerical errors introduced by the machine are controlled through the so-called *interval arithmetic* technique, which is becoming a widespread tool to prove analytical results with the aid of the computer (we refer the reader to [35,17,33] and references therein for details about computer-assisted proofs). We want to stress that computer-assisted techniques are a complementary part in the proof of mathematical results.

The dynamics of the asteroids Iris, Victoria and Renzia have been considered; the application of the (isoenergetic, computer-assisted) KAM theory provides their stability for values of the perturbing parameter less than or equal to 10^{-3}, giving a result in full agreement with the astronomical prediction. As a conclusion, we stress that this result contributes to show that beside solving the stability problem from a purely mathematical point of view, KAM theory turns out to be effective in physical models as well.

3 Unstable Motions

Quasi periodic motions have been investigated in the previous section, where we showed that the essential features are close to the integrable approximation. Within KAM theory, nothing is predicted for the initial conditions which do not satisfy the diophantine inequality (5). In order to have a complete representation of the phase space, we intend to explore the portion in which a resonance condition of the form $\sum_{i=1}^{n} k_i\omega_i = 0$ (with some integers $k_1, \ldots, k_n \in \mathbf{Z}\backslash\{0\}$) is satisfied in the integrable approximation. We define the *Arnold's web* as a suitable neighborhood of resonant orbits, whose size depends on the perturbing parameter and on the order of the resonance, i.e. $\sum_{i=1}^{n} |k_i|$. The trajectories in the Arnold's web can exhibit chaotic features.

The structure of the Arnold's web was clearly explained in [3]; the numerical investigation of its structure was recently performed on mathematical models [37] and on physically interesting systems. We remark that the stability analysis in the frame of KAM theory has touched different fields of physics. To give some examples, we mention the studies on beam-beam interactions [45], asteroids diffusion [48] and galactic models [49]. These works have been based on numerical applications using frequency-map analysis [39]. It is worthwhile to stress that the validity of the numerical investigations is not only explanatory or didactic; indeed, while the KAM theorem describes the persistence of regular orbits, a rigorous proof of the existence of instability or irregularity in the Arnold web is a delicate problem.

In the following sections, we shall review a theoretical result due to Nekhoroshev and we will introduce a numerical tool [25], which allows to detect and to draw the Arnold's web in a very detailed way [21]. Such analysis also allows the study of transition from a stable to a diffusive regime [29]. We conclude by describing a numerical study [41] of the so-called Arnold's diffusion [3].

3.1 Nekhoroshev's Theorem

A breakthrough in the solution of the problem of stability of the Hamiltonian quasi-integrable systems was provided by Nekhoroshev's [47] theorem. Contrary to the KAM theorem, Nekhoroshev's result extends to an open set of initial conditions; the price to pay is that the stability for infinite times is replaced by the stability over *long* times. In particular, the stability time is exponentially long with respect to the inverse of the perturbing parameter. Therefore, the concept of long time stability is often equivalent in physical problems to that of effective stability. To give an example, the infinite time stability of Jupiter has no physical meaning since the solar system will end in some 5 billion years!

Nekhoroshev's theorem applies to quasi-integrable Hamiltonians of the form

$$H(\underline{I}, \underline{\varphi}) = h(\underline{I}) + \varepsilon f(\underline{I}, \underline{\varphi}) \ , \qquad (\underline{I}, \underline{\varphi}) \in B \times \mathbf{T}^n \ , \tag{6}$$

where B is an open subset of $\mathbf{R}^n$ and $\mathbf{T}^n$ is the n-dimensional torus. We assume that h and f are analytic functions and that ε is a small parameter. Moreover, the integrable approximation h is required to satisfy a suitable geometric condition, called *steepness* ([47], see also [30,6]). In the following, we will assume for simplicity that h is a convex function (convex functions satisfy the steepness condition). Under the above hypotheses, the theorem states that, if ε is sufficiently small, any solution $(\underline{I}(t), \underline{\varphi}(t))$ of (6) satisfies the following exponential estimate ([47], see also [7,43,51]):

$$\|\underline{I}(t) - \underline{I}(0)\| \leq I_0\, \varepsilon^a \qquad \text{for} \quad |t| \leq t_0\, e^{(\varepsilon_0/\varepsilon)^b}\ , \tag{7}$$

where $I_0, t_0, \varepsilon_0, a, b$ are suitable positive constants. In the convex case, one can take $a = b = 1/(2n)$ [43,51]. Therefore, for $\varepsilon < \varepsilon_0$, the actions remain close to their initial values up to exponentially long times. Unfortunately, as it often occurs in perturbation theory, the purely analytical estimates of the constants I_0, t_0, ε_0 are quite unrealistic. This remark motivates the numerical study for realistic models of physical interest.

The direct numerical check on the instability of the actions generally fails, since actions are bounded up to an exponentially long time. Therefore, alternative numerical tools have been developed in the last ten years to investigate the problem in an indirect way [36,39,40,15,23].

Actually, the hearth of the Nekhoroshev's theorem is that the long time stability of the actions occurs within a specific structure of the phase space: the core is made up in large part of KAM tori, while the complementary part is structured by the Arnold's web. The numerical experiments are based on checking for such a structure of the phase space.

3.2 Tools for Detecting Chaos and Order

In order to explore the phase space, a careful analysis of a large number of orbits is required. The classical tool for discriminating between chaotic and ordered orbits is the computation of the largest Lyapunov exponent, that we recall as follows. Let us consider the differential equations:

$$\tfrac{d}{dt}\underline{X} \ = \underline{F}(\underline{X}), \quad \underline{X} = (x_1, x_2, ..., x_n) \tag{8}$$

where $\underline{F}$ are suitable regular functions. The Lyapunov exponents are computed by integrating the equations of motion (8), together with the corresponding variational equations:

$$\frac{d\underline{v}}{dt} = (\frac{\partial \underline{F}}{\partial \underline{X}})\, \underline{v}\ ,$$

where $\underline{v}$ is any n-dimensional vector. The largest Lyapunov exponent is defined as

$$\gamma = \lim_{t\to\infty} \ln \frac{||\underline{v}(t)||}{t}\,.$$

If (8) is of Hamiltonian type, the largest Lyapunov exponent is zero for regular motions and it is positive for chaotic orbits; this property has been largely used to discriminate between chaotic and ordered motions. However, among regular motions, the Lyapunov exponent does not distinguish between KAM tori and resonant islands. When computing the Lyapunov exponents, attention is focused on the length of time necessary to get a reliable value of the limit; very little importance has been given to the beginning of the integration, since it was considered a transitory regime, depending mainly on the choice of the initial vector on the tangent manifold.

As already remarked in [25], chaotic, even *weakly* chaotic, and ordered motion can be discriminated by looking at the value of $\log ||\underline{v}(T)||$, where T is relatively small. The authors called the above quantity Fast Lyapunov Indicator (hereafter FLI). It turns out that the FLIs also distinguish regular resonant from non resonant motions (see [29] for an analytic explanation using a Hamiltonian system) although the largest Lyapunov exponent is zero for both cases.

The precise definition of the FLI can be stated as follows: given an initial condition $\underline{X}(0) \in \mathbf{R}^n$ and an initial vector $\underline{v}(0) \in \mathbf{R}^n$, the FLI is provided by the relation [21,29]

$$FLI(\underline{X}(0), \underline{v}(0), T) \equiv \sup_{0<t\leq T} \log ||\underline{v}(t)|| \,, \qquad T \geq 0 \,. \tag{9}$$

For dissipative systems we refer the reader to an important work by Goldhirish et al. [27] concerning a procedure to extrapolate the value of the Lyapunov exponents even on short time data sets. In recent years other numerical tools have been introduced within the Celestial Mechanics community: the frequency map analysis [36,39,37], the sup-map analysis [38,22] and the twist angle [15,23]. The twist angle is well suited for two dimensional area-preserving maps. The frequency map and the sup-map analyses can be applied to a generic system, but they are more expensive in terms of computational time [24]. Moreover, as far as complexity is concerned, other approaches have been proposed in the domains of fluid dynamics and turbulence, for a review paper, the reader can refer to [8]. We remark that the FLI is defined for each kind of orbit, while the frequency is not well defined for chaotic orbits; moreover the FLI provides quantitative information about the measure of chaos.

The Standard Map as a Model Problem

As a model problem, we consider the two-dimensional standard map [19,42,14]:

$$M : \begin{cases} x(n+1) = x(n) + \varepsilon \sin(x(n) + y(n)) & (\text{mod. } 2\pi) \\ y(n+1) = y(n) + x(n) & (\text{mod. } 2\pi) \,. \end{cases} \tag{10}$$

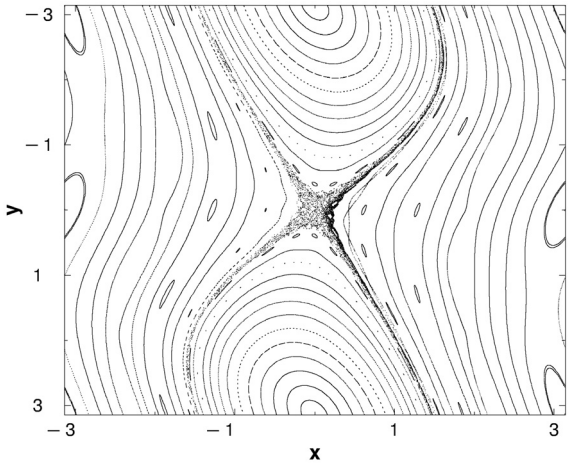

Fig. 2. A set of orbits of the standard map for $\varepsilon = 0.6$

In Fig. 2, we display some orbits of the standard map defined by (10) with $\varepsilon = 0.6$. For $\varepsilon = 0.6$, the majority of the orbits are still invariant tori. Some resonant curves are displayed around elliptic points and a chaotic zone is generated by the existence of the hyperbolic point at the origin.

Figure 3 shows the variation of the FLI (defined in 9) with respect to time for three different kinds of orbits. The upper curve, corresponding to initial conditions $(x, y) = (10^{-4}, 0)$ selected in the previous chaotic zone, shows an exponential increase of the FLI; the upper value of 20 is a threshold that we impose in order to avoid floating overflow. The intermediate curve corresponds to a regular invariant torus with initial conditions $(x, y) = (1, 0)$ and the lowest one describes a resonant trajectory with initial conditions $(x, y) = (0, 1)$.

In [24] the FLI was computed for a set of 1000 initial conditions, regularly spaced on the x-axis in the interval $[0, \pi]$, while $y(0)$ was always set to zero. Figure 4 shows the so-called *FLI-map*, i.e. the value of the FLI after $T = 1000$ iterations against $x(0)$; here, $\underline{v}(0)$ was always taken in the direction of the x-axis. Many orbits appear to have an FLI equal to the logarithm of the integration time, i.e. $\log T = 3$: actually, this value corresponds to invariant tori [24,29].

Values slightly greater than $\log T$ indicate very thin chaotic layers or invariant tori close to very thin chaotic zones. The orbits with an FLI lower than $\log T$ correspond to chains of islands (for an explanation see [29]).

3.3 The Arnold's Web Pictures for a Simple Hamiltonian Model

In this section we give a highly accurate graphical representation of the Arnold's web, obtained by implementing the FLI method. As an example, we

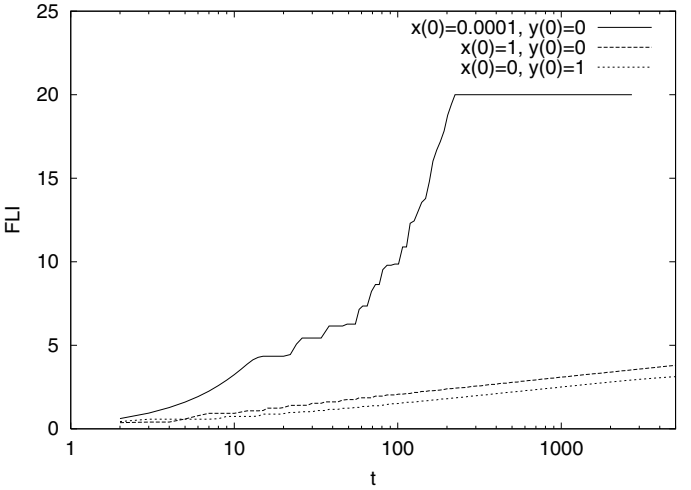

Fig. 3. Variation of the Fast Lyapunov Indicator w.r.t. time for three orbits of the standard map with $\varepsilon = 0.6$. The upper curve corresponds to a chaotic orbit with initial conditions $x(0) = 10^{-4}, y(0) = 0$; the middle one describes a non-resonant orbit with $x(0) = 1, y(0) = 0$, while the lowest one concerns a resonant orbit with $x(0) = 0, y(0) = 1$

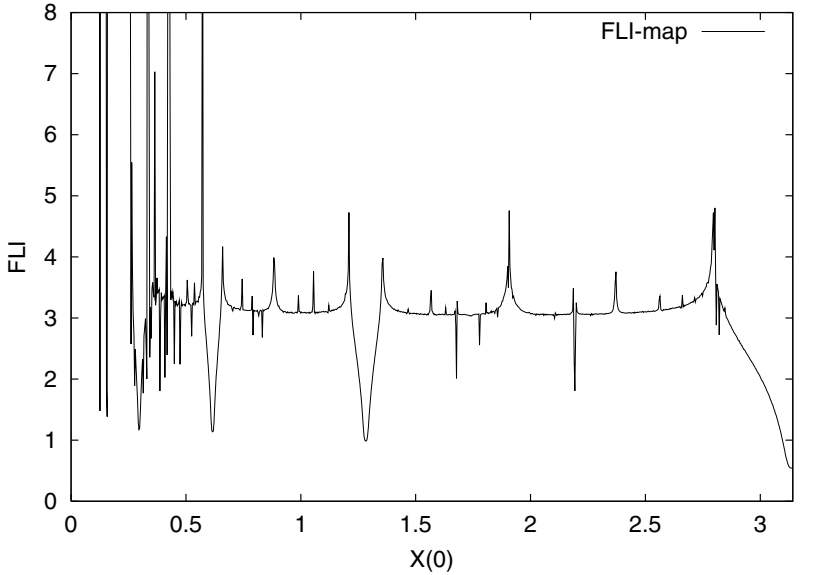

Fig. 4. Variation of the FLI as a function of the initial values $x(0)$ for a set of 1000 orbits regularly spaced in the interval $[0, \pi]$. The initial vector $\underline{v}(0)$ has components $(1, 0)$, i.e. it is almost perpendicular to the corresponding invariant curves

consider the following Hamiltonian function:

$$H_\varepsilon(I_1, I_2, I_3, \varphi_1, \varphi_2, \varphi_3) = \frac{I_1^2}{2} + \frac{I_2^2}{2} + I_3 + \varepsilon \left(\frac{1}{\cos(\varphi_1) + \cos(\varphi_2) + \cos(\varphi_3) + 4} \right),$$

where $I_1, I_2, I_3 \in \mathbf{R}$ and $\varphi_1, \varphi_2, \varphi_3 \in \mathbf{T}$ are canonically conjugated, and ε is a small parameter. The canonical equations of the integrable Hamiltonian H_0 are trivially integrated: I_1, I_2, I_3 stay constant, while the angles $\varphi_1(t) = \varphi_1(0) + I_1 t$, $\varphi_2(t) = \varphi_2(0) + I_2 t$, $\varphi_3(t) = \varphi_3(0) + t$ rotate with constant angular velocity. Therefore, each pair of actions (I_1, I_2) characterizes an invariant torus $\mathbf{T}^3$, on which motions are quasi-periodic with frequencies $\omega_1 = I_1$, $\omega_2 = I_2$, $\omega_3 = 1$. Conversely, for any small ε different from zero, H_ε is not integrable. However, the KAM theory proves the existence of a large volume of invariant tori, embedded in the Arnold's web. Our goal is to numerically determine the structure of the Arnold's web, which can be conveniently represented in the two-dimensional plane (I_1, I_2). Indeed, each point on this plane corresponds to a unique frequency of an unperturbed torus. Moreover, all resonances $k_1\omega_1 + k_2\omega_2 + k_3\omega_3 = 0$ are represented by the straight lines $k_1 I_1 + k_2 I_2 + k_3 = 0$; the set of all resonances is dense on the plane. However, one can expect that irregular orbits surround each resonance line, up to a distance decreasing as $\sqrt{\varepsilon}/|k|^\tau$; we refer to this region as the *resonant zone*. Consequently, the volume of the Arnold's web is of the order of $\sqrt{\varepsilon}$. Let us briefly review the qualitative behaviour of the motions with initial conditions in the Arnold's web. Within resonant zones, both chaotic and regular motions can be observed. Stable periodic orbits are surrounded by islands in which the motion is still quasi-periodic, though the dimension of the torus is strictly smaller than the number of degrees of freedom. Unstable periodic orbits are surrounded by chaotic zones where Nekhoroshev's theorem predicts a diffusion with a velocity exponentially small with respect to $-1/\varepsilon$ in the action space. When increasing the strength of the perturbation, the regular set shrinks until it almost completely disappears. In this case the dynamics is not controlled by Nekhoroshev's theorem anymore. To describe it, we resort to the well-known Chirikov [14] overlapping criterion which allows the resonant chaotic orbits to go from one resonance to the other (eventually giving rise to large-scale diffusion).

From Nekhoroshev to Chirikov

Using the FLI method, we proceed to describe the evolution from a mostly ordered system to a largely chaotic one, i.e. from the Nekhoroshev to the Chirikov regime. To this end, we compute the FLI, using a leap-frog symplectic integrator on a grid of 500×500 initial conditions regularly spaced in the action space (without loss of generality, we set the initial angles $\varphi_1 = \varphi_2 = \varphi_3 = 0$). The initial choice of the tangent vector plays a delicate role. Indeed, it turns

out that resonances which are aligned with the initial tangent vector $(\dot{\varphi}_1, \dot{\varphi}_2)$ are not detected by the FLI method. In order to overcome this problem we have chosen $(\dot{\varphi}_1, \dot{\varphi}_2)$, such that $\dot{\varphi}_1/\dot{\varphi}_2$ is strongly irrational (the other components $\dot{I}_1, \dot{I}_2$ play a minor role and we set them equal to unity). Some results are presented in Fig. 5. In each picture, the initial conditions of I_1 and I_2 are associated to the corresponding FLI value by a different color. The lowest values of the FLI appear in black and they correspond to the resonant islands of the Arnold web; the highest values appear in white and they correspond to chaotic motions, arising at the crossing nodes of resonant lines or arising near the separatrices. The FLIs corresponding to KAM tori have approximately the same value, and therefore the same grey color. Therefore, the resonant lines clearly appear to embed large zones filled by KAM tori (see Fig. 5 (top, left) and the enlargement shown in Fig. 5 (top, right)). Since the perturbation has a full Fourier spectrum, i.e. all harmonics are present at order ε, a high number of resonances is already present for small ε (Fig. 5 (top)); we remark that all resonances should appear just by increasing the integration time. In contrast, in Fig. 5 (middle), which refers to $\varepsilon = 0.01$, the volume of the invariant tori decreases, though the system is still in the Nekhoroshev regime. In these figures, following the Nekhoroshev theorem, the chaotic regions become evident at the crossing of the resonances. In Fig. 5 (bottom), which refers to $\varepsilon = 0.04$, the dynamical regime has completely changed. As outlined before, the majority of invariant tori has disappeared due to resonance overlapping, and a big chaotic connected region has replaced the regularity set.

In conclusion, using a very simple numerical tool, we described the structure of the Arnold's web and its evolution as a function of the perturbing parameter. We observed a transition from an ordered to a chaotic diffusive system, occurring for $0.01 < \varepsilon_0 < 0.04$ (ε_0 being defined by (7)).

Detection of the Diffusion

Following Nekhoroshev's theorem, we expect a very slow diffusion of the actions along the resonances, which is very difficult to detect numerically. The idea of such diffusion was introduced by the pioneering work of Arnold [3], using an ad hoc model. Since then, it is called *Arnold's diffusion*. According to (7), when ε is close to ε_0, i.e. to the transition from the stable regime to the diffusive one, the actions can quickly become unstable. Therefore, in order to measure Arnold's diffusion, we need to start close to ε_0. In the previous section, we found that the transition occurred in the interval $0.01 < \varepsilon_0 < 0.04$ (this interval was confirmed and refined to $0.0300 < \varepsilon_0 < 0.032$ in [29]). We used the FLI pictures to find chaotic initial conditions on a given resonant line, as well as to observe that diffusion occurs along this line.

Figure 6 shows some enlargements of the FLI pictures in the action space, around $I_1 = 0.3$ and $I_2 = 0.14$, for different values of ε. The region between the two white lines corresponds to the resonance associated to $I_1 - 2I_2 = 0$, while the two white lines correspond to its hyperbolic border, where diffusion

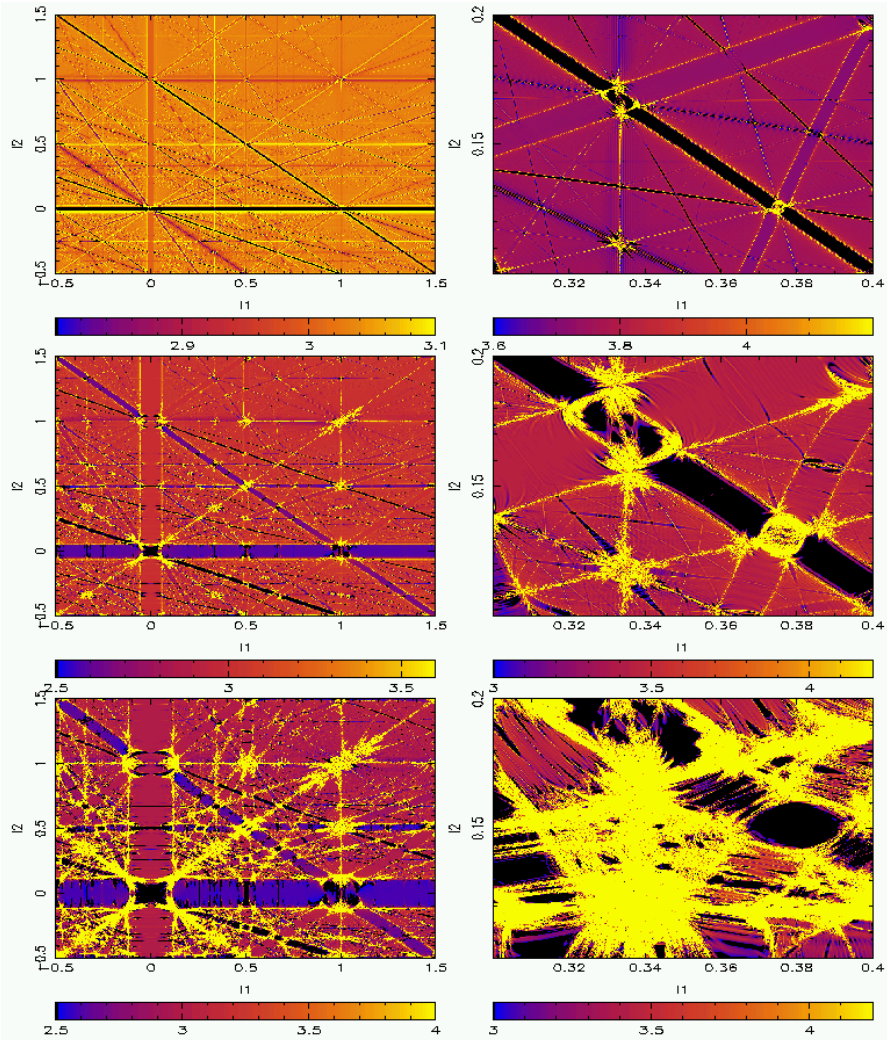

Fig. 5. Evolution of the Arnold web for increasing values of the perturbing parameter ε. The lowest values of the FLI appear in black for the regular resonant islands; the highest values appear in white for chaotic motions arising at the crossing nodes of resonant lines or arising near the separatrix. The FLI corresponding to KAM tori have about the same value, represented in the picture by the same grey color. *Left column:* a large portion of the action plane; *top:* $\varepsilon = 0.001$, $T = 1000$; *middle:* $\varepsilon = 0.01$, $T = 1000$; *bottom:* $\varepsilon = 0.04$, $T = 1000$. *Right column:* enlargement of the figures on the left, obtained with a larger integration time; *top:* $\varepsilon = 0.001$, $T = 4000$; *middle:* $\varepsilon = 0.01$, $T = 2000$; *bottom:* $\varepsilon = 0.04$, $T = 2000$ (from [21])

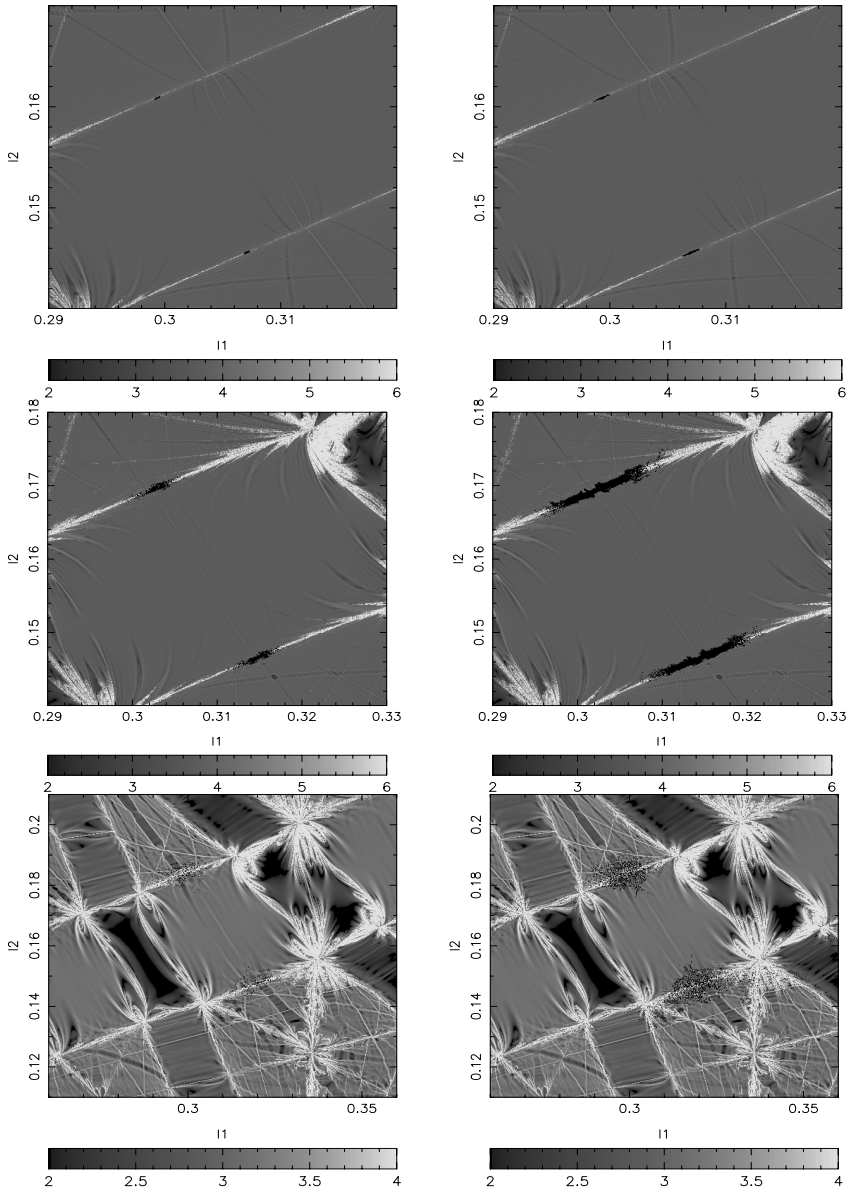

Fig. 6. Diffusion along the resonant line $I_1 = 2I_2$ for $\varepsilon = 0.003$ (*top*), $\varepsilon = 0.007$ (*middle*), $\varepsilon = 0.02$ (*bottom*) of a set of 100 initial conditions in the hyperbolic border of the resonance within the interval $0.303 \leq I_1 \leq 0.304$ and $0.143 \leq I_2 \leq 0.144$. *Black points* correspond to the intersections of the orbits on the double section $|\varphi_1| + |\varphi_2| \leq 0.05$, $\varphi_3 = 0$. The integration times are respectively: 10^7 (*top, left*), 10^8 (*top, right*), 10^6 (*middle, left*), $2.4 \cdot 10^7$ (*middle, right*), $1.6 \cdot 10^4$ (*bottom, left*), $5 \cdot 10^5$ (*bottom, right*). As in Fig. 5 a grey scale, ranging from black to white, is used for identifying different regions (from [41])

is confined. Graphical inspection provides the possibility of choosing initial data in the hyperbolic border. We can follow the evolution of the solution by computing the double section of the solution: $\sigma = |\varphi_1| + |\varphi_2| \leq 0.05$, $\varphi_3 = 0$. This strategy minimizes all projection effects and fast quasi-periodic motions; therefore, we can observe a very slow drift along the border of the resonance.

We considered 100 initial conditions in the interval $0.303 \leq I_1 \leq 0.304$, $0.143 \leq I_2 \leq 0.144$, having FLI values larger than $1.2 \cdot \log(T)$; such initial data generate chaotic orbits at the border of the resonance and they are chosen far from the more stable crossing with other resonances. Let us remark that the points in the double section will appear on both sides of the resonance (in fact the two white lines are connected by a hyperbolic region in the six dimensional phase space).

Figure 6 (top, left) provides the successive intersections with $\sigma \leq 0.05$, $\varphi_3 = 0$, up to a time $t = 10^7$; Fig. 6 (top, right) extends to $t = 10^8$. Even if the perturbing parameter is one order of magnitude lower than the threshold ε_0, the diffusion phenomenon along the resonant line clearly appears. When decreasing ε to the minimum value $\varepsilon = 0.001$, we detected diffusion along the resonance with smaller and smaller speed. For $\varepsilon = 0.007$ (Fig. 6 (middle, left) for $t = 10^6$, Fig. 6 (middle, right) for $t = 2.4 \cdot 10^7$), one obtains similar results, though the speed of diffusion is larger. For $\varepsilon = 0.02$ (Fig. 6 (bottom, left) for $t = 1.6 \cdot 10^4$, Fig. 6 (bottom, right) for $t = 5 \cdot 10^5$), diffusion along the resonance is still evident, although it extends a little in the direction transverse to the resonance. This phenomenon is due to the fact that we are approaching the transition value and higher order resonances intersecting the main one become evident. When approaching the critical value, we expect a large chaotic region to replace the zone of invariant tori, giving rise to fast diffusion.

Measuring the Diffusion Coefficient

Though an exhaustive analytic model does not yet exist, we tried to measure a diffusion coefficient as if the phenomenon was Brownian-like. The numerical experiments are constrained by computational limitations, since we have to find a compromise between the number of initial conditions and the length of the integration time. We observed an average linear increase with time, with slope D, of the mean squared distance from the initial conditions. More precisely we computed the quantity:

$$S(t) = \frac{1}{N} \sum_{j=1}^{N} [(I_{2,j}(t) + 2I_{1,j}(t)) - (I_{2,j}(0) + 2I_{1,j}(0))]^2 \,, \qquad (11)$$

where $I_{1,j}(0)$ and $I_{2,j}(0)$, $j = 1, ..., N$, are the initial conditions of a set of N orbits and $I_{1,j}(t)$ and $I_{2,j}(t)$ are the corresponding values at time t. The term in the square brackets of (11) represents the square of the distances of

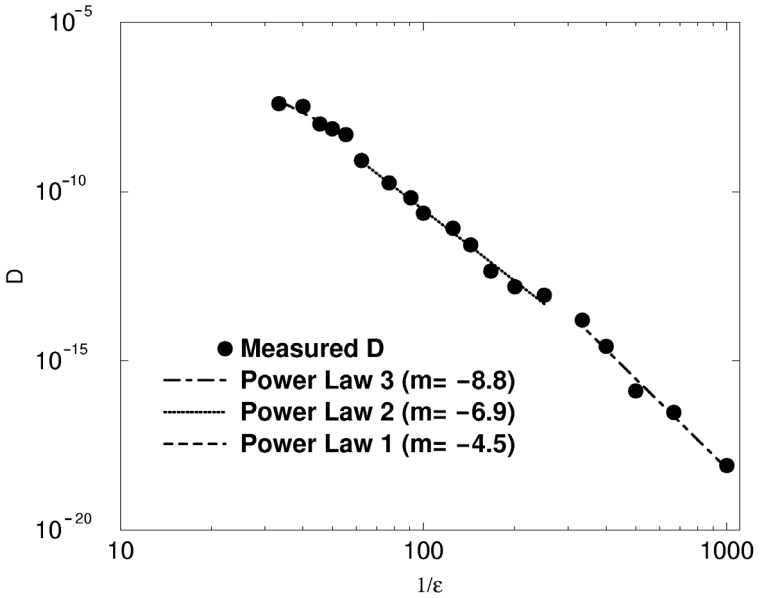

Fig. 7. Computation of the diffusion coefficient as a function of $1/\varepsilon$. The change of the slope of the three power–law fits agrees with the expected exponential decrease of D (from [41])

the actions from the initial values, projected on the resonant line $I_1 = 2I_2$. The estimates of D versus $1/\varepsilon$ are provided in Fig. 7 on a logarithmic scale. Clearly, we can exclude a linear regression, which would correspond to a power law $D(\varepsilon) = C\,(1/\varepsilon)^m$. For some different sets of data, we performed local regression, finding different slopes. In particular, the results are the following: the set containing the values of D for $1/\varepsilon \leq 55$ has slope $m = -4.5$; the set for $62 \leq 1/\varepsilon \leq 250$ has $m = -6.9$; the set with $1/\varepsilon \geq 330$ has $m = -8.8$. The changes in the slopes agree with the expected exponential decrease of D. An exponential fit of the form $D(\varepsilon) = C' \exp(-\kappa/\varepsilon)^\alpha$ (for some parameters κ and α) requires a larger ε-range, which is beyond our computational possibilities. To give en example, the value of D for $1/\varepsilon = 1000$ (see Fig. 7) required 4 months of CPU time on a fast workstation (Compaq AlphaStation XP1000 6/667 Mhz).

Acknowledgments

One of the authors (A.C.) is deeply indebted to L. Chierchia for useful suggestions and discussions.

References

1. V.I. Arnold. Proof of a Theorem by A.N. Kolmogorov on the invariance of quasi-periodic motions under small perturbations of the Hamiltonian. *Russ. Math. Surveys*, **18**, 9 (1963)
2. V.I. Arnold. Small denominators and problems of stability of motion in classical and Celestial Mechanics. *Russian Math. Survey*, **18**, 85 (1963)
3. V.I. Arnold. Instability of dynamical systems with several degrees of freedom. *Sov. Math. Dokl.*, **6**, 581 (1964)
4. V.I. Arnold. *Mathematical methods of classical mechanics.*, Springer–Verlag, New York (1978) (Russian original, Moscow, 1974)
5. V.I. Arnold (editor). *Encyclopaedia of Mathematical Sciences.* Dynamical Systems III, Springer-Verlag **3** (1988)
6. G. Benettin, F. Fassó, M. Guzzo. Nekhoroshev-stability of L_4 and L_5 in the spatial restricted three-body problem. *Regular and Chaotic Dynamics*, **3**, 56 (1988)
7. G. Benettin, G. Gallavotti. Stability of motions near resonances in quasi integrable Hamiltonian systems. *Journ. Stat. Phys.*, **44**, 293 (1986)
8. G. Boffetta, M. Cencini, M. Falcioni, A. Vulpiani. Predictability: a way to characterize complexity. *Physics Reports*, **356**, 367 (2002)
9. A. Celletti. Analysis of resonances in the spin–orbit problem in Celestial Mechanics: The synchronous resonance (Part I). *J. of Appl. Math. and Phys. (ZAMP)*, **41**, 174 (1990)
10. A. Celletti, L. Chierchia. Rigorous estimates for a computer-assisted KAM theory. *J. Math. Phys.*, **28**, 2078 (1987)
11. A. Celletti, L. Chierchia. A Constructive Theory of Lagrangian Tori and Computer–assisted Applications. *Dynamics Reported* (C.K.R.T. Jones, U. Kirchgraber, H.O. Walther Managing Editors), **4** (New Series), Springer–Verlag, 60 (1995)
12. A. Celletti, L. Chierchia. On the stability of realistic three-body problems. *Commun. Math. Phys.*, **186**, 413 (1997)
13. A. Celletti, L. Chierchia. *KAM Stability and Celestial Mechanics.* Preprint (2003)
14. B.V. Chirikov. A universal instability of many dimensional oscillator system. *Phys. Rep.*, **52**, 263 (1979)
15. G. Contopoulos, N. Voglis. A fast method for distinguishing between order and chaotic orbits. *Astron. Astrophys.*, **317**, 73 (1997)
16. C. Delaunay. *Théorie du Mouvement de la Lune.* Mémoires de l'Académie des Sciences **1**, Tome XXVIII, Paris (1860)
17. J.-P. Eckmann, P. Wittwer. *Computer methods and Borel summability applied to Feigenbaum's equation.* Springer Lecture Notes in Physics, **227** (1985)
18. C. Falcolini, R. de la Llave. A rigorous partial justification of Greene's criterion. *J. Stat. Phys.*, **67**, n.3/4, 609 (1992)
19. C. Froeschlé. A numerical study of the stochasticity of dynamical systems with two degrees of freedom. *Astron. Astrophys.*, **9**, 15 (1970)
20. C. Froeschlé. The Lyapunov characteristic exponents and applications. *Journal de Méc. théor. et appl.*, Numero spécial, 101 (1984)
21. C. Froeschlé, M. Guzzo, E. Lega. Graphical evolution of the Arnold's web, from order to chaos. *Science*, **289** N.5487, 2108 (2000)

22. C. Froeschlé, E. Lega. On the measure of the structure around the last KAM torus before and after its break-up. *Celest. Mech. and Dynamical Astron.*, **64**, 21 (1996)
23. C. Froeschlé, E. Lega. Twist angles, a fast method for distinguishing islands, tori and weak chaotic orbits. Comparison with other methods of analysis. *Astron. Astrophys.*, **334**, 355 (1998)
24. C. Froeschlé, E. Lega. On the structure of symplectic mappings. The fast Lyapunov indicator, a very sensitive tool. *Celest. Mech. and Dynamical Astronomy*, **78**, 167 (2000)
25. C. Froeschlé, E. Lega, R. Gonczi. Fast Lyapunov indicators. Application to asteroidal motion. *Celest. Mech. and Dynam. Astron.*, **67**, 41 (1997)
26. G. Gallavotti. *The Elements of Mechanics*. Springer–Verlag, New York (1983)
27. I. Goldhirsch, P-L. Sulem, and S.A. Orszag. Stability and Lyapounov stability of Dynamical Systems: a Differential Approach and a Numerical Method. *Physica D*, **27**, 311 (1987)
28. J.M. Greene. A method for determining a stochastic transition. *J. of Math. Phys.*, **20**, 1183 (1979)
29. M. Guzzo, E. Lega, C. Froeschlé. On the numerical detection of the stability of chaotic motions in quasi–integrable systems. *Physica D*, **163**, 1 (2002)
30. M. Guzzo, A. Morbidelli. Construction of a Nekhoroshev like result for the asteroid belt dynamical system. *Cel. Mech. and Dyn. Astron.*, **66**, 255 (1997)
31. M. Hénon. Explorationes numérique du problème restreint IV: Masses egales, orbites non periodique. *Bullettin Astronomique*, **3**, no. 1, fasc. 2, 49 (1966)
32. M. Hénon, C. Heiles. The applicability of the third integral of motion: Some numerical experiments. *Astron. J.*, **1**:73 (1964)
33. H. Koch, A. Schenkel, P. Wittwer. Computer-Assisted Proofs in Analysis and Programming in Logic: A Case Study. *SIAM Review*, **38**, n. 4, 565 (1996)
34. A.N. Kolmogorov. On the conservation of conditionally periodic motions under small perturbation of the Hamiltonian. *Dokl. Akad. Nauk. SSR*, **98**, 469 (1954)
35. O.E. Lanford III. Computer assisted proofs in analysis. *Physics A*, **124**, 465 (1984)
36. J. Laskar. The chaotic motion of the solar system. A numerical estimate of the size of the chaotic zones. *Icarus*, **88**, 266 (1990)
37. J. Laskar. Frequency analysis for multi-dimensional systems. Global dynamics and diffusion. *Physica D*, **67**, 257 (1993)
38. J. Laskar. Large scale chaos in the solar system. *Astron. Astrophys.*, **287**, L9 (1994)
39. J. Laskar, C. Froeschlé, A. Celletti. The measure of chaos by the numerical analysis of the fundamental frequencies. Application to the standard mapping. *Physica D*, **56**, 253 (1992)
40. E. Lega, C. Froeschlé. *Fast Lyapunov Indicators. Comparison with other chaos indicators. Application to two and four dimensional maps.* In The Dynamical Behaviour of Our Planetary System. Kluwer academ. publ., J. Henrard and R. Dvorak eds. (1997)
41. E. Lega, M. Guzzo, C. Froeschlé. Detection of Arnold diffusion in Hamiltonian systems. *Physica D*, accepted (2003)
42. A.J. Lichtenberg, M.A. Lieberman. *Regular and Stochastic motion.* Springer, Berlin, Heidelberg, New York (1983)
43. P. Lochak. On the conservation of conditionally periodic motions under small perturbation of the Hamiltonian. *Russ. Math. Surv.*, **47**, 57 (1992)

44. R.S. MacKay. Greene's residue criterion. *Nonlinearity*, **5**, 161 (1992)
45. M. Month, J.C. Herrera. *Nonlinear dynamics and the beam-beam interaction.* American Institute of Physics, Month and Errera eds. (1979)
46. J. Moser. On invariant curves of area-preserving mappings of an annulus. *Nach. Akad. Wiss. Göttingen, Math. Phys. Kl. II*, **1**, 1 (1962)
47. N.N. Nekhoroshev. Exponential estimates of the stability time of near-integrable Hamiltonian systems. *Russ. Math. Surveys*, **32**, 1 (1977)
48. D. Nesvorny, S. Ferraz-Mello. On the asteroidal population of the first-order Jovian resonances. *Icarus*, **130**, 247 (1997)
49. Y. Papaphilippou, J. Laskar. Global dynamics of triaxial galactic models through frequency map analysis. *Astronomy and Astrophysics*, **329**, 451 (1998)
50. H. Poincaré. *Les méthodes nouvelles de la méchanique céleste.* Gauthier Villars, Paris (1899)
51. J. Pöschel. Nekhoroshev's estimates for quasi-convex Hamiltonian systems. *Math. Z.*, **213**, 187 (1993)
52. C.L. Siegel, J.K. Moser. *Lectures on Celestial Mechanics.* Springer–Verlag Berlin Heidelberg, New York (1971)
53. V. Szebehely. *Theory of orbits.* Academic Press, New York and London (1967)

Dynamics at the Border of Chaos and Order

Michael Zaks[1] and Arkady Pikovsky[2]

[1] Humboldt University of Berlin, 12489 Berlin Germany,
zaks@physik.hu-berlin.de
[2] Potsdam University, 14469 Potsdam, Germany
pikovsky@stat.physik.uni-potsdam.de

Abstract. We give a brief overview of complex dynamical behavior characterized by singular continuous (fractal) Fourier spectra. After presenting a simple example of an aperiodic symbolic sequence we discuss different dynamical mechanisms of such unusual correlation properties. Examples include hydrodynamical flows and dissipative dynamical systems described by ordinary differential equations.

1 Introduction

In his memoirs A.N. Kolmogorov mentions that, in his treaties on classical mechanics, he was greatly influenced by the works by John von Neumann on the spectral theory of dynamical systems and by the classical work of Bogoljubov and Krylov. In this paper we would like to present recent interesting examples of the spectral properties of dynamical systems; significant progress in this field due in large part to the contributions of the mathematicians belonging to the school of A.N. Kolmogorov.

The term "spectral theory of dynamical systems" has a twofold meaning. The first is rather pragmatic and stems from the statistics and signal processing. Here a dynamical system (e.g., a system of ordinary differential equations) is considered as a "source" of a stationary process – periodic, quasiperiodic, chaotic – and one calculates the Fourier spectrum of this process. Typically, this procedure follows the recipes for the stationary random processes, so that the dynamical system is characterized by the power spectrum of one of its observables. Numerically, one often uses the power spectrum to distinguish a periodic process (whose power spectrum is a series of what a physicist calls "delta"-peaks and a mathematician calls "point spectrum") from a chaotic one, whose spectrum is continuous. As we shall see below, in between of these cases, there is also a possibility of having a fractal power spectrum – the object of our main interest in this paper.

The spectrum in the above sense is obviously a non-invariant characteristic of the system, since it depends on the observable. This drawback is not present in the operator-based approach to a spectrum of a dynamical system, pursued by J. von Neumann after the pioneering work by Koopman. Inspired by the development of quantum mechanics, Koopman and von Neumann suggested the treatment of classical dynamical systems (e.g. systems of ordinary differential equations) by using evolution operators acting on the

M. Zaks and A. Pikovsky, Dynamics at the Border of Chaos and Order, Lect. Notes Phys. **636**, 61–82 (2003)
http://www.springerlink.com/

observables – functions of the phase space variables. In this way the evolution in time is described by a linear operator, called Koopman operator. Now the "spectral theory" means a study of a spectrum of this operator. In this interpretation, the spectrum, like in quantum mechanics, is a set of possible values of frequencies. A particular power spectrum of an observable is in this language a spectral measure, resulting from the spectral expansion of the corresponding function.

Although mathematically deeper, the operator approach to the spectra of dynamical systems is not very helpful in practice, because contrary to the case of the Hamilton operator in quantum mechanics, the spectrum of the Koopman operator generally cannot be found, even numerically. We will therefore mainly use the signal-based approach in our presentation below.

We have already mentioned the two main types of spectra: discrete and continuous. The discrete spectrum includes not only the case of periodic dynamics, but also the quasiperiodic one, where several basic incommensurate frequencies (i.e. frequencies whose ratio is an irrational number) and their combinations are present. In terms of the autocorrelation function (assumed to be normalized, so that the maximum value at zero time is one) of the process, which is the Fourier transform of the power spectrum, the discrete case corresponds to a periodic function which returns to unity for a periodic motion, and to regularly returning almost to unity function for a quasiperiodic motion. Remarkably, the proof of the discreteness of the spectrum for a smooth flow on a 2-torus has been presented by A. N. Kolmogorov in 1953 [1]. Among the conditions which enable the discreteness, formulated by Kolmogorov, is the absence of the equilibrium points; we will discuss the implications that violation of this condition has on the spectrum in the final section of this chapter.

Another popular type of spectrum is a continuous one, here the autocorrelation function decays, which corresponds to mixing – a strong chaotic property. However, such a characterization is not complete. A continuous spectrum, like any continuous measure, can have finite density – and this is the usual case for an absolutely continuous spectrum, but it can also be concentrated on a set of measure zero, e.g. on a fractal. The latter case is called singular continuous, or fractal, spectrum. It corresponds to the dynamics that can be classified as those between order and chaos.

We start with a simple example, the Thue–Morse symbolic sequence, which will allow us to give a clear view of a fractal spectrum and of the corresponding properties of correlations (Sect. 2). In Sect. 3, we demonstrate that this symbolic sequence can be directly applied to the description of certain dissipative dynamical systems at the border between order and chaos. We will show that a nontrivial symbolic representation can be responsible for the complex features of the spectrum; it is also a particular property of trajectories to stick at some places in the phase space. This second mechanism works in our last example – a static incompressible two-dimensional fluid flow (Sect. 4).

2 Thue–Morse Sequence: A Nontrivial Example of Complex Symbolic Coding

Symbolic sequences are the simplest objects allowing one to introduce and study the notions of determinism, randomness, and complexity (see, e.g., [2]). They appear quite naturally as "codes" for the trajectories of deterministic dynamical systems. In the investigations of fractal spectra, the most convenient way to understand the structure of the spectrum and the correlation function is to look at a simple sequence of two symbols, first introduced by A. Thue and M. Morse about a century ago [3,4]. While working on the spectral theory of discrete processes, N. Wiener used this sequence (without calling it by name) as an example of an object where the spectrum was neither discrete nor absolutely continuous [5,6].

The Thue–Morse sequence is constructed with two symbols; we will first denote them with letters A and B. There are several equivalent definitions which describe the same symbolic object. One is based on the repetitive substitutions – inflations – of an initial sequence. Here each element in a sequence is substituted by two according to the following rule:

$$A \;\to\; A\,B\,, \qquad\qquad B \;\to\; B\,A\,.$$

If we start with an initial sequence having one symbol $M_0 = A$, then the substitutions lead to sequences of lengths $2, 4, 8, \ldots$:

$$M_1 = A\,B\,, \qquad M_2 = A\,B\,B\,A\,, \qquad M_3 = A\,B\,B\,A\,B\,A\,A\,B\,, \;\ldots\;. \tag{1}$$

Another definition of the Thue–Morse sequence is based on concatenations, as follows: given a symbolic string M_n of length 2^n we append to it the string $\overline{M_n}$ which is obtained from M_n by exchanging all the symbols : $A \leftrightarrow B$. One can easily check that application of this rule to $M_0 = A$ produces exactly the strings as above.

The construction above suggests that the complexity of the Thue–Morse sequence is small. Kolmogorov has suggested to define the complexity of a finite string as a minimal length of a program (on a universal computer, such as a Turing machine) that produces this string and halts after that [7]. From this definition, it is clear that the algorithm for the generation of the Thue–Morse sequence is compact, and the only additional information needed to proceed to longer strings is that about the length of the string – and it grows as a logarithm of the length. Thus, the complexity per one symbol tends to zero for long pieces of the Thue–Morse sequence (the same is of course true for all other sequences obtained by inflations or concatenations), which corresponds to our intuitive conception of full predictability.

Next we will argue that the predictability of the Thue–Morse sequence is not trivial, since one cannot formulate it as repetition of some of its pieces. Indeed, this sequence is not periodic: one cannot find a repeated subsequence. Moreover, there are no long subsequences which are repeated with minor

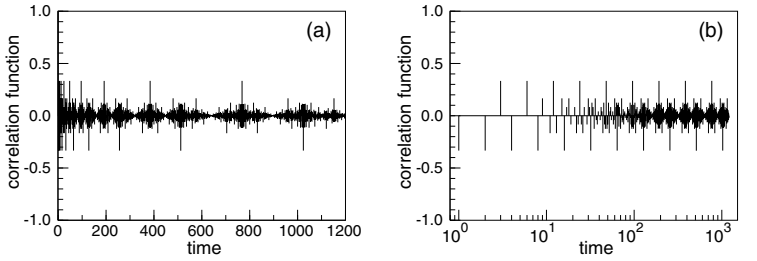

Fig. 1. Autocorrelation function of the Thue–Morse symbolic sequence in two representations of the time scale – linear (**a**) and logarithmic (**b**). In the logarithmic scale, it is evident that the correlations do not decay, though they become more and more rare

modification as for quasiperiodic sequences. To see this, we need to calculate the correlation function and the power spectrum. The first step is in assigning numerical values to the symbols. In the case of two symbols, this assignment is straightforward: we replace A by $+1$ and B by -1. Notice that due to symmetry $A \leftrightarrow B$, the mean value of the resulting sequence x_k vanishes.

Recall that the autocorrelation function of a sequence x_k is defined as

$$C(t) = \langle x_k x_{k+t} \rangle = \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} x_k x_{k+t} \, .$$

The recursive way of constructing the Morse–Thue sequence suggests that the autocorrelation function obeys some recurrences. These can be formulated as [8]

$$C(2t) = C(t) \, , \qquad C(2t+1) = -\frac{C(t) + C(t+1)}{2} \, .$$

These relations, combined with the "initial condition" $C(0) = 1$ allow one to obtain the exact values of $C(t)$ for any t. The autocorrelation function is depicted in Fig. 1. It has peaks of height $-1/3$ at $t = 1, 2, 4, 8, \ldots$ and of height $1/3$ at $t = 3, 6, \ldots$. These peaks characterize some repetitions in the sequence and, because their height is not close to one (as it would be for quasiperiodic processes), the repetitions are not exact, but approximate. However, the correlations do not decay with time as for random and chaotic processes.

A similar recurrence can be written for the power spectrum. It is convenient to represent it as a limit of the spectra obtained from the finite strings. If we take a string of length 2^n, the corresponding approximation to the power spectrum can be written as

$$S_n(\omega) = \frac{1}{2^n} \left| \sum_{k=1}^{2^n} x_k e^{2\pi i k \omega} \right|^2 \, .$$

Now we use the fact that due to the concatenation rule above, the second half-string $x_{2^{n-1}+1} \dots x_{2^n}$ is the first half-string $x_1 \dots x_{2^{n-1}}$ with the opposite sign. Thus

$$\sum_{k=1}^{2^n} x_k e^{2\pi i k\omega} = \left(1 - e^{2\pi i 2^{n-1}\omega}\right) \sum_{k=1}^{2^{n-1}} x_k e^{2\pi i k\omega} .$$

and the approximations to the power spectrum obey

$$S_n(\omega) = (1 - \cos(2^n \pi\omega)) S_{n-1}(\omega) .$$

Iterations of this relation lead to a representation of the power spectrum as an infinite product called the Riesz product. The spectrum has many peaks but no delta-peaks, and therefore has no discrete component. It is fractal, or singular continuous. One gets an impression of this object when considering approximations to S_n, Fig. 2.

Note that although fractality of the spectrum sounds impressive, in practical calculations it is not convenient to look at because the spectrum does not converge to a finite form. In contrast, the autocorrelation function, which is the Fourier transform of the spectrum, is well-defined and one can easily calculate an approximation to it for a finite range of time shifts t. One can therefore formulate a practical approach to the investigation of processes with fractal spectra: one calculates the autocorrelation function, and persisting peaks in it (usually at logarithmically periodically placed values of time) indicate a fractal spectrum.

Moreover, one can use the autocorrelation function to calculate important characteristic of the fractal spectrum: its correlation dimension D_2. To this end one defines the integrated autocorrelation function

$$C_{int}(T) = \frac{1}{T} \sum_{t=0}^{T} C^2(t) .$$

According to Wiener [5], a vanishing $C_{int}(\infty)$ implies the absence of the discrete component in the spectrum. The decay rate of the integrated autocorrelation function for large T is related to the correlation dimension of the spectrum according to the formula [9]

$$C_{int}(T) \sim T^{-D_2} .$$

Remarkably, for the Thue–Morse sequence, this dimension can be calculated exactly [8]

$$D_2 = 3 - \frac{\log(1 + \sqrt{17})}{\log 2} = 0.64298 \dots .$$

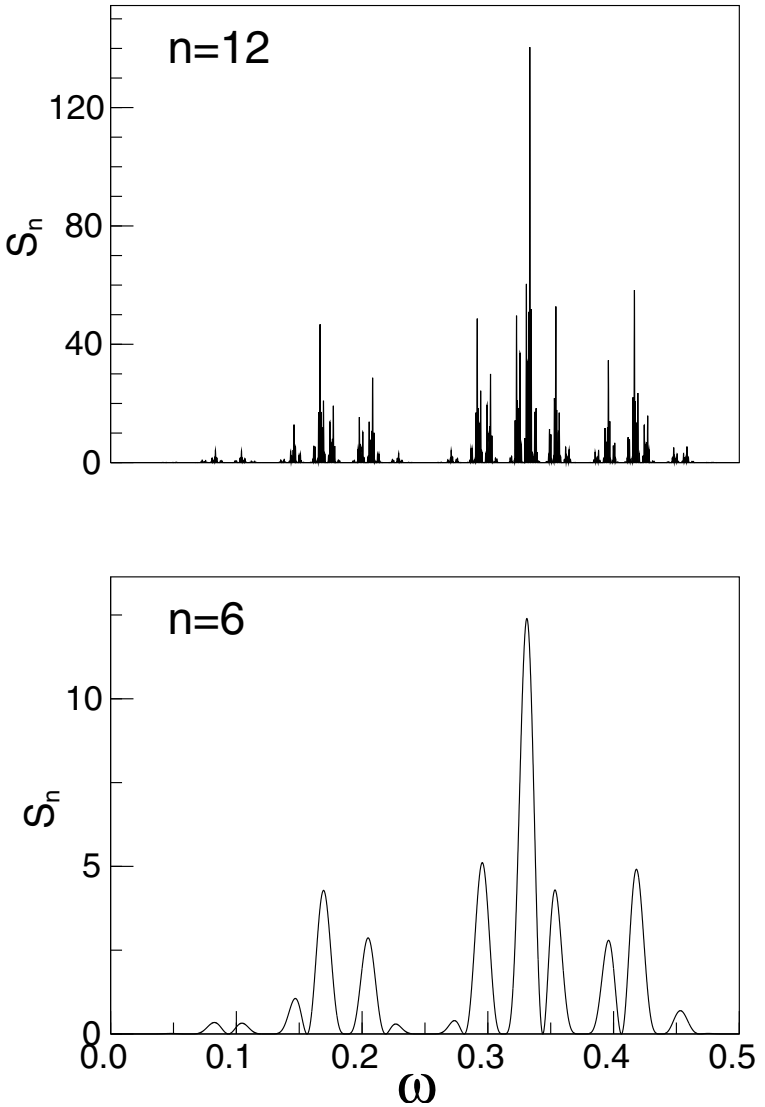

Fig. 2. Two approximations to the fractal spectrum of the Thue–Morse sequence. Notice different vertical scales due to growth of the peaks with increasing n. Both the peaks and the holes are eventually everywhere dense, so that the limiting object is difficult to depict

3 Attractors with Fractal Spectra: From Symbolic Encoding to Singularities of Return Times

Symbolic codes are interesting not only in themselves. As explained in Chap. 4 of this volume which is devoted to entropy, chaos and complexity, strings of

symbols naturally arise in descriptions of trajectories of dynamical systems: phase space is partitioned into regions, each of which is denoted by a symbol, and the trajectory which alternately enters those regions, is encoded by a symbolic string. If a partition of the phase space reflects the properties of the dynamical system, the trajectory and the resulting symbolic sequence share many characteristics.

Keeping this in mind, it would be reasonable to expect that a dynamical system with an attractor encoded by the Thue–Morse sequence possesses a singular-continuous Fourier spectrum [10].

An example of such attractor is delivered by three differential equations:

$$\begin{aligned} \dot{x} &= \sigma(y - x) + \sigma D y(z - R) \\ \dot{y} &= Rx - y - xz \\ \dot{z} &= xy - bz \ , \end{aligned} \tag{2}$$

given appropriate parameter values. A few words about the origin of these equations [11]: They model nonlinear regimes of convection in a layer of non-isothermal fluid subjected to the transverse high-frequency modulations of gravity. It is known that vibrations suppress small-scale fluid motions; it makes sense therefore to truncate the Fourier expansion of the Boussinesq equations of convection, retaining just one large-scale mode (denoted by x) for the velocity field and two modes (y and z) for the temperature distribution, and neglect the higher modes. The parameter D characterizes the intensity of vibration and is given by the squared ratio of the modulation amplitude to its frequency. At $D = 0$ (no vibrations, static gravity field) the system turns into the familiar Lorenz equations [12]. The rest of the parameters have the same meaning as in Lorenz's paper: σ is the Prandtl number of the fluid (ratio of kinematic viscosity to thermal diffusivity), R is the Rayleigh number which measures the strength of the buoyancy force, and b is the geometric parameter defined through the wavelength of the flow.

On part of the border between order and chaos in the parameter space, the attracting set of (3) can be encoded by the Thue–Morse sequence. In order to make sure that this is really the case, it is helpful to consider the whole bifurcation scenario which leads from the steady quiescent state to irregular oscillations [13]. An important property of the equations is their invariance with respect to the transformation $(x, y) \Leftrightarrow (-x, -y)$. Therefore every attracting set in the phase space is either self-symmetric (invariant under the transformation), or has a symmetric "twin". A phase portrait of a time-dependent state is built from alternating rotations "around" (in a some rough sense, which is quite sufficient for us) two unstable fixed points, one of them in the half-space $x > 0$ and the other in the half-space $x < 0$. The partition which we choose is rather simple: we denote each turn of the orbit in the half-space $x > 0$ by the symbol A and each turn in $x < 0$ by B. Obviously, in the case of periodic dynamics, the infinite symbolic sequence is also periodic; if the orbit in the phase space closes after N turns, we call

the initial N letters in its symbolic code a "label". For a starting symbol we take the outermost turn on the (x, z) projection; in the case of two symmetric largest turns, the sequence starts with the "positive" one (i.e. with A).

For a route to chaos in the parameter space of (3), we fix the "traditional" values of $\sigma = 10$ and $b = 8/3$ [12], take some fixed value of D and increase R from zero (physically, this corresponds to the growth of the temperature difference across the fluid layer). If the chosen value of D is small enough, the bifurcation sequence of the Lorenz equations is reproduced: a steady state at the origin O ($x = y = z = 0$) is replaced by two symmetric fixed points; these points lose stability by way of the subcritical Hopf bifurcation and yield to chaos [14]. However, under moderate intensity of vibrations ($0.05 < D < 0.09$), the scenario of transition to chaos is different. Here, the increase of R leads from stable steady states to stable periodic oscillations. In the phase space, such oscillations are presented by periodic orbits (Fig. 3a); their symbolic codes are $AAA\ldots$ and $BBB\ldots$, respectively.

The origin O is a saddle-point in this parameter region. When the value of R is increased, both the amplitude and the period of the oscillations grow (Fig. 3b) until, at a certain bifurcation point, the closed curve "touches" O and forms a so-called homoclinic (bi-asymptotic) orbit to the origin: the solution tends to $x = y = z = 0$ both at $t \to -\infty$ and at $t \to +\infty$. Due to the symmetry, the second closed curve also forms the homoclinic orbit. In this way, two closed curves in the phase space "glue together" (Fig. 3c). When the parameter R is increased beyond this critical value, the homoclinic orbit is destroyed; it leaves a unique stable periodic solution, which consists of two turns in the phase space, and is self-symmetric (Fig. 3d). In some sense this gluing bifurcation can be viewed as a kind of "doubling": doubling of the length of the attracting curve in the phase space. (In contrast to the conventional period-doubling, the temporal period of an orbit is not doubled at this bifurcation. Instead, it grows to infinity and then decreases again). The symbolic code of this new orbit is $ABABAB\ldots$. In terms of the labels, the gluing bifurcation is merely a concatenation: the labels of previously existing orbits A and B are glued together and constitute the new label AB. Further increase in R leads to the symmetry crisis of this state: two new stable periodic orbits bifurcate from it (Fig. 3e). They also consist of two turns in the phase space, and the symmetry transformation maps each orbit onto the other one. For one of the orbits the larger of two turns lies in the half-space $x > 0$; this orbit retains the label AB. Since for the "twin" orbit the situation is opposite, its label is BA.

As R increases further, the outermost turns of these two orbits grow until they touch the origin O. This is another homoclinic bifurcation: the trajectory which starts infinitesimally close to the origin returns back to it, this time after two turns in the phase space (Fig. 3f). Two periodic orbits are now glued together. A further increase of R beyond this bifurcation value leaves in the phase space a unique stable periodic orbit which consists of 4 turns.

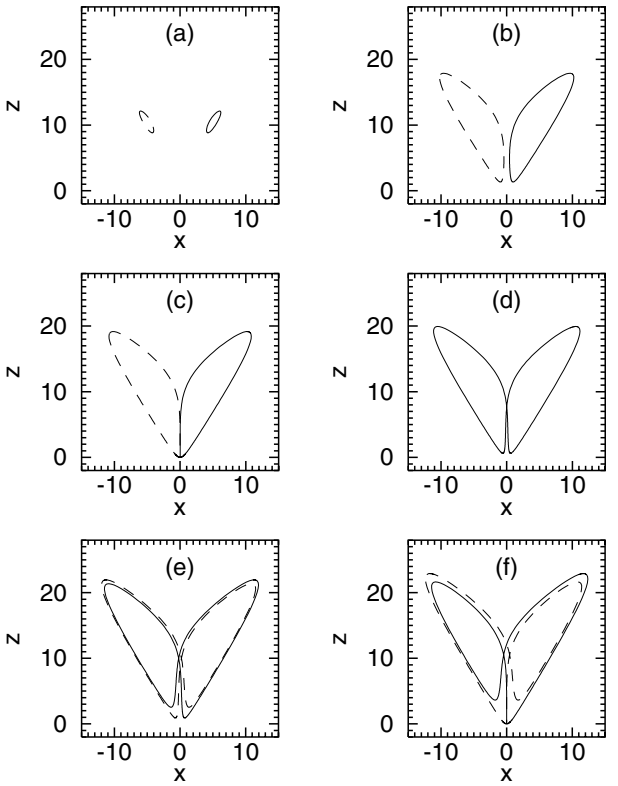

Fig. 3. Evolution of the attracting set. **a,b** two periodic orbits labelled A and B; **c** two homoclinic orbits; **d** single periodic orbit labelled AB; **e** periodic orbits labelled AB and BA; **f** two 2-turn homoclinic orbits. *Solid line:* orbits whose labels start with A; *dashed line:* orbits whose labels start with B

Again, the labels are concatenated: $AB + BA \to ABBA$. This is followed by another symmetry crisis, and so on.

Thus we see that in the parameter space two kinds of bifurcations alternate. The first one is the homoclinic bifurcation which glues two stable periodic solutions into a single self-symmetric one (with doubled length in the phase space), while the second is the symmetry-breaking pitchfork bifurcation: a stable self-symmetric solution gives birth to two stable solutions which are mutually symmetric.

From the point of view of symbolic labels, the first kind of bifurcation is a concatenation of two sequences into a single one. Bifurcations of the second kind can be viewed as a kind of "negation" on the binary alphabet: out of a given symbolic string they produce its mirror image. The evolution of the

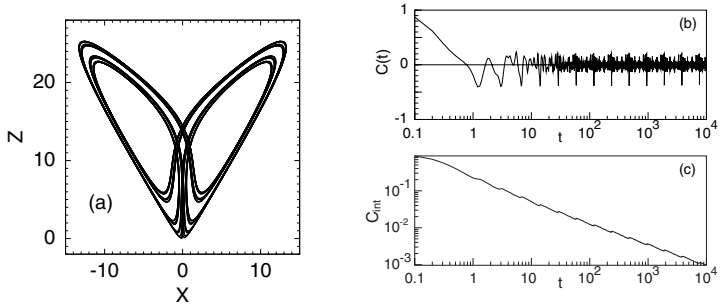

Fig. 4. **a** Attractor with the Thue–Morse symbolic code at the accumulation of the bifurcation scenario; **b** autocorrelation and **c** integrated autocorrelation for the variable x on this attractor

labels of attracting sets proceeds as follows:

$$\begin{matrix} A \\ B \end{matrix} \to AB \to \begin{matrix} AB \\ BA \end{matrix} \to ABBA \to \begin{matrix} ABBA \\ BAAB \end{matrix} \to ABBABAAB \to \dots \ . \qquad (3)$$

We recognize in this description the building rule of the Thue–Morse sequence (see (1)). The attracting set of (3), which emerges at the accumulation point of the bifurcation scenario, is plotted in Fig. 4. This set is vaguely reminiscent of the Lorenz attractor with its two lobes; in contrast to the latter, however, it does not look like a densely filled folded band, but displays a well-pronounced self-similar structure. Of course, the symbolic code of this attractor is nothing else but the Thue–Morse sequence.

Indeed, the computation of the autocorrelation gives evidence that the Fourier spectrum is singular continuous. For the observable $x(t)$, according to Fig. 4b, the autocorrelation does not ultimately decay: it displays an approximate log-periodic pattern like in the case of the Thue–Morse sequence. The largest values of t on this plot correspond to ~ 5000 average durations of one turn in the phase space. Accordingly, the Fourier spectrum cannot be purely absolutely continuous. The integrated autocorrelation, on the other hand, systematically decays by several orders of magnitude within the same time interval (Fig. 4c). This indicates the absence of a discrete component in the spectrum. The only remaining possibility for the spectrum is to be fractal.

Upon closer inspection, however, we notice a remarkable distinction between the autocorrelation of the Thue–Morse sequence and that of the variable x from (3): in the latter case, the largest negative peaks ("anticorrelations") are much more strongly pronounced than their positive counterparts. It appears that the description of dynamics with the help of the symbolic code is somewhat incomplete. Checking for the reason of this minor but nevertheless noticeable discrepancy, we replace the continuous flow (3) by the discrete dynamical system: the return (Poincaré) mapping induced by this flow. As seen in the phase portrait in Fig. 4, the trajectories on the attractor return

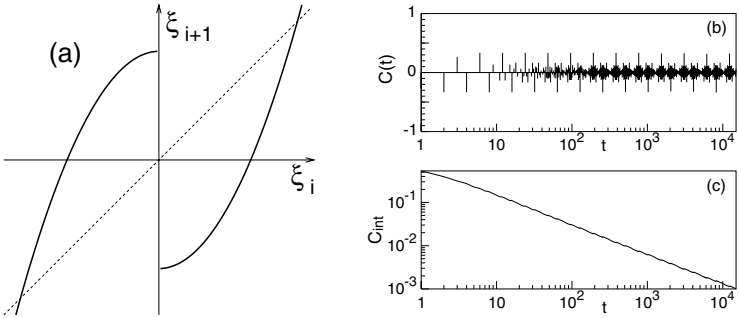

Fig. 5. One-dimensional return mapping (4) with ν=2.0, μ=1.40155189, its correlation function (**b**) and integrated autocorrelation (**c**)

again and again to the region around the origin O. In order to quantify these returns, we position a plane perpendicular to the z-axis close enough to the origin. The return mapping relates the coordinates (x_i, y_i) and (x_{i+1}, y_{i+1}) of two subsequent intersections of this plane by an orbit of the flow. The flow is strongly contracting; within one turn of the orbit (that is the time which separates two intersections of the plane), the phase volume is squeezed so strongly that the image of a rectangle is nearly a one-dimensional curve. Therefore for all practical reasons it is sufficient to consider a one-dimensional return mapping. Before writing it explicitly, we note that in this parameter region the linearization matrix of (3) near the origin has three real eigenvalues; we order them as $\lambda_1 > 0 > \lambda_2 > \lambda_3$. The ratio $\nu = |\lambda_2|/\lambda_1$ is called the saddle index.

It can be rigorously shown that the one-dimensional return mapping is reducible to the form

$$\xi_{i+1} \;=\; (\,|\xi_i|^{\nu} - \mu\,)\,\mathrm{sign}\,\xi_i + \text{ higher order terms}\,. \qquad (4)$$

Here, the coordinate ξ is a linear combination of x and y, whereas μ is the parameter whose variation largely corresponds to the variation of R in (3). The value $\mu = 0$ describes the situation in which a trajectory starting infinitesimally close to the origin O, returns back to O; of course, this is the homoclinic bifurcation shown in Fig. 3c.

A graphical representation of (4) is sketched in Fig. 5. Due to the mirror symmetry of the flow, the mapping is odd: its graph consists of two symmetric branches. The orbits of the flow can leave the neighbourhood of the origin either in the direction of $x > 0$ or in the direction of $x < 0$; in the mapping, this is reflected by the discontinuity at $\xi = 0$. To mimic the bifurcation scenario of Fig. 3, one needs to have $\nu > 1$ (the parameter region $\nu < 1$ corresponds to a different scenario: the Lorenz-type transition to chaos [14]). For a fixed value of $\nu > 1$, the increase of μ leads to a sequence of bifurcations in which either two periodic states are glued together into a symmetric one

(and the period is doubled), or a self-symmetric periodic attractor undergoes a pitchfork bifurcation, giving birth to two mutually symmetric stable periodic orbits. By introducing the simple partition: symbol A for each point of the orbit in the domain $\xi > 0$ and symbol B for each point with $\xi < 0$, we see that each pair of bifurcations results in the concatenation of a symbolic label with its binary "negation". Of course, this bifurcation scenario culminates in the attractor whose symbolic code is the Thue–Morse sequence. It is not surprising then that the autocorrelation function of the observable ξ on this attractor has the familiar log-periodic pattern (Fig. 5b). Moreover, the correspondence is much better than in the above case of the observable from the continuous flow: except for the short initial segment the balance between the largest correlations and anticorrelations is reproduced, and, even quantitatively, the values of $C(t)$ at the highest peaks practically do not decline from the Thue–Morse mark 1/3. This shows that from the spectral point of view, the discrete dynamical system described by the mapping (4) is closer to the Thue–Morse binary sequence than the continuous flow (3): the Fourier spectrum of the latter is also singular continuous but is a bit different from a quantitative point of view.

Before we proceed with an interpretation of this difference, it is worth noting that graphically, the return mapping looks like a symmetric unimodal mapping (e.g. the famous logistic one) whose right branch has been flipped. Indeed, there exists a connection between the bifurcation sequence in (4) and the universal period-doubling scenario [13]. There is also a quantitative distinction: in systems which exhibit period-doubling, the extremum of the return map is generically quadratic. In dynamical systems with homoclinic bifurcations, the saddle index can take any real value and $\nu = 2$ (which corresponds to the "flipped" quadratic map) is by no means singled out. This metric distinction causes differences e.g. in scaling constants like the convergence rate of the bifurcation sequence, etc. For us, however, another difference is more significant: the Fourier spectrum of a system at the accumulation point of the period-doubling bifurcations is purely discrete [15]. Accordingly, this system is well correlated: $C(2^n) \to 1$.

All this has direct consequences on the type of dynamics which we consider. Reduction of (3) to the discrete mapping should not necessarily be done in the above manner. Instead of using for the mapping a combination of variables x and y, both of which participate in the symmetry transformation $(x, y) \Leftrightarrow (-x, -y)$, one can take the remaining variable z. For this, of course, the Poincaré plane $z = \text{const}$ should be replaced by a cylindrical surface (e.g. $z^2 + y^2 = \text{const}$). Alternatively, one can follow the approach of Lorenz [12] and write down the recursion relation for the subsequent maxima of z on the orbit turns. Such mapping can be brought to the form

$$z_{i+1} = A|z_i|^\nu - \mu + \text{ higher order terms.} \tag{5}$$

In contrast to (4), this mapping turns out to be even and continuous. The dynamics of the underlying equations (3) is related to the behavior of (5)

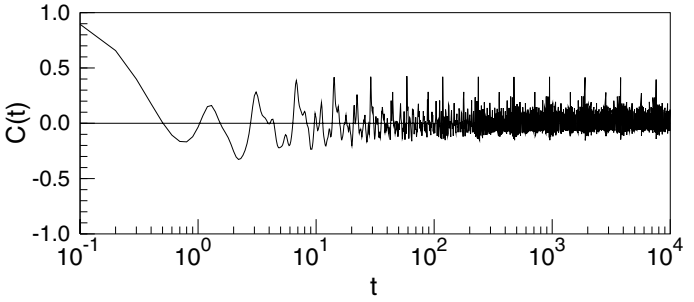

Fig. 6. Autocorrelation for the variable z on the attractor from Fig. 4

in the following way: symmetry-breakings (pitchfork bifurcations) of periodic orbits in (3) correspond to period-doublings in (5), whereas the formation of homoclinic trajectories in (3) corresponds to "superstability" (passage of a periodic point through the extremum) in (5). Accordingly, the attractor from Fig. 4 with its *singular continuous* spectrum is matched in (5) by the period-doubling attractor of the Feigenbaum type, whose spectrum is *discrete*.

This discrepancy cannot be reduced to the difference between the symmetry properties of the observables: by computing the autocorrelation of the variable $z(t)$ from (3), we observe the typical attributes of the fractal spectral component (Fig. 6). This means that the continuous variable $z(t)$ has spectral and correlation properties which are qualitatively different from the characteristics of its discretized counterpart z_i.

At a first sight this situation appears to be unusual: the behavior of the flow is less correlated (in other words, more complicated) than the dynamics of the Poincaré mapping induced by it. This seems to contradict the common practice in which the conclusions for the dynamics of the flow are drawn from observation of the dynamics on the Poincaré system. In fact, there is no contradiction. The necessary condition for adequacy of the mapping is boundedness of time intervals between the returns onto the Poincaré plane. For the attractor from Fig. 4 this condition is violated: the saddle point at the origin belongs to the closure of the attractor, and trajectories pass arbitrarily close by. During such passages, the velocity in the phase space is arbitrarily small: the system "hovers" over the unstable equilibrium, and the duration of such hoverings is unbounded. The nearly constant values of the observables during these long intervals make substantial contributions to the averaged characteristics of the trajectory, such as the autocorrelation or the Fourier spectrum. On the other hand, in terms of the Poincaré mapping these hoverings remain completely unaccounted.

In a sense, iterations of the return mapping are separated by time intervals of variable length. In ergodic theory there exists an efficient tool for modeling such situations. It is known under several different names: special flow, or flow over the mapping, or flow under a function. This is a two-dimensional time-

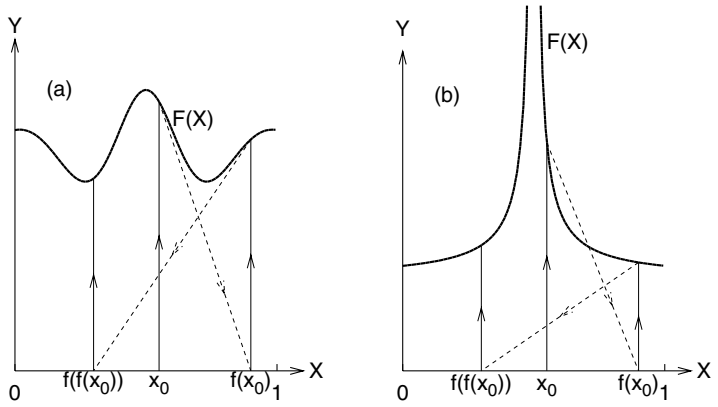

Fig. 7. Geometry and dynamics of special flows. **a** bounded return time, **b** unbounded return time

continuous system in which one of the variables is piecewise-constant; lengths of these "pieces" can be interpreted as durations of time intervals between the iterations of the mapping.

To construct this flow, two functions are needed; we denote them $f(x)$ and $F(x)$, respectively. The special flow is defined on the stripe of 2-dimensional plane $0 \le y \le F(x)$, i.e. between the abscissa and the graph of $F(x)$ (this explains the origin of the name "under the function"). Dynamics of the flow is illustrated by Fig. 7. Take some initial point (x_0, y_0) within this stripe. From there the system moves upwards at unit speed; this means that x remains constant, while y grows uniformly. On reaching the upper boundary at the point $(x_0, F(x_0))$ the system makes an instantaneous jump to the point $(f(x_0), 0)$ on the abscissa. From here the system again moves strictly upwards with unit speed until hitting the boundary at $(f(x_0), F(f(x_0)))$. An instantaneous jump leads from there to the point $(f(f(x_0)), 0)$ on the abscissa, and the process repeats. Now, let the Poincaré section coincide with the abscissa; then $f(x)$ is the return mapping (that's why "flow over the mapping"!), and $F(x)$ is the return time. If $F(x)$ is continuous and bounded, the spectral properties of the flow are adequately represented by characteristics of $f(x)$. However, a singularity in $F(x)$ becomes a source of discrepancy. Von Neumann demonstrated that for $f(x)$ with a discrete spectrum, a discontinuity (a finite "jump") in $F(x)$ can generate the continuous component in the spectrum of the flow [16]. In a system of differential equations with a smooth right hand side, one would expect a divergence of return time rather than a finite discontinuity; this is what happens in (3) for passages close to the saddle point at the origin.

To demonstrate this effect, we compute the spectral characteristics for the special flow over $f(x)$ which corresponds to the return mapping (5) at the

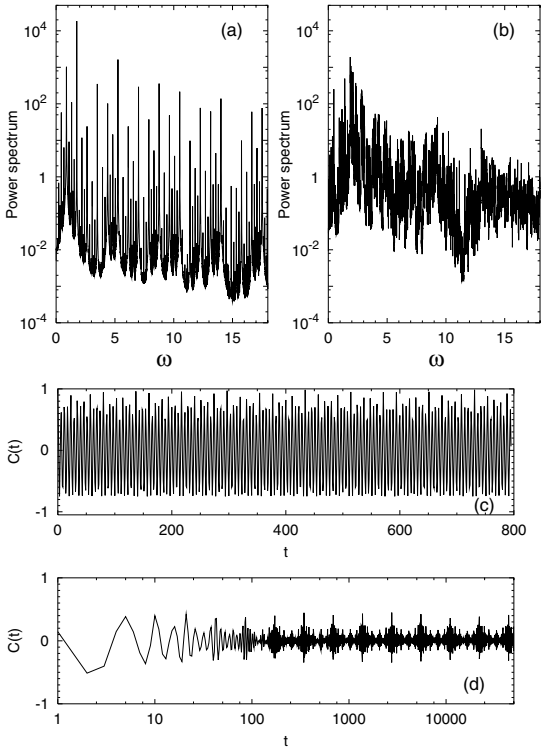

Fig. 8. Power spectra and autocorrelations for the special flow over the mapping $x_{i+1} = 1.40155189 - x_i^2$. Sample length 2^{17}. **a,c** bounded return time, **b,d** unbounded return time

accumulation of period-doubling bifurcations. We use two different functions $F(x)$. The first of them, $F_1(x) = 3 + 0.08x - \frac{1}{2}\sin x$ is bounded on the interval of x which includes the attractor. The second one, $F_2(x) = 2 - \ln|x|$ has a singularity at the point $x = 0$ which belongs to the attractor. The computed estimates of the power spectra and the autocorrelation functions are shown in Fig. 8.

The power spectrum of the flow under the bounded function in the left panel displays the well ordered structure of δ-like peaks at the main frequency, its subharmonics and harmonics. The plot shows the averaged estimate for sample of length 2^{17}; if the length is increased, these peaks become sharper and thinner and the troughs between them deeper and broader. Results for the flow under the function with a singularity are qualitatively different (right panel): the contrast between the highest peaks and the deepest minima is not so sharp. The whole pattern appears to be much more dense and (at least optically) "much more continuous".

The difference is more striking in the time domain, where we can compare autocorrelations for these two different kinds of $F(x)$. Panel (c) of Fig. 8

shows autocorrelation for the special flow without the singularity; the time-scale is conventional, and the returns of autocorrelation arbitrarily close to 1 are distinctly seen: The dynamics is "nearly periodic", exactly as for the mapping (5) itself. In contrast, the special flow with a singularity possesses an autocorrelation (lower panel) whose pattern recalls the autocorrelations of the Thue–Morse sequence and of the continuous variable $z(t)$ from (3) (cf. Fig. 6). This autocorrelation indicates to the presence of the singular continuous spectral component.

This example confirms that the special attractor geometry reducible to the particular (Thue–Morse and alike) symbolic code, is not a necessary precondition for the existence of fractal spectral measure: a singularity in the return times can be responsible for this effect as well. If both mechanisms are acting simultaneously, as in the case of the observable x from (3), the second one distorts the symmetric pattern of autocorrelation typical for the first one. We are in an unexpected situation: the flow described by (3) was brought in as an example of the continuous dynamical system in which the Thue–Morse symbolic code stood behind the fractal spectral component. Now we see that there is some redundancy: not only is this component present at the accumulation of every scenario of gluing bifurcations, it also appears in the absence of the mirror symmetry, when the symbolic codes themselves are different from the Thue–Morse sequence. The only thing which matters are the repeated passages arbitrarily close to the equilibrium point.

4 Fractal Spectra in Laminar Hydrodynamics

Currently we are unaware of examples of time-continuous dynamical systems where the symbolic code is given by the Thue–Morse or similar sequence, but where the return times are bounded. In contrast, examples of flows with singularities of return times (e.g. saddle points embedded into the attractors) are abundant. Indeed, this does not even require the 3-dimensional phase space: two spatial dimensions are quite sufficient. Notably, on the *plane* the repeated alternating passages close by and far away from the same point are forbidden by geometry; however, they are quite possible on the surface of a 2-dimensional *torus*. This brings us back to the problem of the spectral properties of flows on 2-tori, posed and solved by A.N.Kolmogorov in 1953 [1]: now we see that in his proof of the discrete spectrum the requirement of the absence of fixed points is not a mere technical condition. Violation of this condition (the existence of points, where the vector field vanishes) can create the (singular) continuous component in the spectral measure.

In order to demonstrate this, we leave the somewhat abstract geometry of phase spaces and refer to the "physical" geometry of trajectories of the particles carried by the time-independent flow of a viscous fluid. Again, the starting point is the problem which was proposed by Kolmogorov as an exemplary candidate for studying hydrodynamical instabilities and transition

to turbulence: in his seminar of 1959 the participants discussed the motion of a viscous incompressible fluid which is disturbed by the spatially periodic force acting in one direction (i.e. by the force which is directed along the y-axis and is proportional to $\cos x$) [17]. Experimentally such forcing can be (and has been) implemented e.g. in electroconducting fluids by positioning the lattice of electrodes on the bottom. This class of fluid motions, known as the "Kolmogorov flow", remains one of the most often used examples of hydrodynamical transitions.

Here, we consider a modification of this problem [18]. Let the force act along two spatial directions: the y-component of the force is proportional to $\cos x$ whereas its x-component is proportional to $\cos y$. We restrict ourselves to two-dimensional steady (time-independent) flows for which the characteristics (velocity $\boldsymbol{v}$, pressure p) at each physical point remain constant. Further, since the force is periodic, we require the velocity field to have the same periodicity: 2π in each direction. In this way, the problem is formulated as a fluid motion in a square with periodic boundary conditions; equivalently, this can be viewed as the motion on the surface of a 2-torus. We also assume that there is some constant mean drift across the square (ensured e.g. by the constant pressure gradient): its components in the x- and y-direction are β and α, respectively. Accordingly, the direction of the mean flow forms the angle $\arctan \beta/\alpha$ with the x-axis.

The fluid is assumed to be incompressible. Mathematically, this is expressed through the condition $\mathbf{div}\ \boldsymbol{v} = 0$. Physically, this means that the volume (the area in case of the 2-dimensional flow) of an arbitrary element of fluid is time-invariant: the element can move, its shape can be distorted by the flow, but the overall volume remains the same. This condition allows us to relate the velocity to the so-called "stream function" $\Psi(x, y)$: $v_x = \partial\Psi/\partial y,\ v_y = -\partial\Psi/\partial x$. The stream function "visualizes" the flow pattern: in a steady flow the fluid particles move along the isolines of $\Psi(x, y)$.

The velocity field which obeys the hydrodynamical equation reads as

$$v_x = \alpha - \frac{f\ \cos(y + \phi_2)}{\sqrt{\beta^2 + \nu^2}}\ , \qquad v_y = \beta - \frac{f\ \cos(x + \phi_1)}{\sqrt{\alpha^2 + \nu^2}}\ , \tag{6}$$

where f is the amplitude of the force, ν is the viscosity of the fluid, $\phi_1 = \arctan \frac{\nu}{\alpha}$ and $\phi_2 = \arctan \frac{\nu}{\beta}$. At $f = 0$, in the absence of forcing, the velocity is everywhere the same, and the motion of the fluid is simply the homogeneous drift; mathematically, this is a linear flow on the torus with the winding number α/β. If this number is irrational, every streamline is dense on the surface of the torus; in other words, each passive particle transported by the flow eventually passes arbitrarily close to any given place.

An increase in the forcing amplitude f distorts the streamlines (Fig. 9a). Qualitatively however, the flow pattern does not change until, on reaching the threshold value of the forcing amplitude

$$f = f_{cr} = \sqrt{\alpha^2\beta^2 + \nu^2 \max(\alpha^2, \beta^2)}\ , \tag{7}$$

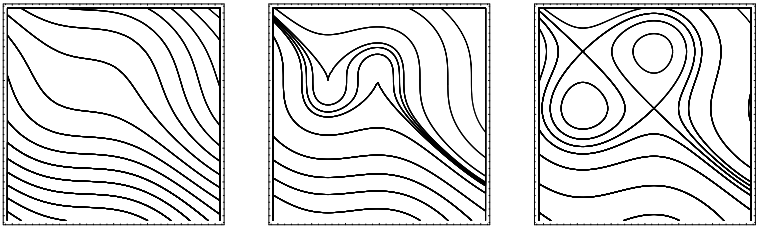

Fig. 9. Flow patterns for different values of the forcing strength. Isolines of the stream function for: **a** $f = 0.5f_{cr}$; **b** $f = f_{cr}$, (c) $f = 1.5f_{cr}$. $\alpha = 1$; $\beta = -0.6018$, $\nu = 1$

two turning points appear on the streamlines (Fig. 9b). Further increase in f results in a qualitative transition, as two symmetric stationary vortices emerge inside the flow. As seen in Fig. 9c, outside the vortices the flow retains the globally drifting component, where each streamline is dense. There is an elliptic stagnation point at the center of each vortex; the boundary of each vortex is formed by the trajectory (called the "separatrix") which starts and ends at the same stagnation point of the saddle type. Of course, from the dynamical point of view this boundary is just a homoclinic trajectory. The time which one trajectory needs for one passage around the torus (we can view this as the return time onto the boundary of the square) diverges when the initial conditions are chosen infinitesimally close to the trajectory which exactly hits the saddle stagnation point. On the other hand, the irrational rotation number ensures that each trajectory of the global component passes arbitrarily close to the stagnation point. Therefore the condition that was discussed in the previous section is fulfilled: there is a region of the phase space where the return time diverges, and there is a mechanism which ensures that the orbits repeatedly visit this region. Therefore, it is reasonable to check for the fractal spectral component.

The choice of the appropriate observable may deserve a short discussion. Apparently, it makesf no sense to measure any characteristic of the flow at a fixed location of the phase space: since the flow is time-independent, at any given place all characteristics remain constant. Instead one can attach the reference frame to the tracer particle floating along the streamlines: with this particle the measuring device will visit the regions of relatively high and relatively low velocity and explore the entire domain outside the vortices. In fluid mechanics such description, associated with the moving fluid particle, is known as "Lagrangian description". For a particle carried by the flow, a convenient observable is either of the velocity components.

The following plots confirm that the birth of vortices and stagnation points at f_{cr} marks a drastic transition in the spectral and correlation properties: as shown in Fig. 10a, for weak forcing the spectrum is apparently discrete whereas for $f > f_{cr}$ (Fig. 10b) it possesses a well-pronounced self-similar structure, typical for fractal sets.

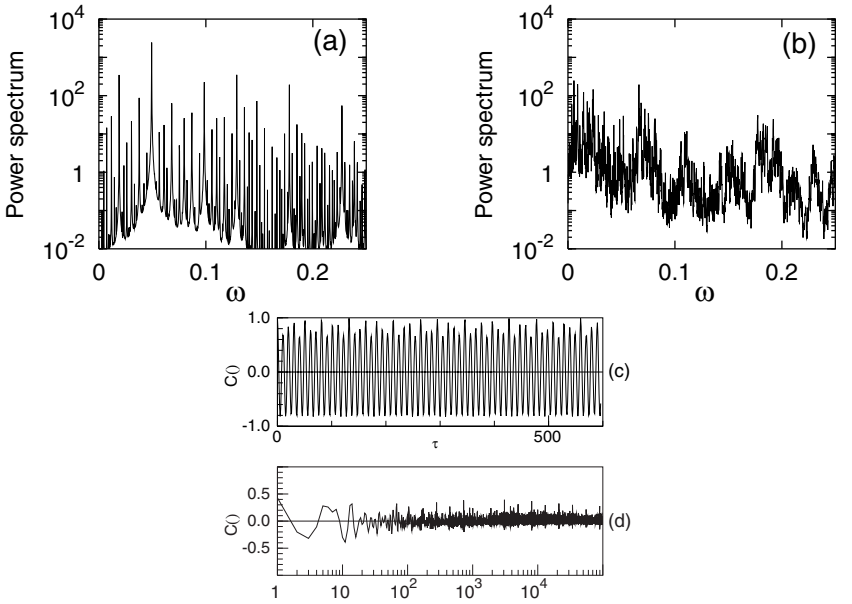

Fig. 10. Estimates of power spectrum and autocorrelation for the tracer velocity. **a,c** $f = 0.5 f_{cr}$; **b,d** $f = 1.5 f_{cr}$

In terms of autocorrelation, the flow without the vortices is characterized by a nearly periodic sequence of peaks with height tending to 1 (Fig. 10c). After the birth of the vortices, this picture is replaced by a lattice where the largest peaks have moderate height and the time intervals to which they correspond form an approximate geometric progression.Thus, the presence of stagnation points, combined with irrational inclination of the mean drift ensures that the Fourier spectrum is fractal.

Extending our excursion into fluid mechanics, let us demonstrate another example of two-dimensional motion of viscous fluid with the stagnation effect: the time-independent flow past a periodic array of solid obstacles [19]; the typical flow pattern is shown in Fig. 11. Due to viscosity, the velocity field identically vanishes along the whole borderline of each obstacle. In comparison with the previous example where the velocity vanished only in isolated stagnation points, we find that the singularity in the return time is much stronger (mathematically, the logarithmic singularity is replaced by the inverse square root). As a result, the spectrum of the tracer velocity is (of course!) fractal, and the autocorrelation function decays: the peaks become progressively smaller and eventually vanish (Fig. 12).

In the real space, the presence of a continuous spectral component has far-reaching implications: this steady laminar flow is mixing ! This property is illustrated by Fig. 13 which shows the evolution of the initially round droplet consisting of 10^4 dye particles. For convenience, here we project the position

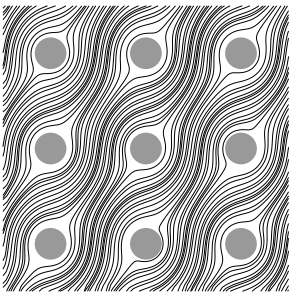

Fig. 11. Steady flow past the lattice of cylinders

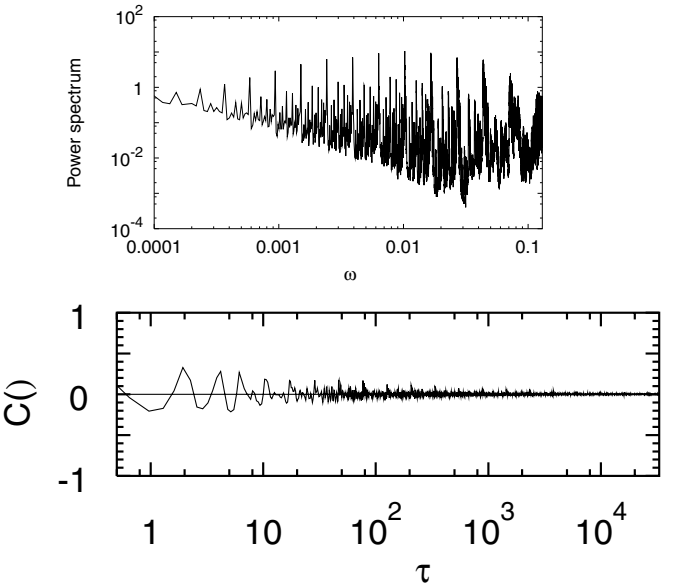

Fig. 12. Power spectrum and autocorrelation for the flow past the lattice of cylinders

of each particle onto a square, "returning" from the plane to the torus. Some particles "stick" for a long time near the obstacle border whereas the other ones pass further from this border and move on much faster. Due to the irrational inclination of the flow, each particle is subjected to repeated slowdowns in the course of its evolution; as a result, after some time the particles get spread over the whole surface of the square (Naturally, they do not *cover* the whole surface of the square: recall the conservation of volume!). Qualitatively, this picture resembles chaotic mixing: a time-periodic two-dimensional flow can possess chaotic Lagrangian trajectories [20,21], and in this case the mixing of an initial droplet of particles is very efficient. The mixing in our

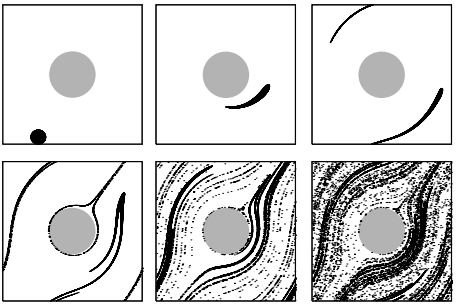

Fig. 13. Mixing in the flow past the lattice of cylinders

example, although producing ultimately the same result, is much slower and less efficient than the mixing in turbulent or chaotic flows – which is not so surprising in view of the full laminarity of stream lines.

5 Conclusion

In this article we have described some nontrivial examples of complex dynamical systems. Complexity in this context does not mean unpredictability and closeness to randomness which characterize true chaos, but rather a nontrivial interplay of dynamical properties allowing one to place such systems between order and chaos. The most visible manifestations of this complexity are the nontrivial spectral properties of the processes: the power spectrum is neither discrete like for the ordered dynamics, nor absolutely continuous like for the chaotic one, instead it is singular continuous, or fractal.

Remarkably, many approaches that are used in the studies of complex dynamics stem from works of Kolmogorov and his school. Of principal importance in the distinction between order and chaos is the concept of the Kolmogorov complexity of symbolic sequences. As we have shown for the Thue–Morse sequence, even simple symbolic sequences can have nontrivial correlations. Another important concept, widely used by Kolmogorov and his scholars (probably introduced first by J. von Neumann), concerns the relation between continuous and discrete time dynamical systems. Mathematically it is represented by the special flow construction. The main idea here is that in the usual Poincaré map construction it is important to trace information on the return times for the trajectories. Solely by changing the properties of the return times, one can destroy correlations in the dynamics. Finally, we have demonstrated that the nontrivial dynamics can be observed in simple two-dimensional noncompressible fluid flows of the type first introduced by Kolmogorov. His original idea was to build the simplest model mimicking transition to turbulence in flows. Rather surprisingly, already in a laminar regime the particles in such a flow can be advected in a quite complex manner.

References

1. A.N. Kolmogorov, On dynamical systems with an integral invariant on a torus, Dokl. Akad. Nauk SSSR Ser. Mat. **93**, 763–766 (1953)
2. R. Badii and A. Politi, *Complexity. Hierarchical structures and scaling in physics*. Cambridge University Press (1997)
3. A. Thue, Über unendliche Zeichenreihen, Norske vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania **7**, 1–22 (1906)
4. M. Morse, Recurrent geodesics on a surface of negative curvature, Trans. A.M.S. **22**, 84–100 (1921)
5. N. Wiener, The spectrum of an array, J. Math. and Phys. **6**, 145–157 (1926)
6. K. Mahler, On the translation properties of a simple class of arithmetical functions, J. Math. and Phys. **6**, 158–163 (1926)
7. A.N. Kolmogorov, Three approaches to the quantitative definition of information, Problems of Information Transmission, **1**, 4 (1965)
8. M.A. Zaks, A.S. Pikovsky, J. Kurths, On the correlation dimension of the spectral measure for the Thue–Morse sequence, J. Stat. Phys. **88**, 1387–1392 (1997)
9. R. Ketzmerick, G. Petschel, T. Geisel, Slow decay of temporal correlations in quantum systems with Cantor spectra, Phys. Rev. Lett. **69**, 695–698 (1992)
10. A.S. Pikovsky, M.A. Zaks, U. Feudel, J. Kurths, Singular continuous spectra in dissipative dynamical systems, Phys. Rev. E **52**, 285–296 (1995)
11. D.V. Lyubimov, M.A. Zaks, Two mechanisms of the transition to chaos in finite-dimensional models of convection, Physica D **9**, 52–64 (1983)
12. E.N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci., **20**, 130–141 (1963)
13. A. Arneodo, P. Coullet, C. Tresser, A possible new mechanism for the onset of turbulence, Phys. Lett. A **81**, 197–201 (1981).
14. C. Sparrow, *The Lorenz equations: Bifurcations, chaos, and strange attractors* (Springer, 1982).
15. M.J. Feigenbaum, The transition to aperiodic behaviour in turbulent systems, Comm.Math.Phys. **77**, 65–86 (1980)
16. J. von Neumann, Zur Operatorenmethode in der klassischen Mechanik, Ann. Math. **33**, 587–642 (1932)
17. V.I. Arnold, L.D. Meshalkin, Seminar led by A.N. Kolmogorov on selected problems of analysis (1958-1959), Usp. Mat. Nauk. **15**, 247 (1960)
18. M.A. Zaks, A.S. Pikovsky, J. Kurths, Steady viscous flow with fractal power spectrum, Phys. Rev. Lett. **77**, pp. 4338–4341 (1996)
19. M.A. Zaks, A.V. Straube. Steady Stokes flow with long-range correlations, fractal Fourier spectrum and anomalous transport, Phys. Rev. Lett. **89**, pp. 244101–244104 (2002)
20. J.M. Ottino, *The Kinematics of Mixing: Stretching, Chaos, and Transport*. Cambridge University Press (1989)
21. T. Bohr, M.H. Jensen, G. Paladin, A. Vulpiani, *Dynamical System Approach to Turbulence*. Cambridge University Press (1998)

Kolmogorov's Legacy about Entropy, Chaos, and Complexity

Massimo Falcioni[1], Vittorio Loreto[2], and Angelo Vulpiani[3]

[1] University of Rome "La Sapienza", Physics Department and INFM Center for Statistical Mechanics and Complexity, P.zzle Aldo Moro, 2, Rome, Italy, 00185, massimo.falcioni@phys.uniroma1.it
[2] University of Rome "La Sapienza", Physics Department and INFM Center for Statistical Mechanics and Complexity, P.zzle Aldo Moro, 2, Rome, Italy, 00185, loreto@pil.phys.uniroma1.it
[3] University of Rome "La Sapienza", Physics Department and INFM Center for Statistical Mechanics and Complexity, P.zzle Aldo Moro, 2, Rome, Italy, 00185, angelo.vulpiani@roma1.infn.it

Abstract. The concept of entropy was initially introduced in thermodynamics in a phenomenological context. Later, mainly by the contribution of Boltzmann, a probabilistic interpretation of the entropy was developed in order to clarify its deep relation with the microscopic structure underlying the macroscopic bodies. In the unrelated field of communication systems, Shannon showed that, with a suitable generalization of the entropy concept, one can establish a mathematically self-consistent information theory. Kolmogorov realized, just after Shannon's work, the conceptual relevance of information theory and the importance of its ideas for the characterization of irregular behaviors in dynamical systems. Going beyond a probabilistic point of view, the Algorithmic Complexity (introduced by Chaitin, Kolmogorov and Solomonoff) allows a formalization of the intuitive notion of randomness of a sequence. In this chapter we discuss the connections among entropy, chaos and algorithmic complexity; in addition we briefly discuss how these concepts and methods can be successfully applied to other unrelated fields: linguistics, bioinformatics, finance.

1 Entropy in Thermodynamics and Statistical Physics

The concept of entropy made its appearance in physics to describe in a mathematical way the irreversible behaviors of the macroscopic bodies, those we experience in everyday life. The goal is achieved by the 2^{nd} principle of the thermodynamics, which can be formulated as follows [16].

There exists a function, the *entropy* S, defined for the equilibrium states of macroscopic systems, such that its difference between (equilibrium) states A and B of a given system satisfies the relation:

$$\Delta S_{AB} \equiv S(B) - S(A) \geq \int_A^B \frac{\delta Q}{T_{\text{source}}}\,, \tag{1}$$

M. Falcioni, V. Loreto, and A. Vulpiani, Kolmogorov's Legacy about Entropy, Chaos, and Complexity, Lect. Notes Phys. **636**, 85–108 (2003)
http://www.springerlink.com/

where the integral is computed along an arbitrary transformation leading from state A to state B, in which the system exchanges heat δQ with external sources at thermodynamic temperatures T_{source}. When the transformation is chosen to be reversible, relation (1) has the equality sign and may be used to compute ΔS_{AB}. In reality, irreversibility transforms relation (1) to an inequality which sets limits to what can happen. In particular, in isolated systems, where $\delta Q = 0$, a transformation $A \to B$ is allowed only if $S(B) > S(A)$.

In the case of n moles of a classical, perfect and mono-atomic gas, contained in a vessel of volume V at temperature T, one is able to write:

$$S(n, U, V) = n\, R\, \ln \left[\gamma_0 \left(\frac{U}{n} \right)^{3/2} \left(\frac{V}{n} \right) \right], \tag{2}$$

where γ_0 is a suitable constant, R is the universal gas constant and $U = \frac{3}{2}\, n\, R\, T$ is the internal energy of the gas.

If we accept the atomistic point of view [5], we put $R = N_A k$, where $N_A = 6.02\ 10^{23}$ is Avogadro's number and $k = 1.38\ 10^{-23} J/K$ is the Boltzmann constant, and we rewrite eq. (2), for $N = n\, N_A$ molecules, as

$$S(N, U, V) = N\, k\, \ln \left[\gamma_1 \left(\frac{U}{N} \right)^{3/2} \left(\frac{V}{N} \right) \right], \tag{3}$$

where γ_1 is another constant. In this formula, the specific volume, V/N, is the mean volume available for the motion of a single molecule, where we assume that molecules are indistinguishable. If we define $V/N = (\Delta x)^3$, Δx may be thought of as the typical variability of the space coordinates for each molecule. For the mean (kinetic) energy U/N one can write:

$$\frac{U}{N} = \frac{1}{2m} \left(\langle p_x^2 \rangle + \langle p_y^2 \rangle + \langle p_z^2 \rangle \right) = \frac{3}{2m} \langle p_x^2 \rangle \equiv \frac{3}{2m} (\Delta p)^2 , \tag{4}$$

where p_x, p_y, p_z are the components of the momentum of one molecule, so that Δp, the root mean-square of one component, gives the order of variability of the single molecule momentum. Then, denoting by γ_2 and γ two suitable constants, we have:

$$S = N\, k\, \ln \left[\gamma_2\, (\Delta x)^3\, (\Delta p)^3 \right] = k\, \ln \left[\frac{(\Delta x)\, (\Delta p)}{\gamma} \right]^{3N} . \tag{5}$$

Here, $[(\Delta x)\, (\Delta p)\, /\gamma]^{3N}$ represents the volume, in units of $(\gamma)^{3N}$, of the $6N$-dimensional phase space region where the motion of the system takes place.

Equation (5) is a special case of the relation between thermodynamic entropy and the extention of the motion in the phase space, the "Boltzmann principle", that one can write as:

$$S = k\, \ln \Gamma\, (N,\, U,\, V) , \tag{6}$$

with

$$\Gamma(N, U, V) = \int\limits_{U-\delta U/2 \le H \le U+\delta U/2} \frac{\mathrm{d}^{3N}x\ \mathrm{d}^{3N}p}{N!\,\mathrm{h}^{3N}} \equiv \int\limits_{U-\delta U/2 \le H \le U+\delta U/2} \mathrm{d}\Gamma_N\ . \tag{7}$$

In (7) $H = \sum_{i=1}^{N} \mathbf{p}_i^2/2m$ is the Hamiltonian function (the energy) of the system, U is its actual value, known within an uncertainty δU, h is a constant that in classical mechanics is indeterminate but that, with the help of quantum mechanics, may be taken to be the Planck's constant. The term $N!$ takes into account the indistinguishability of the molecules of the gas. So $\Gamma(N, U, V)$ measures the amount of freedom the system has in its phase space. This is the starting point for a statistical mechanics description of the thermodynamics. According to the interpretative rules of statistical mechanics,

$$\frac{\mathrm{d}\Gamma_N}{\Gamma(N, U, V)} \tag{8}$$

is the probability to find the system, in equilibrium with energy around U, in a volume $\mathrm{d}^{3N}x\ \mathrm{d}^{3N}p$ of its phase space.

It is interesting to observe that eq. (8) defines $1/\Gamma(N, U, V)$ as the (uniform) probability density of finding a system in equilibrium in anyone of its allowed microscopic states, the micro-canonical distribution:

$$\rho(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_N, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_N) = 1/\Gamma(N, U, V)\ , \tag{9}$$

and we can write

$$\frac{S}{k} = \ln \Gamma = \int \mathrm{d}\Gamma_N\ \rho\ (-\ln \rho) \equiv \langle -\ln \rho \rangle\ . \tag{10}$$

This is a possible and useful definition of the entropy in statistical mechanics, even when the equilibrium states are described by a density ρ, not necessarily the micro-canonical one (Gibbs entropy),

$$S = -k\ \langle \ln \rho \rangle. \tag{11}$$

2 Entropy in Information Theory

The concept of entropy enjoyed a renewed consideration in the scientific community when, at the end of 1940s, in a very different context, C.E. Shannon [22] adressed the problem of an efficient transmission of messages, thus laying the foundations of a mathematical theory of communication. Among the various results obtained by Shannon, the following is of interest for us.

Each particular message to be transmitted is thought to belong to the ensemble of all the messages one source can generate. The source is able to emit M different symbols, constituting the so-called *alphabet*, and every message is a sequence of N letters that we indicate with $(s_1, s_2, s_3, \ldots, s_N)$ or s^N, where each s_i is one of the M symbols. One supposes that every message may be produced with a probability $p(s^N)$. Usually the signal sent through a transmission channel is a sequence of $m\,(\neq M)$ possible physical states; we assume a binary channel, i.e. a channel operating with two states that can be named 0 and 1. At this point one has the coding problem: one needs to transform every M-ary sequence into a binary sequence, in the most economical way, i.e. such that the average length of the coding sequences be minimum (one also wants to recover the original message so the code must be invertible). Shannon was able to show that it is possible to associate each message s^N with a sequence of $\ell(s^N)$ binary symbols, i.e. of $\ell(s^N)$ bits, where

$$\ell(s^N) \approx \log_2 \frac{1}{p(s^N)} \ , \tag{12}$$

thus obtaining a (quasi) optimal coding. So one has that $\log_2 1/p(s^N)$ is a measure of the resources needed to send a particular message. An interesting result is obtained when considering long messages, the $N \to \infty$ limit. In this case one introduces the quantity, called Shannon entropy:

$$h = \lim_{N\to\infty} \frac{1}{N} \langle \ell \rangle_N = \lim_{N\to\infty} \frac{1}{N} \sum_{s^N} p(s^N) \log_2 \frac{1}{p(s^N)} \tag{13}$$

that represents the average number of bits needed to code a source symbol. The importance of h is clarified hereafter. According to the Shannon-McMillan theorem [12], for $N \to \infty$ the ensemble of sequences s^N divides up into two sets:

- the set of sequences with $p(s^N) \approx 2^{-N\,h}$, that are said the typical sequences;
- the set containing all the other sequences.

The set of typical sequences, the ones that are "reasonably probable", has $\mathcal{N}(N) \approx 2^{N\,h}$ elements. Moreover the probability of this set is ≈ 1; while the probability of the remaining set is ≈ 0. This means that the typical sequences almost exhaust the total probability and constitute a set of $\mathcal{N}(N)$ practically equiprobable elements, with $p(s^N) \simeq 1/\mathcal{N}(N)$. The Shannon entropy gives, at the same time, both the number of bits per symbol that are necessary to code each one of these (long) sequences and the rate of growth of their number with the sequence length. From this point of view, writing $h = (1/N) \log \mathcal{N}(N)$ and comparing Eqs. (13) and (11), one realizes that h is the analogue of the entropy per particle in statistical mechanics: $S/N = (1/N)k \ln \Gamma$, hence the name. Therefore, for an efficient transmission of long messages, one must be able to forward h bits per emitted source symbol. This is the same as

identifying, with $N \cdot h$ bits, one among the $\mathcal{N}(N)$ substantially equiprobable messages that one can reasonably expect [22].

The statistical mechanics entropy (Boltzmann entropy) $S = k \ln \Gamma$ and the analog quantity in information theory $h\,N = \log_2 \mathcal{N}(N)$ may be seen as two measures of the amount of uncertainty in the associated system: this uncertainty grows as either Γ (the number of accessible microstates) or $\mathcal{N}(N)$, (the number of messages that can potentially be transmitted) grows. From this point of view, entropy may also be seen as a measure of the quantity of information one obtains after the state of the system has been determined or the message has been transmitted (and received). Given a probabilistic scheme, assigning probabilities p_i $(i = 1, \ldots, M)$ to the M occurrences of an event x, the expression:

$$\sum_{i=1}^{M} p_i \ \log_2 \frac{1}{p_i} = \langle -\log_2 p \rangle \tag{14}$$

defines the entropy or uncertainty of the event, that we indicate with $H(x)$. This quantity indeed can be thought of as the average uncertainty on the event before it happens, or the average amount of information one gets from a single actual realization. It satisfied $0 \leq H \leq \log_2 M$; one has $H = 0$ if all p_i but one are zero, since in this case there is no uncertainty at all (and no new information from observing the certain event). On the other hand the maximum value both of uncertainty and of information, $H = \log_2 M$, is reached when all occurrences are equiprobable: $p_i = 1/M$. In the case when x is a symbol of a source of messages, $x = s^1$,

$$H(x) = \sum_{s^1} p(s^1) \ \log_2 \frac{1}{p(s^1)} \equiv H_1 \tag{15}$$

is the average uncertainty on the emission of one symbol. It is possible to show that the quantity

$$\frac{1}{N} \sum_{s^N} p(s^N) \ \log_2 \frac{1}{p(s^N)}$$

may be read as the mean uncertainty on the emission of the N^{th} symbol of a sequence, given that the preceding $N-1$ are known. Therefore we may interpret h as the mean residual uncertainty on the appearance of an additional symbol in a sequence, after the information contained in the (infinitely long) past history has been acquired. If time correlations exist in the sequences, so that some constraint on the future is given by the past, the mean uncertainty becomes lower. So, in general, we have $h \leq H_1 \leq \log_2 M$.

In the coding problem of information theory, one wants to find an (economical) invertible relation between different sequences that represent the original message and the transmitted one. If one thinks about the sequences

as describing evolutions of dynamical systems (see next subsection), one easily realizes that information theory concepts may play a role in problems of dynamical systems. For instance in the isomorphism problem, where one wants to characterize the conditions under which one is sure that two formally different systems are actually the same. This implies that an invertible relation exists which preserves the probabilistic and dynamical relations among their respective elements. In facing this problem Kolmogorov defined a quantity that is known as the *Kolmogorov-Sinai entropy* of a dynamical system [14,24].

The above results have been obtained by supposing that the message source has well defined probabilistic rules, such that each message has a certain probability to appear. The effect of this probability distribution is to generate a selection among all the M^N *a priori* possible messages, reducing them to the set of the typical sequences that are "reasonable" from a probabilistic point of view. In cases where a message is generated by unknown rules, one can search for the most economical, i.e. the shortest, (binary) code to transmit it. Thus one poses the problem of a "universal coding", depending only on "intrinsic" properties of the symbol sequence. To find the answer, Kolmogorov introduced the idea of *Algorithmic Complexity* (AC) (also known as *Kolmogorov complexity*) [15,6,25].

3 Entropy in Dynamical Systems

Let us now discuss the relevance of information theory in the context of the deterministic systems. The idea is due to A.N.Kolmogorov who, just after the Shannon work, realized the conceptual importance of the information theory beyond its practical use in communications [13,14].

For the sake of self-consistency, we introduce some basic elements on chaotic dynamical systems. A dynamical system is defined as a deterministic rule for the time evolution of the state (described by a d-dimensional vector $\mathbf{x}$). Well-known examples are the ordinary differential equations:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t)) \tag{16}$$

and the discrete time maps:

$$\mathbf{x}(n+1) = \mathbf{g}(\mathbf{x}(n))\,, \tag{17}$$

where $\mathbf{x}, \mathbf{f}, \mathbf{g} \in R^d$. Examples of regular behavior of the solutions of ordinary differential equations are stable fixed points, periodic or quasi-periodic motions.

After the seminal work of Poincaré, and later Lorenz, Hénon and Chirikov (to cite only some of the most eminent ones) the existence of visibly irregular non periodic behaviors, that are called chaotic, is now well established [20].

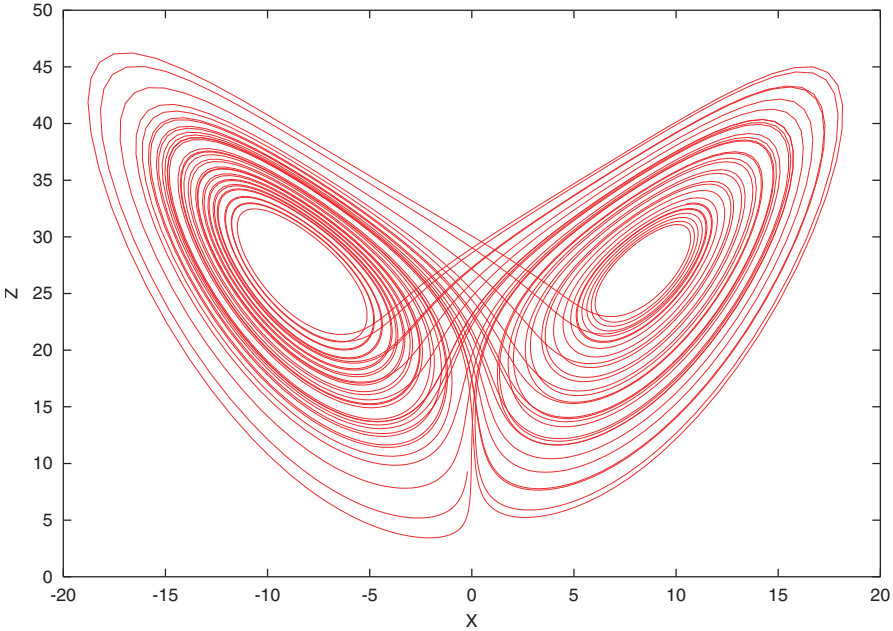

Fig. 1. Lorenz's model: Projection of a trajectory on the $x\,z$ plane for $r = 28$, $\sigma = 10$ and $b = 8/3$

In Fig. 1 one can see the time evolution of the celebrated Lorenz's model [19]:

$$\begin{cases} \dot{x} = -\sigma(x - y) \\ \dot{y} = -xy + rx - y \\ \dot{z} = xy - bz \end{cases} \tag{18}$$

where the dot means the time derivative.

For the purposes of this chapter, the most relevant feature of chaotic dynamics is the so-called sensitive dependence on initial conditions. Consider two trajectories, $\mathbf{x}(t)$ and $\mathbf{x}'(t)$, initially (say, at time $t = 0$) very close; in chaotic systems we find that the distance between them, $|\delta\mathbf{x}(t)| = |\mathbf{x}(t) - \mathbf{x}'(t)|$, increases exponentially in time as $|\delta\mathbf{x}(0)| \to 0$ and $t \to \infty$:

$$|\delta\mathbf{x}(t)| \sim |\delta\mathbf{x}(0)|e^{\lambda_1 t}\,, \tag{19}$$

λ_1 is called first (or maximal) Lyapunov exponent, see Fig. 2. The exponent λ_1 characterizes the typical trajectories of the system, just like the Shannon entropy characterizes the typical sequences of an ensemble of messages. The sensitive dependence on the initial conditions (i.e. $\lambda_1 > 0$) is a practical definition of deterministic chaos, since it implies that, in presence of any uncertainty of the knowledge on the initial state, the system is unpredictable

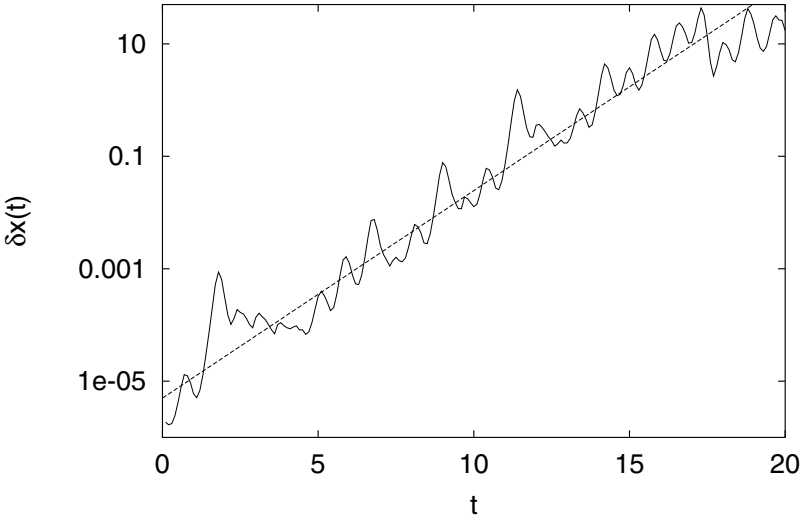

Fig. 2. Lorenz's model: $|\delta\mathbf{x}(t)|$ vs t for $r = 28$, $\sigma = 10$ and $b = 8/3$

for long times. If we know the initial conditions with a finite, but small, precision $\delta_0 = |\delta\mathbf{x}(0)|$ and we want to predict the state with a certain tolerance δ_M then, from eq. (19), we find that our forecast cannot be pushed to times larger than T_P, called predictability time:

$$T_P \sim \frac{1}{\lambda_1} \ln \frac{\delta_M}{\delta_0} \,. \tag{20}$$

Because the logarithm varies in an extremely slow way, T_P is essentially determined by λ_1.

A more complete characterization of the instability properties of the trajectories of a chaotic system is given by the set of Lyapunov exponents $\lambda_1, \lambda_2, ...\lambda_d$ (where conventionally the labeling is such that $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_d$) defined as follows. Consider a small d-dimensional ball of radius ϵ with center $\mathbf{x}(0)$. Under the evolution (16) or (17), at time t the ball will be deformed into a d-dimensional ellipsoid with semi-axes $l_1(t) \geq l_2(t) \geq ... \geq l_d(t)$. The Lyapunov exponents are defined as

$$\lambda_i = \lim_{t\to\infty} \lim_{\epsilon\to 0} \ln \frac{l_i(t)}{\epsilon} \,. \tag{21}$$

An important consequence of the instability with respect to initial conditions is that the time sequences generated by a chaotic system, from the point of view of information theory, have the same properties of genuine stochastic processes. This is another way to view unpredicatability. In order to discuss this idea, let us consider the example of a one-dimensional map, eq. (17) with $d = 1$, the celebrated logistic map at Ulam's point:

$$x(t+1) = 4x(t)(1 - x(t)) \,. \tag{22}$$

It is possible to show that such a system, according to our definition, is chaotic since $\lambda_1 = \ln 2$ [20]. Consider a trajectory $\{x(0), x(1), ..., x(T-1)\}$ and the symbolic sequence $\{i_0, i_1, ..., i_{T-1}\}$ obtained by taking $i_k = 0$ if $x(k) \leq 1/2$, otherwise $i_k = 1$. Let us now study the Shannon entropy h of the ensemble of sequences so generated and its relation with the maximal Lyapunov exponent λ_1. Under rather general conditions, almost always satisfied by the commonly investigated dynamical systems, h can be numerically estimated by analyzing one very long (infinitely long, in principle) sequence as follows. Denote with W_m a m-long subsequence $\{i_j, i_{j+1}, ..., i_{j+m-1}\}$; from its frequency of apparition in $\{i_0, i_1, ..., i_{T-1}\}$, determine its probability $P(W_m)$ and finally compute [1]

$$\lim_{m\to\infty} \frac{1}{m} \sum_{W_m} P(W_m) \ \ln \frac{1}{P(W_m)} = h\,.$$

Note that in dynamical systems theory, entropy is usually defined with the natural logarithm, while in information theory the binary log is preferred. In the above exemple (22), one obtains $h > 0$ and indeed, in this case, one can exactly determine $h = \ln 2$. This coincidence of values for λ_1 and h is not accidental. The results of Sect. 2 show that the number of bits N_T necessary to specify one of these sequences of length T is

$$N_T \simeq T\,\frac{h}{\ln 2}\,. \tag{23}$$

On the other hand, we know that in the presence of chaos, see Eq. (19), $|\delta x(t)| \sim |\delta x(0)| \exp(\lambda_1 t)$. Therefore, in order to be able to specify i_{T-1}, one needs $|\delta x(0)|$ small enough such that $|\delta x(T-1)| \leq 1/2$. Namely $|\delta x(0)| \leq (1/2)\exp(-\lambda_1 T)$, which means that one has to use a number of bits $N_T \simeq \ln(1/|\delta x(0)|)/\ln 2 \simeq T\,\lambda_1/\ln 2$. From these considerations, comparing the above result with Eq. (23), one can conclude that, in a one-dimensional map, the maximal Lyapunov exponent has also an entropic meaning, i.e. it gives (to within a factor $\ln 2$) the number of bits per unit time necessary to specify a trajectory.

Consider again the Lorenz model (which has a unique positive Lyapunov exponent) shown in Fig. 1. The chaotic trajectory evolves turning around two unstable fixed points, C_1 and C_2, with an average period $\langle\tau\rangle$. Let us introduce a binary sequence $\{i_1, i_2, \dots\}$ as follows. For each turn around C_1, put $i = 1$, otherwise $i = 0$. If one computes the Shannon entropy h of such a sequence, one finds the nice relation

$$\frac{h}{\langle\tau\rangle} = \lambda_1\ , \tag{24}$$

[1] Strictly, the estimated probability P depends on T and one would have to let T tend to infinity before letting m tend to infinity. in practise, T is given (finite) and $\lim_{m\to\infty}$ only means taking m large (but still small with respect to T to get a good statistical estimate of $P(W_m)$).

again providing a link between the Lyapunov exponent of chaotic trajectories and the Shannon entropy of symbolic sequences generated by the system.

At first glance one might believe that the above procedure, in which one associates a trajectory (given by real numbers) of a chaotic system to a sequence of integer numbers, is too crude and some relevant features are suppressed. An important result is that, in presence of (unavoidable) finite precision, this reduction may be realized with no suppression of information, in the case of deterministic systems. Previously we saw that, because of the sensitive dependence on the initial conditions, the uncertainty on the initial state implies the practical impossibility of a detailed description of the trajectory, in chaotic systems. Therefore a coarse-grained description appears rather natural.

For the sake of simplicity, we consider a discrete time system (i.e. a map). To model finite precision measurements, we introduce an ϵ-partition of the phase space with $\mathcal{N}$ disjoint sets (hypercubes of edge ϵ): $A_1, A_2, ..., A_{\mathcal{N}}$. From any initial condition $\mathbf{x}(0)$, one has a trajectory $\{\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(T)\}$. This trajectory identifies a symbolic sequence $\{i_1, i_1, ..., i_T\}$, where i_j $(1 \leq i_j \leq \mathcal{N})$ means that $\mathbf{x}(j)$ is in the cell A_{i_j}.

In such a way, for any given ϵ-partition, one can compute the Shannon entropy $h(\epsilon)$, either following the approach discussed in Sect. 2, if one considers the ensemble of the trajectories, or as described above, if one analyzes one single trajectory. The meaning of $h(\epsilon)$ is clear: the average number of bits per unit time necessary to specify the time evolution with a precision ϵ is nothing but $h(\epsilon)/\ln 2$. In the limit $\epsilon \to 0$, we have the so called Kolmogorov-Sinai entropy:

$$h_{KS} = \lim_{\epsilon \to 0} h(\epsilon)\,. \tag{25}$$

For deterministic systems with finite dimensional phase space, $h_{KS} < \infty$. This quantity was introduced by Kolmogorov as a way to distinguish among non isomorphic dynamical systems: h_{KS} is isomorphism invariant, so two systems with different h_{KS} are surely non isomorphic. The hope was that identity of the Kolmogorov-Sinai entropy would entail isomorphism. In general, however, this is not the case.

Just as the Shannon entropy in information theory, the Kolmogorov-Sinai entropy gives the information per unit time necessary to completely specify one typical time evolution of the system. Moreover, as remarked at the end of Sect. 2, it is also a measure of the unavoidable residual randomness in the prediction of the deterministic chaotic evolution, due to the impossibility of perfect measurements. From both points of view, one important result is that a discrete time deterministic system with finite entropy, satisfying fairly general properties, may be represented by (is equivalent to) a random process with a finite number of states; the minimum number of required states, m, depends on the Kolmogorov-Sinai entropy: $\exp(h_{KS}) < m < 1 + \exp(h_{KS})$ (*Krieger's theorem*).

In the above discussion we used a partition with hyper-cubic cells of edge ϵ, and then we took the limit $\epsilon \to 0$, so that the number of elements of the partition went to ∞. It is possible to introduce h_{KS} using more general partitions with a finite number of elements and then to take the supremum over all the partitions.

As suggested by the examples above, there is a close relation between the Kolmogorov-Sinai entropy and the positive Lyapunov exponents [20], which, in many interesting cases, can be written as

$$h_{KS} = \sum_{\lambda_i > 0} \lambda_i \ . \tag{26}$$

If one regards chaos as random behavior, this relation gives a definite and clear meaning to the above definition that a system with $\lambda_1 > 0$ is a chaotic one.

4 Algorithmic Complexity

In Sect. 2 we saw how the Shannon entropy gives a (statistical) measure of complexity by putting a limit on the possibility of compressing an ensemble of sequences. In the Algorithmic Complexity theory (introduced independently by G.Chaitin, A.N.Kolmogorov and R.J.Solomonoff) one goes further, i.e. one treats a single sequence, avoiding to use an ensemble. One can address the problem of compressing a given sequence. However, one soon realizes that a big compression is associated with a "simple structure" (little information is required to reproduce the sequence). On the other hand, a "complicated structure" requires much information to be reproduced, resulting in little compression. So, the posed problem inevitably meets with that of defining the randomness of the sequence.

Consider 3 different sequences of 24 coins flips (0 for heads and 1 for tails):

1) 000000000000000000000000 ;
2) 010101010101010101010101 ;
3) 100001001010011000111010 .

The first and the second sequences appear regular and somehow atypical. In contrast, the last one seems irregular and typical. On the other hand, probability theory does not seem able to express the notion of randomness of an individual sequence. In our case, each sequence has the same occurrence probability, i.e. 2^{-24}.

The notion of Algorithmic Complexity (or Kolmogorov complexity) is a way to formalize the intuitive notion of randomness (and regularity) of a sequence. Consider a binary digit sequence of length T generated by a computer code on some machine $\mathcal{M}$: one defines Algorithmic Complexity $K_{\mathcal{M}}(T)$ of this sequence as the bit length of the shortest program able to

generate the T-sequence and to stop afterwards. Kolmogorov was able to show the existence of universal computers $\mathcal{U}$ such that

$$K_{\mathcal{U}}(T) \leq K_{\mathcal{M}}(T) + c_{\mathcal{M}} \tag{27}$$

where $c_{\mathcal{M}}$ depends only on the computer $\mathcal{M}$. At this point, in the limit $T \to \infty$, we can define the Algorithmic Complexity per symbol with respect to a universal computer (we omit the $\mathcal{U}$-label)

$$\mathcal{C} = \lim_{T\to\infty} \frac{K(T)}{T} \,. \tag{28}$$

Let us note that as consequence of (27), $\mathcal{C}$ does not depend on the machine and it is an intrinsic quantity.

Using the above definition, one easily understands that a T-sequence of all 0 (as in the first series) or a periodic repetition of 01 (as in the second series) can be obtained with a program whose size is $O(\ln T)$ and therefore in the limit $T \to \infty$, the complexity is zero. On the other hand in an irregular sequence one expects that $K(T) \sim T$ and therefore $\mathcal{C}$ is positive and one calls the sequence complex.

Unfortunately Algorithmic Complexity cannot be computed; this impossibility is related to the Gödel incompleteness theorem [6]. Beyond this impossibility, the concept of Algorithmic Complexity has an important role to clarify the vague notion of randomness.

We can wonder whether the sequences generated by a chaotic dynamics are complex (in the sense of the Algorithmic Complexity) or not. In this respect, there exists an important relation between the Shannon entropy and the Algorithmic Complexity K_{W_T} of the T-sequences W_T of an ensemble [3,18]:

$$\lim_{T\to\infty} \frac{1}{T} \langle K_{W_T} \rangle = \lim_{T\to\infty} \frac{1}{T} \sum K_{W_T} P(W_T) = \frac{h}{\ln 2} \,,$$

where $P(W_T)$ is the occurrence probability of W_T. Putting this together with the results of the preceding section, we have that

chaotic systems produce algorithmically complex sequences.

Of course, one has that for periodic sequences $\mathcal{C} = 0$, while for a series of random variables $\mathcal{C} > 0$.

At first glance the above sentence may appear questionable, as one could (erroneously) argue as follows. Since chaos is present also in very simple systems, whose temporal evolution can be obtained with numerical codes of few instructions, a sequence of arbitrary length can be obtained with a finite number of bits, i.e. those necessary to specify the evolution rules, therefore the Algorithmic Complexity is zero. The missing point in the above naive argument is the quantity of information necessary (i.e. how many bits N_T) to specify the initial condition in such a way the code produce a certain

T-sequence. Previously, we saw that in chaotic systems for a typical initial condition one has $N_T/T \simeq \sum_i^+ \lambda_i / \ln 2$ (where $\sum_i^+ \lambda_i = h_{KS}$ is the sum of the positive Lyapunov exponents). Therefore, the Algorithmic Complexity per symbol in chaotic systems is positive and strictly related to the Lyapunov exponents, therefore to the Kolmogorov-Sinai entropy.

Let us clarify the above points with the following simple example [9]. Consider the one-dimensional map (called Bernoulli shift):

$$x(t+1) = 2x(t) \bmod 1\,,$$

with Lyapunov exponent $\ln 2$. Writing the initial condition in binary notation, $x(0) = \sum_{j=1}^{\infty} a_j 2^{-j}$, where the a_j can be 1 or 0, one immediately sees that the action of the Bernoulli shift is nothing but a shift of the binary coordinates: $x(1) = \sum_{j=1}^{\infty} a_{j+1} 2^{-j}$, $x(2) = \sum_{j=1}^{\infty} a_{j+2} 2^{-j}$, and so on. Therefore the Algorithmic Complexity of the sequence $\{i_1, i_2, ..., i_T\}$, obtained by taking $i_k = 0$ if $x(k) \leq 1/2$ otherwise $i_k = 1$, is reduced to the study of the sequence $\{a_1, a_2, ..., a_T\}$ related to the initial condition. Of course there are initial conditions $x(0)$ with zero complexity. On the other hand, a remarkable result, due to Martin-Löf, shows that almost all (in the sense of the Lebesgue measure) binary sequences $\{a_1, a_2, ..., a_T\}$ have maximum complexity, i.e. $K \simeq T$, in agreement with the fact that $\lambda_1 = \ln 2$ [18].

The above results can be summarized with a sort of slogan condensating the deep link between the unpredictability of chaotic systems (i.e. at least one positive Lyapunov exponent) with the impossibility to compress the chaotic trajectories beyond the limit given by the fact that the Kolmogorov-Sinai entropy is positive+[3]:

Complex = Incompressible = Unpredictable.

5 Complexity and Information in Linguistics, Genomics, and Finance

In this section we briefly describe some examples of disciplines for which the concepts discussed so far could represent important tools of investigation.

Let us first study the sequences of characters representing texts written in a given language. This is the most familiar example of sequences of characters which bring and transmit information. It is interesting to see how the previously stated slogan linking complexity, compressibility and predictability applies to language. Consider, without loss of generality, the English language as our case example. The entropy of the English language can be defined as the minimum number of bits per character necessary to encode an ideally infinite message written in English. In order to estimate this quantity one should be able to subtract the unavoidable redundancy which always comes along with any linguistic message. The redundancy can also be seen as the number of constraints (for instance lexical or grammatical rules) imposed on

the English text. For example the fact that a **q** must always be followed by a **u** or the impossibility to have two subsequent **h** are dependencies that make the English language more redundant. Rules of grammar, parts of speech, and the fact that we cannot invent words all make English redundant as well. Redundancy is actually benefical in order to make the message transmission efficient in *noisy* conditions or when only part of a message comes across. For example if one hears "Turn fo th lef!", one can make a fairly good guess as to what the speaker meant.

Redundancy makes language more predictable. Imagine watching a sequence of symbols emitted on a ticker-tape. The question one could ask is how much information will be added by the next symbol s_i once one already knows the sequence $s_1 ... s_{i-1}$. How much information will be gained upon seeing s_i also fixes the amount of surprise we experience. An extreme surprise will convey a large amount of information, while if one can reliably predict the next symbol from context, there will be no surprise and the information gain will be low. The entropy will be highest when you know least about the next symbol, and lowest when you know most. Shannon, with an ingenious experiment [23], showed that the entropy of the English text is something between 0.6 and 1.3 bits per character [21]. He devised an experiment where he was asking subjects to guess letters in a phrase one by one using only the knowledge of the letters previously guessed. At each step, he recorded the total number of guesses taken to reach the correct letter for that position. Shannon looked at the sequence of numbers of guesses as an encoding of the original sentence and used the entropy of the resulting random variable as an estimate for the entropy of an English letter.

More generally, estimating the complexity of a generic symbol sequence is a formidably difficult task which involves the identification and extraction of non-trivial long-range correlations inside the sequences. Kolmogorov proposed the following strategy [28]. Suppose that a subject knows the probability p_i that the next symbol in the sequence is the i-th character of the alphabet, conditioned by the appearance of the previous characters. The subject scans through the text and calculates the running averages of $-\log p_k$ where k indicates the actual character observed at each time. Kolmogorov argued that, if the probabilities p_i were correct, this average would converge to the entropy of the sequence. This approach can be seen as a natural by-product of another very interesting approach to estimate the complexity of a string which was proposed by Kelly [11], shortly after Shannon's seminal paper. This approach is related to the concept of *gambling* and it deserves a short discussion.

5.1 From Gambling to Entropy Estimate

Let us now discuss the deep relation existing between information theory and the optimal strategy in building a portfolio. In particular, compare the growth rate of wealth and the entropy rate of a generic sequence of characters, for

instance one describing the market (i.e. of the prices evolving according probabilistic rules). This apparently surprising result was originally introduced by Kelly [11] who was looking for an interpretation of the information theory outside the context of communication. Later, many authors reconsidered and generalized this approach for the growth-optimal investment strategy.

In order to illustrate the basic idea, we consider the simplest case of a discrete time price of a unique stock:

$$S_{t+1} = u_t S_t \,,$$

where S_t is the price at time t and the gain factors u_t are independent identically distributed random variables, which can assume m possible values. Let us consider an investor, with wealth S_0 at the initial time, who decides to gamble on such a stock market many times. At each time t the agent invests a fraction l_t of the capital in stock, and the rest in risk-less security, i.e. a bank account. For simplicity, we set the case of the risk-less rate to zero. It is easy to realize that at time t, the wealth S_t is given by the following multiplicative random process:

$$S_{t+1} = (1 - l_t)S_t + u_t l_t S_t = [1 + (u_t - 1)l_t]S_t \ .$$

The problem is to find the optimal strategy, i.e. the optimal values of $l_1, l_2, ...$, which maximize the capital at the final time T.

In the financial literature, the standard way to face the problem is to introduce the so-called utility function $U(S_T)$, where U is a monotonic convex function that depends on the preferences of the investor, and to maximize the average of $U(S_T)$ [8]. If T is large, using the large numbers law, one has that the exponential growth of S_T is constant with probability one. That is:

$$\lambda = \lim_{T\to\infty} \frac{1}{T} \ln \frac{S_T}{S_0} = \langle \ln[1 + (u - 1)l] \rangle = \sum_{i=1}^{m} \ln[1 + (u^i - 1)l] p_i \,,$$

where p_i is the probability to have $u = u^i$. The optimal strategy is specified by the value l^* which maximizes $\lambda(l)$. Kelly's strategy corresponds, in terms of the utility function approach, to use $U(S_T) = \ln(S_T)$.

From the maximization procedure, one obtains a new set of probabilities:

$$q_i = \frac{p_i}{1 + (u^i - 1)l^*}$$

and

$$\lambda(l^*) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{q_i} \,.$$

The quantity $\mathcal{I} = \sum_{i=1}^{m} p_i \ln p_i/q_i$ is the relative entropy (or Kullback-Leibler divergence), which is a measure of the statistical distance between two distributions $\{p_i\}$ and $\{q_i\}$ (we shall come back to the definition of relative

entropy in the coming sections). In the "symmetric" case, i.e. $m = 2$ and $u^1 - 1 = 1 - u^2$ one has the very transparent formula

$$\lambda(l^*) = \ln 2 - h$$

where $h = -\sum_i p_i \ln p_i$ is the Shannon entropy of the random process describing the stock market.

The above results show the connection between the information available from the past data and the growth rate of wealth in an optimal strategy: in the case of a process with the maximal entropy, i.e. $h = \ln 2$ the growth rate is zero. The simple case treated by Kelly can be generalized to more realistic cases with many stocks and time dependent stochastic processes, e.g. Markovian process. However, the connection between the growth rate of wealth and the rate entropy of the market is still valid (see [8] Chaps. 6 and 15).

The concept of gambling has also been applied in a linguistic context. Cover and King [7] have proposed a strategy for estimating the entropy of English language which extends Shannon's technique of guessing the next symbol. In particular, they devised a gambling strategy where the gambler bets on the outcome of the next symbol.

By generalising Kelly's strategy, they imagine a gambler with an initial capital $S_0 = 1$. At the generic time step $k+1$, i.e. after k characters $x_1, ..., x_k$ of the string are known, the gambler bets the whole capital over the d possible symbols ($d = 27$ in [7]). On every symbol, he puts a fraction of his capital, given by the probability of occurrence of this symbol conditioned to the knowledge of the k previous symbols:

$$q(x_{k+1}|x_k, ..., x_1), \tag{29}$$

with the condition $\sum_{x_{k+1}} q(x_{k+1}|x_k, ..., x_1) = 1$. This scheme is called *proportional gambling* because the whole capital is bet proportionally to the probabilities of occurrence of the possible symbols, conditioned to the knowledge of the past symbols. For each bet, the maximum possible gain is equal to d, i.e. to the cardinality of the alphabet. The capital changes recursively according to the expression:

$$S_{k+1}(x_1, ..., x_{k+1}) = dq(x_{k+1}|x_k, ..., x_1)S_k(x_1, ..., x_k), \tag{30}$$

where $k = 1, 2, ...$ It can be shown [7] that for any sequential gambling scheme $q(x_{k+1}|x_k, ..., x_1)$, one has

$$(k - E \log_d S_k(x_1, ..., x_k)) \log_2 d \geq K(x_1, ..., x_k), \tag{31}$$

where $E \log_d S_k(x_1, ..., x_k)$ indicates the expectation value of $\log_d S_k$ $(x_1, ..., x_k)$ along the sequential scheme $q(x_{k+1}|x_k, ..., x_1)$ and $K(x_1, ..., x_k)$ is the algorithmic complexity of the string $x_1, ..., x_k$. The equality holds if and only if $q(x_{k+1}|x_k, ..., x_1) = p(x_{k+1}|x_k, ..., x_1)$, $p(x_{k+1}|x_k, ..., x_1)$ being

the true conditional probability of the process. This result (which generalizes Kelly's [11]) states that an optimal gambling scheme allows for an estimate of the Algorithmic Complexity of a finite string. In the limit of an infinite sequence emitted by an ergodic source, this optimal scheme would allow for an optimal estimation of the Shannon entropy of the source. Using this scheme, implemented on humans, Cover and King estimated the entropy of English to a value on the order of 1.3 bits/symbol, in agreement with the earlier estimates by Shannon, also based on humans.

As a by-product of the previous result, it turns out that one can use the gambling scheme as a compressor for finite strings. In particular, using an alphabet of d symbols, it is possible to construct a deterministic compression scheme that allows for saving $\log_2 S_n$ bits in the coding of a sequence of n symbols. This results in a length of the compressed sequence equal to $(1 - 1/n \log_d S_n) \log_2 d$ which, in the limit of optimal gambling scheme, converges to the Algorithmic Complexity of the string. The deterministic compression scheme uses a gambler and an identical twin to the gambler who shares the same thoughts processes of the gambler. The gambler and the twin play the role of encoder and decoder respectively. We refer to [7], which contains an extensive bibliography, for the details of the compression scheme.

5.2 Looking for Relevant Information

One of the most challenging issues is represented by the huge amount of data available nowadays. While this abundance of information and the extreme accessibility to it represents an important cultural advance, one is also faced with the problem of retrieving relevant information. Imagine entering the largest library in the worlds, in search of all the relevant documents on your favorite topic. Without the help of an efficient librarian, this would be a difficult, perhaps hopeless, task [4]. The desired references would likely remain buried under tons of irrelevant volumes. Clearly, the need for effective tools for information retrieval and analysis is becoming more urgent as the databases continue to grow.

To accomplish such an ambitious task, we must determine what is useful or relevant information and where and how it is coded–say, in a written document. This is a non-trivial problem since "information" means different things in different contexts [1,30]. That is, information has no absolute value, but depends on some specific "filters" each observer imposes on his data. Consider a simple coin-toss experiment: The gambler will likely be interested only in the outcome (heads/tails) of the toss. The physicist on the other hand, might be interested in whether the outcomes reveal anything on the nature of the coin—such as whether it is honest or dishonest.

Information extraction takes place *via* a two-step process. First comes the so-called *syntactic* step, where one identifies the structures present in the messages without associating any specific meaning to them. It is only in the

second (or *semantic*) step that comprehension of meaning takes place, by connecting the syntactic information to previous experience and knowledge.

As an example of this two-step process, consider how we might identify the language in which a given text is written. In the first step we scan through the text and identify the syntactic structures: articles, verbs, adjectives, etc. But only one who "knows" the language can carry out the second phase, where the incoherent jumble of syntactic data is summarized in the specific meaning of the sentences. Other examples of this two-step process are how we recognize the subject of a given text, its historical background, and possibly its author.

The very same problem of retrieving relevant information turns out to be crucial for another field that has experienced impressive growth in the last few years: the bioinformatics.

Bioinformatics represents an emblematic example of a situation where too much information comes into play. Since it is impossible to learn everything about all living organisms, biologists *solve* the dilemma by focusing on some model organisms and trying to find out as much as they can about them. On the other hand, all available evidence indicates that the complete information necessary for making organisms is encoded in the genome, i.e. in one or several DNA molecules composed by strings of nucleotides. Having a DNA sequence is like having a text written in an unknown language, coded with an alphabet of four letters (A,C,G,T). The big challenge in this field is that the typical sizes of the few genomes completely sequenced up to now range from 10^6 to 10^{11} bases, numbers which pose two main problems: how and where to store all these data and, more importantly, how to extract relevant information from them.

When analyzing a string of characters, for instance a text, a DNA or a protein sequence, the main question is to extract the information it contains. This is known as sequence analysis or information retrieval. At present, most of bioinformatics and of the activities in the area of the so-called computational linguistics is concerned with sequence analysis.

Here are some of the questions studied in computational linguistics: automatic recognition of the language in which a given text is written, authorship attribution, automatic indexing and classification of large corpora of documents, retrieval of specific documents according to their content, reconstruction of phylogenetic trees for languages, etc..

On the other hand, current trends in bioinformatics include: (i) *gene finding.* One of the most important steps in the analysis of a new DNA sequence is finding out whether or not it contains any genes, and if so, determining exactly where they are. This process is also called segmentation; several algorithms have been devised for this problem and implemented as systems that are widely used today. (ii) *gene function prediction.* Suppose you have identified a gene. What is its role in the biochemistry of its organism? Sequence databases can help us in formulating reasonable hypotheses based for instance on homology considerations. (iii) *protein 3D structure prediction.*

The structure of the protein is directly related to the protein's functionality, probably even determining it. Moreover, 3D structure is more highly conserved than the primary structure (the sequence). (iv) *reconstruction of phylogenetic trees.* A phylogenetic tree gives a graphical representation of the evolution of contemporary species from their possible common ancestor. Typically this can be done by defining and measuring some sort of evolutionary distance and giving an estimate of the time needed for this divergence to occur.

5.3 Relative Entropy and Remoteness of Sequences

One of the most important tools in sequence analysis is the definition of suitable measures of remoteness between pairs of sequences which could be used for information retrieval. This can be done in general by looking for regularities in the sequences. The question is then how to detect such similarities? The example from bioinformatics is again enlightening. In this case a measure of similarities is obtained by means of an *alignment* procedure: roughly speaking, different sequences are made to coincide as much as they can by inserting gaps (corresponding to insertions or deletions with respect to the possible common ancestor sequence) and paying a corresponding *price*. The price to pay to align two sequences is proportional to their remoteness.

Another possible way to define the concept of remoteness between two sequences of characters (used in Linguistics or Bioinformatics) comes from the definition of Kolmogorov complexity. This concept is particularly useful for information retrieval purposes since, as mentioned above, Kolmogorov complexity is a property of a single sequence and does not refer to the existence of a source of messages. Using Algorithmic Complexity one can hope to be able to extract information from individual sequences instead of referring to the sources which are typically not known (think of a text or a DNA sequence).

In this framework an important concept to recall is that of relative entropy or Kullback-Leibler divergence [8] which is a measure of the statistical remoteness between two distributions. Its essence can be easily grasped with the following example. Let us consider two ergodic sources A and B emitting sequences of independent 0 and 1: A emits a 0 with probability p_A and 1 with probability $1-p_A$, while B emits 0 with probability p_B and 1 with probability $1-p_B$. As already described, the compression algorithm applied to a sequence emitted by A will be able to encode the sequence almost optimally, i.e. with an average number of bits per character equal to $-p_A \ln p_A - (1-p_A)\ln(1-p_A)$. This optimal coding will not be the optimal one for the sequence emitted by B. In particular, the entropy per character of the sequence emitted by B in the coding optimal for A will be the cross entropy per character:

$$h(B\|A) = -p_B \ln p_A - (1-p_B) \ln(1-p_A). \tag{32}$$

Similarly, the entropy per character of the sequence emitted by B in its optimal coding is $-p_B \ln p_B - (1-p_B)\ln(1-p_B)$. The number of bits per character wasted to encode the sequence emitted by B with the coding optimal for A is the relative entropy per character of A and B:

$$d(B||A) = -p_B \ln \frac{p_A}{p_B} - (1-p_B) \ln \frac{1-p_A}{1-p_B}. \tag{33}$$

A linguistic example will help clarify the situation: transmitting an Italian text with a Morse code optimized for English will result in the need for transmitting more bits than another coding optimized for Italian: the difference is a measure of the relative entropy.

The concept of relative entropy is crucial in the so-called language modeling. We have already discussed the problem of defining the entropy of a language. In the simplest approximation, one can define a language with a probability distribution $p(x)$ where x indicates a generic sequence of characters (a word). Since the distribution $p(x)$ is typically unknown, one should construct a suitable modeling by making a guess $q(x)$ about what the actual distribution could look like. Now the question is how good an estimate of $p(x)$ is $q(x)$. An answer to this question can be given in terms of the relative entropy. The relative entropy between the $q(x)$ distribution and the actual distribution $p(x)$ decreases as the language modeling improves.

More generally, relative entropy can be considered as a measure of the remoteness between two individual sequences; it represents an important tool for classification and information retrieval purposes. Suppose one wants to estimate the "distance" (or similarity) between texts A and B in terms of their information content. For instance, for two texts written in different languages (e.g. English and Italian), their "distance" is a measure of the difficulty experienced by a typical speaker of tongue A in understanding the text written in language B. The remoteness of one text from the other can be measured by estimating the relative entropy. We should remark that the relative entropy is not a distance (metric) in the mathematical sense: it is neither symmetric, nor does it satisfy the triangle inequality. As we shall see below, in many applications such as phylogenesis, it is vital to define a true metric that measures the actual distance between sequences. In this perspective a very important contribution has been made by the group of Li [17] where a rigorous definition of distance between unaligned sequences was proposed, by using the information theoretical concepts of Kolmogorov complexity [18].

5.4 Data Compression and Measures of Complexity

One last important point is how to obtain approximate measurements of quantities such as the Kolmogorov complexity or the relative entropy. We have already noted the impossibility of computing the Kolmogorov comple-

xity related to Turing's theorem on the halting problem and to Gödel's theorem [18].

Despite the impossibility to compute the Algorithmic Complexity of a sequence, one has to recall that there are algorithms explicitly conceived to provide a good approximation [18]. Since the Algorithmic Complexity of a string fixes the minimum number of bits one need to reproduce it (optimal coding), it is intuitive that a typical compression software (zipper), besides trying to reduce the space occupied on a memory storage device, can be considered as an entropy meter. The better the compression algorithm, the closer the length of the zipped file to the optimal coding limit and the better the estimate of the Algorithmic Complexity provided by the zipper. It should be remarked that the zippers can provide a reliable approximation (always an upper bound) to the Algorithmic Complexity of a typical string produced by a stationary stochastic process with finite memory (finite correlation length). By compressing the sequence of digits of π a compressor would never realize that there is a simple rule to generate it.

We have already discussed how gambling techniques can be used for compression purposes. Now we focus on specific zipping algorithms. A great improvement in this field is represented by the so-called Lempel and Ziv 77 algorithm [29]. This algorithm zips a file by exploiting the existence of repeated sub-sequences of characters in it. Its compression efficiency becomes optimal as the length of the file goes to infinity. This algorithm has been extensively studied and many variants and applications have been drawn from it.

In the field of the so-called computational linguistics, there have been several contributions showing how data compression techniques [27] can be useful in solving different problems: language recognition, authorship recognition or attribution, language classification, classification of large corpora by subject, etc. A detailed description of the state of the art is outside our present scope and we refer the reader to the specialized literature. Just to give an example, we quote one recent work [2] where data compression techniques were used to define a suitable measure of distance between two sequences, to be used for authorship attribution and language phylogenesis (see Fig. 3).

Of course the possibilities of data-compression-based methods go beyond computational linguistics. Another important example is that of genetic problems. Here also there have been important contributions and we refer to [17,10] for a recent overview.

It is evident how the specific features of data compression techniques in measuring entropic quantities make them potentially very important also for fields where human intuition can fail: DNA and protein sequences, as already mentioned, but also geological time series, stock market data or medical monitoring.

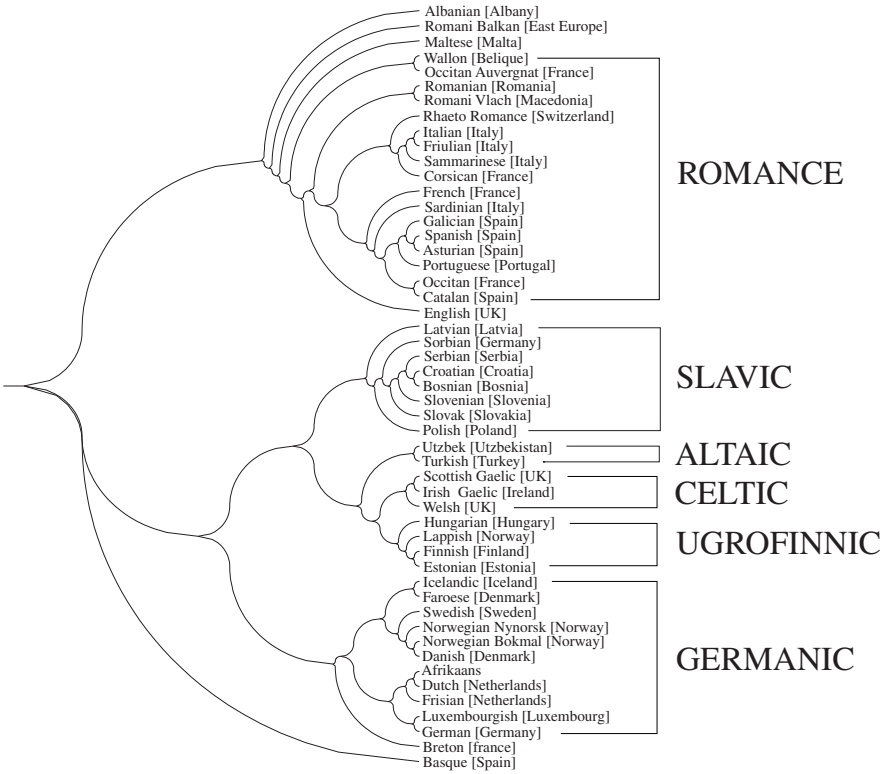

Fig. 3. This figure illustrates the phylogenetic-like tree constructed on the basis of more than 50 different versions of the "The Universal Declaration of Human Rights". The tree is obtained using the Fitch-Margoliash method applied to a symmetrical distance matrix whose elements are computed in terms of the relative entropy between pairs of texts. The tree features essentially all the main linguistic groups of the Euro-Asian continent (Romance, Celtic, Germanic, Ugro-Finnic, Slavic, Baltic, Altaic), as well as few isolated languages such as Maltese, typically considered an Afro-Asian language, and Basque, classified as a non-Indo-European language and whose origins and relationships with other languages are uncertain. The tree is unrooted, i.e. it does not require any hypothesis about common ancestors for the languages. What is important is the relative positions between pairs of languages. The branch lengths do not correspond to the actual distances in the distance matrix. It is important to stress how this tree is not intended to reproduce the current trends in the reconstruction of genetic relations among languages. It is clearly biased by the use of modern texts for its construction. In the reconstruction of genetic relationships among languages, one is typically faced with the problem of distinguishing *vertical* (i.e. the passage of information from parent languages to child languages) from *horizontal* transmission (i.e. which includes all the other pathways in which two languages interact). Horizontal borrowings should be expunged if one is interested in reconstructing the actual genetic relationships between languages

Acknowledgments

The authors are grateful to R. Baviera, D. Benedetto, L. Biferale, G. Boffetta, A. Celani, E. Caglioti, M. Cencini, G. Mantica, A. Puglisi, M. Serva and D. Vergni for the collaboration in the course of the years and for many enlightening discussions; and A. Baronchelli for the preparation of Fig. 3, extracted from his master thesis.

References

1. R. Badii and A. Politi, *Complexity, Hierarchical Structures and Scaling in Physics* (Cambridge University Press, Cambridge 1997)
2. D. Benedetto, E. Caglioti and V. Loreto, "Language trees and zipping", *Phys. Rev. Lett.* **88**, 048702 (2002)
3. G. Boffetta, M. Cencini, M. Falcioni and A. Vulpiani, "Predictability, a way to characterize complexity", *Phys. Rep.* **356**, 367 (2002)
4. J.L. Borges, "La biblioteca de Babel", in *Ficciones*, Primera edición (Buenos Aires, Editorial Sur 1944)
5. C. Cercignani, *Ludwig Boltzmann. The man who trusted atoms* (Oxford University Press, Oxford 1998)
6. G.J. Chaitin, *Information, randomness and incompleteness*, 2nd edition (World Scientific, Singapore 1990)
7. T. Cover and R.C. King, "A convergent gambling scheme of the entropy of english", *IEEE Trans. Inf. Th.* **24**, 413 (1978)
8. T. Cover and J. Thomas, *Elements of information theory*, (Wiley, New York 1991)
9. J. Ford, "How random is a coin tossing?" *Physics Today* **36**, 40 (1983)
10. Proceedings of the *IEEE Computer Society Bioinformatics Conference (CSB'02)*, (IEEE publisher, Palo Alto, California 2002)
11. J. Kelly, "A new interpretation of information rate", *Bell. Sys. Tech. Journal* **35**, 917–926 (1956)
12. A.I. Khinchin, *Mathematical Foundations of Information Theory* (Dover, New York 1957)
13. A.N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals", *IRE Trans. Inf. Theory.* **1**, 102 (1956)
14. A.N. Kolmogorov, "New metric invariant of transitive dynamical systems and auto-morphism of Lebesgue spaces", *Dokl. Akad. Nauk. SSSR* **119**, 861 (1958)
15. A.N. Kolmogorov, "Three Approaches to the Concept of the Amount of Information", *Prob. Info. Trans.* **1**, 1 (1965)
16. K. Huang, *Statistical Mechanics* (Wiley, London 1967)
17. M. Li, J. Badger, X. Chen, S. Kwong, P. Kearney and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny", *Bioinformatics* **17**, 149 (2001)
18. M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*, (Springer, New York 1997)
19. E. Lorenz, "Deterministic Nonperiodic flow", *J. Atmos. Sci* **20**, 130 (1963)
20. E. Ott, *Chaos in Dynamical Systems*, (Cambridge University Press, New York), Chap. 5, pp. 151–183

21. J.R. Pierce, *Introduction to information theory: Symbols, signals and noise*, 2nd edition (Dover Publications, Inc., New York 1980)
22. C.E. Shannon, "A mathematical theory of communication", *The Bell System Technical J.* **27**, 379 and 623 (1948)
23. C. Shannon, "Prediction and Entropy of Printed English", *Bell Systems Technical Journal* **30**, 50 (1951)
24. Y.G. Sinai, "On the concept of entropy for a dynamical system", *Dokl. Akad. Nauk. SSSR* **124**, 768 (1959)
25. R.J. Solomonoff, "A formal theory of inductive inference (part 1 and 2)", *Inform. Contr.* **7**, 1 (part 1) and 224 (part 2) (1964)
26. A.D. Wyner, 1994 Shannon Lecture, "Typical sequences and all that: Entropy, Pattern Matching and Data Compression", *IEEE Information Theory Society Newsletter*, July 1995
27. I.H. Witten, A. Moffat, and T.C. Bell, *Managing Megabytes* (Morgan Kaufmann Publishers, 1999)
28. A.M. Yaglom and J.M. Yaglom, *Probability and Information* (Science House, Leningrad 1973) Chap. IV, Part 3, pp. 236–329.
29. J. Ziv and A. Lempel, "A universal algorithm for sequential data compression", *IEEE Trans. Inf. Th.*, 337 (1977)
30. W.H. Zurek (editor), *Complexity, Entropy and Physics of Information* (Addison-Wesley, Redwood City, 1990)

Complexity and Intelligence

Giorgio Parisi

Dipartimento di Fisica, Sezione INFN, SMC and UdRm1 of INFM, Università di Roma "La Sapienza", Piazzale Aldo Moro 2, Rome, Italy, 00185,
Giorgio.Parisi@roma1.infn.it

Abstract. We discuss the properties of the Algorithmic Complexity, presenting its most relevant properties. The related concept of logical depth is also introduced. These properties will be used to study the problem of learning from example, paying special attention to machine learning. We introduce the propensity of a machine to learn a rule and we use it to define the intelligence of a machine.

1 Algorithmic Complexity

We have already seen in earlier in Part II that Kolmogorov, independently and more or less simultaneously with other people, introduced the concept of algorithmic complexity of a string of characters (or, if we prefer, of an integer number) as the length in bits of the shortest message that is needed to identify such a number. This idea seems similar to Shannon's entropy, but it is deeply different. Shannon's entropy is the shortest length of the message needed to transmit a *generic* (in the sense of measure theory) string of characters belonging to a given ensemble, for example a sentence of given length in Italian. More precisely, given a probability distribution of an ensemble of strings, Shannon's entropy controls the average length of a message obtained by compressing in an optimal way a string extracted with this probability distribution, in the limit where the length of the string goes to infinity. In contrast, Kolmogorov's complexity is the shortest message length needed to transmit a given string of characters: e.g. we want to know the length of the message needed to transmit the Divina Commedia, not to transmit a generic Italian poem of the same length written in Dante's style.

In a slightly different setting, that is nowadays more familiar to most of us, we can define the Kolmogorov complexity $\Sigma(N)$ of a number N as the length of the shortest computer program that computes such a number (or if we prefer, prints that number). This definition of complexity depends on the computer language in which the program is written, so that we should indicate in the definition of $\Sigma(N)$ the language we are using. However, if we are interested to large values of complexity, the choice of the language is irrelevant and it contributes in the worst case to an additive constant. Indeed, it is evident that

$$\Sigma_{Pascal}(N) < \Sigma_{Pascal}(Fortran\ compiler\ written\ in\ Pascal) + \Sigma_{Fortran}(N) \tag{1}$$

G. Parisi, Complexity and Intelligence, Lect. Notes Phys. **636**, 109–121 (2003)
http://www.springerlink.com/

Indeed if we know how to compute something in Fortran and we would like to compute it in Pascal, we can just write a Fortran compiler in Pascal and use it to execute the Fortran code. In the nutshell, we can transmit the Fortran compiler written in Pascal and the program that computes the message in Fortran. Inverting the argument we get

$$|\Sigma_{Fortran}(N) - \Sigma_{Pascal}(N)| < const, \tag{2}$$

where constant does not depend on N. So for very complex messages the choice of the language is not important.

This is common experience also in human languages. If we have a short discussion on the possible evolution of the snow conditions during the coming days, in relation to a trip on a dog-trained sled, it is extremely likely that the conversation should be much shorter is some languages and much longer in others. On the other hand, if we have to explain to somebody the proof of the last Fermat theorem starting from scratch (i.e. from elementary arithmetics), it would take years and in the process we can introduce (or invent) the appropriate mathematical language, so that the dependence on the natural language we use is rather weak. In other words, for messages of small complexity we can profit from the amount of information contained in the language, but this help becomes irrelevant in the case of very complex messages.

It should be clear that in the case of a generic number N of K binary digits, i.e. smaller than 2^K, elementary arguments (and also Shannon's theorem) tell us that the complexity is K. The complexity cannot be larger than K because we can always write a program that is of length K plus a small constant:

```
write  3178216741756235135123939018297137213617617991101654611661.
```

The length of this program is the same as the length of the output of the program (apart for the characters lost in writing "write"); it is rather unlikely (I will pay 1000 Euros for a counterexample) that there is a much shorter one[1]. Of course the program

```
write  11**57
```

is much shorter than

```
write 228761562390246506066453264733492693192365450838991802120171,
```

although they have the same output.

[1] I beg the reader to forget the inefficiency of writing decimal numbers in ASCII.

Let us consider an other example. The program

```
A=2**31-1
B=5**7
I=1
for K=1 to  A-1
    I=mod(I*B,A)
endfor
```

generates a sequence of pseudo random numbers that is rather long (about 8 Gbytes); apart form small variations, it is extremely likely that it is the shortest program that generates this sequence. Of course it is much shorter than the program (of length 8 Gbytes) that explicitly contains the sequence.

2 Some Properties of Complexity and Some Apparent Paradoxes

Although most of the numbers of K bits have complexity K, we have seen that there are notable exceptions. It should be clear that to identify these exceptions and to evaluate their complexity is not a simple task. A number may have a low complexity because it is a power, because it is obtained as the iteration of a simple formula, or just because it coincides with the second million digits of π. A systematic search of these exceptions is not easy.

However one could imagine a simple strategy. We can consider all the programs of length K. Each program will stop and write a string (maybe empty) or it will never stop: A program that never stops is

```
I=1
do forever
    I=I+1
    if(I<0) then
         write I
         stop
    endif
enddo
```

where I is an arbitrary length number.

In this case, it is trivial to show that the program never stops. In other cases it is much more difficult to decide if the program stops. Let us consider the following example:

```
I=1
do forever
     I=I+1
     consider all positive integers a,b,c and n, less than
        I with n>2.
     if(a**n+b**n=c**n) then
         write a,b,c,n
         stop
      endif
  enddo
```

This program stops if and only if Fermat's last theorem is false and we know now that it will never stop. Up to a few years ago, it was possible that it might stop, while writing a gigantic number. A program that stops only if the Goldbach conjecture[2] is false quite likely will never stop, though we do not know for sure.

There cannot be any computer program that can compute if a computer program stops or not. Otherwise we would have a contradiction. A computer program that can find those that stop, would be able to identify those program that are the shortest one and produce the same output: for a given length we can sort them in lexicographic order. If this happens, for example we could identify the first program of length K of this list. The output of this program has complexity K, in the same way as all the programs of this list, however we would identify the first program of the list transmitting only the number K; this can be done using only $\ln_2(K)$ bits and therefore the complexity of the output would be $\ln_2(K)$, i.e. a number much smaller than K and this is a contradiction.

One could try to find a loophole in the argument. Although it is impossible to decide if a program does not stop, because we would need to run it for an infinite amount of time, we can decide if a program stops in M steps. As long as there is a finite number of programs of length K, for each value of K we can define a function $f(K)$ that is equal to the largest value of steps at which a program of length K stops. All programs of length K that do not stop in $f(K)$ must run forever. The previous construction could be done by checking the output of all the programs of length K that stop in less than $f(K)$ steps. Therefore if we could compute the function $f(K)$ or an upper bound to it, we would be able to get the list of all the programs that stop.

A simple analysis shows that no contradiction is present if the function $f(K)$ increases faster than any function that is computable with a program

[2] The Goldbach conjecture states that any even integer n can be written as the sum of two prime numbers: the equation $p_1 + p_2 = n$ always has a solution with prime p_1 and p_2. The conjecture is evidently true: empirically the number of solutions increases, apart logarithmic corrections, linearly with n with fluctuations of order of $n^{1/2}$ with respect to the expected linear behaviour. No proofs are known.

of constant length: e.g. we must have that for large K that

$$f(K) > e^{e^{e^{e^{e^{e^{e^{e^{e^{e^{e^{e^{K}}}}}}}}}}}} . \tag{3}$$

The complexity of finding an upper bound to $f(K)$ must be $O(K)$.

The existence of a function that grows faster than what can be computed by a program of fixed length seems surprising although it is rather natural. Call $M(K)$ the maximum *finite* number[3] that can be printed by a program of length K. Now it is evident that $M(K+1) > M(K)$, therefore the function $M(K)$ cannot be printed by any program with length less that K, so that there is no finite length program that is able to print all the values of $M(K)$. The conclusion is that to find the shortest program (or if we prefer the shorted description) of a number N is something that cannot be done by a program: the complexity is a well defined quantity, but it cannot be computed systematically by a program.

Many variations on the same theme can be made: suppose that we add to our computer an oracle: a specialized hardware that tells the value of $f(K)$) (or directly the complexity of a string). We could now define a complexity $\Sigma_1(N)$ for the system *computer +oracle*. The same arguments as before tell us that the new complexity $\Sigma_1(N)$ cannot be computed for any N by the the system *computer +oracle*. We could introduce a superoracle and so on. The fact that we are unable, also with the use of any kind of oracles, to find a systems that can compute the computational complexity relative to itself recalls intuitively the Gödel theorem: This is not surprising since there are deep relations between these arguments and Gödel's theorem, that have been investigated by Chaitin and presented also in many popular papers.

3 The Logical Depth

The problem of finding the minimal complexity of a number (or of a sequence of numbers) intuitively corresponds to finding the rule by which the number has been generated. This problem is used in intelligence tests. For example one may be asked to find out the rules that generate the following sequences, or equivalently to predict the next numbers of the sequences

```
1 2 3 4 5 6 7 8 9 10 11 12 13
1 2 4 8 16 32 64 128 256 512 1024 2048
1 1 2 3 5 8 13 21 34 55 89 144
1 1 2 2 4 2 4 2 4 6 2 6 4 2 4 6 6 2 6 4 2 6 4 6 8 4 2 4
```

For the first two lines the rule is clear. The third line is a Fibonacci sequence: the next number is the sum of the of the two previous one. The last line

[3] It is trivial to write down a program that never stops and go on by writing an infinite number. Here we need a program that writes a number and then stops.

follows a slightly more complex rule and we leave to the reader the pleasure of finding it.

In the case of different rules, we consider a natural rule the simplest one: for example we could say that the first line is the sequence of natural integer that cannot be written as $a^2 + b^2 + c^2 + d^2$ with a, b c and d all different (the first missing integer would be 30). However this second rule is much longer and it is unnatural. The two rules would produce rather different sequences but if we know only the first elements, the first choice is much more natural.

A related and somewhat complementary concept is the logical depth of a number [10]: roughly speaking it is the amount of CPU time needed to compute it if we use the shortest program that generates it (i.e. the natural definition of the number).

It is evident that if the complexity is given by the upper bound and the program consists of only one write instruction, the execution is very fast, i.e. linear in the number of bits. On the other hand, if a very short program prints a very long sequence, the execution may be still very fast, as in the case of a typical random number generator, or may be very slow, e.g. the computation of the first 10^6 digits of π, or to find the solution of a difficult problem whose instance has been generated using a simple formula. A possible example of this last category is to find the N variables $\sigma(i) = \pm 1$ that minimize

$$\sum_{i=1,N} \left[\sigma(i) - \frac{1}{\sqrt{N}} \sum_{k=1,N} \sin\left(\frac{2\pi i k}{N}\right) \sigma(k) \right]^2 . \tag{4}$$

Although it is not excluded that such a problem has an analytic solution [4], it is quite likely that the shortest program is the trivial one (corresponding to the very definition of the problem): we examine all the 2^N configurations and we determine the minimum.

A careful analysis tells us that only a low complexity sequence may have a large logical depth; moreover the same arguments of the previous section tell us that there must be sequences with extremely large values of the logical depth for any value of the complexity K (i.e. logical depth of order $f(K)$).

These results concern us also because science aims to find out the simplest formulation of a law that reproduces the empirical data. The problem of finding the simplest description of a complicated set of data corresponds to finding the scientific laws of the world. For example both the Newton and Maxwell laws summarize an enormous quantity of empirical data and are likely the shortest descriptions of these data.The ability to find the shortest description is something that cannot be done in general by a computer; it is often taken a sign of intelligence.

Phenomenological explanations with a lot of parameters and not so deep theory are often easy to apply and correspond to rules that have high complexity and low logical depth while simple explanations with few parameters and a lot of computation needed have low complexity and rather high logical depths.

For example the computation of chemical properties using the valence theory and the table of electronegativity of the various elements belongs to the first class, while a first principle computation, starting from basic formulae, i.e. quantum mechanics belongs to the second class. A good scientific explanation is a low complexity theory (i.e. the minimal description of the data) that unfortunately may have an extremely large logical depth. Sometimes this leads to the use of approximated theories with higher complexity and smaller logical depth.

4 Learning from Examples

Scientific learning is just one instance of a general phenomenon: we are able to learn from example and to classify the multitude of external objects into different classes. The problem of how it is possible to learn from example has fascinated thinkers for a long time. In a nutshell the difficulty is the following: if the rule cannot be logically derived from the examples[4] how can we find it? The solution put forward by Plato was that the rule is already contained in the human brain and the examples have the only effect of selecting the good rule among all the admissible ones.

The opposite point of view (Aristotle) claims that the problem is ill posed and that the human brain is a tabula rasa before the experience of the external world.

Plato's point of view has often been dismissed as idealistic and non-scientific. Here we want to suggest that this is not the case and that Platonic ideas are correct, at least in a slightly different context. Indeed the possibility of having machines which learn rules from examples has been the subject of intensive investigations in recent years [5]. A very interesting question is to understand under what conditions the machine is able to generalize from examples, i.e. to learn the whole rule knowing only a few applications of the rule. The main conclusion we reach is that a given machine cannot learn an arbitrary rule; the set of rules that can easily be learned may be determined by analyzing the architecture of the machine. Learning by example consists of selecting the correct rule among those contained in this set. Let us see in the next sections how it may happen and how complexity is related to intelligence.

5 Learning, Generalization, and Propensities

In order to see how this selectionist principle may work, let us start with some definitions. In the following I will consider rules which assign to each

[4] Two examples: (a) the horsiness is not the property of any given horse or set of horses (b) there are many different mathematical rules that generate the same sequence.

input vector of N Boolean variables ($\sigma_i = 0$ or 1 for $i = 1, N$) an output which consists of a single Boolean value. In other words a rule is a Boolean valued function defined on a set of 2^N elements (i.e. the set of all the possible values of the variables σ); it will be denoted by $R[\sigma]$.

The rule may be specified either by some analytic formulae or by explicitly stating the output for all the possible inputs. The number of different rules increases very rapidly with N: it is given by $2^{2^N} \equiv 2^{N_I}$, where N_I is the number of different possible input vectors, i.e. 2^N. In the following we will always consider N to be a large number: terms proportional to 1/N will be neglected.

A learning machine is fully specified if we know its architecture and the learning algorithm.

Le us first define the architecture of the machine. We suppose that the computations that the machine performs depend on M Boolean variables ($J_k,\ k = 1, M$). In a nutshell, the architecture is a Boolean function $A[\sigma, J]$, which gives the response of the machine to the input σ's for each choice of the control parameters J. Typical architectures are the perceptron [6] or a neural network, with discretized synaptic couplings.

For each given rule R and choice of J's, the machine may make some errors with respect to the rule R. The total number of errors $E[R, J]$ depends on the rule R and on the J's; it is given by

$$E[R, J] = \sum_{\{\sigma\}} (R[\sigma] - A[\sigma, J])^2 \ . \tag{5}$$

For a given architecture the machine may learn the rule R without errors if and only if is there exists a set of J's such that $E[R, J] = 0$. Simple counting arguments tell us that there are rules that cannot be learned without errors, if M is smaller that 2^N . In most of the cases 2^N is much larger than M and therefore the number of admissible rules is a very tiny fraction of all the possible rules.

In a learning session we give to the machine the information on the values of $R[\sigma]$ for L instances in the σ's (L is generally much smaller than 2^N). A learning algorithm tries to find the J's which minimize the error on these L instances[5]. Let us denote by J^* the J's found by the learning algorithm.

If the error on the other $2^{\{N\}-L}$ instances has decreased as a consequence of having learned the first L instances, we say that the machine is able to generalize, at least to a certain extent. Perfect generalization is achieved when no error is made on the other $2^{\{N\}-L}$ instances.

For a given machine the propensity to generalize depends on the rule and not all rules will be generalized by the machine. Our aim is to understand how

[5] There are many different learning algorithms and some are faster than others. The choice of the learning algorithm is very important for practical purposes, but we will not investigate this point anymore.

the propensity to learn different rules changes when we change the machine; in this note, we are only interested in the effect of changing the architecture.

It was suggested by Carnevali and Patarnello in a remarkable paper [7] that, if we suppose that the learning algorithm is quite efficient, the propensity of the machine to generalize a given rule (p_R) depends only on the architecture. The propensity (p_R) may be approximated by the number of different control parameters J for which the total number of errors is zero.

In other words we define the propensity[6]:

$$p_R = 2^{-M} \sum_{\{J\}} \delta(E[R, J]) \ , \tag{6}$$

where $E[R, J]$ is given by (1) and obviously depends only on the architecture. The function δ is defined is such a way that $\delta(k) = 1$ for $k = 0$, $\delta(k) = 0$ for $k \neq 0$.

According to Carnevali and Patarnello, rules with very small propensity cannot be generalized, while rules with higher propensity will be easier to generalize. In their approach the propensity of a given architecture in generalizing is summarized by the values of the function p_R for all the 2^{N_I} arguments (the p_R's depend only on the architecture, not on the learning algorithms). Of course the propensity cannot be directly related to the number of examples needed to learn a rule. A more detailed analysis, taking care of the relevance of the presented example, must be done.

6 A Statistical Approach to Propensities

Our aim is to use statistical mechanics techniques [8] to study in detail the properties of p_R.

The p_R are automatically normalized to 1 ($\sum_R p_R = 1$) and it is natural to introduce the entropy of the architecture A:

$$S[A] = -\sum_R p_R \ln(p_R). \tag{7}$$

The entropy $S[A]$ is a non negative number smaller than or equal to $\ln(2) \min(2^N, M)$.

We could say that if the entropy is finite for large N, the machine is able to represent essentially a finite number of rules, while, if the entropy is too large, too many rules are acceptable.

As an example we study the entropy of the perceptron (without hidden unity). In this case (for $N = M$ odd) a detailed computation shows that all the 2^N choices of the J's lead to different rules (i.e. two different J's produce a different output at least in one case) and therefore $S[A] = \ln(2)N$.

[6] The word "propensity" is used with a different and uncorrelated meaning in epistemology.

We note that we could generalize the previous definition of entropy by introducing a partition function $Z(\beta)$ defined as follows

$$Z(\beta) = \sum_R \exp(\beta \ln(p_R)) = \sum_R p_R^\beta. \tag{8}$$

We could introduce the entropy $S(\beta)$ associated with the partition function [8] ($S(\beta) \equiv d\ln[Z(\beta)/\beta]/d\beta$). The previously defined entropy (7) coincides with $S(1)$.

The value of the entropy as a function of β tells us which is the probability distribution of the p'_Rs. There are many unsolved questions whose answer depends on the model: existence of phase transitions, structure of the states at low temperature, breaking of the replica symmetry ...

Many additional questions may be posed if we consider more than one architecture; in particular we would like to find out properties which distinguish between architectures that have similar entropies. For example we could consider two different architectures (i.e. a layered perceptron or a symmetric neural network) with N inputs and one output, and which have the same entropy (this can be achieved for example by adjusting the number of internal layers or hidden neurons). It is natural to ask if these two different architectures are able to generalize the same rules, or if their propensity to generalize is concentrated on rules of quite different nature. Our aim is to define a distance between two architectures, which will help us to compare their different performances.

Let us consider two architectures A and B. A first step may consist of defining the entropy of B relative to A as

$$S[B/A] = -\sum_R p_R(A) \ln[p_R(B)/p_R(A)] \ . \tag{9}$$

It can be shown that $S[B/A]$ is a non-negative quantity that becomes zero if and only if $p_R(A) = p_R(B)$. The relative entropy is not symmetric and we can define the distance (or better the difference) between A and B as

$$d(A,B) = \frac{1}{2}(S[B/A] + S[A/B]) \ . \tag{10}$$

The introduction of a distance allows us to check if two different architectures have the same generalization propensity; do they generalize the same rules, or are the rules they can generalize that different? Unfortunately, the explicit computation of the distance among two architectures may be very long and difficult.

A very interesting and less understood question is how many examples are needed to specify the rule for a given architecture. The results obviously depend on the architecture (people with an high value of serendipity guess the right rule after a few examples) and explicit computations are very difficult, if we exclude rather simple cases.

7 A Possible Definition of Intelligence

Having in our hands a definition of the distance between two architectures, we can now come to a more speculative question: how to define the intelligence of an architecture? One possibility consists of defining an architecture $I(\sigma, J)$, which is the most intelligent by definition; the intelligence of A can be defined as $-d[A, I]/S[A]$ (the factor $S[A]$has been introduced for normalization purposes). The definition of the intelligent architecture I is the real problem.

We suggest that a sequence of the most intelligent architectures is provided by a Turing machine (roughly speaking a general purpose computer) with infinite amount of time for the computation with a code of length L. More precisely the J's are the L bits of a code for the Turing machine (written in a given language) which uses the σ as inputs. The function $I(\sigma, J)$ is 1 if the program coded by J stops after some time, and it is 0, if the the program never stops. With this architecture we can compute the function $s(R)$ (i.e. the simplicity of a rule) [9], defined as $s(R) \equiv p_R(I)$.

It may be possible to estimate the function $s(R)$ using the relation $s(R) \approx 2^{-\Sigma(R)}$, where $\Sigma(R)$ is the algorithmic complexity (introduced in the first section) i.e. the length of the shortest code which computes the function $R[\sigma]$.

If we would like to know the intelligence of an architecture A with entropy S, we should consider a Turing machine with nearly the same entropy and we should compute the distance between the two architectures.

The previous observations imply that algorithms which compute the rule in a relatively short time are probably unable to implement many rules with low algorithmic complexity and high logical depth. In many of the most common algorithms the number of operations needed to compute the output for a given input is proportional to a power of N. For other algorithms (e.g. an asymmetric neural network, in which we require the computation of a limiting cycle) the computer time needed may be much larger (e.g. proportional to 2^L). It is natural to suppose that this last class of algorithms will be more intelligent than the previous one and will learn rules with low algorithmic complexity and high logical depth more easily [11].

The puzzled reader may ask: if a general purpose computer is the most intelligent architecture, why are people studying and proposing for practical applications less intelligent architectures like neural networks? A possible answer is that this definition of intelligence may be useful to disembodied entities that have an unbound amount of time at their disposal. If we have to find a rule and take a decision in real time (one second or one century, it does not matter, what matters is that the time is limited) rules of too low complexity and very large logical depth are completely useless; moreover a computer can be used to define the intelligence but we have seen that a computer is not well suited for finding and executing intelligent rules. The requirement of taking a fast decision may be dramatic in some case, e.g. when

you meet a lion on your path; evolution definitely prefers a living donkey to a dead doctor and it is quite likely that we have not been selected for learning rules with too large logical depth.

There are also other possible inconveniences with rules with too low complexity, i.e. they may be unstable with respect to a small variation of the data they have to reproduce. It is quite possible that the shortest program that pronounce words in Italian may have a completely different structure of the shortest program that pronounces words in Spanish, in spite of the similarity of the two languages. We would like to have a program that may be used for many languages, where we can add a new language without changing too much the core of the program; it is quite likely that such a program would be much longer than the shortest one for a given set of rules. Strong optimization and plasticity are complementary requirements. For example the nervous system of insects is extremely optimized and probably it is able to perform the tasks with a minimal number of neurons, but it certainly lacks the plasticity of mammalian nervous system.

Architectures that are different from a general purpose computer may realize a compromise producing rules that have low, but not too low complexity and high, but not to high logical depth. Moreover these architectures are specialized: some of them (like neural networks) may be very good at working as associative memories but quite bad in doing arithmetics. The problem of finding the architecture that works in the most efficient way for a given task is difficult and fascinating and, in many cases (e.g. reconstructing three dimensional objects from two dimensional images), it also has a very important practical impact.

I hope that this short exposition has convinced the reader of the correctness of the point of view that learning from example can be done only by selecting among already existing rules. This is what typically happens in biology in many cases. The genetic information preselects a large class of behaviors, while external stimuli select the behaviour among the available ones. This procedure can be seen in action in the immune systems. The number of possible different antibodies is extremely high ($O(10^{100})$); each given individual at a given moment produces a much smaller number of different antibodies (e.g. $O(10^8)$). The external antigens select the most active antibodies among those present in the actual repertoire and stimulate their production. Eldeman has stressed that it is quite natural that a similar process happens also in the brain as far as learning is concerned.

In conclusion at the present moment we have only started our exploration of the properties of rules that are implicitly defined by an architecture. I am convinced that the future will bring us many interesting results in this direction. It is amazing how many areas have been affected by Kolmogorov's ideas, and how far reaching the applications.

References

1. A.N. Kolmogorov, *Problemy Peredachi Informatsii* **1**, 3 (1965)
2. G.J. Chaitin, *Journal of ACM* **13**, 547 (1966)
3. G.J. Chaitin, *Journal of ACM* **16**, 145 (1969)
4. E. Marinari, G. Parisi, and F. Ritort, *J. Phys. A: Math and Gen.* **27**, 7647 (1994); M. Degli Esposti, C. Giardinà, S. Graffi, and S. Isola, *J. Stat. Phys.* **102**, 1285 (2001)
5. See for example D.J. Amit, *Modeling Brain Functions*, Cambridge University Press, Cambridge (1989); J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Horward, L. Jackel, and J.J. Hopfield, *Complex Systems* **1**, 877 (1987)
6. M. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, 1969)
7. P. Carnevali and S. Patarnello, *Europhys. Lett.* **4**, 1199 (1987)
8. G. Parisi, *Network* (1990); G. Parisi and F. Slanina, *Europhysics Lett.* **17**, 497 (1992)
9. G.J. Chaitin, *Journal of ACM* **22**, 329 (1975)
10. C.H. Bennet, *Inter. J. Theoretical Physics* **21**, 905 (1982)
11. G. Parisi and F. Slanina, *Europhys. Lett.* **17**, 497–502 (1992)

Information Complexity and Biology

Franco Bagnoli[1], Franco A. Bignone[2], Fabio Cecconi[3], and Antonio Politi[4]

[1] Dipartimento di Energetica "S. Stecco", Università Firenze, Via S. Marta, 3 Firenze, Italy, 50139, franco.bagnoli@unifi.it
[2] Istituto Nazionale per la Ricerca sul Cancro, IST Lr.go Rosanna Benzi 10, Genova, Italy 16132, abignone@unige.it
[3] Università degli Studi di Roma "La Sapienza", INFM Center for Statistical Mechanics and Complexity, P.le Aldo Moro 2, Rome, Italy, 00185, Fabio.Cecconi@roma1.infn.it
[4] Istituto Nazionale di Ottica Applicata, INOA, Lr.go Enrico Fermi 6, Firenze, Italy, 50125, politi@ino.it

Abstract. Kolmogorov contributed directly to Biology in essentially three problems: the analysis of population dynamics (Lotka-Volterra equations), the reaction-diffusion formulation of gene spreading (FKPP equation), and some discussions about Mendel's laws. However, the widely recognized importance of his contribution arises from his work on algorithmic complexity. In fact, the limited direct intervention in Biology reflects the generally slow growth of interest of mathematicians towards biological issues. From the early work of Vito Volterra on species competition, to the slow growth of dynamical systems theory, contributions to the study of matter and the physiology of the nervous system, the first 50–60 years have witnessed important contributions, but as scattered pieces apparently uncorrelated, and in branches often far away from Biology. Up to the 40' it is hard to see the initial loose build up of a convergence, for those theories that will become mainstream research by the end of the century, and connected by the study of biological systems per-se.

The initial intuitions of L. Pauling and E. Schrödinger on life and matter date from this period, and will gave the first initial full fledged results only ten years later, with the discovery of the structure of DNA by J. Watson and F. Crick, and the initial applications of molecular structures to the study of human diseases few years earlier by Pauling. Thus, as a result of scientific developments in Biology that took place after the 50', the work of Kolmogorov on Information Theory is much more fundamental than his direct contributions would suggest. For scientist working in Molecular Biology and Genetics, Information Theory has increasingly become, during the last fifty years, one of the mayor tools in dissecting and understanding basic Biological problems.

After an introductory presentation on algorithmic complexity and information theory, in relation to biological evolution and control, we discuss those aspects relevant for a rational approach to problems arising on different scales. The processes of transcription and replication of DNA which are at the basis of life, can be recasted into an Information theory problem. Proteins and enzymes with their biological functionality contribute to the cellular life and activity. The cell offers an extraordinary example of a highly complex system that is able to regulate its own activity through metabolic network. Then we present an example on the formation of complex structures through cellular division and differentiation in a model organism (C. elegans). Finally we discuss the essential principles that are thought to rule evolution through natural selection (theory of fitness landscapes).

F. Bagnoli, F.A. Bignone, F. Cecconi, and A. Politi, Information Complexity and Biology, Lect. Notes Phys. **636**, 123–146 (2003)
http://www.springerlink.com/

If one were to judge Kolmogorov's contribution to Biology only on the basis of his papers explicitly devoted to the topic, one might conclude that it is of moderate interest, at least in comparison with his direct intervention in turbulence or dynamical systems. However, one should not forget that, in the past, the limited availability of quantitative data in Biology made this subject quite unattractive for mathematicians. It is therefore remarkable that Kolmogorov nevertheless contributed to three different problems: the analysis of population dynamics (Lotka-Volterra equations), the reaction-diffusion formulation of gene spreading (Fisher Kolmogorov Petrovsky Piskunov – FKPP– equation), and some discussions about Mendel's laws. It is however widely recognized that the relevance of Kolmogorov's contribution is connected to his work on algorithmic information. In fact, after the initial intuitions of L. Pauling and E. Schrödinger on life and matter in the '40s and, especially after the discovery, ten years later, of the structure of DNA by J. Watson and F. Crick, it has become increasingly clear that information plays a major role in many biological processes. The penetration of these concepts in Biology has led to the formulation of the central dogma of genetics: the discovery of a one-directional flow of information from DNA and genes to proteins, and from there to morphogenesis, cellular organization and finally to individuals and communities. In this perspective, life is now viewed as the execution of a computer program codified in the DNA sequence. However, in spite of the elegance of this doctrine, one cannot forget the various difficulties that make the final goal of decoding the program much more difficult than one could expect. First of all, in order to understand what the life-code does without running it, it is necessary to know the logic of the underlying "hardware", or "wetware" as it is sometimes called. Wetware is certainly structured in quite a different way from the hardware of ordinary digital computers. Indeed, living system are highly parallel devices, where the parallelism does not only enhance the "computer" performance, but is also there to guarantee the required redundancy for a robust functioning both in the presence of a noisy environment or of significant damages even. As a result, in living systems, information-theoretic aspects are profoundly interlaced with the physico-chemical mechanisms responsible for their functioning and it could not be otherwise, considering that they are the result of a self-assembling process that has evolved over billions of years.

The roadmap of this contribution is as follows. The first section is devoted to an historical account of the connections between biology and the other major scientific disciplines. This allows us to place Kolmogorov's direct contributions (exposed in the next section) in the right context. Next, we give a brief presentation of information and algorithmic-information theories in relation to biological systems. Finally, we discuss the problem of protein folding as an example of how information, physics and dynamics concur to biological processes at an almost microscopic level of description.

1 Historical Notes

For a long time, biology has mainly dealt with definitions, classification and description of objects connected with the evolution of life forms such as bacteria, plants, animals, biochemical substances, proteins, sugars or genes. The great variety of life forms, present and past, has required a large amount of ground work before general theories could be formulated. The Evolution theory of Charles Darwin is a good case in point. Moreover, the dissection of environments, organisms and cells has been practiced by many in the hope that a precise identification of the objects of study is a necessary and perhaps sufficient condition to understand their role: a sort of reductionism, like in physics, with the difference that no simple general laws have ever been identified.

In this continuous search for an appropriate methodology, biologists have been led to emphasize different aspects (ranging from mechanical, to chemical, thermodynamical, and molecular) depending on the current development of science. As an example, in Fig. 1 we report an apparatus invented by an

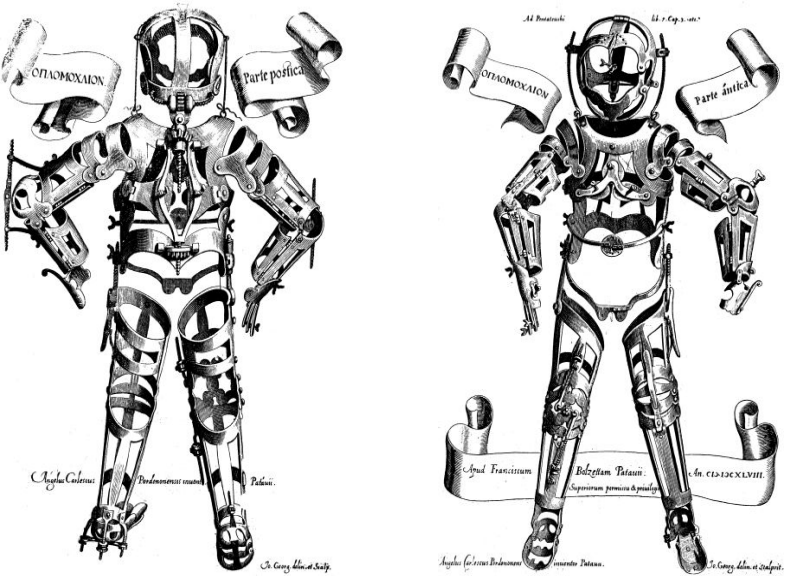

Fig. 1. Girolamus Fabricius ab Acquapendente (1533–1619), machina, Museo Villasneri from [1]. This example of *machina*, a system to bring body parts into correct place–proportions, is due to Fabricius ab Acquapendente, anatomist at the University of Padova. Acquapendente was Professor of Anatomy in Padova during the same period in which Galileo Galilei, who was trained in his youth as a physician, was teaching there (1592–1610). Acquapendente and Galileo were tutors of William Harvey (1578–1657). Harvey, by application of the experimental method, discovered blood circulation, while working at St. Bartholomew's hospital in London [2]

anatomist, Girolamus Fabricius ab Acquapendente at the turn between the 16^{th} and 17^{th} centuries; the machine was meant to be used in order to bring parts of the human body into correct proportions.

The ingenious, naive, primitive, and somehow sadistic mean of fixing body parts through the use of a reference frame is remarkable. The general idea *to fix it* is already there with the means of the period, and has evolved with human knowledge. The present level of technology allows the discussion of atomic-molecular adjustments, but, in many cases, the conceptual aim has not changed much since 1600, aside from a deeper understanding of the inner details involved. The reported picture, nevertheless, warns that the desire to find solutions to biological problems, drawing principles from other sciences, has always been present from the very beginning. Sometimes it has been stretched to the limit.

The limited success of this stratgegy has led in the last century to a strong debate about two alternative approaches, the *holistic* and the *reductionistic* view. The former one assumes that biological systems, in order to be understandable, must be considered and described in their wholeness; the latter one understands that full operational knowledge can be reached only after characterizing all of the single components. While the supporters of the holistic view have, with a few exceptions, fostered more qualitative approaches, the reductionistic attitude has been much more oriented towards quantitative work (for a detailed historical account of this long story, the interested readers are invited to consult the detailed book of B.O. Kuppers, [3], or the historical account of E. Mayr, [4]).

However, in recent years the scenario has started to change and there exists now a chance that the above two points of view can be reconciled within a suitable information-theoretic approach. Indeed, *information processing* represents a truly unifying concept that allows the investigation of seemingly different issues such as the functioning of the brain, the cell cycle, the immune system, or the "simple" task of recognizing food, moving towards it, and so on. Furthermore, one should not forget the problem that has represented a puzzle for centuries: the transmission from one generation to the next of the "plan" for constructing new individuals.

The first results were obtained by G. Mendel (1865) and concerned statistical correlations between phenotypic characters of species. However, the idea that characters are due to elementary units spread about 35 years later, through the work of H. de Vries, C. Correns, and E. Tschermak. Later T. Boveri, W. Sutton, and especially T.H. Morgan and collaborators established the chromosomal theory, the link between characters, genes and the physical existence of chromosomes. Still, all research in genetics up to the beginning of the '40s was done as an inductive reconstruction, through the study of crossing and mutants, with no knowledge of the mechanisms and molecules carrying this information.

All in all, spreading of information-theoretic concepts in Biology occurred by following several convoluted paths. This is true also for the brain: the

system that, more than any other, can be seen as an information-processing unit. It is indeed quite illuminating to cite W.S. McCulloch about the way he came, together with W. Pitts, to the development of his famous simplified model of a brain, in which neurons were treated as boolean interconnected switches [5],

> I came, from a major interest in philosophy and mathematics, into psychology with the problem of how a thing like mathematics could ever arise – what sort of thing it was. For that reason, I gradually shifted into psychology and thence, for the reason that I again and again failed to find significant variables, I was forced into neurophysiology. The attempt to construct a theory in a field like this, so that it can be put to any verification, is tough. Humorously enough, I started entirely at the wrong angle, about 1919, trying to construct a logic for transitive verbs. That turned out to be as mean a problem as modal logic, and it was not until I saw Turing's paper that I began to get going the right way around, and with Pitt's help formulated the required logical calculus.

One of the major obstacles in the development of a theoretical Biology is its nonstationary character: strictly speaking, there are no stationary states – one can at most imagine that, on some scales, quasi-equilibrium is maintained. It is, indeed, since the beginning of the last century that Evolution has been recognized as playing a crucial role and the first models on the time variation of species have been introduced. Initially, the spread of the ideas of C. Darwin strongly depended on the country; in some cases they were partially accepted, allowing evolution but not natural selection – such as, e.g., in France. In others, these ideas were accepted much faster, as in the case of Russia [4]. In the beginning of the century a dichotomy between naturalists and geneticists took place on the way to proceed in order to understand evolution. The former looked more at *final causes*, while the latter, more oriented towards physical and mathematical methods, pursued a strict experimental approach. The major achievement of Genetics in this period was the rejection of the theory of acquired characters – i.e. the pangenesis hypothesis of C. Darwin, or the theories of those biologists who followed J.B. Lamarck [4] –. Without any knowledge of the physical base for the transmission of characters, the demonstration was done by means of statistical methods, and by showing the combinatorial character of traits due to more than one gene (the final experimental demonstration came with work done by Salvatore Luria and Max Delbrück in the '40s).

Attributing a key role to information processing amounts to assuming that the mechanisms through which, e.g., a face or an antigene is recognized, can be understood without the need to characterize in full detail the underlying physical processes and chemical reactions. This is indeed a fruitful hypothesis, formulated already by von Neumann in the book on "The computer and

the brain", that has given rise to the formulation of several more-or-less abstract models introduced in the last 20 years, in the hope of identifying possibly universal mechanisms. One of the most successful models is the Hopfield model [6] that exploited a possible analogy between an associative memory and a spin glass. It shows how information can be robustly stored and retrieved in a context where many connections can be accidentally destroyed, as it is the case of our brains.

Although this is a route that will be useful to pursue in the future as well, one cannot neglect biochemical processes, at least to understand how biological systems can self-assemble. In fact, another discipline that is having an increasing impact on Biology is the theory of dynamical systems. In the last century it has been progressively recognized that most, if not all, processes that are responsible for the functioning of a living system involve nonlinear mechanisms which, in turn, are responsible for the onset of nontrival time dynamics and the onset of spatial patterns. Initial rudimentary attempts to figure out a physico-chemical explanation for the origin of life can already be found in [7], although this remained an isolated attempt, still very qualitative and rooted into J.B. Lamarck ideas. The modelling of oscillations, thanks to the work of A.J. Lotka, where one of the well studied models was introduced, can also be attributed to this path.

Later contributions came thanks to the advances of B.P. Belousov, with his discovery of the chemical reaction that bears his name, the Belousov-Zhabotinsky reaction. Belousov discovered this reaction while attempting to model the Krebs cycle. The Krebs cycle, i.e. the tricarboxylic acid cycle, is the name given to the set of reactions that transforms sugars or lipids into energy. Degradation produces acetyl-CoA, a molecule with two atoms of carbon, which are transformed through the cycle in two molecules of CO_2 while producing energy in the process. He showed that the oxidation of citric acid in acidic bromate, in the presence of Cerium catalysis – [Ce(IV)]/[Ce(III)] –, produces oscillations in the reaction visible through changes in color of the solution. The discovery was made in 1951 but the paper was rejected because the nonequilibrium nature of the thermodynamic process was not understood. He finally published his result in 1958 in the proceedings of a conference. Along this same path, a theoretical paper on the spontaneous formation of patterns in chemical systems was published by Alan Turing in 1952 [8]. However, while Belousov's contribution had an experimental basis, it was not until the beginning of the '90s that Turing's hypothesis was demonstrated experimentally.

The relevance of dynamical system theory in Biology has definitely emerged in the beginning of the '60s in connection with the problem of gene regulation. For instance, in a series of papers, by Jacques Monod, Francois Jacob, and André Lwoff give some hints about the *logic* of living systems and show that regulation of the β-Galactosidase system in bacteria can be seen as a switch. The classical scheme of the Lactose Operon works as a sensor of the presence-absence of Lactose. As a first approximation, a protein

produced in the bacterial cell, the lactose repressor, binds to the operator, a 22 base pairs (bp) long stretch of DNA in front of the genes. This blocks RNA Polymerase that should start transcription of DNA in order to make the mRNAs of the three genes which are part of the Operon (β-Galactosidase, transacetylase and lactose permease). If the lactose is present, it binds to the repressor, unlocking the operator and transcription begins. If the lactose is absent, transcription is blocked. Aside from further complexities, this system, in its baseline, can be considered similar to a boolean switch. These ideas have been pointed out by Jacques Monod and Francois Jacob both in technical papers and in popular articles. Particularly Monod stressed the fact that the logic of living forms follows Boolean algebra, with a series of more or less complex logic circuits at work.

The initial formalization of genes as switches, in a way similar to the modelling of McCulloch and Pitts, is due to M. Sugita in 1961, soon followed by S. Kauffman [9,10]. A similar approach to the study of the dynamics of gene expression was pursued by B. Goodwin [11].

However, recognition that high-level cellular functions are regulated by a plethora of proteins interacting in cells had to wait until the end of the '80s, beginning of the '90s. Since then, it has become increasingly clear that the lower dimensional levels in Biological systems, those of molecules, organelles and cells, are as difficult as the higher dimensional one to solve. Several apparently *simple* functions have revealed a degree of sophistication previously unforeseen. As a result, the currently emerging picture of multicellular systems is much more similar to a *highly regulated society*, than to the simple gene-protein scheme accepted for many years, [12–15].

2 Kolmogorov's Direct Contributions

Before discussing the contributions of Kolmogorov to Biology, it is useful to recall the situation in the Soviet Union. While acceptance of natural selection in the USA and Europe had to overcome political and philosophical barriers, this was not the case in Russia. An important figure in the development of Evolution theories was Sergej S. Chetverikov (1880-1959). In 1906 he published an important study on fluctuations in populations. He was able to demonstrate that not all mutations have a negative impact on fitness: some are almost neutral and, as shown later by Dobzansky, some can even increase the fitness. Moreover, because of heterozygosity – the presence of two copies of each gene – most mutants remain silent within the population, as shown also by R.A. Fisher, and only homozigous individuals will be exposed to selection. Chetverikov demonstrated these facts through back crossing experiments with wild type *Drosophila melanogaster*. His most important result was that the previous idea of the structure of organisms made of independent genes had to be abandoned. No gene has a constant fitness because his expression will depend on the global genetic background. However, his work

was not well-known outside Russia and, after 1926, he had to leave his post for political reasons [4,16].

Independently of the work done within the group of Chetverikov, around 1935, Kolmogorov became interested in some problems of mathematical genetics, probably stimulated by the ongoing debate about Lamarckism occurring in Europe and especially in the Soviet Union.

The first relevant contribution on the subject deals with the propagation of "advantageous genes" (see chapter 9 in this book). The variable of interest is the concentration $0 \leq c(x,t) \leq 1$ of individuals expressing a given gene, at position x and time t. In the absence of spatial dependence, the concentration is assumed to follow a purely logistic growth, $\dot{c} \equiv F(c) = Kc(1-c)$: this dynamics is characterized by two stationary solutions, $c = 0$ and $c = 1$. If $K > 0$, the former one is linearly unstable; any small fraction of individuals carrying the "advantageous" gene tends to grow, but the limited amount of resources put a limit on the growth which converges to the stable solution $c = 1$. In the presence of spatial directions, it is natural to include the possibility of a random movement of the single individuals. As a result, the whole process is described by Fisher's equation [17], proposed in 1937, ten years after the work of Chetverikov,

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} + Kc(1-c). \tag{1}$$

$c = 0$ and $c = 1$ are still meaningful stationary solutions, but now the dynamics of the spatially extended system is determined not only by the evolution of small perturbations, but also by the propagation of one phase into the other. This is a general observation whose relevance goes beyond the original context, since it applies to all pattern-forming dynamical systems. This specific model is important, since it is one of the very few nonlinear reaction-diffusion equations that can be treated analytically. The relevant solutions are front-like ones connecting the two different fixed points (e.g, $c(x,t) \to 1$ for $x \to -\infty$ and $c(x,t) \to 0$ for $x \to \infty$). The relative stability of the two phases is thereby quantified by the front velocity that can be estimated by assuming that the front travels without changing shape, i.e. $c(x,t) \equiv f(x-vt) \equiv f(z)$. By assuming that $f(z)$ decays exponentially to 0, $f(z) \simeq \exp(-\gamma z)$ for large z, one can easily investigate the front propagation by replacing this ansatz into the linearized (1). As a result, one finds that

$$v(\gamma) = \begin{cases} K/\gamma + \gamma D \text{ if } \gamma > \gamma^* \\ \\ 2\sqrt{KD} \qquad \text{if } \gamma \leq \gamma^* \end{cases}$$

where $\gamma^* = \sqrt{K/D}$. If the parameter γ defining the initial profile is smaller than γ^*, then the front propagates with the minimal velocity $v_{min} = v(\gamma^*) = 2\sqrt{KD}$. This is the well-known velocity selection mechanism (see also chapter 9 of this book).

In the same year as Fisher's paper, Kolmogorov, Petrovskii and Piskunov [18] extended the solution of the problem to a fairly general class of local growth-functions $F(c)$, rigorously proving the following expression for the propagation velocity

$$v_{min} = 2\sqrt{F'(0)D} \tag{2}$$

where the prime denotes the derivative w.r.t. the argument (Fisher's result is recovered by noticing that in the logistic case $F'(0) = K$).

There are two other studies by Kolmogorov in genetics. The first one [19] concerns the problem of statistical fluctuations of Mendel's laws. The interest in this subject is mainly historical, since it aimed at refuting a claim by T.D. Lysenko that Mendel's "3:1 ratio"-law is only a statistical regularity, rather than a true biological law. More important is the second contribution which extends the notion of Hardy-Weinberg (HW) equilibrium in population genetics. HW equilibrium refers to the simplest setup, where allele statistics can be studied. It assumes: i) random crossing between individuals; ii) absence of mutations; iii) neutrality, i.e. absence of mechanisms favouring a given allele; iv) closed population in order to avoid exchanges of alleles with the environment; v) an infinite population. Mathematically, the problem can be formulated as follows: Given the probability p ($q = 1 - p$) to observe the allele A (a), the free crossing of genes A and a produces AA, Aa, and aa with probabilities p^2, $2pq$, and q^2, respectively and the frequency of individuals follows a Bernoulli distribution.

The route towards more realistic models requires progressively relaxing the above restrictions. Kolmogorov first investigated the effect of a diffusive coupling in a system of otherwise closed populations and then studied the consequences of selection simulated by a mechanism suppressing the occurrence of the recessive allele a. Here, we briefly summarize the first generalization. Kolmogorov considered a large ensemble of N individuals divided into s populations, each containing the same number n of individuals, ($N = sn$). Like in the HW scheme, random mating (free crossing) is assumed within each population. Moreover, a number of k individuals are allowed to "migrate" towards different populations and thus to contribute to the next generation. As a result of the mutual coupling, a population characterized by a concentration p of the allele A experiences an average drift $F(p)$ towards the equilibrium value p^* (corresponding to the concentration in the total population) with variance σ^2,

$$F(p) = \frac{k}{n}(p^* - p) \qquad \sigma^2(p) = \frac{p(1-p)}{2n}.$$

Altogether, the distribution $\rho(p,t)$ of populations with concentration p satisfies the Fokker-Planck equation,

$$\frac{\partial \rho}{\partial t} = -\frac{\partial F\rho}{\partial p} + \frac{1}{2}\frac{\partial^2 \sigma^2 \rho}{\partial p^2} \quad , \tag{3}$$

whose stationary solution is

$$\rho(p) = \frac{C}{\sigma^2(p)} \exp\left\{2 \int dp \frac{F(p)}{\sigma^2(p)}\right\} = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)},$$

where $\alpha = 4kp^*$, $\beta = 4kq^* = 4k(1-p^*)$ and the Euler beta-function $B(\alpha,\beta)$ accounts for the proper normalization. The frequency of individuals carrying AA, Aa, and aa and, hence deviations from a pure HW-equilibrium can then be estimated by simply averaging p^2, $p(1-p)$, and $(1-p)^2$, respectively over the distribution $\rho(p)$.

Finally, Kolmogorov made some contributions in the modeling of population dynamics, by generalizing the Lotka-Volterra equations. Such a model, on the Volterra side, followed an experimental observation by the Italian biologist Umberto D'Ancona, who discovered a puzzling fact. During the first World War, the Adriatic sea was a dangerous place, so that large-scale fishing effectively stopped. Upon studying the statistics of the fish markets, D'Ancona noticed that the proportion of predators was higher during the war than in the years before and after. V. Volterra, stimulated by his son in law D'Ancona, formulated the problem in terms two coupled differential equations,[1]

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = (\varepsilon_1 - \gamma_1 N_2)N_1, \qquad \frac{\mathrm{d}N_2}{\mathrm{d}t} = (-\varepsilon_2 + \gamma_2 N_1)N_2,$$

N_1 and N_2 being the abundance of preys and predators, respectively. They exhibit periodic behaviour whose amplitude depends on the initial conditions. This feature crucially depends on the form of the proposed equations, because the dynamics admits a conserved quantity E (analogous to the energy in conservative systems)

$$E = \gamma_2 N_1 + \gamma_1 N_2 - \varepsilon_2 \log(N_1) - \varepsilon_1 \log(N_2),$$

while the periodic orbits are level lines of E. However, E has no direct biological meaning. Kolmogorov argued that the term $\gamma_2 N_1$ is too naive, because it implies that the growth rate of predators can increase indefinitely with prey abundance, while it should saturate at the maximum reproductive rate of predators. Accordingly, he suggested the modified model [22]

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = K_1(N_1)N_1 - L(N_1)N_2, \qquad \frac{\mathrm{d}N_2}{\mathrm{d}t} = K_2(N_1)N_2,$$

where $K_1(N_1)$, $K_2(N_1)$ and $L(N_1)$ are suitable functions of the prey abundance and predators are naturally "slaved" to preys. With reasonable assumptions on the form of $K_1(N_1)$, $K_2(N_1)$ and $L(N_1)$, Kolmogorov obtained

[1] The same equations were derived also by A.J. Lotka some years before [20,21] as a possible model for oscillating chemical reactions

a complete phase diagram, showing that a two-species predator-prey competition may lead to either extinction of predators, stable coexistence of prey and predator, or, finally, oscillating cycles. He also generalized the differential equation to more than two species[2], introducing most of the phenomenology nowadays known in population dynamics.

Moreover, Kolmogorov pointed to the strong character of the assumptions behind an approach based on differential equations. In particular, populations are composed of individuals and statistical fluctuations may not be negligible, especially for small populations. In practice, there exists a fourth scenario: at the minimum of a large oscillation, fluctuations can extinguish the prey population, thereby causing the extinction of predators too. It is remarkable to notice how Evolution has developed mechanisms to reduce "accidental" extintions. In most species, the birth of individuals takes place during a very short time interval. In some cases, such as for example for herbivores like the gnus – *Connochaetes taurinus* –, living in herds, the birth of puppies is limited to a time span as short as one-two weeks. This mechanism helps in preserving the species since the number of newborns highly exceeds the possibility of killing by predators.

3 Information and Biology

Information is one of those technical words that can also be encountered within natural languages. C.E. Shannon, who was mainly interested in signal transmission, succeeded in formalizing the concept of information by deliberately discarding semantic aspects. He states in the beginning of [24]

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently, the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.

In fact, before discussing the meaning of single messages, it is necessary to distinguish among them. In this sense, the information becomes the number of independent specifications one needs in order to identify a single message x in an ensemble of N possible choices. Given, for instance, an ensemble of N equiprobable messages x_k, the unambiguous identification of a specific message x requires taking $\log_2 N$ binary decisions. One can indeed split the initial ensemble into two subsets and identify the one containing x. This operation involves the minimal amount of information, a "bit", and it must be recursively repeated until the remaining set contains no more than one element. Accordingly, the amount of information is $I = I(x) = \log_2 N$ bits.

[2] See for instance [23]; the generalized version is sometimes referred as the *Kolmogorov model.*

If messages do not have the same probability, it is natural to define the information of a single message as

$$I_k = -\log_2 p(x_k) \quad , \tag{4}$$

and accordingly introduce the average information

$$H = \sum_k p(x_k) I_k = -\sum_k p(x_k) \log_2 p(x_k) \quad . \tag{5}$$

The quantity H was defined as an entropy by Shannon, since it can also be interpreted as an uncertainty about the actual message.

In many contexts, the object of investigation is an ideally infinite sequence $s_1 s_2 \dots s_n \dots$ of symbols (s_i belonging to an alphabet with 2^b letters) that can be viewed as a collection of strings S_i of length n with probability $p(S_i, n)$. In this case, the information is written as $H(n) = \sum_{i=1}^{2^{bn}} p(S_i, n) \log_2 p(S_i, n)$ and the sum extends to the set of all possible strings. The difference $h_n = H(n+1) - H(n)$ is the information needed to specify the $(n+1)$st symbol given the previous n, while $h = \lim_{n\to\infty} h_n$ is the Kolmogorov-Sinai entropy of the signal. The maximum value, $h = b$ is attained for random sequences of equally probable symbols, while $h = 0$ is a distinctive feature of regular messages.

Another useful indicator is mutual information

$$M(k) = \sum p(s_j, s_{j+k}) \log_2 \frac{p(s_j, s_{j+k})}{p(s_j) p(s_{j+k})}, \tag{6}$$

measuring the statistical dependence between two variables; $M(k) = 0$ if and only if the two variables are mutually independent.

While the concept of information was being formalized, crucial progress was made in Biology that led to the discovery of the DNA double-helix structure by J. Watson and F. Crick [25] in 1953. This was made possible by the development of methods for the analysis of chemical structures based on X-ray scattering, mostly by William and Lawrence Bragg together with the intuitions of L. Pauling for the application of the method to the study of protein and DNA structure [26][3]. One should not however forget also the impact of the book of E. Schrödinger on atoms and life [27], where he argued about the existence of a disordered solid as the medium hiding the secrets of life.

The crucial role of information within genetics has become increasingly clear with the discovery that DNA and proteins are essentially string-like objects, composed by a sequence of different units (bases and amino acids,

[3] The story goes that the interest in protein structure was aroused in Pauling by Warren Weaver, head of the Natural Sciences division of the Rockefeller Foundation, who convinced him to work on the problem, financed by the Rockefeller's funds.

Table 1. Genetic code, translating the codon triplets into amino acids, e.g. UUU and UUC both corresponds to amino acid Phenylalanine (Phe), while Leucine (Leu) is encoded by six possibilities UUA, UUG, CUU, CUC, CUA, CUG. Notice that the symbol "T" is replaced by "U" since the translation codons → aminacids actually involves RNA and not directly DNA. It is evident that most of the redundancy in the code is due to the third base of each codon. Triplets UAA, UGA and UAG are the stop-codons; they do not encode any amino acid but locate the end of the protein

First Position	Second Position				Third Position
	U	C	A	C	
U	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	**Stop**	**Stop**	A
U	Leu	Ser	**Stop**	Trp	G
C	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
C	Leu (Met)	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
A	Ile	Thr	Lys	Arg	A
A	Met (Start)	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
G	Val (Met)	Ala	Glu	Gly	G

respectively) linked together by covalent chemical bonds which ensure a chain structure. The properties of DNA, RNA and proteins are briefly recalled below in a specific box; for further details, we recommend that the reader consult any modern textbook on molecular biology [28,29].

The information contained in the DNA is first transferred to RNA and eventually to proteins. In the latter step there is a loss of information because the $4^3 = 64$ possible different triplets of nucleotides – codons – are mapped onto only 20 amino acids (see Table 1). This is therefore an irreversible process and there is no way to go back from the proteins to DNA since different nucleotide sequences can code for the same protein.

After the discovery of DNA's structure and function, a shift of focus took place in genetics and biology: all relevant properties became traced back to the information stored at the molecular level. The DNA sequence is viewed as "the fundamental issue" and the pursuit of the *sequencing projects* for several organisms has been the main direct consequence of this view. The centrality of DNA is so undisputed that *human-like* behavioral characteristics are occasionally attributed to the chemical properties of this molecule.[4] The

[4] The use of the catchy metaphor of the *selfish gene* by R. Dawkins is a good example.

reasons for this are several, the main one being the appealing interpretation of living organisms as complex computing machines. The DNA then represents the code that the program actually runs. However, the relative ease with which DNA molecules can be studied has contributed to this view, since the DNA is relatively constant for a certain organism. Technological advances during the last three decades have made the process sequencing routine, through the use of automatic machines.

Given its naturally symbolic structure, DNA can be directly investigated by means of information-theoretic tools. The zero-*th* order question is whether DNA can be treated as a stationary process, since the very computation of probabilities requires this. Several studies have revealed that a slow drift in the composition may be present and must be carefully accounted for. Once this is made clear, one can proceed by computing the probability $p(S_i, n)$ of each sequence of length n and thus determine the information $H(n)$. If the DNA was a purely random sequence of equally probable bases (four), $H(n)$ would attain its maximum value $H(n) = 2n$. This is almost true for $n \leq 4 \div 5$: for instance $H(1) = 1.95$ in the human chromosome 22 [30], the 0.05 difference from 2 being an indication of the slightly uneven distribution of the four nucleotides. However, upon increasing n, h_n decreases and for $n = 10$ it has already decreased down to 1.7. Going to larger n-values is basically impossible, since one would need such large samples to reliably determine the exponentially small probabilities, that even $10^7 - 10^8$ bp are no longer sufficient. One partial way to go around the problem is by looking for low-order correlations. A straightforward solution consists in studying the standard correlation $C(k) = \langle s_j s_{j+k} \rangle - \langle s_j \rangle^2$. Starting with [31], several studies performed on different living organisms have revealed a slow correlation-decay, $C(k) \simeq k^{-\gamma}$. Such observations have been confirmed by studying also the mutual information $M(k)$ (see (6)) which only requires computing probabilities of pairs of symbols, k bp apart from each other. For instance, in [30], a decay with $\gamma \approx 1/4$ was found up to $k = 10^5$. Many factors seem to be responsible for such a slow decay on different scales, but none of them prevails. For instance, it is known that many almost-equal repeats are interspersed within DNA and some are even as long as 300 bp, but they are responsible for correlations only up to $k \approx 10^2$,

3.1 Algorithmic Information

As soon it was realized that DNA is simply a long message possibly containing the instructions for the development of a living being, algorithmic issues immediately became relevant. In the decade across '60s and '70s, R. Solomonoff, A. Kolmogorov, and G. Chaitin, [32–37], independently set the basis of what is now known as *algorithmic information theory*. Consider the sequences

$$S_a = ATGCATGCATGCATGCATGCATGCATGCATGCATGC$$
$$S_b = AATAGATACAAACATGTCGACTTGACACATTTCCTA,$$

it is clear that S_a is somehow simpler than S_b. Suppose indeed that we have to describe them; while the former string is fully characterized by the statement `8 times ATGC`, the latter cannot be better described than enumerating the individual symbols. However, in order to make more quantitative our considerations about "simplicity", it is necessary to formalize the concept of description of a given string. The Turing machine is a tool to answer this question: it is a general purpose computer which, upon reading a series of instructions and input data (altogether representing the program), produces the required string S. It is therefore natural to consider the program length as a measure of the "complexity" of S. As proven by Solomonoff, Kolmogorov and Chaitin this is an objective definition, provided that the shortest code is first identified. In fact, on the one hand, there exist the so-called universal Turing machines (UTMs) that are able to emulate any other machine. On the other hand, there is no need to refer to a specific UTM, since the unavoidable differences among the lengths of minimal codes corresponding to different UTMs are independent of the sequence length N. More precisely, the Kolmogorov-Chaitin algorithmic complexity $K(S)$, i.e. the minimal code length, is known within a machine-dependent constant and $\kappa(S) = \lim_{N\to\infty} K(S)/N$ is an objective quantity. Unfortunately one consequence of the undecidability theorem, proved by Kurt Gödel, is that there is no general algorithm to determine $\kappa(S)$ which thereby turns out to be an uncomputable quantity.

While information deals with ensembles of strings, algorithmic information aims at measuring properties of single sequences. In spite of this striking difference, there is a close analogy to the extent that we now often speak of "algorithmic information". In fact, one may want to determine the probability $P(S)$ that a given UTM generates a string S when fed with a sequence of independent, equally probable bits. Since Chaitin proved that $K(S) = -\log_2 P$, one can interpret $K(S)$ as the logarithm of the probability that the minimal code is randomly assembled. This observation is particularly suited to discuss the role of chance within biological evolution. Indeed, if the DNA sequence is a randomly selected program, even imagining the Earth as a gigantic parallel processor performing independent tests every cubic millimiter each nanosecond,[5] the probability $P(DNA)$ should be larger than 10^{-50} and, accordingly, $K(DNA) < 200$. In other words, it should be possible to compress the DNA

[5] This is reminescent of an episode of *the hitchiker's guide to the galaxy* by Douglas Noel Adams (whose acronym is DNA)
`http://www.bbc.co.uk/cult/hitchhikers/`: Some time ago a group of hyperintelligent pan dimensional beings decided to finally answer the great question of Life, The Universe and Everything. To this end they built an incredibly powerful computer, Deep Thought. After the great computer programme had run seven and a half million years, the answer was "42". The great computer kindly

sequence down to less than 200 bits (or, equivalently, 100 bp). We cannot exclude that this is the case, but it is hard to believe that all instructions for the development, of e.g. humans, can be compressed within such a short length!

An observation that helps to close the gap is that only part of the genome is transcribed and then translated: according to the most recent results, less than 2% of human DNA is transformed into proteins! A small fraction of the remaining 98% contributes to the regulation of metabolic processes, but the vast majority seems to be only accidentally there. This is so true that onion-DNA contains 3 times more bp than human-DNA! Whatever the algorithmic content of this so-called "junk" DNA, we are clearly left with the crucial problem of discovering the language used to store information in DNA. Several researchers have investigated the DNA structure in the hope of identifying the relevant building blocks. In natural and artificial languages, words represent the minimal blocks; they can be easily identified because words are separated by the special "blank" character. But how to proceed in partitioning an unknown language, if all blanks have been removed? Siggia and collaborators [38] have proposed to construct a dictionary recursively. Given, e.g., the sequence $S_n = s_1 \dots s_n$, it is extended to $S_{n+1} = s_1 \dots s_n s'$, if the probability of S_{n+1} turns out to be larger than the probability of S_n multiplied that of the symbol s'. In the opposite case, the process is stopped and S_n is identified as a new word. The rationale behind this approach is that when a word is completed, a sudden uncertainty arises due to the ignorance about the newly starting one. The above approach has been successfully implemented, allowing the recognition of several regulatory motifs.

Extracting information from the sole knowledge of the DNA sequence seems however to be an exceedingly hard problem, since the products of DNA translation interact with each other and with the DNA itself. In order to gain some insight about living matter, it is therefore useful, if not necessary, to look directly at the structure of the "final product" in the hope of identifying the relevant ingredients. In this philosophy, many researchers have pointed out that living matter is characterized by non-trivial relationships among its constituents. Chaitin [39], in particular, suggested that the algorithmic equivalent of mutual information represents the right setup for quantifying the degree of organization of a sequence S. More precisely, he introduced the d-diameter complexity $K_d(S)$ as the minimum number of bits needed to describe S as the composition of separate parts S_i, each of diameter not greater than d,

$$K_d(S) = min \left[K(\alpha) + \sum_i K(S_i) \right], \tag{7}$$

pointed out that what the problem really was that no-one knew the question. Accordingly, the computer designed its successor, the Earth, to find the question to the ultimate answer.

where $K(S_i)$ is the algorithmic information of the single ith piece and $K(\alpha)$ accounts for the reassembling processes needed to combine the various pieces. If $d > N$, $K_d(S) = K(S)$ and $K_d(S)$ increases as d decreases. The faster the difference $\delta K(S) = K_d(S) - K(S)$ increases, the more structured and organized S is. The beauty of this approach is in the fact that no definition of the constituents is required: they are automatically identified by determining the partition that minimizes the d-diameter complexity.

In the case of a perfect crystal (i.e. a periodic self-repeating sequence), $K(S)$ is very low and $K_d(S)$ remains low, even when S is broken into various pieces, since there is no connection between the different cells. The same is true in the opposite limit of a gas-like (purely random) sequence. In this case, $K(S)$ is maximal and remains large when S is partitioned in whatever way, as all bits are, by definition, uncorrelated with each other.

3.2 DNA → RNA → Proteins

DNA (deoxyribonucleic acid) is a double-stranded polymer made of four elementary components called nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). Nucleotides are small molecules consisting of a phosphate group linked to a pentose (a sugar with 5 carbon atoms) which is in turn bound to one of the bases. The two strands interact via hydrogen bonds linking the pairs A-T and C-G. In its native state, the two DNA-strands spiral around each other and assume the well-known double helix conformation, as proposed by Watson and Crick in 1953 [25]. DNA is the carrier of the genetic information required to build a living organism. Such information is organized in units named genes which, from a molecular point of view, are sequences of DNA nucleotides capable of synthesizing a functional polypeptide. Roughly speaking a gene is a portion of DNA which encodes a protein.

RNA (ribonucleic acid) is generally a single strand polymer made of the same nucleotides as DNA, except for the replacement of Thymine with Uracil (U). RNA is the outcome of DNA transcription, and is copied by using a given portion of DNA as a template. RNAs which carry information to be translated into proteins are called messenger RNAs, mRNA. Other RNAs, such as rRNA and tRNA are involved in the translation of mRNA into proteins.

Amino acids are the proteins' building blocks; even if their number is potentially limitless, only twenty types of amino acid are involved in natural proteins. Amino acids share a common structure: each of them is made by at least one amino group -NH_2 and a carboxyl group -COOH, both linked to a central carbon atom C_α (α-carbon) which is in turn bound to a side chain (functional group or residue). It is the chemical nature of the side chains that differentiate amino acids from one another, conferring to them a structure with chemical and physical specificity. Amino acids are connected together to form the protein chain through peptide bonds, which are established by the chemical reaction between the -NH_2 group of one amino acid and the -COOH

group of another. The sequence of amino acids determines the properties and the function of a protein.

4 Proteins: A Paradigmatic Example of Complexity

In addition to being a technical term introduced in the theory of computation, *complexity* is a word widely invoked in many different contexts ranging from turbulence, to networks of any kind, spin glasses, chaotic dynamics and so on. In spite of the great diffusion of this term, no clear definition of complexity has yet been given. This will presumably remain so in the near future, since it is unlikely that so many different problems share some well-defined properties. Nevertheless, if there exists a scientific discipline where complexity is to be used, it is Biology, both for the diversity of structures existing over a wide range of scales and for the combination of several mutual interactions among the very many constituents.

Proteins have been selected for their mixed digital and analog nature and since they represent the point where the initial set of genetic instructions is transformed into a working device capable of processing information. A protein is uniquely defined by a sequence of amino acids, which in turn follows from the translation of a string of DNA. However, after a protein is linearly assembled by the ribosomes of the cell, it begins to twist and bend until it attains a three-dimensional compact structure – the native configuration – that is specific of each protein and is crucial for its biological function. Because of thermal fluctuations, the final shape is, however, not exactly determined so that a protein can be seen as a digitally assembled analog device. Once assembled, proteins represent the "working molecules", supporting and controlling the life of an organism. Structural proteins, for instance, are the basic constituents of cells and tissues; other proteins store and transport electrons, ions, molecules, and other chemical compounds. Moreover, some proteins perform a catalytic function (enzymes), while others control and regulate cell activity. Most of these processes involve many proteins of the same type at once, so that it is tempting to draw an analogy with statistical mechanics, with a microscopic level (that of the single molecules) and a macroscopic one, characterized by a few effective variables (e.g., concentrations and currents). Accordingly, biological problems are akin to non-equilibrium statistical mechanics and the relevant questions concern how the definition of specific microscopic rules translates into a given macroscopic behaviour.

The lack of theoretical tools for dealing with such systems prevents us of finding general solutions, but analogies with known problems can sometimes be invoked and many help in making substantial progress (examples are the statistical mechanics of disordered systems and reaction-diffusion equations). Moreover, modern microscopic techniques allow the visualization and manipulation of single molecules, so that it is now possible to study proteins experimentally and clarify their mutual interactions.

The first problem one is faced with is to understand how each protein finds its native configuration. Indeed, consider a protein molecule with N amino acids, and assume that there are q preferred orientations of each monomer with respect to the previous one along the chain. Then, there exist q^N local minima of the energy that can a priori be meaningful "native" states. Moreover, one can imagine that the true native configuration is that corresponding to the most stable minimum. If this picture were correct, it is hard to imagine a polymeric chain exploring the whole set of minima in a few milliseconds (the folding time can indeed be so short) to identify the absolute minimum: for a chain with $N = 100$ amino acids and $q = 3$, no more than $10^{-50}s$ should be dedicated to each minimum! This is basically the famous Levinthal paradox [40], which strongly indicates that protein folding can neither be the result of an exhaustive search nor of a random exploration of the phase space.

How can proteins find their native state within such a huge number of possible configurations? A significant step towards the answer was made by Anfinsen and coworkers. They discovered, in an *in vitro* experiment – i.e. outside the cell environment – that the enzyme ribonuclease, previously denaturated, was able to spontaneously refold into its native state when the physiological conditions for the folding were restored. This work [41], that won Anfinsen the Nobel Prize, demonstrated that the folding of protein molecules is a self-assembly process determined, at a first approximation, only by the amino acids sequence. It is not assisted by the complex cell machinery nor by enzymatic activity[6]. This great conceptual simplification in the folding problem gave great stimulus to its study. The activity was not restricted to the fields of Biology and Biochemistry, but was tackled with the methods of Physics, specifically of statistical mechanics. After Anfinsen's work, attention soon shifted towards the so-called "folding code", i.e. the basic rules through which the information stored in a one dimensional structure, the amino acid sequence (also called primary structure), encode the three-dimensional protein structure (tertiary structure). From an information-theoretic point of view, if one is to specify the native configuration out of the above mentioned ensemble of possibilities, the required information is on the order of $N \log_2 q$. This is compatible with the information contained in the DNA sequence, equal to $6N$ bits. However, no algorithm has been found to predict the tertiary structure, given the primary one: this seems to belong to the class of hard computational problems. If a shortcut exists, a solution of the *folding problem* will be more likely found by studying the underlying physics and understanding its dynamics.

The first insight into the folding mechanisms came from the observation of protein shapes. Experimental data on protein structures, collected through

[6] Actually, futher studies revealed that large-protein folding can be assisted by molecular chaperons (chaperonins) and other helper enzymes to prevent protein self-aggregation and possibly dangerous misfolding. However the role of the chaperonins that are themselves proteins is still non fully elucidated.

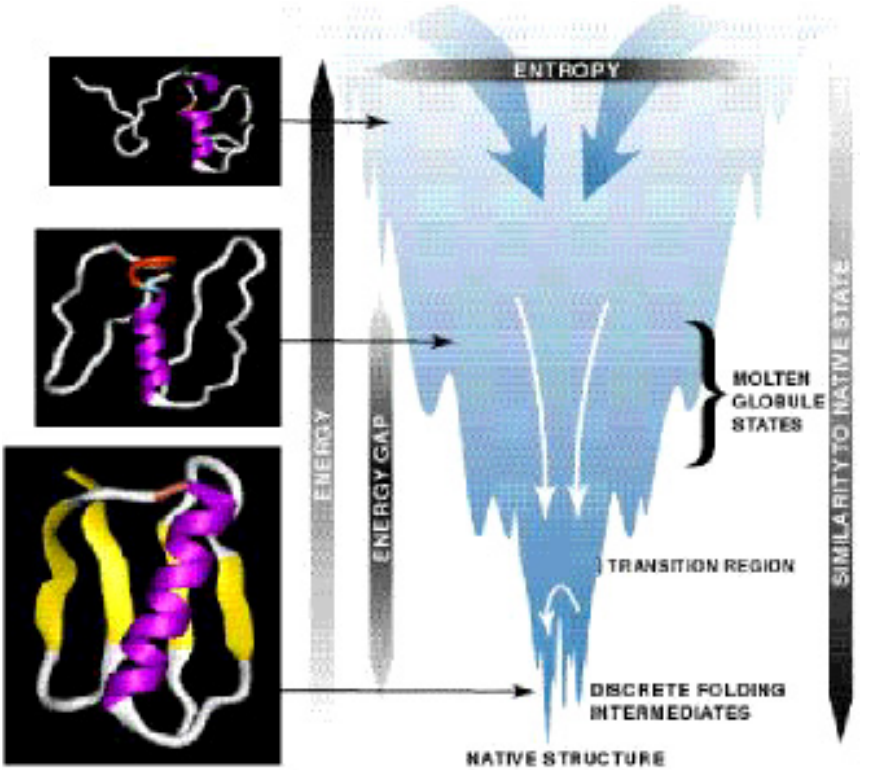

Fig. 2. A possible folding funnel scenario with the corresponding interpretation of folding stages. In the horizontal axis, protein conformations are parametrized by conformational entropy, while the energy is on vertical axis. On the side, typical protein conformations corresponding to states in the funnel.

X-ray spectroscopy and nuclear magnetic resonance (NMR), show that folded proteins are not random arrangements of atoms, but present recurrent motifs. Such motifs, forming the *secondary structure* of proteins, consist of α-helices (L. Pauling), β-sheets and loops (see Fig. 2). The secondary structure formation plays a crucial role in the folding process, since it introduces severe steric and topological contraints that strongly influence the way the native state can be reached.

Another hint about the rules that govern the folding comes from the analysis of the amino acid properties. The twenty natural amino acids can be grouped into two classes: hydrophobic and polar. While polar amino acids are preferentially exposed to water molecules, hydrophobic ones avoid contact with water; this is possible by grouping them together. As a result, most of the hydrophobic residues are buried inside the native structure, while the polar ones are located near the surface. In 1959, Kauzmann [42] realized that the hydrophobic effect is the principal driving force of the folding. However,

even if the hydrophobic collapse has a prominent role in directing the folding, it is not a sufficiently precise criterion to predict the protein structure from the knowledge of the amino acid sequence.

Earlier theoretical efforts to understand protein folding were directly aimed at bypassing Levinthal's paradox. For instance, it was proposed that a protein, during folding, follows a precise sequence of steps (pathway) to the native state without exploring the whole configurational space. This ensures a fast and large decrease of conformational entropy and justifies the relatively short folding times. However, even though it must be true that only a subset of the phase space is explored, several works on the folding kinetics revealed that folding of a given protein does not always follows the same route. The pathway scenario implies also the concept of intermediate states, i.e. states with partially folded domains that favour the correct aggregation of the rest of the protein. However the determination of intermediates is critical because they are metastable states with a relatively short lifetime.

A general theory of the protein folding requires the combination of polymer theory and the statistical mechanics of disordered systems. In fact, several features of the folding process can be understood from the properties of random heteropolymers and spin-glasses. However, there is a great difference between random heteropolymers and proteins: proteins have an (almost) unique ground state, while random heteropolymers have, in general, many degenerate ground states. In other words, proteins correspond to specific amino acid sequences that have been carefully selected by evolution in such a way that they can always fold in the same "native" configuration.

Many of these features have been investigated in what is perhaps the simplest model of a protein, the HP model [43]. It amounts to schematizing a protein as a chain of two kinds of amino acids, hydrophobic (H) and polar (P) ones, lying on a three-dimensional cubic lattice. Accordingly, the primary structure reduces to a binary sequence such as, e.g., $HPPHHHPHH\ldots$. Moreover, pairwise interactions are assigned so as to energetically favour neighbouring of H monomers in real space. Investigation of HP-type models and of more realistic generalizations has led to the "folding funnel" theory [44] which provides the currently unifying picture for the folding process.

This picture, generally referred to in the literature as the "new view", is based on the concept of free-energy landscape. This landscape neither refers to the real space nor to the phase-space, but to the space identified by the order parameter(s). In order to construct such a picture, it is first necessary to identify the proper parameters; the study of spin glasses has shown that this may not be an easy task in disordered systems. In the case of protein models, it was suggested that the proper coordinate is the fraction of correct contacts, i.e. the number of monomer pairs that are nearest neighbours both in the given and the "native" configuration. Theoretical considerations [44] based on simplified models, such as the HP model, suggest that the landscape of proteins is funnel-shaped with some degree of ruggedness (see Fig. 3). The local energy oscillations are a manifestation of frustration, a typical

Fig. 3. Ribbon representation of the protein chemo-trypsin-inhibitor (CI2), showing the characteristic secondary motifs alpha-helix and β-sheets

property of many disordered systems, here induced by the conflicting polar and hydrophobic interactions.

The funnel structure is the essential property ensuring an efficient collapse, because it naturally drives the system towards the minimum of free energy. Moreover, the protein can be temporarily trapped into the deepest relative minima, which correspond to the intermediates observed in kinetics experiments. Accordingly, the funnel scenario is able to reconcile the thermodynamic and kinetic features of the folding process.

References

1. http://www.unipd.it/musei/vallisneri/uomo/13.html
2. W. Harvey, *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus.* Frankfurt am Main, Wilhelm Fitzer, 1628
3. B.O. Kuppers, *Information and the origin of life.* The MIT Press, Cambridge, Mass., 1990
4. E. Mayr, *The growth of biological thought. Diversity, evolution and inheritance.* The Belknap Press of Harvard University Press, Cambridge, Mass, 1982
5. W.S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115 (1943)
6. J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. USA* **79**, 2554 (1982)
7. S. Leduc, *Théorie physico-chimique de la vie, et générations spontanées.* Poinant Éditeur, Paris, 1910

8. A.M. Turing, The chemical basis of morphogenesis. *Philos. Trans. R. Soc. London B*, **237**, 37 (1952)
9. M. Sugita, Functional analysis of chemical systems *in vivo* using a logical circuit equivalent. *J. Theoret. Biol.* **1**, 415 (1961)
10. S.A. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437 (1969)
11. B.C. Goodwin, *Temporal organization in cells.* Academic Press, 1963
12. M. Eigen, Self-organization of matter and the evolution of biological macromolecules, *Naturwissenschaften* **58**, 465 (1971)
13. M. Eigen, J. McCaskill, and P. Schuster, The molecular quasi-species. *Adv. Chem. Phys.* **75**, (1989)
14. D. Bray, Protein molecules as computational elements in living cells, *Nature*, **376**, 307 (1995)
15. T.S. Shimizu, N. Le Novere, M.D. Levin, A.J. Beavil, B.J. Sutton, and D. Bray, Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis. *Nat Cell Biol* **2**, 792 (2000)
16. S.S. Chetverikov, *On certain aspects of the evolutionary process from the standpoint of modern genetics. Zhurnal Ekspeimental'noi Biologii*, **A2**, 3 (1926). Translated by M. Barker, I.M. Lerner Editor, *Proc. Am. Phil. Soc.* **105**, 167 (1961)
17. R.A. Fisher, The wave of advance of advantageous genes, *Ann. Eugenetics* **7**, 353 (1937)
18. A.N. Kolmogorov, I.G. Petrovskii, and N.S. Piskunov, Étude de l'équation de la diffusion avec crissance de la quantité de matière e son application à un problème de biologie. *Moskow Univ. Bull. Math.* **1**, 1 (1937)
19. A.N. Kolmogorov, On a new confirmation of Mendel's Laws. *Dokl. Akad. Nauk. SSSR* **27**, 38 (1940)
20. A.J. Lotka, Undamped oscillations derived from the law of mass action. *J. Phys. Chem.*, **14**, 271 (1920)
21. A.J. Lotka, *Elements of physical biology.* Williams and Wilkins, Baltimore, 1925
22. A.N. Kolmogorov, Sulla teoria di Volterra della lotta per l'esistenza. *Giorn. Ist. Ital. Attuar.* **7**, 74 (1936)
23. J.D. Murray, *Mathematical Biology.* Springer, Berlin, 1993
24. C.E. Shannon, A mathematical teory of communication. *The Bell System Technical Journal* **27**, 379 and 623 (1948)
25. J.D. Watson and F.H. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**, 737 (1953). Classical article, *Ann. N.Y. Acad. Sci.* **758**, 13 (1995)
26. L. Pauling and R.B. Corey, Two hydrogen-bonded spiral configurations of the polypeptide chain. *J. Am. Chem. Soc.* **72**, 5349 (1950)
27. E. Schrödinger, *What is Life.* Cambridge University Press, Cambridge, 1992. original appeared in 1944
28. T.E. Creighton, *Proteins: Structure and Molecular Properties.* W.H. Freeman and Company, New York, 2000
29. H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology.* W.H. Freeman and Company, New York (2000)
30. D. Holste, I. Grosse, and H. Herzel, Statistical analysis of the DNA sequence oh human chromosome 22. *Phys. Rev. E* **64**, 041917 (2001)

31. C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, Long-range correlations in nucleotide sequences, *Nature* **356**, 168 (1992)
32. R.J. Solomonoff, A formal theory of inductive inference, part i. *Inform. Contr.* **7**,1 (1964)
33. R.J. Solomonoff, A formal theory of inductive inference, part ii. *Inform. Contr.* **7**, 224 (1964)
34. A.N. Kolmogorov, Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii* **1**, 3 (1965). Original in russian, translated in: Three approaches to the quantitative definition of information. *Probl. Inform. Trans.* **1**, 1 (1965)
35. A.N. Kolmogorov, Logical basis for information theory and probability theory. *IEEE Trans. Inform. Theory*, **14**, 663 (1968)
36. G.J. Chaitin, Information-theoretic computational complexity. *IEEE Trans. Inform. Theory* **20**, 10 (1974)
37. G.J. Chaitin, Randomness and mathematical proof. *Scientific American* **232**, 47 (1975)
38. H.J. Bussemaker, Hao Li, and E.D. Siggia, Building a dictionary for genomes: identification for presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* **97**, 10096 (2000)
39. G.J. Chaitin, *Toward a mathematical definition of "Life"* in R.D. Levine and M. Tribus, The Maximum Entropy Formalism, MIT Press, 1979, 477
40. C. Levinthal, Are there Pathways for Protein Folding? *J. Chem. Phys.* **65**, 44 (1968)
41. C.B. Anfinsen, Principles that govern the folding of protein chains. *Science* **161**, 223 (1973)
42. W. Kauzmann, Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1 (1959)
43. H.S. Chan and K.A. Dill, Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins* **30**, 2 (1998)
44. P.G. Wolynes, J.N. Onuchic, and D. Thirumalay, Navigating the folding routes. *Science* **267**, 1619 (1995)

Fully Developed Turbulence

Luca Biferale[1], Guido Boffetta[2], and Bernard Castaing[3]

[1] Dept. of Physics and INFM, University of Tor Vergata, Via della Ricerca Scientifica 1, Rome, Italy, 00133, Luca.Biferale@roma2.infn.it
[2] Dept. of General Physics and INFM, University of Torino, Via P.Giuria 1, Torino, Italy, 10125, boffetta@to.infn.it
[3] Ecole Normale Superieure de Lyon, 46 Allee d'Italie, Lyon, France, 69364 Lyon Cedex 07, bcastain@ens-lyon.fr

Abstract. The contribution of ANK to the foundation of a statistical theory of fully developed turbulence cannot be overestimated. His fundamental papers radically changed physicists' approach to the study of turbulence and are still source of inspiration for recent developments in the fields.

The intention of this paper is to give an up-to-date vision of our knowledge about fully developed turbulence, with particular emphasis on the evolution of Kolmogorov's pioneering ideas. We start from the discussion of what is now called the "Kolmogorov 1941" theory of turbulence and its relation with the state of the art at that time. The second important contribution of Kolmogorov (the "refined similarity hypothesis" of 1962) is discussed in relation to the Landau objections to the original 1941 theory. The last part of the review is devoted to recent developments and gives a modern view of concepts, such as isotropy and universality of small scale fluctuations, already present in the original papers by ANK.

1 Richardson Cascade

The Navier-Stokes equations describing the spatio/temporal evolution of incompressible three-dimensional flows are known since Navier (1823):

$$\begin{aligned} &\partial_t \boldsymbol{v} + \boldsymbol{v} \cdot \boldsymbol{\partial} \boldsymbol{v} = -\boldsymbol{\partial} P + \nu \Delta \boldsymbol{v} \\ &\boldsymbol{\partial} \cdot \boldsymbol{v} = 0 \end{aligned} \tag{1}$$

where with $\boldsymbol{v}(\boldsymbol{x}, t), P(\boldsymbol{x}, t), \nu$ we denote the local instantaneous velocity field, the pressure field and the molecular viscosity, respectively. Nevertheless, as noted by Frisch [1] this is one of those interesting situations where the perfect knowledge of the deterministic equation of motion does not help very much in the understanding of the global and local behavior of the flow, neither for short nor long times. Very few rigorous results are known on the behavior of (1), especially in the *fully developed turbulent* regime, i.e. when the external forcing mechanism injects enough energy to produce a chaotic *unpredictable* flow evolution. Fully developed turbulence is described by the Navier-Stokes equations (1) in the regime of high Reynolds numbers Re, i.e. in the limit when the ratio between the non-linear terms and the viscous, linear, contribution become large: $Re \sim \frac{v \partial v}{\nu \Delta u} \to \infty$.

L. Biferale, G. Boffetta, and B. Castaing, Fully Developed Turbulence, Lect. Notes Phys. **636**, 149–172 (2003)
http://www.springerlink.com/

In this chapter we review the seminal contributions of A.N. Kolmogorov on the phenomenological, theoretical and analytical understanding of turbulent flows, starting from the pioneering works in the 1939-41 years [4,5] and ending with the paper published in the occasion of the *Marseille Conference* in 1962 [2]. We also discuss the importance and fallout of A.N. Kolmogorov's ideas on modern turbulence research, in both experiments and numerical simulations.

The basic brick of the entire work of Kolmogorov was the celebrated Richardson energy cascade, as he himself stated in the 1962 paper [2]:

> *"[...] The hypothesis concerning the local structure of turbulence at high Reynolds number, developed in the years 1939–41 were based physically on the Richardson's idea of the existence in the turbulent flow of vortices on all possible scales $\eta < r < L$ between the external scale L and the internal scale η and of certain uniform mechanism of energy transfer the coarser-scaled vortices to the finer".*

The Richardson cascade is a pictorial and phenomenological way to explain the transfer of fluctuations from the largest scale in the system, L to the dissipative scale η where turbulent fluctuations are overwhelmed by the viscous term. Richardson proposed that the energy transfer from the pumping scale to the dissipative scales happens through a multi-step process, where eddies break up to smaller eddies, small-eddies break up to smaller and smaller eddies and so on "to viscosity". The main assumption is the locality of interactions in Fourier space, i.e. there is not any important direct transfer of energy connecting very large scales with very small scales. Moreover, respecting the scale-invariance of NS equations, in the limit of vanishing viscosity, the mechanism is supposed to happen in a self-similar way ; typical velocities, scales and times of eddies of size ℓ can be estimated on the basis of dimensional analysis without the introduction of any *outer* or *inner* lengths. Namely, by denoting with $\delta \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\ell}, t) = \boldsymbol{v}(\boldsymbol{x} + \boldsymbol{\ell}) - \boldsymbol{v}(\boldsymbol{x})$ the velocity increments over a scale ℓ, the typical local eddy turn-over time, needed for a turbulent fluctuations of scale ℓ to change its energy content, becomes: $t_\ell = \ell / \delta v(\ell)$. Richardson's scenario predicts that typical eddy-turn-over times of eddies of size ℓ become faster and faster by going to smaller and smaller scales: $t_{\ell'} < t_\ell$, if $\ell' < \ell$. In the cascade process, small-scale velocity fluctuations may therefore expect to loose the memory of the large-scale fluctuations. This is a qualitative explanation of the supposed – and experimentally observed – *universality* of small-scale turbulent fluctuations. By *universality*, we mean the independence of large-scale forcing mechanism used to maintain the flow. Universality of small-scale fluctuations is also strictly linked with the so-called *return-to-isotropy*, i.e. to the assumption that small-scale turbulent statistics is dominated by the isotropic component. Both self-similar behavior and small-scale universality may hold only far enough from the boundaries. Near the boundaries on the other hand, non-homogeneous effects and the presence of solid walls may strongly perturb the flow at all scales.

In the years 1939-41, A.N. Kolmogorov published a series of papers where the above phenomenological scenario reached a level of solidity such as, later on, people refer to them as *"the 1941 Kolmogorov theory"* or *K41* for short. First, A.N. Kolmogorov listed the minimal set of hypotheses required to deduce in a systematic way the *self-similar* and *universal* character of small-scale fluctuations. Second, he determined quantitative information on the statistical properties of velocity fluctuations at small scales. Third, he derived the 4/5-law, the only known exact result describing the scaling properties of the energy flux at large Reynolds numbers.

Another important contribution of A.N. Kolmogorov was to propose the most suitable set of correlation functions able to highlight small-scale fluctuations, the so-called structure functions. A generic velocity structure function of order p, at scale ℓ, is a p-rank tensor defined in terms of the correlation function between p simultaneous velocity difference components, $\delta\boldsymbol{v}(\boldsymbol{x},\boldsymbol{\ell},t)$, at scales $\boldsymbol{\ell}$:

$$T^{\{\alpha\}}(\boldsymbol{\ell}) = \langle \delta v_{\alpha_1}(\boldsymbol{\ell}) \dots \delta v_{\alpha_p}(\boldsymbol{\ell}) \rangle, \tag{2}$$

where by the shorthand notation, $\{\alpha\}$, we denote the set of p vectorial indices, $\alpha_1, \dots, \alpha_p$. In (2) we have dropped both dependencies on time, t, and on the spatial location, $\boldsymbol{x}$, because we assume to be in a stationary and homogeneous statistical ensemble. A simpler set of observables is made of the *longitudinal structure functions*, i.e. the projections of the p-th rank tensor (2) on the unit separation vector, $\hat{\boldsymbol{\ell}}$:

$$S^{(p)}(\boldsymbol{\ell}) = \langle [\delta\boldsymbol{v}(\boldsymbol{\ell}) \cdot \hat{\boldsymbol{\ell}}]^p \rangle \tag{3}$$

which, for an isotropic ensemble, are scalar functions that depend only on the modulus of the separation, $\ell = |\boldsymbol{\ell}|$.

This contribution is organized as follow. In Sect. 2, we review the main bricks of the 1941 theory, focusing in particular on (i) the idea that turbulent flows are characterized by a statistical recovery of symmetries at scales small-enough, (ii) the experimental evidence for the existence of a *dissipative anomaly*, i.e. that the energy dissipation tends to a non-vanishing constant in the limit of high Reynolds numbers; (iii) the importance of the 4/5 law and, finally, on the universality of small scale statistics. In Sect. 3 we review the important step forward made by A.N. Kolmogorov in 1962, in response to the criticisms made by Landau and Obhukov of the 1941 theory. There, A.N. Kolmogorov further elaborated his 1941 theory to include also the possibility to describe *intermittency* and *anomalous scaling* of small scale statistics. In Sect. 4 we discuss the modern consequences of Kolmogorov's contributions, including recent issues raised by experimental and numerical evidence in – apparent – contrast with some of the hypotheses at the basis of 1941 and 1962 theories.

2 Kolmogorov 1941 Theory

Fully developed turbulence is a spatio-temporal disordered motion involving a wide range of scales. Large scales, of the order of the flow domain, are the most important from an engineering point of view; they contain most of the energy and dominate the transport of momentum, mass and heat. Small turbulent scales, including the so-called inertial scales and dissipative scales, are more interesting for a fundamental approach as they display properties which are universal with respect to the flow configuration and/or forcing mechanism. The fundamental contribution of A.N. Kolmogorov to the theory of turbulence is based on a statistical description of the fluctuations of these small scales, leading to very simple and universal predictions. His first work is contained in three short papers which appeared in the USSR in 1941 and which constitute what is now called the K41 theory.

2.1 Symmetries for Navier–Stokes

At the basis of the concept of universality is the idea that small scale turbulence, at sufficiently high Reynolds numbers, is statistically independent of the large scales and can thus locally recover homogeneity and isotropy. The concept of homogeneous and isotropic turbulence was already introduced by Taylor [6] for describing grid generated turbulence. The important step made by A.N. Kolmogorov in 1941 was to postulate that small scales are statistically isotropic, no matter how the turbulence is generated. This hypothesis is based on intrinsic properties of the dynamics, i.e. the invariance of the Navier-Stokes equations (1) under the following set of symmetries [1]:

- space translations: $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{r}$
- space rotations: $(\boldsymbol{x}, \boldsymbol{v}) \to (A\boldsymbol{x}, A\boldsymbol{v})$ with $A \in SO(3)$
- scaling: $(t, \boldsymbol{x}, \boldsymbol{v}) \to (\lambda^{1-h} t, \lambda \boldsymbol{x}, \lambda^h \boldsymbol{v})$ for any h and $\lambda > 0$.

The first two symmetries are consequences of space homogeneity and isotropy and are broken in the presence of boundaries or a forcing mechanism. As for the scaling symmetry, a classical example is the so-called similarity principle of fluid mechanics which states that two flows with the same geometry and the same Reynolds number are similar. The similarity principle is at the basis of laboratory modeling of engineering and geophysical flows.

The idea at the basis of Kolmogorov's treatment of small scale turbulence is the hypothesis that, in the limit of high Reynolds numbers and far from boundaries, the symmetries of Navier-Stokes equation are restored for statistical quantities. To be more precise, let us consider the velocity increment $\delta \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\ell})$ over the scales $\ell \ll L$. Restoring the homogeneity in a statistical sense requires $\delta \boldsymbol{v}(\boldsymbol{x} + \boldsymbol{r}, \boldsymbol{\ell}) \stackrel{law}{=} \delta \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\ell})$, where equality in law means that the probability distribution function (PDFs) of $\delta \boldsymbol{v}(\boldsymbol{x} + \boldsymbol{r}, \boldsymbol{\ell})$ and $\delta \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\ell})$ are identical. Similarly, statistical isotropy, also used by Kolmogorov in his 1941

papers, requires $\delta \boldsymbol{v}(A\boldsymbol{x}, A\boldsymbol{\ell}) \stackrel{law}{=} \delta A\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\ell})$ where A is a rotation matrix. The issue of restoring the small scale isotropy will be discussed in details in Sect. 4.

In the limit of large Reynolds number, the second Kolmogorov similarity hypothesis [4] states, that for separation in the inertial range of scales $\eta \ll \ell \ll L$, the PDF of $\delta \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\ell})$ becomes independent on viscosity ν. As a consequence, in this limit and range of scales, scaling invariance is statistically recovered with a larger freedom in the values of the scaling exponent:

$$\delta \boldsymbol{v}(\boldsymbol{x}, \lambda \boldsymbol{\ell}) \stackrel{law}{=} \lambda^h \delta \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\ell}). \tag{4}$$

The values of the scaling exponent, h, are now limited only by requiring that the velocity fluctuations do not break incompressibility, so that $h \geq 0$ [1]. In this section we discuss the self-similar $K41$ theory, limiting our solves to the case where the turbulent flow possesses a global scaling invariance with a unique scaling exponent $h = 1/3$. In Sects. 3.3 we relax the requirement of global scaling invariance and we consider a possible extension of the theory to the case of local scaling invariance.

2.2 Dissipative Anomaly

Because we are interested in the statistical properties of small scale turbulence far from boundaries, we consider in the following the simple geometry without boundary of a periodic box of size L. This geometry is also very convenient for numerical simulations which can profit of the Fourier decomposition of the velocity field. In the inviscid limit $\nu = 0$, the Navier-Stokes equation (1) conserves the kinetic energy

$$E = \langle \frac{1}{2} v^2 \rangle \equiv \frac{1}{V} \int_V v^2(\boldsymbol{x}, t) d^3x, \tag{5}$$

where the brackets denote the average over the periodic box of volume $V = L^3$. Indeed, using (1) in (5), one obtains:

$$\frac{dE}{dt} = -\langle \frac{1}{2} \sum_{ij} \nu(\partial_i v_j + \partial_j v_i)^2 \rangle \tag{6}$$

and thus the mean energy dissipation ε vanishes for $\nu = 0$. Expression (6) introduces the most important experimental fact in the theory of fully developed turbulence: In the limit $Re \to \infty$ (which is equivalent to the limit $\nu \to 0$) the mean energy dissipation attains a finite limit ε [3,7,8]. This is actually true far from boundaries, in addition to boundary layers in general possessing a Re dependence. Formally, we have

$$\lim_{\nu \to 0} \frac{dE}{dt} = -\varepsilon \neq \frac{dE}{dt} \|_{\nu=0} \tag{7}$$

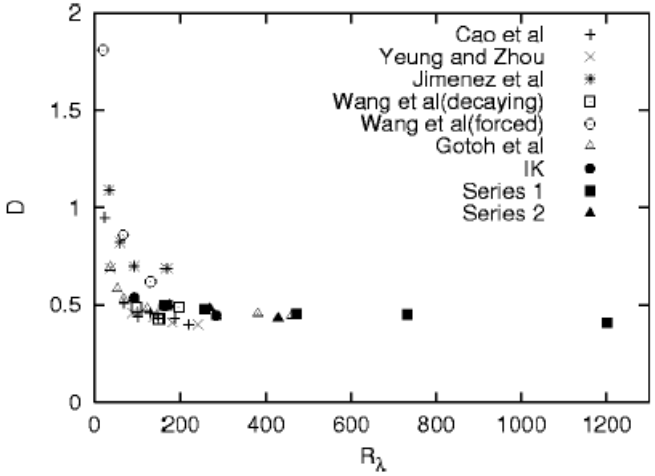

Fig. 1. Dimensionless mean energy dissipation $D = \varepsilon L/U^3$ as a function of $R_\lambda = \sqrt{15Re}$ from different direct numerical simulations (from [8])

which is the essence of the so-called *dissipative anomaly*. In Fig. 1 we present a collection of numerical data on the Reynolds number dependence of mean energy dissipation, supporting the existence of a finite asymptotic for $Re \to \infty$.

The observation that kinetic energy is asymptotically dissipated at a constant rate has the important consequence that the flow develops small scales in order to compensate the $\nu \to 0$ limit in (6). This picture was already in the mind of Richardson about twenty years before Kolmogorov [9] when he developed a qualitative theory of turbulence based on the idea of a turbulent cascade in which the energy is transferred from the largest scales (where energy is injected by mechanical forcing) down to the smallest scales (where it is dissipated by molecular viscosity). The intermediate range of scales at which energy is simply transferred, the inertial range, can be expected to display universal statistics, independent of the forcing mechanism and of the Reynolds number. Kolmogorov 1941 theory is the first attempt at a quantitative theory of inertial scale statistics.

2.3 The 4/5 Law and Self-Similarity

In the first of his 1941 papers, A.N. Kolmogorov made the fundamental assumption that in the inertial range of scales the distribution function for $\delta \boldsymbol{v}(\boldsymbol{\ell})$ depends on $\boldsymbol{\ell}$ and ε only (and not on viscosity). As a consequence, dimensional analysis leads, assuming self-similarity, to the power-law behavior for the structure functions in the inertial range

$$S^{(p)}(\ell) = C_p \varepsilon^{p/3} \ell^{p/3}, \tag{8}$$

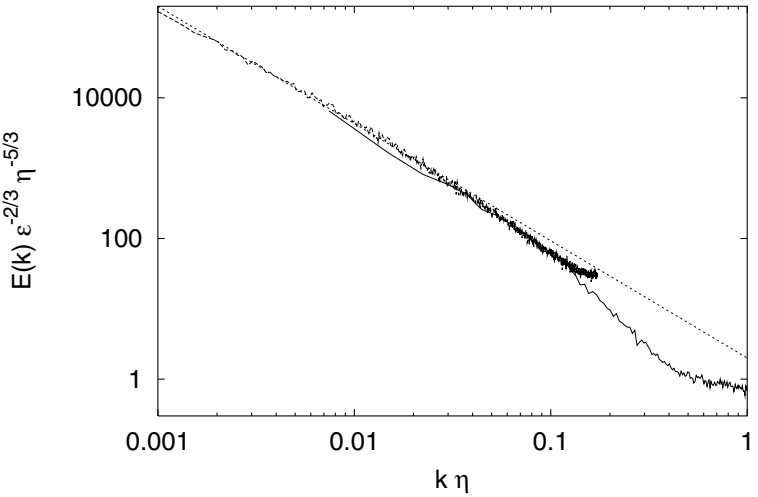

Fig. 2. Velocity power spectra at $R_\lambda = 230$ (*thin line*) and $R_\lambda = 1840$ (*thick line*) normalized with Kolmogorov scale and velocity. The *dashed line* represents the Kolmogorov spectrum, $E(k) = Ck^{-5/3}$ with $C = 2$ (from [11])

where the C_p are dimensionless, universal and constant. In particular he made the prediction for the second order structure function $S^{(2)}(\ell)$ or, equivalently [3], for the energy spectrum

$$E(k) = C\varepsilon^{2/3}k^{-5/3}. \tag{9}$$

Since the work of Grant et al. [10] in a tidal channel, there have been many experiments (both in field and in laboratory) and numerical simulations showing an energy spectrum following Kolmogorov's prediction (9) with constant $C \simeq 2.0$ [3,1]. We now know that the spectrum (9) is not exact, as intermittency induces some deviations from the Kolmogorov exponent $-5/3$. Nevertheless, these effects are small and their existence has been definitively accepted only very recently (see the discussion in Sect. 4). For this reason, K41 theory is still important not only for historical reasons but also for many applications in modeling engineering and geophysical flows.

In Fig. 2, we present an example of experimental energy spectrum obtained from a water jet experiment. Note that the inertial range, showing the Kolmogorov scaling (9), increases with Reynolds number.

In his third 1941 paper, A.N. Kolmogorov was able to obtain, starting from Navier-Stokes equation, the exact expression for the third-order longitudinal structure function, called the "4/5 law" after the numerical prefactor:

$$S^{(3)}(\ell) = -\frac{4}{5}\varepsilon\ell. \tag{10}$$

This is still the only exact dynamical result on the behavior of velocity differences in the inertial range. The fact that the third moment of velocity

differences does not vanish is a consequence of the directional transfer (from large to small scales) of energy on average. An important consequence, which will be discussed in detail in connection with the intermittency issue, is that the PDF of velocity differences cannot be Gaussian. In the following, we will give a dimensional argument for the K41 predictions. The interested reader can find in [1] a complete derivation of the 4/5 law.

Let us consider the typical velocity fluctuation v_ℓ at scale ℓ, for example the rms of the velocity difference $\delta v(\ell)$. The eddy turnover time, i.e. the characteristic time of variation of v_ℓ, can be estimated dimensionally as $t_\ell \sim \ell/v_\ell$. This is also the time for the transfer of kinetic energy to smaller scales in the cascade. The flux of kinetic energy from scale ℓ to smaller scales is thus

$$\Pi_\ell \sim \frac{v_\ell^2}{t_\ell} \sim \frac{v_\ell^3}{\ell}. \tag{11}$$

Because in the inertial range the energy is neither injected nor dissipated, the energy flux (11) has to be independent of the scale ℓ and thus it must balance the mean energy dissipation ε. From (11) one obtains the prediction

$$v_\ell^3 \sim \varepsilon \ell, \tag{12}$$

which is the dimensional analog of the four-fifths law (10). We observe that in this dimensional derivation (as in the complete derivation of (10) from Navier–Stokes equation) we have only made use of the hypothesis of finite energy dissipation in the $Re \to \infty$ limit. Under the additional assumption of self-similarity, the scaling exponent in (4) is forced to the value $h = 1/3$ by (12) and thus one obtains the K41 prediction for structure function exponents, $S^{(p)}(\ell) = C_p \ell^{\zeta(p)}$ with $\zeta(p) = p/3$ (8).

The inertial range has a viscous small scale cutoff at the Kolmogorov scale η. This scale can be dimensionally estimated as the scale at which the local Reynolds number $v_\ell \ell/\nu$ (which measures the relative importance of the inertial and viscous term in (1)) is of order unity. Using the Kolmogorov scaling one obtains

$$\eta \simeq LRe^{-3/4}. \tag{13}$$

Below the dissipative scale, the energy spectrum presents an exponential or more than exponential decay as a consequence of the smoothness of the velocity field. The extension of the inertial range, L/η, thus increases as $Re^{3/4}$.

We conclude this section by introducing the problem of intermittency. It is now accepted that K41 theory is not exact because higher order structure functions display unambiguous departure from the scaling exponents (8):

$$S^{(p)}(\ell) = C_p \left(\frac{\ell}{L}\right)^{\zeta(p)} \quad \text{with} \quad \zeta(p) \neq p/3. \tag{14}$$

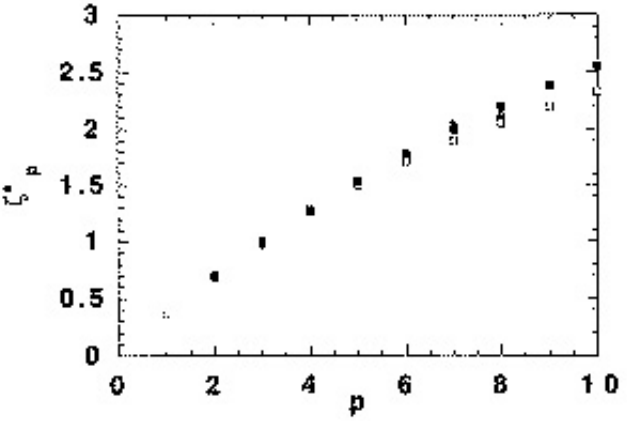

Fig. 3. Velocity structure function exponents ζ_p for different experimental conditions. From Arneodo et al., *Europhys. Lett.* **34**, 411 (1996)

In Fig. 3 we reproduce a famous collection of scaling exponents $\zeta(p)$ extracted from several experiments [30] using the so-called ESS procedure [31]. Let us recall that the scaling exponents are not completely free, since (11) still requires $\zeta(3) = 1$. Under very general hypotheses, one can also demonstrate that ζ_p has to be a concave and nondecreasing function of p [1]. From Fig. 3 it is evident that the $\zeta(p)$ exponents are firstly *universal* and secondarly *anomalous*, i.e. they are expressed by a non-linear function of p. This means that the PDFs of velocity differences $\delta \boldsymbol{v}(\ell)$ vary as a function of length scale ℓ and the skewness of the velocity differences increases without bound as we go to small scales (similarly, skewness based on longitudinal gradients increases as a function of Reynolds number, instead of staying constant as predicted by the K41 theory).

3 Kolmogorov 1962 Theory

In September 1961, a Colloquium on turbulence was organized by the CNRS in Marseille, France. Organizers asked A.N. Kolmogorov to give a talk on the state of the knowledge. He took this opportunity to emphasize a contemporary contribution of Oboukhov, gave the new hypotheses which could justify it, and its consequences. Due to their importance, this contribution and that of Oboukhov were translated into English and published by the Journal of Fluid Mechanics ("under the Editor responsibility") in two short papers [2]. They constitute the basis of the Kolmogorov-Oboukhov 62 (KO62) theory. At that time, there were no strong experimental motivations to call for an improvement of the 1941 theory. The main criticisms were theoretical. The motivation invoked by Kolmogorov himself in his paper lays in a remark by L. Landau [2]:

"But quite soon after [the K41 hypotheses] originated, Landau noticed that they did not take into account a circumstance which arises directly from the essentially accidental and random character of the mechanism of transfer of energy from the coarser vortices to the finer".

Indeed, when looking at Landau's various remarks, as reported for instance by Frisch [1], one can see that he emphasizes two points. First, that the constants, such as the "Kolmogorov constant" C_2 in the K41 expression for the second order longitudinal structure function:

$$S^{(2)}(\ell) = C_2 \epsilon^{2/3} \ell^{2/3}, \tag{15}$$

cannot be universal. As $\langle \epsilon^{2/3} \rangle$ differs from $\langle \epsilon \rangle^{2/3}$, C_2 must depend on the distribution of ϵ at large scales, close to the integral scale, which cannot be universal. The first point is not even discussed by Kolmogorov. There will be no universality for constants. The second point is the one that Kolmogorov emphasizes, and it is directly connected to the issue of intermittency introduced in the previous section: vorticity, and thus probably dissipation too, is known through direct observations to be concentrated in tiny regions of the flow. This may lead to *anomalous* values for the scaling exponents, $\zeta(p)$, of velocity structure functions in the inertial range (14). To take into account this point, Kolmogorov emphasizes the role of the local dissipation. Namely, by denoting with

$$\epsilon_\ell(\mathbf{x}, t) = \frac{1}{4/3\pi\ell^3} \int_{|y|<\ell} d\boldsymbol{y}\, \epsilon(\boldsymbol{x} + \boldsymbol{y}, t) \tag{16}$$

the coarse grained energy dissipation on a ball of radius ℓ centered on $\mathbf{x}$, and with $\delta v(\ell)$ the velocity difference over a scale ℓ, Kolmogorov postulated that the non dimensional ratio:

$$\frac{\delta v(\ell)}{\epsilon_\ell^{1/3} \ell^{1/3}} \tag{17}$$

has a probability distribution independent of the local Reynolds number $Re_\ell = \delta v(\ell)\ell/\nu$, in the limit $Re_\ell \to \infty$. This is the Refined Similarity Hypothesis (RSH). It links the scaling laws of velocity structure functions with the scaling properties of the energy dissipation:

$$S^{(p)}(\boldsymbol{\ell}) = \langle [\delta \boldsymbol{v}(\boldsymbol{\ell}) \cdot \hat{\boldsymbol{\ell}}]^p \rangle = C_p \langle \epsilon_\ell^{p/3} \rangle \ell^{p/3}. \tag{18}$$

In a second group of hypotheses, Kolmogorov assumes that $\ln(\epsilon_\ell)$ has a Gaussian distribution, whose variances behaves as:

$$\sigma_\ell^2 = A + 9\mu \ln(L/\ell), \tag{19}$$

where experiments give $\mu \simeq 0.025$ [45,46]. We shall first address the consequences of these hypotheses, discussing intermittency and its physical interpretations. Then we shall go to the experimental verifications, and the discussions around the hypotheses.

3.1 Intermittency and Anomalous Scaling

From the log-Gaussian assumption it is easy to derive a parabolic shape for the behavior of the scaling exponents $\zeta(p)$. In particular, the behavior of the flatness factor $F(\ell)$, which characterizes the shape of the distribution of $\delta v(\ell)$ is given by:

$$F(\ell) = \frac{\langle \delta v^4(\ell) \rangle}{\langle \delta v^2(\ell) \rangle^2} \sim \ell^{\zeta(4)-2\zeta(2)} \sim \ell^{-4\mu}. \tag{20}$$

For a Gaussian distribution, $F(\ell) = 3$, independent of the scale. Larger values of $F(\ell)$ correspond to thicker wings for the distribution compared to its center. Velocity differences much larger than their rms are more probable as ℓ becomes smaller. This evolution of the PDF of velocity differences is called *intermittency*. It is well observed [47], both along the scales in the inertial range ($\eta < \ell < L$) and *via* its consequences for the PDF of velocity gradients [48,15]. Let us now present a simple phenomenological interpretation of this PDF evolution.

3.2 Multiplicative Cascade

As suggested by Kolmogorov himself, the laws he proposed can follow from very simple arguments [22,23]. Let us consider a set of reference scales $\ell_n = Lb^{-n}$ in the inertial range with b, the inter-scale ratio, usually chosen equal to 2. The simplest relation one may define among the PDFs of velocity differences at scales ℓ_n and ℓ_{n+1} is given by a multiplicative convolution:

$$\delta v(\ell_{n+1}) = \alpha_{n+1,n} \delta v(\ell_n), \tag{21}$$

where $\alpha_{n+1,n}$ denotes a suitable stochastic variable. The logarithm of $\alpha_{n+1,n}$, $\phi = \ln(\alpha_{n+1,n})$, has a probability distribution $G_{n+1,n}[\phi]$ which defines the statistical dependence between eddies of different sizes. Intermittent scaling laws are then easily obtained through three assumptions:

- Scale invariance: $G_{n+1,n} = G_b$
- $\alpha_{n+1,n}$ and $\delta v(\ell_n)$ are uncorrelated
- $\alpha_{n+1,n}$ and $\alpha_{n+2,n+1}$ are uncorrelated.

Then one obtains for the scaling of longitudinal structure functions:

$$S^{(p)}(\ell) = C_p \left(\frac{\ell}{L} \right)^{\zeta(p)}, \tag{22}$$

with the scaling exponents uniquely fixed in terms of the basic distribution $G_b[\phi]$: $\zeta(p) = \log_b \langle \exp(p\phi) \rangle$. In (22) the constants C_p depend on the –

non universal – velocity distribution at the integral scale, L. By setting $\alpha_{n+1,n} = \epsilon_{\ell_{n+1}}^{1/3} \ell_{n+1}^{1/3} / \epsilon_{\ell_n}^{1/3} \ell_n^{1/3}$, the multiplicative model recovers the Refined Similarity Hypothesis (RSH) of Kolmogorov. The log-normal distribution proposed by Kolmogorov in the 1962 is the only one possible choice for the basic distribution G_b. The above model ensures the growth of the variance of $\ln(\epsilon_\ell)$ in agreement with (19). The Central Limit Theorem is of no help since the evaluation of $\langle \epsilon_\ell^{p/3} \rangle$ involves parts of the distribution which are out of its range.

Note that the presentation of the above multiplicative cascade model can be made without using the discretized inter-scale ratio b [47,14]. An attempt was made to characterize the multiplicative stochastic process in terms of *structures* in the flow [13]. Also, important developments have been achieved using deterministic dynamical models for the turbulent energy cascade (Shell Models), originating from the pioneering ideas of the Russian school [24,1,25,26].

Through the Refined Similarity Hypothesis, the intermittency built by multiplicative models for velocity differences can be directly translated to the scaling properties of the coarse-grained energy dissipation, $\epsilon(\ell)$. The Refined Kolmogorov Hypothesis can therefore be read as a statement about the multifractal properties of three-dimensional measure defined by the local energy dissipation.

3.3 Multifractal

The above process yields the concentration of the dissipation on small spots, which corresponds to the Landau remark [1]. This in turn suggests a geometrical, rather than dynamical, interpretation of intermittency: intense spots surrounded by calm regions. This has been generalized and formalized by Parisi and Frisch [50], in the multifractal picture of turbulence. The starting point is the invariance of the Navier-Stokes equations under scaling transformation (4). If no characteristic scale appears in the inertial range, there should exist locally self-similar solutions where velocity differences scale with ℓ^h. In general, different exponents h are possible, but, again due to scale invariance, the subset of points where the exponent is h should be self-similar, with fractal dimension $D(h)$.

Covering the flow, whose size is L, with balls of size ℓ requires $(L/\ell)^3$ balls. The regions that contain a point where the exponent is h add to $(L/\ell)^{D(h)}$. Therefore the probability to find an exponent h at size ℓ is:

$$P(h) = (L/\ell)^{D(h)-3}. \tag{23}$$

In the neighborhood of such a point, velocity differences at scale ℓ should be of order $\delta v(L)(\ell/L)^h$. Thus an estimate of $\langle \delta v^p(\ell) \rangle$ is given by:

$$\langle \delta v^p(\ell) \rangle \simeq \int \delta v^p(L)(\ell/L)^{ph} P(h) dh \simeq \delta v^p(L) \int (\ell/L)^{3+ph-D(h)} dh. \tag{24}$$

Such an integral, in the limit $\ell/L \to 0$, can be estimated by the saddle-point method:

$$\langle \delta v^p(\ell) \rangle \sim (\ell/L)^{\zeta(p)} \text{ where } \zeta(p) = min_h(3 + ph - D(h)). \tag{25}$$

The fractal dimension of singularities $D(h)$ is not completely free, since (25) must be consistent with the 4/5 law (10), i.e. $\zeta(3) = 1$. This imposes the inequality $D(h) \leq 3h+2$ and that there exist a value h for which the equality holds.

In spite of apparent differences, the present formalism is close to the previous one. In particular, it can be verified that the log-normal hypothesis corresponds to a parabolic shape for $D(h)$:

$$D(h) = 3 - \frac{(h - h_o)^2}{2\mu}. \tag{26}$$

The original K41 theory is recovered assuming perfect self-similarity, i.e. a single scaling exponent h_0 on the whole domains, i.e. $D(h_0) = 3$. The above inequality on $D(h)$ then imposes $h_0 = 1/3$.

3.4 Tests of Kolmogorov Hypotheses

A large body of work in the past decades has been devoted to test the various hypotheses implicitely or explicitely proposed by Kolmogorov.

The Refined Similarity Hypothesis

An experimental validation of the RSH requires large conditional statistics which were possible only recently. Even though attempts had been proposed earlier, a systematic study was made in [51], where a one-dimensional record of turbulent velocity measured by a hot wire was used. By using Taylor's Hypothesis, the estimation of the probability to have a coarse-grained energy fluctuation $\epsilon(\ell)$ for a given velocity increment, $P[\epsilon(\ell)|\delta v(\ell)]$ gave a Gaussian distribution at large ℓ and a bimodal distribution at small ℓ. This result was corrected by Gagne et al. [52], where Gaussian statistics at all scales with a variance depending both on the scale and $\epsilon(\ell)$ was obtained, in agreement with Kolmogorov's law. But for the test to have a physical meaning, it should not give positive result with any random signal. Indeed, applying the same procedure to a temperature signal instead of velocity gave the same Gaussian conditional statistics, while it should not [53]. The uncontestable success of this test could then be due to a statistical effect and to the long-range correlation of dissipation (see later).

The Log-Normal Distribution

The log-normal hypothesis for ϵ_ℓ was addressed [54]. The quantity which plays the role of $\epsilon(\ell)$ is not easy to measure. But even within the general frame of

multiplicative cascades, without refering to dissipation, theorems show that a pure Gaussian form for G_b would yield inconsistencies in the far tails of velocity difference distributions. Moreover, log-normal distribution violates a condition for the boundedness of the velocity field for large enough Reynolds numbers [1]. However, these objections need not be taken into account, since they concern incredibly small levels of probability. To take an example, the Fourier heat law allows (small) signals to propagate faster than the velocity of light, but nobody would discard it for this reason. So far, all serious attempts to demonstrate deviations from the log-normal predictions have failed.

The Fokker-Planck Approach

Some recent approaches [60,61], looking directly at multipliers statistics, could test their correlations. They experimentally verified the Markovian properties of the cascade (independence between successive steps). They obtained a Fokker-Planck equation for the velocity difference PDF $P_\ell[\delta v(\ell)]$:

$$\frac{\partial P_\ell}{\partial \ln(L/\ell)} = \frac{\partial (D_1 P_\ell)}{\partial \delta v(\ell)} + \frac{\partial^2 (D_2 P_\ell)}{\partial \delta v(\ell)^2}, \tag{27}$$

where the functions $D_1[\delta v(\ell)]$ and $D_2[\delta v(\ell)]$ can be experimentally determined. A pure multiplicative cascade would give D_1 linear and D_2 purely quadratic. Deviations appear at large scales, but these functions seem to converge on this behaviour as ℓ decreases.

Scale Invariance

An extension of scale invariance has been proposed in the following way [47,57]: instead of using the scale separation, b, as a scaling factor, is it possible to find scaling laws in terms of more general scaling functions? Is it possible to absorb finite-Reynolds number effects with a suitable redefinition of the scaling? One possibility is to use the third order longitudinal structure function, $S^{(3)}(\ell)$ as a reference function. This is because we know from the 4/5 law that it is exactly scale invariant in the limit of infinite Reynolds number. Then, we may look at the scaling of different orders of structure functions with respect to the third-order function. This is the Extended Self Similarity (ESS), discovered independently by Benzi and coworkers [31]. Confirmation of this extension of scaling was found also when looking at the flatness $\langle \delta v(\ell)^4 \rangle / \langle \delta v(\ell)^2 \rangle^2$ *versus* $\langle \delta v(\ell)^2 \rangle$ instead of ℓ. However, these effects have another possible explanation [45,58]. In the inertial range, typical velocity differences $\delta v(\ell)$ scale as $\ell^{1/3}$ (K41 theory). In the dissipative range, where the local Reynolds number is smaller than 1, they behave as ℓ. Then the decrease of all the $\langle \delta v(\ell)^p \rangle$ is faster when in the dissipative range than in the inertial one. However large energy events contribute more in $\langle \delta v(\ell)^4 \rangle$ than in $\langle \delta v(\ell)^2 \rangle$. $\langle \delta v(\ell)^4 \rangle$, thus enters the dissipative range at smaller ℓ than $\langle \delta v(\ell)^2 \rangle$

[62], and the flatness $\langle \delta v(\ell)^4 \rangle / \langle \delta v(\ell)^2 \rangle^2$ rapidly increases. For high Re, this growth is much faster than the decrease of $\langle \delta v(\ell)^2 \rangle$ and the self-similarity cannot be extended in this "Intermediate Dissipative Range" [62]. However, for moderate Re, the two behaviours are accidentally similar, which gives the illusion of ESS.

However, ESS cannot improve the precision on ζ_p, since it fails to extrapolate to high Re [45,58].

Two Points Correlation

It was soon recognized [3] that the Kolmogorov hypotheses implied long-range correlation between local dissipations, therefore velocity differences. This was checked only recently, by J. Delour et al. [46]. They considered the "magnitude" (in the astronomical sense) $M(\boldsymbol{x}) = \ln |\delta v(\ell_c)|$ of an interval of length ℓ_c centered on $\boldsymbol{x}$, and the correlation of this magnitude between two points distant of $\boldsymbol{\ell}$:

$$\langle M(\boldsymbol{x}) M(\boldsymbol{x} + \boldsymbol{\ell}) \rangle - \langle M(\boldsymbol{x}) \rangle^2. \tag{28}$$

with $\ell > \ell_c$. After a rapid variation due to the short-range correlation of the signs of $\delta v(\ell_c)$, they observe a slow decrease, with the correlation proportional to $\ln(L/\ell)$.

This is in perfect agreement with a multiplicative cascade model. Using the notations above, let us assume that $\ell_c = \ell_n = b^{-n} L$ and $\ell = \ell_j = b^{-j} L$. Then, in the expression of $\delta v(\ell_c)$:

$$\delta v(\ell_c) = \alpha_{n,n-1} \cdots \alpha_{j,j-1} \cdots \alpha_{1,0} L, \tag{29}$$

all multipliers $\alpha_{j,j-1}, \cdots, \alpha_{1,0}$ are common to the two velocity differences considered, which reduces the variance of the logarithm of their ratio. Starting from this observation, they could produce a "Multiplicative Random Walk" having this long-range correlation, with the intermittency properties of real velocity records.

4 Kolmogorov's Legacy on Modern Turbulence

As already stated, Kolmogorov's theories are founded on two major statements. First, that for high enough Reynolds numbers – and far from the boundaries – the small scales of turbulent flows are dominated by isotropic fluctuations. Second, that small-scale fluctuations are *universal*, i.e. independent of the large-scale set up. Both properties are strongly linked. Anisotropic fluctuations are injected in the flow only through the large-scale forcing (or by boundary effects). Therefore, any anisotropic fluctuations left at small scales must be seen as the legacy of the large-scale physics. Moreover, the total amount of small-scale anisotropy cannot be universal, being the direct effect of the particular forcing used to maintain that particular flow.

On the whole, experimental tests of Kolmogorov's theories ran into increasing difficulties when the data was analyzed with greater detail. The first systematic attempt to check the isotropic scaling (8) for high Reynolds number turbulence was done by Anselmet et al. [17]. The authors performed, for the first time, a high-order statistical test of K41 theory by going beyond the usual two-point correlation. They also measured structure functions of higher order, reaching the conclusion that *anomalous* deviations to the $p/3$ scaling exponents existed. At that time, and for many year later, the situation was very controversial both theoretically and experimentally. Many claims that the observed deviations were due to sub-leading finite-Reynolds number effects appeared. One should not underestimate the difficulties of getting reliable estimates of the scaling exponents. First, one must expect finite Reynolds number corrections to *strongly* reduce the extent of the inertial range extension where scaling laws are expected and/or introduce anisotropic corrections to the isotropic K41 predictions. Both effects are usually present in all experiments and numerical simulations. Nowadays, state-of-the-art experiments of turbulence in controlled geometry reach a maximum Reynolds numbers, measured on the gradient scales, of $R_\lambda \sim 5000$, which can be pushed to $R_\lambda \sim 10000$ for atmospheric boundary flows (though highly anisotropic!). The situation of Direct Numerical Simulations (DNS) is more complex, the best resolution ever reached up to now being 4096^3, corresponding to a $R_\lambda \sim 800$. DNS allow a minimization of the anisotropic corrections, thanks to the possibility of implementing periodic boundary conditions and fully isotropic forcing, something which is out of reach in any real experiments. However, even in DNS the discrete symmetries induced by the finite lattice-spacing do not allow for a perfect isotropic statistics. We thus either have high Reynolds number experiments which are strongly perturbed by anisotropic effects, or DNS isotropic flow at moderate Reynolds numbers. Therefore, one has to face the problem of how to disentangle isotropic from anisotropic fluctuations and how to extract information on the asymptotic scaling with a finite – often short – inertial-range extension. Only recently, after many experimental and numerical confirmations of the results of [17], did the situation become clearer [31]. We may affirm now with some degree of certainty that the isotropic scaling exponents are *anomalous*, the K41 prediction $\zeta(p) = p/3$ is wrong, except for $p = 3$ which is fixed to be $\zeta(3) = 1$ by the exact $4/5$ law. More recently, the possibility of analytically showing the existence of anomalous scaling in turbulent advection [30] eliminated the arguments supporting the *impossibility* of having a Reynolds independent anomalous scaling in any hydrodynamical system.

As stated above, according to the Refined Similarity Hypothesis, the anomalous scaling of isotropic structure functions is connected to the multifractal properties of the three-dimensional measure defined in terms of the energy dissipation [1]. It should be noted however that RSH related the inertial range scaling to the scaling of dissipative quantities, and delicate issues connected to small distance expansions and fusion rules are being disregarded

here [27–29]. At any rate, the RSH did not advance the calculation of the scaling exponents beyond crude phenomenology.

4.1 Universality of Small-Scales Fluctuations

Universality of small-scale forced turbulence is at the forefront of both theoretical and experimental investigation of real turbulent flows [1]. The problem is to identify those statistical properties which are robust against changes of the large-scale physics, that is against changes in the boundary conditions and the forcing mechanisms. The brute force method to check universality quickly run into bad problems. In nature or in labs, one finds an enormous variety of turbulent flows with different large-scale physics, as for example channel flows, convective flows, flows maintained by counter-rotating disks to list only a few. The problem is that all of these different experimental set-ups suffer also from very different anisotropic corrections. Therefore it may be very difficult on the experimental data to clearly disentangle the isotropic component. A systematic study of small-scale universality is therefore limited to some degree of uncertainty (see below on the anisotropic fluctuations). Still, the beautiful connection and comparison of data coming from more than 10 different experiments with different large-scale set-ups and with different Reynolds numbers presented in [30] certainly supports the *universal* picture (see also figure 3).

More recently, a different proposal to test small-scale universality was made [12]. The idea is to relate the small-scale universal properties of forced turbulent statistics to those of short-time decay for an ensemble of initial configurations. An immediate remark is that one cannot expect universal behavior for all statistical observables, since the very existence of anomalous scaling is the signature of the memory of the boundaries and/or the external forcing throughout all scales (due to the appearance of the *outer* scale in the expression for structure functions). Indeed, the most one may expect is that the scaling of small-scale fluctuations is universal, at least for forcing concentrated at large scales. The prefactors are not expected to be so. There is therefore no reason to expect that quantities such as the skewness, the kurtosis and in fact the whole PDF of velocity increments or gradients be universal.

This is the same behavior as for the passive transport of scalar and vector fields (see [36] and references therein). For those systems both the existence and the origin of the observed anomalous scaling laws have been understood and even calculated analytically for some cases in the special class of Kraichnan flows [35]. However, carrying over the analytical knowledge developed for linear hydrodynamical problems involves some nontrivial, yet missing, steps. For the Navier-Stokes dynamics, linear equations of motion only appear at the functional level of the whole set of correlation functions.

In schematic form:

$$\partial_t C^{(p)} = \Gamma^{(p+1)} C^{(p+1)} + \nu D^{(p)} C^{(p)} + F^{(p)}, \tag{30}$$

where $\Gamma^{(p+1)}$ is the integro-differential linear operator coming from the inertial and pressure terms and $C^{(p+1)}$ is a shorthand notation for a generic $(p+1)$-point correlation. The molecular viscosity is denoted by ν and $D^{(p)}$ is the linear operator describing dissipative effects. Finally, $F^{(p)}$ is the correlation involving increments of the large-scale forcing $\boldsymbol{f}$ and of the velocity field. Balancing inertial and injection terms gives a dimensional scaling, and anomalously scaling terms must therefore have a different source. A natural possibility is that a mechanism similar to the one identified in linear transport problems may be at work in the Navier-Stokes case as well. The anomalous contributions to the correlation would then be associated with the statistically stationary solutions of the unforced equations (30). The scaling exponents would *a fortiori* be independent of the forcing and therefore be universal. As for the prefactors, the anomalous scaling exponents are positive and thus the anomalous contributions grow at infinity. They should then be matched at the large scales with the contributions coming from the forcing to ensure that the resulting combination vanishes at infinity, as required for correlation functions. Proof of the previous points is still out of analytical reach. Instead, one may check the most obvious catch: the Navier-Stokes equations being integro-differential, non-local contributions may directly couple inertial and injection scales and invalidate the argument. In order to investigate the previous point, a comparison between the behavior of *decaying turbulence* with two different ensembles of initial conditions was made in [12]. The first ensemble contained initial conditions taken from a forced stationary run, while the second ensemble was made up of random initial conditions. The fact that only the decaying experiment with the initial conditions picked from the first ensemble does not decay for times up to the largest eddy turn over time can be seen as direct support for the statement that *forced* small-scale turbulence is *universal*.

4.2 Anisotropic Turbulence

Although the phenomenological and experimental frameworks of fully-developed turbulent flows seem well founded, there are still many problems which lack clear experimental validation and/or a theoretical understanding. Most of them have to do with anisotropic fluctuations.

For example, the central issue of K41 phenomenology is the assumption of *return-to-isotropy* for smaller and smaller scales. Recently, some detailed analysis of small-scale isotropy on experimental and numerical shear turbulence have been performed [40,32,44,37]. The ideal experimental set-up to test such a problem is the case of a *homogeneous* shear flow. In this flow, the shear is spatially homogeneous and points in one direction, i.e. the large-scale

mean-velocity has the perfect linear profile: $\boldsymbol{V} = (V_0 y, 0, 0)$. The shear is given by $\mathcal{S}_{ij} = \partial_i V_j = \delta_{iy}\delta_j x V_0$. It is easy to see that we have a homogeneous but anisotropic flow, the ideal case to study the influence of anisotropies on small-scale statistics. Small scales must be compared to the characteristic shear length, $L_S = \epsilon^{1/3}/\mathcal{S}$; only for $r \ll L_\mathcal{S}$ may we expect that anisotropic fluctuations are secondary with respect to the isotropic ones. The case $r \gg L_\mathcal{S}$ is of some interest in all situations where the shear becomes very intense, e.g. very close to the walls in bounded flows. In the latter case, different physics than the K41 phenomenology must be expected [16]. Fortunately, it is not difficult to design ad-hoc experiments or DNS possessing an almost perfect linear profile such as homogeneous shear [40,39,44]. A popular way to measure small-scale anisotropies is to focus on the Reynolds number dependencies of isotropic and anisotropic observables built in terms of velocity gradients. For example, due to the symmetries of the mean flow, gradients of the stream-wise component in the shear direction, $\partial_y v_x$, may have a skewed distribution only due to the anisotropic fluctuations; they should have a perfectly symmetric PDF in a perfectly isotropic flow. A natural way to measure the residual anisotropy at small scales as a function of the Reynolds numbers is to build mixed generalized Skewness based on gradients:

$$M^{(2p+1)}(R_\lambda) = \frac{\langle (\partial_y v_x)^{2p+1} \rangle}{\langle (\partial_y v_x)^2 \rangle^{\frac{2p+1}{2}}}. \tag{31}$$

The above generalized Skewness should be exactly zero in an isotropic ensemble, because its numerator vanishes. Of course, for a finite Reynolds number one cannot expect that the large-scale anisotropy introduced by the shear has completely decayed on the gradient scale. Therefore, a direct measure of the rate of decay of (31) as a function of Reynolds number is a quantitative indication on the rate of decay of anisotropy at small scales, i.e. a direct check of the assumption of *local isotropy* made by Kolmogorov for Reynolds large enough. Lumley set up a dimensional argument for anisotropic fluctuations predicting, as a function of the Reynolds numbers based on the Taylor scale, R_λ:

$$M^{(2p+1)}(R_\lambda) \sim R_\lambda^{-\frac{1}{2}} \tag{32}$$

independently of the order of the moment, n. Surprisingly, both numerical [44] (at very low Reynolds numbers) and experimental tests (up to $R_\lambda \sim 1000$) showed a clear discrepancy from the dimensional prediction (32), see Fig. 4.

For example in [39] the authors quote an almost constant behavior as a function of Reynolds number for the fifth order, $M^{(5)}(R_\lambda) \sim O(1)$ and an *increasing* behavior for the seventh order $M^{(7)}(R_\lambda)$, although in the latter case some problems of statistical convergence may have contaminated the result. These results have cast severe doubts on the basis of the K41 and K62 theories.

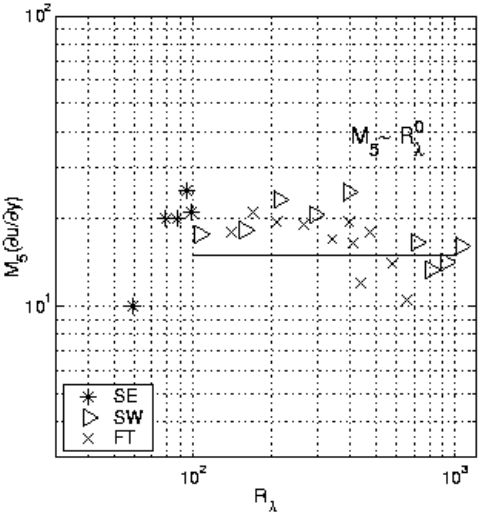

Fig. 4. A collection of experimental and numerical results for the fifth order skewness $M^{(5)}(R_\lambda)$ as a function of Reynolds numbers taken from Fig. 2 of J. Schumacher et al., *Phys. Fluids* **15**, 84 (2003). Notation in the legend stands for SE reference [41]; FT reference [43], SW reference [42]

In recent years, net advancements in the understanding of anisotropic turbulence have been obtained thanks to systematic decomposition of velocity correlation functions in the basis of the rotational operator, the SO(3) group [18–20,32,33]. These works defined a clear theoretical, experimental and numerical basis from which one may start to attack the *return-to-isotropy* discussed before. Moreover, the net improvement in the scaling quality measured after the SO(3) decomposition allowed for a better quantitative understanding of both isotropic and anisotropic turbulence. For scalar objects, such as the longitudinal structure functions, the SO(3) decomposition reduces to the projection on the spherical harmonics:

$$S^{(p)}(\boldsymbol{\ell}) = \sum_{j=0}^{\infty} \sum_{m=-j}^{j} s_{jm}^{(p)}(\ell) Y_{jm}(\hat{\boldsymbol{\ell}}). \tag{33}$$

As customary, the indices (j, m) label the total angular momentum and its projection on a reference axis, respectively. Our interest here is concentrated on the behavior of the projection on the isotropic sector, $s_{j=0m=0}^{(p)}(\ell)$, or any of the anisotropic sectors, $s_{jm}^{(p)}(\ell)$ with $j > 0$. Theoretical and phenomenological reasonings suggest that each projections has its own scaling behavior, $s_{jm}^{(p)}(\ell) \sim \ell^{\zeta^j(p)}$, with the exponent depending on the j sector only. Now, the *recovery of isotropy* can be restated in terms of the values of the scaling exponents. The *return-to-isotropy* hypothesis made by Kolmogorov implies that

the scaling exponents are organized in a hierarchical way:

$$\zeta^{j=0}(p) < \zeta^{j}(p), \tag{34}$$

for any order p. All numerical [33,37], experimental [32,38] and phenomenological [34] works on this subject confirm the existence of the hierarchy (34). Any turbulent flow should become more and more isotropic by going to smaller and smaller scales. The persistence of small-scale anisotropy, as measured by the generalized Skewness defined in (31), can simply be explained by noticing that it is a balance of isotropic and anisotropic effects of two different order of velocity correlation functions, the $(2p + 1)$th order in the numerator and the 2nd order in the denominator. It is easy to see that, due to the presence of intermittency in the anisotropic sectors as well, one may have a recovery of isotropy in the sense that the hierarchy (34) holds. One may also still observe the persistence of anisotropies based on the gradient statistics, as shown in Fig. 4 [37].

This paper is made of three different contributions. G.B. wrote Sect. 2, B.C. Sect. 3 and L.B. Sect. 4.

Acknowledgments

It is a pleasure for us to thank the numerous friends and collaborators which have shared with us their knowledge and physical insight on turbulence. In particular, we cite R. Benzi, A. Celani, M. Cencini, L. Chevillard, U. Frisch, G. Lacorata, E. Leveque, A. Lanotte, S. Musacchio, F. Toschi, J.F. Pinton, I. Procaccia, I.M. Sokolov, M. Vergassola and A. Vulpiani.

References

1. U. Frisch, *Turbulence – The legacy of A.N. Kolmogorov*, Cambridge. Univ. Press (1995)
2. A.N. Kolmogorov, A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number, J. Fluid. Mech. **13** 82–85 (1962)
3. A.S. Monin, A.M. Yaglom, *Statistical fluid dynamics: Mechanics of turbulence* MIT Press, (1987)
4. A.N. Kolmogorov, The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers, Dokl. Akad. Nauk SSSR **30**, 301 (1941)
5. A.N. Kolmogorov, On degeneration of isotropic turbulence in an incompressible viscous liquid, Dokl. Akad. Nauk SSSR **31**, 538 (1941)
6. G.I. Taylor, Statistical theory of turbulence, Proc. Roy. Soc. A **151**, 421 (1935)
7. K.R. Sreenivasan, An update on the energy dissipation rate in isotropic turbulence, Phys. Fluids **10**, 528 (1998)
8. Y. Kaneda, T. Ishihara, M. Yokokawa, K. Itakura and A. Uno, Energy dissipation rate and energy spectrum in high resolution direct numerical simulations of turbulence in a periodic box, Phys. Fluids **15**, L21 (2003)

9. L.F. Richardson, *Weather Prediction by Numerical Process*, Cambridge University Press, 1922
10. H.L. Grant, R.W. Stewart, and A. Moilliet, Turbulent spectra from a tidal channel, J. Fluid Mech. **12**, 241 (1962)
11. G. Boffetta and G.P. Romano, Structure functions and energy dissipation dependence on Reynolds number, Phys. Fluids **14**, 3453 (2002)
12. L. Biferale, G. Boffetta, A. Celani, A. Lanotte, F. Toschi, and M. Vergassola, The decay of anisotropic turbulence, *Phys. Fluids* **15**, 2105 (2003)
13. R. Benzi, G. Paladin, G. Parisi, and A. Vulpiani, On the multifractal nature of fully developed turbulence and chaotic systems, J. Phys. A **17** 3521 (1984)
14. B. Castaing, Y. Gagne, On the inertial range in developed turbulence, in Turbulence in spatially extended systems (R. Benzi, S. Ciliberto, C. Basdevant Eds, Les Houches, 1992)
15. R. Benzi, L. Biferale, G. Paladin, A. Vulpiani and M. Vergassola, Multifractality in the statistics of the velocity gradients in turbulence, Phys. Rev. Lett. **67**, 2299 (1991)
16. P. Gualtieri, C.M. Casciola, R. Benzi, G. Amati, and R. Piva, *Phys. Fluids* **14**, 583 (2002)
17. F. Anselmet, Y. Gagne, E.J. Hopfinger, and R.A. Antonia, High order velocity structure functions in turbulent shear flow, J. Fluid Mech. **140** 63 (1984)
18. I. Arad, V.S. L'vov and I. Procaccia, Correlation functions in isotropic and anisotropic turbulence: the role of the symmetry group, Phys. Rev. E **59**, 6753 (1999)
19. I. Arad, B. Dhruva, S. Kurien, V.S. L'vov, I. Procaccia, and K.R. Sreenivasan, Extraction of anisotropic contributions in turbulent flows, Phys. Rev. Lett. **81**, 5330 (1998)
20. I. Arad, L. Biferale, I. Mazzitelli, and I. Procaccia, Disentangling scaling properties in anisotropic and inhomogeneous turbulence, Phys. Rev. Lett. **82**, 5040 (1999)
21. J.L. Lumley, Similarity and the turbulent energy spectrum, Phys. Fluids **10** 855 (1967)
22. E.A. Novikov and R.W. Stewart, "The intermittency of turbulence and the spectrum of energy dissipation" Izv. Akad. Nauk SSSR Geofiz. III, 408 (1964)
23. R.H. Kraichnann, On Kolmogorov's inertial-range theories' J. Fluid Mech. **62**, 305 (1974)
24. V.N. Desnyansky and E.A. Novikov, The evolution of turbulence spectra to the similarity regime, Izv. Akad. Nauk SSSR Fiz. Atmos. Okeana **10**, 127 (1974)
25. T, Bohr, M.H. Jensen, G. Paladin, and A. Vulpiani, *Dynamical systems approach to turbulence* (Cambridge University Press, Cambridge, UK)
26. L. Biferale, Shell Models of Energy Cascade in Turbulence, Annu. Rev. Fluid Mech. **35**, 441 (2003)
27. G. Eyink, Lagrangian field-theory, multifractals and universal scaling in turbulence, Phys. Lett. A **172**, 355 (1993)
28. V.S. L'vov and I. Procaccia, Fusion Rules in Turbulent Systems with Flux Equilibrium Phys. Rev. Lett **76**, 2898 (1996)
29. R. Benzi, L. Biferale, and F. Toschi, Multiscale correlation functions in turbulence. Phys. Rev. Lett. **80**, 3244 (1998)
30. A. Arneodo, C. Baudet, F. Belin et al., Structure functions in turbulence, in various flow configurations at Reynolds number between 30 and 5000, using extended self-similarity, EuroPhys. Lett. **34**, 411–416 (1996)

31. R. Benzi, S. Ciliberto, R. Tripiccione, C. Baudet, F. Massaioli, and S. Succi, Extended self-similarity in turbulent flows, Phys. Rev. E **48**, R29 (1993)
32. S. Kurien and K.R. Sreenivasan, Anisotropic scaling contributions to high-order structure functions in high-Reynolds-number turbulence, Phys. Rev. E **62**, 2206 (2000)
33. L. Biferale and F. Toschi, Anisotropic homogeneous turbulence: hierarchy and intermittency of scaling exponents in the anisotropic sectors, Phys. Rev. Lett. **86** 4831 (2001)
34. L. Biferale, I. Daumont, A. Lanotte, and F. Toschi, Anomalous and dimensional scaling in anisotropic turbulence, Phys. Rev. E. **66**, 056306 (2002)
35. R.H. Kraichnan, Anomalous scaling of a randomly advected passive scalar, Phys. Rev. Lett. **72**, 1016 (1994)
36. G. Falkovich, K. Gawędzki, and M. Vergassola, Particles and fields in fluid turbulence, Rev. Mod. Phys. **73**, 913 (2001)
37. L. Biferale and M. Vergassola, Isotropy *vs* anisotropy in small-scale turbulence, Phys. Fluids **13**, 2139 (2001)
38. X. Shen and Z. Warhaft, Longitudinal and transvers structure functions in sheared and unsheared wind-tunnel turbulence, Phys. Fluids **14**, 370 (2002)
39. X. Shen and Z. Warhaft, The Anisotropy of the small scale structure in high Reynolds number ($R_\lambda \sim 1000$) turbulent shear flow, Phys. Fluids **14**, 2432 (2002)
40. S. Garg and Z. Warhaft, Phys. Fluids **10**, 662 (1998)
41. J. Schumacher and B. Eckhardt, On statistically stationary homogeneous shear turbulence, Europhys. Lett. **52**, 627 (2000)
42. X. Shen and Z. Warhaft, The anisotropy of the small scale structure in high Reynolds number turbulent shear flow, Phys. Fluids **12**, 2976 (2000)
43. M. Ferchichi and S. Tavoularis, Reynolds number effects on the fine structure of uniformly sheared turbulence, Phys. Fluids **12**, 2942 (2000)
44. A. Pumir, Turbulence in homogeneous shear flows, Phys. Fluids **8**, 3112 (1996)
45. O. Chanal, B. Chabaud, B. Castaing, and B. Hébral, Eur. Phys. J. **B17**, 309 (2000)
46. J. Delour, J.F. Muzy, and A. Arneodo, Eur. Phys. J. B **23**, 243 (2001); J. Delour, PHD Thesis Université de Bordeaux, (2001) unpublished
47. B. Castaing, Y. Gagne, and E.J. Hopfinger, Velocity probability density functions of high Reynolds number turbulence, Physica D **46**, 177–200, (1990)
48. K.R. Sreenivasan and R.A. Antonia, The phenomenology of small-scale turbulence, Annu. Rev. Fluid Mech. **29**, 435–472 (1997)
49. B. Castaing, The temperature of turbulent flows, J. Phys. II **6**, 105–114 (1996)
50. G. Parisi and U. Frish, On the singularity structure of fully developed turbulence, in Turbulence and Predictability in Geophysical Fluid Dynamics, Proceed. Intern. School of Physics 'Enrico Fermi', 1983, Varenna, Italy, 84–87 (eds. M. Ghil, R. Benzi, G. Parisi, North-Holland, Amsterdam, 1985)
51. G. Stolovisky and K.R. Sreenivasan, Kolmogorov refined similarity hypothesis for turbulence and general stochastic processes, Rev. Mod. Phys. **66** 229–240 (1994)
52. Y. Gagne, M. Marchand, and B. Castaing, Conditional velocity pdf in 3D Turbulence, J. Phys. II **4**. 1–8 (1994)
53. Y. Gagne, Private communication
54. E.A. Novikov, Intermittency and scale similarity of the structure of turbulent flow, Prikl. Math. Mekh. **35**, 266–277 (1970)

55. Z.S. She and E. Leveque, Universal scaling laws in fully developed turbulence, Phys. Rev. Lett. A **72**, 336–339 (1994)
56. B. Dubrulle, Intermittency in fully developed turbulence: log-Poisson statistics and generalised covariance, Phys. Rev. Lett. A **73**, 959–962 (1994)
57. Y. Malecot, C. Auriault, H. Kahalerras, Y. Gagne, O. Chanal, B. Chabaud, and B. Castaing, A statistical estimator of turbulence intermittency in physical and numerical experiments, Eur. Phys. J. B **16**, 549 (2000)
58. E. Leveque and L. Chevillard, Energy cascade acceleration close to the dissipative scale in turbulence, to be published
59. R. Benzi, L. Biferale, S. Ciliberto, M.V. Struglia, and R. Tripiccione, Generalized scaling in fully developed turbulence; Physica **96**, 162 (1996)
60. R. Freidrich, J. Peinke, Description of a turbulent cascade by a Fokker-Planck equation, Phys. Rev. Lett. **78**, 863–866 (1997)
61. P. Marcq and A. Naert, A Langevin equation for turbulent velocity increments, Phys. Fluids, **13**, 2590–2595 (2001)
62. U. Frisch and M. Vergassola, A prediction of the multifractal model: the intermediate dissipation range, EuroPhys. Lett. **14**, 439–444 (1991)

Turbulence and Stochastic Processes

Antonio Celani[1], Andrea Mazzino[2], and Alain Pumir[3]

[1] CNRS, INLN, 1361 Route des Lucioles, 06560 Valbonne, France, celani@inln.cnrs.fr
[2] ISAC-CNR, Lecce Section, Lecce, Italy, 73100, and Department of Physics, Genova University, Genova, Italy, 16146, mazzino@fisica.unige.it
[3] CNRS, INLN, 1361 Route des Lucioles, 06560 Valbonne, France, Alain.Pumir@inln.cnrs.fr

Abstract. In 1931 the monograph *Analytical Methods in Probability Theory* appeared, in which A.N. Kolmogorov laid the foundations for the modern theory of Markov processes [1]. According to Gnedenko: "In the history of probability theory it is difficult to find other works that changed the established points of view and basic trends in research work in such a decisive way". Ten years later, his article on fully developed turbulence provided the framework within which most, if not all, of the subsequent theoretical investigations have been conducted [2] (see e.g. the review by Biferale et al. in this volume [3]). Remarkably, the greatest advances made in the last few years towards a thorough understanding of turbulence developed from the successful marriage between the theory of stochastic processes and the phenomenology of turbulent transport of scalar fields. In this article we will summarize these recent developments which expose the direct link between the intermittency of transported fields and the statistical properties of particle trajectories advected by the turbulent flow (see also [4], and, for a more thorough review, [5]). We also discuss the perspectives of the Lagrangian approach beyond passive scalars, especially for the modeling of hydrodynamic turbulence.

1 Passive Scalar Turbulence

In his 1941 paper, Kolmogorov postulated that in a turbulent flow at very large Reynolds number, governed by the Navier-Stokes equation

$$\partial_t \boldsymbol{v} + \boldsymbol{v} \cdot \boldsymbol{\nabla} \boldsymbol{v} = -\boldsymbol{\nabla} p + \nu \nabla^2 \boldsymbol{v} \,, \tag{1}$$

the statistics of velocity differences $\delta_r v = [\boldsymbol{v}(\boldsymbol{x}+\boldsymbol{r},t) - \boldsymbol{v}(\boldsymbol{x},t)] \cdot \hat{\boldsymbol{r}}$ should depend only on the size of the spatial separation between the two measurement points r and on the average energy input ϵ_v [1]. This assumption should be valid across a distance $\boldsymbol{r}$ neither too large, in order to safely disregard the

[1] The dependence on the position $\boldsymbol{x}$ and on the orientation of $\boldsymbol{r}$ can be omitted by resorting to the hypothesis of statistical homogeneity and isotropy, i.e. to the statistical invariance under translations and rotations, which is believed to be correct over sufficiently small regions of space, and far enough from the boundaries.

A. Celani, A. Mazzino, and A. Pumir, Turbulence and Stochastic Processes, Lect. Notes Phys. **636**, 173–186 (2003)
http://www.springerlink.com/

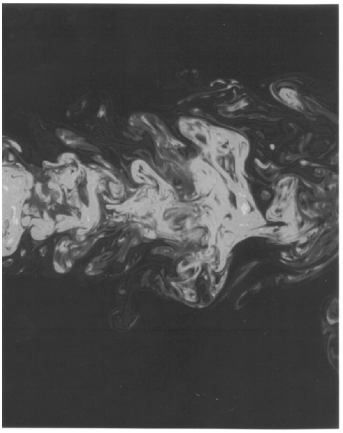

Fig. 1. The concentration of fluoresceine, a passive scalar, transported by a turbulent jet flow (from Shraiman and Siggia, Nature **405** (2000) 639)

effect of the boundaries, nor too small, so that the effect of viscous dissipation can be neglected.

It is important to point out that the Kolmogorov assumption is a statement about the universality of small-scale turbulence. Indeed the only quantity that matters is the average energy supplied to the system per unit time, ϵ_v, irrespective of the detailed mechanism of injection and dissipation. Dimensional arguments[2] then yield the celebrated Kolmogorov scaling law (see, e.g., [6])

$$\delta_r v \sim \epsilon_v^{1/3} r^{1/3} \ . \tag{2}$$

The same theoretical framework can be applied to the statistics of a passive scalar – for example, a dilute nonreacting tracer (see Fig. 1), or temperature under appropriate experimental conditions – governed by the advection-diffusion equation

$$\partial_t \theta + \boldsymbol{v} \cdot \boldsymbol{\nabla} \theta = \kappa \nabla^2 \theta + f \ , \tag{3}$$

where f is an external source of scalar fluctuations. For scalar differences $\delta_r \theta = [\theta(\boldsymbol{x} + \boldsymbol{r}, t) - \theta(\boldsymbol{x}, t)]$ one obtains the law[3]

$$\delta_r \theta \sim \epsilon_v^{-1/6} \epsilon_\theta^{1/2} r^{1/3} \tag{4}$$

independently proposed by Obukhov (1949) and Corrsin (1951), [7], where ϵ_θ is the rate of injection of squared scalar fluctuations.

[2] It should be noted that Kolmogorov derived from the Navier-Stokes equations the exact relation $\langle (\delta_r v)^3 \rangle = -\frac{4}{5} \epsilon_v r$. Notice that an exact relation can be drived only for the third-order moment of velocity differences.

[3] Also in this case an exact relation has been obtained by Yaglom: $\langle \delta_r v (\delta_r \theta)^2 \rangle = -\frac{4}{3} \epsilon_\theta r$.

At scales much smaller than the viscous lengthscale $\eta = (\nu^3/\epsilon_v)^{1/4}$ (also called the Kolmogorov scale) the velocity field is smooth, and the scalar field obeys the Batchelor law $\langle \theta(\boldsymbol{x})\theta(\boldsymbol{x}+\boldsymbol{r})\rangle \sim \epsilon_\theta \sigma^{-1} \log(\eta/r)$, where σ is the typical velocity gradient at scale η [8]. Such behavior can be derived by exploiting ideas borrowed from dynamical systems theory, without invoking phenomenological assumptions. The key ingredient for the Batchelor law to emerge is the presence of Lagrangian chaos, that is the exponential separation of initially nearby fluid particle trajectories. For details on Lagrangian chaos readers can refer to [9].

During the 50's evidence started to accumulate that small yet significant deviations from Kolmogorov scaling (2) were present. Experimental results were consistent with a power law scaling of velocity differences, although with an exponent depending on the order of the moment considered: $\langle (\delta_r v)^n \rangle \sim r_n^\sigma$ where σ_n is not linear with n, but rather a strictly concave function of the order. This phenomenon goes under the name of anomalous scaling, or intermittency[4].

The failure of Kolmogorov's theory poses an intriguing problem. Anomalous scaling requires the introduction of at least one external lengthscale in order to match the physical dimensions. It could be the large scale L where energy is injected, or the small scale η where the energy is dissipated, or both. In any event, can we still expect to find some universality, or is every observable affected by the details of energy supply and removal ? In the latter case, fully developed turbulence would exit the realm of theoretical physics to enter the domain of engineering: one could not speak about "turbulence" anymore, but rather about a collection of turbulent flows, to be studied each *per se*. Although the answer to the above question is not yet known for hydrodynamic turbulence, the situation appears to be under control in the context of passive scalar turbulence: universality persists, yet in a narrower sense than the one proposed by Kolmogorov. We will now recall some of the steps that led to that conclusion.

In the 70's, accurate experimental measurements of temperature fluctuations showed that the Obukhov-Corrsin scaling (4) does not hold true, i.e. the scalar field is intermittent as well. One observes indeed the scaling $\langle (\delta_r \theta)^n \rangle \sim r^{\zeta_n}$, where again the graph of ζ_n versus n is concave. Surprisingly enough, the deviations in the passive scalar case appear to be more pronounced than those observed for velocity (see Fig. 2), culminating in the saturation of scaling exponents with the order [10]. Those substantial deviations at large orders are due to the existence of intense structures in the scalar field: the "cliffs". On the two sides of a cliff the scalar can experience variations as large as ten times the typical fluctuation across a very small

[4] The name intermittency originates from the fact that going to smaller and smaller scales the probability of weak excursions increases and, simultaneously, very intense events become more likely. This reflects in the "bursty" behavior observed in experimental time series of velocity differences.

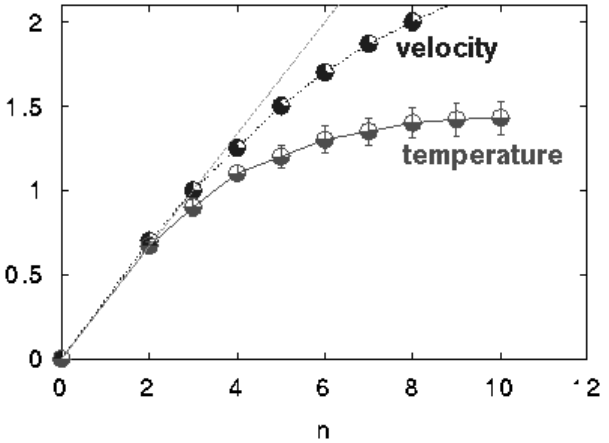

Fig. 2. Scaling exponents of velocity and temperature obtained from experiments in low temperature helium gas (from H. Willaime et al., Eur. Phys. J. B **18**, 363 (2000); F. Moisy et al., Phys. Rev. Lett. **86** (2001) 4827)

distance (see [11] and references therein). A crucial observation is that the presence of cliffs appears to be independent of the experimental flow, and that those structures are present even in synthetic velocity fields generated in computer simulations of passive scalar transport [12,13] (see Fig. 3). This suggests the possibility that passive scalar intermittency can be present and studied even in the context of simplified models of advection, among which the most important has been the Kraichnan model.

2 The Kraichnan Model and Beyond

One of the simplest models of passive transport by a turbulent flow was introduced by Kraichnan in 1968 [14]. The advecting velocity is prescribed: at any given time it is an incompresssible Gaussian vector field whose increments scale as $\delta_r v \sim r^{\xi/2}$: no intermittency for the velocity statistics. This is a reasonable representation of an instantaneous snapshot of a turbulent flow (for $\xi = 4/3$, and neglecting the deviations from Kolmogorov scaling). However, the Kraichnan velocity field changes from time to time without keeping any memory of its past states, a highly unphysical assumption. The asset of such an unrealistic choice is that it allows a mathematical treatment that would be otherwise impossible for realistic, finite-time correlated flows. As an example, it is easy to derive that the scaling exponent of the second-order moment of scalar differences is $\zeta_2 = 2 - \xi$, i.e. the value expected by dimensional arguments.

In 1994, Kraichnan, under a supplementary special assumption on the scalar statistics, calculated all the scaling exponents ζ_n. Those of order larger than 2 turned out to be anomalous, despite the simple statistics of the

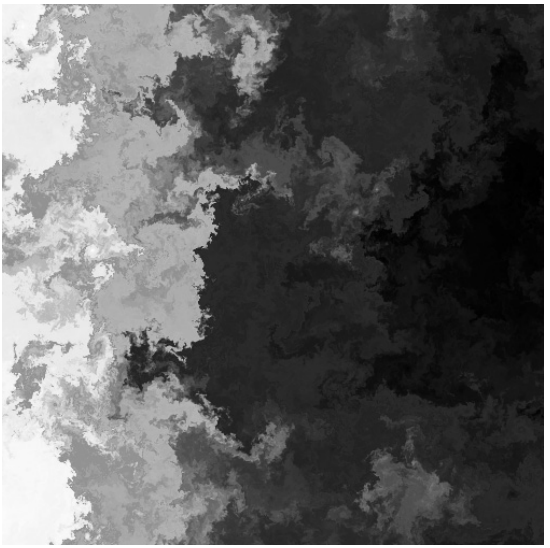

Fig. 3. A snapshot of a passive scalar field from computer simulations of two-dimensional turbulences (from A. Celani et al., Phys. Rev. Lett. **84** (2000) 2385)

advecting flow. Just one year later, three independent groups arrived at the same qualitative conclusion. However, they used perturbative methods that do not require additional hypothesis, albeit being viable only in some restricted region of parameters (e.g. for $\xi \ll 1$ or $2 - \xi \ll 1$) [16–18]. Yet, the latter values computed from first principles were different from those predicted by Kraichnan. The controversy was then settled in 1998 by the computer simulations by Frisch, Mazzino and Vergassola, who confirmed the presence of intermittency in the range of parameters $0 < \xi < 2$ [19]. The data were in agreement with perturbation theory (See Fig. 4). For our present purposes, a relevant aspect of [19] is the use of a particle-based (i.e. Lagrangian) method to obtain information on the (Eulerian) statistics of the scalar field. The basic observation is that the passive scalar equation (3) can be solved in terms of Lagrangian trajectories. Particles are transported by the flow and are subject to molecular diffusion, according to the stochastic differential equation[5]

$$d\boldsymbol{X} = \boldsymbol{v}(\boldsymbol{X}, t)dt + \sqrt{2\kappa}\, d\boldsymbol{W}(t)\ , \tag{5}$$

[5] It is worth recalling that, for $\kappa = 0$, (5) becomes a dynamical system which is conservative for incompressible flows. In two dimensions, the introduction of the stream-function ψ ($v_1 = \partial\psi/\partial x_2$, $v_2 = -\partial\psi/\partial x_1$), reduces (5) to a Hamiltonian system with ψ playing the role of Hamiltonian. Important consequences of such identifications (e.g., the Hamiltonian chaos and the celebrated Kolmogorov-Arnold-Moser (KAM) theorem) can be found in the reviews by Livi et al. [20] and by Celletti et al. [21] in this volume).

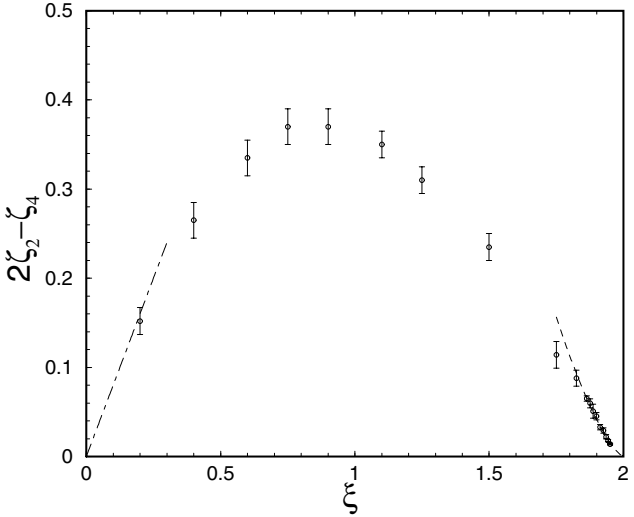

Fig. 4. The deviation from dimensional scaling for the fourth-order moment of scalar differences, measured in the Kraichnan model. The *dash-dotted* and *dashed lines* are the predictions of perturbation theory in ξ and $2-\xi$, respectively (from U. Frisch, A. Mazzino, and M. Vergassola, Phys. Rev. Lett. **80** (1998) 5532)

where $\boldsymbol{W}(t)$ is the multi-dimensional Wiener process. The link between Eulerian and Lagrangian observables is given by the relation

$$\theta(\boldsymbol{x},t) = \int d\boldsymbol{y} \int_{-\infty}^{t} ds\, f(\boldsymbol{y},s) p(\boldsymbol{y},s|\boldsymbol{x},t)\,, \tag{6}$$

where $p(\boldsymbol{y},s|\boldsymbol{x},t)$ is the probability to observe a particle at $\boldsymbol{y}$ at time s, given that it is in $\boldsymbol{x}$ at time t. The propagator p evolves according to the Kolmogorov equations

$$-\partial_s p(\boldsymbol{y},s|\boldsymbol{x},t) - \boldsymbol{\nabla}_y \cdot [\boldsymbol{v}(\boldsymbol{y},s) p(\boldsymbol{y},s|\boldsymbol{x},t)] = \kappa\, \nabla_y^2 p(\boldsymbol{y},s|\boldsymbol{x},t)\,, \tag{7}$$

$$\partial_t p(\boldsymbol{y},s|\boldsymbol{x},t) + \boldsymbol{\nabla}_x \cdot [\boldsymbol{v}(\boldsymbol{x},t) p(\boldsymbol{y},s|\boldsymbol{x},t)] = \kappa\, \nabla_x^2 p(\boldsymbol{y},s|\boldsymbol{x},t)\,. \tag{8}$$

It is therefore clear that there is a perfect duality between the description of passive scalar transport in terms of fields and in terms of particles. Relevant observables as the n-point correlation functions[6] $\langle \theta(\boldsymbol{x}_1,t)\cdots\theta(\boldsymbol{x}_n,t)\rangle$ are then simply expressed in terms of averages over the ensemble of trajectories of N particles simultaneously transported by the flow.

Since every concept based on Eulerian observables must have its Lagrangian counterpart, one is naturally led to ask: what is the interpretation of

[6] Note that the n-th moment of scalar differences $\langle(\delta_r\theta)^n\rangle$ is just a linear combination of n-point correlation functions taken in $n+1$ special configurations.

intermittency in the language of particle trajectories ? The answer to this question is once more most easily given in the framework of the Kraichnan model. Indeed, the time-decorrelation of the velocity field induces a great simplification in the description of the N-particle statistics. One has that the statistical evolution of trajectories described by the velocity averaged probability

$$P(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n; s|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n; t) = \langle p(\boldsymbol{y}_1, s|\boldsymbol{x}_1, t) \cdots p(\boldsymbol{y}_n, s|\boldsymbol{x}_n, t) \rangle_v \quad (9)$$

is a Markov process in the space of particle configurations. As a consequence, P itself obeys a Kolmogorov equation

$$\partial_t P = \mathcal{M} P, \quad (10)$$

where $\mathcal{M}$ is a second order partial differential operator that describes the diffusion of particles in the space of configurations $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$.

In 1998, Bernard, Gawędzki and Kupiainen [22] showed that, within the Kraichnan model, the appearance of anomalously scaling terms in the n-point correlation function of the scalar field is intimately related to the existence of peculiar homogeneous functions Z of the n-particle configuration[7],

$$Z(\lambda \boldsymbol{x}_1, \dots, \lambda \boldsymbol{x}_n) = \lambda^{\zeta_n} Z(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n), \quad (11)$$

where ζ_n is the Eulerian anomalous exponent. These functions are *preserved on average* by the Lagrangian evolution

$$\int P(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n; s|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n; t) Z(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)\, d\boldsymbol{y}_1 \cdots d\boldsymbol{y}_n = Z(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)\,, \quad (12)$$

although dimensional arguments would predict a growth in time for the l.h.s of (12) as $|t-s|^{\zeta_n/(2-\xi)}$.

The Lagrangian interpretation of intermittency in terms of statistically invariant functions lends itself to a straightforward generalization, since (12) is well defined for any statistical ensemble of fluid velocities, including realistic ones. In the latter case, instead of proving that the functions Z are responsible for anomalous scaling, one can proceed in the reverse direction: first extract from numerical experiments the anomalous part of the $n-$point correlation, and then check if this function is statistically preserved on average. A major problem of this procedure is that one has to sample the correlation function over a high-dimensional space: for example, to specify the geometry of four points in three dimensions, one needs five "angular" degrees of freedom plus the overall size of the configuration[8]. However, in two dimensions a triangle

[7] These are called also "zero-modes" since they satisfy the equation $\mathcal{M}Z = 0$ in the limit of vanishing molecular diffusivity $\kappa \to 0$.

[8] The symmetries under translations and rotations reduce the degrees of freedom from the initial number of twelve to the final six. The size of the configuration R is usually defined by $R^2 = \frac{1}{N(N-1)} \sum_{i,j} |\boldsymbol{r}_i - \boldsymbol{r}_j|^2$.

is identified by only two angles, and this makes the problem tractable: Celani and Vergassola [23] have indeed shown that the anomalous three-point correlation function of a passive scalar advected by two-dimensional Navier-Stokes turbulence is a statistically preserved function.

Another important connection between intermittency and stochastic processes was pointed out by Gat and Zeitak in [24]. They considered the evolution of the shape of a configuration, of n particles transported by the Kraichnan flow: replacing time by the ratio of the sizes R_2/R_1 of the configurations they were able to conclude that this was a (discontinuous) Markov process in the space of shapes. As the size-ratio increases, the probability density function over the angular degrees of freedom approaches a stationary distribution[9] with a decay rate $(R_2/R_1)^{-\zeta_n}$, where the dependence on the shape of the configuration is specified by the corresponding function Z. The asymptotic distribution in the space of shapes has been studied in realistic flows by Pumir et al. [25]. However, the direct measurement of Z from particle evolution seems to be a much more difficult task [26].

A direct consequence of the identification of Z functions with the anomalous part of the correlation is that the scaling exponents are independent of the choice of the forcing, since the latter does not appear in the definition (12). Note, however, that the ζ_n still depend on the statistics of the velocity field. The numerical prefactors that appear in front of the moments of scalar differences are nonuniversal and change according to the details of f. This is a narrower universality than the one prospected by Kolmogorov, as anticipated earlier: it is an open question whether this picture applies to Navier-Stokes turbulence as well [27].

3 Towards Navier-Stokes Turbulence

The study of the passive scalar problem has emphasized the importance of considering the full multipoint correlation function. In addition, the use of Lagrangian concepts has been crucial in identifying the key features responsible for intermittency. The purpose of the present section is to present an extension of the ideas discussed above, to address the problem of hydrodynamic turbulence itself.

Decades of research have shown the importance of the dynamics of the velocity gradient tensor, $m_{ab} \equiv \partial_a v_b$, to understand many aspects of turbulent flows [28]. Traditionally, the tensor m is decomposed as a sum of a symmetric part, $s_{ab} = \frac{1}{2}(m_{ab} + m_{ba})$, the rate of strain tensor, and of an antisymmetric part, $\Omega_{ab} \equiv \frac{1}{2}(m_{ab} - m_{ba}) = \epsilon_{abc}\omega_c$, where $\omega = \nabla \wedge v$ is the vorticity, and ϵ_{abc} is the completely antisymmetric tensor. In the absence of viscosity, vorticity is stretched as a material line by the flow. This effect,

[9] The stationary distribution of shapes F can be obtained by seeking a self-similar solution to the equation $\partial_t F = \mathcal{M}F$, i.e. a function with the property $F(t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = F(1, \boldsymbol{x}_1/t^{1/(2-\xi)}, \ldots, \boldsymbol{x}_n/t^{1/(2-\xi)})$.

known as vortex stretching, is responsible for the production of small-scale fluctuations in turbulent flows. The equation of evolution for the tensor m, deduced from the Navier-Stokes equations (1), reads:

$$\frac{dm_{ab}}{dt} \equiv \partial_t m_{ab} + (v \cdot \nabla) m_{ab} = -m_{ab}^2 - \partial_{ab} p + \nu \nabla^2 m_{ab}. \tag{13}$$

The evolution equation for m is thus nonlinear. It involves a nonlocal contribution due to the pressure Hessian term ($\partial_{ab} p$), which ensures that the flow is incompressible ($tr(m) = 0$).

Turbulent flows notoriously involve many length scales. The goal here is to develop a theory, based on the essential ingredients of Navier-Stokes turbulence, that describes the statistical properties of the velocity field as a function of scale. We note in passing that the structure functions, defined as the moments of the velocity differences between to fixed spatial points, are inappropriate for our purpose. The first step in our construction consists in extending the notion of the velocity derivative tensor, corresponding formally to the scale $r = 0$, to a 'size' $r \neq 0$. We consider a homogeneous situation, so the location of the center of mass of the tetrahedron $\boldsymbol{\rho}_0 = (\sum_i \boldsymbol{r}_i)/4$, as well as its velocity $\boldsymbol{u}_0 = (\sum_i \boldsymbol{v}_i)/4$ are not important for our purpose.

To this end, we consider a set of four points, a tetrahedron, separated by a scale r. We define the "finite difference" tensor, M, based on the knowledge of the velocities, $(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4)$ at the four points $(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4)$. In practice, the matrix M is constructed by defining:

$$\begin{aligned}
\boldsymbol{\rho}_1 &\equiv (\boldsymbol{r}_2 - \boldsymbol{r}_1)/\sqrt{2}, \\
\boldsymbol{\rho}_2 &\equiv (2\boldsymbol{r}_3 - \boldsymbol{r}_2 - \boldsymbol{r}_1)/\sqrt{6}, \\
\boldsymbol{\rho}_3 &\equiv (3\boldsymbol{r}_4 - \boldsymbol{r}_3 - \boldsymbol{r}_2 - \boldsymbol{r}_1)/\sqrt{12}
\end{aligned} \tag{14}$$

$$\begin{aligned}
\boldsymbol{u}_1 &\equiv (\boldsymbol{v}_2 - \boldsymbol{v}_1)/\sqrt{2}, \\
\boldsymbol{u}_2 &\equiv (2\boldsymbol{v}_3 - \boldsymbol{v}_2 - \boldsymbol{v}_1)/\sqrt{6}, \\
\boldsymbol{u}_3 &\equiv (3\boldsymbol{v}_4 - \boldsymbol{v}_3 - \boldsymbol{v}_2 - \boldsymbol{v}_1)/\sqrt{12},
\end{aligned} \tag{15}$$

and:

$$M \equiv (\rho^{-1})_i^a u_i^b \tag{16}$$

where ρ^{-1} is the inverse of the tensor ρ, and a, b refer to spatial directions. The recent development of flow diagnostic methods, such as the holographic particle image velocimetry (PIV), now permits direct measurements of the matrix M, and the study of its statistical properties in real laboratory flows [29].

In order to theoretically investigate the statistical properties of the matrix M as a function of scale r, we introduce a phenomenological approach. Namely, we construct a stochastic model, with the proper non linear dynamics of strain and vorticity coarse grained over the lagrangian tetrahedron. More precisely, we decompose the velocity of the set of particles $\boldsymbol{r}_i$ into a coherent

component, $\rho \cdot M$, and a fluctuation residual, ζ. This leads to the following evolution equation for the variable $\boldsymbol{\rho}_i$:

$$\frac{d\rho_i^a}{dt} = \rho_i^b \cdot M_{ba} + \zeta_i^a. \tag{17}$$

The rapidly varying fluctuations are modeled by a Gaussian random term, white in time:

$$\langle \zeta_i^a(t)\zeta_j^b(0)\rangle = C_u(\delta_{ij}\rho^2\delta_{ab} - \rho_i^a\rho_j^b)\sqrt{tr(MM^T)}\ \delta(t). \tag{18}$$

The evolution equation for the tensor M results in principle from the evolution equation for the velocity, written in the Lagrangian frame:

$$\frac{dV_i}{dt} = -\nabla p_i. \tag{19}$$

where p_i is the pressure at point r_i. Taking finite differences of (19), one obtains formally an equation of the form:

$$\frac{dM}{dt} + M^2 = \text{pressure Hessian}. \tag{20}$$

The key question is then how does the pressure Hessian correlate with the velocity difference tensor, M. Here, we propose the following stochastic dynamics:

$$\frac{dM}{dt} + \alpha(M^2 - \kappa_{ab}tr(M^2)) = \xi_{ab}. \tag{21}$$

The factor α "renormalize" the nonlinear dynamics, as a result of the pressure term. This has been justified by studying the correlation between the pressure Hessian and the velocity gradient, m, or the finite difference tensor M [30,31]. A tetrad anisotropy tensor, $\kappa_{ab} \equiv [\sum_i (\rho^{-1})_i^a (\rho^{-1})_i^b]/[\sum_{j,c} (\rho^{-1})_j^c (\rho^{-1})_j^c]$ was introduced into the pressure hessian to make sure that the local pressure drops out of the energy balance equation. The rest of the pressure Hessian term, attributed to the nonlocal contribution, is modeled by a random Gaussian term white in time:

$$\langle \xi^a(t)\xi^b(0)\rangle = C_M \varepsilon \rho^{-2}\ \delta(t) \tag{22}$$

obeying the Kolmogorov scaling. Here $\rho^{-2} = 1/tr(\rho \cdot \rho^T)$, and ε is the energy dissipation term.

Our model is thus formulated as a set of stochastic differential equations. The corresponding Fokker-Planck equation for the probability distribution function $P(M, \rho, t)$ is

$$\partial_t P = \mathcal{L}P, \tag{23}$$

where the operator $\mathcal{L}$ is a second-order, diffusion-like operator, in the (18-dimensional !) space (M, ρ). The Eulerian probability distribution function $P(M, \rho)$ satisfies $\mathcal{L}P = 0$, with the normalization condition $\int dM P(M, \rho) = 1$. We also use the well-documented fact that the statistics of the velocity field at the integral scale is Gaussian. The stationary solution thus can be expressed in terms of a path integral,

$$P(M, \rho) = \int dM' \int dt\, G_t(M, \rho | M', \rho') P(M', |\rho'| = L). \tag{24}$$

where G is the Green's function of the problem. It can be expressed formally by integrating over all trajectories in the (M, ρ) space, with the prescribed boundary conditions, and with a statistical weight equal to $\sim \exp[-S]$, where S is the classical action [32]. In our problem it reads

$$G_t(M, \rho | M', \rho') = \int \mathcal{D}M \int \mathcal{D}\rho \exp[-S(M, \rho)], \tag{25}$$

with the boundary conditions $M(0) = M', \rho(0) = \rho'$ and $M(t) = M, \rho(t) = \rho$. The action S is defined by:

$$S = \int_0^t dt' \Big(\frac{||\dot{M} + \alpha(M^2 - \kappa tr(M^2))||^2}{C_M \varepsilon \rho^{-2}} + \frac{||\dot{\rho} - \rho \cdot M||^2}{C_u ||M|| \rho^2} \Big). \tag{26}$$

This formal solution of the problem relates the probability of having M at ρ to the (known) probability of having M' at scale L. It leads to appealing approximations, especially in the semiclassical limit ($C_M,\ C_u \to 0$), where the trajectories contributing to the integral can be found around the minimum of the action. A coarser approximation consists of taking the zero action solution, i.e., the most probable solution passing through (M, ρ) [31].

The knowledge of the entire probability distribution function would allow us to compute many interesting physical quantities. We restrict ourselves here to a the invariants of the tensor which characterize the topology of the flow. Specifically, the geometric invariants of the tensor M at a scale ρ are:

$$Q \equiv -\frac{1}{2} tr(M^2) \quad \text{and} \quad R \equiv -\frac{1}{3} tr(M^3) \tag{27}$$

(remember that $tr(M) = 0$, by incompressibility). These two invariants provide a characterization of the local topology of the flow. The eigenvalues of M are the roots of

$$\lambda^3 + \lambda Q + R = 0. \tag{28}$$

The zero discriminant line: $\Delta \equiv 27Q^3 + 4R^2 = 0$, separates the (R, Q) plane into two regions. For $\Delta > 0$, the flow is elliptic, with locally swirling streamlines. For $\Delta < 0$, strain dominates and the flow is locally hyperbolic.

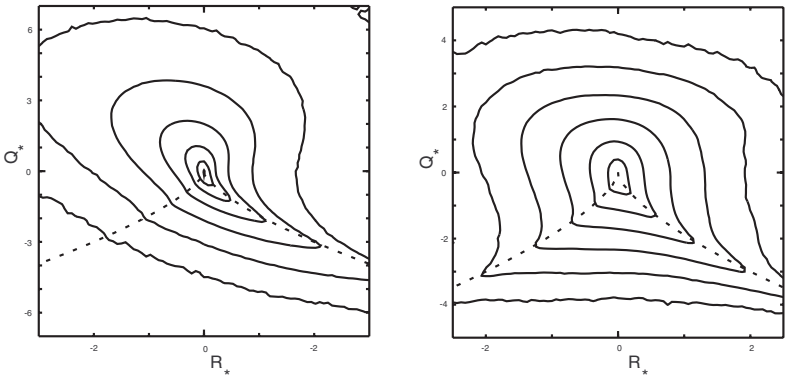

Fig. 5. The PDF of Q_*, R_* invariants normalized to the variance of strain, $Q_* \equiv Q/ < s^2 >$ and $R_* \equiv R/ < s^2 >^{3/2}$ ("star" denotes normalization), obtained from DNS at $R_\lambda = 85$, measured in the dissipation range $\rho = 2\eta$ (left) and in the inertial range $\rho = L/2 = 32\eta$ (right). The isoprobability contours are logarithmically spaced, and are separated by factors of 10. *The dashed line* corresponds to the separatrix: $4Q^3 + 27R^2 = 0$

Examples of probabilily distribution functions (PDF) of (R, Q) obtained by direct numerical simulations (DNS) of homogeneous isotropic turbulent flows are shown in Fig. 5. ($R_\lambda = 85$). The Reynolds number is moderate at $R_\lambda = 85$. The results are shown for $\rho \approx 2\eta$, corresponding to the dissipative range, and for $\rho \approx L/2$, in the inertial range, close to the integral scale.

In the dissipation range, the PDF shows a very skewed shape, with a large probability along the separatrix $\Delta = 0$, $R \geq 0$, and in the upper left quadrant. At higher values of ρ, the PDF is more symmetric with respect to the $R = 0$ line. The systematic evolution of the shape of the PDF as a function of ρ can be easily inferred from these figures.

An example of a PDF computed from the stochastic model computed in the zero action approximation is shown in Fig. 6.

In spite of the significant differences with the solution obtained from DNS of the Navier-Stokes equations, this approximated solution of the model reproduces the strong skewness of the numerical PDF, in particular the accumulation of probability along the zero discriminant line, for $R > 0$. A better approximation, such as the semiclassical approximation, is expected to produce a much better agreement with the DNS solutions.

The analysis of the model also sheds light on various aspects related to the energy transfer occuring at scale r [33]. This is potentially very useful, especially in the context of large eddy simulation (LES) of turbulent flows, a method consisting in simulating the large scale of the flow and treating the small scales by adequately parametrizing the energy transfer.

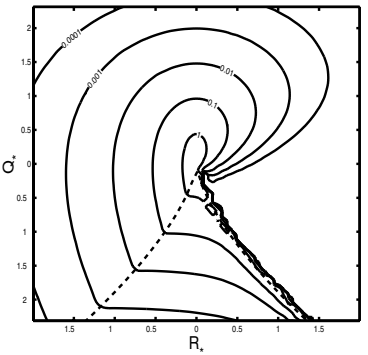

Fig. 6. PDF of Q_*, R_* invariants (normalized as in Fig. 5) calculated for the tetrad model in the deterministic approximation for $\rho/L = .5$

4 Conclusions

The recent advances in the understanding of passive scalar turbulence and the promising avenues that start to appear in hydrodynamic turbulence result from the fertile cross-breeding between two different subjects: the physical phenomenology of turbulence and the mathematical theory of stochastic processes. These bear the indelible mark of the genius of Andrei Nikolaievich Kolmogorov, to whom this review is dedicated.

We acknowledge numerous fruitful discussions with G. Boffetta, P. Castiglione, M. Chertkov, B. Shraiman, and M. Vergassola. This work has been supported by the EU under the contract HPRN-CT-2002-00300 and by Cofin 2001, prot. 2001023848.

References

1. A.N. Kolmogorov, "Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung" Math. Ann. **104** (1931) 415. English translation "On analytical methods in probability theory" in *Selected works of A.N. Kolmogorov* Vol. II (ed. A.N. Shiryayev) Kluwer Acad. Publ. (1992) 62
2. A.N. Kolmogorov, Dokl. Akad. Nauk SSSR **30** (1941) 3201. English translation "The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers" Proc. R. Soc. Lond. A **434** (1991) 9
3. L. Biferale, G. Boffetta, and B. Castaing, "Eulerian turbulence", this volume, Chap. 7
4. B.I. Shraiman and E.D. Siggia, Nature **405**, 639 (2000)
5. G. Falkovich, K. Gawędzki, and M. Vergassola, Rev. Mod. Phys. **73**, 913 (2001)
6. U. Frisch, *Turbulence: the legacy of A.N. Kolmogorov*, (1995) Cambridge university Press, Cambdridge, UK
7. A.M. Obukhov, C. R. Acad. Sci. URSS, Geogr. Geophys., **13** (1949) 58; S. Corrsin, J. Appl. Phys. **22** (1951) 469

8. G.K. Batchelor, J. Fluid Mech. **5** (1959) 113
9. T. Bohr, G. Paladin, M.H. Jensen, and A. Vulpiani, *Dynamical Systems Approach to Turbulence*, (1998) Cambridge university Press, Cambdridge, UK
10. A. Celani, A. Lanotte, A. Mazzino, and M. Vergassola, Phys. Rev. Lett. **84** (2000) 2385
11. K.R. Sreenivasan, Proc. Roy. Soc. London A **434** (1991) 165
12. M. Holzer and E. Siggia, Phys. Fluids **6** (1994) 1820
13. A. Pumir, Phys. Fluids **6** (1994) 2118
14. R.H. Kraichnan, Phys. Fluids **11** (1968) 945
15. R.H. Kraichnan, Phys. Rev. Lett. **72** (1994) 1016
16. K. Gawędzki and A. Kupiainen, Phys. Rev. Lett. **75** (1995) 3834
17. M. Chertkov, G. Falkovich, I. Kolokolov, and V. Lebedev, Phys. Rev. E **2** (1995) 4924
18. B.I. Shraiman and E.D. Siggia, C. R. Acad. Sci. Paris, série II **321** (1995) 279
19. U. Frisch, A. Mazzino, and M. Vergassola, Phys. Rev. Lett. **80** (1998) 5532
20. R. Livi, S. Ruffo, and D.L. Shepelyansky, "Kolmogorov pathways from integrability to chaos and beyond", this volume, Chap. 1
21. A. Celletti, C. Froeschle, and E. Lega, "From stable to chaotic motions through the work of Kolmogorov" this volume, Chap. 2.
22. D. Bernard, K. Gawędzki, and A. Kupiainen, J. Stat. Phys. **90** (1998) 519
23. A. Celani and M. Vergassola, Phys. Rev. Lett. **86** (2001) 424
24. O. Gat and R. Zeitak, Phys. Rev. E **57** (1998) 5511
25. A. Pumir, B. I. Shraiman, and M. Chertkov, Phys. Rev. Lett. **85** (2000) 5324; P. Castiglione and A. Pumir, Phys. Rev. E **64** (2001) 056303
26. O. Gat, R. Zeitak, and I. Procaccia, Phys. Rev. Lett **80** (1998) 5536
27. L. Biferale, G. Boffetta, A. Celani, A. Lanotte, F. Toschi, and M. Vergassola, Phys. Fluids **15** (2003) 2105; see also http://xxx.arxiv.org/abs/nlin.CD/0301040
28. H. Tennekes and J.L. Lumley, *A first course in turbulence* , MIT Press (1983)
29. B. To, J. Katz, and C. Meneveau, J. Fluid Mech. **457**, 35 (2002)
30. V. Borue and S.A. Orszag, J. Fluid Mech. **336**, 1 (1998)
31. M. Chertkov, A. Pumir, and B. Shraiman, Phys. Fluids **11**, 2394 (1999)
32. R.P. Feynman and A.R. Hibbs, "Quantum Mechanics and Path Integrals", Mc Graw Hill (1965)
33. A. Pumir, B. Shraiman, and M. Chertkov, Europhys. Lett. **56** 379 (2001)

Reaction-Diffusion Systems: Front Propagation and Spatial Structures

Massimo Cencini[1], Cristobal Lopez[2], and Davide Vergni[3]

[1] INFM Center for Statistical Mechanics and Complexity, Dipartimento di Fisica di Roma "La Sapienza", P.zzle Aldo Moro, 2 Rome, Italy, 00185, Massimo.Cencini@roma1.infn.it
[2] Instituto Mediterraneo de Estudios Avanzados (IMEDEA) CSIC-UIB, Campus Universitat Illes Balears Palma de Mallorca, Spain, 07122, clopez@imedea.uib.es
[3] Istituto Applicazioni del Calcolo, IAC-CNR V.le del Policlinico, 137 Rome, Italy, 00161, Davide.Vergni@roma1.infn.it

Abstract. After the pioneering works of Kolmogorov, Petrovskii and Piskunov [1] and Fisher [2] in 1937 on the nonlinear diffusion equation and its traveling wave solutions, scientists from many different disciplines have been captivated by questions about structure, formation and dynamics of patterns in reactive media. Combustion, spreading of epidemics, diffusive transport of chemicals in cells and population dynamics are just a few examples bearing witness of the influence of those works in different areas of modern science.

1 Introduction

In many natural phenomena we encounter propagating fronts separating different phases. An unfortunately familiar example is the front separating burnt from unburnt trees in forest fires. Similarly, propagating fronts play an important role in the speed of epidemics, in population dynamics, or in the propagation of flames and chemical reactions. Most of these, at first glance disparate phenomena find their common denominator in the presence of *diffusion* (allowing the agent of an epidemic or a chemical substance to spread), and *reaction* (that is the specific way in which different phases or chemical components react); they are generically referred to as *reaction diffusion* (RD) systems.

The prototypical model for RD systems is the nonlinear diffusion equation

$$\frac{\partial}{\partial t}\theta(x,t) = D\frac{\partial^2}{\partial x^2}\theta(x,t) + F(\theta)\,, \tag{1}$$

introduced[1] in 1937 in the seminal contributions of R.A. Fisher [2] and A.N. Kolmogorov, together with I.G. Petrovskii and N.S. Piskunov [1] (hereafter referred to as FKPP), as a model to describe the spreading of an advantageous gene. (1) describes the spatio-temporal evolution of a population

[1] As mentioned in Murray (see p. 277 in [3]), (1) was already introduced in 1906 by Luther.

M. Cencini, C. Lopez, and D. Vergni, Reaction-Diffusion Systems: Front Propagation and Spatial Structures, Lect. Notes Phys. **636**, 187–210 (2003)
http://www.springerlink.com/

(concentration), $\theta(x,t)$, of individuals which diffuse with diffusion coefficient D, and grow according to a specific rule $F(\theta)$. In FKPP it was shown that (1) admits uniformly translating solutions – traveling waves.

Similar types of propagation phenomena are ubiquitous in Nature. The concepts and mathematical tools developed in [1,2] stand at the foundation of a still increasing number of applications of the reaction diffusion equations in biology, chemistry and physics (see [3–6] and references therein).

The present knowledge on reaction diffusion systems is so vast that it cannot be presented here in a comprehensive and systematic way. Therefore, our discussion will be limited to introductory material. The first part of this chapter is devoted to (1) in one spatial dimension, providing the reader with the main concepts and simplest mathematical tools necessary to understand its behavior. In the second part we enlarge the discussion to generalizations of (1) in moving media with more than one reacting species and, to dimension higher than one.

2 Front Propagation in the Nonlinear Diffusion Equation

Perhaps the best way to start our discussion on (1) is to motivate it as originally proposed in FKPP. Consider an area populated by individuals of the same species. Suppose that $\theta(x,t)(\in[0,1])$ is the concentration of the subset of these individuals which possess a particular genotype that makes them favored in the struggle for survival. In particular, assume that the survival probability of individuals with that character is $1+\alpha$ ($\alpha>0$) times larger than that of individuals without it. Then the evolution of the concentration θ is ruled out by the standard logistic growth model

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = F(\theta) = \alpha\theta(1-\theta)\,. \tag{2}$$

The above equation implies that starting from $\theta\approx 0$ there is an initial exponential growth $\theta\sim\exp(\alpha t)$ followed by a saturation at $\theta=1$ due to nonlinearities. Hence $\theta=0$ is an unstable state and $\theta=1$ a stable one.

If, during one generation (the period between birth and reproduction), individuals move randomly in any direction, the concentration evolution is given by (1) with $F(\theta)$ as in (2).

Now if the concentration of individuals with the advantageous genotype is initially different from zero only in a small region, it is natural to ask how it will spread over the space. Specifically, following Kolmogorov et al., let us assume that at $t=0$ there is a localized region in which the density is different from 0 and 1, and on the left of this region $\theta=1$ while on the right $\theta=0$. By means of the combined effect of diffusion and reaction, the region of density close to 1 will expand, moving from left to right. In other words,

at long times, $\theta(x,t)$ can be expressed as

$$\theta(x,t) = \Theta_v(x - vt)\,, \tag{3}$$

meaning that the concentration behaves as a wave with propagation velocity v and shape Θ_v.

The problem is to find the limiting shape of the density profile and the limiting rate of its motion. Nowadays, after the efforts of many scientists who extended and generalized Kolmogorov results to different classes of nonlinear terms and generic initial conditions, this problem is well understood (see [5,7–9] and references therein). In the following, we present the modern understanding of it, trying to remain at an intuitive level of discussion.

First of all let us consider the general equation (1), rewritten here for convenience

$$\frac{\partial}{\partial t}\theta(x,t) = D\frac{\partial^2}{\partial x^2}\theta(x,t) + F[\theta(x,t)]\,. \tag{4}$$

Without specifying the shape of $F(\theta)$, we assume two steady states, an unstable one ($\theta = 0$) and a stable one ($\theta = 1$), i.e. $F(\theta)$ satisfies the conditions

$$\begin{aligned} &F(0) = F(1) = 0\,; \\ &F(\theta) > 0 \quad \text{if} \quad 0 < \theta < 1\,. \end{aligned} \tag{5}$$

Pulled versus Pushed Fronts

Within the assumptions (5), we can distinguish two classes of nonlinear terms. The first one, often indicated as FKPP-like, is characterized by having the maximum slope of $F(\theta)$ for $\theta = 0$ (as for the logistic growth model (2), see Fig. 1a). This is the case of the so-called *pulled* fronts, for which the front dynamics can be understood by linear analysis since it is essentially

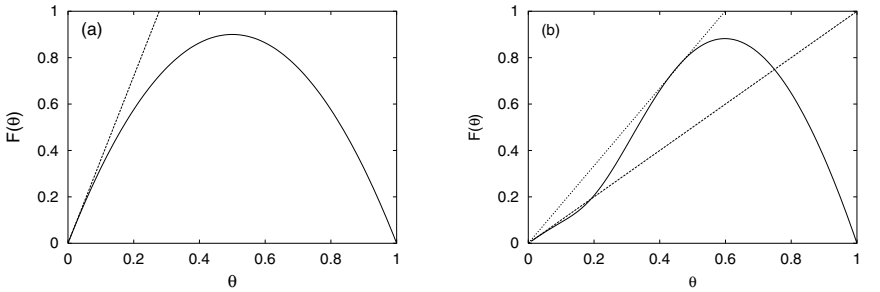

Fig. 1. a A typical FKPP-like production term (pulled dynamics). **b** A production term which produces a pushed dynamics. The *dashed* and the *dotted straight lines* display the linear behaviors $F'(0)\cdot\theta$ and $(\sup_\vartheta \{F(\vartheta)/\vartheta)\})\cdot\theta$, respectively. See text for explanation

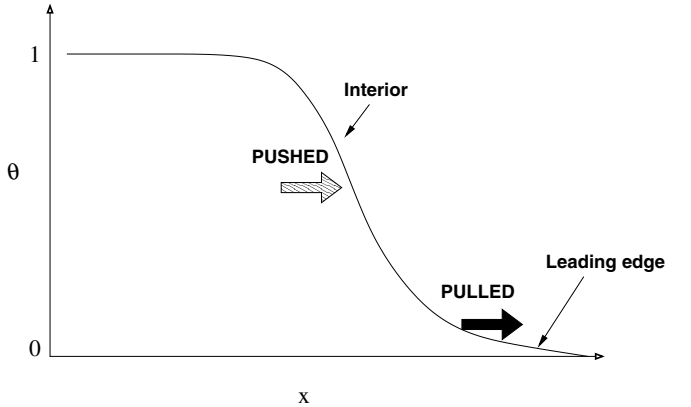

Fig. 2. Pictorial representation of FKPP (pulled) and non FKPP (pushed) fronts

determined by the $\theta(x,t) \approx 0$ region (so the front is pulled by its leading edge, see Fig. 2). In the second class, non FKPP-like, the maximal growth rate is not realized at $\theta = 0$ but at some finite value of θ, see Fig. 1b, where the details of the nonlinearity of $F(\theta)$ are important. In this case front dynamics is often referred as *pushed*, meaning that the front is pushed by its (nonlinear) interior, Fig. 2. In contrast with the previous case, a detailed non linear analysis of (4) is now required to determine the front speed.

In both pushed and pulled fronts there exists a one parameter family of traveling wave solutions, Θ_v, characterized by their velocity, v. In their work of 1937, Kolmogorov et al. proved that not all velocities are allowed for pulled fronts. Indeed the following inequality has to be satisfied

$$v \geq v_0 = 2\sqrt{DF'(0)}\,.$$

Furthermore, the equality $v = v_0$ is always realized for localized initial conditions as the one mentioned above. This result was lately generalized by Aronson and Weinberger [7] to generic production terms $F(\theta)$. They showed that the minimal allowed front speed, v_{min}, is bounded by (see Fig. 1)

$$2\sqrt{DF'(0)} \leq v_{min} < 2\sqrt{D \sup_{\theta} \left\{ \frac{F(\theta)}{\theta} \right\}}\,. \tag{6}$$

Note that $F(\theta)/\theta$ is a measure of the growth rate, and that for FKPP dynamics $\sup_\theta \{F(\theta)/\theta\} = F'(0)$ which implies Kolmogorov's result $v_{min} = v_0$.

Equation (6) bounds the minimal front speed but, in principle, front solutions with $v \geq v_{min}$ are allowed. Therefore, it is important to determine the velocity that is actually selected for a given initial condition. This is the so-called velocity selection problem.

FKPP-Like Reaction Terms

Here, like Kolmogorov et al., we assume that $F(\theta)$ fulfills the conditions (5) supplemented by

$$F'(0) > 0 \quad \text{and} \quad F'(\theta) < F'(0) \quad \text{for all} \quad 0 < \theta \leq 1\,, \tag{7}$$

ensuring that $F'(0) \equiv \sup_\theta \{F(\theta)/\theta\}$. Note that assumptions (5) and (7) are quite reasonable in biological problems and in some chemical reactions. Then from (4) by choosing the frame of reference moving with the front, i.e. with the variable change $z = x - vt$, one obtains the equation for the limiting front profile

$$D\frac{\mathrm{d}^2}{\mathrm{d}z^2}\Theta_v(z) + v\frac{\mathrm{d}}{\mathrm{d}z}\Theta_v(z) + F(\Theta_v) = 0\,, \tag{8}$$

with boundary conditions $\Theta_v(-\infty) = 1$ and $\Theta_v(+\infty) = 0$. In the case of localized initial conditions, Kolmogorov and coworkers rigorously demonstrated, using a very interesting constructive proof, that (8) has a positive definite[2] solution with speed

$$v_0 = 2\sqrt{DF'(0)}\,. \tag{9}$$

Such a solution exists and is unique, apart from a linear transformation $x' = x + c$ which does not modify the front profile.

The fact that many solutions with different velocities appear and that v_0 is the minimal one can be inferred by linearizing (4) around $\theta = 0$ (which is the important region in the pulled regime)

$$\frac{\partial}{\partial t}\theta(x,t) = D\frac{\partial^2}{\partial x^2}\theta(x,t) + F'(0)\,\theta\,. \tag{10}$$

In the neighborhood of $\theta(x,t) \approx 0$, i.e. in the leading edge region, it is reasonable to expect an exponential profile (this is actually always observed) so that for $x \to \infty$, while t is large but finite, one can write

$$\theta(x,t) \sim e^{-a(x-vt)}\,, \tag{11}$$

where $1/a$ is a measure of the flatness/steepness of the profile. Substituting the last expression in (10) one finds the link between asymptotic front shape and speed,

$$v = Da + \frac{F'(0)}{a}\,, \tag{12}$$

[2] Notice that if the initial concentration is non negative, $\theta \geq 0$, it will remain so under the dynamics (4). This follows immediately by interpreting (4) as the heat equation with heat source $F(\theta)$, which by (5) is never negative.

which is the so-called dispersion relation. This equation implies that profiles with different velocities are allowed and that there exists a minimal velocity, v_{min}, realized for

$$a^* = \sqrt{\frac{F'(0)}{D}}\,, \tag{13}$$

corresponding to $v_{min} = 2\sqrt{F'(0)D}$, which is Kolmogorov's result (9).

It is possible to show that front solutions with $a > a^*$ are unstable [8–10]. This is a crucial point which is at the core of the speed selection problem. In fact, one observes that for steep enough initial conditions ($a \geq a^*$), the front always relaxes to the one of minimal speed. On the other hand, if the initial conditions are sufficiently flat, $a < a^*$, the front propagates with a rate given by (12), thus the broader the front ($a \to 0$) the faster it moves. We will comment further on the selection mechanism at the end of this section.

Let us now show how, for localized initial conditions, the front always converges to the minimal speed solution [10]. Writing $\theta(x,t) = \exp[F'(0)t]\,\phi(x,t)$, (10) reduces to the heat equation for the new variable ϕ ($\partial_t \phi = D\partial_x^2 \phi$) which can be easily solved. In terms of θ the solution reads

$$\theta(x,t) = \exp(F'(0)t) \int_{-\infty}^{\infty} \mathrm{d}y\, \theta(y,0)\, \frac{\exp\left[-\frac{(x-y)^2}{4Dt}\right]}{\sqrt{4\pi D t}}\,. \tag{14}$$

Now introducing the coordinate $z = x - v_0 t$ with v_0 given by (9), and assuming that $\theta(y,0)$ is different from 0 only in a small region (say around the origin) one obtains

$$\theta(x,t) = \Theta_v(z) \propto \frac{\exp(-z\sqrt{F'(0)/D} - z^2/4Dt)}{\sqrt{t}}\,. \tag{15}$$

This equation tells us that the front shape asymptotically approaches an exponential profile with steepness $\xi \propto 1/a^* = \sqrt{D/F'(0)}$, and that the front speed v_0 is reached as

$$v(t) - v_0 \propto \frac{1}{t}\,, \tag{16}$$

i.e. in an algebraically slow way[3].

Pushed Fronts

For pushed fronts, conditions (7) are not satisfied. This implies that the maximal growth rate is not realized at $\theta = 0$ but in the non linear interior of the

[3] Actually it has been shown that the prefactor $\frac{1}{t}$ is universal in FKPP-like fronts [10]. Here we just stress that the algebraic convergence comes out from the $1/\sqrt{t}$ prefactor characteristic of the Gaussian propagator.

front (see Figs. 1b and 2). As a relevant example, we mention thermally activated chemical reactions (such as combustion) which are typically modeled using the Arrhenius production term,

$$F(\theta) = (1-\theta)e^{-\theta_c/\theta}\,, \tag{17}$$

where below the activation concentration, θ_c, essentially no reaction takes place.

Contrary to pulled fronts, where the front speed can be determined by linear analysis, a fully nonlinear treatment is required here. However, there still exists a minimal velocity v_{min} below which no solutions are allowed. The point is that now $v_{min} > v_0$, with v_0 given by (9).

A simple way to understand how a minimal velocity larger than v_0 appears can be found in [9]; here we report the main idea. We have seen that the leading edge is always exponential, so that for large $z = x - vt$

$$\Theta_v(z) = A_v^F \exp[-a_F(v)z] + A_v^S \exp[-a_S(v)z], \tag{18}$$

where $a_F(v)$ and $a_S(v)$ ($> a_F(v)$) are the flat and the steep modes, respectively. In the above analysis (see (11)) to derive the dispersion relation (12), we considered only the flat decreasing mode, because asymptotically it is the leading one. However, seeing that (8) is a second order equation, a superposition like (18) is expected in general. The constants A_v^F and A_v^S depend on the velocity v, and on the nonlinear part of the front through matching conditions with the interior. As before, allowed front solutions should be positive, meaning that at least A_v^F should be positive. For large enough v, front solutions are allowed since both A_v^S and A_v^F are positive. If, for $v = v_{min}$, the amplitude of the leading mode A_v^F goes to zero, then, for continuity, A_v^F will become negative for $v < v_{min}$, by continuity. As a consequence, the corresponding Θ_v is not an allowed solution. Precisely at $v = v_{min}$ (18) reduces to the single fast exponential decrease. Note that for pulled fronts at $v = v_0$, $a_S(v_0) = a_F(v_0)$; see [10] for a discussion about this point.

Also in this case, depending on the flatness/steepness of the initial profile, the asymptotic front speed may be larger than v_{min} or may relax to the minimal one.

Velocity Selection

For initial conditions steep enough (including localized initial conditions), the front dynamics is always attracted by the minimal speed solution which, for pulled fronts, corresponds to the linear prediction (9) and in general satisfies the bounds (6). The detailed proof of this statement requires a non trivial analysis which depends crucially on the simplicity of the model under consideration. However, while remaining at the level of a general discussion, it is interesting here to briefly recall the ideas at the basis of the modern way in which the speed selection problem is understood.

The crucial concept is that of *stability in the moving frame* or *marginal stability condition* [5,8,9], which may be formulated as follows. If one adds a small perturbation to the leading edge of a moving front, then the front is stable if it outruns the perturbation (it is left behind and readsorbed) and unstable if the perturbation persists at long times. The natural front is self-sustained, i.e., the growth of the perturbation from the unstable to the stable state should be the cause of front propagation. Therefore, the selected front should be the one which is marginally stable with respect to the perturbation. In this framework, the Aronson and Weinberger result [7] is equivalent to the statement that the front with the lowest velocity is marginally stable with respect to local perturbations of the state $\theta = 0$. This criterion can be generalized also to the case of pushed fronts [8,9].

2.1 Multiple Steady States

Up to now, we have considered reaction terms having only two steady states. However, in a broad class of problems in nonlinear chemistry and population dynamics, such as enzymatic reactions or insects spreading [3], multiple steady states may be present, meaning that the production term have $N \geq 3$ zeros in $[0, 1]$

$$F(\theta_i) = 0\,, \qquad \text{for} \qquad i = 1, ..., N\,.$$

These fixed points can be stable or unstable and more complicated propagation phenomena can appear.

In order to provide the reader with some basic ideas, let us introduce a simple and instructive description of the front propagation problem exploiting an analogy with the dynamics of a point particle [5,8]. To make it evident let us rewrite (8) as

$$D\ddot{y} + v\dot{y} + F(y) = 0\,, \tag{19}$$

where $y \equiv \Theta_v$ and the dots indicate derivatives with respect to the variable $z = x - vt$, which here represents time. The reader will recognize that this is the equation for a classical particle moving in a potential

$$V(y) = \int^{y} \mathrm{d}y'\, F(y')\,, \tag{20}$$

and damped with a friction coefficient v.

By using this analogy, the existence of a minimal velocity below which no uniformly translating fronts exist has a clear interpretation [10]. Let us reconsider, for the sake of simplicity, the case of pulled fronts in the framework of linear analysis (10). We assume a parabolic potential $V(y) = -F'(0)y^2/2$. Due to the friction term, at sufficiently long times, an exponential decay, $y(t) \sim \exp(-at)$ is expected (i.e. an exponential front profile at large distances). Substituting this behaviour in (19) one obtains that

$$a(v) = \frac{v \pm \sqrt{v^2 - 4F'(0)D}}{2D}\,.$$

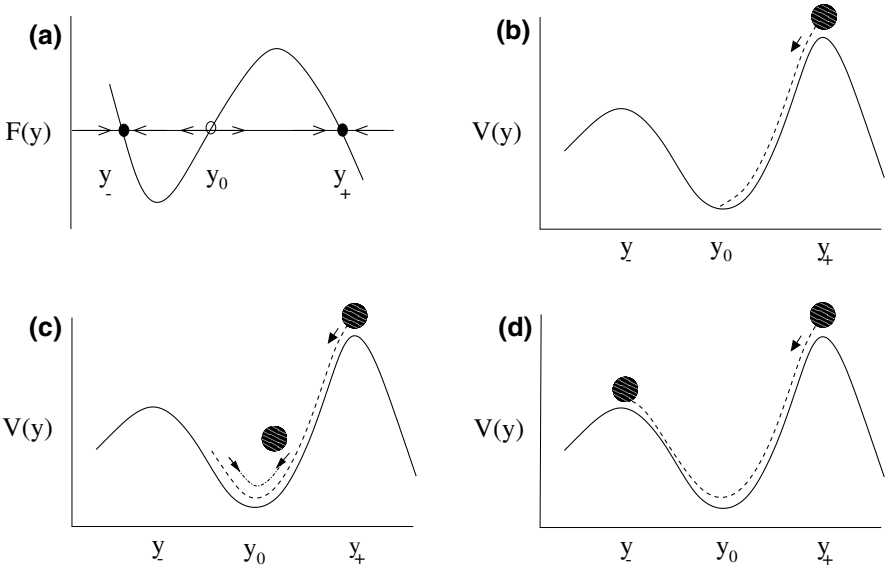

Fig. 3. Pictorial representation of the particle in the potential analogy. **a** Generic production term with three steady states, two stable (filled circles) and one unstable (empty circle). **b,c,d** The potential obtained with (20), see text for explanation on the different situations

Now if the damping is not strong enough ($v < 2\sqrt{DF'(0)}$) the particle will reach $y=0$ in a finite time implying that the front becomes negative, which is not allowed. Therefore, $v=2\sqrt{DF'(0)}$, is the minimal friction ensuring that the particle will asymptotically reach $y=0$, and so the front remains positive and monotonic.

This analogy becomes very useful in the presence of many steady states. For example, let us consider a generic function $F(y)$ having three zeros [5] (see Fig. 3a): an unstable state at y_0 and two stable states at $y_\pm$, corresponding to the minimum and the two maxima of the potential $V(y)$, respectively. A front (in general any structure connecting two different states) is a trajectory connecting one of the maxima with the minimum, e.g. y_+ with y_0.

For the parabolic potential previously examined, for large enough v (say $v \geq v_1$), the damping is efficient and y_0 is reached at $t \to \infty$, i.e. the front is monotonic (see Fig. 3b). Below the critical damping v_1, there is an overshoot before reaching y_0 and the particle will pass y_0 going uphill toward y_- before ending in y_0. Below another critical value $v_2 < v_1$, the approach to the minimum may be underdamped the particle oscillate for ever in the valley (Fig. 3c); that is, the leading edge of the front is oscillatory. There also exists a critical value v_3 ($< v_2 < v_1$) for which the particle lands precisely at y_-, the front joins two stable states (Fig. 3d). For $v < v_{min}$ the orbit goes to $-\infty$, which does not represent a finite solution of (19). Notice that contrary

to fronts propagating from an unstable state to a stable one, for those joining two stable states there exists a unique solution and not a family of solutions – only one speed value is allowed.

3 Reaction Diffusion Systems in Physics, Chemistry, and Biology

In the previous sections we have examined the one-dimensional reaction diffusion equation. Now, after a brief overview on the wide range of applicability of this equation in different areas of science, we will focus on some specific issues such as multicomponent chemical reactions, combustion and an ecological problem concerning the distribution of plankton in the ocean.

Chemical Reactions

The most natural application of the nonlinear diffusion equation is the study of chemical reactions taking place in the environment or in living organisms. In multicomponent chemical reactions, one considers generalizations of (1) where many species with their interrelations and diffusion constants are present. Moreover, depending on the media where the reaction takes place, one can have either an advection term (reaction in fluid flows), or spatial a dependence in the diffusion coefficient [11] (reaction in heterogeneous media). In the presence of many species, the problem becomes much more difficult. Indeed a large range of behaviors, from oscillations to chaos [12], can be found. As it will become clear in Sect. 3.1, this gives rise to much more complex spatio-temporal propagation phenomena than in the simple one-dimensional case (see [3,4,13] and references therein).

Combustion Theory

Among the many chemical reactions, for its theoretical and practical importance, we mention the problem of combustion [14], which has been the first application [15] of concepts and tools originally introduced in FKPP. Combustion problems are complicated not only by the presence of many reactants, but also by the fact that the burning of combustible takes place in a moving medium, usually a fluid. Hence one has to include in (1) the advection by the fluid velocity, and we speak about *reaction advection diffusion systems*. This increases enormously the complexity and difficulty of the problem because the fluid motion is usually very complex due to turbulence [16], which is another fundamental aspect of Kolmogorov's interests (see Chaps. 7 and 8). In Sec. 3.2, we will discuss in details this problem.

Population Dynamics and Ecology

The contributions of Fisher and Kolmogorov on the reaction diffusion equation had a prominent role in the development of mathematical tools in *population dynamics* and *ecology* (see, e.g., [3,6]). Indeed (1) and its generalizations

are at the basis of many studies ranging from the rate of advance of invading species [3] to human genetics and expansion [17]. Closely related to these works, and building upon them, has been the development of models to explain patchiness in the distribution of organisms [6], which is an important issue in the study of plankton and zooplankton in oceans [18]. In Sec. 3.2 we will provide the reader with a brief overview on this ongoing research area.

Pattern Formation and Developmental Biology

Finally, it is important to mention that RD systems do not only give rise to propagation phenomena but also to standing-patterns. Steady heterogeneous spatial structures, or patterns, appear in nature running from the very small scales, like in colonies of bacteria, to astronomical ones, like the spiral structure of some galaxies [3,5]. The interest is then in understanding *pattern formation*.

A central role in pattern formation studies was played by another great scientist, namely A. Turing[4] who, in a classic paper [19], showed that pattern forming instabilities may arise when considering RD mechanisms. Even if there is no room here to properly treat this problem, we mention that already from the early work of Turing it was realized the potentiality of RD modeling for developmental biology [3]. Probably the most striking example in this context is offered by *morphogenesis*, i.e., the development of patterns and forms from the initially homogeneous mass of cells in the embryo of many animals. For instance, think of the richness and beauty of patterns in animal coats or butterflies leaves [3].

Nowadays, many different formation mechanisms have been identified, and pattern formation in RD systems is a very vast and important area in the study of non-equilibrium physical and chemical systems [5].

3.1 Multi-components Reaction Diffusion Systems

The general mathematical expression for a multicomponent reaction diffusion system is just an extension of (4) to an N-components vector field $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$,

$$\frac{\partial \theta_i}{\partial t} = D_i \nabla^2 \theta_i + F_i(\theta_1, \ldots, \theta_N), \tag{21}$$

where $F_i(\theta_1, \ldots, \theta_N)$ is the reaction term for the i-th species, and D_i its diffusivity. Very complex behaviors appear now depending on the intrinsic reaction dynamics given by the F_i's. To illustrate the phenomenology of these systems, we will use two paradigmatic examples: the celebrated Belusov-Zhabotinskii (BZ) chemical reaction [12], and the predator-prey (PP) systems [6], with many applications in populations dynamics and ecology.

[4] From a historical point of view it is interesting to know that Turing was not aware of the works of Kolmogorov, Petrovskii and Piskunov [1], which was poorly known in the West for many years.

The BZ reaction is probably the most widely studied, both theoretically and experimentally, oscillating (it can also have excitable or chaotic behavior) chemical reaction. It involves more than 40 elementary reactions which result in changes of several dozens of intermediate substances. The basic mechanism consists of the oxidation of malonic acid by bromate ions, catalyzed by, e.g., ferroin and ferriin. For some values of the reagent concentrations, periodic oscillations are observed in these catalysts, producing a periodic color change oscillating between red (ferroin) and blue (ferriin). More details can be found, for example, in [12]. Concerning the PP systems, the simplest model involves two components, predators and preys (the concentrations of which are denoted by v and u, respectively), and can be written

$$\frac{du}{dt} = ru(1 - \frac{u}{u_0}) - cvf(u) + D\nabla^2 u, \tag{22}$$

$$\frac{dv}{dt} = avf(u) - bv + D\nabla^2 v, \tag{23}$$

where r, c, a, b, u_0 are positive parameters and D is the diffusivity. The first term on the rhs of (22) indicates the intrinsic birth-death of the preys; in the second term, f is the prey consumption function per predator. Analogously, the first term on the rhs of (23) is the benefit from predation and the second one models predators' death.

It is important to remark on the universality of behavior in this class of systems: similar wave patterns to those found in the BZ reaction (see below Figs. 4 and 7) or the PP model appear also in many other reaction diffusion systems having the *same* dynamical behavior.

Let us start with the case in which the F_i's give rise to an oscillatory dynamics, inducing a periodic evolution of the Θ_i fields in time. A front may develop, for example, if the oscillation of any part of the system is perturbed, and traveling wave *trains* move through the system. In the context of predator-prey systems, these periodic traveling waves can be originated by the invasion of a predator population into the prey population. In one dimension, wave train solutions are of the form

$$\theta_i(x,t) = \Theta_i(\omega t - kx), \tag{24}$$

where ω is the frequency, k the wavenumber, and Θ_i is a periodic function of the phase. Therefore the advancing front leaves behind it a spatially periodic pattern. In two spatial dimensions, these wave trains are concentric circles, referred to as *target patterns* (see Fig. 4 for an example appearing in the BZ reaction).

Other interesting behaviors appear in *excitable systems*, which are characterized by the presence of *activator* and *inhibitor* components. The activator has a catalytic effect both on itself (autocatalysis) and on the inhibitor which, in turn, depletes the activator production. At the end, the system has a stable fixed point as the only attractor for the dynamics. Examples of excitable systems may be found in semiconductor lasers with optical feedback [20], neural

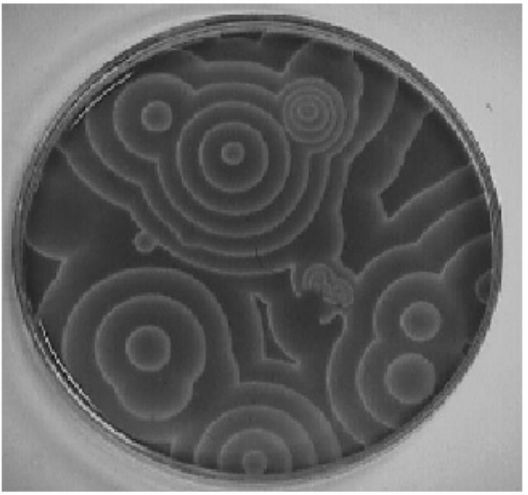

Fig. 4. Target patterns for the BZ reaction

communications [3], and populations dynamics [3]. Moreover, for some values of the parameters, the BZ reaction and PP models also behave as excitable systems.

The main feature of this kind of systems is the way in which they respond to perturbations. Typically there exists a *threshold* value such that if the perturbation goes above it the system reaches the fixed point only after a long excursion in the phase space. This behavior usually appears when the activator has a temporal response much faster than the inhibitor, so that it takes some time before stopping the growth of the activator. The threshold property is characteristic of cubic nonlinearities in the reaction term, as exemplified by the Fitzhugh-Nagumo (FN) equations,

$$\begin{aligned} \frac{\partial u}{\partial t} &= u(a-u)(u-1) - v + D\frac{\partial^2 u}{\partial x^2}\,, \\ \frac{\partial v}{\partial t} &= bu - \gamma v\,, \end{aligned} \tag{25}$$

originally introduced as a mathematical model for neural activity [21], where u is the activator, v the inhibitor, and a, b, γ are positive parameters.

The threshold property can be understood by studying the nullclines of the system (Fig. 5), which are obtained by equating to zero the rhs of (25) with $D=0$. If the value of u is smaller than the threshold a, u quickly returns to the origin (the stable fixed point) and the spatial perturbation dies out. On the contrary, if the perturbation is larger than the threshold, the fixed point is reached after a large excursion in both u and v passing through the points $0BCD0$.

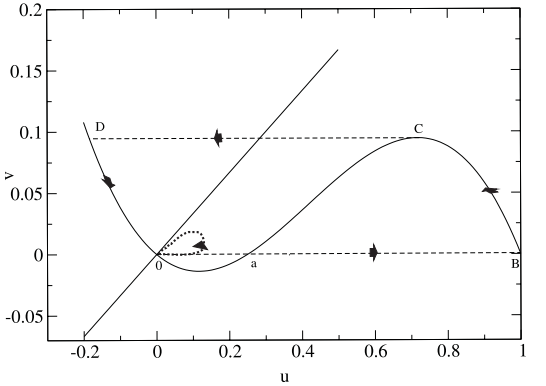

Fig. 5. Phase trajectories for u and v depending on whether the perturbation is larger than or smaller than the threshold a. *Solid lines* are the nullclines

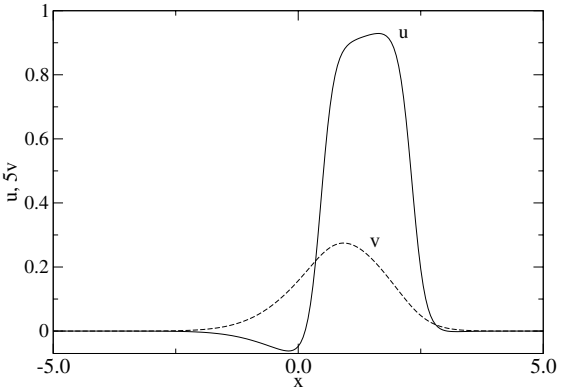

Fig. 6. Pulse-like front solution for a one-dimensional excitable medium

The propagation of the excitation through neighboring points, coupled diffusively, generates traveling *pulses* as the one shown in Fig. 6. In two dimensions we have circular propagating waves. Pulse waves have been shown to exist in generic excitable models, but the values of the propagation velocity and the shape of the pulse depend on the specific reaction term. In particular for the FN system, one can show [3] that, in the limit of small b and γ, the wave speed is given by $c = \sqrt{D/2}(1 - 2a)$. In two dimensions, when a propagating pulse is broken at a point, it begins to rotate around the ends, producing *spiral waves* (see Fig. 7 for a typical example in the BZ reaction). There are also many other relevant occurrences of spiral waves in natural systems. Just to name a few, let us mention fibrillating hearts, where small regions contract independently and the spreading through the cortex of damaged brains [3].

Fig. 7. Spiral waves observed in the BZ reaction

However, the phenomenology can be much more complicated. For example, target patterns can also be formed in an extended excitable medium if the pulses are emitted periodically, and spiral waves can be formed by breaking target waves by stirring the medium, or by noise-induced effects.

Let us now briefly comment the case of chaotic reaction dynamics (see [4,5] for more details). An interesting case, widely observed in predator-prey systems, appears when periodic wave trains become highly disordered, loosing their periodicity. In this case, very irregular spatial patterns appear behind the front. Moreover, spiral waves may become highly disordered and organize themselves in chaotic sets that continuously form and decay (the so-called transition to spatio-temporal chaos).

3.2 Advection Reaction Diffusion Systems

Reaction diffusion processes taking place in moving media such as fluids are of considerable importance, e.g. in combustion, atmospheric chemistry, and ecological problems. As we saw in Chap. 8, even passive scalars, substances which are simply transported by the flow, display a very complex behavior both in laminar and turbulent flows. When reaction is taken into account the problem becomes even more complex. In fact the heat release associated with chemical reactions will affect the velocity field, and transport is not passive any more. However, even when the feedback of the advected scalar on the moving medium can be neglected (like in certain aqueous autocatalytic reactions), the dynamics of the reacting species is strongly complicated by the presence of the velocity field.

Similarly to passive scalars, one can adopt two complementary points of view. A first possibility is to consider particles modeling reagents (or individuals, in ecological problems), which move under the effect of a velocity field and thermal noise (diffusion), and reacting when they come into contact – this is the *Lagrangian* viewpoint (see also Chap. 8). Alternatively, adopting an *Eulerian* viewpoint (see also Chap. 7), one considers a field of concentration which evolves according to the *advection reaction diffusion* (ARD) equation, which for one species reads

$$\frac{\partial \theta}{\partial t} + \mathbf{u} \cdot \boldsymbol{\nabla} \theta = D \nabla^2 \theta + F(\theta), \tag{26}$$

where $\mathbf{u}$ is the velocity field; we wrote (26) for an incompressible flow ($\boldsymbol{\nabla} \cdot \boldsymbol{u}$= 0) for simplicity. In the most general formulation of the problem, one also has to consider the Navier-Stokes equation for $\mathbf{u}$ with a term accounting for the feedback of θ on the velocity field. The two points of view can be related through an elegant mathematical formulation in terms of the Feynman-Kac formula [22].

Chemical Processes in Fluid Flows

Among the many chemical reactions that take place in fluids, one of the most important is combustion [14]. The general problem is very difficult due to the presence of many components which react in a complicated way and which modify the flow via heat release, thus enhancing the complexity of the flow generating turbulence. Turbulence itself plays an important role in increasing the mixing of reagents (see Chap. 8) and therefore improving the efficiency of combustion processes [16]. For example, in the spark-ignition engine, fuel and oxidizer are firstly mixed by turbulence before the spark ignites the mixture.

For a complete mathematical formulation of the general problem, one has to consider N reacting species θ_i which evolve according to (26)

$$\frac{\partial \theta_i}{\partial t} + \mathbf{u} \cdot \boldsymbol{\nabla} \theta_i = D_i \nabla^2 \theta_i + F_i(\theta_1, \ldots, \theta_N, T), \tag{27}$$

with their own diffusivity constant, D_i, and reaction kinetics, F_i, that depends on the temperature, T. The temperature itself is transported by the flow and modifies it through buoyancy, so that the Navier-Stokes equation for the velocity field with back-reaction should be considered. Usually the dependence of F_i on the temperature is of the Arrhenius type (17) [15]. It is obvious that this set of equations is too complicated for us to give a satisfactory non-technical treatment of the problem.

However, some new phenomena which arise because of the presence of the advecting velocity field can be appreciated, even considering a single reacting species and neglecting the back-reaction on the fluid, i.e. remaining at the level of the ARD equation (26). For the sake of simplicity, let us consider

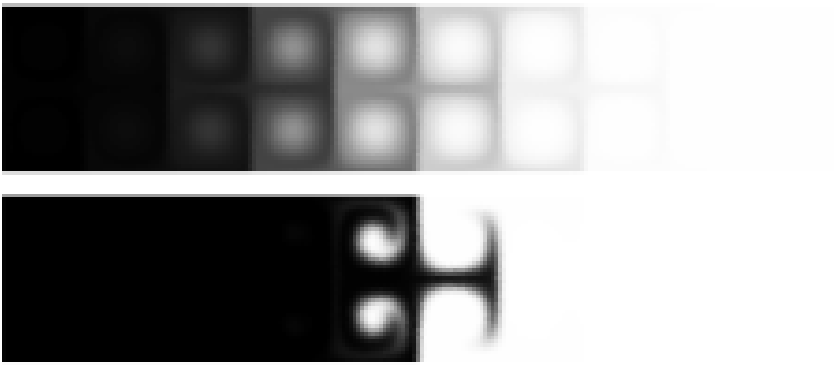

Fig. 8. Front propagation in a pipe-like geometry in two-dimension. The boundary conditions are $\theta(-\infty, y) = 1$ and $\theta(+\infty, y) = 0$. The grey scale indicates a concentration going from fresh material ($\theta = 0$, white) to burnt material ($\theta = 1$, black). The advecting velocity field is a cellular flow, $u_x = U \sin x \cos y$ and $u_y = -U \sin y \cos x$, where U is the stirring intensity. The upper image is for low Da, i.e. slow chemistry; note the thickness of the reaction zone, ξ, which extends over many velocity characteristic length scales, L (here the transversal length). The lower image is for high Da, fast reaction, here $\xi \ll L$

(26) in a pipe-like geometry with a given velocity field (see Fig. 8). For our illustrative purpose, the velocity field is left unspecified. We only assume that it is characterized by its intensity, U (the root mean square velocity) and its typical length scale, L (the correlation length, or in the pipe-problem the transverse length). In addition, let us assume that $F(\theta)$ is of the FKPP-type with $\theta = 0$ and $\theta = 1$ representing, the unstable (unburnt) and stable (burnt) states, respectively. Let us rewrite $F(\theta) = f(\theta)/\tau$ where τ is the typical temporal scale of the chemical kinetics, and f is the rescaled production term with $f'(0) = 1$. Suppose that at the initial time on the left there is burnt material and the rest of the pipe is filled with fresh unburnt reagent. This can be seen as an extremely simplified burning process or chemical reaction. As time advances the front separating burnt from unburnt material will advance from left to right with a speed v_f. Now the question is how the front speed and shape will be modified by the presence of the velocity field, $\mathbf{u}$.

In a fluid at rest, $\mathbf{u} = 0$, we saw that the propagation speed is given by $v_0 = 2\sqrt{D/\tau}$ (9), the laminar front speed in combustion jargon. Moreover, the thickness of the region where the reaction takes place, ξ, is roughly given by $\xi \sim \sqrt{D\tau}$ (see (13)). In a moving fluid it is natural to expect that the front will propagate with an average (turbulent) speed v_f greater than v_0. The turbulent front speed v_f will be the result of the interplay among the flow characteristics, L and U, the diffusivity D, and the chemical time scale τ. The analysis can be simplified by introducing two non-dimensional num-

bers: the *Damköhler number* $Da = L/(U\tau)$, the ratio of advective to reactive time scales, and the *Peclet number* $Pe = UL/D$, the ratio of diffusive to advective time scales. Usually one is interested in the limit of high Pe number, when advection is stronger than diffusion. The front speed is expected to be expressed as $v_f = v_0\phi(Da, Pe)$ which is in general larger than v_0 [22].

The study of the detailed dependence of v_f on Da and Pe, $\phi(Da, Pe)$, is non-trivial. However, some limiting cases can be identified. A particularly simple case is when the reaction is very slow, $Da \ll 1$. In this regime the thickness of the reaction zone is much larger than the typical velocity length scale, $\xi \gg L$ (see Fig. 8). On length scales larger than L the transport properties of an advected scalar (or of particles in the Lagrangian viewpoint) are known to be well described by an effective diffusion constant, D_{eff}, usually much larger than the molecular diffusivity, D (see [23] and references therein). As a consequence, the reaction zone behaves as if the diffusion coefficient is D_{eff}. In other words, on scales much larger than L (26) reduces to (1) with $D \to D_{eff}$. So that the theory discussed in Sect. 2 applies [22,24] with

$$v_f \approx 2\sqrt{D_{eff}/\tau}\,. \tag{28}$$

Apart from slow biological reactions, or when the stirring by the velocity field is extremely intense, most of reactions of interest have time scales comparable or faster than the advection time, L/U (fast reaction). Therefore the previous result cannot be applied. However, it is interesting to note that the rhs of (28) is a rigorous upper bound to the turbulent front speed v_f [22]. A possible approach in the case of fast reactions is to renormalize both the diffusion constant and the chemical time scales. But while the computation of the renormalized diffusion coefficient is based on powerful, well-established mathematical methods [23], the renormalization of τ can only be approached phenomenologically [22].

Another limit is when $D, \tau \to 0$, while remaining the ratio D/τ constant; here, the reaction zone thickness $\xi \sim \sqrt{D\tau}$ shrinks to zero, while the laminar front speed v_0 stays finite). In this case, the ARD equation reduces to the so-called G-equation [16]

$$\frac{\partial G}{\partial t} + \mathbf{u}\cdot\boldsymbol{\nabla} G = v_0|\boldsymbol{\nabla} G|\,. \tag{29}$$

The iso-scalar surface (line in two dimension), say $G = 0$, represents the front position. Equation (29) has a simple geometrical interpretation: in the absence of stirring ($\mathbf{u} = \mathbf{0}$) the front evolves according to the Huygens principle, i.e., a point $\mathbf{x}$ belonging to the front moves with a velocity $\mathbf{v}(\mathbf{x}) = v_0\hat{\mathbf{n}}(\mathbf{x})$, $\hat{\mathbf{n}}(\mathbf{x})$ being the perpendicular direction to the front surface in $\mathbf{x}$. The effect of the velocity field is to wrinkle the front, increasing its area and thereby its speed [16]. Indeed the front speed in this limit is linked to the amount of material which is burnt per unit time, which increases as the front area increases. Assuming a velocity field with a turbulent spectrum, Yakhot [25]

proposed that at large flow intensities ($U \gg v_0$) $V_f \propto U/\sqrt{\ln U}$. We do not know whether this prediction is correct or not, although the fact that v_f has an almost linear behavior with U (here corrected by $\sqrt{\ln U}$) seems to be a generic feature in laboratory and numerical experiments up to moderately high intensities.

Plankton Patchiness

The large importance of plankton distributions in the sea must not be underestimated. They are at the lowest level of the ocean food chain, and among the most important ingredients for understanding the interchange of CO_2 between the atmosphere and the oceans and, consequently, the phenomenon of global warming [26].

A characteristic that is well known since the earliest *in situ* observations, recently verified by satellite remote sensing and detailed numerical simulations [27], is *plankton patchiness*, i.e., the inhomogeneity of plankton spatial distributions. These analyses identify filaments, irregular patches, sharp gradients, and other complex structures involving a wide range of spatial scales in the concentration patterns, which typically extend from medium scales ($\sim$ 10 km) to very large ones ($\sim$ 1000 km), associated with the major ocean currents and gyres (Fig. 9).

Traditionally, patchiness has been variously attributed to the interplay of diffusion and biological growth, oceanic turbulence, diffusive Turing-like instabilities, and nutrient or biological inhomogeneities [28]. Advection by unsteady fluid flows and predator-prey interactions are only recently emerging as two key ingredients able to reproduce the main qualitative features of plankton patchiness [18]. Therefore, the proper mathematical framework for this problem is that of advection reaction diffusion systems with many components (27).

The reaction term usually takes into account three different trophic levels and their interactions: nutrients (N), phytoplankton (P) and zooplankton (Z). The nutrients are inorganic materials dissolved in water that can be assimilated by the phytoplankton organisms; the zooplankton grazes on the latter. The interactions among N, P and Z are schematically sketched in Fig. 10. As one can see they are of the predator-prey type with competition for resources and mortality. Moreover, the uptake of nutrients by phytoplankton, and the grazing of these by the zooplankton are also taken into account. Regarding the advection by oceanic flows, the mechanism which is now emerging as a key feature in explaining the observed filament-like structures of the concentration patterns is *chaotic advection* [29], i.e. the presence of highly chaotic trajectories of the fluid particles even in relatively simple Eulerian flows. In fact the continuous stretching and folding of fluid elements induced by the flow is considered to be one of the basic ingredients for the generation of patchiness, see [27] for a recent review.

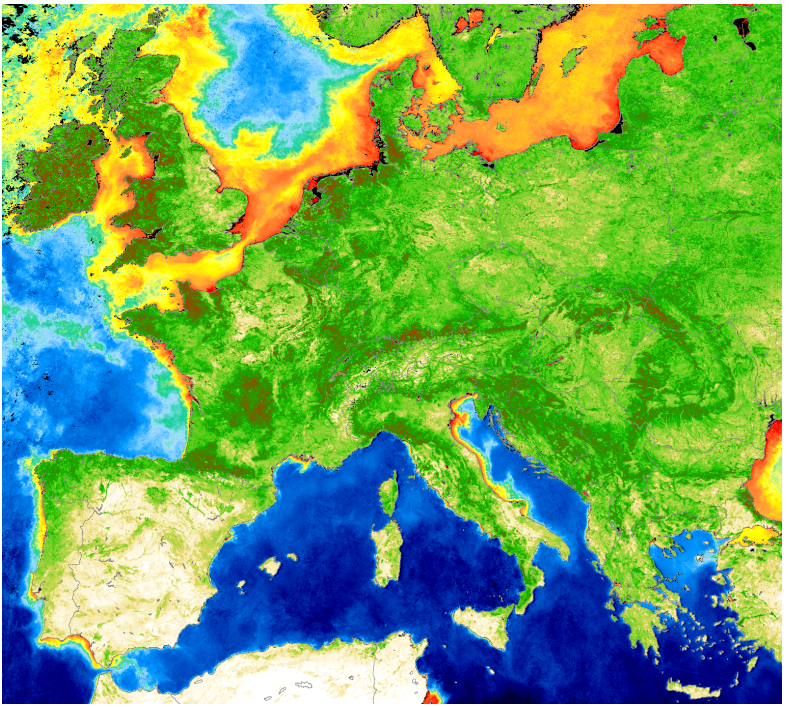

Fig. 9. Satellite image of phytoplankton pigment (chlorophyll) concentration in Western Europe (Courtesy by Marine Environment Unit, image from SeaWiFS Images Archive: http://www.me.sai.jrc.it)

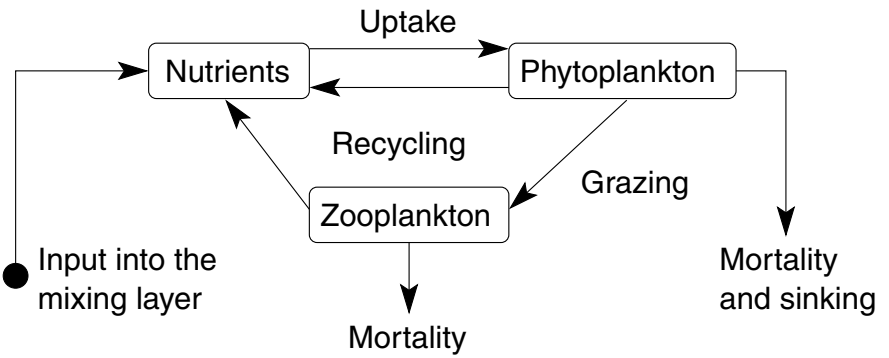

Fig. 10. The processes in the NPZ models

Here we limit the discussion to a simple model [30] constituted by three trophic levels where, instead of considering explicitly the nutrients, we introduce the carrying capacity, C, defined as the maximum phytoplankton content that a parcel of water can support in the absence of grazing. Considering a relaxational dynamics for C, and assuming equal diffusivities (which is reasonable because mixing in plankton communities is largely due to seawater turbulence), the model is

$$\begin{aligned} \frac{\partial C}{\partial t} + \mathbf{u}\cdot\nabla C &= \alpha(C - C_0(\mathbf{x})) + D\nabla^2 C\,, \\ \frac{\partial P}{\partial t} + \mathbf{u}\cdot\nabla P &= P(1 - P/C) + D\nabla^2 P\,, \\ \frac{\partial Z}{\partial t} + \mathbf{u}\cdot\nabla Z &= PZ - \delta Z^2 + D\nabla^2 Z\,, \end{aligned} \tag{30}$$

where α describes the relaxation of C onto an imposed spatially dependent carrying capacity, $C_0(\mathbf{x})$ (see [30] for more details), and δ is the Z mortality. The velocity field $\mathbf{u}(\mathbf{x}, t)$ is incompressible and it is assumed to give rise to chaotic advection, which implies that the separation between two fluid particles, $|\delta\mathbf{x}(t)|$, initially close ($|\delta\mathbf{x}(0)| \ll 1$) typically diverges in time at a rate given by the Lyapunov exponent of the flow $\lambda_F > 0$,

$$|\delta\mathbf{x}(t)| \propto |\delta\mathbf{x}(0)| e^{\lambda_F t}\,. \tag{31}$$

In the absence of the flow and with $D = 0$, the dynamics is attracted by the stable fixed point of (30): $C^* = C_0(\mathbf{x})$, $P^* = C_0\delta/(\delta + C_0)$, and $Z^* = P^*/\delta$. Thus the *chemical* Lyapunov exponent[5] λ_C is negative. This simple model displays an interesting transition depending on the value of λ_F and λ_C. If $|\lambda_C| > \lambda_F$ the plankton distribution is smooth, while if $|\lambda_C| < \lambda_F$, i.e. when the flow is enough chaotic to overcome the stability of plankton dynamics, the asymptotic spatial distribution of plankton has fractal properties.

Another remarkable feature of this model is its ability to reproduce a well-known experimental fact related to the behavior of the power spectrum, $\Gamma(k)$ (k is the wavenumber), of the species distributions. Specifically, analysis of transects taken by oceanographic ships have shown that the power spectra of zooplankton and phytoplankton have a power law behavior characterized by different scaling exponents [18]: $\Gamma_P(k) \propto k^{-\beta_P}$ and $\Gamma_Z(k) \propto k^{-\beta_Z}$, with $\beta_P \neq \beta_Z$, indicating the different distributions of P and Z. Furthermore, β_P and β_Z seem to be different from 1, the scaling exponent expected for passive scalars (such as temperature and salinity). Therefore, P, Z and the temperature field are distributed in a very different manner, Z being much more irregularly distributed than P, i.e., $1 < \beta_P < \beta_Z$ [18,30]. In the model (30) the power spectrum scaling exponents can be computed in terms of

[5] That is the Lyapunov exponent of the dynamical systems obtained by (30) with $\mathbf{u} = \mathbf{0}$ and $D = 0$.

the Lyapunov exponents λ_F and λ_C as $\beta_P = \beta_Z = 1 + 2|\lambda_C|/\lambda_F < 1$, which partially reproduces the observations. However, a slightly more sophisticated model has obtained the result $\beta_P < \beta_Z$ (see article on page 470 of [27]), which is closer to observations.

The complete characterization of plankton patchiness requires the introduction of more refined observables than the power spectrum. For instance, one can define the qth structure functions of, say, phytoplankton, as

$$S_q(\delta r) = \langle |P(\mathbf{x} + \delta\mathbf{x}, t) - P(\mathbf{x}, t)|^q \rangle, \tag{32}$$

where the bracket represents averaging over locations $\mathbf{x}$, $\delta r = |\delta\mathbf{x}|$ and q is a positive number. Interestingly, as observed in turbulence and passive scalars (see Chaps. 7, 8), in the limit $\delta r \to 0$, structure functions have a power law behavior given by $S_q \propto \delta r^{\zeta_q}$. Moreover, the exponents ζ_q display a non-trivial dependence on q, namely they deviate from the linear dimensional estimation $\zeta_q = q(1-\beta)/2$ (where β is the power spectrum scaling exponent). These deviations are the signature of the multifractal behavior of plankton distribution. Remarkably, multifractality naturally arises in the framework of models like (30) due to the fluctuations of finite-time Lyapunov exponents (see [30] for a detailed discussion on this point).

The most important lesson one can learn from this simple model is that from the interplay of a smooth flow, which accounts for the *physics*, and a (simplified) stable interacting dynamics, the biology, one can have a very irregular (multifractal) spatial distribution of population concentrations. Moreover, the relevant quantities describing the inhomogeneities of these distributions, such as the power spectrum or structure function scaling exponents, can be expressed in terms of the Lyapunov exponents that characterize, separately, the dynamics of the flow and of the plankton populations.

References

1. A.N. Kolmogorov, I. Petrovskii and N. Piskunov, "A study of the diffusion equation with increase in the amount of substance and its application to a biology problem", in *Selected works of A.N. Kolmogorov*, ed. V.M. Tikhomirov, Vol. I, p. 242, Kluwer Academic Publishers, London (1991); original work *Bull. Univ. Moscow, Ser. Int.* A, **1**, 1 (1937)
2. R.A. Fisher, "The wave of advance of advantageous genes", *Ann. Eugenics*, **7**, 353 (1937)
3. J.D. Murray, *Mathematical biology*, Springer-Verlag, Berlin (1993)
4. Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence*, Springer-Verlag, Berlin (1984)
5. M.C. Cross and P.C. Hohenberg, "Pattern formation outside equilibrium", *Rev. Mod. Phys.*, **65**, 851 (1993)
6. A. Okubo and S.A. Levin, *Diffusion and ecological problems*, Springer-Verlag, Berlin (2001)
7. D.G. Aronson and H.F. Weinberger, "Multidimensional nonlinear diffusion arising in population genetics", *Adv. Math.*, **30**, 33 (1978)

8. E. Ben-Jacob, H.R. Brand, G. Dee, L. Kramer, and J.S. Langer, "Pattern propagation in nonlinear dissipative systems", *Physica* D, **14**, 348 (1985)
9. W. van Saarloos, "Front propagation into unstable states: Marginal stability as a dynamical mechanism for velocity selection", *Phys. Rev.* A, **37**, 211 (1988).; W. van Saarloos, "Front propagation into unstable states. II. Linear versus nonlinear marginal stability and rate of convergence", *Phys. Rev.* A, **39**, 6367 (1989)
10. U. Ebert and W. van Saarloos, "Front propagation into unstable states: universal algebraic convergence towards uniformly translating pulled fronts", *Physica* D, **146**, 1 (2000)
11. J. Xin, "Front propagation in heterogeneous media", *SIAM Review*, **42**, 161 (2000)
12. S.K. Scott,*Oscillations, Waves and Chaos in Chemical Kinetics*, Oxford University Press, (1994)
13. J. Ross, S.C. Müller, and C. Vidal, "Chemical waves", *Science*, **240**, 460 (1988)
14. F.A. Williams, *Combustion Theory*, Benjamin-Cummings, Menlo Park (1985)
15. Y.B. Zel'dovich and D.A. Frank-Kamenetskii, "A Theory of thermal propagation of flame", *Acta Physicochimica U.R.S.S.*, Vol. XVII, **1–2**, 42 (1938)
16. N. Peters, *Turbulent combustion*, Cambridge University Press, Cambridge (2000)
17. A.J. Ammerman and L.L. Cavalli-Sforza, *The neolithic transition and the genetics of population in Europe*, Princeton University Press, Princeton (1984)
18. E.R. Abraham, "The generation of plankton patchiness by turbulent stirring", *Nature* **391**, 577 (1998)
19. A.M. Turing, "The chemical basis of morphogenesis", *Phil. Trans. R. Soc. London* B, **237**, 37 (1952)
20. M. Giudici, C. Green, G. Giacomelli, U. Nespolo, and J. Tredicce, "Andronov bifurcation and excitability in semiconductor lasers with optical feedback", *Phys. Rev. E*, **55**, 6414 (1997)
21. R. Fitzhugh, "Impulses and physiological states in theoretical models of nerve membranes", *Biophys. J.*, **1**, 445 (1961). J. Nagumo, S. Arimoto, and S. Yoshizawa, "An active pulse transmission line simulating 1214-nerve axons", *Proc. IRL*, **50**, 2061 (1960)
22. M. Abel, A. Celani, D. Vergni, and A. Vulpiani, "Front propagation in laminar flows", *Phys. Rev.* E, **64**, 046307 (2001)
23. A.J. Majda, and P. R. Kramer, "Simplified models for turbulent diffusion: Theory, numerical modeling and physical phenomena", *Phys. Rep.*, **314**, 237 (1999).
24. P. Constantin, A. Kiselev, A. Oberman, and L. Ryzhik, "Bulk Burning Rate in Passive - Reactive Diffusion", *Arch. Rational Mechanics*, **154**, 53 (2000)
25. V. Yakhot, "Propagation velocity of premixed turbulent flame", *Combust. Sci. Technol.*, **60**, 191 (1988)
26. P.M. Cox, R.A. Betts, C.D. Jones, S.A. Spall, and I.J. Totterdell, "Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model", *Nature*, **408**, 184 (2000)
27. Focus Issue on Active Chaotic Flow, *Chaos*, **12**, 372 (2002), editors Z. Toroczkai and T. Tél
28. D.L. Mackas, K.L. Denman, and M.R. Abbott, "Plankton patchiness: biology in the physical vernacular", *Bull. Mar. Sci.*, **37**, 652 (1985)

29. J.M. Ottino *The Kinematics of Mixing: Stretching, Chaos and Transport*, Cambridge University Press, Cambridge (1989)
30. C.López, Z. Neufeld, E. Hernández-García, and P.H. Haynes, "Chaotic advection of reacting substances: Plankton dynamics on a meandering jet", *Phys. Chem. Earth*, **B 26**, 313 (2001)

Self-Similar Random Fields: From Kolmogorov to Renormalization Group

Giovanni Jona-Lasinio

Dipartimento di Fisica, Universitá "La Sapienza" and INFN, Piazza A. Moro 2, Roma, Italy, 00185, Gianni.Jona@roma1.infn.it

Abstract. The first example of a self-similar random process, the Wiener process, appeared long ago. An abstract geometrical approach to such processes was initiated by Kolmogorov and pursued by the school of Russian probabilists. An obvious generalization, that is processes dependent on a multidimensional argument (usually called random fields) were introduced by Kolmogorov in his phenomenological theory of turbulence, a main contribution to physics.

Self-similar random fields reappeared independently thirty years ago in an entirely different context: the area of critical phenomena in phase transitions, with interesting developments. The new approach to the subject came through the combination of the physicist renormalization group with an extension of the concept of stable distribution, a class of distributions that plays an important role in limit theorems for independent random variables.

The present paper is a short guide through these topics.

1 Historical Sketch

The idea of self-similarity of physical phenomena plays a fundamental role in fluid mechanics: for example the stationary plane motions in a viscous incompressible fluid of geometrically similar objects, that is having the same shape, but different sizes, depend only on two dimensionless quantities, the angle formed by the object with the direction of motion, and the Reynolds number. This allows the scientist to reduce the study of the motion of a very large body to experiments in a laboratory by acting on the values of the velocity, the fluid density, the viscosity and the size in such a way as to keep the Reynolds number constant. For a rich exemplification of self-similar situations in fluid dynamics we refer the reader to the interesting book by Sedov [1].

In turbulence, which is part of fluid dynamics, stochasticity appears and statistical averages seem to depend, over certain spatial scales, only on special combinations of the physical quantities. The notion of self-similar random field seems natural in this context and was introduced by Kolmogorov in a rather informal way as a phenomenological tool. However, earlier, he had laid the basis for a mathematical theory.

In two short papers published in the Dok. Akad. Nauk SSSR in 1940 [2,3] Kolmogorov initiated the study of classes of stochastic processes with invariance properties under certain transformations. In particular in [3] he

G. Jona-Lasinio, Self-Similar Random Fields: From Kolmogorov to Renormalization Group, Lect. Notes Phys. **636**, 213–227 (2003)
http://www.springerlink.com/

introduced a class of processes homogeneous under a rescaling of the independent variable. One of them, characterized by a special value of the degree of homogeneity, he called the Wiener spiral. The approach in [2,3] is geometrical in terms of curves in Hilbert space and the probabilistic interpretation is introduced as a special realization. Later it was generalized and developed in a 1955 article by Pinsker [4].

One year later, his famous articles on the theory of turbulence [5–7] appeared where random fields, i.e. stochastic processes depending on a multidimensional space variable, and possessing scaling properties, played a crucial role. A systematic analysis of the processes involved in Kolmogorov theory was developed in 1957 in a paper by A.M. Yaglom [8]. It is worth noting that in his physical theory Kolmogorov did not quote his mathematical papers. May be because in absence of a theory for the origin of stochasticity in developed turbulence he considered these fields only as descriptive concepts without a mathematical foundation on more primitive notions.

In 1962 Lamperti [9], apparently unaware of the work of the Russians, introduced essentially the same processes as [3]. He called these processes semi-stable, with the aim of generalizing the idea of stable distribution, a class of probability distributions which have remarkable scaling properties [10].

Self-similar random fields occupy a central place in the modern theory of the critical point of phase transitions. Here the situation is very different from turbulence. Randomness is introduced at the microscopic level by Gibbs probability distributions on which statistical mechanics is founded. We deal with a well defined problem which consists in understanding how the non scaling invariant original Gibbs distribution becomes self-similar over large distances at the critical point. The qualitative explanation is that at the critical point a thermodynamic system develops statistical correlations of infinite range so that the nature of atomic constituents and the typical interatomic distances become irrelevant. The physicist approach is via the so called renormalization group (RG) which, in one of its formulations, has a probabilistic interpretation in terms of limit theorems for random fields. In perspective one can see in this formulation the melting of different ideas originated earlier in different contexts.

The RG theory of the critical point leads to the study of self-similar random fields which are obtained as limits of stationary random fields. A characteristic feature of the critical point is the non-integrability of the correlation function, a property that probabilistically denotes a strong dependence among the variables.

In the early sixties Rosenblatt [11] constructed a stationary asymptotically self-similar process of discrete argument, which, although not connected with problems in physics, later provided an interesting example in the probabilistic interpretation of the renormalization group approach to phase transitions.

The probabilistic version of the RG was developed independently of Kolmogorov's and Lamperti's work and the connections with its ancestors were

understood afterwards. The first informal definition of translation invariant, i.e. stationary, self similar random field, was given in [12] and formalized in [13]. One year later a similar definition was given independently in [14]. These papers were concerned with systems on a lattice. The extension to the continuous case and a systematic analysis from the standpoint of generalized random fields was developed by Dobrushin [16,17].

In the following we shall briefly review the mathematical concepts involved in the story outlined above emphasizing the shift in point of view which has taken place with the renormalization group.

2 The Wiener Spiral and Related Processes

In [2] probability is not even mentioned: the title of the note is *"Curves in Hilbert Space which are Invariant under a One-parameter Group of Motions"*. In [3] the probabilistic interpretation of the Hilbert space structure is defined via the expectation value, that is, given a random variable ξ, the square of its norm as an element of the Hilbert space $\mathbf{H}$ is $\mathbb{E}(|\xi|^2)$ and the scalar product between two real variables $< \xi, \eta >= \mathbb{E}(\xi\eta)$. The random processes considered ξ_t are curves in $\mathbf{H}$ generated as follows

$$\xi_t = a_t + U_t\xi = K_t\xi \tag{1}$$

where $a_t, \xi \in \mathbf{H}$, U_t is a unitary group and $K_{t+s} = K_tK_s$ which implies $a_{t+s} = a_t + U_ta_s = a_s + U_sa_t$ with $a_0 = 0$. Clearly if we take $a_t = 0$ the processes generated are stationary in the wide sense while for $a_t \neq 0$ their increments $\xi_{t+h} - \xi_t$ have this property. A process is stationary in the wide sense if the expectations $\mathbb{E}(\xi_t)$, $\mathbb{E}(|\xi_t|^2)$ do not depend on t and $\mathbb{E}(\xi_t\xi_s)$ depends only on $t - s$. It is a simple calculation to verify that $\mathbb{E}([\xi_{t_1+h_1} - \xi_{t_1}][\xi_{t_2+h_2} - \xi_{t_2}])$ depends only on $t_2 - t_1$ for any fixed h_1, h_2.

A similarity transformation in $\mathbf{H}$ is defined by the operator

$$S\xi = \bar{a} + q\bar{U}\xi \tag{2}$$

where $\bar{a}$ is again an element of $\mathbf{H}$, q a real number and $\bar{U}$ a unitary transformation.

The class of self-similar random processes is defined by the property

$$\xi_{\lambda t} = S_\lambda\xi_t \tag{3}$$

for all t, where S_λ is a similarity transformation.

A first theorem proved by Kolmogorov shows that the expectation of the product of two increments at the same time t of a self-similar process has the form

$$\mathbb{E}([\xi_{t+h_1} - \xi_t][\xi_{t+h_2} - \xi_t]) = c[|h_1|^\alpha + |h_2|^\alpha - |h_1 - h_2|^\alpha] \tag{4}$$

where c is a positive constant and $0 \leq \alpha \leq 2$. The scaling factor q_λ in S_λ is therefore $\lambda^{\alpha/2}$. By definition the Wiener spiral corresponds to $\alpha = 1$. This terminology stems from the following definition: a curve $x(t)$ in a vector space is a spiral if the distance $||x(t) - x(s)||$ depends only on $|t - s|$.

A second theorem proves that a necessary and sufficient condition for a self-similar process to be a Wiener spiral is the vanishing of the correlation between two increments referring to non overlapping intervals of time.

The Wiener process is a special case of the Wiener spiral. About spirals and geometry of the Wiener process see also [18].

In [9] Lamperti considers the following situation: given a process η_t suppose that the limit exists in the sense of convergence of finite dimensional joint distributions

$$\lim_{\lambda\to\infty} \frac{\eta_{\lambda t} + g(\lambda)}{f(\lambda)} = \xi_t \tag{5}$$

for an appropriate choice of $g(\lambda)$ and $f(\lambda)$. Then he proves that ξ_t is a self-similar process with $f(\lambda) = \lambda^{\alpha/2}L(\lambda)$ and $g(\lambda) = \omega(\lambda)\lambda^{\alpha/2}L(\lambda)$. The function $\omega(\lambda)$ has a limit when λ tends to infinity and $L(\lambda)$ is slowly varying.

An important example of this kind is provided by the Donsker invariance principle [19] which states the following: consider an infinite sequence of independent Gaussian variables ξ_i of unit variance and form the sums $S_k = \sum_1^k \xi_i$. For each integer n construct the process

$$X_t^n = \frac{1}{\sqrt{n}}(S_{[nt]} + (nt - [nt])\xi_{[nt]+1}) \tag{6}$$

where $[nt]$ is the integer part. Then X_t^n converges in distribution as $n \to \infty$ to the Wiener process.

This type of limit theorems for processes is akin, but much simpler, to the situations we shall consider later in connection with the renormalization group.

The objects appearing in Kolmogorov theory of turbulence are random fields. The term random field is used when a process depends on a multidimensional variable like the coordinates of a point in space or space-time. A homogeneous field is the analog of a stationary process: this means that the joint probability distributions are invariant under translations of the arguments. A locally homogeneous field is the analog of a process with stationary increments. The increments are defined by $\xi(\underline{x} + \underline{h}) - \xi(\underline{x})$. In [5] the locally homogeneous random field is the velocity field but the quantities of interest are the increments which in certain regions of their arguments are assumed to satisfy self-similarity conditions. It is easy to show that self-similar stationary processes or fields cannot exist as ordinary functions and must be interpreted as distributions, that is linear functionals on an appropriate space of test functions. The argument is as follows. The probability distributions of ξ_t and $\lambda^{-\alpha/2}\xi_{\lambda t}$ cannot coincide unless $\xi_t = \xi_0 = 0$. In fact stationarity implies that we can replace $\xi_{\lambda t}$ with ξ_0 from which the statement is obvious.

This difficulty does not arise if we consider an ordinary process as a generalised process and choose properly the test functions. If ξ_t is viewed as a generalised process only expressions of the form

$$\xi(\phi) = \int \xi_t \phi(t) dt \tag{7}$$

are considered where we take ϕ of compact support. Then by translational invariance $\int \lambda^{-\alpha/2} \xi_{\lambda t} \phi(t) dt$ is equivalent to $\int \lambda^{-\alpha/2} \xi_0 \phi(t) dt$ which is zero if $\int \phi(t) dt = 0$. Therefore self-similarity can be implemented in a space of test functions of vanishing integral. To see the effect of smearing a non stationary self similar process let us consider the case of the Wiener process w_t. Our test functions are of compact support on $[0, \infty)$ with vanishing integral. Let us consider the correlation function

$$\mathbb{E}(w(\phi)w(\psi)) = \int_0^\infty dt \int_0^\infty dt' \min(t, t') \phi(t) \psi(t') \tag{8}$$

A simple calculation shows that this correlation function is invariant under forward time translations. Therefore the process $w(\phi)$, being gaussian and therefore determined by its correlation function, is invariant under these translations.

The self-similar random fields considered in turbulence can be viewed as generalised processes in a space of test functions with vanishing integral [8].

3 The Renormalization Group: General Ideas

Statistical mechanics describes macroscopic systems in terms of an underlying microscopic structure whose configurations are the arguments of a probability distribution, an ensemble in the terminology of physicists. Therefore statistical mechanics deals with random fields depending on a discrete argument for systems on a lattice, or a continuous argument, e.g. in the case of a classical gas.

When a system approaches a critical point large islands of a new phase appear so that correlations among the microscopic constituents extend over macroscopic distances. One characterizes this situations by introducing a correlation length which measures the extension of such correlations. At the critical point this length becomes infinite and typically correlations decay with a non-integrable power law as opposed to an exponential decrease away from criticality. The exponents in these power laws exhibit a remarkable degree of universality because the same exponents appear for physically different systems such as a gas and a ferromagnet.

The renormalization group (RG) is both a way of thinking about critical phenomena and a calculational tool. It was introduced to understand the two aspects mentioned above, i.e. the appearance of power laws and the universality of the exponents. There exist two rather different versions of RG both

originated in quantum field theory: one, called the multiplicative RG or the Green's function RG, is an exact generalized scaling relation satisfied by the Green's functions in quantum field theory and by the correlation functions in statistical mechanics; the second, called Wilson's or sometimes Kadanoff-Wilson's RG [20], is based on a progressive elimination of the degrees of freedom of the system and properly speaking is a semigroup. The multiplicative RG was the first to be applied in the theory of the critical point [21] but it is the second one which has an interesting probabilistic interpretation. However the two are structurally connected as we shall illustrate later.

The qualitative argument mentioned in Sect. 1 to figure out why scaling properties should be expected at the critical point was put forward by Kadanoff [22]. If correlations extend over macroscopic distances it must be irrelevant whether we consider our system constituted by the original microscopic objects or by blocks containing a large number of constituents. In the limit when the correlations extend to infinity the size of the blocks should not matter and this leads immediately to homogeneity properties for the correlation functions and the thermodynamic potentials. A mathematical implementation of the idea came later with the application of Wilson's renormalization transformation to criticality. Technically Wilson's transformation does not involve blocks and is defined in Fourier space but conceptually represents a way of realizing Kadanoff's point of view. The probabilistic version that we shall discuss on the other hand can be considered as the mathematically most faithful implementation of Kadanoff's idea.

Forming blocks of stochastic variables is common practice in probability, the central limit theorem (CLT) being the prototype of such a way of reasoning. The crucial point is that when we sum many random variables we have to normalize properly the sum in order to obtain a regular probability distribution. In the case of the CLT, the correct normalization is the square root of the number of variables. When we deal with processes which have long range correlations and we substitute the original variables with sums, in addition to the normalization it is necessary to redefine the unit of distance in order to get in the limit of infinite blocks a reasonable stochastic process. In summary the probabilistic RG consists of three steps: forming blocks, normalizing them, redefining the unit of space distances. The infinite iteration of this procedure, which is the same as taking the limit of infinite blocks, will provide, if convergent, a self-similar random field.

Universality of critical exponents has a very natural interpretation. In analogy to the case of the CLT there will be different random fields that under the RG will converge to the same self-similar field. A physical universality class will correspond to a subset of the domain of attraction of such a field. In general we expect to be a subset because not all fields in the domain of attraction will admit a physical interpretation.

4 The Renormalization Group: A Probabilistic View

The theory of generalized random fields was initiated by Gelfand and a systematic exposition is contained in[23]. A generalized random field $\xi(\phi)$ is a linear continuous functional over a space of test functions defined on a euclidean space of d dimensions. A probability measure P can be specified by a collection of one-dimensional distributions $P \equiv \{P_\phi\}$ defined by

$$P_\phi(B) = P(\xi(\phi) \in B) \tag{9}$$

where B is a Borel set on the real line.

We define the scale transformation on test functions

$$(S_{\alpha,\lambda}\phi)(x) = \lambda^{-d\alpha/2-d}\phi(\lambda^{-1}x) \tag{10}$$

with $0 \leq \alpha \leq 2$. $S_{\alpha,\lambda}$ induces a transformation on probability measures according to

$$(S^*_{\alpha,\lambda}P)_\phi = P_{S_{\alpha,\lambda}\phi} \tag{11}$$

A random field is called self-similar if

$$S^*_{\alpha,\lambda}P = P \tag{12}$$

The transformation $S^*_{\alpha,\lambda}$ is also called a continuous renormalization transformation. It is clear from (10) and (11) that for increasing λ we are integrating the field over increasingly larger scales and blocks are defined by the choice of the test functions. The transformation *renormalizes*, i.e. rescales, the field by the factor $\lambda^{-d\alpha/2}$ and the effective unit of distance is fixed by λ. These are the three steps discussed above. The critical point of a phase transition requires $\alpha > 1$ meaning that the central limit theorem fails. A parallel situation is encountered in euclidean quantum field theory where the interesting limit is λ small. This is the small scale or ultraviolet limit in Fourier space.

The randon fields appearing in the theory of the critical point or in quantum field theory are not self-similar over all scales but only asymptotically at large scales in the first case or at small scales in the second. As we remarked previously, a new interesting chapter in the theory of limit theorems was opened in this way. The limit theorems involved refer to variables characterized by a strong dependence and explore a domain complementary to the central limit theorem.

The notion of self similar random field of discrete argument was introduced in statistical mechanical models to provide a proper mathematical setting for the notion of RG *a la* Kadanoff-Wilson (block-spin transformation).

Let $\mathbf{Z}^d$ be a lattice in d-dimensional space and j a generic point of $\mathbf{Z}^d$, $j = (j_1, j_2, ..., j_d)$ with integer coordinates j_i. We associate to each site a

centered random variable ξ_j and define a new random field on a rescaled lattice

$$\xi_j^n = (\mathcal{R}_{\alpha,n}\xi)_j = n^{-d\alpha/2} \sum_{s \in V_j^n} \xi_s, \tag{13}$$

where

$$V_j^n = \{s : j_k n - n/2 < s_k \leq j_k n + n/2\} \tag{14}$$

and $1 \leq \alpha < 2$. The transformation (13) induces a transformation on the corresponding probability measure according to

$$(\mathcal{R}_{\alpha,n}^* P)(A) = P'(A) = P(\mathcal{R}_{\alpha,n}^{-1} A), \tag{15}$$

where A is a measurable set and $\mathcal{R}_{\alpha,n}^*$ has the semi-group property

$$\mathcal{R}_{\alpha,n_1}^* \mathcal{R}_{\alpha,n_2}^* = \mathcal{R}_{\alpha,n_1+n_2}^*. \tag{16}$$

A measure P is called self-similar if

$$\mathcal{R}_{\alpha,n}^* P = P \tag{17}$$

and the corresponding field is called a self-similar random field. Let us briefly discuss the choice of the parameter α. It is natural to take $1 \leq \alpha < 2$. In fact $\alpha = 2$ corresponds to the law of large numbers so that the block variable (13) will tend for large n to zero in probability. The case $\alpha > 1$ means that we are considering random systems which fluctuate more than a collection of independent variables and $\alpha = 1$ corresponds to the CLT. Mathematically the lower bound is not natural but it becomes so when we restrict ourselves to the consideration of ferromagnetic-like systems.

A general theory of self similar random fields does not exist yet and presumably is very difficult. However Gaussian fields are completely specified by their correlation function and self similar Gaussian fields can be constructed explicitly [14,15]. It is expedient to represent the correlation function in terms of its Fourier transform

$$\mathbb{E}(\xi_i \xi_j) = \int_0^1 \prod_1^d d\lambda_k \rho(\lambda_1, \ldots, \lambda_d) e^{2\pi i \sum_k \lambda_k (i-j)_k}. \tag{18}$$

The prescription to construct ρ in such a way that the corresponding Gaussian field satisfies (17) is as follows. Take a positive homogeneous function $f(\lambda_1, ..., \lambda_d)$ with homogeneity exponent $d(1+\alpha)$, that is

$$f(c\lambda_1, ..., c\lambda_d) = c^{d(1+\alpha)} f(\lambda_1, ..., \lambda_d) \tag{19}$$

Next we construct a periodic function $g(\lambda_1, ..., \lambda_d)$ by taking an average over the lattice Z^d

$$g(\lambda_1, ..., \lambda_d) = \sum_{i_k} \frac{1}{f(\lambda_1 + i_1, \ldots, \lambda_d + i_d)}. \tag{20}$$

If we define

$$\rho(\lambda_1, \dots, \lambda_d) = \prod_k |1 - e^{2\pi i \lambda_k}|^2 g(\lambda_1, \dots, \lambda_d), \tag{21}$$

it is not difficult to see that the corresponding Gaussian measure satisfies (17). The periodicity of ρ insures translational invariance.

For $d = 1$ there is only one, apart from a multiplicative constant, homogeneous function and one can show that the above construction exhausts all possible Gaussian self similar distributions. For $d > 1$ it is not known whether a similar conclusion holds.

The search of non-gaussian self-similar fields is considerably more difficult. A reasonable question is whether such fields exist in the neighborhood of a gaussian one. An approach to this problem has been developed by physicists, the so called ε expansion, and in our context can be interpreted as follows.

Consider a small deformation $P_G(1+h)$ of a Gaussian self-similar measure P_G and apply to it $\mathcal{R}^*_{\alpha,n}$. It is easily seen that

$$\mathcal{R}^*_{\alpha,n} P_G h = \mathbb{E}(h|\{\xi^n_j\})\mathcal{R}^*_{\alpha,n} P_G = \mathbb{E}(h|\{\xi^n_j\})P_G(\{\xi^n_j\}). \tag{22}$$

The conditional expectation on the right hand side of (22) will be called the linearization of the RG at P_G and we want to study its stability as a linear operator. This means that we want to understand the properties of the transformation $\mathcal{R}^*_{\alpha,n}$ near its Gaussian fixed point. For this purpose we have to find the eigenvectors and eigenvalues of $\mathbb{E}(h|\{\xi^n_j\}$. These have been calculated by Sinai. The eigenvectors are appropriate infinite dimensional generalizations of Hermite polynomials H_k which are described in full detail in [15]. They satisfy the eigenvalue equation

$$\mathbb{E}(H_k|\{\xi^n_j\}) = n^{[k(\alpha/2-1)+1]d} H_k(\{\xi^n_j\}). \tag{23}$$

We see immediately that H_2 is always unstable while H_4 becomes unstable when α crosses from below the value $3/2$. Bifurcation theory predicts, generically, an exchange of stability between two fixed points, so we should look for the new one in the direction which has just become unstable. By introducing the parameter $\varepsilon = \alpha - 3/2$, one can construct a non Gaussian fixed point using ε as a perturbation parameter. The formal construction is explained in Sinai's book [15] where one can also find a discussion of questions, mostly still unsolved, arising in connection to this problem. Different rigorous constructions of non Gaussian fixed points, for $d = 4$ and $d = 3$ have been made recently by Brydges, Dimock and Hurd [24] and by Brydges, Mitter and Scoppola [25].

There is a simple relationship between the renormalization group in the discrete and in the continuous case. The idea is roughly the following. We say that a space of test functions can be discretized if it contains the indicator functions $\chi_{b,c}$ of the sets

$$V_{b,c} = \{x; b_k < x_k \leq c_k\} \tag{24}$$

We define the discretization of the random field $\xi(\phi)$ as

$$\xi_j = \xi(\chi_j) \tag{25}$$

where χ_j is the indicator function of

$$V_j = \{x; j_k - 1/2 < x_k \leq j_k + 1/2\} \tag{26}$$

Conversely every discrete field can be thought as the discretization of some continuous field. Indeed

$$\xi(\phi) = \sum_j \xi_j \int_{V_j} \phi(x) dx \tag{27}$$

defines a continuous random field whose discretization is ξ_j. It can be shown that if two distributions are related in the continuous case by

$$P^2 = S^*_{\alpha,n} P^1 \tag{28}$$

their discretizations, i.e. the distributions of the discretized fields, that we denote $\tilde{P}^1$ and $\tilde{P}^2$, are related by

$$\tilde{P^2} = \mathcal{R}^*_{\alpha,n} \tilde{P^1} \tag{29}$$

For more details we refer the mathematically inclined reader to [16].

5 A Property of Critical Self-Similar Random Fields

We have already characterized the critical point as a situation of strongly dependent random variables due to the non integrability of the correlation function and, as a consequence, of the failure of the CLT. We want to give here a characterization which refers to the random field globally. This is based on the concept of *strong mixing* introduced by Rosenblatt [11]. Consider the cylinder sets in the product space of the variables ξ_i, that is the sets of the form

$$\{\xi_{i_1} \in A_1, \ldots, \xi_{i_n} \in A_n\}, \tag{30}$$

with $i_1, \ldots, i_n \in \Lambda$, Λ being an arbitrary finite region in $\mathbf{Z}^d$ and with A_i measurable sets in the space of the variables ξ_i. We denote with Σ_Λ the σ-algebra generated by such sets. We say that the variables ξ_i are *weakly dependent* or that they are a *strong mixing* random field if the following holds. Given two finite regions Λ_1 and Λ_2 separated by a distance

$$d(\Lambda_1, \Lambda_2) = \min_{i \in \Lambda_1, j \in \Lambda_2} |i - j|, \tag{31}$$

where $|i-j|$ is for example the Euclidean distance, define

$$\tau(\Lambda_1, \Lambda_2) = \sup_{A \in \Sigma_{\Lambda_1}, B \in \Sigma_{\Lambda_2}} |\mu(A \cap B) - \mu(A)\mu(B)|. \tag{32}$$

Then $\tau(\Lambda_1, \Lambda_2) \to 0$ when $d(\Lambda_1, \Lambda_2) \to \infty$.

Intuitively the strong mixing idea is that one cannot compensate for the weakening of the dependence of the variables due to an increase of their space distance, by increasing the size of the sets.

This situation is typical when one has exponential decay of correlations. This has been proved for a wide class of random fields such as ferromagnetic non critical spin systems [26].

The situation is entirely different at the critical point where one expects the correlations to decay as an inverse power of the distance. In this connection the following result has been proved in [27]: a ferromagnetic translational invariant system with pair interactions and with correlation function

$$C(i) = \mathbb{E}(\xi_0 \xi_i) - \mathbb{E}(\xi_0)\mathbb{E}(\xi_i) \tag{33}$$

such that

$$\lim_{L \to \infty} \frac{\sum_{L(s_k-1) \le i_k < L(s_k+1)} C(i)}{\sum_{0 \le i_k < L} C(i)} \neq 0 \tag{34}$$

for arbitrary s_k, does not satisfy the strong mixing condition.

This theorem implies in particular that a critical 2-dimensional Ising model violates strong mixing. Therefore violation of strong mixing seems to provide a reasonable characterization of the type of strong dependence encountered in critical phenomena. On the other hand, under very general conditions, the one-block distribution satisfies the CLT [31] if strong mixing holds.

The first example of a non Gaussian self-similar process violating strong mixing was constructed by Rosenblatt in [11]: we shall describe it in the last section in connection with limit theorems and the failure of the CLT.

6 Multiplicative Structure

In this section we show that there is a natural multiplicative structure associated with transformations on probability distributions like those induced by the RG. This multiplicative structure is related to the properties of conditional expectations [28]. Suppose we wish to evaluate the conditional expectation

$$\mathbb{E}(h|\{\xi_j^n\}), \tag{35}$$

where the collection of block variables ξ_j^n indexed by j is given a fixed value. Here h is a function of the individual spins ξ_i. It is an elementary property of conditional expectations that

$$\mathbb{E}(\mathbb{E}(h|\{\xi_j^n\})|\{\xi_j^{nm}\}) = \mathbb{E}(h|\{\xi_j^{nm}\}). \tag{36}$$

Let P be the probability distribution of the ξ_i and $\mathcal{R}^*_{\alpha,n}P$ the distribution obtained by applying to it the RG transformation, i.e. the distribution of the block variables ξ^n_j. By specifying in (36) the distribution with respect to which expectations are taken we can rewrite it as

$$\mathbb{E}_{\mathcal{R}^*_{\alpha,n}P}(\mathbb{E}_P(h|\{\xi^n_j\})|\{\xi^{nm}_j\}) = \mathbb{E}_P(h|\{\xi^{nm}_j\}). \tag{37}$$

This is the basic equation of this section and we want to work out its consequences. For this purpose we generalize the eigenvalue equation (23) to the case in which the probability distribution is not a fixed point of the RG. In analogy with the theory of dynamical systems we interpret the conditional expectation as a linear transformation from the linear space tangent to P to the linear space tangent to $\mathcal{R}^*_{\alpha,n}P$ and we assume that in each of these spaces there is a basis of vectors H^P_k, $H_k^{\mathcal{R}^*_{\alpha,n}P}$ connected by the following generalized eigenvalue equation [29]

$$\mathbb{E}_P(H^P_k|\{\xi^n_j\}) = \lambda_k(n,P)H_k^{\mathcal{R}^*_{\alpha,n}P}(\{\xi^n_j\}). \tag{38}$$

Equation (37) implies that the λ's must satisfy the relationship

$$\lambda_k(m,\mathcal{R}^*_{\alpha,n}P)\lambda_k(n,P) = \lambda_k(mn,P). \tag{39}$$

From (38) and (39) we find that the λ_k are given by the following expectations

$$\lambda_k(n,P) = \mathbb{E}(\bar{H}_k^{\mathcal{R}^*_{\alpha,n}P}(\{\xi^n_j\})H^P_k(\{\xi_j\})), \tag{40}$$

where $\bar{H}^P_k$ are dual to H^P_k according to the orthogonality relation $\int \bar{H}^P_k H^P_j dP = \delta_{kj}$. The λ_k are therefore special correlation functions.

If P is self-similar (39) implies that the λ's are powers of n. An example is provided by (23). In the theory of the critical point the corresponding eigenvectors are called the scaling fields.

For those familiar with the multiplicative renormalization group of quantum field theory and statistical mechanics we emphasize its structural similarity with (39). The usual Green's function RG is associated to a very simple transformation of the probability distribution such that its form is unchanged and only the values of its parameters are rescaled together with the random variables; no block variable is introduced.

RG equations akin to the Green's function multiplicative group have been used in the study of turbulence in the scaling invariant Kolmogorov regime, however their interpretation from first principles is not as obvious as in critical phenomena [30].

7 Limit Theorems and Universality of Critical Phenomena

The transformations $S^*_{\alpha,\lambda}$ and $\mathcal{R}^*_{\alpha,n}$ can be considered as dynamical systems, with continuous and discrete time respectively, in spaces of probability dis-

tributions. We can therefore take the limits

$$\lim_{\lambda\to\infty} S^*_{\alpha,\lambda} P = P_\infty \tag{41}$$

and

$$\lim_{n\to\infty} \mathcal{R}^*_{\alpha,n} P = P_\infty \tag{42}$$

In both cases the limit distribution is self-similar according to our previous definitions.

In the study of limit theorems the crucial question is to determine the domain of attraction of a limit distribution, in our case the class of distributions for which the above limits exist and is equal to a given P_∞. From the stand point of dynamical systems this is the stable manifold of P_∞. Such a calculation is in general extremely difficult: only the neighborhood of self-similar Gaussian fields has been explored so far. In the example discussed in Sect. 4 the stable manifold near P_G for $\alpha < 3/2$ consists of those probability distributions whose projection on the second generalized Hermite polynomial vanishes.

A more limited problem is whether we can describe the structure of the limit one-block distributions appearing at the critical point beside the Gaussian, that is the distributions of the single variable

$$\xi^n_j = (\mathcal{R}_{\alpha,n}\xi)_j = n^{-d\alpha/2} \sum_{s\in V^n_j} \xi_s, \tag{43}$$

as n tends to ∞. It was shown in [27], building on previous results by Newman, that for ferromagnetic systems the Fourier transform (characteristic function in probabilistic language) of the limit distribution must be of the form

$$\mathbb{E}(e^{it\xi}) = e^{-bt^2} \prod_j (1 - t^2/\alpha_j^2) \tag{44}$$

with $\sum_j 1/\alpha_j^2 < \infty$. In the probabilistic literature these distributions belong to a class called the D-class [32]. The Gaussian is the only infinitely divisible distribution belonging to this class. In the case of independent random variables only infinitely divisible distributions can appear as limit distributions. These are characterized by the possibility of decomposing them as the convolution of an arbitrary number of identical distributions.

Another case in which we are able to determine the structure of the one block distribution is the Rosenblatt process. The construction of this process is as follows. We start from a doubly infinite sequence η_k of independent random variables normally distributed with unit variance. Let

$$\xi_i = \sum_{-\infty}^{-1} |k|^{-a} \eta_{k+i} \tag{45}$$

with a to be chosen. ξ_i is a stationary Gaussian process characterised by a correlation function $\mathbb{E}(\xi_i\xi_j)$ which can be explicitely computed. Asymptotically for large $|i-j|$ we have

$$\mathbb{E}(\xi_i\xi_j) \asymp |i-j|^{1-2a} = |l|^{1-2a} \tag{46}$$

For $1/2 < a < 1$ correlations decay at infinity but they are not integrable, i.e. $\sum_l |l|^{1-2a} = \infty$. The Gaussian process (45) is therefore critical and the variables are strongly dependent according to the definition of Sect. 5. Violation of strong mixing follows from a theorem of Kolmogorov and Rozanov [31,33]. The limit distribution of the block variables ξ_j^n is Gaussian but the normalization $n^{3/2-a}$ is anomalous. Consider now the new process, the Rosenblatt process,

$$\Xi_i = \xi_i^2 - \mathbb{E}(\xi_i^2) \tag{47}$$

It is non Gaussian and it follows from (46) that its correlation function, $2(\mathbb{E}(\xi_i\xi_j))^2$, is non integrable for $a < 3/4$. Therefore for $1/2 < a < 3/4$ the process (43) is also critical. The limit distribution of the one block variable can be computed explicitely and turns out to be non Gaussian and non infinitely divisible [11,31].

The principal merit of the probabilistic view of RG in my opinion resides in unveiling the statistical nature of universality in critical phenomena. It is also interesting that a new class of limit theorems has arisen from a very important area in physics. The mathematics involved at this stage of development is difficult and one should look for simplifications.

8 Conclusion

We conclude with the following natural question: do the self-similar fields appearing in the theory of turbulence admit a deeper interpretation in terms of an underlying more fundamental probability distribution? During the last twenty five years there has been a considerable amount of work on the possible structure of the probability distributions relevant for the description of turbulence. The input has come from the study of chaotic dynamical systems and their invariant measures [34] but the problem is far from being solved. An answer to the above question therefore is not available: it should however provide an insight on the nature of the famous Kolmogorov exponents. For a systematic dynamical systems approach to turbulence we refer the reader to the book by Bohr, Jensen, Paladin and Vulpiani [35].

Finally some suggestions for further reading. Self-similar random fields, often in connection with fractals, in the last decades have become pervasive of many areas in the natural sciences. On these aspects we recommend the book by Mandelbrot [36] and the collection of essays [37]. More recently self-similar processes have found applications outside the natural sciences in particular in finance for which we refer the reader to Mandelbrot [38] and to the article by Bouchaud and Muzy in this volume.

References

1. L. Sedov, *"Similitude et Dimensions en Mechanique"* Editions Mir, Moscou, 1977
2. A.N. Kolmogorov, Dokl. Akad. Nauk. SSSR **26**, 6 (1940)
3. A.N. Kolmogorov, Dokl. Akad. Nauk. SSSR **26**, 115 (1940)
4. M.S. Pinsker, Izv. Akad. Nauk. SSSR **19**, 319 (1955)
5. A.N. Kolmogorov, Dokl. Akad. Nauk. SSSR **30**, 299 (1941)
6. A.N. Kolmogorov, Dokl. Akad. Nauk. SSSR **31**, 538 (1941)
7. A.N. Kolmogorov, Dokl. Akad. Nauk. SSSR **32**, 19 (1941)
8. A.M. Yaglom, Theory of Prob. and Appl. **2**, 273 (1957)
9. J. Lamperti, Trans. Am. Math. Soc. **104**, 62 (1962)
10. B. Gnedenko, A.N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Cambridge 42, Mass. (1954)
11. M. Rosenblatt, Proc. 4th Berkeley Symp. Math. Stat. and Prob., 431 (1961)
12. G. Jona-Lasinio, Nuovo Cim. B **26**, 9 (1975)
13. G. Gallavotti, G. Jona-Lasinio, Comm. Math. Phys. **41**, 301 (1975)
14. Y.G. Sinai, Theory of Prob. and Appl. **21**, 273 (1976)
15. Y.G. Sinai, *Theory of Phase Transitions: Rigorous Results*, Pergamon Press, London (1982)
16. R.L. Dobrushin, in *Multicomponent Random Systems*, Dekker, N.Y. (1978)
17. R.L. Dobrushin, Ann. of Prob. **7**, 1 (1979)
18. J.P. Kahane, *Some Random Series of Functions*, Cambridge U.P. (1985)
19. M. Donsker, Mem. Am. Math. Soc. **6**, (1951)
20. K. Wilson, J. Kogut, Phys. Rep. **12C**, 75 (1974)
21. C. Di Castro, G. Jona-Lasinio, Phys. Letts. **29 A**, 322 (1969)
22. L.P. Kadanoff, Physics (N.Y.) **2**, 263 (1966)
23. I. M. Gelfand, N. Ya. Vilenkin, *Theory of Distributions vol. IV*, Academic Press (1964)
24. D. Brydges, J. Dimock, T. R. Hurd, Comm. Math Phys. **198**, 111 (1998)
25. D. Brydges, P.K. Mitter, B. Scoppola, *Critical* $(\Phi^4)_{3,\epsilon}$, to appear in Comm. Math Phys.
26. G. Hegerfeldt, C. Nappi, Comm.Math Phys. **53**, 1 (1977)
27. M. Cassandro, G. Jona-Lasinio, in *Many Degrees of Freedom in Field Theory*, L. Streit ed., Plenum Press (1971)
28. M. Cassandro, G. Jona-Lasinio, Adv. Phys. **27**, 913 (1978)
29. V.I. Oseledec, Trans. Moscow Math. Soc. **19**, 197 (1968)
30. U. Frisch, *Turbulence*, Cambridge U.P. (1995)
31. I.A. Ibragimov, Y.V. Linnik, *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff, Groningen (1971)
32. Y.V. Linnik, I.V. Ostrovski, *Decomposition of Random Variables and Vectors*, Transl. of Math. Monographs, AMS (1977)
33. I.A. Ibragimov, Y.A. Rozanov, *Processus Aleatoires Gaussiens*, Editions Mir, Moscou (1974)
34. J.-P. Eckmann, D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985)
35. T. Bohr, M.H. Jensen, G. Paladin, A. Vulpiani, *"Dynamical Systems Approach to Turbulence"*, Cambridge U.P. (1998)
36. B.B. Mandelbrot, *The Fractal Geometry of Nature*, W.H. Freeman and Co. N.Y. (1983)
37. J. Feder, A. Aharony eds., *Fractals in Physics – Essays in Honour of B. Mandelbrot*, North Holland (1990)
38. B.B. Mandelbrot, *Fractal and Scaling in Finance*, Springer N.Y. (1997)

Financial Time Series: From Batchelier's Random Walks to Multifractal 'Cascades'

Jean-Philippe Bouchaud[1] and Jean-François Muzy[2]

[1] Service de Physique de l'État Condensé, Centre d'études de Saclay, Orme des Merisiers, Gif-sur-Yvette Cedex, France, 91191, and Science & Finance, Capital Fund Management, rue Victor-Hugo 109-111, Levallois, France, 92532, bouchau@drecam.saclay.cea.fr

[2] Laboratoire SPE, CNRS UMR 6134, Université de Corse, Corte, France, 20250, muzy@univ-corse.fr

Abstract. Kolmogorov's pioneering work on turbulence is at the heart of most modern concepts and models proposed to account for the so called "intermittency phenomenon", where quiet periods are interrupted by intense bursts of activity. Recent findings in empirical finance suggest that these concepts, and more precisely the framework of multiplicative cascades, might be relevant to model the main statistical features of financial time series, in particular the intermittent nature of the volatility.

1 Introduction

Kolmogorov is undoubtedly one of the most influential scientist of the 20^{th} century. His seminal ideas pervade many fields of science: probability and statistics, dynamical systems, fluid mechanics and turbulence, front dynamics and phase ordering, to list only a few. As often in science, genuinely innovative ideas turn out to be of interest far beyond the particular context in which they are developed. In that respect, his work on stochastic processes theory and his penetrating contribution to the phenomenology of turbulence (see the contributions by Biferale et al. and Celani et al. in this book) will probably also be of fundamental importance in mathematical and empirical finance. This will be the subject of the present tribute to a rare mathematician who dared to confront himself with 'dirty' issues of the real world, and managed to turn some of them into true scientific gems.

Financial time series represent an extremely rich and fascinating source of questions. Here, a trace of human activity is recorded and stored in a quantitative way, sometimes over hundreds of years. These time series, perhaps surprisingly, turn out to reveal a very rich and non trivial statistical structure, which is to some degree universal across different assets (stocks, stock indices, currencies, etc.), regions (U.S., European, Asian) and epochs. Statistical models that describe these fluctuations have a long history, which dates back to Bachelier's "Brownian walk" model for speculative prices in 1900. Much more sophisticated models are however needed to describe more faithfully empirical data. Many recent empirical studies have shown that financial data share

J.-P. Bouchaud and J.-F. Muzy, Financial Time Series: From Batchelier's Random Walks to Multifractal 'Cascades', Lect. Notes Phys. **636**, 229–246 (2003)
http://www.springerlink.com/

many statistical properties with the "intermittent" fluctuations of turbulent velocity. In that respect, as we shall discuss below, the phenomenology of turbulence as initiated by Kolmogorov, has provided new concepts and tools to analyze market fluctuations and inspired a particularly elegant family of models that accounts for the main observed statistical properties.

The paper is organized as follows: In Sect. 2, we give some basic definitions and describe the main empirical features of financial time series. The "intermittent" or "multifractal" nature of return fluctuations are discussed in Sect. 3. We then introduce the notion of multiplicative cascade and compare intermittency in turbulence, as described in Kolmogorov's 1962 theory, with multiscaling in financial data. In Sect. 4, we review a recently introduced multifractal stochastic volatility model and its link with turbulent cascades. Conclusions and prospects are provided in Sect. 5.

2 Universal Features of Return Time Series

The modeling of random fluctuations of asset prices is of primary importance in finance, with many applications to risk control, derivative pricing and systematic trading. During the last decade, the availability of huge data sets of high frequency time series has promoted intensive statistical studies that lead to invalidate the classic and popular "Brownian walk" model, and to uncover many new and robust features. In this section, we briefly review the main statistical properties of asset prices which can be considered as universal, in the sense that they are common across most markets and epochs [1,4].

Let us first define some basic notions. If one denotes $p(t)$ the price of an asset at time t, the *return* $R_\tau(t)$, at time t and scale τ is simply the relative variation of the price from t to $t+\tau$: $R_\tau(t) = [p(t+\tau) - p(t)]\,/p(t)$. Financial mathematics mainly focuses on returns because only the relative variations are meaningful for investors. Moreover, one can empirically check that return time series have attractive statistical properties such that stationarity and ergodicity: see, e.g. Fig. 2 (but see the discussion in [2]). If τ is small enough, one has approximately: $R_\tau(t) \simeq \ln p(t+\tau) - \ln p(t)$. Let us therefore define the *continuous compound returns* (hereafter simply denoted as returns) as the variations of the logarithm of the price, $x(t) = \ln p(t)$:

$$r_\tau(t) = x(t+\tau) - x(t). \tag{1}$$

Continuous compound returns are often preferred to simple returns because returns at large time scale are obtained from small scale return by a simple aggregation of small scale returns:

$$r_{n\tau}(t) = \sum_{i=1}^{n} r_\tau(t_i) \tag{2}$$

In Fig. 1 the daily "price" of Dow-Jones index over the last century is shown. One can see that price fluctuations grow around a mean exponential

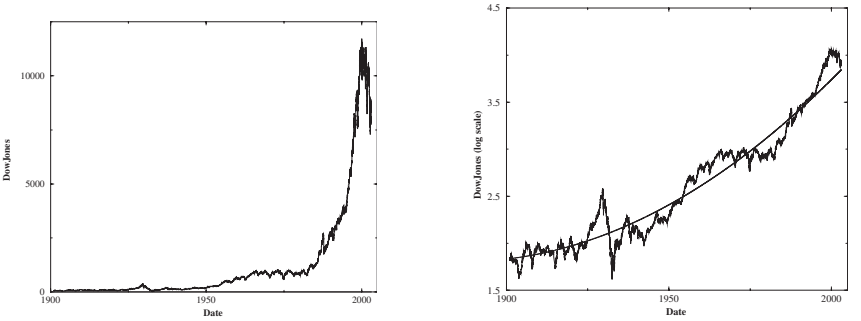

Fig. 1. a Evolution of the Dow-Jones index price over the last century (1900-2003). **b** Same data in a linear-log (base 10) representation. The full line is a parabolic fit, which shows that the average annual return has actually increased with time

trend. In Fig. 1b, we plot the logarithmic price time series $x(t) = \ln p(t)$: In this case, the fluctuations are seen to be stationary around a mean return where the drift m is around 5% per year, but has slowly increased during the whole century. Note that the current level of the Dow-Jones (after the "Internet crash") is, perhaps anecdotally, very close to its historical extrapolation.

The simplest "universal" feature of financial time series is the roughly linear growth of the variance of the return fluctuations with time. More precisely, if $m\tau$ is the mean return at scale τ, the following property holds to a good approximation:

$$\langle (r_\tau(t) - m\tau)^2 \rangle_e \simeq \sigma^2 \tau, \tag{3}$$

where $\langle ... \rangle_e$ denotes the empirical average. This behaviour typically holds for τ between a few tens of minutes and a few years, and is equivalent to the statement that relative price changes are, to a good approximation, *uncorrelated* beyond a time scale on the order of tens of minutes (on liquid markets). Very long time scales (beyond a few years) are difficult to investigate, in particular because the average drift itself m becomes time dependent. The absence of linear correlations in financial time series is often related to the so-called "market efficiency" according to which one cannot make anomalous profits by predicting future price values.

The variance σ^2 in the above equation is called, in financial economics, the *volatility*. Volatility is the simplest quantity that measures the amplitude of return fluctuations. It therefore quantifies the risk associated with some given asset. Linear growth of the variance of the fluctuations with time is typical of the Brownian motion, which was proposed as a model of return fluctuations by Bachelier (in which case the price process is called geometric Brownian motion). In the Bachelier model, returns are not only uncorrelated, as mentioned above, but actually *independent* and identical Gaussian random variables. However, this model completely fails to capture other statistical

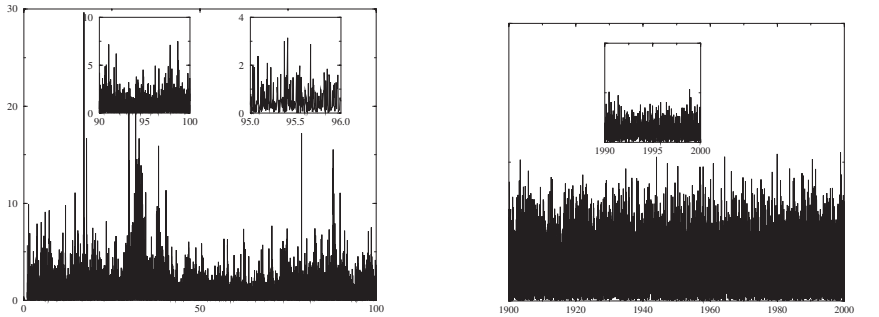

Fig. 2. a: Absolute value of the daily price returns for the Dow-Jones index over a century (1900–2000), and zoom on different scales (1990–2000 and 1995). Note that the volatility can remain high for a few years (like in the early 1930's) or for a few days. This volatility clustering can also be observed on high frequency (intra-day) data. **b** Same plot for a Brownian random walk, which shows a featureless pattern in this case

features of financial markets that even a rough analysis of empirical data allows one to identify, at least qualitatively:

(i) The distribution of returns is in fact strongly non Gaussian and its shape continuously depends on the return period τ: for τ large enough (around few months), one observes quasi-Gaussian distributions. For small τ values, the return distributions have a strong kurtosis (see Fig. 3). Several studies actually suggest that these distributions can be characterized by Pareto (power-law) tails $|\delta x|^{-1-\mu}$ with an exponent μ close to 3 even for liquid markets such as the US stock index, major currencies, or interest rates [5,6,1,7,8]. In such a case, the kurtosis would diverge. Emerging markets have even more extreme tails, with an exponent μ that can be less than 2 – in which case the volatility is infinite.

(ii) Another striking feature is the intermittent and correlated nature of return amplitudes. At some given time period τ, the volatility is a quantity that can be defined locally in various ways: the simplest ones being the square return $r_\tau(t)^2$ or the absolute return $|r_\tau(t)|$, or a local moving average of these quantities. The volatility signals are characterized by self-similar outbursts (see Fig. 2) that are reminiscent of intermittent variations of dissipation rate in fully developed turbulence [11]. The occurrence of such bursts are strongly correlated and high volatility periods tend to persist in time. This feature is known as *volatility clustering* [9,10,3,2]. This effect can be analyzed more quantitatively: the temporal correlation function of the (e.g. daily) volatility can be fit with an inverse power of the lag, with a rather small exponent in the range $0.1 - 0.3$ [10,12,13,22].

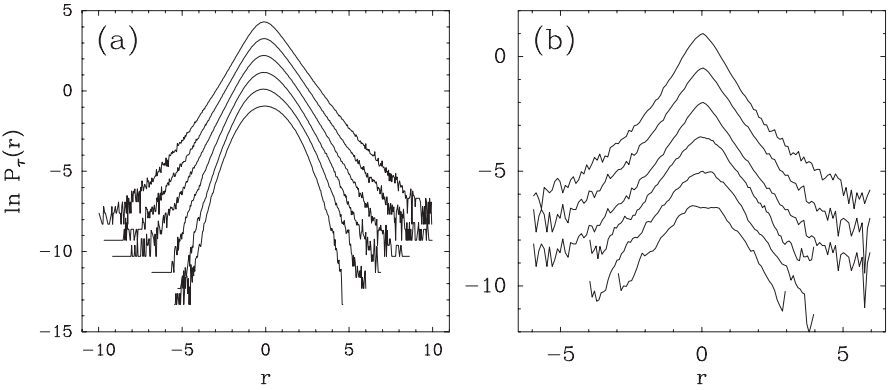

Fig. 3. Continuous deformation of turbulent velocity increments and financial returns distributions from small (*top*) to large (*bottom*) scales. **a** Standardized probability distribution functions of spatial velocity increments at different length scales in a high Reynolds number wind tunnel turbulence experiment. The distributions are plotted in logarithmic scale so that a parabola corresponds to a Gaussian distribution. **b** Standardized p.d.f. of S&P 500 index returns at different time scales from few minutes to one month. Notice that because of sample size limitation, the noise amplitude is larger than in turbulence

(iii) Past price changes and future volatilities are negatively correlated – this is the so called *leverage effect*, which reflects the fact that markets become more active after a price drop, and tend to calm down when the price rises. This correlation is most visible on stock indices [14]. This leverage effect leads to an anomalous negative skewness in the distribution of price changes as a function of time.

The most important message of these empirical studies is that prices behave very differently from simple geometric Brownian motion: extreme events are much more probable, and interesting non linear correlations (volatility-volatility and price-volatility) are observed. These "statistical anomalies" are very important for a reliable estimation of financial risk and for quantitative option pricing and hedging (see, e.g. [2]), for which one often requires an accurate model that captures the statistical features of the return for different time horizons τ. It is rather amazing to remark that empirical properties (i), (ii) and (iii) are, to some extent, also observed on experimental velocity data in fully developed turbulent flows (see Fig. 3). The framework of scaling theory and multifractal analysis, initially proposed to characterize turbulent signals [11], is therefore well suited to further characterize statistical properties of price changes over different time periods [25].

3 From Multifractal Scaling to Cascade Processes

3.1 Multiscaling of Asset Returns

For geometric Brownian motion, the return distribution is identical (on the scale of $\sigma\sqrt{\tau}$) and Gaussian for all τ. As emphasized in the previous section, the distribution of real returns, on the other hand, is *not* scale invariant, and progressively deforms from a highly kurtic shape on short time scales to a nearly Gaussian shape for larger time intervals (Fig. 3). Much the same phenomenon is observed for the distribution of velocity differences in a turbulent flow: small scale distributions depart from a Gaussian law while large scale distributions are closer to a Gaussian law (Fig. 3). A way to characterize this dependence is to study the unconditional absolute centered moments of the price returns, defined as:

$$M_q(\tau) = \langle |r_\tau(t) - m\tau|^q \rangle \,. \tag{4}$$

The question is whether these moments, when rescaled by the root mean square $[M_2(\tau)]^{q/2}$, depend in a non trivial way on τ. For a Brownian random walk, these rescaled moments would be given by τ independent constants corresponding to the moments of the Gaussian distribution. On the other hand, multifractal (or multi-affine) process corresponds to the case when:

$$M_q(\tau) \simeq A_q \tau^{\zeta_q} \tag{5}$$

where $\zeta_q \neq \frac{q}{2}\zeta_2$ is a non-linear function that can be shown to be concave ($\zeta''(q) \leq 0$). This power-law behaviour holds in a scaling regime $\tau_0 \ll \tau \ll T$. In this case, the rescaled moments scale as a (q-dependent) power-law of τ. Such a behavior for price (or log-price) changes has indeed been reported by many authors [22,27,29,30,28]. Figure 4 illustrates the empirical multifractal analysis of the S&P 500 index return. As one can see in Fig. 4b, the scaling behavior (5), corresponding to a linear dependence in a log-log representation of the absolute moments versus the time scale τ, is well verified over some range of time scales (typically 3 decades).

The multifractal nature of the index fluctuations can be verified in Fig. 4c, where one sees that the moment ratios strongly depend on the scale τ. The estimated ζ_q spectrum (Fig. 4d) has a concave shape that is well fitted by the parabola: $\zeta_q = q(1/2 + \lambda^2) - \lambda^2 q^2/2$ with $\lambda^2 \simeq 0.03$. The coefficient λ^2 that quantifies the curvature of the ζ_q function is called, in the framework of turbulence theory, the *intermittency coefficient*. The most natural way to account for the multiscaling property (5) is through the notion of cascade from coarse to fine scales.

3.2 The Cascade Picture

As previous noted for the geometric Brownian motion, the return probability distributions at different time periods τ are Gaussian and thus differ only by

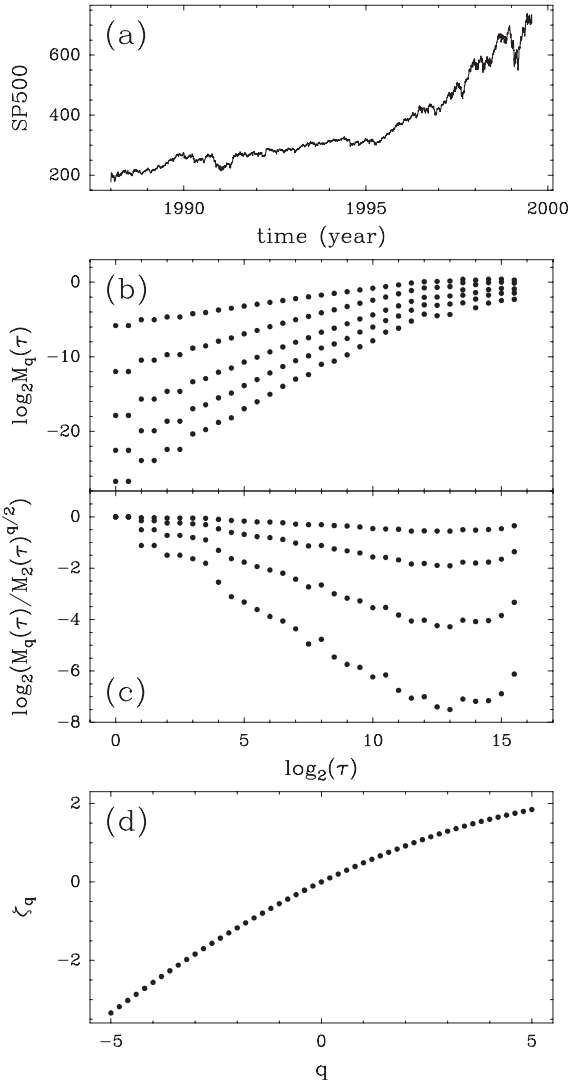

Fig. 4. Multifractal scaling analysis of S&P 500 returns. **a** S&P 500 index price series sampled at a 5 minutes rate. **b** First five absolute moments of the index (defined in (5)) as a function of the time period τ in double logarithmic representation. For each moment, a linear fit of the small scale behavior provides an estimate of ζ_q. **c** Moment ratios $M_q(\tau)/M_2(\tau)^{q/2}$ in log-log representation. Such curves would be flat for a geometric Brownian process. **d** ζ_q spectrum estimate versus q. Negative q values are obtained using a wavelet method as defined in e.g. [15]. This function is well fitted by a Kolmogorov log-normal spectrum (see text)

their width that is proportional to $\sqrt{\tau}$. If $x(t)$ is Brownian, this property can be written as

$$r_\tau(t) =_{law} \sigma_\tau \varepsilon(t) \tag{6}$$
$$\sigma_{s\tau} = s^{1/2}\sigma_\tau, \tag{7}$$

where $=_{law}$ means that the two quantities have the same probability distribution. Here, $\varepsilon(t)$ is a standardized Gaussian white noise and σ_τ is the volatility at scale τ. When going from some scale τ to the scale $s\tau$, the return volatility is simply multiplied by $s^{1/2}$. The cascade model also assumes such a multiplicative rule but the multiplicative factor is now a random variable. The volatility itself becomes a random process $\sigma_\tau(t)$:

$$r_\tau(t) =_{law} \sigma_\tau(t)\varepsilon(t) \tag{8}$$
$$\sigma_{s\tau}(t) =_{law} W_s \sigma_\tau(t), \tag{9}$$

where the law for W_s depends only on the scale ratio s and is independent of $\sigma_\tau(t)$. Let T be some coarse time period and let $s < 1$. Then, by setting $W_s = e^{\xi_s}$, and by iterating equation (9) n times, one obtains:

$$\sigma_{s^n T}(t) \underset{law}{=} W_{s^n}\sigma_T(t) \underset{law}{=} e^{\sum_{i=1}^{n}\xi_s}\sigma_T(t). \tag{10}$$

Therefore, the *logarithm* of the volatility at some fixed scale $\tau = s^n T$, can be written as a sum of an arbitrarily large number n of independent, identically distributed random variables. Mathematically, this means that the logarithm of the volatility (and hence ξ_s, the logarithm of the "weights" W_s) belongs the class of the so-called infinitely divisible distributions [16]. The simplest such distribution (often invoked using the central limit theorem) is the Gaussian law. In that case, the volatility is a log-normal random variable. As explained in the next subsection, this is precisely the model introduced by Kolmogorov in 1962 to account for intermittency turbulence. It can be proven that the random cascade equations (8) and (9) directly lead to the deformation of return probability distribution functions as observed in Fig. 3. Using the fact that ξ_s is a Gaussian variable of mean $\mu \ln(s)$ and variance $\lambda^2 \ln(s)$, one can compute the absolute moment of order q of the returns at scale $\tau = s^n T$ ($s < 1$). One finds:

$$\langle |r_\tau(t)|^q \rangle = A_q \left(\frac{\tau}{T}\right)^{\mu q - \lambda^2 q^2/2}, \tag{11}$$

where $A_q = \langle |r_T(t)|^q \rangle$.

Using a simple multiplicative cascade, we have thus recovered the empirical findings of previous sections. For log-normal random weights W_s, the return process is multifractal with a spectrum of ζ_q that scales as a parabolic: $\zeta_q = q\mu - \lambda^2 q^2/2$ where the parameter μ is related to the mean of $\ln W_s$. The

curvature λ^2 (the intermittency coefficient) is then related to the variance of $\ln W_s$ and therefore to the variance of the log-volatility $\ln(\sigma_\tau)$:

$$\langle (\ln \sigma_\tau(t))^2 \rangle - \langle \ln \sigma_\tau(t) \rangle^2 = -\lambda^2 \ln(\tau) + V_0. \tag{12}$$

The random character of $\ln W_s$ is therefore directly related to the intermittency of the returns.

3.3 Kolmogorov's Legacy, Turbulence, and Finance

The first log-normal theory of intermittency was presented by Kolmogorov in 1961 at the Colloque International de la Mécanique de la Turbulence in Marseille and published in the famous Kolmogorov 1962 paper [18]. In this work, motivated by a remark by Landau [11] closely related to another work by Obhukov (a student of Kolmogorov) [19], Kolmorogov proposed a "refinement" of his own 1941 dimensional theory of turbulence [17] in order to account for the spatial fluctuations of the dissipation of energy at the origin of intermittency. Kolmogorov-Obhukov intermittency theory of turbulence can be summarized as follows [18,19]: If $e_\ell(\boldsymbol{x}, t)$ is the local dissipation rate of energy at scale ℓ, and $\delta_\ell v(\boldsymbol{x}, t)$ the local longitudinal velocity field increment at scale ℓ, one assumes that $\ln(e_\ell)$ is normally distributed, and that:

$$\delta_\ell v(\boldsymbol{x}, t) =_{law} e_\ell(\boldsymbol{x}, t)^{1/3} \ell^{1/3} \varepsilon(\boldsymbol{x}, t) \tag{13}$$

$$\langle (\ln(e_\ell))^2 \rangle - \langle \ln(e_\ell) \rangle^2 = -\lambda^2 \ln(\ell) + V_0 \tag{14}$$

where $\varepsilon(\boldsymbol{x}, t)$ is a stochastic field that does not depend of the scale ℓ nor on e_ℓ. The similarity with (8), (9) and (12) is obvious. The (log-normal) multiscaling properties of turbulent velocity fields can be directly deduced along the same lines as in previous computation. Besides their own interest for turbulence, Kolmogorov's ideas on intermittency have inspired many further developments (like Mandelbrot's cascades [20] or the Parisi and Frisch multifractal formalism [21]) that are still at the heart of active research in many fields of applied science and mathematics.

The quantitative similarity between turbulence and finance was first suggested by Ghashghaie et al. [25] who directly checked the relevance of (12) on high frequency FX rate data. These authors suggested that the analog of the energy cascade in turbulence could be an "information" cascade from large time periods to small ones. However, this notion is rather fuzzy and there is still no explicit model where a causal information cascade can be constructed and given a precise meaning (see also the discussion in the conclusion). The empirical study of [25] has however been supported by many other works, in particular by the multifractal scaling analysis described above. The analogy can be summarized in the following table:

Turbulence (Euler)	Finance
δv_l : Velocity increment at scale ℓ	r_τ: Return at time scale τ
e_ℓ: Local energy dissipation rate	σ_ℓ^2: Local volatility
Kolmogorov 1941 model: $\delta v_\ell \sim \ell^{1/3}$	Bachelier 1900 model: $r_\tau \sim \tau^{1/2}$
$\delta v_\ell = e_l^{1/3}\ell^{1/3}\epsilon$: Kolmogorov hypothesis	$r_\tau = \sigma_\ell\epsilon$: Stochastic volatility model
Kolmogorov 1962 energy cascade	Log-normal volatility cascade
$\Rightarrow$ Intermittency	$\Rightarrow$ Multifractality

Let us now discuss how the multifractal properties of asset returns and the framework of cascade models can lead to an explicit dynamical model for financial time series.

4 The Multifractal Random Walk

The cascade picture assumes that the volatility can be constructed as a *product* of random variables associated with different time scales. In the previous section, we exclusively focused on return probability distribution functions (mono-variate laws) and scaling laws associated with such models. Explicit constructions of stochastic processes whose marginals satisfy a random multiplicative rule were first introduced by Mandelbrot [20]. They are known as Mandelbrot's cascades or random multiplicative cascades. The construction of a Mandelbrot cascade, illustrated in Fig. 5, always involves a discrete scale ratio s (generally $s = 1/2$). One begins with the volatility at the coarsest scale and proceeds in a recursive manner to finer resolution: The n-th step of the volatility construction corresponds to scale 2^{-n} and is obtained from the $(n-1)$-th step by multiplication with a positive random process W, the max of which does not depend on n. More precisely, the volatility in each subinterval is the volatility of its parent interval multiplied by an independent copy of W.

Mandelbrot cascades are considered to be the paradigm of multifractal processes. They have been extensively used for modeling scale-invariance properties in many fields, in particular statistical finance [27,28]. However, this class of models possess several drawbacks: (i) they involve a preferred scale ratio s, (ii) they are not stationary and (iii) they violate causality. In that respect, it is difficult to see how such models could arise from a realistic (agent based) description of financial markets.

Recently, Bacry, Muzy and Delour (BMD) [22,23] introduced a model that does not possess any of the above limitations and captures the essential features of cascades through their correlation structure. Let us first show, from a general point of view, how volatility fluctuations and correlations can induce multiscaling. We will then discuss the BMD model where multifractality is indeed exact, and for which ζ_q can be computed for all q.

As mentioned in Sect. 2, the empirical volatility correlation function decays slowly, as a power law. More precisely, if the correlation function of the *square returns* δx^2 (which serves as a proxy for the true volatility) decays as

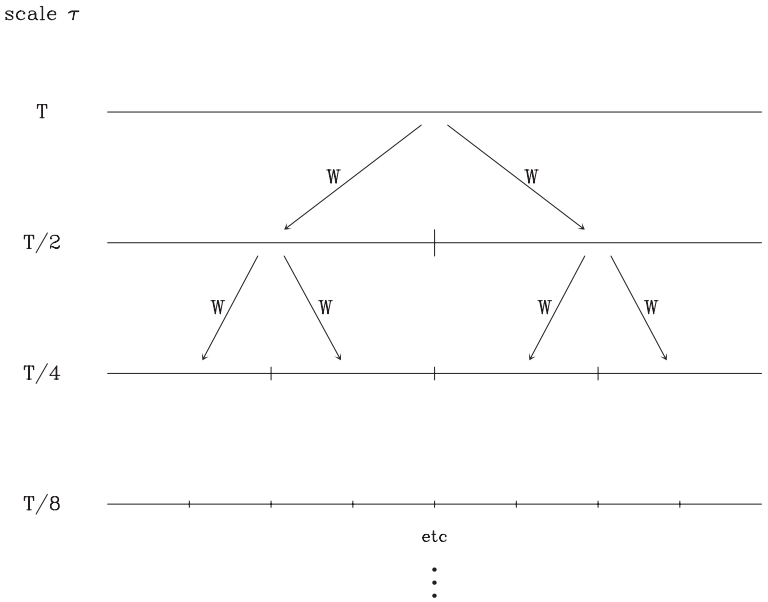

Fig. 5. Multiplicative construction of a Mandelbrot cascade. One starts at the coarsest scale T and constructs the volatility fluctuations at fine scales using a recursive multiplication rule. The variables W are independent copies of the same random variable

$\tau^{-\nu}$ with $\nu < 1$, it is quite easy to obtain explicitly the fourth moment of the price return for large τ:

$$M_4(\tau) \approx \sigma^4 \tau^2 \left(1 + A\tau^{-\nu}\right), \tag{15}$$

where A measures the amplitude of the long range part of the square volatility correlation. The fourth moment of the price difference therefore behaves as the *sum* of two power-laws, not as a unique power-law as in the case of a multifractal process. However, when ν is small and in a restricted range of τ, this sum of two power-laws is indistinguishable from a unique power-law with an effective exponent $\zeta_{4,eff}$ somewhere between 2 and $2 - \nu$; therefore $\zeta_{4,eff} < 2\zeta_2 = 2$.

In the BMD model, the key ingredient is the volatility correlation shape that mimics cascade features. Indeed, as remarked in ref. [26], the tree like structure underlying a Mandelbrot cascade implies that the volatility logarithm covariance decreases very slowly, as a logarithm function, i.e.,

$$\langle \ln(\sigma_\tau(t) \ln(\sigma_\tau(t + \Delta t)\rangle - \langle \ln(\sigma_\tau(t)\rangle^2 = C_0 - \lambda^2 \ln(\Delta t + \tau). \tag{16}$$

This equation can be seen as a generalization of the Kolmogorov equation (12) that describes only the behavior of the log-volatility variance (corresponding to the log-dissipation variance in Kolmogorov's paper). It is important to note

that such a logarithmic behaviour of the covariance has indeed been observed for empirically estimated log-volatilities in various stock market data [26][1].

The BMD model involves (16) within the continuous time limit of a discrete stochastic volatility model. One first discretizes time in units of an elementary time step τ_0 and sets $t \equiv i\tau_0$. The volatility σ_i at "time" i is a log-normal random variable such that $\sigma_i = \sigma_0 \exp \xi_i$, where the Gaussian process ξ_i has the same covariance as in (16):

$$\langle \xi_i \rangle = -\lambda^2 \ln\left(\frac{T}{\tau_0}\right) \equiv \mu_0; \qquad \langle \xi_i \xi_j \rangle - \mu_0^2 = \lambda^2 \ln\left(\frac{T}{\tau_0}\right) - \lambda^2 \ln(|i-j|+1), \tag{17}$$

for $|i-j|\tau_0 \leq T$. Here T is a large cut-off time scale beyond which the volatility correlation vanishes. In the above equation, the brackets stand for the mathematical expectation. The choice of the mean value μ_0 is such that $\langle \sigma^2 \rangle = \sigma_0^2$. As before, the parameter λ^2 measures the intensity of volatility fluctuations (called in the finance jargon the 'vol of the vol'), and corresponds to the intermittency parameter.

Now, the price returns are constructed as:

$$x\left((i+1)\tau_0\right) - x\left(i\tau_0\right) = r_{\tau_0}(i) \equiv \sigma_i \varepsilon_i = \sigma_0 e^{\xi_i} \varepsilon_i, \tag{18}$$

where the ε_i are a set of independent, identically distributed random variables of zero mean and variance equal to τ_0. One also assumes that the ε_i and the ξ_i are independent (but see [32]). In the original BMD model, ε_i's are Gaussian, and the continuous time limit $\tau_0 = \mathrm{d}t \to 0$ is taken. Since $x = \ln p$, where p is the price, the exponential of a sample path of the BMD model plotted in Fig. 6a can be compared to the real price charts of Figs. 1a and 3a.

The multifractal scaling properties of this model can be computed explicitly. Moreover, using the properties of multivariate Gaussian variables, one can get closed expressions for all even moments $M_q(\tau)$ ($q = 2k$). In the case $q = 2$ one trivially finds:

$$M_2(\tau = \ell\tau_0) = \sigma_0^2 \, \ell\tau_0 \equiv \sigma_0^2 \tau, \tag{19}$$

independently of λ^2. For $q \neq 2$, one has to distinguish between the cases $q\lambda^2 < 1$ and $q\lambda^2 > 1$. For $q\lambda^2 < 1$, the corresponding moments are finite, and one finds, in the scaling region $\tau_0 \ll \tau \leq T$, a *true* multifractal behaviour [22,23]:

$$M_q(\tau) = A_q \tau^{\zeta_q} \tag{20}$$

where

$$\begin{cases} \zeta_q &= q(1/2+\lambda^2) - q^2\lambda^2/2 \\ A_q &= T^{q/2}\sigma^q (q-1)!! \prod_{k=0}^{q/2-1} \frac{\Gamma(1-2\lambda^2 k)^2 \Gamma(1-2\lambda^2(k+1))}{\Gamma(2-2\lambda^2(q/2+k-1))\Gamma(1-2\lambda^2)} \text{ for } q \text{ even} \end{cases} \tag{21}$$

[1] In fact, other functional forms, motivated by some simple agent based models, are also possible [35].

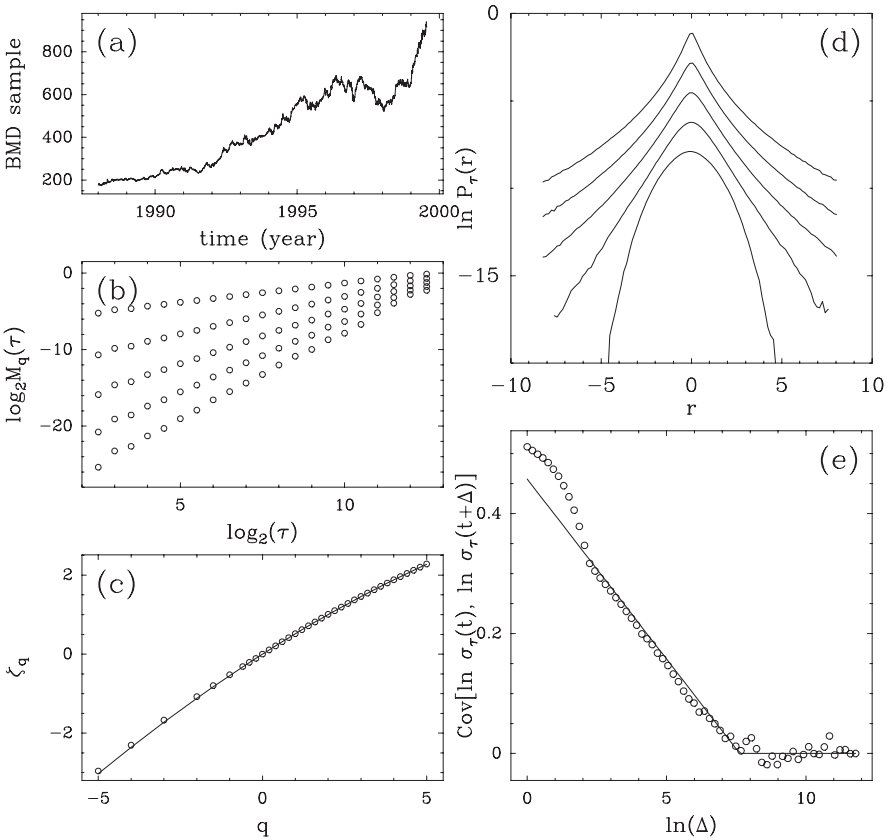

Fig. 6. Multifractal properties of BMD model. **a** Exponential of a BMD process realization. The parameters have been adjusted in order to mimic the features of S&P 500 index reported in Fig. 4. **b** Multifractal scaling of BMD return moments for $q = 1, 2, 3, 4, 5$. **c** Estimated ζ_q spectrum ($\circ$) compared with the log-normal analytical expression (21) (*solid line*). **d** Evolution of the return probability distributions across scales, from nearly Gaussian at coarse scale (bottom) to fat tailed law at small scales (*top*). **e** Log-volatility covariance as a function of the logarithm of the lag τ

For $q\lambda^2 > 1$, on the other hand, the moments diverge, suggesting that the unconditional distribution of $x(t+\tau) - x(t)$ has power law tails with an exponent $\mu = 1/\lambda^2$ (possibly multiplied by some slow function). These multifractal scaling properties that BMD processes are numerically checked in Figs. 6a and 6b, where one recovers the same features as in Fig. 4 for the S&P 500 index. Since volatility correlations are absent for $\tau \gg T$, the scaling becomes that of a standard random walk, for which $\zeta_q = q/2$. The corresponding distribution of price returns thus becomes progressively Gaussian. An illustration of the progressive deformation of the distributions as τ increases

in the BMD model is reported in Fig. 6c. This figure can be directly compared to Fig. 3. As shown in Fig. 6e, this model also reproduces the empirically observed logarithmic covariance of log-volatility (16). This functional form, introduced at a "microscopic" level (at scale $\tau_0 \to 0$), is therefore stable across finite time scales τ.

To summarize, the BMD process is attractive for modeling financial time series since it reproduces most "stylized facts" reviewed in Sect. 2 and has exact multifractal properties as described in Sect. 3. Moreover, this model has stationary increments, does not exhibit any particular scale ratio and can be formulated in a purely causal way: the log volatility ξ_i can be expressed as a sum over "past" random shocks, with a memory kernel that decays as the inverse square root of the time lag [36]. It would be interesting to give a precise economic justification to this causal construction.

It is useful to relate the above results of the BMD model to the general discussion relating volatility correlations and multifractality, given at the beginning of this section. It is easy to show that the correlation function of the square returns is, within the BMD model, proportional to $(T/\tau)^\nu$ with $\nu \equiv 4\lambda^2$. We can choose λ small enough such that $\nu < 1$ and thus place our analysis in the regime discussed above, where (15) holds. The interesting point here is that for large T, the constant A appearing in (15) is also large: $A = T^\nu/(2-\nu)(1-\nu)$. Therefore, the second term in (15) is dominant whenever $\tau \ll T$. In the regime $\tau_0 \ll \tau \ll T$, one indeed finds a *unique* power-law for $M_4(\tau)$, with $\zeta_4 \equiv 2 - \nu = 2(1 - 2\lambda^2)$, in agreement with the general expression of ζ_q in the BMD model.

Let us end this section with a few remarks.

- Direct studies of the empirical distribution of the volatility is indeed compatible with the assumption of log-normality, although an inverse gamma distribution also fits the data very well [40,2].
- One of the predictions of the BMD model is the equality between the intermittency coefficient estimated from the curvature of the ζ_q function and the slope of the log-volatility covariance logarithmic decrease. The direct study of various empirical log-volatility correlation functions shows that they can indeed be fit by a logarithm of the time lag, with a slope that is roughly equal to the corresponding intermittency coefficient λ^2. These empirical studies also suggest that the integral time T is around one or few years.
- On the other hand, the empirical tail of the distribution of price increments is described by a power-law with an exponent μ in the range $3-5$ [5,6,1], much smaller than the value $\mu = 1/\lambda^2 \sim 10-100$ predicted by the BMD model. This suggests that the random variable ε itself is non Gaussian and further fattens the tails. Other laws for the stochastic volatility could also lead to fatter tails [24].
- Finally, one can extend the above multifractal model to account for a skewed distribution of returns and the return-volatility correlations mentioned in Sect. 2 (see also [14]). In the BMD model, all the odd moments

of the process vanish by symmetry. A simple possibility, recently investigated in [32], is to correlate negatively the variable ξ_i with 'past' values of the variables ε_j, $j < i$, through a kernel that decays as a power-law. In this case, the multifractality properties of the model are preserved, though the expression for ζ_q is different for q even and q odd.

5 Conclusions

Kolmogorov's pioneering work on the scaling properties of fully developed turbulence has been the seed for considerable theoretical developments and has influenced many domains, far beyond the statistical theory of turbulence. In this paper, we discussed a somewhat unexpected outcome of Kolmogorov's ideas: models for financial time series. The interest of the physics community in financial economics problems has grown considerably over the past few years. Using scaling concepts and (multi-)fractal analysis, physicists (but not only physicists [37,28]) have shed new light on financial time series analysis. As we have emphasized, the analogy between the apparently remote fields of turbulence and finance has led to very interesting results; many exciting developments can still be expected. Applications of the BMD model (or some other related model) to risk management, volatility prediction [37] and option pricing [32] have already been considered. The slow relaxation of the volatility after a shock can also be accurately modeled within the BMD framework [36]. Let us mention that the BMD model, originally designed to reproduce price fluctuations, has found an application "back" in Lagrangian turbulence [33]. It also has deep connections with the theory of disordered systems [38].

The description of financial data using cascade-like ideas is however still only phenomenological. An important theoretical issue is to understand the underlying mechanisms that can give rise to such a remarkable structure of the volatility correlations. One path could be to justify why the volatility response to a shock decays as the inverse square root of time, a property implying the logarithmic decay of the correlation function which is at the heart of multifractal scaling. Another possibility, discussed in [35], is that although a logarithmic form provides a very good fit to the data, it does not exclude other functional forms. For example, a stretched exponential form actually fits the data equally well, and is suggested by a simple model where the activity of agents is subordinated to a random walk signal [35]. Finally, it is interesting to mention that direct evidence of different time scales in the trading activity has been presented in [39].

Let us conclude by quoting Kolmogorov himself when he explained his interest for turbulence [34]:

> *"...It was clear to me from the very beginning that the main mathematical instrument in this study must be the theory of random functions of several variables (random fields) which had only then originated. Moreover, it soon became clear to me that there was no chance of*

developing a purely closed mathematical theory, it was necessary to use some hypotheses based on the results of the treatment of the experimental data...”

This sentence clearly applies equally well to statistical finance.

Acknowledgments

We thank E. Bacry, J. Delour and B. Pochart for sharing their results with us.

References

1. D.M. Guillaume, M.M. Dacorogna, R.D. Davé, U.A. Müller, R.B. Olsen, and O.V. Pictet, *Finance and Stochastics* **1** 95 (1997); M. Dacorogna, R. Gençay, U. Müller, R. Olsen, and O. Pictet, *An Introduction to High-Frequency Finance* (Academic Press, London, 2001)
2. J.-P. Bouchaud and M. Potters, *Théorie des Risques Financiers*, Aléa-Saclay, 1997; *Theory of Financial Risks*, Cambridge University Press, 2000, and 2003
3. R. Mantegna and H.E. Stanley, *An Introduction to Econophysics*, Cambridge University Press, 1999
4. R. Cont, *Empirical properties of asset returns: stylized facts and statistical issues*, Quantitative Finance, **1**, 223 (2001)
5. V. Plerou, P. Gopikrishnan, L.A. Amaral, M. Meyer, and H.E. Stanley, *Scaling of the distribution of price fluctuations of individual companies*, Phys. Rev. **E60** 6519 (1999); P. Gopikrishnan, V. Plerou, L.A. Amaral, M. Meyer, and H.E. Stanley, *Scaling of the distribution of fluctuations of financial market indices*, Phys. Rev. E **60** 5305 (1999)
6. T. Lux, *The stable Paretian hypothesis and the frequency of large returns: an examination of major German stocks*, Applied Financial Economics, **6**, 463, (1996)
7. F. Longin, *The asymptotic distribution of extreme stock market returns*, Journal of Business **69** 383 (1996)
8. J. Nuyts and I. Platten, *Phenomenology of the term structure of interest rates with Padé approximants*, Physica A **299**, 528 (2001)
9. A. Lo, *Long term memory in stock market prices*, Econometrica **59**, 1279 (1991)
10. Z. Ding, C.W.J. Granger, and R. F. Engle, *A long memory property of stock market returns and a new model*, J. Empirical Finance **1**, 83 (1993)
11. U. Frisch, *Turbulence: The Legacy of A. Kolmogorov*, Cambridge University Press (1997)
12. M. Potters, R. Cont and J.-P. Bouchaud, *Financial Markets as Adaptive Ecosystems*, *Europhys. Lett.* **41**, 239 (1998)
13. Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H.E. Stanley, *Correlations in Economic Time Series*, Physica A **245**, 437 (1997)
14. J.P. Bouchaud, A. Matacz, and M. Potters, *The leverage effect in financial markets: retarded volatility and market panic*, Physical Review Letters **87**, 228701 (2001)

15. J.F. Muzy, E. Bacry, and A. Arneodo, *The multifractal formalism revisited with wavelets*, Int. J. of Bifurcat. and Chaos **4**, 245 (1994)
16. W. Feller, *An introduction to probability theory and is applications*, Vol. 2, John Wiley & Sons (1971)
17. (a) A.N. Kolmogorov, *The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers*, Dokl. Akad. Nauk SSSR **30**, 301 (1941). (b) A.N. Kolmogorov, *On degeneration of isotropic turbulence in an incompressible viscous liquid*, Dokl. Akad. Nauk SSSR **31**, 338 (1941)
18. A.N. Kolmogorov, *A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number*, J. Fluid. Mech. **13**, 82 (1962)
19. A.M. Obhukov, *Some specific features of atmospheric turbulence*, J. Fluid. Mech. **13**, 77 (1962)
20. B.B. Mandelbrot, *Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier*, J. Fluid Mech. **62**, 331 (1974)
21. G. Parisi and U. Frisch, *Fully developed turbulence and intermittency*, Proc. of Int. School on *"Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics"*, M. Ghil, R. Benzi and G. Parisi editors, (North-Holland, Amsterdam, 1985) p. 84
22. J.-F. Muzy, J. Delour, and E. Bacry, *Modelling fluctuations of financial time series: from cascade process to stochastic volatility model*, Eur. Phys. J. B **17**, 537–548 (2000)
23. E. Bacry, J. Delour, and J.F. Muzy, *Multifractal random walk*, Phys. Rev. E **64**, 026103 (2001)
24. J.F. Muzy and E. Bacry, *Multifractal stationary random measures and multifractal random walks with log-infinitely divisible scaling laws*, Phys. Rev. E **66**, 056121 (2002)
25. S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, Y. Dodge, *Turbulent cascades in foreign exchange markets*, *Nature* **381**, 767 (1996)
26. A. Arneodo, J.-F. Muzy, and D. Sornette, *Causal cascade in the stock market from the 'infrared' to the 'ultraviolet'*, Eur. Phys. J. B **2**, 277 (1998)
27. B.B. Mandelbrot, *Fractals and Scaling in Finance*, Springer, New York, 1997; A. Fisher, L. Calvet, and B.B. Mandelbrot, *Multifractality of DEM/$ rates*, Cowles Foundation Discussion Paper 1165; B.B. Mandelbrot, *A multifractal walk down Wall street*, Scientific American, Feb. (1999)
28. T. Lux, *Turbulence in financial markets: the surprising explanatory power of simple cascade models*, Quantitative Finance **1**, 632, (2001)
29. F. Schmitt, D. Schertzer, S. Lovejoy, *Multifractal analysis of Foreign exchange data*, Applied Stochastic Models and Data Analysis **15** 29 (1999)
30. M.-E. Brachet, E. Taflin, and J.M. Tchéou, *Scaling transformation and probability distributions for financial time series*, e-print cond-mat/9905169
31. J.P. Bouchaud, M. Potters, and M. Meyer, *Apparent multifractality in financial time series*, Eur. Phys. J. B **13**, 595 (1999)
32. B. Pochart and J.P. Bouchaud, *The skewed multifractal random walk with applications to option smiles*, Quantitative Finance, **2**, 303 (2002)
33. N. Mordant, J. Delour, E. Lévêque, A. Arneodo, and J.F. Pinton, *Long time correlations in lagrangian dynamics: a key to intermittency*, Phys. Rev. Lett. **89**, 254502 (2002)
34. A.N. Kolmogorov, *Selected works of A.N. Kolmogorov, Vol. I, Mathematics and mechanics*, V.M. Tikhomirov ed., Kluwer (1991)

35. J.P. Bouchaud, I. Giardina, and M. Mézard, *On a universal mechanism for long ranged volatility correlations*, Quantitative Finance **1**, 212 (2001); I. Giardina and J.P. Bouchaud, *Bubbles, Crashes and Intermittency in agent based market models*, Eur. Phys. J. B **31**, 421 (2003)
36. D. Sornette, Y. Malevergne, and J.F. Muzy, *What causes crashes*, Risk Magazine, 67 (Feb. 2003)
37. L. Calvet and A. Fisher, *Forecasting multifractal volatility*, Journal of Econometrics **105**, 27, (2001)
38. D. Carpentier, P. Le Doussal, *Glass transition of a particle in a random potential, front selection in nonlinear renormalisation group, and entropic phenomena in Liouville and sinh-Gordon model*, Phys. Rev. E **63**, 026110 (2001)
39. G. Zumbach and P. Lynch, *Heterogeneous volatility cascade in financial markets*, Physica A **298**, 521, (2001)
40. S. Miccichè, G. Bonanno, F. Lillo, and R.N. Mantegna, *Volatility in financial markets: stochastic models and empirical results*, e-print cond-mat/0202527

Index