SPRINGER REFERENCE

Daniel ben-Avraham
Italo Capuzzo Dolcetta
Giuseppe Gaeta
Shlomo Havlin
Mohamed A. Helal
Matthias Kawski
Bryna Kra
Ferdinand Verhulst
Section Editors

Robert A. Meyers
Editor-in-Chief

Mathematics of Complexity and Dynamical Systems

Selected entries from the Encyclopedia of Complexity and Systems Science

Springer

Copyrighted material

Robert A. Meyers (Ed.)

Mathematics of Complexity and Dynamical Systems

With 489 Figures and 25 Tables



Copyrighted material

Robert A. Meyers, Ph. D.
Editor-in-Chief
RAMTECH LIMITED
122 Escalle Lane
Larkspur, CA 94939
USA
robert.meyers@ramtechlimited.org

Library of Congress Control Number: 2011939016

ISBN: 978-1-4614-1806-1
This publication is available also as:
Print publication under ISBN: 978-1-4614-1805-4 and
Print and electronic bundle under ISBN 978-1-4614-1807-8

© 2012 SpringerScience+Business Media, LLC.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

This book consists of selections from the *Encyclopedia of Complexity and Systems Science* edited by Robert A. Meyers, published by Springer New York in 2009.

springer.com

Printed on acid free paper

Copyrighted material

Table of Contents

Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory
Abdul-Majid Wazwaz 1

Anomalous Diffusion on Fractal Networks
Igor M. Sokolov 13

Biological Fluid Dynamics, Non-linear Partial Differential Equations
Antonio DeSimone, François Alouges, Aline Lefebvre 26

Catastrophe Theory
Werner Sanns 32

Center Manifolds
George Osipenko 48

Chaos and Ergodic Theory
Jérôme Buzzi 63

Chronological Calculus in Systems and Control Theory
Matthias Kawski 88

Control of Non-linear Partial Differential Equations
Fatiha Alabau-Boussouira, Piermarco Cannarsa 102

Diagrammatic Methods in Classical Perturbation Theory
Guido Gentile 126

Discrete Control Systems
Taeyoung Lee, Melvin Leok, Harris McClamroch 143

Dispersion Phenomena in Partial Differential Equations
Piero D'Ancona 160

Dynamics on Fractals
Raymond L. Orbach 175

Dynamics of Parametric Excitation
Alan Champneys 183

Entropy in Ergodic Theory
Jonathan L. F. King 205

Ergodicity and Mixing Properties
Anthony Quas 225

Ergodic Theorems
Andrés del Junco 241

Ergodic Theory: Basic Examples and Constructions
Matthew Nicol, Karl Petersen 264

Ergodic Theory: Fractal Geometry
Jörg Schmeling 288

Copyrighted material

Ergodic Theory on Homogeneous Spaces and Metric Number Theory
Dmitry Kleinbock 302
Ergodic Theory: Interactions with Combinatorics and Number Theory
Tom Ward 313
Ergodic Theory, Introduction to
Bryna Kra 327
Ergodic Theory: Non-singular Transformations
Alexandre I. Danilenko, Cesar E. Silva 329
Ergodic Theory: Recurrence
Nikos Frantzikinakis, Randall McCutcheon 357
Ergodic Theory: Rigidity
Viorel Nițică 369
Existence and Uniqueness of Solutions of Initial Value Problems
Gianne Derks 383
Finite Dimensional Controllability
Lionel Rosier 395
Fractal Geometry, A Brief Introduction to
Armin Bunde, Shlomo Havlin 409
Fractal Growth Processes
Leonard M. Sander 429
Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks
Sidney Redner 446
Fractal and Multifractal Time Series
Jan W. Kantelhardt 463
Fractals in Biology
Sergey V. Buldyrev 488
Fractals and Economics
Misako Takayasu, Hideki Takayasu 512
Fractals in Geology and Geophysics
Donald L. Turcotte 532
Fractals Meet Chaos
Tony Crilly 537
Fractals and Multifractals, Introduction to
Daniel ben-Avraham, Shlomo Havlin 557
Fractals and Percolation
Yakov M. Strelniker, Shlomo Havlin, Armin Bunde 559
Fractals in the Quantum Theory of Spacetime
Laurent Nottale 571
Fractal Structures in Condensed Matter Physics
Tsuneyoshi Nakayama 591
Fractals and Wavelets: What Can We Learn on Transcription and Replication from Wavelet-Based Multifractal Analysis of DNA Sequences?
Alain Arneodo, Benjamin Audit, Edward-Benedict Brodie of Brodie, Samuel Nicolay, Marie Touchon, Yves d'Aubenton-Carafa, Maxime Huvet, Claude Thermes 606
Fractal and Transfractal Scale-Free Networks
Hernán D. Rozenfeld, Lazaros K. Gallos, Chaoming Song, Hernán A. Makse 637

Copyrighted material

Hamiltonian Perturbation Theory (and Transition to Chaos)
Henk W. Broer, Heinz Hanßmann 657
Hamilton–Jacobi Equations and Weak KAM Theory
Antonio Siconolfi 683
Hybrid Control Systems
Andrew R. Teel, Ricardo G. Sanfelice, Rafal Goebel 704
Hyperbolic Conservation Laws
Alberto Bressan 729
Hyperbolic Dynamical Systems
Vitor Araújo, Marcelo Viana 740
Infinite Dimensional Controllability
Olivier Glass 755
Inverse Scattering Transform and the Theory of Solitons
Tuncay Aktosun 771
Isomorphism Theory in Ergodic Theory
Christopher Hoffman 783
Joinings in Ergodic Theory
Thierry de la Rue 796
Kolmogorov–Arnold–Moser (KAM) Theory
Luigi Chierchia 810
Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the
Yu-Jie Ren 837
Korteweg–de Vries Equation (KdV), History, Exact N-Soliton Solutions and Further Properties of the
Yi Zang 884
Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the
Doğan Kaya 890
Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the
Mustafa Inc 908
Learning, System Identification, and Complexity
M. Vidyasagar 924
Lyapunov–Schmidt Method for Dynamical Systems
André Vanderbauwhede 937
Maximum Principle in Optimal Control
Velimir Jurdjevic 953
Measure Preserving Systems
Karl Petersen 964
Mechanical Systems: Symmetries and Reduction
Jerrold E. Marsden, Tudor S. Ratiu 981
Navier–Stokes Equations: A Mathematical Analysis
Giovanni P. Galdi 1009
***n*-Body Problem and Choreographies**
Susanna Terracini 1043
Nekhoroshev Theory
Laurent Niederman 1070
Non-linear Dynamics, Symmetry and Perturbation Theory in
Giuseppe Gaeta 1082

Copyrighted material

Non-linear Internal Waves
Moustafa S. Abou-Dina, Mohamed A. Helal . . . 1102
Non-linear Ordinary Differential Equations and Dynamical Systems, Introduction to
Ferdinand Verhulst . . . 1111
Non-linear Partial Differential Equations, Introduction to
Italo Capuzzo Dolcetta . . . 1113
Non-linear Partial Differential Equations, Viscosity Solution Method in
Shigeaki Koike . . . 1115
Non-linear Stochastic Partial Differential Equations
Giuseppe Da Prato . . . 1126
Nonsmooth Analysis in Systems and Control Theory
Francis Clarke . . . 1137
Normal Forms in Perturbation Theory
Henk W. Broer . . . 1152
Numerical Bifurcation Analysis
Hil Meijer, Fabio Dercole, Bart Oldeman . . . 1172
Observability (Deterministic Systems) and Realization Theory
Jean-Paul André Gauthier . . . 1195
Partial Differential Equations that Lead to Solitons
Doğan Kaya . . . 1205
Periodic Orbits of Hamiltonian Systems
Luca Sbano . . . 1212
Periodic Solutions of Non-autonomous Ordinary Differential Equations
Jean Mawhin . . . 1236
Perturbation Analysis of Parametric Resonance
Ferdinand Verhulst . . . 1251
Perturbation of Equilibria in the Mathematical Theory of Evolution
Angel Sánchez . . . 1265
Perturbation of Systems with Nilpotent Real Part
Todor Gramchev . . . 1276
Perturbation Theory
Giovanni Gallavotti . . . 1290
Perturbation Theory in Celestial Mechanics
Alessandra Celletti . . . 1301
Perturbation Theory, Introduction to
Giuseppe Gaeta . . . 1314
Perturbation Theory and Molecular Dynamics
Gianluca Panati . . . 1317
Perturbation Theory for Non-smooth Systems
Marco Antônio Teixeira . . . 1325
Perturbation Theory for PDEs
Dario Bambusi . . . 1337
Perturbation Theory in Quantum Mechanics
Luigi E. Picasso, Luciano Bracci, Emilio d'Emilio . . . 1351
Perturbation Theory, Semiclassical
Andrea Sacchetti . . . 1376

Copyrighted material

Perturbative Expansions, Convergence of
Sebastian Walcher 1389
Phase Transitions on Fractals and Networks
Dietrich Stauffer 1400
Philosophy of Science, Mathematical Models in
Zoltan Domotor 1407
Pressure and Equilibrium States in Ergodic Theory
Jean-René Chazottes, Gerhard Keller 1422
Quantum Bifurcations
Boris Zhilinskií 1438
Reaction Kinetics in Fractals
Ezequiel V. Albano 1457
Relaxation Oscillations
Johan Grasman 1475
Robotic Networks, Distributed Algorithms for
Francesco Bullo, Jorge Cortés, Sonia Martínez 1489
Scaling Limits of Large Systems of Non-linear Partial Differential Equations
D. Benedetto, M. Pulvirenti 1505
Shallow Water Waves and Solitary Waves
Willy Hereman 1520
Smooth Ergodic Theory
Amie Wilkinson 1533
Soliton Perturbation
Ji-Huan He 1548
Solitons and Compactons
Ji-Huan He, Shun-dong Zhu 1553
Solitons: Historical and Physical Introduction
François Marin 1561
Solitons Interactions
Tarmo Soomere 1576
Solitons, Introduction to
Mohamed A. Helal 1601
Solitons, Tsunamis and Oceanographical Applications of
M. Lakshmanan 1603
Spectral Theory of Dynamical Systems
Mariusz Lemańczyk 1618
Stability and Feedback Stabilization
Eduardo D. Sontag 1639
Stability Theory of Ordinary Differential Equations
Carmen Chicone 1653
Stochastic Noises, Observation, Identification and Realization with
Giorgio Picci 1672
Symbolic Dynamics
Brian Marcus 1689
System Regulation and Design, Geometric and Algebraic Methods in
Alberto Isidori, Lorenzo Marconi 1711

Copyrighted material

Systems and Control, Introduction to
Matthias Kawski 1724
Topological Dynamics
Ethan Akin 1726
Vehicular Traffic: A Review of Continuum Mathematical Models
Benedetto Piccoli, Andrea Tosin 1748
Water Waves and the Korteweg–de Vries Equation
Lokenath Debnath 1771
List of Glossary Terms 1811
Index 1819

Copyrighted material

Contributors

Abou-Dina, Moustafa S.
Cairo University
Giza
Egypt

Akin, Ethan
The City College
New York City
USA

Aktosun, Tuncay
University of Texas at Arlington
Arlington
USA

Alabau-Boussouira, Fatiha
Université de Metz
Metz
France

Albano, Ezequiel V.
Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA) CCT La Plata
La Plata
Argentina

Alouges, François
Université Paris-Sud
Orsay cedex
France

Araújo, Vitor
CMUP
Porto
Portugal
IM-UFRJ
Rio de Janeiro
Brazil

Arneodo, Alain
ENS-Lyon CNRS
Lyon Cedex
France

Audit, Benjamin
ENS-Lyon CNRS
Lyon Cedex
France

Bambusi, Dario
Università degli Studi di Milano
Milano
Italia

ben-Avraham, Daniel
Clarkson University
Potsdam
USA

Benedetto, D.
Dipartimento di Matematica, Università di Roma 'La Sapienza'
Roma
Italy

Bracci, Luciano
Università di Pisa
Pisa
Italy
Sezione di Pisa
Pisa
Italy

Bressan, Alberto
Penn State University
University Park
USA

Brodie of Brodie, Edward-Benedict
ENS-Lyon CNRS
Lyon Cedex
France

Broer, Henk W.
University of Groningen
Groningen
The Netherlands

Copyrighted material

BULDYREV, SERGEY V.
Yeshiva University
New York
USA

BULLO, FRANCESCO
University of California
Santa Barbara
USA

BUNDE, ARMIN
Justus-Liebig-Universität
Giessen
Germany

BUZZI, JÉRÔME
C.N.R.S. and Université Paris-Sud
Orsay
France

CANNARSA, PIERMARCO
Università di Roma "Tor Vergata"
Rome
Italy

CELLETTI, ALESSANDRA
Università di Roma Tor Vergata
Roma
Italy

CHAMPNEYS, ALAN
University of Bristol
Bristol
United Kingdom

CHAZOTTES, JEAN-RENÉ
CNRS/École Polytechnique
Palaiseau
France

CHICONE, CARMEN
University of Missouri-Columbia
Columbia
USA

CHIERCHIA, LUIGI
Università "Roma Tre"
Roma
Italy

CLARKE, FRANCIS
Institut universitaire de France et Université de Lyon
Lyon
France

CORTÉS, JORGE
University of California
San Diego
USA

CRILLY, TONY
Middlesex University
London
UK

D'ANCONA, PIERO
Unversità di Roma "La Sapienza"
Roma
Italy

DANILENKO, ALEXANDRE I.
Ukrainian National Academy of Sciences
Kharkov
Ukraine

DA PRATO, GIUSEPPE
Scuola Normale Superiore
Pisa
Italy

D'AUBENTON-CARAFA, YVES
CNRS
Gif-sur-Yvette
France

DEBNATH, LOKENATH
University of Texas – Pan American
Edinburg
USA

DE LA RUE, THIERRY
CNRS – Université de Rouen
Saint Étienne du Rouvray
France

D'EMILIO, EMILIO
Università di Pisa
Pisa
Italy
Sezione di Pisa
Pisa
Italy

DERCOLE, FABIO
Politecnico di Milano
Milano
Italy

Copyrighted material

DERKS, GIANNE
University of Surrey
Guildford
UK

DESIMONE, ANTONIO
SISSA-International School for Advanced Studies
Trieste
Italy

DOLCETTA, ITALO CAPUZZO
Sapienza Universita' di Roma
Rome
Italy

DOMOTOR, ZOLTAN
University of Pennsylvania
Philadelphia
USA

FRANTZIKINAKIS, NIKOS
University of Memphis
Memphis
USA

GAETA, GIUSEPPE
Università di Milano
Milan
Italy

GALDI, GIOVANNI P.
University of Pittsburgh
Pittsburgh
USA

GALLAVOTTI, GIOVANNI
Università di Roma I "La Sapienza"
Roma
Italy

GALLOS, LAZAROS K.
City College of New York
New York
USA

GAUTHIER, JEAN-PAUL ANDRÉ
University of Toulon
Toulon
France

GENTILE, GUIDO
Università di Roma Tre
Roma
Italy

GLASS, OLIVIER
Université Pierre et Marie Curie
Paris
France

GOEBEL, RAFAL
Loyola University
Chicago
USA

GRAMCHEV, TODOR
Università di Cagliari
Cagliari
Italy

GRASMAN, JOHAN
Wageningen University and Research Centre
Wageningen
The Netherlands

HANSSMANN, HEINZ
Universiteit Utrecht
Utrecht
The Netherlands

HAVLIN, SHLOMO
Bar–Ilan University
Ramat–Gan
Israel

HE, JI-HUAN
Donghua University
Shanghai
China

HELAL, MOHAMED A.
Cairo University
Giza
Egypt

HEREMAN, WILLY
Colorado School of Mines
Golden
USA

HOFFMAN, CHRISTOPHER
University of Washington
Seattle
USA

HUVET, MAXIME
CNRS
Gif-sur-Yvette
France

Copyrighted material

INC, MUSTAFA
Fırat University
Elazığ
Turkey

ISIDORI, ALBERTO
University of Rome
La Sapienza
Italy

JUNCO, ANDRÉS DEL
University of Toronto
Toronto
Canada

JURDJEVIC, VELIMIR
University of Toronto
Toronto
Canada

KANTELHARDT, JAN W.
Martin-Luther-University Halle-Wittenberg
Halle
Germany

KAWSKI, MATTHIAS
Arizona State University
Tempe
USA

KAYA, DOĞAN
Firat University
Elazig
Turkey

KELLER, GERHARD
Universität Erlangen-Nürnberg
Erlangen
Germany

KING, JONATHAN L. F.
University of Florida
Gainesville
USA

KLEINBOCK, DMITRY
Brandeis University
Waltham
USA

KOIKE, SHIGEAKI
Saitama University
Saitama
Japan

KRA, BRYNA
Northwestern University
Evanston
USA

LAKSHMANAN, M.
Bharathidasan University
Tiruchirapalli
India

LEE, TAEYOUNG
University of Michigan
Ann Arbor
USA

LEFEBVRE, ALINE
Université Paris-Sud
Orsay cedex
France

LEMAŃCZYK, MARIUSZ
Nicolaus Copernicus University
Toruń
Poland

LEOK, MELVIN
Purdue University
West Lafayette
USA

MAKSE, HERNÁN A.
City College of New York
New York
USA

MARCONI, LORENZO
University of Bologna
Bologna
Italy

MARCUS, BRIAN
University of British Columbia
Vancouver
Canada

MARIN, FRANÇOIS
Laboratoire Ondes et Milieux Complexes, Fre CNRS 3102
Le Havre Cedex
France

MARSDEN, JERROLD E.
California Institute of Technology
Pasadena
USA

Copyrighted material

MARTÍNEZ, SONIA
University of California
San Diego
USA

MAWHIN, JEAN
Université Catholique de Louvain
Maryland
USA

MCCLAMROCH, HARRIS
University of Michigan
Ann Arbor
USA

MCCUTCHEON, RANDALL
University of Memphis
Memphis
USA

MEIJER, HIL
University of Twente
Enschede
The Netherlands

NAKAYAMA, TSUNEYOSHI
Toyota Physical and Chemical Research Institute
Nagakute
Japan

NICOLAY, SAMUEL
Université de Liège
Liège
Belgium

NICOL, MATTHEW
University of Houston
Houston
USA

NIEDERMAN, LAURENT
Université Paris
Paris
France
IMCCE
Paris
France

NIŢICĂ, VIOREL
West Chester University
West Chester
USA
Institute of Mathematics
Bucharest
Romania

NOTTALE, LAURENT
Paris Observatory and Paris Diderot University
Paris
France

OLDEMAN, BART
Concordia University
Montreal
Canada

ORBACH, RAYMOND L.
University of California
Riverside
USA

OSIPENKO, GEORGE
State Polytechnic University
St. Petersburg
Russia

PANATI, GIANLUCA
Università di Roma "La Sapienza"
Roma
Italy

PETERSEN, KARL
University of North Carolina
Chapel Hill
USA

PICASSO, LUIGI E.
Università di Pisa
Pisa
Italy
Sezione di Pisa
Pisa
Italy

PICCI, GIORGIO
University of Padua
Padua
Italy

PICCOLI, BENEDETTO
Consiglio Nazionale delle Ricerche
Rome
Italy

PULVIRENTI, M.
Dipartimento di Matematica, Università di Roma
'La Sapienza'
Roma
Italy

Copyrighted material

QUAS, ANTHONY
University of Victoria
Victoria
Canada

RATIU, TUDOR S.
École Polytechnique Fédérale de Lausanne
Lausanne
Switzerland

REDNER, SIDNEY
Boston University
Boston
USA

REN, YU-JIE
Shantou University
Shantou
People's Republic of China
Dalian Polytechnic University
Dalian
People's Republic of China
Beijing Institute of Applied Physics and Computational Mathematics
Beijing
People's Republic of China

ROSIER, LIONEL
Institut Elie Cartan
Vandoeuvre-lès-Nancy
France

ROZENFELD, HERNÁN D.
City College of New York
New York
USA

SACCHETTI, ANDREA
Universitá di Modena e Reggio Emilia
Modena
Italy

SÁNCHEZ, ANGEL
Universidad Carlos III de Madrid
Madrid
Spain
Universidad de Zaragoza
Zaragoza
Spain

SANDER, LEONARD M.
The University of Michigan
Ann Arbor
USA

SANFELICE, RICARDO G.
University of Arizona
Tucson
USA

SANNS, WERNER
University of Applied Sciences
Darmstadt
Germany

SBANO, LUCA
University of Warwick
Warwick
UK

SCHMELING, JÖRG
Lund University
Lund
Sweden

SICONOLFI, ANTONIO
"La Sapienza" Università di Roma
Roma
Italy

SILVA, CESAR E.
Williams College
Williamstown
USA

SOKOLOV, IGOR M.
Humboldt-Universität zu Berlin
Berlin
Germany

SONG, CHAOMING
City College of New York
New York
USA

SONTAG, EDUARDO D.
Rutgers University
New Brunswick
USA

SOOMERE, TARMO
Tallinn University of Technology
Tallinn
Estonia

STAUFFER, DIETRICH
Cologne University
Köln
Germany

Copyrighted material

STRELNIKER, YAKOV M.
Bar–Ilan University
Ramat–Gan
Israel

TAKAYASU, HIDEKI
Sony Computer Science Laboratories Inc
Tokyo
Japan

TAKAYASU, MISAKO
Tokyo Institute of Technology
Tokyo
Japan

TEEL, ANDREW R.
University of California
Santa Barbara
USA

TEIXEIRA, MARCO ANTÔNIO
Universidade Estadual de Campinas
Campinas
Brazil

TERRACINI, SUSANNA
Università di Milano Bicocca
Milano
Italia

THERMES, CLAUDE
CNRS
Gif-sur-Yvette
France

TOSIN, ANDREA
Consiglio Nazionale delle Ricerche
Rome
Italy

TOUCHON, MARIE
CNRS
Paris
France
Université Pierre et Marie Curie
Paris
France

TURCOTTE, DONALD L.
University of California
Davis
USA

VANDERBAUWHEDE, ANDRÉ
Ghent University
Gent
Belgium

VERHULST, FERDINAND
University of Utrecht
Utrecht
The Netherlands

VIANA, MARCELO
IMPA
Rio de Janeiro
Brazil

VIDYASAGAR, M.
Software Units Layout
Hyderabad
India

WALCHER, SEBASTIAN
RWTH Aachen
Aachen
Germany

WARD, TOM
University of East Anglia
Norwich
UK

WAZWAZ, ABDUL-MAJID
Saint Xavier University
Chicago
USA

WILKINSON, AMIE
Northwestern University
Evanston
USA

ZANG, YI
Zhejiang Normal University
Jinhua
China

ZHILINSKIÍ, BORIS
Université du Littoral
Dunkerque
France

ZHU, SHUN-DONG
Zhejiang Lishui University
Lishui
China

Copyrighted material

Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory

ABDUL-MAJID WAZWAZ
Department of Mathematics, Saint Xavier University, Chicago, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Adomian Decomposition Method and Adomian Polynomials
Modified Decomposition Method and Noise Terms Phenomenon
Solitons, Peakons, Kinks, and Compactons
Solitons of the KdV Equation
Kinks of the Burgers Equation
Peakons of the Camassa–Holm Equation
Compactons of the K(n,n) Equation
Future Directions
Bibliography

Glossary

Solitons Solitons appear as a result of a balance between a weakly nonlinear convection and a linear dispersion. The solitons are localized highly stable waves that retains its identity: shape and speed, upon interaction, and resemble particle like behavior. In the case of a collision, solitons undergo a phase shift.

Types of solitons *Solitary waves*, which are localized traveling waves, are asymptotically zero at large distances and appear in many structures such as solitons, kinks, peakons, cuspons, and compactons, among others. *Solitons* appear as a bell-shaped sech profile. *Kink waves* rise or descend from one asymptotic state to another. *Peakons* are peaked solitary-wave solutions. *Cuspons* exhibit cusps at their crests. In the peakon structure, the traveling-wave solutions are smooth except for a peak at a corner of its crest. Peakons are the points at which spatial derivative changes sign so that peakons have a finite jump in the first derivative of the solution $u(x, t)$. Unlike peakons, where the derivatives at the peak differ only by a sign, the derivatives at the jump of a cuspon diverge.

Compactons are solitons with compact spatial support such that each compacton is a soliton confined to a finite core or a soliton without exponential tails. Compactons are generated due to the delicate interaction between the effect of the genuine nonlinear convection and the genuinely nonlinear dispersion.

Adomian method The Adomian decomposition method approaches linear and nonlinear, and homogeneous and inhomogeneous differential and integral equations in a unified way. The method provides the solution in a rapidly convergent series with terms that are elegantly determined in a recursive manner. The method can be used to obtain closed-form solutions, if such solutions exist. A truncated number of the obtained series can be used for numerical purposes. The method was modified to accelerate the computational process. The noise terms phenomenon, which may appear for inhomogeneous cases, can give the exact solution in two iterations only.

Definition of the Subject

Nonlinear phenomena play a significant role in many branches of applied sciences such as applied mathematics, physics, biology, chemistry, astronomy, plasma, and fluid dynamics. Nonlinear dispersive equations that govern these phenomena have the genuine soliton property. Solitons are pulses that propagate without any change of its identity, i. e., shape and speed, during their travel through a nonlinear dispersive medium [1,5,34]. Solitons resemble properties of a particle, hence the suffix *on* is used [19,20].

Solitons exist in many scientific branches, such as optical fiber photonics, fiber lasers, plasmas, molecular systems, laser pulses propagating in solids, liquid crystals, nonlinear optics, cosmology, and condensed-matter physics. Based on its importance in many fields, a huge amount of research work has been conducted during the last four decades to make more progress in understanding the soliton phenomenon. A variety of very powerful algorithms has been used to achieve this goal. The Adomian decomposition method introduced in [2,6,21,22,23,24,25,26], which will be used in this work, is one of the reliable methods that has been used recently.

Introduction

The aim of this work is to apply the Adomian decomposition method to derive specific types of soliton solutions. Solitons were discovered *experimentally* by John Scott Russell in 1844. Korteweg and de Vries in 1895 investigated *analytically* the soliton concept [10], where they derived the pioneer equation of solitons, well known as the KdV equation, that models the height of the surface of shallow water in the presence of solitary waves. Moreover,

Copyrighted material

Zabusky and Kruskal [35] investigated this phenomenon *analytically* in 1965. Since 1965, a huge number of research works have been conducted on nonlinear dispersive and dissipative equations. The aim of these works has been to study thoroughly the characteristics of soliton solutions and to study various types of solitons that appear as a result of these equations. Several reliable methods were used in the literature to handle nonlinear dispersive equations. Hirota's bilinear method [8,9] has been used for single- and multiple-soliton solutions. The inverse scattering method [1] has been widely used. For single-soliton solutions, several methods, such as the tanh method [13,14,15], the pseudospectral method, and the truncated Painlevé expansion, are used.

However, in this work, the decomposition method, introduced by Adomian in 1984, will be applied to derive the desired types of soliton solutions. The method approaches all problems in a unified way and in a straightforward manner. The method computes the solution in a fast convergent series with components that are elegantly determined. Unlike other methods, the initial or boundary conditions are necessary for the use of the Adomian method.

Adomian Decomposition Method and Adomian Polynomials

The Adomian decomposition method, developed by George Adomian in 1984, has been receiving much attention in applied mathematics in general and in the area of initial value and boundary value problems in particular. Moreover, it is also used in the area of series solutions for numerical purposes. The method is powerful and effective and can be used in linear or nonlinear, ordinary or partial differential equations, and integral equations. The decomposition method demonstrates fast convergence of the solution and provides numerical approximation with a high level of accuracy. The method handles applied problems directly and in a straightforward manner without using linearization, perturbation, or any other restrictive assumption that might change the physical behavior of the physical model under investigation.

The method is effectively addressed and thoroughly used by many researchers in the literature [1,21,22,23,24,25,26]. It is important to indicate that well-known methods, such as Backlund transformation, inverse scattering method, Hirota's bilinear formalism, the tanh method, and many others, can handle problems without using initial or boundary value conditions. For the Adomian decomposition method, the initial or boundary conditions are necessary to conduct the determination of the components recursively. However, some of the standard methods require huge calculations, whereas the Adomian method minimizes the volume of computational work.

The Adomian decomposition method consists in decomposing the unknown function $u(x, t)$ of any equation into a sum of an infinite number of components given by the decomposition series

$$u(x, t) = \sum_{n=0}^{\infty} u_n(x, t) , \tag{1}$$

where the components $u_n(x, t), n \geq 0$ are to be determined in a recursive manner. The decomposition method concerns itself with the determination of the components $u_0(x, t), u_1(x, t), u_2(x, t), \ldots$ individually. The determination of these components can be obtained through a recursive relation that usually involves evaluation of simple integrals.

We now give a clear overview of Adomian decomposition method. Consider the linear differential equation written in an operator form by

$$Lu + Ru = f , \tag{2}$$

where L is the lower-order derivative, which is assumed to be invertible, R is a linear differential operator of order greater than L, and f is a source term. We next apply the inverse operator L^{-1} to both sides of Eq. (2) and use the given initial or boundary condition to get

$$u(x, t) = g - L^{-1}(Ru) , \tag{3}$$

where the function g represents the terms arising from integrating the source term f and from using the given conditions; all are assumed to be prescribed. Substituting the infinite series of components

$$u(x, t) = \sum_{n=0}^{\infty} u_n(x, t) \tag{4}$$

into both sides of (3) yields

$$\sum_{n=0}^{\infty} u_n = g - L^{-1}\left(R\left(\sum_{n=0}^{\infty} u_n\right)\right) . \tag{5}$$

The decomposition method suggests that the zeroth component u_0 is usually defined by all terms not included under the inverse operator L^{-1}, which arise from the initial data and from integrating the inhomogeneous term. This in turn gives the formal recursive relation

$$\begin{aligned} u_0(x, t) &= g , \\ u_{k+1}(x, t) &= -L^{-1}\left(R(u_k)\right) , \quad k \geq 0 , \end{aligned} \tag{6}$$

Copyrighted material

or, equivalently,

$$\begin{aligned} u_0(x,t) &= g\,, \\ u_1(x,t) &= -L^{-1}\left(R(u_0)\right)\,, \\ u_2(x,t) &= -L^{-1}\left(R(u_1)\right)\,, \\ u_3(x,t) &= -L^{-1}\left(R(u_2)\right)\,, \\ &\vdots\,. \end{aligned} \tag{7}$$

The differential equation under consideration is now reduced to integrals that can be easily evaluated. Having determined these components $u_0(x,t), u_1(x,t), u_2(x,t), \ldots$, we then substitute the obtained components into (4) to obtain the solution in a series form. The determined series may converge very rapidly to a closed-form solution, if an exact solution exists. For concrete problems, where a closed-form solution is not obtainable, a truncated number of terms is usually used for numerical purposes. Few terms of the truncated series give an approximation with a high degree of accuracy. The convergence concept of the decomposition series is investigated thoroughly in the literature.

Several significant studies were conducted to compare the performance of the Adomian method with other methods, such as Picard's method, Taylor series method, finite differences method, perturbation techniques, and others. The conclusions emphasized the fact that the Adomian method has many advantages and requires less computational work compared to existing techniques.

The Adomian method and many others applied this method to many deterministic and stochastic problems. However, the Adomian method, like some other methods, suffers if the zeroth component $u_0(x,t) = 0$ and makes the integrand of the right side in (7) $u_1(x,t) = 0$. If the right side integrand in (7) does not vanish, if it is of a form such as e^{u_0} or $\ln(\alpha + u_0), \alpha > 0$, then the method works effectively.

As stated before, the Adomian method decomposes the unknown function $u(x,t)$ into an infinite number of components. However, for nonlinear functions of $u(x,t)$ such as $u^2, u^3, \ln(1+u), \cos u, e^u, uu_x$, etc., a special representation for nonlinear terms was developed by Adomian and others. Adomian introduced a formal algorithm to establish the proper representation for all forms of nonlinear functions of $u(x,t)$. This representation of nonlinear terms is necessary to apply the Adomian method properly. Several alternative algorithms have been introduced in the literature by researchers to calculate Adomian polynomials. However, the Adomian algorithm remains the most commonly applied because it is simple and practical; therefore, it will be used in this work.

Adomian assumed that the nonlinear term $F(u)$ can be expressed by an infinite series of the so-called Adomian polynomials A_n given in the form

$$F(u) = \sum_{n=0}^{\infty} A_n(u_0, u_1, u_2, \ldots, u_n)\,, \tag{8}$$

where A_n can be evaluated for all forms of nonlinearity.

Adomian polynomials A_n for the nonlinear term $F(u)$ can be evaluated using the following expression:

$$A_n = \frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}\left[F\left(\sum_{i=0}^{n}\lambda^i u_i\right)\right]_{\lambda 0}\,, n = 0, 1, 2, \cdots. \tag{9}$$

The general formula (9) can be simplified as follows. Assuming that the nonlinear function is $F(u)$, using (9), Adomian polynomials are given by

$$\begin{aligned} A_0 &= F(u_0)\,, \\ A_1 &= u_1 F'(u_0)\,, \\ A_2 &= u_2 F'(u_0) + \frac{1}{2!}u_1^2 F''(u_0)\,, \\ A_3 &= u_3 F'(u_0) + u_1 u_2 F''(u_0) + \frac{1}{3!}u_1^3 F'''(u_0)\,, \\ A_4 &= u_4 F'(u_0) + \left(\frac{1}{2!}u_2^2 + u_1 u_3\right)F''(u_0) \\ &\quad + \frac{1}{2!}u_1^2 u_2 F'''(u_0) + \frac{1}{4!}u_1^4 F^{(iv)}(u_0)\,, \\ A_5 &= u_5 F'(u_0) + (u_2 u_3 + u_1 u_4)F''(u_0) \\ &\quad + \left(\frac{1}{2!}u_1 u_2^2 + \frac{1}{2!}u_1^2 u_3\right)F'''(u_0) \\ &\quad + \frac{1}{3!}u_1^3 u_2 F^{(iv)}(u_0) + \frac{1}{5!}u_1^5 F^{(v)}(u_0)\,. \end{aligned} \tag{10}$$

Other polynomials can be generated in a similar manner. It is clear that A_0 depends only on u_0, A_1 depends only on u_0 and u_1, A_2 depends only on u_0, u_1, and u_2, and so on.

For $F(u) = u^2$, we find

$$\begin{aligned} A_0 &= F(u_0) = u_0^2\,, \\ A_1 &= u_1 F'(u_0) = 2u_0 u_1\,, \\ A_2 &= u_2 F'(u_0) + \frac{1}{2!}u_1^2 F''(u_0) = 2u_0 u_2 + u_1^2\,, \\ A_3 &= u_3 F'(u_0) + u_1 u_2 F''(u_0) + \frac{1}{3!}u_1^3 F'''(u_0) \\ &= 2u_0 u_3 + 2u_1 u_2\,. \end{aligned} \tag{11}$$

Modified Decomposition Method and Noise Terms Phenomenon

A reliable modification of the Adomian decomposition method [22] was developed by Wazwaz in 1999. The modification will further accelerate the convergence of the se-

Copyrighted material

ries solution. As presented earlier, the standard decomposition method admits the use of the recursive relation

$$\begin{aligned} u_0 &= g\,, \\ u_{k+1} &= -L^{-1}(Ru_k)\,, \quad k \geq 0\,. \end{aligned} \tag{12}$$

The modified decomposition method introduces a slight variation to the recursive relation (12) that will lead to the determination of the components of u in a faster and easier way. For specific cases, the function g can be set as the sum of two partial functions, g_1 and g_2. In other words, we can set

$$g = g_1 + g_2\,. \tag{13}$$

This assumption gives a slight qualitative change in the formation of the recursive relation (12). To reduce the size of calculations, we identify the zeroth component u_0 by one part of g, that is, g_1 or g_2. The other part of g can be added to the component u_1 among other terms. In other words, the modified recursive relation can be defined as

$$\begin{aligned} u_0 &= g_1\,, \\ u_1 &= g_2 - L^{-1}(Ru_0)\,, \\ u_{k+1} &= -L^{-1}(Ru_k)\,, \quad k \geq 1\,. \end{aligned} \tag{14}$$

The change occurred in the formation of the first two components u_0 and u_1 only [22,24,25]. Although this variation in the formation of u_0 and u_1 is slight, it plays a major role in accelerating the convergence of the solution and in minimizing the size of calculations. It is interesting to point out that by selecting the parts g_1 and g_2 properly, the exact solution $u(x, t)$ may be obtained by using very few iterations, and sometimes by evaluating only two components. Moreover, if g consists of one term only, the standard decomposition method should be employed in this case.

Another useful feature of the Adomian method is the noise terms phenomenon. The noise terms may appear for inhomogeneous problems only. This phenomenon was addressed by Adomian in 1994. In 1997, Wazwaz investigated the necessary conditions for the appearance of noise terms in the decomposition series. **Noise terms** are defined as identical terms with opposite signs [21] that arise in the components u_0 and u_1 particularly, and in other components as well. By canceling the noise terms between u_0 and u_1, even though u_1 contains other terms, the remaining noncanceled terms of u_0 may give the exact solution of the equation. Therefore, it is necessary to verify that the noncanceled terms of u_0 satisfy the equation. The noise terms, if they exist in the components u_0 and u_1, will provide the solution in a closed form with only two successive iterations.

It was formally proved by Wazwaz in 1997 that a necessary condition for the appearance of the noise terms is required. The conclusion made is that the zeroth component u_0 must contain the exact solution u, among other terms [21]. Moreover, it was shown that the nonhomogeneity condition does not always guarantee the appearance of the noise terms.

Solitons, Peakons, Kinks, and Compactons

There are many types of *solitary waves*. *Solitons*, which are localized traveling waves, are asymptotically zero at large distances [1,5,7,21,22,23,24,25,26,27,28,29,30,31,32,33]. Solitons appear as a bell-shaped sech profile. Soliton solution $u(\xi)$ results as a balance between nonlinearity and dispersion, where $u(\xi), u'(\xi), u''(\xi), \cdots \to 0$ as $\xi \to \pm\infty$, where $\xi = x - ct$, and c is the speed of the wave prorogation. The soliton solution either decays exponentially as in the KdV equation, or it converges to a constant at infinity such as the kinks of the sine-Gordon equation. This means that the soliton solutions appear as sech^{α} or $\arctan(e^{\alpha(x-ct)})$. Moreover, one soliton interacts with other solitons preserving its permanent form.

Another type of solitary wave is the *kink wave*, which rises or descends from one asymptotic state to another [18]. The Burgers equation and the sine-Gordon equation are examples of nonlinear wave equations that exhibit kink solutions. It is to be noted that the kink $u(\xi)$ converges to $\pm\alpha$, where α is a constant. However, $u'(\xi), u''(\xi), \cdots \to 0$ as $\xi \to \pm\infty$.

Peakons are peaked solitary-wave solutions and are another type of solitary-wave solution. In this structure, the traveling wave solutions are smooth except for a peak at a corner of its crest. Peakons are the points at which a spatial derivative changes signs so that peakons have a finite jump in the first derivative of the solution $u(x, t)$ [3,4,11,12,32]. This means that the first derivative of $u(x, t)$ has identical values with opposite signs around the peak.

A significant type of soliton is the *compacton*, which is a soliton with compact spatial support such that each compacton is a soliton confined to a finite core or a soliton without exponential tails [16,17]. Compactons were formally derived by Rosenau and Hyman in 1993 where a special kind of the KdV equation was used to derive this new discovery. Unlike a soliton that narrows as the amplitude increases, a compacton's width is independent of the amplitude [16,17,27,28,29,30]. Classical solitons are analytic solutions, whereas compactons are nonanalytic solutions [16,17]. As will be shown by a graph below, compactons are solitons that are free of exponential wings or

Copyrighted material

Using the decomposition identification for the linear and nonlinear terms in (35) yields

$$\sum_{n=0}^{\infty} u_n(x,t) = \frac{2c}{(1+\mathrm{e}^{\frac{cx}{\nu}})} + L_t^{-1}\left(\nu \sum_{n=0}^{\infty} u_n(x,t)\right)_{xx} - L_t^{-1}\left(\sum_{n=0}^{\infty} A_n\right). \quad (36)$$

The Adomian method allows for the use of the recursive relation

$$\begin{aligned} u_0(x,t) &= \frac{2c}{(1+\mathrm{e}^{\frac{cx}{n}})}, \\ u_{k+1}(x,t) &= L_t^{-1}\left(\nu u_{k_{xxx}}\right) - L_t^{-1}(A_k), \quad k \geq 0. \end{aligned} \quad (37)$$

The components u_n, $n \geq 0$ can be elegantly calculated by

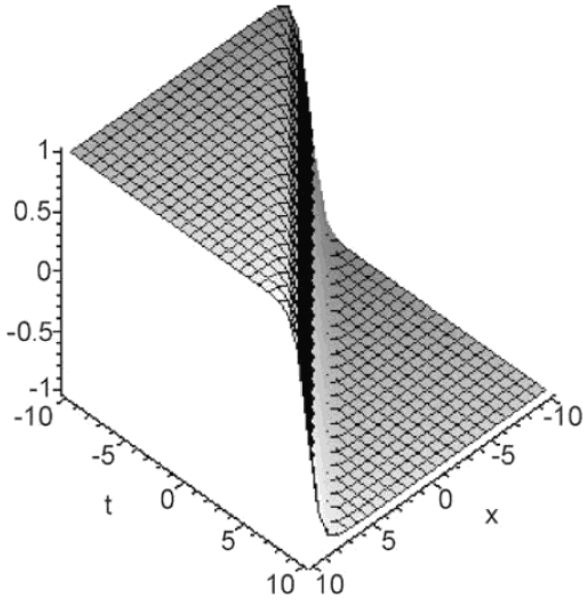
$$\begin{aligned} u_0(x,t) &= \frac{2c}{(1+\mathrm{e}^{\frac{cx}{\nu}})}, \\ u_1(x,t) &= L_t^{-1}\left(\nu u_{0_{xxx}}\right) - L_t^{-1}(A_0) \\ &= \frac{2c^3\mathrm{e}^{\frac{cx}{\nu}}}{\nu(1+\mathrm{e}^{\frac{cx}{\nu}})^2}\, t, \\ u_2(x,t) &= L_t^{-1}\left(\nu u_{1_{xxx}}\right) - L_t^{-1}(A_1) \\ &= \frac{c^5\mathrm{e}^{\frac{cx}{\nu}}(\mathrm{e}^{\frac{cx}{\nu}}-1)}{\nu^2(1+\mathrm{e}^{\frac{cx}{\nu}})^3}\, t^2, \\ u_3(x,t) &= L_t^{-1}\left(\nu u_{2_{xxx}}\right) - L_t^{-1}(A_2) \\ &= \frac{c^7\mathrm{e}^{\frac{cx}{\nu}}(\mathrm{e}^{\frac{2cx}{\nu}} - 4\mathrm{e}^{\frac{cx}{\nu}} + 1)}{3\nu^3(1+\mathrm{e}^{\frac{cx}{\nu}})^4}\, t^3, \\ u_4(x,t) &= L_t^{-1}\left(\nu u_{3_{xxx}}\right) - L_t^{-1}(A_3) \\ &= \frac{c^9\mathrm{e}^{\frac{cx}{\nu}}(\mathrm{e}^{\frac{3cx}{\nu}} - 11\mathrm{e}^{\frac{2cx}{\nu}} + 11\mathrm{e}^{\frac{cx}{\nu}} - 1)}{12\nu^4(1+\mathrm{e}^{\frac{cx}{\nu}})^5}\, t^4, \end{aligned} \quad (38)$$

and so on. The series solution is thus given by

$$\begin{aligned} u(x,t) = {} & \frac{2c\mathrm{e}^{\frac{cx}{\nu}}}{(1+\mathrm{e}^{\frac{cx}{\nu}})} \\ & \times \left(\mathrm{e}^{-\frac{cx}{\nu}} + \frac{c^2}{\nu(1+\mathrm{e}^{\frac{cx}{\nu}})}\, t + \frac{c^4(\mathrm{e}^{\frac{cx}{\nu}}-1)}{2\nu^2(1+\mathrm{e}^{\frac{cx}{\nu}})^2}\, t^2 \right. \\ & \left. + \frac{c^6(\mathrm{e}^{\frac{2cx}{\nu}} - 4\mathrm{e}^{\frac{cx}{\nu}} + 1)}{6\nu^3(1+\mathrm{e}^{\frac{cx}{\nu}})^3}\, t^3 + \cdots\right), \end{aligned} \quad (39)$$

and in a closed form by

$$u(x,t) = \frac{2c}{\left(1+\mathrm{e}^{\frac{c}{\nu}(x-ct)}\right)}, \quad \nu \neq 0, \quad (40)$$



Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory, Figure 2

Figure 2 shows the kink graph $u(x,t) = \tanh(x - ct)$, $c = 1$, $-10 \leq x$, $t \leq 10$

or equivalently

$$u(x,t) = c\left(1 - \tanh\left[\frac{c}{2\nu}(x-ct)\right]\right). \quad (41)$$

Figure 2 shows a kink graph.

The graph shows that the kink converges to ± 1 as $\xi \to \pm\infty$.

Peakons of the Camassa–Holm Equation

Camassa and Holm [4] derived in 1993 a completely integrable wave equation

$$u_t + 2ku_x - u_{xxt} + 3uu_x = 2u_x u_{xx} + uu_{xxx} \quad (42)$$

by retaining two terms that are usually neglected in the small-amplitude shallow-water limit [3]. The constant k is related to the critical shallow-water wave speed. This equation models the unidirectional propagation of water waves in shallow water. The CH Eq. (42) differs from the well-known regularized long wave (RLW) equation only through the two terms on the right-hand side of (42). Moreover, this equation has an integrable bi-Hamiltonian structure and arises in the context of differential geometry, where it can be seen as a reexpression for geodesic flow on an infinite-dimensional Lie group. The CH equation admits a second-order isospectral problem and allows for peaked solitary-wave solutions, called peakons.

Copyrighted material

In this work, we will apply the Adomian method to the CH equation

$$u_t - u_{xxt} + 3uu_x = 2u_x u_{xx} + uu_{xxx}, \quad u(x,0) = ce^{-|x|}, \tag{43}$$

or equivalently

$$u_t = u_{xxt} - 3uu_x + 2u_x u_{xx} + uu_{xxx}, \quad u(x,0) = ce^{-|x|}, \tag{44}$$

where we set $k = 0$.

Proceeding as before, the CH Eq. (44) in an operator form is as follows:

$$L_t u = u_{xxt} - 3uu_x + 2u_x u_{xx} + uu_{xxx}, \quad u(x,0) = ce^{-|x|}, \tag{45}$$

where the differential operator L_t is as defined above. It is important to point out that the CH equation includes three nonlinear terms; therefore, we derive the following three sets of Adomian polynomials for the terms $uu_x, u_x u_{xx}$, and uu_{xxx}:

$$\begin{aligned} A_0 &= u_0 u_{0_x}, \\ A_1 &= u_{0_x} u_1 + u_0 u_{1_x}, \\ A_2 &= u_{0_x} u_2 + u_{1_x} u_1 + u_{2_x} u_0, \end{aligned} \tag{46}$$

$$\begin{aligned} B_0 &= u_{0_x} u_{0_{xx}}, \\ B_1 &= u_{0_{xx}} u_{1_x} + u_{0_x} u_{1_{xx}}, \\ B_2 &= u_{0_{xx}} u_{2_x} + u_{1_x} u_{1_{xx}} + u_{0_x} u_{2_{xx}}, \end{aligned} \tag{47}$$

and

$$\begin{aligned} C_0 &= u_0 u_{0_{xxx}}, \\ C_1 &= u_{0_{xxx}} u_1 + u_0 u_{1_{xxx}}, \\ C_2 &= u_{0_{xxx}} u_2 + u_1 u_{1_{xxx}} + u_0 u_{2_{xxx}}. \end{aligned} \tag{48}$$

Applying the inverse operator L_t^{-1} to both sides of (45) and using the initial condition we find

$$u(x,t) = ce^{-|x|} + L_t^{-1}\left(u_{xxt} - 3uu_x + 2u_x u_{xx} + uu_{xxx}\right). \tag{49}$$

Case 1. For $x > 0$, the initial condition will be $u(x,0) = e^{-x}$. Using the decomposition series (1) and Adomian polynomials in (49) yields

$$\sum_{n=0}^{\infty} u_n(x,t) = ce^{-x} + L_t^{-1}\left(\sum_{n=0}^{\infty} u_n(x,t)\right)_{xxt} + L_t^{-1}\left(-3\sum_{n=0}^{\infty} A_n + 2\sum_{n=0}^{\infty} B_n + \sum_{n=0}^{\infty} C_n\right). \tag{50}$$

The Adomian scheme allows the use of the recursive relation

$$\begin{aligned} u_0(x,t) &= ce^{-x}, \\ u_{k+1}(x,t) &= L_t^{-1}\left(u_k(x,t)\right)_{xxt} \\ &\quad + L_t^{-1}\left(-3A_k + 2B_k + C_k\right), \quad k \geq 0. \end{aligned} \tag{51}$$

The components $u_n, n \geq 0$ can be elegantly computed by

$$\begin{aligned} u_0(x,t) &= ce^{-x}, \\ u_1(x,t) &= L_t^{-1}\left(u_0(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_0 + 2B_0 + C_0\right) \\ &= c^2 e^{-x} t, \\ u_2(x,t) &= L_t^{-1}\left(u_1(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_1 + 2B_1 + C_1\right) \\ &= \frac{1}{2!} c^3 e^{-x} t^2, \\ u_3(x,t) &= L_t^{-1}\left(u_2(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_2 + 2B_2 + C_2\right) \\ &= \frac{1}{3!} c^4 e^{-x} t^3, \\ u_4(x,t) &= L_t^{-1}\left(u_3(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_3 + 2B_3 + C_3\right) \\ &= \frac{1}{4!} c^4 e^{-x} t^4, \end{aligned} \tag{52}$$

and so on. The series solution for $x > 0$ is given by

$$u(x,t) = ce^{-x} \times \left(1 + ct + \frac{1}{2!}(ct)^2 + \frac{1}{3!}(ct)^3 + \frac{1}{4!}(ct)^4 + \cdots\right), \tag{53}$$

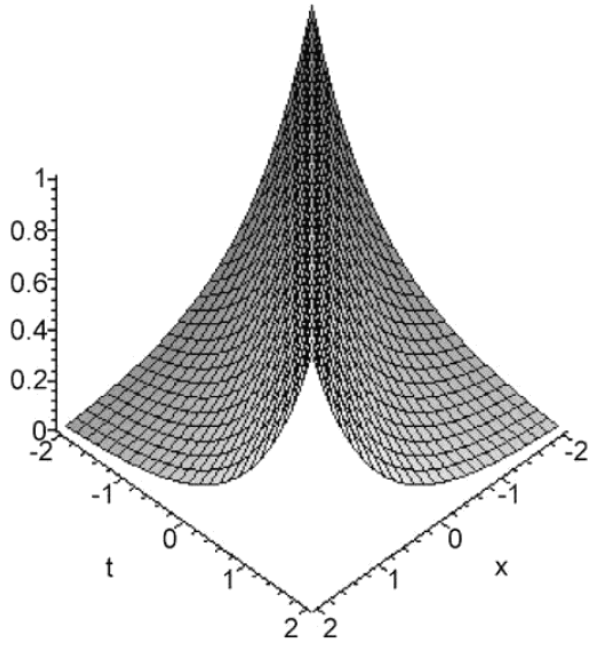
and in a closed form by

$$u(x,t) = ce^{-(x-ct)}. \tag{54}$$

Case 2. For $x < 0$, the initial condition will be $u(x,0) = e^{x}$. Proceeding as before, we obtain

$$\begin{aligned} u_0(x,t) &= ce^{x}, \\ u_1(x,t) &= L_t^{-1}\left(u_0(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_0 + 2B_0 + C_0\right) \\ &= -c^2 e^{x} t, \\ u_2(x,t) &= L_t^{-1}\left(u_1(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_1 + 2B_1 + C_1\right) \\ &= \frac{1}{2!} c^3 e^{x} t^2, \\ u_3(x,t) &= L_t^{-1}\left(u_2(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_2 + 2B_2 + C_2\right) \\ &= -\frac{1}{3!} c^4 e^{x} t^3, \\ u_4(x,t) &= L_t^{-1}\left(u_3(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_3 + 2B_3 + C_3\right) \\ &= \frac{1}{4!} c^4 e^{x} t^4, \end{aligned} \tag{55}$$

Copyrighted material



Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory, Figure 3
Figure 3 shows the peakon graph $u(x,t) = e^{-|(x-ct)|}$, $c = 1$, $-2 \le x,\ t \le 2$

and so on. The series solution for $x < 0$ is given by

$$u(x,t) = ce^{x} \times \left(1 - ct + \frac{1}{2!}(ct)^2 - \frac{1}{3!}(ct)^3 + \frac{1}{4!}(ct)^4 + \cdots\right), \tag{56}$$

and in a closed form by

$$u(x,t) = ce^{(x-ct)}. \tag{57}$$

Combining the results for both cases gives the peakon solution

$$u(x,t) = ce^{-|x-ct|}. \tag{58}$$

Figure 3 shows a peakon graph.

The graph shows that a peakon with a peak at a corner is generated with equal first derivatives at both sides of the peak, but with opposite signs.

Compactons of the K(n,n) Equation

The K(n,n) equation [16,17] was introduced by Rosenau and Hyman in 1993. This equation was investigated experimentally and analytically. The K(m,n) equation is a genuinely nonlinear dispersive equation, a special type of the KdV equation, named K(m,n), of the form

$$u_t + (u^m)_x + (u^n)_{xxx} = 0, \quad n > 1. \tag{59}$$

Compactons, which are solitons with compact support or strict localization of solitary waves, have been investigated thoroughly in the literature. The delicate interaction between the effect of the genuine nonlinear convection $(u^n)_x$ and the genuinely nonlinear dispersion of $(u^n)_{xxx}$ generates solitary waves with exact compact support that are called *compactons*. It was also discovered that solitary waves may compactify under the influence of nonlinear dispersion, which is capable of causing deep qualitative changes in the nature of genuinely nonlinear phenomena [16,17]. Unlike solitons that narrow as the amplitude increases, a compacton's width is independent of the amplitude. Compactons such as drops do not possess infinite wings; hence they interact among themselves only across short distances.

Compactons are nonanalytic solutions, whereas classical solitons are analytic solutions. The points of nonanalyticity at the edge of a compacton correspond to points of genuine nonlinearity for the differential equation and introduce singularities in the associated dynamical system for the traveling waves [16,17,27,28,29,30]. Compactons were proved to collide elastically and vanish outside a finite core region. This discovery was studied thoroughly by many researchers who were involved with identical nonlinear dispersive equations. It is to be noted that solutions were obtained only for cases where $m = n$. However, for $m \neq n$, solutions have not yet been determined.

Without loss of generality, we will examine the special K(3,3) initial value problem

$$u_t + (u^3)_x + (u^3)_{xxx} = 0, \quad u(x,0) = \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right). \tag{60}$$

We first write the K(3,3) Eq. (60) in an operator form:

$$L_t u = -(u^3)_x - (u^3)_{xxx}, \quad u(x,0) = \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right), \tag{61}$$

where the differential operator L_t is as defined above. The differential operator L_t is defined by

$$L_t = \frac{\partial}{\partial t}. \tag{62}$$

Applying L_t^{-1} to both sides of (61) and using the initial condition we find

$$u(x,t) = \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right) - L_t^{-1}\left((u^3)_x + (u^3)_{xxx}\right). \tag{63}$$

Copyrighted material

Notice that the right-hand side contains two nonlinear terms $(u^3)_x$ and $(u^3)_{xxx}$. Accordingly, the Adomian polynomials for these terms are given by

$$\begin{aligned} A_0 &= (u_0^3)_x \,, \\ A_1 &= (3u_0^2u_1)_x \,, \\ A_2 &= (3u_0^2u_2 + 3u_0u_1^2)_x \,, \\ A_3 &= (3u_0^2u_3 + 6u_0u_1u_2 + u_1^3)_x \,, \end{aligned} \tag{64}$$

and

$$\begin{aligned} B_0 &= (u_0^3)_{xxx} \,, \\ B_1 &= (3u_0^2u_1)_{xxx} \,, \\ B_2 &= (3u_0^2u_2 + 3u_0u_1^2)_{xxx} \,, \\ B_3 &= (3u_0^2u_3 + 6u_0u_1u_2 + u_1^3)_{xxx} \,, \end{aligned} \tag{65}$$

respectively.

Proceeding as before, Eq. (63) becomes

$$\sum_{n=0}^{\infty} u_n(x,t) = \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right) - L_t^{-1}\left(\sum_{n=0}^{\infty} A_n + B_n\right). \tag{66}$$

This gives the recursive relation

$$\begin{aligned} u_0(x,t) &= \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right) \,, \\ u_{k+1}(x,t) &= -L_t^{-1}\left(L_t^{-1}(A_k + B_k)\right) \,, \quad k \geq 0 \,. \end{aligned} \tag{67}$$

The components $u_n, n \geq 0$ can be recursively determined as

$$\begin{aligned} u_0(x,t) &= \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right) \,, \\ u_1(x,t) &= \frac{c\sqrt{6c}}{6}\sin\left(\frac{1}{3}x\right) t \,, \\ u_2(x,t) &= -\frac{c^2\sqrt{6c}}{36}\cos\left(\frac{1}{3}x\right) t^2 \,, \\ u_3(x,t) &= -\frac{c^3\sqrt{6c}}{324}\sin\left(\frac{1}{3}x\right) t^3 \,, \\ u_4(x,t) &= \frac{c^4\sqrt{6c}}{3888}\cos\left(\frac{1}{3}x\right) t^4 \,, \\ u_5(x,t) &= \frac{c^5\sqrt{6c}}{58320}\sin\left(\frac{1}{3}x\right) t^5 \,, \end{aligned} \tag{68}$$

and so on. The series solution is thus given by

$$\begin{aligned} &u(x,t) \\ &= \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right)\left(1 - \frac{1}{2!}\left(\frac{ct}{3}\right)^2 + \frac{1}{4!}\left(\frac{ct}{3}\right)^4 + \cdots\right) \\ &\quad + \frac{\sqrt{6c}}{2}\sin\left(\frac{1}{3}x\right)\left(\frac{ct}{3} - \frac{1}{3!}\left(\frac{ct}{3}\right)^3 + \frac{1}{5!}\left(\frac{ct}{5}\right)^5 + \cdots\right). \end{aligned} \tag{69}$$

This is equivalent to

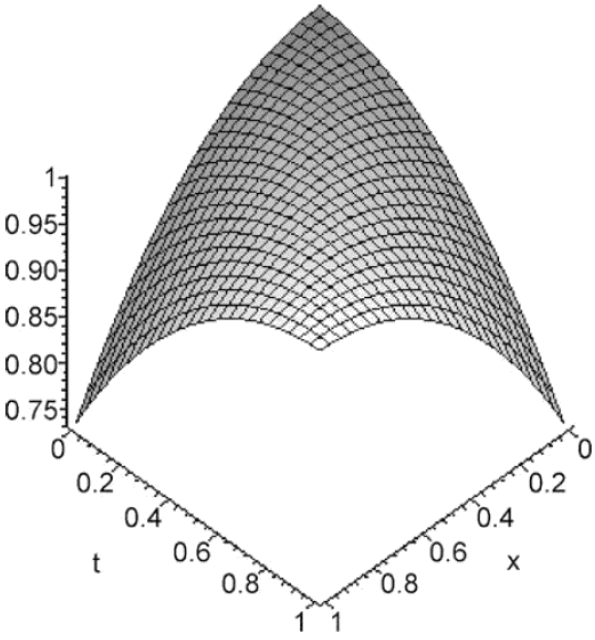
$$\begin{aligned} &u(x,t) \\ &= \frac{\sqrt{6c}}{2}\left(\cos\left(\frac{1}{3}x\right)\cos\left(\frac{1}{3}ct\right) + \sin\left(\frac{1}{3}x\right)\sin\left(\frac{1}{3}ct\right)\right), \end{aligned} \tag{70}$$

or equivalently

$$u(x,t) = \begin{cases} \left\{\frac{\sqrt{6c}}{2}\cos\left[\frac{1}{3}(x-ct)\right]\right\}, & |\, x - ct \,| \leq \frac{3\pi}{2} \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{71}$$

This in turn gives the compacton solution of the K(3,3) equation. Figure 4 shows a compacton confined to a finite core without exponential wings.

The graph shows a compacton: a soliton free of exponential wings.



Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory, Figure 4

Figure 4 shows the compacton graph $u(x,t) = \cos^{\frac{1}{2}}(x - ct)$, $c = 1$, $0 \leq x, t \leq 1$

Copyrighted material

It is interesting to point out that the generalized solution of the K(n,n) equation, $n > 1$, is given by the compactons

$$u(x,t) = \begin{cases} \left\{ \frac{2cn}{(n+1)} \cos^2 \left[\frac{n-1}{2n}(x-ct) \right] \right\}^{\frac{1}{n-1}}, & |x-ct| \le \frac{\pi}{2\mu}, \\ 0 & \text{otherwise}, \end{cases} \tag{72}$$

and

$$u_3(x,t) = \begin{cases} \left\{ \frac{2cn}{(n+1)} \sin^2 \left[\frac{n-1}{2n}(x-ct) \right] \right\}^{\frac{1}{n-1}}, & |x-ct| \le \frac{\pi}{\mu}, \\ 0 & \text{otherwise}. \end{cases} \tag{73}$$

Future Directions

The most significant advantage of the Adomian method is that it attacks any problem without any need for a transformation formula or any restrictive assumption that may change the physical behavior of the solution. For singular problems, it was possible to overcome the singularity phenomenon and to attain practically a series solution in a standard way. Moreover, for problems on unbounded domains, combining the obtained series solution by the Adomian method with Padé approximants provides a promising tool to handle boundary value problems. The Padé approximants, which often show superior performance over series approximations, provide a promising tool for use in applied fields.

As stated above, unlike other methods such as Hirota and the inverse scattering method, where solutions can be obtained without using prescribed conditions, the Adomian method requires such conditions. Moreover, these conditions must be of a form that does not provide zero for the integrand of the first component $u_1(x,t)$. Such a case should be addressed to enhance the performance of the method. Another aspect that should be addressed is Adomian polynomials. The existing techniques require tedious work to evaluate it. Most importantly, the Adomian method guarantees only one solution for nonlinear problems. This is an issue that should be investigated to improve the performance of the Adomian method compared to other methods.

On the other hand, it is important to further examine the N-soliton solutions by simplified forms, such as the form given by Hereman and Nuseir in [7]. The bilinear form of Hirota is not that easy to use and is not always attainable for nonlinear models.

Bibliography

Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolution equations and inverse scattering. Cambridge University Press, Cambridge
2. Adomian G (1984) A new approach to nonlinear partial differential equations. J Math Anal Appl 102:420–434
3. Boyd PJ (1997) Peakons and cashoidal waves: travelling wave solutions of the Camassa–Holm equation. Appl Math Comput 81(2–3):173–187
4. Camassa R, Holm D (1993) An integrable shallow water equation with peaked solitons. Phys Rev Lett 71(11):1661–1664
5. Drazin PG, Johnson RS (1996) Solitons: an introduction. Cambridge University Press, Cambridge
6. Helal MA, Mehanna MS (2006) A comparison between two different methods for solving KdV-Burgers equation. Chaos Solitons Fractals 28:320–326
7. Hereman W, Nuseir A (1997) Symbolic methods to construct exact solutions of nonlinear partial differential equations. Math Comp Simul 43:13–27
8. Hirota R (1971) Exact solutions of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27(18):1192–1194
9. Hirota R (1972) Exact solutions of the modified Korteweg–de Vries equation for multiple collisions of solitons. J Phys Soc Jpn 33(5):1456–1458
10. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary waves. Philos Mag 5th Ser 36:422–443
11. Lenells J (2005) Travelling wave solutions of the Camassa–Holm equation. J Differ Equ 217:393–430
12. Liu Z, Wang R, Jing Z (2004) Peaked wave solutions of Camassa–Holm equation. Chaos Solitons Fractals 19:77–92
13. Malfliet W (1992) Solitary wave solutions of nonlinear wave equations. Am J Phys 60(7):650–654
14. Malfliet W, Hereman W (1996) The tanh method: I. Exact solutions of nonlinear evolution and wave equations. Phys Scr 54:563–568
15. Malfliet W, Hereman W (1996) The tanh method: II. Perturbation technique for conservative systems. Phys Scr 54:569–575
16. Rosenau P (1998) On a class of nonlinear dispersive-dissipative interactions. Phys D 230(5/6):535–546
17. Rosenau P, Hyman J (1993) Compactons: solitons with finite wavelengths. Phys Rev Lett 70(5):564–567
18. Veksler A, Zarmi Y (2005) Wave interactions and the analysis of the perturbed Burgers equation. Phys D 211:57–73
19. Wadati M (1972) The exact solution of the modified Korteweg–de Vries equation. J Phys Soc Jpn 32:1681–1687
20. Wadati M (2001) Introduction to solitons. Pramana J Phys 57(5/6):841–847
21. Wazwaz AM (1997) Necessary conditions for the appearance of noise terms in decomposition solution series. Appl Math Comput 81:265–274
22. Wazwaz AM (1998) A reliable modification of Adomian's decomposition method. Appl Math Comput 92:1–7
23. Wazwaz AM (1999) A comparison between the Adomian decomposition method and the Taylor series method in the series solutions. Appl Math Comput 102:77–86

Copyrighted material

from the linear dependence of the mean squared displacement on time. Many other properties, like return probabilities, mean number of different visited nodes etc. also behave in a way different from the one they behave in normal diffusion. Understanding these properties is of primary importance for theoretical description of transport in such systems, chemical reactions in them, and in devising efficient search algorithms or in understanding ones used in natural search, as taking place in gene expression or incorporated into the animals' foraging strategies.

Introduction

The theory of diffusion processes was initiated by A. Fick who was interested in the nutrients' transport in a living body and put down the phenomenological diffusion equation (1855). The microscopic, probabilistic discussion of diffusion started with the works of L. Bachelier on the theory of financial speculation (1900), of A. Einstein on the motion of colloidal particles in a quiescent fluid (the Brownian motion) (1905), and with a short letter of K. Pearson to the readers of Nature in the same year motivated by his work on animal motion. These works put forward the description of diffusion processes within the framework of random walk models, the ones describing the motion as a sequence of independent bounded steps in space requiring some bounded time to be completed. These models lead to larger scales and longer times to a typical diffusion behavior as described by Fick's laws. The most prominent property of this so-called normal diffusion is the fact that the mean squared displacement of the diffusing particle from its initial position at times much longer than the typical time of one step grows as $\langle x^2(t)\rangle \propto t$. Another prominent property of normal diffusion is the fact that the distribution of the particle's displacements tends to a Gaussian (of width $\langle x^2(t)\rangle^{1/2}$), as a consequence of the independence of different steps.

The situations described by normal diffusion are abundant, but looking closer into many other processes (which still can be described within a random walk picture, however, with only more or less independent steps, or with steps showing a broad distribution of their lengths or times) revealed that these processes, still resembling or related to normal diffusion, show a vastly different behavior of the mean squared displacement, e. g. $\langle x^2(t)\rangle \propto t^\alpha$ with $\alpha \neq 1$. In this case one speaks about anomalous diffusion [11]. The cases with $\alpha < 1$ correspond to *subdiffusion*, the ones with $\alpha > 1$ correspond to *superdiffusion*. The subdiffusion is typical for the motion of charge carriers in disordered semiconductors, for contaminant transport by underground water or for proteins' motion in cell membranes. The models of subdiffusion are often connected either with the continuous-time random walks or with diffusion in fractal networks (the topic of the present article) or with combinations thereof. The superdiffusive behavior, associated with divergent mean square displacement per complete step, is encountered e. g. in transport in some Hamiltonian models and in maps, in animal motion, in transport of particles by flows or in diffusion in porous media in Knudsen regime.

The diffusion on fractal networks (typically subdiffusion) was a topic of extensive investigation in 1980s–1990s, so that the most results discussed here can be considered as well-established. Therefore most of the references in this contribution are given to the review articles, and only few of them to original works (mostly to the pioneering publications or to newer works which didn't find their place in reviews).

Random Walks and Normal Diffusion

Let us first turn to normal diffusion and concentrate on the mean squared displacement of the random walker. Our model will correspond to a random walk on a regular lattice or in free space, where the steps of the random walker will be considered independent, identically distributed random variables (step lengths $\mathbf{s}_i$) following a distribution with a given symmetric probability density function (PDF) $p(\mathbf{s})$, so that $\langle \mathbf{s}\rangle = 0$. Let a be the root mean square displacement in one step, $\langle \mathbf{s}_i^2\rangle = a^2$. The squared displacement of the walker after n steps is thus

$$\langle r_n^2\rangle = \left\langle \left(\sum_{i=1}^{n} \mathbf{s}_i\right)^2 \right\rangle = \sum_{i=1}^{n} \langle \mathbf{s}_i^2\rangle + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \langle \mathbf{s}_i \mathbf{s}_j\rangle \,.$$

The first sum is simply na^2, and the second one vanishes if different steps are uncorrelated and have zero mean. Therefore $\langle r_n^2\rangle = a^2 n$: The mean squared displacement in a walk is proportional to the number of steps.

Provided the mean time τ necessary to complete a step exists, this expression can be translated into the temporal dependence of the displacement $\langle r^2(t)\rangle = (a^2/\tau)t$, since the mean number of steps done during the time t is $n = t/\tau$. This is the time-dependence of the mean squared displacement of a random walker characteristic for normal diffusion. The prefactor a^2/τ in the $\langle r^2(t)\rangle \propto t$ dependence is connected with the usual diffusion coefficient K, defined through $\langle r^2(t)\rangle = 2dKt$, where d is the dimension of the Euclidean space. Here we note that in the theory of random walks one often distinguishes between the situations termed as genuine "walks" (or velocity models) where the particle moves, say, with a constant velocity and

Copyrighted material

changes its direction at the end of the step, and "random flight" models, when steps are considered to be instantaneous ("jumps") and the time cost of a step is attributed to the waiting time at a site before the jump is made. The situations considered below correspond to the flight picture.

The are three important postulates guaranteeing that random walks correspond to a normal diffusion process:

- The existence of the mean squared displacement per step, $\langle \mathbf{s}^2 \rangle < \infty$
- The existence of the mean time per step (interpreted as a mean waiting time on a site) $\tau < \infty$ and
- The uncorrelated nature of the walk $\langle \mathbf{s}_i \mathbf{s}_j \rangle = \langle \mathbf{s}^2 \rangle \delta_{ij}$.

These are exactly the three postulates made by Einstein in his work on Brownian motion. The last postulate can be weakened: The persistent (i. e. correlated) random walks still lead to normal diffusion provided $\sum_{j=i}^{\infty} \langle \mathbf{s}_i \mathbf{s}_j \rangle < \infty$.

In the case when the steps are independent, the central limit theorem immediately states that the limiting distribution of the displacements will be Gaussian, of the form

$$P(\mathbf{r}, t) = \frac{1}{(4\pi K t)^{d/2}} \exp\left(-\frac{\mathrm{d}r^2}{4Kt}\right) .$$

This distribution scales as a whole: rescaling the distance by a factor of λ, $x \to \lambda x$ and of the time by the factor of λ^2, $t \to \lambda^2 t$, does not change the form of this distribution. The moments of the displacement scale according to $\langle \mathbf{r}^k \rangle = \int P(r, t) \mathbf{r}^k d^d \mathbf{r}^k = \mathrm{const}(k) t^{k/2}$. Another important property of the the PDF is the fact that the return probability, i. e. the probability to be at time t at the origin of the motion, $P(\mathbf{0}, t)$, scales as

$$P(\mathbf{0}, t) \propto t^{-d/2} , \tag{1}$$

i. e. shows the behavior depending on the substrate's dimension.

The PDF $P(\mathbf{r}, t)$ is the solution of the Fick's diffusion equation

$$\frac{\partial}{\partial t} P(\mathbf{r}, t) = k \Delta P(\mathbf{r}, t) , \tag{2}$$

where Δ is a Laplace operator. Equation 2 is a parabolic partial differential equation, so that its form is invariant under the scale transformation $t \to \lambda^2 t, r \to \lambda r$ discussed above. This invariance of the equation has very strong implications. If follows, for example, that one can invert the scaling relation for the mean squared displacement

$$R^2 \propto t$$

and interpret this inverse one

$$T \propto r^2$$

as e. g. a scaling rule governing the dependence of a characteristic passage time from some initial site 0 to the sites at a distance r from this one. Such scaling holds e. g. for the mean time spent by a walk in a prescribed region of space (i. e. for the mean first passage time to its boundary). For an isotropic system Eq. (2) can be rewritten as an equation in the radial coordinate only:

$$\frac{\partial}{\partial t} P(r, t) = K \frac{1}{r^{d-1}} \frac{\partial}{\partial r} \left[r^{d-1} \frac{\partial}{\partial r} P(r, t) \right] . \tag{3}$$

Looking at the trajectory of a random walk at scales much larger than the step's length one infers that it is self-similar in statistical sense, i. e. that its whatever portion (of the size considerably larger then the step's length) looks like the whole trajectory, i. e. it is a statistical, random fractal. The fractal (mass) dimension of the trajectory of a random walk can be easily found. Let us interpret the number of steps as the "mass" M of the object and $\langle \mathbf{r}_n^2 \rangle^{1/2}$ as its "size" L. The mass dimension D is then defined as $M \propto L^D$, it corresponds to the scaling of the mass of a solid D-dimensional object with its typical size. The scaling of the mean squared displacement of a simple random walk with its number of steps suggests that the corresponding dimension $D = d_\mathrm{w}$ for a walk is $d_\mathrm{w} = 2$. The large-scale properties of the random walks in Euclidean space are universal and do not depend on exact form of the PDF of waiting times and of jumps' lengths (provided the first possesses at least one and the second one at least two moments).

This fact can be used for an interpretation of several known properties of simple random walks. Thus, random walks on a one-dimensional lattice ($d = 1$) are recurrent, i. e. each site earlier visited by the walk is revisited repeatedly afterwards, while for $d \geq 3$ they are nonrecurrent. This property can be rationalized through arguing that an intrinsically two-dimensional object cannot be "squeezed" into a one-dimensional space without self-intersections, and that it cannot fill fully a space of three or more dimensions. The same considerations can be used to rationalize the behavior of the mean number of different sites visited by a walker during n steps $\langle S_n \rangle$. $\langle S_n \rangle$ typically goes as a power law in n. Thus, on a one-dimensional lattice $\langle S_n \rangle \propto n^{1/2}$, in lattices in three and more dimensions $\langle S_n \rangle \propto n$, and in a marginal two-dimensional situation $\langle S_n \rangle \propto n/\log n$. To understand the behavior in $d = 1$ and in $d = 3$, it is enough to note that in a recurrent one-dimensional walk all sites within the span of the walk are visited by a walker at least once, so that the mean number of different sites visited by a one-dimensional walk ($d = 1$) grows as a span of the walk, which itself is proportional to a mean squared displacement, $\langle S_n \rangle \propto L = \langle \mathbf{r}_n^2 \rangle^{1/2}$, i. e.

Copyrighted material

$\langle S_n \rangle \propto n^{1/2} = n^{1/d_w}$. In three dimensions ($d = 3$), on the contrary, the mean number of different visited sites grows as the number of steps, $\langle S_n \rangle \propto n$, since different sites are on the average visited only finite number of times. The property of random walks on one-dimensional and two-dimensional lattices to visit practically all the sites within the domain of the size of their mean square displacement is called the *compact visitation* property.

The importance of the number of different sites visited (and its moments like $\langle S_n \rangle$ or $\langle S(t) \rangle$) gets clear when one considers chemical reactions or search problems on the corresponding networks. The situation when each site of a network can with the probability p be a trap for a walker, so that the walker is removed whenever it visits such a marked site (reaction scheme A + B → B, where the symbol A denotes the walker and symbol B the "immortal" and immobile trap) corresponds to a trapping problem of reaction kinetics. The opposite situation, corresponding to the evaluation of the survival probability of an immobile A-particle at one of the network's nodes in presence of mobile B-walkers (the same A + B → B reaction scheme, however now with immobile A and mobile B) corresponds to the target problem, or scavenging. In the first case the survival probability $\Phi(N)$ for the A-particle after N steps goes for p small as $\Phi(t) \propto \langle \exp(-pS_N) \rangle$, while for the second case it behaves as $\Phi(t) \propto \exp(-p\langle S_N \rangle)$. The simple A + B → B reaction problems can be also interpreted as the ones of the distribution of the times necessary to find one of multiple hidden "traps" by one searching agent, or as the one of finding one special marked site by a swarm of independent searching agents.

Anomalous Diffusion

The Einstein's postulates leading to normal diffusion are to no extent the laws of Nature, and do not have to hold for a whatever random walk process. Before discussing general properties of such anomalous diffusion, we start from a few simple examples [10].

Simple Geometric Models Leading to Subdiffusion

Let us first discuss two simple models of geometrical structures showing anomalous diffusion. The discussion of these models allows one to gain some intuition about the emergence and properties of anomalous diffusion in complex geometries, as opposed to Euclidean lattices, and discuss mathematical notions appearing in description of such systems. In all cases the structure on which the walk takes place will be considered as a lattice with the unit spacing between the sites, and the walk corresponds to jumps from a site to one of its neighbors. Moreover, we consider the random walk to evolve in discrete time; the time t and the number of steps n are simply proportional to each other.

Diffusion on a Comb Structure The simplest example of how the subdiffusive motion can be created in a more or less complex (however yet non-fractal) geometrical structure is delivered by the diffusion on a comb. We consider the diffusion of a walker along the backbone of a comb with infinite "teeth" (dangling ends), see Fig. 1. The walker's displacement in x-direction is only possible when it is on the backbone; entering the side branch switches off the diffusion along the backbone. Therefore the motion on the backbone of the comb is interrupted by waiting times. The distribution of these waiting times is given by the power law $\psi(t) \propto t^{-3/2}$ corresponding to the first return probability of a one-dimensional random walk in a side branch to its origin. This situation – simple random walk with broad distribution of waiting times on sites – corresponds to a continuous-time random walk model. In this model, for the case of the power-law waiting time distributions of the form $\psi(t) \propto t^{-1-\alpha}$ with $\alpha < 1$, the mean number of steps $\langle n(t) \rangle$ (in our case: the mean number of steps over the backbone) as a function of time goes as $\langle n(t) \rangle \propto t^{\alpha}$. Since the mean squared displacement along the backbone is proportional to this number of steps, it is also governed by $\langle x^2(t) \rangle \propto t^{\alpha}$, i. e. in our case we have to do with strongly subdiffusive behavior $\langle x^2(t) \rangle \propto t^{1/2}$.

The PDF of the particle's displacements in CTRW with the waiting time distribution $\psi(t)$ is governed by a non-Markovian generalization of the diffusion equation

$$\frac{\partial}{\partial t} P(x,t) = a^2 \frac{\partial}{\partial t} \Delta \int_0^{\infty} M(t-t') P(x,t') \, ,$$

with the memory kernel $M(t)$ which has a physical meaning of the time-dependent density of steps and is given by the inverse Laplace transform of $\tilde{M}(u) = \tilde{\psi}(u)/[1-\tilde{\psi}(u)]$ (with $\tilde{\psi}(u)$ being the Laplace-transform of $\psi(t)$). For the



Anomalous Diffusion on Fractal Networks, Figure 1
A comb structure: Trapping in the dangling ends leads to the anomalous diffusion along the comb's backbone. This kind of anomalous diffusion is modeled by CTRW with power-law waiting time distribution and described by a fractional diffusion equation

Copyrighted material

case of a power-law waiting-time probability density of the type $\psi(t) \propto t^{-1-\alpha}$ the structure of the right-hand side of the equation corresponds to the *fractional Riemann–Liouvolle derivative* ${}_0D_t^{\beta}$, an integro-differential operator defined as

$$ {}_0D_t^{\beta} f(t) = \frac{1}{\Gamma(1-\beta)} \frac{\mathrm{d}}{\mathrm{d}t} \int_0^t \mathrm{d}t' \frac{f(t')}{(t-t')^{\beta}} , \tag{4} $$

(here only for $0 < \beta < 1$), so that in the case of the diffusion on a backbone of the comb we have

$$ \frac{\partial}{\partial t} P(x,t) = K_0 D_t^{1-\alpha} \Delta P(x,t) , \tag{5} $$

(here with $\alpha = 1/2$). An exhaustive discussion of such equations is provided in [13,14]. Similar fractional diffusion equation was used in [12] for the description of the diffusion in a percolation system (as measured by NMR). However one has to be cautious when using CTRW and corresponding fractional equations for the description of diffusion on fractal networks, since they may capture some properties of such diffusion and fully disregard other ones. This question will be discussed to some extent in Sect. "Anomalous Diffusion on Fractal Structures".

Equation (5) can be rewritten in a different but equivalent form

$$ {}_0^*D_t^{\alpha} \frac{\partial}{\partial t} P(x,t) = K \Delta P(x,t) , $$

with the operator ${}_0^*D_t^{\alpha}$, a Caputo derivative, conjugated to the Riemann–Liouville one, Eq. (4), and different from Eq. (4) by interchanged order of differentiation and integration (temporal derivative inside the integral). The derivation of both forms of the fractional equations and their interpretation is discussed in [21]. For the sake of transparency, the corresponding integrodifferential operators ${}_0D_t^{\beta}$ and ${}_0^*D_t^{\beta}$ are simply denoted as a fractional partial derivative $\partial^{\beta}/\partial t^{\beta}$ and interpreted as a Riemann–Liouville or as a Caputo derivative depending whether they stand on the left or on the right-hand side of a generalized diffusion equation.

Random Walk on a Random Walk and on a Self-Avoiding Walk Let us now turn to another source of anomalous diffusion, namely the nondecaying correlations between the subsequent steps, introduced by the structure of the possible ways on a substrate lattice.

The case where the anomalies of diffusion are due to the tortuosity of the path between the two sites of a fractal "network" is illustrated by the examples of random walks on polymer chains, being rather simple, topologically one-dimensional objects. Considering lattice models for the chains we encounter two typical situations, the simpler one, corresponding to the random walk (RW) chains and a more complex one, corresponding to self-avoiding ones. A conformation of a RW-chain of N monomers corresponds to a random walk on a lattice, self-intersections are allowed. In reality, RW chains are a reasonable model for polymer chains in melts or in θ-solutions, where the repulsion between the monomers of the same chain is compensated by the repulsion from the monomers of other chains or from the solvent molecules, so that when not real intersections, then at least very close contacts between the monomers are possible. The other case corresponds to chains in good solvents, where the effects of steric repulsion dominate. The end-to-end distance in a random walk chain grows as $R \propto l^{1/2}$ (with l being its contour length). In a self-avoiding-walk (SAW) chain $R \propto l^{\nu}$, with ν being a so-called Flory exponent (e. g. $\nu \approx 3/5$ in $3d$). The corresponding chains can be considered as topologically one-dimensional fractal objects with fractal dimensions $d_{\mathrm{f}} = 2$ and $d_{\mathrm{f}} = 1/\nu \approx 5/3$ respectively.

We now concentrate on an walker (excitation, enzyme, transcription factor) performing its diffusive motion on a chain, serving as a "rail track" for diffusion. This chain itself is embedded into the Euclidean space. The contour length of the chain corresponds to the *chemical coordinate* along the path, in each step of the walk at a chain the chemical coordinate changes by ± 1. Let us first consider the situation when a particle can only travel along this chemical path, i. e. cannot change its track at a point of a self-intersection of the walk or at a close contact between the two different parts of the chain. This walk starts with its first step at $\mathbf{r} = 0$. Let K be the diffusion coefficient of the walker along the chain. The typical displacement of a walker along the chain after time t will then be $l \propto t^{1/2} = t^{1/d_{\mathrm{w}}}$. For a RW-chain this displacement along the chain is translated into the displacement in Euclidean space according to $R = \langle \mathbf{R}^2 \rangle^{1/2} \propto l^{1/2}$, so that the typical displacement in Euclidean space after time t goes as $R \propto t^{1/4}$, a strongly subdiffusive behavior known as one of the regimes predicted by a repetition theory of polymer dynamics. The dimension of the corresponding random walk is, accordingly, $d_{\mathrm{w}} = 4$. The same discussion for the SAW chains leads to $d_{\mathrm{w}} = 1/2\nu$.

After time t the displacement in the chemical space is given by a Gaussian distribution

$$ p_{\mathrm{c}}(l,t) = \frac{1}{\sqrt{4\pi K t}} \exp\left(-\frac{l^2}{4Kt}\right) . $$

The change in the contour length l can be interpreted as the number of steps along the random-walk "rail", so that the Euclidean displacement for the given l is distributed

Copyrighted material

according to

$$p_{\mathrm{E}}(\mathbf{r}, l) = \left(\frac{d}{2\pi a^2 |l|}\right) \exp\left(-\frac{dr^2}{2a^2|l|}\right) . \quad (6)$$

We can, thus, obtain the PDF of the Euclidean displacement:

$$P(\mathbf{r}, l) = \int_0^\infty p_{\mathrm{E}}(\mathbf{r}, l) p_{\mathrm{c}}(l, t) \mathrm{d}l . \quad (7)$$

The same discussion can be pursued for a self-avoiding walk for which

$$p_{\mathrm{E}}(\mathbf{r}, l) \simeq \left(\frac{r}{l^\nu}\right)^{\frac{\gamma-1}{\nu}} \exp\left[-\mathrm{const}\left(\frac{r}{l^\nu}\right)^{1/(1-\nu)}\right] , \quad (8)$$

with γ being the exponent describing the dependence of the number of realizations of the corresponding chains on their lengths. Equation (6) is a particular case of Eq. (8) with $\gamma = 1$ and $\nu = 1/2$. Evaluating the integral using the Laplace method, we get

$$P(\mathbf{r}, t) \simeq \exp\left[-\mathrm{const}\left(\frac{r}{t^{1/d_{\mathrm{w}}}}\right)^{(1-1/d_{\mathrm{w}})^{-1}}\right]$$

up to the preexponential. We note that in this case the reason for the anomalous diffusion is the tortuosity of the chemical path. The topologically one-dimensional structure of the lattice allowed us for discussing the problem in considerable detail.

The situation changes when we introduce the possibility for a walker to jump not only to the nearest-neighbor in chemical space by also to the sites which are far away along the chemical sequence but close to each other in Euclidean space (say one lattice site of the underlying lattice far). In this case the structure of the chain in the Euclidean space leads to the possibility to jump very long in chemical sequence of the chain, with the jump length distribution going as $p(l) \propto l^{-3/2}$ for an RW and $p(l) \propto l^{-2.2}$ for a SAW in $d = 3$ (the probability of a close return of a SAW to its original path). The corresponding distributions lack the second moment (and even the first moment for RW chains), and therefore one might assume that also the diffusion in a chemical space of a chain will be anomalous. It indeed shows considerable peculiarities. If the chain's structure is frozen (i. e. the conformational dynamics of a chain is slow compared to the diffusion on the chain), the situation in both cases corresponds to "paradoxical diffusion" [22]: Although the PDF of displacements in the chemical space lacks the second (and higher) moments, the width of the PDF (described e. g. as its interquartile distance) grows, according to $W \propto t^{1/2}$. In the Euclidean space we have here to do with the diffusion on a static fractal structure, i. e. with a network of fractal dimension d_{f} coinciding with the fractal dimension of the chain. Allowing the chain conformation changing in time (i. e. considering the opposite limiting case when the conformation changes are fast enough compared to the diffusion along the chain) leads to another, superdiffusive, behavior, namely to Lévy flights along the chemical sequence. The difference between the static, fixed networks and the transient ones has to be borne in mind when interpreting physical results.

Anomalous Diffusion on Fractal Structures

Simple Fractal Lattices: the Walk's Dimension

As already mentioned, the diffusion on fractal structures is often anomalous, $\langle x^2(t)\rangle \propto t^\alpha$ with $\alpha \neq 1$. Parallel to the situation with random walks in Euclidean space, the trajectory of a walk is typically self similar (on the scales exceeding the typical step length), so that the exponent α can be connected with the fractal dimension of the walk's trajectory d_{w}. Assuming the mean time cost per step to be finite, i. e. $t \propto n$, one readily infers that $\alpha = 2/d_{\mathrm{w}}$. The fractal properties of the trajectory can be rather easily checked, and the value of d_{w} can be rather easily obtained for simple regular fractals, the prominent example being a Sierpinski gasket, discussed below. Analytically, the value of d_{w} is often obtained via scaling of the first passage time to a boundary of the region (taken to be a sphere of radius r), i. e. using not the relation $R \propto t^{1/d_{\mathrm{w}}}$ but the inverse one, $T \propto r^{d_{\mathrm{w}}}$. The fact that both values of d_{w} coincide witnesses strongly in favor of the self-similar nature of diffusion on fractal structures.

Diffusion on a Sierpinski Gasket Serpinski gasket (essentially a Sierpinski lattice), one of the simplest regular fractal network is a structure obtained by iterating a generator shown in Fig. 2. Its simple iterative structure and the ease of its numerical implementation made it a popular toy model exemplifying the properties of fractal networks. The properties of diffusion on this network were investigated in detail, both numerically and analytically. Since the structure does not possess dangling ends, the cause of the diffusion anomalies is connected with tortuosity of the typical paths available for diffusion, however, differently from the previous examples, these paths are numerous and non-equivalent.

The numerical method allowing for obtaining exact results on the properties of diffusion is based on exact enumeration techniques, see e. g. [10]. The idea here is to calculate the displacement probability distribution based on the exact number of ways $W_{i,n}$ the particle, starting at

Copyrighted material

A different discussion of the same problem based on the Einstein's relation between diffusion coefficient and conductivity, and on crossover arguments can be found in [10,20].

The Spectral Dimension

In mathematical literature spectral dimension is defined through the return probability of the random walk, i. e. the probability to be at the origin after n steps or at time t. We consider a node on a network and take it to be an origin of a simple random walk. We moreover consider the probability $P(\mathbf{0}, t)$ to be at this node at time t (a return probability). The spectral dimension defines then the asymptotic behavior of this probability for n large:

$$P(\mathbf{0}, t) \propto t^{-d_s/2} , \quad \text{for } t \to \infty . \tag{13}$$

The spectral dimension d_s is therefore exactly a quantity substituting the Euclidean dimension d is Eq. (1). In the statistical case one should consider an average of $p(t)$ over the origin of the random walk and over the ensemble of the corresponding graphs.

For fixed graphs the spectral dimension and the fractal (Hausdorff) dimension are related by

$$\frac{2d_f}{1+d_f} \le d_s \le d_f , \tag{14}$$

provided both exist [5]. The same relation is also shown true for some random geometries, see e. g. [6,7] and references therein. The bounds are optimal, i. e. both equalities can be realized in some structures.

The connection between the walk dimension d_w and the spectral dimension d_s of a network can be easily rationalized by the following consideration. Let the random walk have a dimension d_w, so that after time t the position of the walker can be considered as more or less homogeneously spread within the spatial region of the linear size $R \propto n^{1/d_w}$ or $R \propto t^{1/d_w}$, with the overall number of nodes $N \propto R^{d_f} \propto t^{d_f/d_w}$. The probability to be at one particular node, namely at $\mathbf{0}$, goes then as $1/N$, i. e. as t^{-d_f/d_w}. Comparing this with the definition of the spectral dimension, Eq. (13) we get exactly Eq. (12). The lower bound in the inequality Eq. (14) follows from Eq. (13) and from the the observation that the value of ζ in Eq. (11) is always smaller or equal to one (the value of $\zeta = 1$ would correspond to a one-dimensional wire without shunts). The upper bound can be obtained using Eq. (12) and noting that the walks' dimension never gets less than 2, its value for the regular network: the random walk on a whatever structure is more compact than one on a line.

We note that the relation Eq. (12) relies on the assumption that the spread of the walker within the r-domain is homogeneous, and needs reinterpretation for strongly anisotropic structures like combs, where the spread in teeth and along the spine are very different (see e. g. [2]). For the planar comb with infinite teeth $d_f = 2$, and d_s as calculated through the return time is $d_s = 3/2$, while the walk's dimension (mostly determined by the motion in teeth) is $d_w = 2$, like for random walks in Euclidean space. The inequality Eq. (14) is, on the contrary, universal.

Let us discuss another meaning of the spectral dimension, the one due to which it got its name. This one has to do with the description of random walks within the master equation scheme. From the spectral (Laplace) representation of the solution of the master equation, Eq. (9), we can easily find the probability $P(\mathbf{0}, t)$ that the walker starting at site $\mathbf{0}$ at $t = 0$ is found at the same site at time t. This one reads

$$P(\mathbf{0}, t) = \sum_{i=1}^{\infty} a_i \exp(-\epsilon_i t) ,$$

where ϵ_i is the ith eigenvalue of the matrix $\mathbf{W}$ and a_i is the amplitude of its ith eigenvector at site $\mathbf{0}$. Considering the lattice as infinite, we can pass from discrete eigenvalue decomposition to a continuum

$$P(\mathbf{0}, t) = \int_0^{\infty} \mathcal{N}(\epsilon) a(\epsilon) \exp(-\epsilon t) \mathrm{d}\epsilon .$$

For long times, the behavior of $P(\mathbf{0}, t)$ is dominated by the behavior of $\mathcal{N}(\epsilon)$ for small values of ϵ. Here $\mathcal{N}(\epsilon)$ is the density of states of a system described by the matrix $\mathbf{W}$. The exact forms of such densities are well known for many Euclidean lattices, since the problem is equivalent to the calculating of spectrum in tight-binding approximation used in the solid state physics. For all Euclidean lattices $\mathcal{N}(\epsilon) \propto \epsilon^{d/2-1}$ for $\epsilon \to 0$, which gives us the forms of famous van Hove singularities of the spectrum. Assuming $a(\epsilon)$ to be nonsingular at $\epsilon \to 0$, we get $P(\mathbf{0}, t) \propto t^{-d/2}$. For a fractal structure, the value of d changed for the one of the spectral dimension d_s. This corresponds to the density of states $\mathcal{N}(\epsilon) \propto \epsilon^{d_s/2-1}$ which describes the properties of spectrum of a fractal analog of a Laplace operator. The dimension d_s is also often called fracton dimension of the structure, since the corresponding eigenvectors of the matrix (corresponding to eigenstates in tight-binding model) are called fractons, see e. g. [16].

In the examples of random walks on quenched polymer chains the dimension d_w was twice the fractal dimension of the underlying structures, so that the spectral dimension of the corresponding structures (without intersections) was exactly 1 just like their topological dimension.

Copyrighted material

The spectral dimension of the network governs also the behavior of the mean number of different sites visited by the random walk. A random walk of n steps having a property of compact visitation typically visits all sites within the radius of the order of its typical displacement $R_n \propto n^{1/d_w}$. The number of these sites is $S_n \propto R_n^{d_f}$ where d_f is the fractal dimension of the network, so that $S_n \propto n^{d_s/2}$ (provided $d_s \leq 2$, i. e. provided the random walk is recurrent and shows compact visitation). The spectral dimension of a structure plays important role in many other applications [16]. This and many other properties of complex networks related to the spectrum of $\mathbf{W}$ can often be easier obtained by considering random walk on the lattice than by obtaining the spectrum through direct diagonalization.

Percolation Clusters

Percolation clusters close to criticality are one of the most important examples of fractal networks. The properties of these clusters and the corresponding fractal and spectral dimensions are intimately connected to the critical indices in percolation theory.

Thus, simple crossover arguments show that the fractal dimension of an incipient infinite percolation cluster in $d \leq 6$ dimensions is $d_f = d - \beta/\nu$, where β is the critical exponent of the density of the infinite cluster, $P_\infty \propto (p - p_c)^\beta$ and ν is the critical exponent of the correlation length, $\xi \propto |p - p_c|^\nu$, and stagnates at $d_f = 4$ for $d > 6$, see e. g. [23].

On the other hand, the critical exponent t, describing the behavior of the resistance of a percolation system close to percolation threshold, $\rho \propto (p - p_c)^{-t}$, is connected with the exponent ζ via the following crossover argument. The resistance of the system is the one of the infinite cluster. For $p > p_c$ this cluster can be considered as fractal at scales $L < \xi$ and as homogeneous at larger scales. Our value for the fractal dimension of the cluster follows exactly from this argument by matching the density of the cluster $P(L) \propto L^{(d_f - d)}$ at $L = \xi \propto (p - p_c)^{-\nu}$ to its density $P_\infty \propto (p - p_c)^\beta$ at larger scales. The infinite cluster for $p > p_c$ is then a dense regular assembly of subunits of size ξ which on their turn are fractal. If the typical resistivity of each fractal subunit is R_ξ then the resistivity of the overall assembly goes as $R \propto R_\xi (L/\xi)^{2-d}$. This is a well-known relation showing that the resistivity of a wire grows proportional to its length, the resistivities of similar flat figures are the same etc. This relation holds in the homogeneous regime. On the other hand, in the fractal regime $R_L \propto L^\zeta$, so that overall $R \propto \xi^{\zeta + d - 2} \propto (p - p_c)^{-\nu(\zeta + d - 2)}$ giving the value of the critical exponent $t = \nu(\zeta + d - 2)$. The value of ζ can in its turn be expressed through the values of spectral and fractal dimensions of a percolation cluster.

The spectral dimension of the percolation cluster is very close to 4/3 in any dimension larger then one. This surprising finding lead to the conjecture by Alexander and Orbach that this value of $d_s = 4/3$ might be exact [1] (it is exact for percolation systems in $d \geq 6$ and for trees). Much effort was put into proving or disproving the conjecture, see the discussion in [10]. The latest, very accurate simulations of two-dimensional percolation by Grassberger show that the conjecture does not hold in $d = 2$, and the prediction $d_s = 4/3$ is off by around 2% [9]. In any case it can be considered as a very useful mnemonic rule.

Anomalous diffusion on percolation clusters corresponds theoretically to the most involved case, since it combines all mechanisms generating diffusion anomalies. The infinite cluster of the percolation system consists of a backbone, its main current-carrying structure, and the dangling ends (smaller clusters on all scales attached to the backbone through only one point). The anomalous diffusion on the infinite cluster is thus partly caused by trapping in this dangling ends. If one considers a shortest (chemical) way between the two nodes on the cluster, this way is tortuous and has fractal dimension larger than one.

The Role of Finite Clusters

The exponent t of the conductivity of a percolation system is essentially the characteristic of an infinite cluster only. Depending on the physical problem at hand, one can consider the situations when only the infinite cluster plays the role (like in the experiments of [12], where only the infinite cluster is filled by fluid pressed through the boundary of the system), and situations when the excitations can be found in infinite as well as in the finite clusters, as it is the case for optical excitation in mixed molecular crystals, a situation discussed in [24].

Let us concentrate on the case $p = p_c$. The structure of large but finite clusters at lower scales is indistinguishable from those of the infinite cluster. Thus, the mean squared displacement of a walker on a cluster of size L (one with $N \propto L^{d_f}$ sites) grows as $R^2(t) \propto t^{2/d_w}$ until is stagnates at the value of R of the order of the cluster's size, $R^2 \propto N^{2/d_f}$. At a given time t we can subdivide all clusters into two classes: those whose size is small compared to t^{1/d_w} (i. e. the ones with $N < t^{d_s/2}$), on which the mean square displacement stagnates, and those of the larger size, on which it still grows:

$$R^2(t) \propto \begin{cases} t^{2/d_w}, & t^{1/d_w} < N^{1/f_t} \\ N^{2/d_f}, & \text{otherwise.} \end{cases}$$

Copyrighted material

The characteristic size $N_{\mathrm{cross}}(t)$ corresponds to the crossover between these two regimes for a given time t. The probability that a particle starts at a cluster of size N is proportional to the number of its sites and goes as $P(N) \propto Np(N) \propto N^{1-\tau}$ with $\tau = (2d\nu-\beta)/(d\nu-\beta)$, see e. g. [23]. Here $p(N)$ is the probability to find a cluster of N sites among all clusters. The overall mean squared displacement is then given by averaging over the corresponding cluster sizes:

$$\langle r^2(t)\rangle \propto \sum_{N=1}^{N_{\mathrm{cross}}(t)} N^{2/d_{\mathrm{f}}} N^{1-\tau} + \sum_{N_{\mathrm{cross}}(t)}^{\infty} t^{2/d_{\mathrm{w}}} N^{1-\tau} .$$

Introducing the expression for $N_{\mathrm{cross}}(t)$, we get

$$\langle r^2(t)\rangle \propto t^{2/d_{\mathrm{w}}+(2-\tau)d_{\mathrm{f}}/d_{\mathrm{w}}} = t^{(1/d_{\mathrm{w}})(2d_{\mathrm{f}}-\beta/\nu)} .$$

A similar result holds for the number of distinct sites visited, where one has to perform the analogous averaging with

$$S(t) \propto \begin{cases} t^{d_{\mathrm{s}}/2} , & t^{1/d_{\mathrm{w}}} < N^{1/f_{\mathrm{f}}} \\ N , & \text{otherwise} \end{cases}$$

giving $S(t) \propto t^{(d_{\mathrm{s}}/2)(1-\beta\nu/d_{\mathrm{f}})}$. This gives us the effective spectral dimension of the system $\tilde{d}_{\mathrm{s}} = (d_{\mathrm{s}}/2)(1-\beta\nu/d_{\mathrm{f}})$.

Scaling of PDF and Diffusion Equations on Fractal Lattices

If the probability density of the displacement of a particle on a fractal scales, the overall scaling form of this displacement can be obtained rather easily: Indeed, the typical displacement R during the time t goes as $r \propto t^{1/d_{\mathrm{w}}}$, so that the corresponding PDF has to scale as

$$P(r,t) = \frac{r^{d_{\mathrm{f}}-d}}{t^{d_{\mathrm{s}}/2}} f\left(\frac{r}{t^{1/d_{\mathrm{w}}}}\right) . \tag{15}$$

The prefactor takes into account the normalization of the probability density on a fractal structure (denominator) and also the scaling of the density of the fractal structure in the Euclidean space (enumerator). The simple scaling assumes that all the moments of the distance scale in the same way, so that $\langle \mathbf{r}^k\rangle = \int P(r,t)\mathbf{r}^k d^d\mathbf{r}^k = \mathrm{const}(k)t^{k/d_{\mathrm{w}}}$. The overall scaling behavior of the PDF was confirmed numerically for many fractal lattices like Sierpinski gaskets and percolation clusters, and is corroborated in many other cases by the coincidence of the values of d_{w} obtained via calculation of the mean squared displacement and of the mean first passage time (or resistivity of the network). The existence of the scaling form leads to important consequences e. g. for quantification of anomalous diffusion in biological tissues by characterizing the diffusion-time dependence of the magnetic resonance signal [17], which, for example, allows for differentiation between the healthy and the tumor tissues. However, even in the cases when the scaling form Eq. (15) is believed to hold, the exact form of the scaling function $f(x)$ is hard to get; the question is not yet resolved even for simple regular fractals.

For applications it is often interesting to have a kind of a phenomenological equation roughly describing the behavior of the PDF. Such an approach will be analogous to putting down Richardson's equation for the turbulent diffusion. Essentially, the problem here is to find a correct way of averaging or coarse-graining the microscopic master equation, Eq. (9), to get its valid continuous limit on larger scales. The regular procedure to do this is unknown; therefore, several phenomenological approaches to the problem were formulated.

We are looking for an equation for the PDF of the walker's displacement from its initial position and assume the system to be isotropic on the average. We look for an analog of the classical Fick's equation, $\frac{\partial}{\partial t}P(r,t) = \nabla K(r)\nabla P(r,t)$ which in spherical coordinates and for $K = K(r)$ takes the form of Eq. (3):

$$\frac{\partial}{\partial t}P(r,t) = \frac{1}{r^{d-1}}\frac{\partial}{\partial r}\left[r^{d-1}K(r)\frac{\partial}{\partial r}P(r,t)\right] .$$

On fractal lattices one changes from the Euclidean dimension of space to the fractal dimension of the lattice d_{f}, and takes into account that the effective diffusion coefficient $K(r)$ decays with distance as $K(r) \simeq Kr^{-\theta}$ with $\theta = d_{\mathrm{w}}-2$ to capture the slowing down of anomalous diffusion on a fractal compared to the Euclidean lattice situation. The corresponding equation put forward by O'Shaughnessy and Procaccia [18] reads:

$$\frac{\partial}{\partial t}P(r,t) = K\frac{1}{r^{d_{\mathrm{f}}-1}}\frac{\partial}{\partial r}\left[r^{d_{\mathrm{f}}-d_{\mathrm{w}}+1}\frac{\partial}{\partial r}P(r,t)\right] . \tag{16}$$

This equation was widely used in description of anomalous diffusion on fractal lattices and percolation clusters, but can be considered only as a rough phenomenological approximation. Its solution

$$P(r,t) = A\frac{r^{d_{\mathrm{f}}-d}}{t^{d_{\mathrm{s}}/2}}\exp\left(-B\frac{r^{d_{\mathrm{w}}}}{t}\right) ,$$

(with $B = 1/Kd_{\mathrm{w}}^2$ and A being the corresponding normalization constant) corresponds exactly to the type of scaling given by Eq. (15), with the scaling function $f(x) =$

Copyrighted material

$\exp(x^{d_w})$. This behavior of the scaling function disagrees e. g. with the results for random walks on polymer chains, for which case we had $f(x) = \exp(x^{\mu})$ with $\mu = (1 - 1/d_w)^{-1}$ for large enough x.

In literature, several other proposals (based on plausible assumptions but to no extent following uniquely from the models considered) were made, taking into account possible non-Markovian nature of the motion. These are the fractional equations of the type

$$\frac{\partial^{1/d_w}}{\partial t^{1/d_w}} P(r,t) = -K_1 \frac{1}{r^{(d_s-1)/2}} \frac{\partial}{\partial r}\left[r^{(d_s-1)/2} P(r,t)\right], \quad (17)$$

resembling "half" of the diffusion equation and containing a fractional derivative [8], as well as the "full" fractional diffusion equation

$$\frac{\partial^{2/d_w}}{\partial t^{2/d_w}} P(r,t) = K_2 \frac{1}{r^{d_s-1}} \frac{\partial}{\partial r}\left[r^{d_s-1} \frac{\partial}{\partial r} P(r,t)\right], \quad (18)$$

as proposed in [15]. All these equation are invariant under the scale transformation

$$t \to \lambda t\,, \quad r \to \lambda^{1/d_w} r\,,$$

and lead to PDFs showing correct overall scaling properties, see [19]. None of them reproduces correctly the PDF of displacements on a fractal (i. e. the scaling function $f(x)$) in the whole range of distances. Ref. [19] comparing the corresponding solutions with the results of simulation of anomalous diffusion an a Sierpinski gasket shows that the O'Shaughnessy–Procaccia equation, Eq. (16), performs the best for the central part of the distribution (small displacements), where Eq. (18) overestimates the PDF and Eq. (17) shows an unphysical divergence. On the other hand, the results of Eqs. (17) and (18) reproduce equally well the PDF's tail for large displacements, while Eq. (16) leads to a PDF decaying considerably faster than the numerical one. This fact witnesses favor of strong non-Markovian effects in the fractal diffusion, however, the physical nature of this non-Markovian behavior observed here in a fractal network without dangling ends, is not as clear as it is in a comb model and its more complex analogs. Therefore, the question of the correct equation describing the diffusion in fractal system (or different correct equations for the corresponding different classes of fractal systems) is still open. We note also that in some cases a simple fractional diffusion equation (with the corresponding power of the derivative leading to the correct scaling exponent d_w) gives a reasonable approximation to experimental data, as the ones of the model experiment of [12].

Further Directions

The physical understanding of anomalous diffusion due to random walks on fractal substrates may be considered as rather deep and full although it does not always lead to simple pictures. For example, although the spectral dimension for a whatever graph can be rather easily calculated, it is not quite clear what properties of the graph are responsible for its particular value. Moreover, there are large differences with respect to the degree of rigor with which different statements are proved. Mathematicians have recognized the problem, and the field of diffusion on fractal networks has become a fruitful field of research in probability theory. One of the examples of recent mathematical development is the deep understanding of spectral properties of fractal lattices and a proof of inequalities like Eq. (14).

A question which is still fully open is the one of the detailed description of the diffusion within a kind of (generalized) diffusion equation. It seems clear that there is more than one type of such equations depending on the concrete fractal network described. However even the classification of possible types of such equations is still missing.

All our discussion (except for the discussion of the role of finite clusters in percolation) was pertinent to infinite structures. The recent work [4] has showed that finite fractal networks are interesting on their own and opens a new possible direction of investigations.

Bibliography

Primary Literature

1. Alexander S, Orbach RD (1982) Density of states on fractals–fractons. J Phys Lett 43:L625–L631
2. Bertacci D (2006) Asymptotic behavior of the simple random walk on the 2-dimensional comb. Electron J Probab 45:1184–1203
3. Christou A, Stinchcombe RB (1986) Anomalous diffusion on regular and random models for diffusion-limited aggregation. J Phys A Math Gen 19:2625–2636
4. Condamin S, Bénichou O, Tejedor V, Voituriez R, Klafter J (2007) First-passage times in complex scale-invariant media. Nature 450:77–80
5. Coulhon T (2000) Random Walks and Geometry on Infinite Graphs. In: Ambrosio L, Cassano FS (eds) Lecture Notes on Analysis on Metric Spaces. Trento, CIMR, (1999) Scuola Normale Superiore di Pisa
6. Durhuus B, Jonsson T, Wheather JF (2006) Random walks on combs. J Phys A Math Gen 39:1009–1037
7. Durhuus B, Jonsson T, Wheather JF (2007) The spectral dimension of generic trees. J Stat Phys 128:1237–1260
8. Giona M, Roman HE (1992) Fractional diffusion equation on fractals – one-dimensional case and asymptotic-behavior. J Phys A Math Gen 25:2093–2105; Roman HE, Giona M, Fractional diffusion equation on fractals – 3-dimensional case and scattering function, ibid., 2107–2117

Copyrighted material

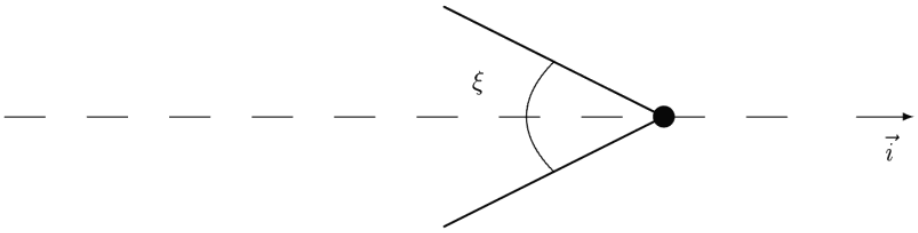
9. Grassberger P (1999) Conductivity exponent and backbone dimension in 2-*d* percolation. Physica A 262:251–263
10. Havlin S, Ben-Avraham D (2002) Diffusion in disordered media. Adv Phys 51:187–292
11. Klafter J, Sokolov IM (2005) Anomalous diffusion spreads its wings. Phys World 18:29–32
12. Klemm A, Metzler R, Kimmich R (2002) Diffusion on random-site percolation clusters: Theory and NMR microscopy experiments with model objects. Phys Rev E 65:021112
13. Metzler R, Klafter J (2000) The random walk's guide to anomalous diffusion: a fractional dynamics approach. Phys Rep 339:1–77
14. Metzler R, Klafter J (2004) The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. J Phys A Math Gen 37:R161–R208
15. Metzler R, Glöckle WG, Nonnenmacher TF (1994) Fractional model equation for anomalous diffusion. Physica A 211:13–24
16. Nakayama T, Yakubo K, Orbach RL (1994) Dynamical properties of fractal networks: Scaling, numerical simulations, and physical realizations. Rev Mod Phys 66:381–443
17. Özarslan E, Basser PJ, Shepherd TM, Thelwall PE, Vemuri BC, Blackband SJ (2006) Observation of anomalous diffusion in excised tissue by characterizing the diffusion-time dependence of the MR signal. J Magn Res 183:315–323
18. O'Shaughnessy B, Procaccia I (1985) Analytical solutions for diffusion on fractal objects. Phys Rev Lett 54:455–458
19. Schulzky C, Essex C, Davidson M, Franz A, Hoffmann KH (2000) The similarity group and anomalous diffusion equations. J Phys A Math Gen 33:5501–5511
20. Sokolov IM (1986) Dimensions and other geometrical critical exponents in the percolation theory. Usp Fizicheskikh Nauk 150:221–255 (1986) translated in: Sov. Phys. Usp. 29:924
21. Sokolov IM, Klafter J (2005) From diffusion to anomalous diffusion: A century after Einstein's Brownian motion. Chaos 15:026103
22. Sokolov IM, Mai J, Blumen A (1997) Paradoxical diffusion in chemical space for nearest-neighbor walks over polymer chains. Phys Rev Lett 79:857–860
23. Stauffer D (1979) Scaling theory of percolation clusters. Phys Rep 54:1–74
24. Webman I (1984) Diffusion and trapping of excitations on fractals. Phys Rev Lett 52:220–223

Additional Reading

The present article gave a brief overview of what is known about the diffusion on fractal networks, however this overview is far from covering all the facets of the problem. Thus, we only discussed unbiased diffusion (the effects of bias may be drastic due to e. g. stronger trapping in the dangling ends), and considered only the situations in which the waiting time at all nodes was the same (we did not discuss e. g. the continuous-time random walks on fractal networks), as well as left out of attention many particular systems and applications. Several review articles can be recommended as a further reading, some of them already mentioned in the text. One of the best-known sources is [10] being a reprint of the text from the "sturm und drang" period of investigation of fractal geometries. A lot of useful information on random walks models in general an on walks on fractals is contained in the review by Haus and Kehr from approximately the same time. General discussion of the anomalous diffusion is contained in the work by Bouchaud and Georges. The classical review of the percolation theory is given in the book of Stauffer and Aharony. Some additional information on anomalous diffusion in percolation systems can be found in the review by Isichenko. A classical source on random walks in disordered systems is the book by Hughes.

Haus JW, Kehr K (1987) Diffusion in regular and disordered lattices. Phys Rep 150:263–406
Bouchaud JP, Georges A (1990) Anomalous diffusion in disordered media – statistical mechanisms, models and physical applications. Phys Rep 195:127–293
Stauffer D, Aharony A (2003) Introduction to Percolation Theory. Taylor & Fransis, London
Isichenko MB (1992) Percolation, statistical topography, and transport in random-media. Rev Mod Phys 64:961–1043
Hughes BD (1995) Random Walks and random Environments. Oxford University Press, New York

Copyrighted material



Biological Fluid Dynamics, Non-linear Partial Differential Equations, Figure 1
A mirror-symmetric scallop or an axisymmetric octopus

independent of c because of translational invariance. The coefficient of $\dot{c}$ in (3) represents the drag force corresponding to a rigid translation along the symmetry axis at unit speed, and it never vanishes. Thus (3) can be solved for $\dot{c}$, and we obtain

$$\dot{c} = \sum_{i=1}^{N} V_i(\xi^{(1)}, \ldots, \xi^{(N)})\dot{\xi}^{(i)} = V(\xi) \cdot \dot{\xi} \,. \tag{4}$$

Equation (4) links positional changes to shape changes through shape-dependent coefficients. These coefficients encode all hydrodynamic interactions between Ω and the surrounding fluid due to shape changes with rates $\dot{\xi}^{(1)}, \ldots, \dot{\xi}^{(N)}$.

A stroke is a closed path γ in the space S of admissible shapes given by $[0, T] \ni t \mapsto (\xi^{(1)}, \ldots \xi^{(N-1)})$. Swimming requires that

$$0 \neq \Delta c = \int_0^T \sum_{i=1}^{N} V_i \dot{\xi}^{(i)} \mathrm{d}t \tag{5}$$

i. e., that the differential form $\sum_{i=1}^{N} V_i \, \mathrm{d}\xi^{(i)}$ is not exact.

The Scallop Theorem Proved

Consider a swimmer whose motion is described by a parametrized curve in two dimensions ($N = 1$), so that (4) becomes

$$\dot{c}(t) = V(\xi(t))\dot{\xi}(t) \,, \qquad t \in \mathbb{R} \,, \tag{6}$$

and assume that $V \in L^1(S)$ is an integrable function in the space of admissible shapes and $\xi \in W^{1,\infty}(\mathbb{R}; S)$ is a Lipschitz-continuous and T-periodic function for some $T > 0$, with values in S.

Figure 1 is a sketch representing concrete examples compatible with these hypotheses. The axisymmetric case consists of a three-dimensional cone with axis along $\vec{\imath}$ and opening angle $\xi \in [0, 2\pi]$ (an *axisymmetric octopus*). A non-axisymmetric example is also allowed in this discussion, consisting of two rigid parts (*valves*), always maintaining mirror symmetry with respect to a plane (containing $\vec{\imath}$ and perpendicular to it) while swiveling around a joint contained in the symmetry plane and perpendicular to $\vec{\imath}$ (a *mirror-symmetric scallop*), and swimming parallel to $\vec{\imath}$.

Among the systems that are *not* compatible with the assumptions above are those containing helical elements with axis of rotation $\vec{\imath}$, and capable of rotating around $\vec{\imath}$ always in the same direction (call θ the rotation angle). Indeed, a monotone function $t \mapsto \theta(t)$ is not periodic.

The celebrated "scallop theorem" [8] states that, for a system like the one depicted in Fig. 1, the net displacement of the center of mass at the end of a periodic stroke will always vanish. This is due to the linearity of Stokes equation (which leads to symmetry under time reversals), and to the low dimensionality of the system (a one-dimensional periodic stroke is necessarily reciprocal). Thus, whatever forward motion is achieved by the scallop by closing its valves, it will be exactly compensated by a backward motion upon reopening them. Since the low Re world is unaware of inertia, it will not help to close the valves quickly and reopen them slowly. A precise statement and a rigorous short proof of the scallop theorem are given below.

Theorem 1 *Consider a swimmer whose motion is described by*

$$\dot{c}(t) = V(\xi(t))\dot{\xi}(t) \,, \quad t \in \mathbb{R} \,, \tag{7}$$

with $V \in L^1(S)$. *Then for every T-periodic stroke* $\xi \in W^{1,\infty}(\mathbb{R}; S)$, *one has*

$$\Delta c = \int_0^T \dot{c}(t)dt = 0 \,. \tag{8}$$

Proof Define the primitive of V by

$$\Psi(s) = \int_0^s V(\sigma)\,\mathrm{d}\sigma \tag{9}$$

Copyrighted material

so that $\Psi'(\xi) = V(\xi)$. Then, using (7),

$$\begin{aligned}\Delta c &= \int_0^T V(\xi(t))\dot{\xi}(t)\mathrm{d}t \\ &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\Psi(\xi(t))\mathrm{d}t \\ &= \Psi(\xi(T)) - \Psi(\xi(0)) = 0\end{aligned}$$

by the T-periodicity of $t \mapsto \xi(t)$.

Optimal Swimming

A classical notion of swimming efficiency is due to Lighthill [6]. It is defined as the inverse of the ratio between the average power expended by the swimmer during a stroke starting and ending at the shape $\xi_0 = (\xi_0^{(1)}, \ldots, \xi_0^{(N)})$ and the power that an external force would spend to translate the system rigidly at the same average speed $\bar{c} = \Delta c/T$:

$$\mathrm{Eff}^{-1} = \frac{\frac{1}{T}\int_0^T \int_{\partial\Omega} \sigma n \cdot v}{6\pi\eta L \bar{c}^2} = \frac{\int_0^1 \int_{\partial\Omega} \sigma n \cdot v}{6\pi\eta L(\Delta c)^2} \tag{10}$$

where η is the viscosity of the fluid, $L = L(\xi_0)$ is the effective radius of the swimmer, and time has been rescaled to a unit interval to obtain the second identity. The expression in the denominator in (10) comes from a generalized version of Stokes formula giving the drag on a sphere of radius L moving at velocity $\bar{c}$ as $6\pi\eta L\bar{c}$.

Let DN: $H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$ be the Dirichlet to Neumann map of the outer Stokes problem, i. e., the map such that $\sigma n = DNv$, where σ is the stress in the fluid, evaluated on $\partial\Omega$, arising in response to the prescribed velocity v on $\partial\Omega$, and obtained by solving the Stokes problem outside Ω. The expended power in (10) can be written as

$$\int_{\partial\Omega} \sigma n \cdot v = \int_{\partial\Omega} DN(v) \cdot v\,. \tag{11}$$

At a point $p \in \partial\Omega$, the velocity $v(p)$ accompanying a change of state of the swimmer can be written as a linear combination of the $u^{(i)}$

$$v(p) = \sum_{i=1}^{N+1} \mathcal{V}_i(p,\xi)u^{(i)} \tag{12}$$

$$= \sum_{i=1}^{N} \mathcal{W}_i(p,\xi)\dot{\xi}^{(i)}\,. \tag{13}$$

Indeed, the functions $\mathcal{V}_i$ are independent of c by translational invariance, and (4) has been used to get (13) from the line above.

Substituting (13) in (11), the expended power becomes a quadratic form in $\dot{\xi}$

$$\int_{\partial\Omega} \sigma n \cdot v = (G(\xi)\dot{\xi}, \dot{\xi}) \tag{14}$$

where the symmetric and positive definite matrix $G(\xi)$ is given by

$$G_{ij}(\xi) = \int_{\partial\Omega} DN(\mathcal{W}_i(p,\xi)) \cdot \mathcal{W}_j(p,\xi)\mathrm{d}p\,. \tag{15}$$

Strokes of maximal efficiency may be defined as those producing a given displacement Δc of the center of mass with minimal expended power. Thus, from (10), maximal efficiency is obtained by minimizing

$$\int_0^1 \int_{\partial\Omega} \sigma n \cdot v = \int_0^1 (G(\xi)\dot{\xi}, \dot{\xi}) \tag{16}$$

subject to the constraint

$$\Delta c = \int_0^1 V(\xi) \cdot \dot{\xi} \tag{17}$$

among all closed curves $\xi\colon [0,1] \to S$ in the set S of admissible shapes such that $\xi(0) = \xi(1) = \xi_0$.

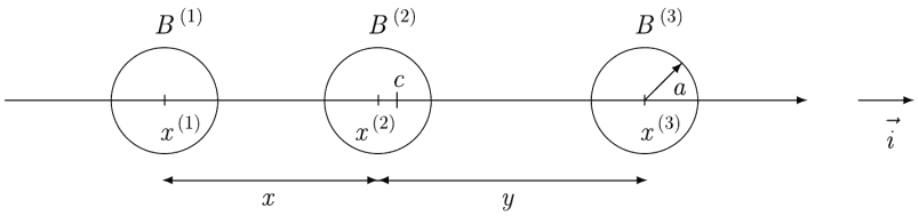
The Euler–Lagrange equations for this optimization problem are

$$-\frac{\mathrm{d}}{\mathrm{d}t}(G\dot{\xi}) + \frac{1}{2}\begin{pmatrix} \left(\frac{\partial G}{\partial \xi^{(1)}}\dot{\xi}, \dot{\xi}\right) \\ \vdots \\ \left(\frac{\partial G}{\partial \xi^{(N)}}\dot{\xi}, \dot{\xi}\right) \end{pmatrix} + \lambda\left(\nabla_\xi V - \nabla_\xi^t V\right)\dot{\xi} = 0 \tag{18}$$

where $\nabla_\xi V$ is the matrix $(\nabla_\xi V)_{ij} = \partial V_i/\partial\xi_j$, $\nabla_\xi^t V$ is its transpose, and λ is the Lagrange multiplier associated with the constraint (17).

Given an initial shape ξ_0 and an initial position c_0, the solutions of (18) are in fact sub-Riemannian geodesics joining the states parametrized by (ξ_0, c_0) and $(\xi_0, c_0 + \Delta c)$ in the space of admissible states $\mathcal{X}$, see [1]. It is well known, and easy to prove using (18), that along such geodesics $(G(\gamma)\dot{\gamma}, \dot{\gamma})$ is constant. This has interesting consequences, because swimming strokes are often divided into a power phase, where $|G(\gamma)|$ is large, and a recovery phase, where $|G(\gamma)|$ is smaller. Thus, along optimal strokes, the recovery phase is executed quickly while the power phase is executed slowly.

Copyrighted material



Biological Fluid Dynamics, Non-linear Partial Differential Equations, Figure 2
Swimmer's geometry and notation

The Three-Sphere Swimmer

For the three-sphere-swimmer of Najafi and Golestanian [7], see Fig. 2, Ω is the union of three rigid disjoint balls $B^{(i)}$ of radius a, shape is described by the distances x and y, the space of admissible shapes is $S = (2a, +\infty)^2$, and the kinematic relation (1) takes the form

$$\begin{aligned} u^{(1)} &= \dot{c} - \frac{1}{3}(2\dot{x} + \dot{y}) \\ u^{(2)} &= \dot{c} + \frac{1}{3}(\dot{x} - \dot{y}) \\ u^{(3)} &= \dot{c} + \frac{1}{3}(2\dot{y} + \dot{x})\,. \end{aligned} \tag{19}$$

Consider, for definiteness, a system with $a = 0.05\,\mathrm{mm}$, swimming in water. Calling $f^{(i)}$ the total propulsive force on ball $B^{(i)}$, we find that the following relation among forces and ball velocities holds

$$\begin{pmatrix} f^{(1)} \\ f^{(2)} \\ f^{(3)} \end{pmatrix} = R(x, y) \begin{pmatrix} u^{(1)} \\ u^{(2)} \\ u^{(3)} \end{pmatrix} \tag{20}$$

where the symmetric and positive definite matrix R is known as the resistance matrix. From this last equation, using also (19), the condition for self-propulsion $f^{(1)} + f^{(2)} + f^{(3)} = 0$ is equivalent to

$$\dot{c} = V_x(x, y)\dot{x} + V_y(x, y)\dot{y}\,, \tag{21}$$

where

$$V_x(x, y) = \frac{Re_c \cdot (e_c \times e_y)}{Re_c \cdot (e_x \times e_y)} \tag{22}$$

$$V_y(x, y) = -\frac{Re_c \cdot (e_c \times e_x)}{Re_c \cdot (e_x \times e_y)}. \tag{23}$$

Moreover, $e_x = (-1, 1, 0)^t$, $e_y = (0, -1, 1)^t$, $e_c = (1/3, 1/3, 1/3)^t$.

Given a stroke $\gamma = \partial\omega$ in the space of admissible shapes, condition (5) for swimming reads

$$\begin{aligned} 0 \neq \Delta c &= \int_0^T \left(V_x\dot{x} + V_y\dot{y}\right) \mathrm{d}t \\ &= \int_\omega \mathrm{curl}\, V(x, y)\mathrm{d}x\mathrm{d}y \end{aligned} \tag{24}$$

Biological Fluid Dynamics, Non-linear Partial Differential Equations, Table 1
Energy consumption (10^{-12} J) for the three strokes of Fig. 3 inducing the same displacement $\Delta c = 0.01$ mm in $T = 1$ s

Optimal stroke	Small square stroke	Large square stroke
0.229	0.278	0.914

which is guaranteed, in particular, if curl V is bounded away from zero. Strokes of maximal efficiency for a given initial shape (x_0, y_0) and given displacement Δc are obtained by solving Eq. (18). For $N = 2$, this becomes

$$-\frac{\mathrm{d}}{\mathrm{d}t}(G\dot{\gamma}) + \frac{1}{2}\begin{pmatrix} (\partial_x G\dot{\gamma}, \dot{\gamma}) \\ (\partial_y G\dot{\gamma}, \dot{\gamma}) \end{pmatrix} + \lambda \mathrm{curl}\, V(\gamma)\dot{\gamma}^\perp = 0 \tag{25}$$

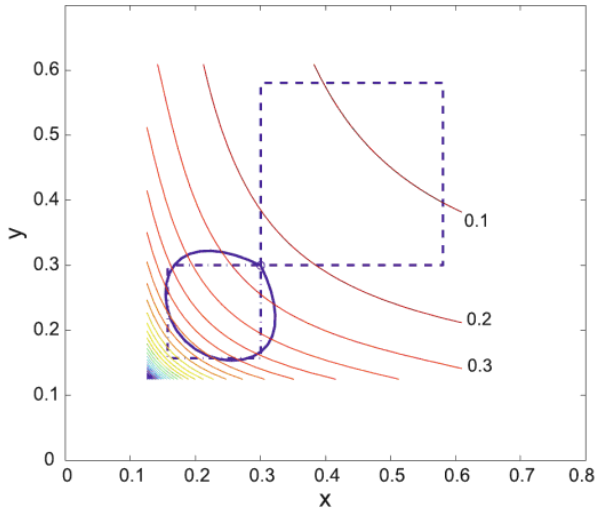
where $\partial_x G$ and $\partial_y G$ stand for the x and y derivatives of the 2×2 matrix $G(x, y)$.

It is important to observe that, for the three-sphere swimmer, all hydrodynamic interactions are encoded in the shape dependent functions $V(x, y)$ and $G(x, y)$. These can be found by solving a two-parameter family of outer Stokes problems, where the parameters are the distances x and y between the three spheres. In [1], this has been done numerically via the finite element method: a representative example of an optimal stroke, compared to two more naive proposals, is shown in Fig. 3.

Future Directions

The techniques discussed in this article provide a head start for the mathematical modeling of microscopic swimmers, and for the quantitative optimization of their strokes. A complete theory for axisymmetric swimmers is already available, see [2], and further generalizations to arbitrary shapes are relatively straightforward. The combination of numerical simulations with the use of tools from sub-Riemannian geometry proposed here may prove extremely valuable for both the question of adjusting the stroke to global optimality criteria, and of optimizing the stroke of complex swimmers. Useful inspiration can come from the sizable literature on the related field dealing with control of swimmers in a perfect fluid.

Copyrighted material



Biological Fluid Dynamics, Non-linear Partial Differential Equations, Figure 3

Optimal stroke and square strokes which induce the same displacement $\Delta c = 0.01\,\text{mm}$ in $T = 1\,\text{s}$, and equally spaced level curves of curl V. The small circle locates the initial shape $\xi_0 = (0.3\,\text{mm}, 0.3\,\text{mm})$

Bibliography

Primary Literature

1. Alouges F, DeSimone A, Lefebvre A (2008) Optimal strokes for low Reynolds number swimmers: an example. J Nonlinear Sci 18:277–302
2. Alouges F, DeSimone A, Lefebvre A (2008) Optimal strokes for low Reynolds number axisymmetric swimmers. Preprint SISSA 61/2008/M
3. Avron JE, Kenneth O, Oakmin DH (2005) Pushmepullyou: an efficient micro-swimmer. New J Phys 7:234-1–8
4. Becker LE, Koehler SA, Stone HA (2003) On self-propulsion of micro-machines at low Reynolds numbers: Purcell's three-link swimmer. J Fluid Mechanics 490:15–35
5. Berg HC, Anderson R (1973) Bacteria swim by rotating their flagellar filaments. Nature 245:380–382
6. Lighthill MJ (1952) On the Squirming Motion of Nearly Spherical Deformable Bodies through Liquids at Very Small Reynolds Numbers. Comm Pure Appl Math 5:109–118
7. Najafi A, Golestanian R (2004) Simple swimmer at low Reynolds numbers: Three linked spheres. Phys Rev E 69:062901-1–4
8. Purcell EM (1977) Life at low Reynolds numbers. Am J Phys 45:3–11
9. Tan D, Hosoi AE (2007) Optimal stroke patterns for Purcell's three-link swimmer. Phys Rev Lett 98:068105-1–4
10. Taylor GI (1951) Analysis of the swimming of microscopic organisms. Proc Roy Soc Lond A 209:447–461

Books and Reviews

Agrachev A, Sachkov Y (2004) Control Theory from the Geometric Viewpoint. In: Encyclopaedia of Mathematical Sciences, vol 87, Control Theory and Optimization. Springer, Berlin

Childress S (1981) Mechanics of swimming and flying. Cambridge University Press, Cambridge

Happel J, Brenner H (1983) Low Reynolds number hydrodynamics. Nijhoff, The Hague

Kanso E, Marsden JE, Rowley CW, Melli-Huber JB (2005) Locomotion of Articulated Bodies in a Perfect Fluid. J Nonlinear Sci 15: 255–289

Koiller J, Ehlers K, Montgomery R (1996) Problems and Progress in Microswimming. J Nonlinear Sci 6:507–541

Montgomery R (2002) A Tour of Subriemannian Geometries, Their Geodesics and Applications. AMS Mathematical Surveys and Monographs, vol 91. American Mathematical Society, Providence

Copyrighted material

Catastrophe Theory

WERNER SANNS
University of Applied Sciences,
Darmstadt, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Example 1: The Eccentric Cylinder on the Inclined Plane
Example 2: The Formation of Traffic Jam
Unfoldings
The Seven Elementary Catastrophes
The Geometry of the Fold and the Cusp
Further Applications
Future Directions
Bibliography

Glossary

Singularity Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable map defined in some open neighborhood of the point $p \in \mathbb{R}^n$ and $J_p f$ its Jacobian matrix at p, consisting of the partial derivatives of all components of f with respect to all variables. f is called singular in p if rank $J_p f < \min\{n, m\}$. If rank $J_p f = \min\{m, n\}$, then f is called regular in p. For $m = 1$ (we call the map $f\colon \mathbb{R}^n \to \mathbb{R}$ a differentiable function) the definition implies: A differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is singular in $p \in \mathbb{R}^n$, if grad $f(p) = 0$. The point p, where the function is singular, is called a singularity of the function. Often the name "singularity" is used for the function itself if it is singular at a point p. A point where a function is singular is also called critical point. A critical point p of a function f is called a degenerated critical point if the Hessian (a quadratic matrix containing the second order partial derivatives) is singular at p; that means its determinant is zero at p.

Diffeomorphism A diffeomorphism is a bijective differentiable map between open sets of $\mathbb{R}^n$ whose inverse is differentiable, too.

Map germ Two continuous maps $f\colon U \to \mathbb{R}^k$ and $g\colon V \to \mathbb{R}^k$, defined on neighborhoods U and V of $p \in \mathbb{R}^n$, are called equivalent as germs at p if there exists a neighborhood $W \subset U \cap V$ on which both coincide. Maps or functions, respectively, that are equivalent as germs can be considered to be equal regarding local features. The equivalence classes of this equivalence relation are called germs. The set of all germs of differentiable maps $\mathbb{R}^n \to \mathbb{R}^k$ at a point p is named $\varepsilon_p(n, k)$. If p is the origin of $\mathbb{R}^n$, one simply writes $\varepsilon(n, k)$ instead of $\varepsilon_0(n, k)$. Further, if $k = 1$, we write $\varepsilon(n)$ instead of $\varepsilon(n, 1)$ and speak of function germs (also simplified as "germs") at the origin of $\mathbb{R}^n$. $\varepsilon(n, k)$ is a vector space and $\varepsilon(n)$ is an algebra, that is, a vector space with a structure of a ring. The ring $\varepsilon(n)$ contains a unique maximal ideal $\mu(n) = \{f \in \varepsilon(n) | f(0) = 0\}$. The ideal $\mu(n)$ is generated by the germs of the coordinate functions $x_1, \ldots, x_n$. We use the form $\mu(n) = \langle x_1, \ldots, x_n \rangle \varepsilon(n)$ to emphasize, that this ideal is generated over the ring $\varepsilon(n)$. So a function germ in $\mu(n)$ is of the form $\sum_{i=1}^{n} a_i(x) \cdot x_i$, with certain function germs $a_i(x) \in \varepsilon(n)$.

r-Equivalence Two function germs $f, g \in \varepsilon(n)$ are called r-equivalent if there exists a germ of a local diffeomorphism $h\colon \mathbb{R}^n \to \mathbb{R}^n$ at the origin, such that $g = f \circ h$.

Unfolding An unfolding of a differentiable function germ $f \in \mu(n)$ is a germ $F \in \mu(n+r)$ with $F|\mathbb{R}^n = f$ (here $|$ means the restriction. The number r is called an unfolding dimension of f. An unfolding F of a germ f is called universal if every other unfolding of f can be received by suitable coordinate transformations, "morphisms of unfoldings" from F, and the number of unfolding parameters of F is minimal (see "codimension").

Unfolding morphism Suppose $f \in \varepsilon(n)$ and $F\colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ and $G\colon \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ be unfoldings of f. A right-morphism from F to G, also called unfolding morphism, is a pair (Φ, α), with $\Phi \in \varepsilon(n + k, n + r)$ and $\alpha \in \mu(k)$, such that:

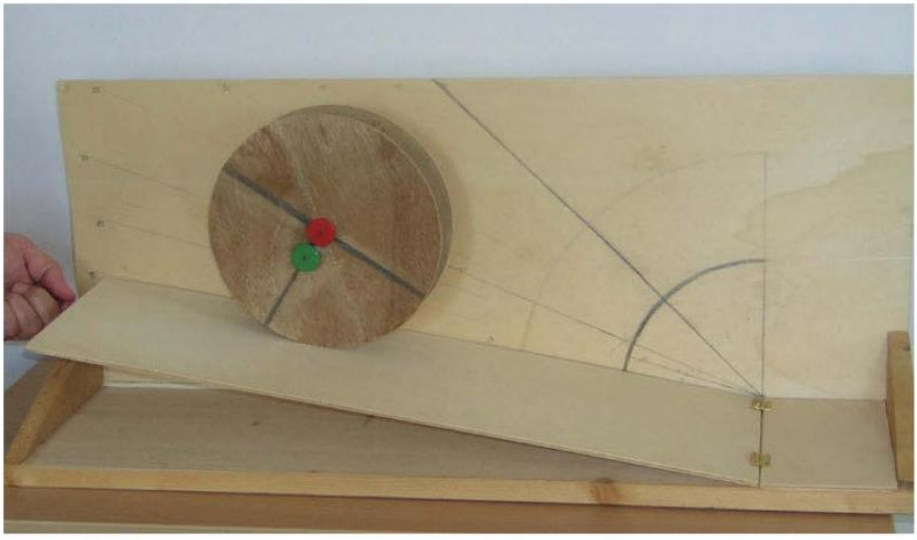
1. $\Phi|\mathbb{R}^n = \mathrm{id}(\mathbb{R}^n)$, that is $\Phi(x, 0) = (x, 0)$,
2. If $\Phi = (\phi, \psi)$, with $\phi \in \varepsilon(n + k, n)$, $\psi \in \varepsilon(n + k, r)$, then $\psi \in \varepsilon(k, r)$,
3. For all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^k$ we get $F(x, u) = G(\Phi(x, u)) + \alpha(u)$.

Catastrophe A catastrophe is a universal unfolding of a singular function germ. The singular function germs are called organization centers of the catastrophes.

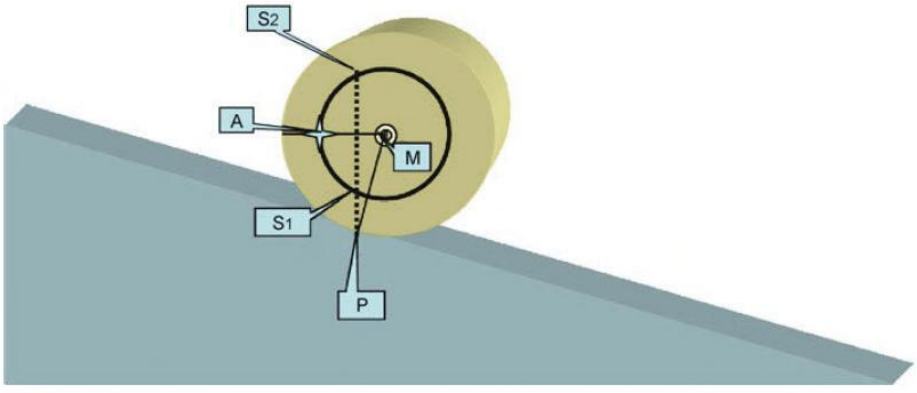
Codimension The codimension of a singularity f is given by $\mathrm{codim}(f) = \dim_{\mathbb{R}} \mu(n)/\langle \partial_x f \rangle$ (quotient space). Here $\langle \partial_x f \rangle$ is the Jacobian ideal generated by the partial derivatives of f and $\mu(n) = \{f \in \varepsilon(n) | f(0) = 0\}$. The codimension of a singularity gives of the minimal number of unfolding parameters needed for the universal unfolding of the singularity.

Potential function Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable map (that is, a differentiable vector field). If there exists a function $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ with the property that

Copyrighted material



Catastrophe Theory, Figure 2
The eccentric cylinder on the inclined plane. The inclination of the plane can be adjusted by hand. The position of the disc center and the center of gravity are marked by colored buttons

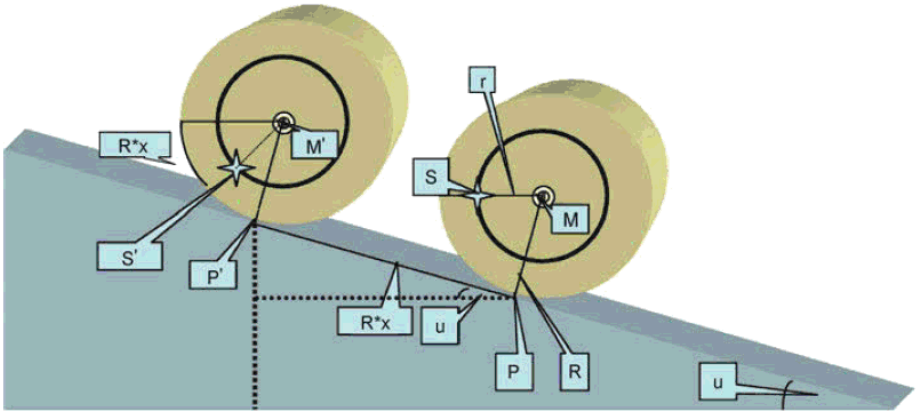


Catastrophe Theory, Figure 3
Two equilibrium positions (S_1 and S_2) are possible with this configuration. If the point A lies "too close" to M, there is no equilibrium position possible at all when releasing the wheel, since the center of gravity can never lie vertically over P

with some sticks of about 5 cm in length. The eccentric center of gravity is generated by fixing a heavy piece of metal between the two discs at some distance from the axes of rotation. Draw some lines on the disks, where you can read or estimate the angle of rotation. (Hint: It could be interesting for students to experimentally determine the center of gravity of the construction.)

If mass and radius are fixed, there are two parameters which determine the system: The position of the cylinder can be specified by the angle x against a zero level (that is, against the horizontal plane; Fig. 4). This angle is the "inner" variable of the system, which means that it is the "answer" of the cylinder to the experimenter's changes of the inclination u of the plane, which is the "outer" parameter. In this example you should determine experimentally the equilibrium positions of the cylinder, that is, the angle of rotation where the cylinder stays at rest on the inclined plane (after releasing it cautiously and after letting the system level off a short time). We are interested in the stable positions, that is, the positions where the cylinder returns after slight disturbances (slightly nudging it or slightly changing the inclination of the plane) without rolling down the whole plane. The search for the equilibrium position x varies with the inclination angle u. You can make a table and a graphic showing the dependence of x and u.

Copyrighted material



Catastrophe Theory, Figure 4
The movement of S is a composite of the movement of the cylinder represented by the movement of point M and the rotation movement with angle *x*

The next step is to find an analytical approach. Refer to Fig. 3: Point P is the supporting point of the cylinder on the plane. Point A marks the starting position of the center of gravity. When the wheel moves and the distance between M and A is big enough, this center can come into two equilibrium positions S_1 and S_2, which lie vertically over P.

Now, what does the potential function look like? The variables are:

***R*:** radius of the discs,

***u*:** angle of the inclined plane against the horizontal plane measured in radians,

***x*:** angle of rotation of the cylinder measured against the horizontal plane in radians,

***r*:** distance between the center of gravity S of the cylinder from the axes marked by point M,

x_0 (resp. u_0): angles where the cylinder suddenly starts to roll down the plane.

Refer to Fig. 4. If the wheel is rolled upward by hand from the lower position on the right to the upper position, the center of gravity moves upward with the whole wheel, or as the midpoint M or the point P. But point S also moves downward by rotation of the wheel with an amount of $r \cdot \sin(x)$.

The height gained is

$$h = R \cdot x \cdot \sin(u) - r \cdot \sin(x) \; . \qquad (1)$$

From physics one knows the formula for the potential energy E_{pot} of a point mass. It is the product $m \cdot g \cdot h$, where m = mass, g = gravitational constant, h = height, where the height is to be expressed by the parameters x and u and the given data of the cylinder.

To a fixed angle u_1 of the plane, by formula (1), there corresponds the following potential function:

$$f(x) = E_{\text{pot}} = m \cdot g \cdot h = m \cdot g \cdot (R \cdot x \cdot \sin(u_1) - r \cdot \sin(x)) \; .$$

If the inclination of the plane is variable, the angle u of inclination serves as an outer parameter and we obtain a whole family of potential functions as a function F in two variables x and u:

$$F(x, u) = F_u(x) = m \cdot g \cdot (R \cdot x \cdot \sin(u) - r \cdot \sin(x)) \; .$$

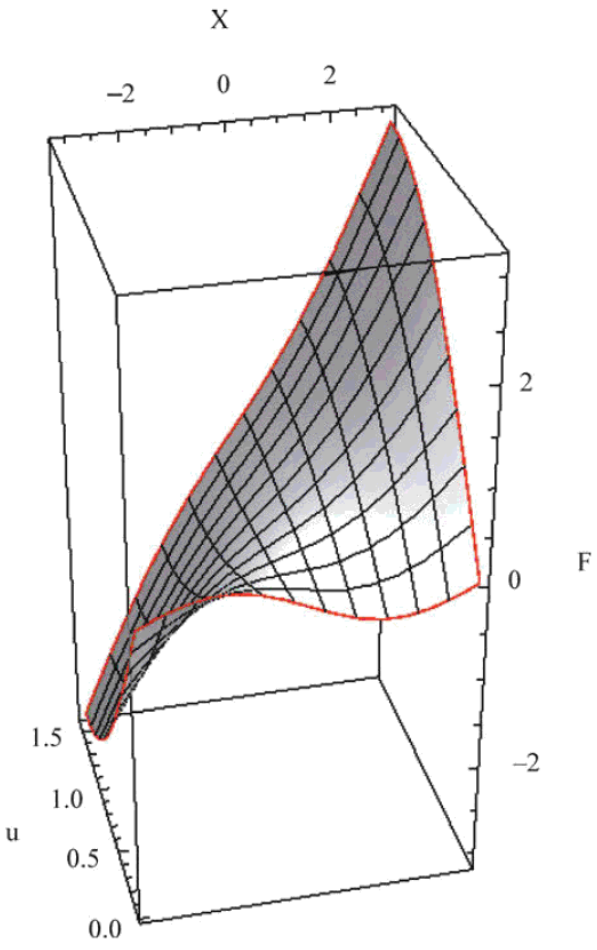
Thus, the behavior of the eccentric cylinder is described by a family of potential functions. The extrema of the single potentials characterize the equilibria of the system. The saddle point characterizes the place where the cylinder suddenly begins to roll down due to the slightest disturbance, that is, where the catastrophe begins (see Fig. 5).

Example. The data are $m = 1\,\text{kg}$, $g = 10\,\text{m/sec}^2$, $R = 0.1\,\text{m}$, $r = 0.025\,\text{m}$.

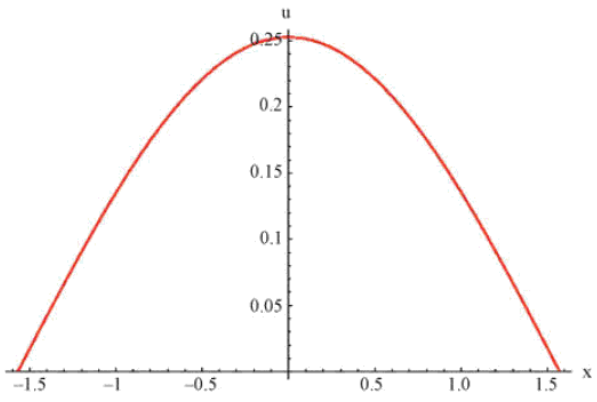
For constant angle of inclination u, each of the functions $f(x) = F(x, u = \text{const})$ is a section of the function F. To determine the minima of each of the section graphs, we calculate the zeros of the partial derivative of F with respect to the inner variable x. In our particular example above, the result is $\partial_x F = -0.25\cos(x) + \sin(u)$. The solution for $\partial_x F = 0$ is $u = \arcsin(0.25\cos(x))$ (see Fig. 6).

Calculating the point in x–u-space at which both partial derivatives $\partial_x F$ and $\partial_{xx} F$ vanish (saddle point of the section) and using only positive values of angle u gives the point $(x_0, u_0) = (0, 0.25268)$. In general: $x_0 = 0$ and $u_0 = \arcsin(r/R)$. In our example this means the catastrophe begins at an inclination angle of $u_0 = 0.25268$ (about

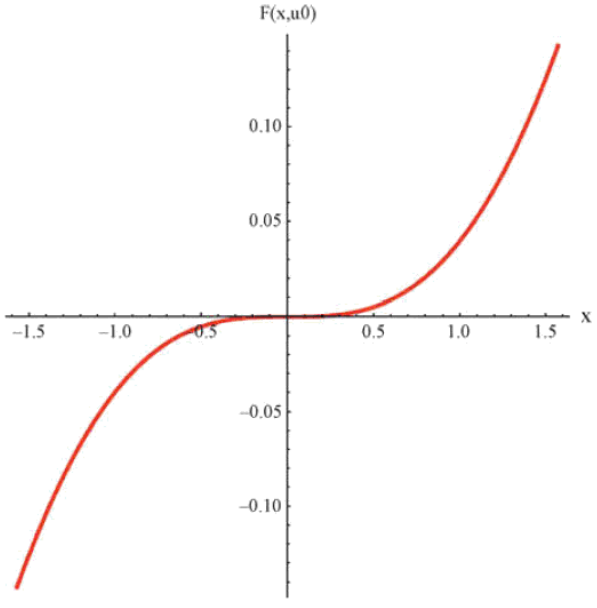
Copyrighted material



Catastrophe Theory, Figure 5
Graph of $F(x, u)$ for a concrete cylinder



Catastrophe Theory, Figure 6
Curve in x–u-space, for which $F(x, u)$ has local extrema



Catastrophe Theory, Figure 7
Graph of the section $F(x, u_0)$

14.5 degrees). At that angle, sections of F with $u > u_0$ do not possess any minimum where the system could stably rest. Making u bigger than u_0 means that the cylinder must leave its stable position and suddenly roll down the plane. The point u_0 in the u-parameter space is called catastrophe point.

What is the shape of the section of F at u_0? Since in our example

$$f(x) = F(x, u_0) = 0.25x - 0.25\sin(x) ,$$

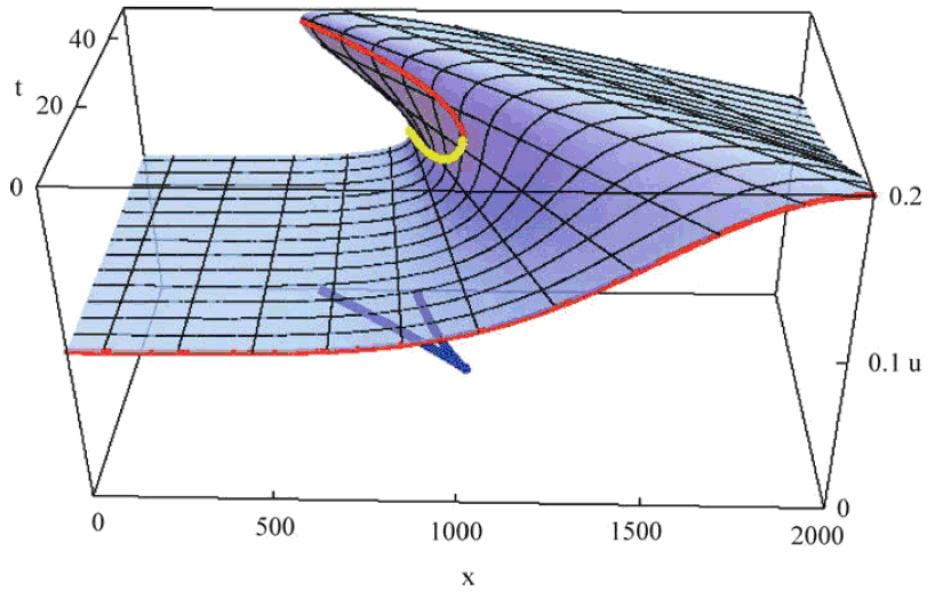
the graph appears as shown in Fig. 7.

The point $x = 0$ is a degenerated critical point of $f(x)$, since $f'(x) = f''(x) = 0$. If we expand f into its Taylor polynomial $t(x)$ around that point we get

$$t(x) = 0.0416667x^3 - 0.00208333x^5 + 0.0000496032x^7 - 6.88933 \times 10^{-7}x^9 + O[x]^{11} .$$

Catastrophe theory tries to simplify the Taylor polynomials of the potentials locally about their degenerated critical points. This simplification is done by coordinate transformations, which do not change the structure of the singularities. These transformations are called diffeomorphisms (see Sect. "Glossary"). A map $h\colon U \to V$, with U and V open in $\mathbb{R}^n$, is a local diffeomorphism if the Jacobian determinant $\det(J_p h) \neq 0$ at all points p in U. Two functions f and g, both defined in some neighborhood of the origin of $\mathbb{R}^n$, are called r-equivalent if there exists a local diffeomorphism $h\colon U \to V$ at the origin, such that $g = f \circ h$.

Copyrighted material



Catastrophe Theory, Figure 8
Characteristic surface *S*, initial density, fold curve and cusp curve

The Taylor polynomial $t(x)$ in our example starts with a term of 3rd order in x. Thus, there exists a diffeomorphism (a non-singular coordinate transformation), such that in the new coordinates t is of the form x^3. To see this, we write $t(x) = x^3 \cdot a(x)$, with $a(0) \neq 0$. We define $h(x) = x \cdot a(x)^{1/3}$. Note that the map h is a local diffeomorphism near the origin since $h'(0) = a(0)^{1/3} \neq 0$. (Hint: Modern computer algebra systems, such as Mathematica, are able to expand expressions like $a(x)^{1/3}$ and calculate the derivative of h easily for the concrete example.)

Thus for the eccentric cylinder the intersection of the graph of the potential $F(x, u)$ with the plane $u = u_0$ is the graph of a function, whose Taylor series at its critical point $x = 0$ is $t = x^3 \circ h$, in other words: t is r-equivalent to x^3.

Before we continue with this example and its relation to catastrophe theory, let us introduce another example which leads to a potential of the form x^4. Then we will have two examples which lead to the two simplest catastrophes in René Thom's famous list of the seven elementary catastrophes.

Example 2: The Formation of Traffic Jam

In the first example we constructed a family F of potential functions directly from the geometric properties of the model of the eccentric cylinder. We then found its critical points by calculating the derivatives of F. In this second example the method is slightly different. We start with a partial differential equation which occurs in traffic flow modeling. The solution surface of the initial value problem will be regarded as the surface of zeros of the derivative of a potential family. From this, we then calculate the family of potential functions. Finally, we will see that the Taylor series of a special member of this family, the one which belongs to its degenerated critical point, is equivalent to x^4.

Note: When modeling a traffic jam we use the names of variables that are common in literature about this subject. Later we will rename the variables into the ones commonly used in catastrophe theory.

Let $u(x, t)$ be the (continuous) density of the traffic at a street point x at time t and let $a(u)$ be the derivative of the traffic flow $f(u)$. The mathematical modeling of simple traffic problems leads to the well known traffic flow equation:

$$\frac{\partial u(x,t)}{\partial t} + a(u)\frac{\partial u(x,t)}{\partial x} = 0 \,.$$

We must solve the Cauchy problem for this quasilinear partial differential equation of first order with given initial density $u_0 = u(x, 0)$ by the method of characteristics. The solution is constant along the (base) characteristics, which are straight lines in xt-plane with slope $a(u_0(y))$ against the t-axes, if y is a point on the x-axes, where the characteristic starts at $t = 0$. For simplicity we write "a" instead of $a(u_0(y))$. Thus the solution of the initial value problem is $u(x, t) = u_0(x - a \cdot t)$ and the characteristic surface $S = \{(x, t, u) | u - u_0(x - a \cdot t) = 0\}$ is a surface in xtu-space. Under certain circumstances S may not lie uniquely above the xt-plane but may be folded as is shown in Fig. 8.

Copyrighted material

There is another curve to be seen on the surface: The border curve of the fold. Its projection onto the xt-plane is a curve with a cusp, specifying the position and the beginning time of the traffic jam.

The equations used in the example, in addition to the traffic flow equation, can be deduced in traffic jam modeling from a simple parabolic model for a flow-density relation (see [5]). The constants in the following equations result from considerations of maximal possible density and an assumed maximal allowed velocity on the street. In our example a model of a road of 2 km in length and maximal traffic velocity of 100 km/h is chosen. Maximal traffic density is $u_{\max} = 0.2$. The equations are $f(u) = 27.78u - 138.89u^2$ and $u(x,0) = u_0(x) = 0.1 + 0.1 \cdot \mathrm{Exp}(-((x - 2000)/700)^2)$. The latter was constructed to simulate a slight increase of the initial density u_0 along the street at time $t = 0$. The graph of u_0 can be seen as the front border of the surface.

To determine the cusp curve shown in Fig. 8, consider the following parameterization Φ of the surface S and the projection map $\pi : S \to \mathbb{R}^2$:

$$\Phi : \mathbb{R} \times \mathbb{R}_{\geq 0} \to S \subset \mathbb{R}^3$$
$$(x,t) \mapsto (x + a(u_0(x)) \cdot t,\, t,\, u_0(x)) \,.$$

The Jacobian matrix of $\pi \circ \Phi$ is

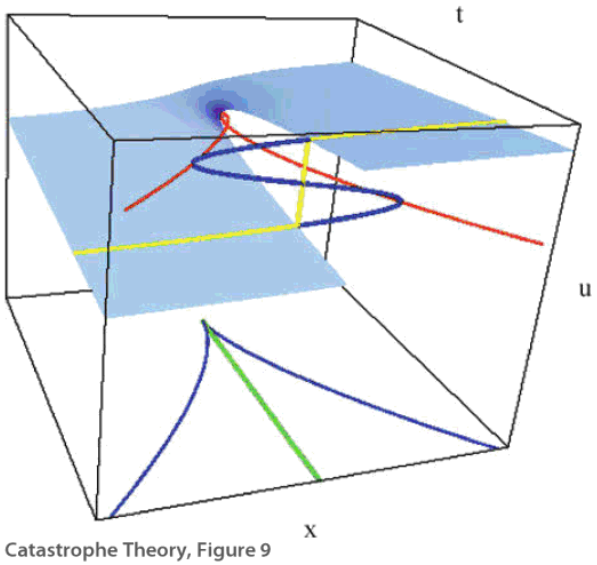
$$J(\pi \circ \Phi)(x,t) = \begin{pmatrix} \frac{\partial(\pi \circ \Phi)_1}{\partial x} & \frac{\partial(\pi \circ \Phi)_1}{\partial t} \\ \frac{\partial(\pi \circ \Phi)_2}{\partial x} & \frac{\partial(\pi \circ \Phi)_2}{\partial t} \end{pmatrix} = \begin{pmatrix} 1 + a(u_0(x))' \cdot t & a(u_0(x)) \\ 0 & 1 \end{pmatrix} .$$

It is singular (that is, its determinate vanishes) if $1 + (a(u_0(x)))' \cdot t = 0$. From this we find the curve $c : \mathbb{R} \to \mathbb{R}^2$

$$c(x) = \left(x, -\frac{1}{a'(u_0(x)) \cdot u_0'(x)} \right)$$

which is the pre-image of the cusp curve and the cusp curve in the Fig. 8 is the image of c under $\pi \circ \Phi$.

Whether the surface folds or not depends on both the initial density u_0 and the properties of the flow function $f(u)$. If such folding happens, it would mean that the density would have three values at a given point in xt-space, which is a physical impossibility. Thus, in this case, no continuous solution of the flow equation can exist in the whole plane. "Within" the cusp curve, which is the boundary for the region where the folding happens, there exists a curve with its origin in the cusp point. Along this curve the surface must be "cut" so that a discontinuous solution occurs. The curve within the cusp region is called



Catastrophe Theory, Figure 9
The solution surface is "cut" along the shock curve, running within the cusp region in the xt-plane

a shock curve and it is characterized by physical conditions (jump condition, entropy condition). The density makes a jump along the shock.

The cusp, together with the form of the folding of the surface, may give an association to catastrophe theory. One has to find a family of potential functions for this model, that is, functions $F_{(x,t)}(u)$, whose negative gradient, that is, its derivative by u (the inner parameter), describes the characteristic surface S by its zeroes. The family of potentials is given by

$$F(x,t,u) = -t \cdot \left(u \cdot a(u) - \int a(u)\mathrm{d}u \right) - \int_0^{x - a(u) \cdot t} u_0(x)\mathrm{d}x$$

(see [6]). Since

$$-\mathrm{grad} F_{x,t}(u) = -\frac{\partial F}{\partial u} = 0 \Leftrightarrow u - u_0(x - a(u) \cdot t) = 0$$

the connection of F and S is evident. One can show that a member f of the family F, the function $f(u) = F(\xi, \tau, u)$, is locally in a traffic jam formation point (ξ, τ) of the form $f(u) = u^4$ (after Taylor expansion and after suitable coordinate transformations, similar to the first example).

The complicated function terms of f and F thus can be replaced in qualitative investigations by simple polynomials. From the theorems of catastrophe theory, more properties of the models under investigation can be discovered.

Now we want to use the customary notation of catastrophe theory. We shall rename the variables: u is the inner variables, which in catastrophe theory usually is x, and x, t (the outer parameters) get the names u and v respectively.

Copyrighted material

The Seven Elementary Catastrophes

Elementary catastrophe theory works with families (unfoldings) of potential functions

$$F\colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R} \quad (\text{with } k \le 4)$$
$$(x,u) \mapsto F(x,u)\ .$$

$\mathbb{R}^n$ is called the state space, $\mathbb{R}^k$ is called the parameter space or control space. Accordingly $x = (x_1,\dots,x_n)$ is called a state variable or endogenous variable and $u = (u_1,\dots,u_k)$ is called a control variable (exogenous variable, parameter).

We are interested in the singularities of F with respect to x and the dependence of F with respect to the control variables u. We regard F as the unfolding of a function germ f and want to classify F.

In analogy to the r-equivalence of function germs, we call two unfoldings F and G of the same function germ r-equivalent as unfoldings if there exists an unfolding-morphism (see Sect. "Glossary") from F to G. An unfolding F of f is called versal (or stable), if each other unfolding of f is right-equivalent as unfolding to F, that is, if there exists a right-morphism between the unfoldings. A versal unfolding with minimal unfolding dimensions is called universal unfolding of f.

Thus we have the following theorem (all proofs of the theorems cited in this article can be found in [1,2] or [12].)

Theorem 1 (Theorem on the existence of a universal unfolding) *A singularity f has a universal unfolding iff $codim(f) = k < \infty$.*

Examples (see also the following theorem with its examples):

$$\text{If } f(x) = x^3 \quad \text{it follows}$$
$$F(x,u,v) = x^3 + ux^2 + vx \quad \text{is versal,}$$
$$F(x,u) = x^3 + ux \quad \text{is universal,}$$
$$F(x,u) = x^3 + ux^2 \quad \text{is not versal .}$$

The following theorem states how one can find a universal unfolding of a function germ:

Theorem 2 (Theorem on the normal form of universal unfoldings) *Let $f \in \mu(n)$ be a singularity with $codim(f) = k < \infty$. Let $u = (u_1,\dots,u_k)$ be the parameters of the unfolding. Let $b_i(x)$, $i = 1,\dots,k$, be the elements of $\mu(n)$, whose cosets modulo $\langle \partial_x f\rangle$ generate the vector space $\mu(n)/\langle \partial_x f\rangle$. Then*

$$F(x,u) = f(x) + \sum_{i=1}^{k} u_i b_i(x)$$

is a universal unfolding of f.

Examples:

1. Consider $f(x) = x^4$. Here $n = 1$ and $\mu(n) = \mu(1) = \langle x\rangle$. The derivative of f is $4x^3$, so $\langle \partial_x f\rangle = \langle x^3\rangle$ and we get $\langle x\rangle/\langle x^3\rangle = \langle x, x^2\rangle$. Thus $F(x,u_1,u_2) = x^4 + u_1x^2 + u_2x$ is the universal unfolding of f.
2. Consider $f(x,y) = x^3 + y^3$. Here $n = 2$ and $\mu(n) = \mu(2) = \langle x, y\rangle$. The partial derivatives of f with respect to x and y are $3x^2$ and $3y^2$, thus $\langle \partial_x f\rangle = \langle x^2, y^2\rangle$ and we get $\langle x,y\rangle/\langle x^2,y^2\rangle = \langle x, xy, y\rangle$. Therefore $F(x,y,u_1,u_2,u_3) = x^3 + y^3 + u_1\cdot x\cdot y + u_2\cdot x + u_3\cdot y$ is the universal unfolding of f.

Hint: There exists a useful method for the calculation of the quotient space which you will find in the literature as "Siersma's trick" (see for example [7]).

We now present René Thom's famous list of the seven elementary catastrophes.

Classification Theorem (Thom' s List)

Up to addition of a non degenerated quadratic form in other variables and up to multiplication by ± 1 a singularity f of codimension k ($1 \le k \le 4$) is right-equivalent to one of the following seven:

f	codim f	universal unfolding	name
x^3	1	$x^3 + ux$	fold
x^4	2	$x^4 + ux^2 + vx$	cusp
x^5	3	$x^5 + ux^3 + vx^2 + wx$	swallowtail
$x^3 + y^3$	3	$x^3 + y^3 + uxy + vx + wy$	hyperbolic umbilic
$x^3 - xy^2$	3	$x^3 - xy^2 + u(x^2 + y^2) + vx + wy$	elliptic umbilic
x^6	4	$x^6 + ux^4 + vx^3 + wx^2 + tx$	butterfly
$x^2y + y^4$	4	$x^2y + y^4 + ux^2 + vy^2 + wx + ty$	parabolic umbilic

In a few words: The seven elementary catastrophes are the seven universal unfoldings of singular function germs of codimension k ($1 \le k \le 4$). The singularities itself are called organization centers of the catastrophes.

We will take a closer look at the geometry of the simplest of the seven elementary catastrophes: the fold and the cusp. The other examples are discussed in [9].

The Geometry of the Fold and the Cusp

The fold catastrophe is the first in Thom's list of the seven elementary catastrophes. It is the universal unfolding of the singularity $f(x) = x^3$ which is described by the equation $F(x,u) = x^3 + u\cdot x$. We are interested in the set of singular points of F relative to x, that is, for those points where the first partial derivative with respect to x (the sys-

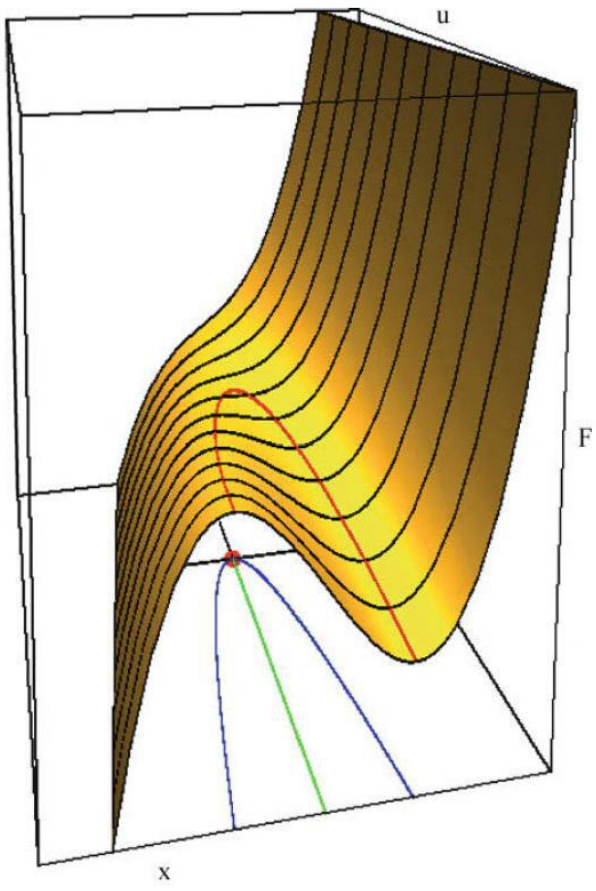
Copyrighted material

tem variable) vanishes. This gives the parabola $u = -3x^2$. Inserting this into F gives a space curve.

Figure 13 shows the graph of $F(x, u)$. The curve on the surface joins the extrema of F and the projection curve represents their x, u values. There is only one degenerated critical point of F (at the vertex of the Parabola), that is, a critical point where the second derivative by x also vanishes. The two branches of the parabola in the xu-plane give the positions of the maxima and the minima of F, respectively. At the points belonging to a minimum, the system is stable, while in a maximum it is unstable. Projecting x–u space, and with it the parabola, onto parameter space (u-axis), one gets a straight line shown within the interior of the parabola, which represents negative parameter values of u. There, the system has a stable minimum (and an unstable maximum, not observable in nature). For parameter values $u > 0$ the system has no stability points at all. Interpreting the parameter u as time, and "walking along" the u-axis, the point $u = 0$ is the beginning or the end of a stable system behavior. Here the system shows catastrophic behavior. According to the interpretation of the external parameter as time or space, the morphology of the fold is a "beginning," an "end" or a "border", where something new occurs.

What can we say about the fold catastrophe and the modeling of our first example, the eccentric cylinder? We have found that the potential related to the catastrophe point of the eccentric cylinder is equivalent to x^3, and $F(x, u) = x^3 + ux$ is the universal unfolding. This is independent of the size of our 'machine'. The behavior of the machine should be qualitatively the same as that of other machines which are described by the fold catastrophe as a mechanism. We can expect that the begin (or end) of a catastrophe depends on the value of a single outer parameter. Stable states are possible (for $u < 0$ in the standard fold. In the untransformed potential function of our model of the eccentric cylinder, this value is $u = 0.25268$) corresponding to the local minima of $F(., u)$. This means that it is possible for the cylinder to stay at rest on the inclined plane. There are unstable equilibria (local maxima of $F(., u)$ and the saddle point, too). The cylinder can be turned such that the center of gravity lies in the upper position over the supporting point (see Fig. 3). But a tiny disturbance will make the system leave this equilibrium.

Let us now turn to the cusp catastrophe. The cusp catastrophe is the universal unfolding of the function germ $f(x) = x^4$. Its equation is $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$. In our second example (traffic jam), the function f is equivalent to the organization center of the cusp catastrophe, the second in René Thom's famous list of the seven elementary catastrophes. So the potential family is equivalent to

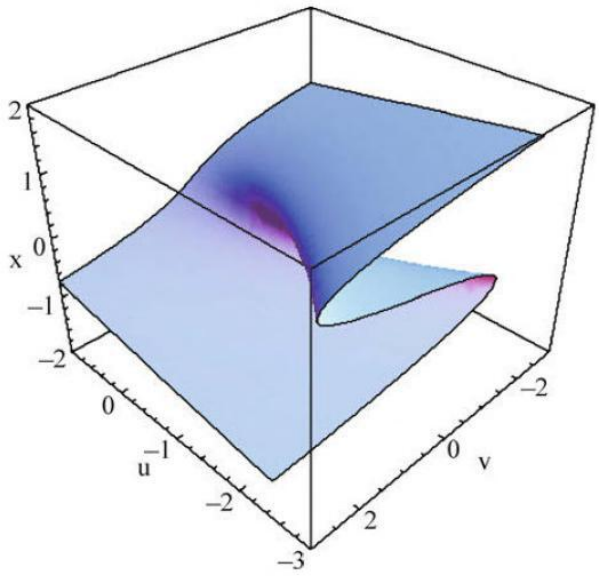


Catastrophe Theory, Figure 13
Graph of $F(x, u) = x^3 + u \cdot x$

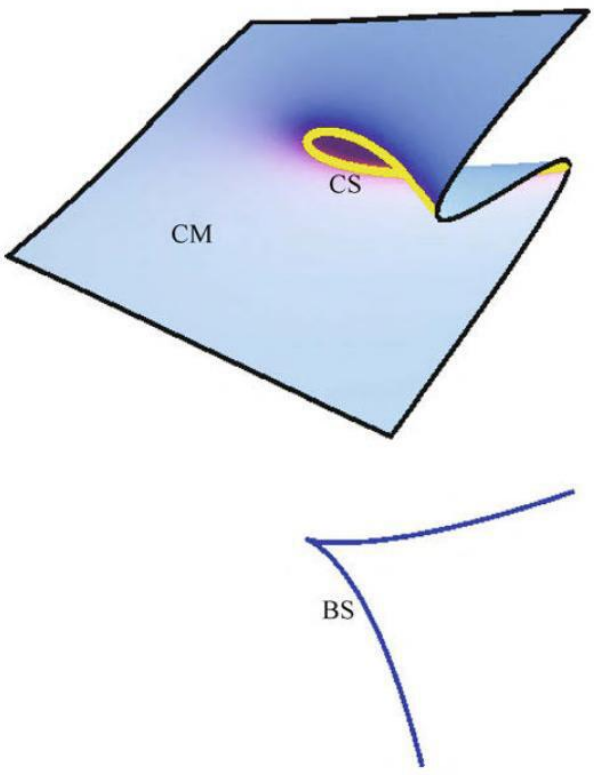
the universal unfolding $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$ of the function x^4. We cannot draw such a function in three variables in three-space, but we can try to draw sections giving u or v certain constant values. We want to look at its catastrophe surface S, that is, the set of those points in (x, u, v)-space, where the partial derivative of F with respect to the system variable x is zero. It is a surface in three-space called catastrophe manifold, stability surface or equilibrium surface, since the surface S describes the equilibrium points of the system.

If S is not folded over a point in u, v-space there is exactly one minimum of the potential F. If it is folded in three sheets, there are two minima (corresponding to the upper and lower sheet of the fold) and a maximum (corresponding to the middle sheet). Stable states of the system belong to the stable minima, that is, to the upper and lower sheets of the folded surface. Other points in three-space that do not lie on the surface S correspond to the states of the system that are not equilibrium. The system does not rest there. Catastrophes do occur when the

Copyrighted material



Catastrophe Theory, Figure 14
The catastrophe manifold of the cusp catastrophe



Catastrophe Theory, Figure 15
The catastrophe manifold (CM), catastrophe set (CS) and bifurcation set (BS) of the cusp catastrophe

system jumps suddenly from one stable state to another stable state, that is, from one sheet to the other. There are two principal possibilities, called "conventions", where this jump can happen. The "perfect delay convention" says that jumps happen at the border of the folded surface. The "Maxwell convention" says that jumps can happen along a curve in uv-space (a "shock-curve") which lays inside the cusp area. This curve consists of those points (u, v) in parameter space, where $F(., u, v)$ has two critical points with the same critical value. In our traffic jam model a shock curve is a curve in xt-space that we can interpret dynamically as the movement of the jam formation front along the street (see Fig. 9). The same figure shows what we call the morphology of the cusp: according to the convention, the catastrophe manifold with its fold or its cut can be interpreted as a fault or slip (as in geology) or a separation (if one of the parameter represents time). In our traffic jam model, there is a "line" (actually a region) of separation (shock wave) between the regions of high and low traffic density.

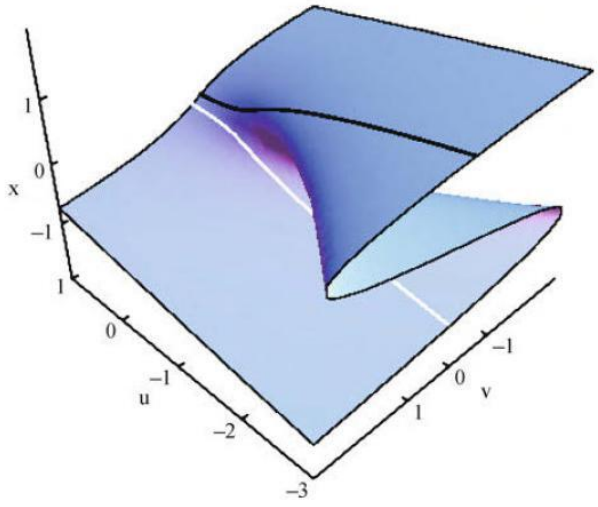
Besides the catastrophe manifold (catastrophe surface) there are two other essential terms: catastrophe set and bifurcation set (see Fig. 15). The catastrophe set can be viewed as the curve which goes along the border of the fold on the catastrophe manifold. It is the set of degenerated critical points and is described mathematically as the set $\{(x, u, v) | \partial_x F = \partial_x^2 F = 0\}$. Its projection into the uv-parameter space is the cusp curve, which is called the bifurcation set of the cusp catastrophe.

The cusp catastrophe has some special properties which we discuss now with the aid of some graphics. If the system under investigation shows one or more of these properties, the experimenter should try to find a possible cusp potential that accompanies the process.

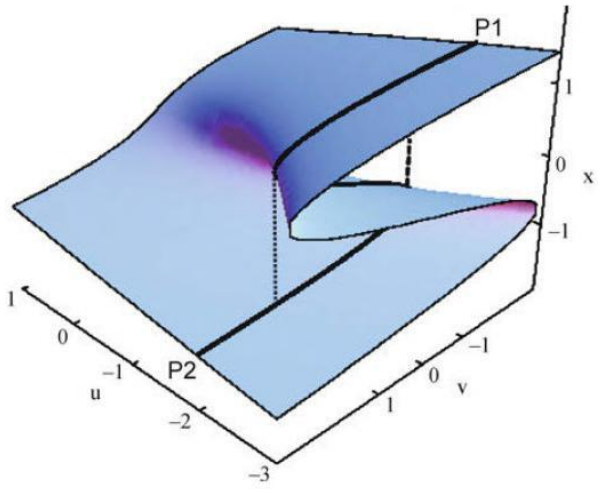
The first property is called divergence. This property can be shown by the two curves (black and white) on the catastrophe manifold which describes the system's equilibrium points. Both curves initially start at nearby positions on the surface, that is, they start at nearby stable system states. But the development of the system can proceed quite differently. One of the curves runs on the upper sheet of the surface while the other curve runs on the lower sheet. The system's stability positions are different and thus its behavior is different.

The next property is called hysteresis. If the system development is running along the path shown in Fig. 17 from P1 to P2 or vice versa, the jumps in the system behavior, that is, the sudden changes of internal system variables, occur at different parameter constellations, depending on the direction the path takes, from the upper to the lower sheet or vice versa. Jumps upward or downward happen

Copyrighted material



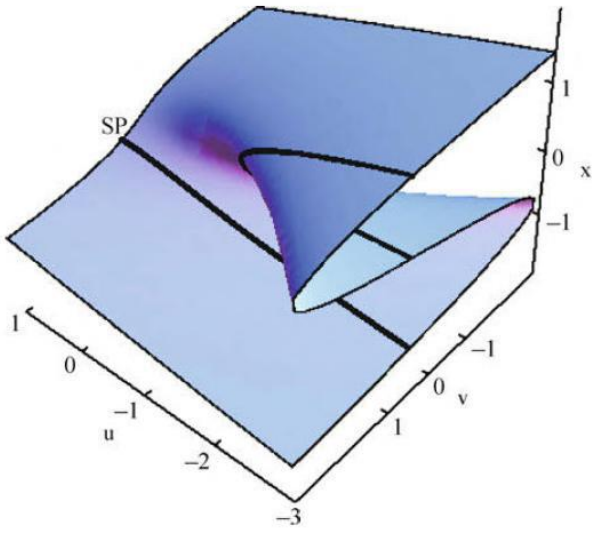
Catastrophe Theory, Figure 16
The divergence property of the cusp catastrophe



Catastrophe Theory, Figure 17
The hysteresis property of the cusp catastrophe

at the border of the fold along the dotted lines. The name of this property of the cusp comes from the characteristic hysteresis curve in physics, occurring for example at the investigation of the magnetic properties of iron.

Figure 18 shows the bifurcation property of the cusp catastrophe: the number of equilibrium states of the system splits from one to three along the path beginning at the starting point (SP) and running from positive to negative u values as shown. If the upper and the lower sheet of the fold correspond to the local minima of F and the middle sheet corresponds to the local maxima, then along the path shown, one stable state splits into two stable states and one unstable state.



Catastrophe Theory, Figure 18
The bifurcation property of the cusp catastrophe

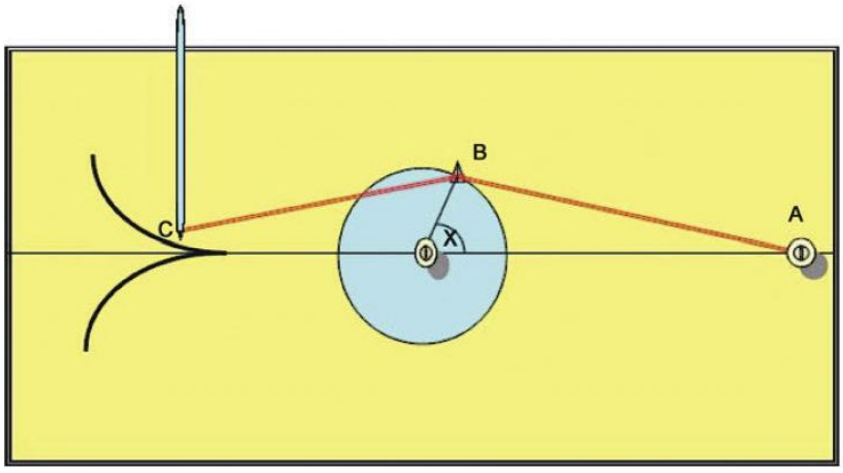
Further Applications

Many applications of catastrophe theory are attributable to Christopher Zeeman. For example, his catastrophe machine, the "Zeeman wheel," is often found in literature. This simple model consists of a wheel mounted flat against a board and able to turn freely. Two elastics are attached at one point (Fig. 19, point B) close to the periphery of the wheel. One elastic is fixed with its second end on the board (point A). The other elastic can be moved with its free end in the plane (point C).

Moving the free end smoothly, the wheel changes its angle of rotation smoothly almost everywhere. But at certain positions of C, which can be marked with a pencil, the wheel suddenly changes this angle dramatically. Joining the marked points where these jumps occur, you will find a cusp curve in the plane. In order to describe this behavior, the two coordinates of the position of point C serve as control parameters in the catastrophe model. The potential function for this model results from Hook's law and simple geometric considerations. Expanding the potential near its degenerated critical point into its Taylor series, and transforming the series with diffeomorphisms similar to the examples we have already calculated, this model leads to the cusp catastrophe. Details can be found in the book of Poston and Stewart [7] or Saunders (1986).

Another example of an application of catastrophe theory is similar to the traffic jam model we used here. Catastrophe theory can describe the formation of shockwaves in hydrodynamics. Again the cusp catastrophe describes

Copyrighted material



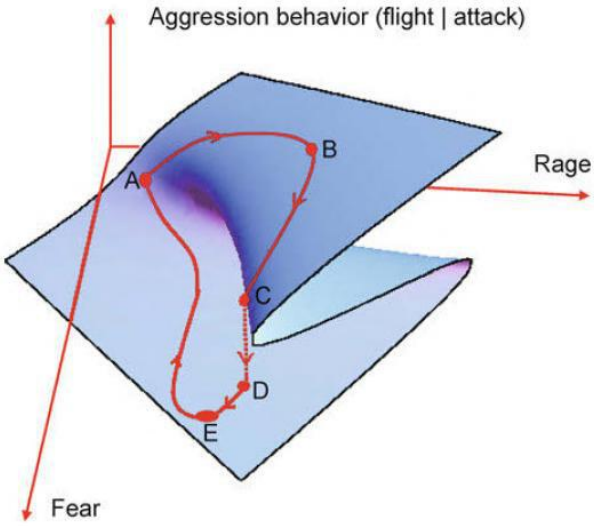
Catastrophe Theory, Figure 19
The Zeeman wheel

this phenomenon. The mathematical frame is given in the works of Lax [6], Golubitsky and Schaeffer[3] and Guckenheimer [4].

In geometrical optics light rays are investigated which are reflected on smooth surfaces. Caustics are the envelopes of the light rays. They are the bright patterns of intensive light, which can be seen, for example, on a cup of coffee when bright sunlight is reflected on the border of the cup. Catastrophe theory can be applied to light caustics, because the light rays obey a variational principle. According to Fermat's principle, light rays travel along geodesics and the role of the potential functions in catastrophe theory is played by geodesic distance functions. Caustics are their bifurcation sets. The calculations are given for example in the work of Sinha [11].

A purely speculative example, given by Zeeman, is an aggression model for dogs. Suppose that fear and rage of a dog can be measured in some way from aspects of its body language (its attitude, ears, tail, mouth, etc.). Fear and rage are two conflicting parameters, so that a catastrophe in the dog's behavior can happen. Look at Fig. 20.

The vertical axis shows the dog's aggression potential, the other axes show its fear and rage. Out of the many different ways the dog can behave, we choose one example. Suppose we start at point A on the catastrophe manifold (behavioral surface). Maybe the dog is dozing in the sun without any thought of aggression. The dog's aggression behavior is neutral (point A). Another dog is coming closer and our dog, now awakened, is becoming more and more angry by this breach of his territory. Moving along the path on the behavior surface, the dog is on a trajectory to attack his opponent. But it suddenly notices that the enemy is very big (point B). So its fear is growing, its



Catastrophe Theory, Figure 20
A dog's aggression behavior depending on its rage and fear

rage diminishes. At point C the catastrophe happens: Attack suddenly changes to flight (point D). This flight continues until the dog notices that the enemy does not follow (point E). Fear and rage both get smaller until the dog becomes calm again.

The example can be modified a little bit to apply to conflicts between two states. The conflicting factors can be the costs and the expected gain of wars, and so on.

Among the variety of possible examples, those worthy of mention include investigations of the stability of ships, the gravitational collapse of stars, the buckling beam (Euler beam), the breaking of ocean waves, and the development of cells in biology. The latter two problems are examples for the occurrence of higher catastrophes, for example

Copyrighted material

for all $p \in R^n$. In this case the mapping $X_t(p) = X(t,p)$ under fixed t is a diffeomorphism of R^n. Thus, F generates the smooth flow $X_t\colon R^n \to R^n$, $X_t(p) = X(t,p)$. The differential $DX_t(0)$ is a fundamental matrix of the linearized system $\dot{v} = DF(0)v$. A trajectory (an orbit) of the system (3) through x_0 is the set $T(x_0) = \{x = X(t,x_0), t \in R\}$.

Definition 2 A manifold M is said to be invariant under (3) if for any $p \in M$ the trajectory through p lies in M. A manifold M is said to be locally invariant under (3) if for every $p \in M$ there exists, depending on p, an interval $T_1 < 0 < T_2$, such that $X_t(p) \in M$ as $t \in (T_1,\ T_2)$.

It means that the invariant manifold is formed by trajectories of the system and the locally invariant manifold consists of arcs of trajectories. An equilibrium is 0-dimensional invariant manifold and a periodic orbit is 1-dimensional one. The concept of invariant manifold is a useful tool for simplification of dynamical systems.

Definition 3 A manifold M is said to be invariant in a neighborhood U if for any $p \in M \bigcap U$ the moving point $X_t(p)$ remains on M as long as $X_t(p) \in U$. In this case the manifold M is locally invariant.

Near an equilibrium point O the system (3) can be rewritten in the form

$$\dot{x} = Ax + f(x)\,, \tag{4}$$

where $O = \{x = 0\}$, $A = DF(O)$, f is second-order at the origin, i. e. $f(0) = 0$ and $Df(0) = 0$. Let us show that the system (4) near O can be considered as a perturbation of the linearized system

$$\dot{x} = Ax\,. \tag{5}$$

For this we construct a C^k-smooth mapping g which coincides with f in a sufficiently small neighborhood of the origin and is C^1-close to zero.

To construct the mapping g one uses the C^∞-smooth cut-off function $\alpha\colon R^+ \to R^+$

$$\alpha(r) = \begin{cases} 1, & r < 1/2\,, \\ > 0, & 1/2 \le r \le 1\,, \\ 0, & r > 1\,. \end{cases}$$

Then set $g(x) = \alpha(|x|/\epsilon)f(x)$ where ε is a parameter. If $|x| > \epsilon$, $g(x) = 0$ and if $|x| < \epsilon/2$, $g(x) = f(x)$ and g is C^1-close to zero. Consider the system

$$\dot{x} = Ax + g(x)\,. \tag{6}$$

It is evident that the dynamics of (4) and (6) is the same on $U = \{|x| < \epsilon/2\}$. If the constructed system (6) has an invariant manifold M then $M \bigcap U$ is locally invariant manifold for the initial system (4) or, more precisely, the manifold M is invariant in U. It should be noted that for the case of infinite-dimensional phase space the described construction is more delicate (see details below).

Let us consider a C^k-smooth mapping ($k \ge 1$) $G\colon R^n \to R^n$ which has the inverse C^k-smooth mapping G^{-1}. The mapping G generates the discrete dynamical system of the form

$$x_{m+1} = G(x_m)\,, \quad m = \ldots, -2, -1, 0, 1, 2, \ldots\,. \tag{7}$$

Each continuous system $X(t,p)$ gives rise to the discrete system $G(x) = X(1,x)$ which is the shift operator on unit time. Such a system preserves the orientation of the phase space,where as an arbitrary discrete system may change it.

Hence, the space of discrete systems is more rich than the space of continuous ones. Moreover, usually investigation of a discrete system is, as a rule, simpler than study of a differential equations one. In many instances the investigation of a differential equation may be reduced to study of a discrete system by Poincaré (first return) mapping. An orbit of (7) is the set $T = \{x = x_m,\ m \in \mathbb{Z}\}$, where x_m satisfies (7).

Near a fixed point O the mapping $x \to G(x)$ can be rewritten in the form

$$x \to Ax + f(x)\,, \tag{8}$$

where $O = \{x = 0\}$, $A = DG(O)$, f is second-order at the origin. It is shown above that near O (8) may be considered as a perturbation of the linearized system

$$x \to Ax\,. \tag{9}$$

Motivational Example

Consider the discrete dynamical system

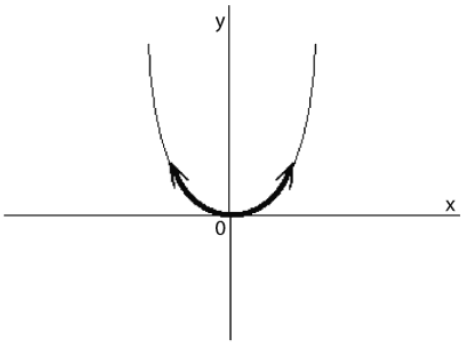
$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} X(x,y) = x + xy \\ Y(x,y) = \frac{1}{2}y + \frac{1}{2}x^2 + 2x^2y + y^3 \end{pmatrix}. \tag{10}$$

The origin $(0,0)$ is fixed point. Our task is to examine the stability of the fixed point. The linearized system at 0 is defined by the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$$

which has two eigenvalues 1 and 1/2. Hence, the first approximation system contracts along y-axis, whereas its action along x-axis is neutral. So the stability of the fixed

Copyrighted material



Center Manifolds, Figure 1
Dynamics on the invariant curve $y = x^2$

point depends on nonlinear terms. Let us show that the curve $y = x^2$ is invariant for the mapping (10). It is enough to check that if a point (x, y) is on the curve then the image (X, Y) is on the curve, i. e. $Y = X^2$ as $y = x^2$. We have

$$\begin{aligned} Y|_{y=x^2} &= \tfrac{1}{2}y + \tfrac{1}{2}x^2 + 2x^2y + y^3|_{y=x^2} \\ &= x^2 + 2x^4 + x^6 , \\ X^2|_{y=x^2} &= (x + xy)^2|_{y=x^2} \\ &= x^2 + 2x^4 + x^6 . \end{aligned} \tag{11}$$

Thus, the curve $y = x^2$ is an invariant one-dimensional manifold which is a center manifold W^c for the system (10), see Fig. 1. The fixed point O is on W^c. The restriction of the system on the manifold is

$$x \to x + xy|_{y=x^2} = x(1 + x^2) . \tag{12}$$

It follows that the fixed point 0 is unstable and $x_m \to \infty$ as $m \to \infty$, see Fig. 1. Hence, the origin is unstable fixed point for the discrete system (10). It turns out the system (10) near O is topologically equivalent to the system

$$\begin{aligned} &x \to x(1 + x^2) , \\ &y \to \tfrac{1}{2}y \end{aligned} \tag{13}$$

which is simpler than (10). The center manifold $y = x^2$ is tangent to the x-axis at the origin and W^c near 0 can be considered as a perturbed manifold of $\{y = 0\}$.

Introduction

In his famous dissertation "General Problem on Stability of Motion" [41] published at 1892 in Khar'kov (Ukraine) A.M. Lyapunov proved that the equilibrium O of the system (4) is stable if all eigenvalues of matrix A have negative real parts and O is unstable if there exists an eigenvalue with positive real part. He studied the case when some eigenvalues of A have negative real part and the rest of them have zero real one. Lyapunov proved that if the matrix A has a pair of pure imaginary eigenvalues and the other eigenvalues have negative real parts then there exists a two-dimensional invariant surface M through O, and the equilibrium O is stable if O is stable for the system restricted on M. Speaking in modern terms, Lyapunov proved the existence of "center manifold" and formulated "reduction principle" under the described conditions. Moreover, he found the stability condition by using power series expansions, which is an extension of his first method to evaluate stability of systems whose eigenvalues have zero real part. Now we use the same method to check stability [10,25,31,36,43,76]. The general reduction principle in stability theory was established by V. Pliss [59] in 1964, the term "center manifold" was introduced by A. Kelley at 1967 in the paper [38] where the existence of a family of invariant manifolds through equilibrium was proved in general case.

As we see the center manifold can be considered as a perturbation of the center subspace of linearized system. H. Poincaré [61] was probably the first who perceived the importance of the perturbation problem and began to study conditions ensuring the preservation of the equilibriums and periodic orbits under a perturbation of differential equations. Hadamard [26] and Perron [57] proved the existence of stable and unstable invariant manifolds for a hyperbolic equilibrium point and, in fact, showed their preservation. The conditions necessary and sufficient for the preservation of locally invariant manifolds passing through an equilibrium have been obtained in [51]. Many results on the center manifold follow from theory of normal hyperbolicity [77] which studies dynamics of a system near a compact invariant manifold. They will be considered below. The books by Carr [10], Guckenheimer and Holmes [25], Marsden and McCracken [43], Iooss [36], Hassard, Kazarinoff and Wan [31], and Wiggins [76] are popular sources of the information about center manifolds.

Center Manifold in Ordinary Differential Equations

Consider a linear system of differential equations

$$\dot{v} = Lv , \tag{14}$$

where $v \in R^n$ and L is a matrix. Divide the eigenvalues of the matrix L into three parts: stable σ_s = {eigenvalues with negative real part}, unstable σ_u = {eigenvalues with positive real part} and central (neutral, critical) σ_c = {eigenvalues with zero real part}. If the matrix does not have the eigenvalues with zero real part, the system is called

Copyrighted material

hyperbolic. The matrix L has three eigenspaces: the stable subspace E^s, the unstable subspace E^u, and the central subspace E^c correspond to these parts of spectrum. The subspaces E^s, E^u, and E^c are invariant under the system (14), that is, the solution starting on the subspace stays on it [33]. The solutions on E^s have exponential tending to 0, the solutions on E^u have exponential growth. The subspaces E^s, E^c, E^u meet in pairs only at the origin and $E^s + E^c + E^u = R^n$, i. e. we have the invariant decomposition $R^n = E^s \oplus E^c \oplus E^u$. There exists a linear change of coordinates (see [33]) transforming (14) to the form

$$\begin{aligned} \dot{x} &= Ax\,, \\ \dot{y} &= By\,, \\ \dot{z} &= Cz\,, \end{aligned} \tag{15}$$

where the matrix A has eigenvalues with zero real part, the matrix B has eigenvalues with negative real part, and the matrix C has eigenvalues with positive real part. So, (14) decomposes into three independent systems of differential equations. It is known [1,23,24] that the system (15) is topologically equivalent to the system

$$\begin{aligned} \dot{x} &= Ax\,, \\ \dot{y} &= -y\,, \\ \dot{z} &= z\,. \end{aligned} \tag{16}$$

Thus, the dynamics of the system (14) is determined by the system $\dot{x} = Ax$ which is the restriction of (14) on the center subspace E^c. Our goal is to justify a similar "reduction principle" for nonlinear systems of differential equations.

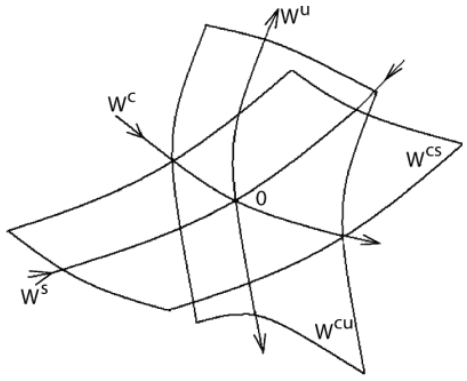
Summarizing the results of V. Pliss [59,60], A. Kelley [38], N. Fenichel [19], and M. Hirsch, C. Pugh, M. Shub [34] we obtain the following theorem.

Theorem 1 (Existence Theorem) *Consider a C^k-smooth ($1 \le k < \infty$) system of differential equations $\dot{v} = F(v)$, $F(O) = 0$ and $L = DF(O)$. Let E^s, E^u and E^c be the stable, unstable, and central eigenspaces of L.*

Then near the equilibrium O there exist the following five C^k-smooth invariant manifolds (see Fig. 2):

- *The center manifold W^c, tangent to E^c at O;*
- *The stable manifold W^s, tangent to E^s at O;*
- *The unstable manifold W^u, tangent to E^u at O;*
- *The center-stable manifold W^{cs}, tangent to $E^c + E^s$ at O;*
- *The center-unstable manifold W^{cu}, tangent to $E^c + E^u$ at O;*

The stable and unstable manifolds are unique, but center, center-stable, and center-unstable manifolds are not necessarily unique.



Center Manifolds, Figure 2
The classical five invariant manifolds

Solutions on W^s have exponential tending to O as $t \to \infty$. Solutions on W^u have exponential tending to O as $t \to -\infty$.

Remark 1 The expression "near the equilibrium O" means that there exists a neighborhood $U(O)$, depending on F and k, where the statements of the Existence Theorem hold.

Remark 2 The linearized system $\dot{v} = DF(O)v$ has the invariant subspace $E^s + E^u$ whereas the complete system $\dot{v} = F(v)$ may not have any smooth invariant manifolds with the tangent space $E^s + E^u$ at O, see [5,30].

Representation of the Center Manifold

At first we suppose that our system does not have the unstable eigenspace and is transformed by the linear change of coordinates mentioned above to the form

$$\begin{aligned} \dot{x} &= Ax + f(x, y)\,, \\ \dot{y} &= By + g(x, y)\,, \end{aligned} \tag{17}$$

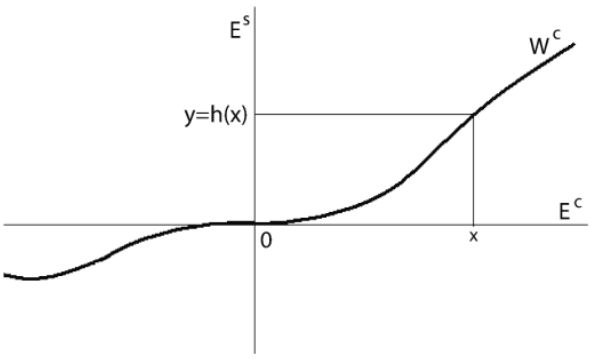
where $E^c = \{(x, 0)\}$, $E^s = \{(0, y)\}$. Since the center manifold W^c is tangent to $E^c = \{(x, 0)\}$ at the origin $O = (0, 0)$, it can be represented near O in the form

$$W^c = \{(x, y) \,|\, |x| < \varepsilon,\ y = h(x)\}\,.$$

In other words the center manifold is represented as the graph of a smooth mapping $h\colon V \subset R^c \to R^s$, where c and s are dimensions of the center and stable subspaces, see Fig. 3. Since W^c goes through the origin and is tangent to E^s, we have the equalities

$$h(0) = 0\,, \quad Dh(0) = 0\,. \tag{18}$$

Copyrighted material



Center Manifolds, Figure 3
Representation of a center manifold

The invariance of W^c means that if an initial point (x, y) is in W^c, i. e. $y = h(x)$ then the solution $(X(t, x, y), Y(t, x, y))$ of (17) is in W^c, i. e. $Y(t, x, y)) = h(X(t, x, y))$. Thus, we get the invariance condition

$$Y(t, x, h(x))) = h(X(t, x, h(x))) \,. \tag{19}$$

Differentiating (19) with respect to t and putting $t = 0$ we get

$$Bh(x) + g(x, h(x)) = Dh(x)(Ax + f(x, h(x))) \,. \tag{20}$$

Thus, the mapping h have to be a solution of the partial differential equation (20) and satisfy (18).

Suppose that the system has the form

$$\begin{aligned} \dot{x} &= Ax + f(x, y, z) \,, \\ \dot{y} &= By + g_1(x, y, z) \,, \\ \dot{z} &= Cz + g_2(x, y, z) \,, \end{aligned} \tag{21}$$

where A has the eigenvalues with zero real part, B has the eigenvalues with negative real part, C has the eigenvalues with positive real part. Analogously to the previous case we can show that the center manifold is represented in the form

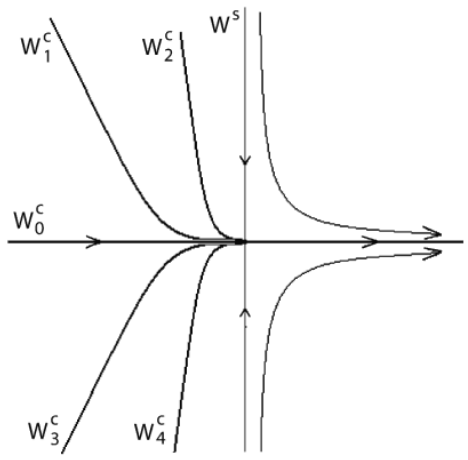
$$W^c = \{(x, y, z) \,|\, |x| < \epsilon,\ y = h_1(x),\ z = h_2(x)\}$$

where the mappings h_1 and h_2 satisfy the equations

$$\begin{aligned} &Bh_1(x) + g_1(x, h_1(x), h_2(x)) \\ &\qquad = Dh_1(x)(Ax + f(x, h_1(x), h_2(x))) \\ &Ch_2(x) + g_2(x, h_1(x), h_2(x)) \\ &\qquad = Dh_2(x)(Ax + f(x, h_1(x), h_2(x))) \end{aligned} \tag{22}$$

and the conditions

$$\begin{aligned} h_1(0) &= 0 \,, \quad Dh_1(0) = 0 \\ h_2(0) &= 0 \,, \quad Dh_2(0) = 0 \,. \end{aligned} \tag{23}$$



Center Manifolds, Figure 4
Nonuniqueness of the center manifold

Uniqueness and Nonuniqueness of the Center Manifold

The center manifold is nonunique in general.

Let us consider the illustrating example from [38]. The system has the form

$$\begin{aligned} \dot{x} &= x^2 \,, \\ \dot{y} &= -y \,, \end{aligned} \tag{24}$$

where $(x, y) \in R^2$. It is obviously that the origin $(0, 0)$ is the equilibrium with single stable manifold $W^s = \{(0, y)\}$. There is a center manifold of the form $W^c = \{(x, 0)\}$. Moreover, there are the center manifolds that can be obtained when solving the equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{-y}{x^2} \,. \tag{25}$$

The solution of (24) has the form

$$y(x) = a \exp\left(\frac{1}{x}\right) \tag{26}$$

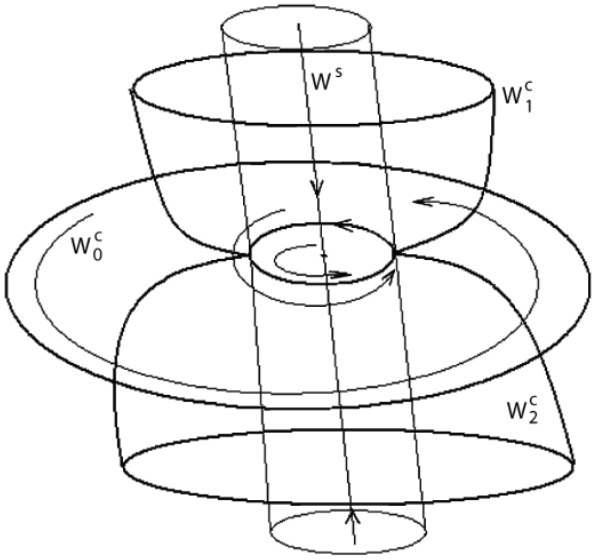
for $x \neq 0$ and any constant $a \in R$. It follows that

$$W^c(a) = \{(x, y) | y = a \exp\left(\frac{1}{x}\right) \quad \text{as } x < 0 \text{ and } y = 0 \text{ as } x \geq 0\}$$

is center manifold for each a, being C^∞-smooth and pasting together the curve $\{y = a \exp(1/x)\}$ as $x < 0$ and the straight line $\{y = 0\}$ as $x \geq 0$., see Fig. 4.

However, the center manifolds possess the following weak uniqueness property. Let U be a neighborhood of the W^c. It turns out that the maximal invariant set I in U

Copyrighted material



Center Manifolds, Figure 5
Weak uniqueness of the center manifold

must be in W^c, that follows from the Reduction Theorem, see below. Consequently, all center manifolds contact on I. In this case the center manifold is called locally maximal. Thus, the center manifolds may differ on the trajectories leaving the neighborhood as $t \to \infty$ or $-\infty$.

For example, suppose that a center manifold is two-dimensional and a limit cycle is generated from the equilibrium through the Hopf bifurcation [31,35]. In this case the invariant set I is a disk bounded by the limit cycle and all center manifolds will contain this disk, see Fig. 5.

Smoothness

M. Hirsch, C. Pugh, M. Shub [34] proved that the center manifold is C^k-smooth if the system is C^k-smooth and $k < \infty$. Moreover if the kth derivative of the vector field is α-Hölder or Lipschitz mapping, the center manifold is the same.

However, if the system is C^∞ or analytic then the center manifold is not necessary the same. First consider the analytic case. It is clear that if the analytic center manifold exists then it is uniquely determined by applying the Taylor power series expansion. Consider the illustrating example [31]

$$\begin{aligned} \dot{x} &= -x^2 \,, \\ \dot{y} &= -y + x^2 \end{aligned} \tag{27}$$

which does not have an analytic center manifold. In fact, applying the Taylor power series expansions we obtain that the center manifold has to be given by the series $y = \sum_{n=2}^{\infty} (n-1)! x^n$ which diverges when $x \neq 0$.

If the system is C^∞-smooth then the Existence Theorem guaranties the existence of a C^k-smooth center manifold for any $k < \infty$. However, the center manifold may not be C^∞-smooth. Van Strien S.J. [71] showed that the system

$$\begin{aligned} \dot{x} &= -x^2 + \mu^2 \,, \\ \dot{y} &= -y - (x^2 - \mu^2) \,, \\ \dot{\mu} &= 0 \,. \end{aligned} \tag{28}$$

does not have a C^∞-smooth center manifold. In fact if the system is C^∞-smooth then for each k there is a neighborhood $U_k(O)$ of a C^k-smooth central manifold W^c. There are the systems for which the sequence $\{U_k\}$ can shrink to O as $k \to \infty$ [25,71]. The results of the papers [19,34] show that the smoothness of the invariant manifold depends on the relation between Lyapunov exponents on the center manifold and on the normal subspace (stable plus unstable subspaces). This relation σ is included in the concept of the normal hyperbolicity. At an equilibrium $\sigma = \min(\mathrm{Re}\lambda_n/\mathrm{Re}\lambda_c)$, where $\mathrm{Re}\lambda_n/\mathrm{Re}\lambda_c > 0$, λ_c is the eigenvalue on the center subspace and λ_n is the eigenvalue on the normal subspace. The condition $\sigma > 1$ is necessary for a persistence of smooth invariant manifold. One can guarantee degree of smoothness k of the manifold provided $k < \sigma$. Moreover, there exists examples showing that the condition $k < \sigma$ is essential. It means that the system has to contract (expand) along $E^s(E^u)k$-times stronger than along the center manifold.

If a center manifold to the equilibrium O exists then $\mathrm{Re}\,\lambda_c = 0$ and $\sigma = \infty$ at O, but near O may be other equilibrium (or other orbits) on the center manifold where $\sigma < \infty$, and as a consequence, the center manifold may not be C^∞-smooth. Let us consider the illustrating example [25]

$$\begin{aligned} \dot{x} &= \mu x - x^3 \,, \\ \dot{y} &= y + x^4 \,, \\ \dot{\mu} &= 0 \,. \end{aligned} \tag{29}$$

The point $O = (0,0,0)$ is equilibrium, the system linearized at O has the form

$$\begin{aligned} \dot{x} &= 0 \,, \\ \dot{y} &= y \,, \\ \dot{\mu} &= 0 \,. \end{aligned} \tag{30}$$

The system (30) has the following invariant subspace: $E^s = \{0,0,0\}$, $E^c = \{x,0,\mu\}$, $E^u = \{0,y,0\}$. The μ-axis

Copyrighted material

The orbits on W^s have exponential tending to O as $m \to \infty$. The orbits on W^u have exponential tending to O as $m \to -\infty$.

There exists a linear change of coordinates transforming (46) to the form

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \to \begin{pmatrix} Ax + f(x,y,z) \\ By + g_1(x,y,z) \\ Cz + g_2(x,y,z) \end{pmatrix}, \tag{48}$$

where A has eigenvalues with modulus equal 1, B has eigenvalues with modulus less than 1, C has eigenvalues with modulus greater than 1. The system (48) at the origin has the following invariant eigenspaces: $E^c = \{(x,0,0)\}$, $E^s = \{(0,y,0)\}$, and $E^u = \{(0,0,z)\}$. Since the center manifold W^c is tangent to E^c at the origin $O = (0,0)$, it can be represented near O in the form

$$W^c = \{(x,y,z) \,|\, |x| < \epsilon,\ y = h_1(x), z = h_2(x)\},$$

where h_1 and h_2 are second-order at the origin, i. e. $h_{1,2}(0) = 0$, $Dh_{1,2}(0) = 0$. The invariance of W^c means that if a point $(x,y,z) \in W^c$, i. e. $y = h_1(x)$, $z = h_2(x)$, then $(Ax + f(x,y,z), By + g_1(x,y,z), Cz + g_2(x,y,z)) \in W^c$, i. e. $By + g_1(x,y,z) = h_1(Ax + f(x,y,z))$, $Cz + g_2(x,y,z) = h_2(Ax + f(x,y,z))$. Thus, we get the invariance property

$$\begin{aligned} &Bh_1(x) + g_1(x, h_1(x), h_2(x)) \\ &\qquad = h_1(Ax + f(x, h_1(x), h_2(x))\,, \\ &Ch_2(x) + g_2(x, h_1(x), h_2(x)) \\ &\qquad = h_2(Ax + f(x, h_1(x), h_2(x))\,. \end{aligned} \tag{49}$$

Results on smoothness and uniqueness of center manifold for discrete systems are the same as for continuous systems.

Theorem 5 (Reduction Theorem [55,62,63,64]) *The discrete system (48) near the origin is topologically equivalent to the system*

$$\begin{aligned} &x \to Ax + f(x, h_1(x), h_2(x))\,, \\ &y \to By\,, \\ &z \to Cz\,. \end{aligned} \tag{50}$$

The first mapping is the restriction of the system on the center manifold. The theorem reduces the investigation of dynamics near a nonhyperbolic fixed point to the study of the system on the center manifold.

Normally Hyperbolic Invariant Manifolds

As it is indicated above, a center manifold can be considered as a perturbed invariant manifold for the center subspace of the linearized system. The normal hyperbolicity conception arises as a natural condition for the persistence of invariant manifold under a system perturbation. Informally speaking, an f-invariant manifold M, where f-diffeomorphism, is normally hyperbolic if the action of Df along the normal space of M is hyperbolic and dominates over its action along the tangent one. (the rigorous definition is below). Many of the results concerning center manifold follows from the theory of normal hyperbolic invariant manifolds. In particular, the results related to center eigenvalues of Df with $|\lambda| \neq 1$ may be obtained using this theory.

The problem of existence and preservation for invariant manifolds has a long history. Initial results on invariant (integral) manifolds of differential equations were obtained by Hadamard [26] and Perron [57], Bogoliubov, Mitropolskii [8] and Lykova [48], Pliss [60], Hale [28], Diliberto [16] and other authors (for references see [39,77]). In the late 1960s and throughout 1970s the theory of perturbations of invariant manifolds began to assume a very general and well-developed form. Sacker [66] and Neimark [50] proved (independently and by different methods) that a normally hyperbolic compact invariant manifold is preserved under C^1 perturbations. Final results on preservation of invariant manifolds were obtained by Fenichel [19,20] and Hirsch, Pugh, Shub [34]. The linearization theorem for a normally hyperbolic manifold was proved in [62].

Let $f\colon R^n \to R^n$ be a C^r-diffeomorphism, $1 \le r < \infty$, and a compact manifold M is f-invariant.

Definition 4 An invariant manifold M is called r-normally hyperbolic if there exists a continuous invariant decomposition

$$TR^n|_M = TM \oplus E^s \oplus E^u \tag{51}$$

of the tangent bundle TR^n into a direct sum of subbundles TM, E^s, E^u and constants a, $\lambda > 0$ such that for $0 \le k \le r$ and all $p \in M$,

$$\begin{aligned} &|D^s f^n(p)|\ |(D^0 f^n(p))^{-1}|^k \le a \exp(-\lambda n) \quad \text{for } n \ge 0\,, \\ &|D^u f^n(p)|\ |(D^0 f^n(p))^{-1}|^k \le a \exp(\lambda n) \quad \text{for } n \le 0\,. \end{aligned} \tag{52}$$

Here $D^0 f$, $D^s f$ and $D^u f$ are restrictions of Df on TM, E^s and E^u, respectively.

The invariance of the bundle E^* means that for every $p \in M_0$, the image of the fiber $E^*(p)$ at p under $Df(p)$

Copyrighted material

is the fiber $E^*(q)$ at $q = f(p)$. The bundles E^s and E^u are called stable and unstable, respectively. In other words, M is normally hyperbolic if the differential Df contracts (expands) along the normal (to M) direction and this contraction (expansion) is stronger in r times than a conceivable contraction (expansion) along M. Summarizing the results of [19,20,34,50,62,66] we obtain the following theorem.

Theorem 6 *Let a C^r diffeomorphism f be r-normally hyperbolic on a compact invariant manifold M with the decomposition $TR^n|_M = TM \oplus E^s \oplus E^u$. Then*

- *there exist invariant manifolds W^s and W^u near M, which are tangent at M to $TM \oplus E^s$ and $TM \oplus E^u$;*
- *the manifolds M, W^s and W^u are C^r-smooth;*
- *if g is another C^r-diffeomorphism C^r-close to f, then there exists the unique manifold M_g which is invariant and r-normally hyperbolic for g;*
- *near M f is topologically equivalent to $Df|_{E^s \oplus E^u}$, which in local coordinates has the form*

$$(x, y, z) \to (f(x), Df(x)y, Df(x)z)\,, \qquad x \in M\,, \quad y \in E^s(x)\,, \quad z \in E^u(x)\,. \tag{53}$$

Similarly result takes place for flows. Mañé [44] showed that a locally unique, preserved, compact invariant manifold is necessarily normally hyperbolic. Local uniqueness means that near M, an invariant set I of the perturbation g is in M_g. Conditions for the preservation of locally nonunique compact invariant manifolds and linearization were found by G. Osipenko [52,53].

Applications

Stability

As it was mentioned above, it was Lyapunov who pioneered in applying the concept of center manifold to establish the stability of equilibrium. V. Pliss [59] proved the general reduction principle in stability theory. According to this principle the equilibrium is stable if and only if it is stable on the center manifold. We have considered the motivating example where the center manifold concept was used. The system

$$\begin{aligned} \dot{x} &= xy\,, \\ \dot{y} &= -y + ax^2\,, \end{aligned} \tag{54}$$

has the center manifold $y = h(x)$. Applying Theorem 3, we obtain

$$h(x) = ax^2 + \cdots\,.$$

The reduced system is of the form

$$\dot{x} = ax^3 + \ldots\,.$$

Thus, if $a < 0$, the origin is stable. The book [25] contains other examples of investigation of stability.

Bifurcations

Consider a parametrized system of differential equations

$$\dot{x} = A(\mu)x + f(x, \mu)\,, \tag{55}$$

where $\mu \in R^k$, f is C^1-small when μ is small. To study the bifurcation problem near $\mu = 0$ it is useful to deal with the extended system of the form

$$\begin{aligned} \dot{x} &= A(\mu)x + f(x, \mu)\,, \\ \dot{\mu} &= 0\,. \end{aligned} \tag{56}$$

Suppose that (55) has a n-dimensional center manifold for $\mu = 0$. Then the extended system (56) has a $n + k$-dimensional center manifold. The reduction principle guarantees that all bifurcations lie on the center manifold. Moreover, the center manifold has the form $y = h(x, \mu_0)$ for every μ_0 which is a solution of the equation $\dot{\mu} = 0$ and the mapping h may be presented as a power series in μ. Further interesting information regarding to recent advancements in the approximation and computation of center manifolds is in [37]. The book [31] deals with Hopf bifurcation and contains a lot of examples and applications. Invariant manifolds and foliations for nonautonomous systems are considered in [3,13]. Partial linearization for noninvertible mappings is studied in [4].

Center Manifold in Infinite-Dimensional Space

Center manifold theory is a standard tool to study the dynamics of discrete and continuous dynamical systems in infinite dimensional phase space. We start with the discrete dynamical system generated by a mapping of a Banach space. The existence of the center manifold of the mappings in Banach space was proved by Hirsch, Pugh and Shub [34]. Let $T\colon E \to E$ be a linear endomorphism of a Banach space E.

Definition 5 A linear endomorphism T is said to be ρ-hyperbolic, $\rho > 0$, if no eigenvalue of T has modulus ρ, i. e., its spectrum Spect(T) has no points on the circle of radius ρ in the complex plane C. A linear 1-hyperbolic isomorphism T is called hyperbolic.

For a ρ-hyperbolic linear endomorphism there exists a T-invariant splitting of $E = E_1 \oplus E_2$ such that the spectrum of the restriction $T_1 = T|_{E_1}$ lies outside of the disk of

Copyrighted material

the radius ρ, where as the spectrum of $T_2 = T|_{E_2}$ lies inside it. So T_1 is automorphism and T_1^{-1} exists. It is known (for details see [34]) that one can define norms in E_1 and E_2 such that associated norms of T_1^{-1} and T_2 may be estimated in the following manner

$$|T_1^{-1}| < \frac{1}{\rho}\,, \quad |T_2| < \rho\,. \tag{57}$$

Conversely, if T admits an invariant splitting $T = T_1 \oplus T_2$ with $|T_1^{-1}| \equiv a$, $|T_2| \equiv b$, $ab < 1$ then Spect(T_1) lies outside the disk $\{\lambda \in C\colon |\lambda| < 1/a\}$, and Spect$(T_2)$ lies in the disk $\{\lambda \in C\colon |\lambda| \le b\}$. Thus, T is ρ-hyperbolic with $b < \rho < 1/a$.

Theorem 7 *Let T be a ρ-hyperbolic linear endomorphism, $\rho \le 1$. Assume that $f\colon E \to E$ is C^r, $r \ge 1$, $f = T + g$, $f(0) = 0$ and there is $\delta > 0$ such that if g is a Lipschitz mapping and*

$$\mathrm{Lip}(g) = \mathrm{Lip}(f - T) \le \delta\,, \tag{58}$$

then there exists a manifold W_f, which is a graph of a C^1 map $\varphi\colon E_1 \to E_2$, i. e.

$$W_f = \{x + y \,|\, y = \varphi(x),\ x \in E_1,\ y \in E_2\}\,,$$

with the following properties:

- *W_f is f-invariant;*
- *if $\|T_1^{-1}\|^j \|T_2\| < 1$, $j = 1, \dots, r$ then W_f is C^r and depends continuously on f in the C^r topology;*
- *$W_T = E_1$;*
- *for $x \in W_f$*

$$|f^{-1}(x)| \le (a + 2\varepsilon)\,|x|\,, \tag{59}$$

  where $a < 1/\rho$, and ε is small provided δ is small;
- *if $Df(0) = T$ then W_f is tangent to E_1 at 0.*

Suppose that the spectrum of T is contained in $A_1 \bigcup A_2$ where

$$A_1 = \{z \in C,\ |z| \ge 1\}\,, \quad A_2 = \{z \in C,\ |z| \le a < 1\}\,.$$

Corollary 1 *If $f\colon E \to E$ is C^r, $1 \le r < \infty$, $f(0) = 0$ and $\mathrm{Lip}(f - T) \le \epsilon$ is small, then there exists a center-unstable manifold $W^{cu} = W_f$ which is the graph of a C^r function $E_1 \to E_2$. The center-unstable manifold is attractor, i. e. for any $x \in E$ the distance between $f^n(x)$ and W^{cu} tends to zero as $n \to \infty$. The manifold $W_f = W^c$ is center manifold when $A_1 = \{z \in C,\ |z| = 1\}$.*

Apply this theorem to a mapping of the general form

$$u \to Au + f(u)\,, \tag{60}$$

where E is a Banach space, $u \in E$, $A\colon E \to E$ is a linear operator, f is second-order at the origin. The question arises of whether there exists a smooth mapping g such that g coincides with f in a sufficiently small neighborhood of the origin and g is C^1-close to zero? The answer is positive if E admits C^∞-norm, i. e. the mapping $u \to \|u\|$ is C^∞-smooth for $u \ne 0$. It is performed for Hilbert space. The desired mapping g may be constructed with the use of a cut-off function. To apply Theorem 7 to (60) we have to suppose *Hypothesis on C^∞-norm*: the Banach space E admits C^∞-norm.

Thus, Theorem 7 together with Hypothesis on C^∞-norm guarantee an existence of center manifold for the system (60) in a Banach space. A. Reinfelds [63,64] prove a reduction theorem for homeomorphism in a metric space from which an analog of Reduction Theorem 5 for Banach space follows.

Flows in Banach Space

Fundamentals of center manifold theory for flows and differential equations in Banach spaces can be found in [10,21,43,73]. In Euclidean space a flow is generated by a solution of a differential equation and the theory of ordinary differential equations guarantees the clear connection between flows and differential equations. In infinite dimensional (Banach) space this connection is more delicate and at first, we consider infinite dimensional flows and then (partial) differential equations.

Definition 6 A flow (semiflow) on a domain D is a continuous family of mappings $F_t\colon D \to D$ where $t \in R$ $(t \ge 0)$ such that $F_0 = I$ and $F_{t+s} = F_t F_s$ for $t, s \in R$ $(t, s \ge 0)$.

Theorem 8 (Center Manifold for Infinite Dimensional Flows [43]) *Let the Hypothesis on C^∞-norm be fulfilled and $F_t\colon E \to E$ be a semiflow defined near the origin for $0 \le t \le \tau$, $F_t(0) = 0$. Suppose that the mapping $F_t(x)$ is C^0 in t and C^{k+1} in x and the spectrum of the linear semigroup $DF_t(0)$ is of the form $e^{t(\sigma_1 \cup \sigma_2)}$, where $e^{t(\sigma_1)}$ is on the unit circle (i. e. σ_1 is on the imaginary axis) and $e^{t(\sigma_2)}$ is inside the unit circle with a positive distance from one as $t > 0$ (i. e. σ_2 is in the left half-plane). Let $E = E_1 \oplus E_2$ be the $DF_t(0)$-invariant decomposition corresponding to the decomposition of the spectrum $e^{t(\sigma_1 \cup \sigma_2)}$ and $dim E_1 = d < \infty$.*

Then there exists a neighborhood U of the origin and C^k-submanifold $W^c \subset U$ of the dimension d such that

- *W^c is invariant for F_t in U, $0 \in W^c$ and W^c is tangent to E_1 at 0;*
- *W^c is attractor in U, i. e. if $F_t^n(x) \in U$, $n = 0, 1, 2, \dots$ then $F_t^n(x) \to W^c$ as $n \to \infty$.*

Copyrighted material

In infinite dimensional dynamics it is also popular notion of "inertial manifold" which is an invariant finite dimensional attracting manifold corresponding to W^c of the Theorem 8.

Partial Differential Equations

At the present the center manifold method takes root in theory of partial differential equations. Applying center manifold to partial differential equations leads to a number of new problems. Consider an evolution equation

$$\dot{u} = Au + f(u)\,, \tag{61}$$

where $u \in E$, E is a Banach space, A is a linear operator, f is second-order at the origin. Usually A is a differential operator defined on a domain $D \subset E$, $D \neq E$. To apply Theorem 8 we have to construct a semiflow F_t defined near the origin of E. As a rule, an infinite dimensional phase space E is a functional space where the topology may be chosen by the investigator. A proper choice of functional space and its topology may essentially facilitate the study. As an example we consider a nonlinear heat equation

$$u'_t = \Delta u + f(u)\,, \tag{62}$$

where u is defined on a bounded domain $\Omega \subset R^n$ with Dirichlet condition on the boundary $u|_{\partial\Omega} = 0$, $\Delta u = (\partial^2/\partial x_1^2 + \cdots + \partial^2/\partial x_n^2)u$ is the Laplace operator defined on the set

$$C_0^2 = \{u \in C^2(\overline{\Omega}),\ u = 0 \text{ on } \partial\Omega\}\,.$$

According to [49] we consider a Hilbert space $E = L^2(\Omega)$ with the inner product $< u, v >= \int_\Omega uv dx$. The operator Δ may be extended to a self-conjugate operator $A\colon D_A \to L^2(\Omega)$ with the domain D_A that is the closure of C_0^2 in $L^2(\Omega)$. Consider the linear heat equation $u'_t = \Delta u$ and its extension

$$u'_t = Au, \quad u \in D_A\,. \tag{63}$$

What do we mean by a flow generated by (63)?

Definition 7 A linear operator A is the infinitisemal generator of a continuous semigroup $U(t)$, $t \geq 0$ if:
$U(t)$ is a linear mapping for each t;
$U(t)$ is a semigroup and $\|U(t)u - u\| \to 0$ as $t \to +0$;

$$Au = \lim_{t \to +0} \frac{U(t)u - u}{t} \tag{64}$$

whenever the limit exists.

For Laplace operator we have $< \Delta u, u >\leq 0$ for $u \in C_0^2$, hence the operator A is dissipative. It guarantees [6] the existence of the semigroup $U(t)$ with the infinitesimal generator A. The mapping $U(t)$ is the semigroup $DF_t(0)$ in Theorem 8.

The next problem is the relation between the linearized system $u' = Au$ and full nonlinear Equation (61). To consider the mapping f in (62) as C^1-small perturbation, the Hypothesis on C^∞-norm has to be fulfilled. To prove the existence of the solution of the nonlinear equation with $u(0) = u_0$ one uses the Duhamel formula

$$u(t) = U_t u_0 + \int_0^t U_{t-s} f(u(s)) \mathrm{d}s$$

and the iterative Picard method. By this way we construct the semiflow required to apply Theorem 8. The relation between the spectrum of a semigroup $U(t)$ and the spectrum of its infinitisemal generator gives rise to an essential problem. All known existence theorems are formulated in terms of the semigroup as in Theorem 8. The existence of the spectrum decompositions of a semigroup $U(t)$ and the corresponding estimations on the appropriate projections are supposed. This framework is inconvenient for many applications, especially in partial differential equations. In finite dimensions, a linear system $\dot{x} = Ax$ has a solution of the form $U(t)x = \mathrm{e}^{At}x$ and λ is an eigenvalue of A if and only if $\mathrm{e}^{\lambda t}$ is an eigenvalue of $U(t)$. We have the spectral equality

$$\mathrm{Spect}(U(t)) = \mathrm{e}^{\mathrm{Spect}(A)t}\,.$$

In infinite dimensions, relating the spectrum of the infinitesimal generator A to that of the semigroup $U(t)$ is a spectral mapping problem which is often nontrivial. The spectral inclusion

$$\mathrm{Spect}(U(t)) \subset \mathrm{e}^{\mathrm{Spect}(A)t}$$

always holds and the inverse inclusion is a problem that is solved by the spectral mapping theorems [12,18,22,65].

Often an application of center manifold theory needs to prove an appropriate version of the reduction theorem [7,11,12,14,15,18]. Pages 1–5 of the book [40] give an extensive list of the applications of center manifold theory to infinite dimensional problems. Here we consider a few of typical applications. Reaction-diffusion equations are typical models in chemical reaction (Belousov–Zhabotinsky reaction), biological systems, population dynamics and nuclear reaction physics. They have the form

$$u'_t = (K(\lambda) + D\Delta)u + f(u, \lambda)\,, \tag{65}$$

where K is a matrix depending on a parameter, D is a symmetric, positive semi-definite, often diagonal matrix, Δ is the Laplace operator, λ is a control parameter. The pa-

Copyrighted material

pers [2,27,31,67,78] applied center manifold theory to the study of the reaction-diffusion equation. Invariant manifolds for nonlinear Schrodinger equations are studied in [22,40]. In the elliptic quasilinear equations the spectrum of the linear operator is unbounded in unstable direction and in this case the solution does not generate even a semiflow. Nevertheless, center manifold technique is successfully applied in this case as well [45]. Henry [32] proved the persistence of normally hyperbolic closed linear subspace for semilinear parabolic equations. It follows the existence of the center manifold which is a perturbed manifold for the center subspace. I. Chueshov [14,15] considered a reduction principle for coupled nonlinear parabolic-hyperbolic partial differential equations that has applications in thermoelastisity. Mielke [47] has developed the center manifold theory for elliptic partial differential equations and has applied it to problems of elasticity and hydrodynamics. The results for non-autonomous system in Banach space can be found in [46]. The reduction principle for stochastic partial differential equations is considered in [9,17,75].

Future Directions

Now we should realize that principal problems in the center manifolds theory of finite dimensional dynamics have been solved and we may expect new results in applications. However the analogous theory for infinite dimensional dynamics is far from its completion. Here we have few principal problems such as the reduction principle (an analogue of Theorem 2 or 5) and the construction of the center manifold. The application of the center manifold theory to partial differential equations is one of promising and developing direction. In particular, it is very important to investigate the behavior of center manifolds when Galerkin method is applied due to a transition from finite dimension to infinite one. As it was mentioned above each application of center manifold methods to nonlinear infinite dimensional dynamics is nontrivial and gives rise to new directions of researches.

Bibliography

Primary Literature

1. Arnold VI (1973) Ordinary Differential Equations. MIT, Cambridge
2. Auchmuty J, Nicolis G (1976) Bifurcation analysis of reaction-diffusion equation (III). Chemical Oscilations. Bull Math Biol 38:325–350
3. Aulbach B (1982) A reduction principle for nonautonomous differential equations. Arch Math 39:217–232
4. Aulbach B, Garay B (1994) Partial linearization for noninvertible mappings. J Appl Math Phys (ZAMP) 45:505–542
5. Aulbach B, Colonius F (eds) (1996) Six Lectures on Dynamical Systems. World Scientific, New York
6. Balakrishnan AV (1976) Applied Functional Analysis. Springer, New York, Heidelberg
7. Bates P, Jones C (1989) Invariant manifolds for semilinear partial differential equations. In: Dynamics Reported 2. Wiley, Chichester, pp 1–38
8. Bogoliubov NN, Mitropolsky YUA (1963) The method of integral manifolds in non-linear mechanics. In: Contributions Differential Equations 2. Wiley, New York, pp 123–196
9. Caraballo T, Chueshov I, Landa J (2005) Existence of invariant manifolds for coupled parabolic and hyperbolic stochastic partial differential equations. Nonlinearity 18:747–767
10. Carr J (1981) Applications of Center Manifold Theory. In: Applied Mathematical Sciences, vol 35. Springer, New York
11. Carr J, Muncaster RG (1983) Applications of center manifold theory to amplitude expansions. J Diff Equ 59:260–288
12. Chicone C, Latushkin Yu (1999) Evolution Semigroups in Dynamical Systems and Differential Equations. Math Surv Monogr 70. Amer Math Soc, Providene
13. Chow SN, Lu K (1995) Invariant manifolds and foliations for quasiperiodic systems. J Diff Equ 117:1–27
14. Chueshov I (2007) Invariant manifolds and nonlinear master-slave synchronization in coupled systems. In: Applicable Analysis, vol 86, 3rd edn. Taylor and Francis, London, pp 269–286
15. Chueshov I (2004) A reduction principle for coupled nonlinesr parabolic-hyperbolic PDE. J Evol Equ 4:591–612
16. Diliberto S (1960) Perturbation theorem for periodic surfaces I, II. Rend Cir Mat Palermo 9:256–299; 10:111–161
17. Du A, Duan J (2006) Invariant manifold reduction for stochastic dynamical systems. http://arXiv:math.DS/0607366
18. Engel K, Nagel R (2000) One-parameter Semigroups for Linear Evolution Equations. Springer, New York
19. Fenichel N (1971) Persistence and smoothness of invariant manifolds for flows. Ind Univ Math 21:193–226
20. Fenichel N (1974) Asymptotic stability with rate conditions. Ind Univ Math 23:1109–1137
21. Gallay T (1993) A center-stable manifold theorem for differential equations in Banach space. Commun Math Phys 152:249–268
22. Gesztesy F, Jones C, Latushkin YU, Stanislavova M (2000) A spectral mapping theorem and invariant manifolds for nonlinear Schrodinger equations. Ind Univ Math 49(1):221–243
23. Grobman D (1959) Homeomorphism of system of differential equations. Dokl Akad Nauk SSSR 128:880 (in Russian)
24. Grobman D (1962) The topological classification of the vicinity of a singular point in *n*-dimensional space. Math USSR Sbornik 56:77–94; in Russian
25. Guckenheimer J, Holmes P (1993) Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vectors Fields. Springer, New York
26. Hadamard J (1901) Sur l'etaration et les solution asymptotiques des equations differentielles. Bull Soc Math France 29:224–228
27. Haken H (2004) Synergetics: Introduction and Advanced topics. Springer, Berlin
28. Hale J (1961) Integral manifolds of perturbed differential systems. Ann Math 73(2):496–531
29. Hartman P (1960) On local homeomorphisms of Euclidean spaces. Bol Soc Mat Mex 5:220

Copyrighted material

Chaos and Ergodic Theory

JÉRÔME BUZZI
C.N.R.S. and Université Paris-Sud,
Orsay, France

Article Outline

Glossary
Definition of the Subject
Introduction
Picking an Invariant Probability Measure
Tractable Chaotic Dynamics
Statistical Properties
Orbit Complexity
Stability
Untreated Topics
Future Directions
Acknowledgments
Bibliography

Glossary

For simplicity, definitions are given for a continuous or smooth self-map or diffeomorphism T of a compact manifold M.

Entropy, measure-theoretic (or: metric entropy) For an ergodic invariant probability measure μ, it is the smallest exponential growth rates of the number of orbit segments of given length, with respect to that length, after restriction to a set of positive measure. We denote it by $h(T, \mu)$. See ▶ Entropy in Ergodic Theory and Subsect. "Local Complexity" below.

Entropy, topological It is the exponential growth rates of the number of orbit segments of given length, with respect to that length. We denote it by $h_{\text{top}}(f)$. See ▶ Entropy in Ergodic Theory and Subsect. "Local Complexity" below.

Ergodicity A measure is ergodic with respect to a map T if given any measurable subset S which is invariant, i. e., such that $T^{-1}S = S$, either S or its complement has zero measure.

Hyperbolicity A measure is hyperbolic in the sense of Pesin if at almost every point no Lyapunov exponent is zero. See ▶ Smooth Ergodic Theory.

Kolmogorov typicality A property is *typical in the sense of Kolmogorov* for a topological space $\mathcal{F}$ of parametrized families $f = (f_t)_{t\in U}$, U being an open subset of $\mathbb{R}^d$ for some $d \geq 1$, if it holds for f_t for Lebesgue almost every t and topologically generic $f \in \mathcal{F}$.

Lyapunov exponents The Lyapunov exponents (▶ Smooth Ergodic Theory) are the limits, when they exist, $\lim_{n\to\infty} \frac{1}{n} \log \|(T^n)'(x).v\|$ where $x \in M$ and v is a non zero tangent vector to M at x. The Lyapunov exponents of an ergodic measure is the set of Lyapunov exponents obtained at almost every point with respect to that measure for all non-zero tangent vectors.

Markov shift (topological, countable state) It is the set of all infinite or bi-infinite paths on some countable directed graph endowed with the left-shift, which just translates these sequences.

Maximum entropy measure It is a measure μ which maximizes the measured entropy and, by the variational principle, realized the topological entropy.

Physical measure It is a measure μ whose basin, $\{x \in M\colon \forall \phi\colon M \to \mathbb{R}$ continuous $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(f^k x) = \int \phi \, d\mu\}$ has nonzero volume.

Prevalence A property is *prevalent* in some complete metric, separable vector space X if it holds outside of a set N such that, for some Borel probability measure μ on X: $\mu(A + v) = 0$ for all $v \in X$. See [76,141,239].

Sensitivity on initial conditions T has *sensitivity to initial conditions* on $X' \subset X$ if there exists a constant $\rho > 0$ such that for every $x \in X'$, there exists $y \in X$, arbitrarily close to x, and $n \geq 0$ such that $d(T^n y, T^n x) > \rho$.

Sinai–Ruelle–Bowen measures It is an invariant probability measure which is absolutely continuous along the unstable foliation (defined using the unstable manifolds of almost every $x \in M$, which are the sets $W^u(x)$, of points y such that $\lim_{n\to\infty} \frac{1}{n} \log d(T^{-n} y, T^{-n} x) < 0$).

Statistical stability T is statistically stable if the physical measures of nearby deterministic systems are arbitrarily close to the convex hull of the physical measures of T.

Stochastic stability T is stochastically stable if the invariant measures of the Markov chains obtained from T by adding a suitable, smooth noise with size $\epsilon \to 0$ are arbitrarily close to the convex hull of the physical measures of T.

Structural stability T is structurally stable if any S close enough to T is topologically the same as T: there exists a homeomorphism $h\colon M \to M$ such that $h \circ T = S \circ h$ (orbits are sent to orbits).

Subshift of finite type It is a closed subset Σ_F of $\Sigma = \mathcal{A}^{\mathbb{Z}}$ or $\Sigma = \mathcal{A}^{\mathbb{N}}$ where $\mathcal{A}$ is a finite set satisfying: $\Sigma_F = \{x \in \Sigma\colon \forall k < \ell\colon x_k x_{k+1} \dots x_\ell \notin F\}$ for some finite set F.

Copyrighted material

Topological genericity Let X be a Baire space, e. g., a complete metric space. A property is *(topologically) generic* in a space X (or holds for the (topologically) generic element of X) if it holds on a nonmeager set (or set of second Baire category), i. e., on a dense G_δ subset.

Definition of the Subject

Chaotic dynamical systems are those which present unpredictable and/or complex behaviors. The existence and importance of such systems has been known at least since Hadamard [126] and Poincaré [208], however it became well-known only in the sixties. We refer to [36,128,226,236] and [80,107,120,125,192,213] for the relevance of such dynamics in other fields, mathematical or not (see also ▶ Ergodic Theory: Interactions with Combinatorics and Number Theory, ▶ Ergodic Theory: Fractal Geometry). The numerical simulations of chaotic dynamics can be difficult to interpret and to plan, even misleading, and the tools and ideas of mathematical dynamical system theory are indispensable. Arguably the most powerful set of such tools is ergodic theory which provides a statistical description of the dynamics by attaching relevant probability measures. In opposition to single orbits, the statistical properties of chaotic systems often have good stability properties. In many cases, this allows an understanding of the complexity of the dynamical system and even precise and quantitative statistical predictions of its behavior. In fact, chaotic behavior of single orbits often yields global stability properties.

Introduction

The word *chaos*, from the ancient Greek $\chi\alpha o\sigma$, "shapeless void" [131] and "raw confused mass" [199], has been used [*$\chi\alpha o\sigma$ also inspired Van Helmont to create the word "gas" in the seventeenth century and this other thread leads to the molecular chaos of Boltzmann in the nineteenth century and therefore to ergodic theory itself*.] since a celebrated paper of Li and Yorke [169] to describe evolutions which however deterministic and defined by rather simple rules, exhibit unpredictable or complex behavior.

Attempts at Definition

We note that, like many ideas [237], this is not captured by a single mathematical definition, despite several attempts (see, e. g., [39,112,158,225] for some discussions as well as the monographs on chaotic dynamics [15,16,58,60,78,104,121,127,215,250,261]). Let us give some of the most well-known definitions, which have been given mostly from the topological point of view, i. e., in the setting of a self-map $T\colon X \to X$ on a compact metric space whose distance is denoted by d:

T has *sensitivity to initial conditions* on $X' \subset X$ if there exists a constant $\rho > 0$ such that for every $x \in X'$, there exists $y \in X$, arbitrarily close to x with a finite separating time:

$$\exists n \geq 0 \quad \text{such that} \quad d(T^n y, T^n x) > \rho\,.$$

In other words, any uncertainty on the exact value of the initial condition x makes $T^n(x)$ completely unknown for n large enough. If X is a manifold, then *sensitivity to initial conditions in the sense of Guckenheimer* [120] means that the previous phenomenon occurs for a set X' with nonzero volume.

T is *chaotic in the sense of Devaney* [94] if it admits a dense orbit and if the periodic points are dense in X. It implies sensitivity to initial conditions on X.

T is *chaotic in the sense of Li and Yorke* [169] if there exists an uncountable subset $X' \subset X$ of points, such that, for all $x \neq y \in X'$,

$$\liminf_{n\to\infty} d(T^n x, T^n y) = 0 \quad \text{and}$$
$$\limsup_{n\to\infty} d(T^n x, T^n y) > 0\,.$$

T has *generic chaos* in the sense of Lasota [206] if the set

$$\{(x,y) \in X \times X\colon \liminf_{n\to\infty} d(T^n x, T^n y) = 0 < \limsup_{n\to\infty} d(T^n x, T^n y)\}$$

is topologically generic (*see glossary.*) in $X \times X$.

Topological chaos is also sometimes characterized by *nonzero topological entropy* (▶ Entropy in Ergodic Theory): there exist exponentially many orbit segments of a given length. This implies chaos in the sense of Li and Yorke by [39].

As we shall see ergodic theory describes a number of chaotic properties, many of them implying some or all of the above topological ones. The main such property for a smooth dynamical system, say a $C^{1+\alpha}$-diffeomorphism of a compact manifold, is the existence of an invariant probability measure which is:

1. Ergodic (cannot be split) and aperiodic (not carried by a periodic orbit);
2. Hyperbolic (nearby orbits converge or diverge at a definite exponential rate);
3. Sinai–Ruelle–Bowen (as smooth as it is possible).

(For precise definitions we refer to ▶ Smooth Ergodic Theory or to the discussions below.) In particular such

Copyrighted material

a situation implies nonzero entropy and sensitivity to initial condition of a set of nonzero Lebesgue measure (i. e., positive volume).

Before starting our survey in earnest, we shall describe an elementary and classical example, the full tent map, on which the basic phenomena can be analyzed in a very elementary way. Then, in Sect. "Picking an Invariant Probability Measure", we shall give some motivations for introducing probability theory in the description of chaotic but deterministic systems, in particular the unpredictability of their individual orbits. We define two of the most relevant classes of invariant measures: the physical measures and those maximizing entropy. It is unknown in which generality these measures exist and can be analyzed but we describe in Sect. "Tractable Chaotic Dynamics" the major classes of dynamics for which this has been done. In Sect. "Statistical Properties" we describe some of the finer statistical properties that have been obtained for such good chaotic systems: sums of observables along orbits are statistically undistinguishable from sums of independent and identically distributed random variables. Sect. "Orbit Complexity" is devoted to the other side of chaos: the complexity of these dynamics and how, again, this complexity can be analyzed, and sometimes classified, using ergodic theory. Sect "Stability" describes perhaps the most striking aspect of chaotic dynamics: the unstability of individual orbit is linked to various forms of stability of the global dynamics.

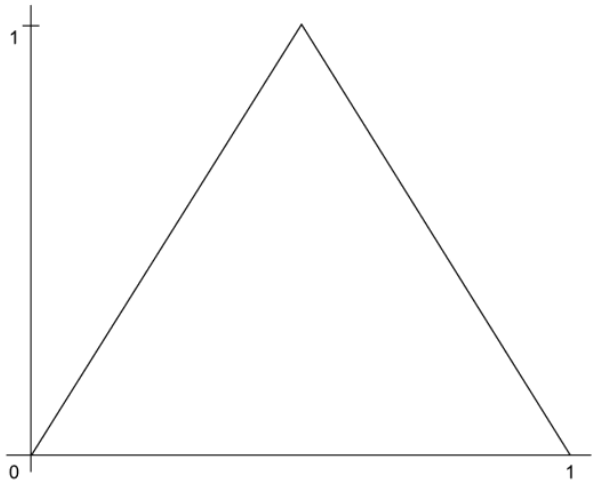
Finally we conclude by mentioning some of the most important topics that we could not address and we list some possible future directions.

Caveat. The subject-matter of this article is somewhat fuzzy and we have taken advantage of this to steer our path towards some of our favorite theorems and to avoid the parts we know less (some of which are listed below). We make no pretense at exhaustivity neither in the topics nor in the selected results and we hope that our colleagues will excuse our shortcomings.

Remark 1 In this article we only consider *compact, smooth and finite-dimensional* dynamical systems *in discrete time*, i. e., defined by self-maps. In particular, we have omitted the natural and important variants applying to flows, e. g., evolutions defined by ordinary differential equations but we refer to the textbooks (see, e. g.,[15,128,148]) for these.

Elementary Chaos: A Simple Example

We start with a toy model: the full tent map T of Fig. 1. Observe that for any point $x \in [0,1]$, $T^{-n}(x) = \{(\sigma(k,n)x+k)\cdot 2^{-n}\colon k = 0,1,\dots,2^n-1\}$, where $\sigma(k,n) = \pm 1$.



Chaos and Ergodic Theory, Figure 1
The graph of the full tent map $f(x) = 1 - |1 - 2x|$ over $[0, 1]$

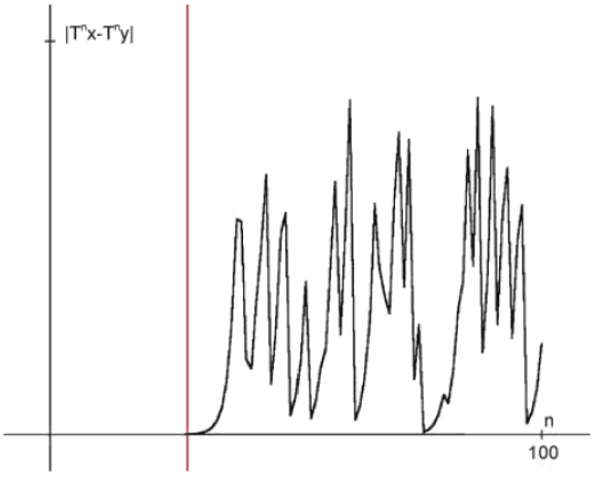
Hence $\bigcup_{n\geq 0} T^{-n}(x)$ is dense in $[0,1]$. It easily follows that T exhibits *sensitive dependence to initial conditions*. Even worse in this example, the qualitative asymptotic behavior can be completely changed by this arbitrarily small perturbation: x may have a dense orbit whereas y is eventually mapped to a fixed point! This is Devaney chaos [94].

This kind of unstability was first discovered by J. Hadamard [126] in his study of the geodesic flow (i. e., the frictionless movement of a point mass constrained to remain on a surface). At that time, such an unpredictability was considered a purely mathematical pathology, necessarily devoid of any physical meaning [*Duhem qualified Hadamard's result as "an example of a mathematical deduction which can never be used by physics" (see pp. 206–211 in* [103])!].

Returning to out tent map, we can be more quantitative. At any point $x \in [0,1]$ whose orbit never visits $1/2$, the *Lyapunov exponent* $\lim_{n\to\infty} \frac{1}{n}\log|(T^n)'(x)|$ is $\log 2$. (*See the glossary.*) Such a positive Lyapunov exponent corresponds to infinitesimally close orbits getting separated exponentially fast. This can be observed in Fig. 2. Note how this exponential speed creates a rather sharp transition.

It follows in particular that experimental or numerical errors can grow very quickly to size 1, [*For simple precision arithmetic the uncertainty is* 10^{-16} *which grows to size 1 in 38 iterations of* T.] i. e., the approximate orbit may contain after a while *no information about the true orbit*. This casts a doubt on the reliability of simulations. Indeed, a simulation of T on most computers will suggest that all orbits quickly converge to 0, which is completely false [*Such a collapse to 0 does really occurs but only for a countable*

Copyrighted material



Chaos and Ergodic Theory, Figure 2
$|T^n(x) - T^n(y)|$ for $T(x) = 1 - |1 - 2x|$, $|x - y| = 10^{-12}$ and $0 \le n \le 100$. The vertical red line is at $n = 28$ and shows when $|T^n x - T^n y| \ge 0.5 \cdot 10^{-4}$ for the first time

subset of initial conditions in $[0, 1]$ *whereas the points with dense orbit make a subset of* $[0, 1]$ *with full Lebesgue measure (see below). This artefact comes from the way numbers are represented – and approximated – on the computer: multiplication by even integers tends to "simplify" binary representations. Thus the computations involved in drawing Fig. 2 cannot be performed too naively.*]. Though somewhat atypical in its dramatic character, this failure illustrates the *unpredictability* and *unstability* of individual orbits in chaotic systems.

Does this mean that all quantitative predictions about orbits of T are to be forfeited? Not at all, *if we are ready to change our point of view* and look beyond a single orbit. This can be seen easily in this case. Let us start with such a global analysis from the topological point of view. Associate to $x \in [0, 1]$, a sequence $i(x) = i = i_0 i_1 i_2 \dots$ of 0s and 1s according to $i_k = 0$ if $T^k x \le 1/2$, $i_k = 1$ otherwise. One can check that [*Up to a countable set of exceptions.*] $\{i(x)\colon x \in [0, 1]\}$ is the set $\Sigma_2 := \{0, 1\}^{\mathbb{N}}$ of *all* infinite sequences of 0s and 1s and that at most one $x \in [0, 1]$ can realize a given sequence as $i(x)$.

Notice how the transformation f becomes trivial in this representation:

$$i(f(x)) = i_1 i_2 i_3 \dots \quad \text{if } i(x) = i_0 i_1 i_2 i_3 \dots$$

Thus f is represented by the simple and universal "left-shift" on sequences, which is denoted by σ. This representation of a rather general dynamical system by the left-shift on a space of sequences is called *symbolic dynamics* (▶ Symbolic Dynamics), [171].

This can be a very powerful tool. Observe for instance how here it makes obvious that we have complete *combinatorial freedom* over the orbits of T: one can easily build orbits with various asymptotic behaviors: if a sequence of Σ_2 contains all the finite sequences of 0s and 1s, then the corresponding point has a dense orbit; if the sequence is periodic, then the corresponding point is itself periodic, to give two examples of the *richness* of the dynamics.

More quantitatively, the number of distinct subsequences of length n appearing in sequences $i(x)$, $x \in [0, 1]$, is 2^n. It follows that the *topological entropy* (▶ Entropy in Ergodic Theory) of T is $h_{\text{top}}(T) = \log 2$. [*For the coincidence of the entropy and the Lyapunov exponent see below.*] The positivity of the topological entropy can be considered as the signature of the complexity of the dynamics and considered as the definition, or at least the stamp, of a topologically chaotic dynamics.

Let us move on to a probabilistic point of view. Pick $x \in [0, 1]$ randomly according to, say, the uniform law in $[0, 1]$. It is then routine to check that $i(x)$ follows the $(1/2, 1/2)$-Bernoulli law: the probability that, for any given k, $i_k(x) = 0$ is $1/2$ and the i_ks are independent. Thus $i(x)$, seen as a sequence of random 0 and 1 when x is subject to the uniform law on $[0, 1]$, is *statistically undistinguishable from coin tossing!* This important remark leads to quantitative predictions. For instance, the *strong law of large numbers* implies that, for Lebesgue-almost every $x \in [0, 1]$ (i. e., for all $x \in [0, 1]$ except in a set of Lebesgue measure zero), the fraction of the time spent in any dyadic interval $I = [k \cdot 2^{-N}, \ell \cdot 2^{-N}] \subset [0, 1]$, $k, \ell, N \in \mathbb{N}$, by the orbit of x,

$$\lim_{n\to\infty} \frac{1}{n} \#\{0 \le k < n\colon T^k x \in I\} \tag{1}$$

exists and is equal to the length, 2^{-N}, of that interval. [*Eq. (1) in fact holds for any interval I. This implies that the orbit of almost every* $x \in [0, 1]$ *visits all subintervals of* $[0, 1]$, *i. e., the orbit is dense: in complete contradiction with the above mentioned numerical simulation!*] More generally, we shall see that, if $\phi\colon [0, 1] \to \mathbb{R}$ is any continuous function, then, for Lebesgue almost every-x,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(T^k x) \text{ exists and is equal to } \int \phi(x)\, dx\,. \tag{2}$$

Using strong *mixing properties* (▶ Ergodicity and Mixing Properties) of the Lebesgue measure under T, one can prove further properties, e. g., sensitivity on initial conditions in the sense of Guckenheimer [*The Lebesgue measure is* weak-mixing*: Lebesgue-almost all couples of points* $(x, y) \in [0, 1]^2$ *get separated. Note that it is not true*

Copyrighted material

of every couple off the diagonal: Counter-examples can be found among couples (x, y) *with* $T^n x = T^n y$ *arbitrarily close to 1.*] and study the fluctuations of the averages $\frac{1}{n}\sum_{k=0}^{n-1}\phi(Tx)$ by the way of limit theorems.

The above analysis relied on the very special structure of T but, as we shall explain, the ergodic theory of differentiable dynamical systems shows that all of the above (and much more) holds in some form for rather general classes of chaotic systems. The different chaotic properties are independent in general (e. g., one may have topological chaos whereas the asymptotic behavior of almost all orbits is periodic) and the proofs can become much more difficult. We shall nonetheless be rewarded for our efforts by the discovery of unexpected links between chaos and stability, complexity and simplicity as we shall see.

Picking an Invariant Probability Measure

One could think that dynamical systems, such as those defined by self-maps of manifolds, being *completely deterministic*, have nothing to do with probability theory. There are in fact several motivations for introducing various invariant probability measures.

Statistical Descriptions

An abstract goal might be to enrich the structure: a smooth self-map is a particular case of a Borel self-map, hence one can canonically attach to this map its set of all invariant Borel probability measures [*From now on all measures will be Borel probability measures except if it is explicitly stated otherwise.*], or just the set of *ergodic* [*A measure is* ergodic *if all measurable invariant subsets have measure 0 or 1. Note that arbitrary invariant measures are averages of ergodic ones, so many questions about invariant measures can be reduced to ergodic ones.*] ones. By the Krylov–Bogoliubov theorem (see, e. g., [148]), this set is non-empty for any continuous self-map of a compact space.

By the following fundamental theorem ▶ Ergodic Theorems, each such measure is the statistical description of some orbit:

Birkhoff Pointwise Ergodic Theorem *Let* $(X, \mathcal{F}, \mu)$ *be a space with a* σ*-field and a probability measure. Let* $f\colon X \to X$ *be a measure-preserving map, i. e.,* $f^{-1}(\mathcal{F}) \subset \mathcal{F}$ *and* $\mu \circ f^{-1} = \mu$*. Assume ergodicity of* μ *(See the glossary.) and (absolute) integrability of* $\phi\colon X \to \mathbb{R}$ *with respect to* μ*. Then for* μ*-almost every* $x \in X$*,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \phi(f^k x) \quad \textit{exists and is} \quad \int \phi \, \mathrm{d}\mu \,.$$

This theorem can be interpreted as saying that "time averages" coincide almost surely with "ensemble averages" (or "phase space average"), i. e., that Boltzmann's Ergodic Hypothesis of statistical mechanics [110] holds for dynamical systems that cannot be split in a measurable and non trivial way. [*This indecomposability is however often difficult to establish. For instance, for the hard ball model of a gas it is known only under some generic assumption (see* [238] *and the references therein).*] We refer to [161] for background.

Remark 2 One should observe that the existence of the above limit is not at all obvious. In fact it often fails from other points of view. One can show that for the full tent map $T(x) = 1 - |1 - 2x|$ analyzed above and many functions ϕ, the set of points for which it fails is large both *from the topological point of view* (it contains a dense G_δ set) and *from the dimension point of view* (it has Hausdorff dimension 1 [28]). This is an important point: the introduction of invariant measures allows one to avoid some of the wilder pathologies.

To illustrate this let us consider the full tent map $T(x) = 1 - |1 - 2x|$ again and the two ergodic invariant measures: δ_0 (the Dirac measure concentrated at the fixed point 0) and the Lebesgue measure $\mathrm{d}x$. In the first case, we obtain a complex proof of the obvious fact that the time average at $x = 0$ (some set of full measure!) and the ensemble average with respect to δ_0 are both equal to $f(0)$. In the second case, we obtain a very general proof of the above Eq. (2).

Another type of example is provided by the contracting map $S\colon [0, 1] \to [0, 1]$, $S(x) = x/2$. S has a unique invariant probability measure, δ_0. For Birkhoff theorem the situation is the same as that of T and δ_0: it asserts only that the orbit of 0 is described by δ_0.

One can understand Birkhoff theorem as a (first and rather weak) *stability result*: the time averages are independent of the initial condition, *almost surely with respect to* μ.

Physical Measures

In the above silly example S, much more is true than the conclusion of Birkhoff Theorem: *all* points of $[0, 1]$ are described by δ_0. This leads to the definition of the *basin* of a probability measure μ for a self-map f of a space M:

$$\mathcal{B}(\mu) := \left\{ x \in M\colon \forall \phi\colon M \to \mathbb{R} \text{ continuous } \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \phi(f^k x) = \int \phi \, \mathrm{d}\mu \right\}.$$

Copyrighted material

The classical approach to the uniformly hyperbolic dynamics [52,222,233] is through symbolic dynamics and coding. Under the above hypothesis one can build a finite partition of M which is tailored to the dynamics (a *Markov partition*) so that the corresponding symbolic dynamics has a very simple structure: it is a full shift $\{1,\dots,d\}^{\mathbb{Z}}$, like in the example of the full tent map, or a subshifts of finite type. The above problems can then be solved using the thermodynamical formalism inspired from the statistical mechanics of one-dimensional ferromagnets [217]: ergodic properties are obtained through the spectral properties of a suitable transfer operator acting on some space of regular functions, e. g., the Hölder-continuous functions defined over the symbolic dynamics with respect to the distance $d(x,y) := \sum_{n\in\mathbb{Z}} 2^{-n} 1_{x_n\neq y_n}$ where $1_{s\neq t}$ is 1 if $s \neq t$, 0 otherwise.

A recent development [24,43,116] has been to find suitable Banach spaces to apply the transfer operator technique directly in the smooth setting, which not only avoids the complication of coding (or rather replace them with functional analytic preliminaries) but allows the use of the smoothness beyond Hölder-continuity which is important for finer ergodic properties.

Uniform expansion or hyperbolicity can easily be obstructed in a given system: a "bad" point (a critical point or a periodic point with multiplier with an eigenvalue on the unit circle) is enough. This leads to the study of other systems and has motivated many works devoted to relaxing the uniform hyperbolicity assumptions [51].

Pesin Theory

The most general such approach is *Pesin theory*. Let f be a $C^{1+\alpha}$-diffeomorphism [*It is an important open problem to determine to which extent Pesin theory could be generalized to the C^1 setting.*] f with an ergodic invariant measure μ. By Oseledets Theorem ▶ Smooth Ergodic Theory, for almost every x with respect to any invariant measure, the behavior of the differential $T_x f^n$ for n large is described by the Lyapunov exponents $\lambda_1,\dots,\lambda_d$, at x. Pesin is able to build charts around almost every orbit in which this asymptotic linear behavior describes that of f at the first iteration. That is, there are diffeomorphisms $\Phi_x : U_x \subset M \to V_x \subset \mathbb{R}^d$ with a "reasonable dependence on x" such that the differential of $\Phi_{f^n x} \circ f^n \circ \Phi_x^{-1}$ at any point where it is defined is close to a diagonal matrix with entries $(e^{(\lambda_1\pm\epsilon)n}, e^{(\lambda_2\pm\epsilon)n},\dots,e^{(\lambda_d\pm\epsilon)n})$.

In this full generality, one already obtains significant results:

- The entropy is bounded by the expansion: $h(f,\mu) \leq \sum_{i=1}^{d} \lambda_i^+(\mu)$ [219]
- At almost every point x, there are strong stable resp. unstable manifolds $W^{ss}(x)$, resp. $W^{uu}(x)$, coinciding with the sets of points y such that $d(T^n x, T^n y) \to 0$ exponentially fast when $n \to \infty$, resp. $n \to -\infty$. The corresponding holonomies are *absolutely continuous* (see, e. g., [59]) like in the uniform case. This allows Ledrappier's definition of Sinai–Ruelle–Bowen measures [166] in that setting.
- Equality in the above formula holds if and only if μ is a Sinai–Ruelle–Bowen measure [167]. More generally the entropy can be computed as $\sum_{i=1}^{d} \gamma_i(\mu)\lambda_i^+(\mu)$ where the γ_i are some fractal dimensions related to the exponents.

Under the only assumption of hyperbolicity (i. e., no zero Lyapunov exponent almost everywhere), one gets further properties:

- Existence of an hyperbolic measure which is not periodic forces $h_{\text{top}}(f) > 0$ [*However $h(f,\mu)$ can be zero.*] by [147].
- μ is *exact dimensional* [29,256]: the limit $\lim_{r\to 0} \log \mu(B(x,r))/\log r$ exists μ-almost everywhere and is equal to the Hausdorff dimension of μ (the infimum of the Hausdorff dimension of the sets with full μ-measure). This is deduced from a more technical "asymptotic product structure" property of any such measure.

For hyperbolic Sinai–Ruelle–Bowen measures μ, one can then prove, e. g.,:

- *local ergodicity* [202]: μ has at most countably many ergodic components and μ-almost every point has a neighborhood whose Lebesgue-almost every point are contained in the basin of an ergodic component of μ.
- *Bernoulli* [198]: Each ergodic component of μ is conjugate in a measure-preserving way, up to a period, to a Bernoulli shift, that is, a full shift $\{1,\dots,N\}^{\mathbb{Z}}$ equipped with a product measure. This in particular implies mixing and sensitivity on initial conditions for a set of positive Lebesgue measure.

However, establishing even such a weak form of hyperbolicity is rather difficult. The fragility of this condition can be illustrated by the result [44,45] that the topologically generic area-preserving surface C^1-diffeomorphism is either uniformly hyperbolic or has Lebesgue almost everywhere vanishing Lyapunov exponents, hence is never non-uniformly hyperbolic! (but this is believed to be very specific to the very weak C^1 topology). Moreover, such weak hyperbolicity is not enough, with the current techniques, to build Sinai–Ruelle–Bowen measures or ana-

Copyrighted material

lyze maximum entropy measures only assuming non-zero Lyapunov exponents. Let us though quote two conjectures. The first one is from [251] [*We slightly strengthened Viana's statement for expository reasons.*]:

Conjecture 1 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact manifold. If Lebesgue-almost every point x has well-defined Lyapunov exponents in every direction and none of these exponents is zero, then there exists an absolutely continuous invariant σ-finite positive measure.*

The analogue of this conjecture has been proved for C^3 interval maps with unique critical point and negative Schwarzian derivative by Keller [150], but only partial results are available for diffeomorphisms [168].

We turn to measures of maximum entropy. As we said, C^∞ smoothness is enough to ensure their existence but this is through a functional-analytic argument (allowed by Yomdin theory [254]) which says nothing about their structure. Indeed, the following problem is open:

Conjecture 2 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact surface. If the topological entropy of f is nonzero then f has at most countably many ergodic invariant measures maximizing entropy.*

The analogue of this conjecture has been proved for $C^{1+\epsilon}$ interval maps [64,66,70]. In the above setting a classical result of Katok shows the existence of uniformly hyperbolic compact invariant subsets with topological entropy arbitrarily close to that of f implying the existence of many periodic points:

$$\limsup_{n\to\infty} \frac{1}{n} \log \#\{x \in M : f^n(x) = x\} \geq h_{\mathrm{top}}(f) .$$

The previous conjecture would follow from the following one:

Conjecture 3 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact manifold. There exists an invariant subset $X \subset M$, carrying all ergodic measures with maximum entropy, such that the restriction $f|X$ is conjugate to a countable state topological Markov shift (See the glossary.).*

Systems with Discontinuities

We now consider stronger assumptions to be able to build the relevant measures.

The simplest step beyond uniformity is to allow discontinuities, considering piecewise expanding maps. The discontinuities break the rigidity of the uniformly expanding situation. For instance, their symbolic dynamics are usually no longer subshifts of finite type though they still retain some "simplicity" in good cases (see [68]).

To understand the problem in constructing the absolutely continuous invariant measures, it is instructive to consider the pushed forwards of a smooth measure. Expansion tends to keep the measure smooth whereas discontinuities may pile it up, creating non-absolute continuity in the limit. One thus has to check that expansion wins. In dimension 1, a simple fact resolves the argument: under a high enough iterate, one can make the expansion arbitrarily large everywhere, whereas a small interval can be chopped into at most two pieces.

Lasota and Yorke [165] found a suitable framework. They considered C^2 interval maps with $|f'(x)| \geq \text{const} > 1$ except at finitely many points. They used the Ruelle transfer operator directly on the interval. Namely they studied

$$(L\phi)(x) = \sum_{y \in T^{-1}x} \frac{\phi(y)}{|T'(y)|}$$

acting on functions $\phi : [0, 1] \to \mathbb{R}$ with bounded variation and obtained the invariant density as the eigenfunction associated to the eigenvalue 1. One can then prove a *Lasota–Yorke inequality* (which might more accurately be called Doeblin–Fortet since it was introduced in the theory of Markov chains much earlier):

$$\|L\phi\|_{\mathrm{BV}} \leq \alpha \|\phi\|_{\mathrm{BV}} + \beta \|\phi\|_1 \tag{3}$$

where $\|\cdot\|_{\mathrm{BV}}, \|\cdot\|_1$ are a strong and a weak norm, respectively and $\alpha < 1$ and $\beta < \infty$. One can then apply general theorems [143] or [196] (see [21] for a detailed presentation of this approach and its variants). Here α can essentially be taken as 2 (reflecting the locally simple discontinuities) divided by the minimum expansion: so $\alpha < 1$, perhaps after replacing T with an iterate. In particular, the existence of a finite statistical description then follows (see [61] for various generalizations and strengthenings of this result on the interval).

The situation in higher dimension is more complex for the reason explained above. One can obtain inequalities such as (3) on suitable if less simple functional spaces (see, e. g., [231]) but proving $\alpha < 1$ is another matter: discontinuities can get arbitrarily complex under iteration. [67,241] show that indeed, in dimension 2 and higher, piecewise uniform expansion (with a finite number of pieces) is not enough to ensure a finite statistical description if the pieces of the map have only finite smoothness. In dimension 2, resp. 3 or more, piecewise real-analytic, resp. piecewise affine, is enough to exclude such examples [65,240], resp. [242]. [82] has shown that, for any $r > 1$, an open and dense subset of piecewise C^r and expanding maps have a finite statistical description.

Copyrighted material

Piecewise hyperbolic diffeomorphisms are more difficult to analyze though several results (conditioned on technical assumptions that can be checked in many cases) are available [22,74,230,257].

Interval Maps with Critical Points

A more natural but also more difficult situation is a map for which the uniformity of the expansion fails because of the existence of critical points. [*Note that, by a theorem of Mañé a circle map without critical points or indifferent periodic point is either conjugate to a rotation or uniformly expanding* [181].]

A class which has been completely analyzed at the level of the above conjecture is that of *real-analytic families of maps of the interval* $f_t\colon [0,1] \to [0,1]$, $t \in I$, with a unique critical point, the main example being the quadratic family $Q_t(x) = tx(1-x)$ for $0 \le t \le 4$.

It is not very difficult to find quadratic maps with the following two types of behavior:

(stable) the orbit of Lebesgue-almost every $x \in [0,1]$ tends to an attracting periodic orbit;

(chaotic) there is an absolutely continuous invariant probability measure μ whose basin contains Lebesgue-almost every $x \in [0,1]$.

To realize the first it is enough to arrange the critical point to be periodic. One can easily prove that this stable behavior occurs on an open set of parameters – thus it is stable with respect to the parameter or the dynamical system. The second occurs for Q_4 with $\mu = \mathrm{d}x/\sqrt{\pi x(1-x)}$. It is much more difficult to show that this chaotic behavior occurs for a set of parameters of positive Lebesgue measure. This is a theorem of Jakobson [145] for the quadratic family (see for a recent variant [265]). Let us sketch two main ingredients of the various proofs of this theorem. The first is *inducing*: around Lebesgue-almost every point $x \in [0,1]$ one tries to find a time $\tau(x)$ and an interval $J(x)$ such that $f^{\tau(x)}\colon J(x) \to f^{\tau(x)}(J(x))$ is a map with good expansion and distortion properties. This powerful idea appears in many disguises in the non-uniform hyperbolic theory (see for instance [133,262]). The second ingredient is *parameter exclusion*: one removes the parameters at which a good inducing scheme cannot be built. More precisely one proceeds inductively, performing simultaneously the inducing and the exclusion, the good properties of the early stage of the inducing allowing one to control the measure of the parameters that need to be excluded to continue [30,145]. Indeed, the expansion established at a given stage allows to transfer estimates in the dynamical space to the parameter space.

Using methods from complex analysis and renormalization theory one can go much further and prove the following difficult theorems (actually the product of the work of many people, including Avila, Graczyk, Kozlovski, Lyubich, de Melo, Moreira, Shen, van Strien, Swiatek), which in particular solves Palis conjecture in this setting:

Theorem 2 [117,160,178] *Stable maps (that is, such that Lebesgue almost every orbit converges to one of finitely many periodic orbits) form an open and dense set among C^r interval maps, for any $r \ge 2$. [In fact this is even true for polynomials.].*

The picture has been completed in the *unimodal case* (that is, with a unique critical point):

Theorem 3 [19,20,117,159,178] *Let $f_t\colon [0,1] \to [0,1]$, $t \in [t_0, t_1]$, be a real-analytic family of unimodal maps. Assume that it is not degenerate [f_{t_0} and f_{t_1} are not conjugate]. Then:*

- *The set of t such that f_t is chaotic in the above sense has positive Lebesgue measure;*
- *The set of t such that f_t is stable is open and dense;*
- *The remaining set of parameters has zero Lebesgue measure. [This set of "strange parameters" of zero Lebesgue measure has however positive Hausdorff dimension according to work of Avila and Moreira. In particular each of the following situations is realized on a set of parameters t of positive Hausdorff dimension: non-existence of the Birkhoff limit at Lebesgue-almost every point, the physical measure is δ_p for p a repelling fixed point, the physical measure is non-ergodic.]*

We note that the theory underlying the above theorem yields much more results, including a very paradoxical rigidity of typical analytic families as above. See [19].

Non-uniform Expansion/Contraction

Beyond the dimension 1, only partial results are available. The most general of those assume uniform contraction or expansion along some direction, restricting the non-uniform behavior to an invariant sub-bundle often one-dimensional, or "one-dimensional-like".

A first, simpler situation is when there is a *dominated decomposition* with a uniformly expanding term: there is a continuous and invariant splitting of the tangent bundle, $T_\Lambda M = E^{\mathrm{uu}} \oplus E^{\mathrm{cs}}$, over some Λ an attracting set: for all unit vectors $v^u \in E^{\mathrm{uu}}$, $v^c \in E^{\mathrm{cs}}$,

$$\|(f^n)'(x).v^u\| \ge C\lambda^n \quad \text{and}$$
$$\|(f^n)'(x).v^c\| \le C\mu^n \|(f^n)'(x).v^u\|\,.$$

Standard techniques (pushing the Riemannian volume of a piece of unstable leaf and taking limits) allow the construction of Gibbs u-states as introduced by [205].

Copyrighted material

Theorem 4 (Alves–Bonatti–Viana [8]) *[A slightly different result is obtained in [49].] Let $f: M \to M$ be a C^2 diffeomorphism with an invariant compact subset Λ. Assume that there is a dominated splitting $T_\Lambda M = E^u \oplus E^{cs}$ such that, for some $c > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \prod_{k=0}^{n-1} \|f'(f^x)|E^{cs}\| \le -c < 0$$

on a subset of Λ of positive Lebesgue measure. Then this subset is contained, up to a set of zero Lebesgue measure, in the union of the basins of finitely many ergodic and hyperbolic Sinai – Ruelle – Bowen measures.

The non-invertible, purely expansive version of the above theorem can be applied in particular to the following maps of the cylinder ($d \ge 16$, a is properly chosen close to 2 and α is small):

$$f(\theta, x) = (\mathrm{d}x \bmod 2\pi, a - x^2 + \epsilon \sin(\theta))$$

which are natural examples of maps with multidimensional expansion and critical lines considered by Viana [249]. A series of works have shown that the above maps fit in the above non-uniformly expanding setting with a proper control of the critical set and hence can be thoroughly analyzed through variants of the above theorem [4,249] and the references in [5]. For a, b properly chosen close to 2 and small ϵ, the following maps should be even more natural examples:

$$f(x, y) = (a - x^2 + \epsilon y, b - y^2 + \epsilon x)\,. \tag{4}$$

However the inexistence of a dominated splitting has prevented the analysis of their physical measures. See [66,70] for their maximum entropy measures.

Cowieson and Young have used completely different techniques (thermodynamical formalism, Ledrappier–Young formula and Yomdin theory on entropy and smoothness) to prove the following result (see [13] for related work):

Theorem 5 (Cowieson–Young [83]) *Let $f: M \to M$ be a C^∞ diffeomorphism of a compact manifold. Assume that f admits an attractor $\Lambda \subset M$ on which the tangent bundle has an invariant continuous decomposition $T_\Lambda M = E^+ \oplus E^-$ such that all vectors of $E^+ \setminus \{0\}$, resp. $E^- \setminus \{0\}$, have positive, resp. negative, Lyapunov exponents. Then any zero-noise limit measure μ of f is a Sinai–Ruelle–Bowen measure and therefore, if it is ergodic and hyperbolic, a physical measure.*

One can hope, that typically, the latter ergodicity and hyperbolicity assumptions are satisfied (see, e. g., [27]).

By pushing classical techniques and introducing new ideas for generic maps with one expanding and one weakly contracting direction, Tsujii has been able to prove the following generic result (which can be viewed as a 2-dimensional extension of some of the one-dimensional results above: one adds a uniformly expanding direction):

Theorem 6 (Tsujii [243]) *Let M be a compact surface. Consider the space of C^{20} self-maps $f: M \to M$ which admits directions that are uniformly expanded [More precisely, there exists a continuous, forward invariant cone field which is uniformly expanded under the differential of f.]*

Then existence of a finite statistical description is both topologically generic and prevalent in this space of maps.

Hénon-Like Maps and Rank One Attractors

In 1976, Hénon [130] observed that the diffeomorphism of the plane

$$H_{a,b}(x, y) = (1 - ax^2 + y, bx)$$

seemed to present a "strange attractor" for $a = 1.4$ and $b = 0.3$, that is, the points of a numerically simulated orbit seemed to draw a set looking locally like the product of a segment with a Cantor set. This attractor seems to be supported by the unstable manifold $W^u(P)$ of the hyperbolic fixed point P with positive abscis. On the other hand, the Plykin classification [207] excluded the existence of a uniformly hyperbolic attractor for a dissipative surface diffeomorphism.

For almost twenty years the question of the existence of such an attractor (by opposition to an attracting periodic orbit with a very long period) remained open. Indeed, one knew since Newhouse that, for many such maps, there exist infinitely many such periodic orbits which are very difficult to distinguish numerically. But in 1991 Benedicks and Carleson succeeded in proposing an argument refining (with considerable difficulties) their earlier proof of Jakobson one-dimensional theorem and established the first part of the following theorem:

Theorem 7 (Benedicks–Carleson [31]) *For any $\epsilon > 0$, for $|b|$ small enough, there is a set A with $\mathrm{Leb}(A) > 0$ satisfying: for all $a \in A$, there exists $z \in W^u(P)$ such that*

- *The orbit of z is dense in $W^u(P)$;*
- $\liminf_{n\to\infty} \frac{1}{n} \log \|(f^n)'(z)\| > 0$.

Further properties were then established, especially by Benedicks, Viana, Wang, Young [32,34,35,264]. Let us quote the following theorem of Wang and Young which includes the previous results:

Copyrighted material

Theorem 8 [264] *Let* $T_{ab}: S^1 \times [-1,1] \to S^1 \times [-1,1]$ *be such that*

- $T_{a0}(S^1 \times [-1,1]) \subset S^1 \times \{0\}$;
- *For* $b > 0$, T_{ab} *is a diffeomorphism on its image with*

$$c^{-1}b \leq |\det T_{ab}(x,y)| \leq c \cdot b$$

for some $c > 1$ *and all* $(x,y) \in S^1 \times [-1,1]$ *and all* (a,b).

Let $f_a: S^1 \to S^1$ *be the restriction of* T_{a0}. *Assume that* $f = f_0$ *satisfies:*

- *Non-degenerate critical points:* $f'(c) = 0 \Rightarrow f''(c) \neq 0$;
- *Negative Schwarzian derivative: for all* $x \in S^1$ *non-critical,* $f'''(x)/f'(x) - 3/2(f''(x)/f'(x))^2 < 0$;
- *No indifferent or attracting periodic point, i. e., x such that* $f^n(x) = x$ *and* $|(f)'(x)| \leq 1$;
- *Misiurewicz condition:* $d(f^n c, d) > c > 0$ *for all* $n \geq 1$ *and all critical points* c, d.

Assume the following transversality condition on f *at* $a = 0$: *for every critical point* c, $(\mathrm{d}/da)(f_a(c_a) - p_a) \neq 0$ *if* c_a *is the critical point of* f_a *near* c *and* p_a *is the point having the same itinerary under* f_a *as* $f(c)$ *under* c. *Assume the following non-degeneracy of* T: $f_0'(c) = 0 \implies \partial T_{00}(c,0)/\partial y \neq 0$.

- T_{ab} *restricted to a neighborhood of* $S^1 \times \{0\}$ *has a finite statistical description by a number of hyperbolic Sinai–Ruelle–Bowen measures bounded by the number of critical points of* f;
- *There is exponential decay of correlations and a Central Limit Theorem (see below) – except, in an obvious way, if there is a periodic interval with period* > 1;
- *There is a natural coding of the orbits that remains for ever close to* $S^1 \times \{0\}$ *by a closed invariant subset of a full shift.*

Very importantly, the above dynamical situation has been shown to occur near typical homoclinic tangencies: [190] proved that there is an open and dense subset of the set of all C^3 families of diffeomorphisms unfolding a first homoclinic tangency such that the above holds. However [201] shows that the set of parameters with a Henon-like attractor has zero Lebesgue density at the bifurcation itself, at least under an assumption on the so-called stable and unstable Hausdorff dimensions. [95] establishes positive density for another type of bifurcation. Furthermore [191] has related the Hausdorff dimensions to the abundance of uniformly hyperbolic dynamics near the tangency.

[248] is able to treat situations with more than one contracting direction. More recently [266] has proposed a rather general framework, with easily checkable assumptions in order to establish the existence of such dynamics in various applications. See also [122,252] for applications.

Statistical Properties

The ergodic theorem asserts that time averages of integrable functions converge to phase space averages for any ergodic system. The speed of convergence is quite arbitrary in that generality [161] (only upcrossing inequalities seem to be available [38,132]), however many results are available under very natural hypothesis as we are going to explain in this section. The underlying idea is that for sufficiently chaotic dynamics T and reasonably smooth observables ϕ, the time averages

$$A_n\phi(x) := \frac{1}{n}\sum_{k=0}^{n-1} \phi \circ T^k(x)$$

should behave as averages of independent and identically distributed random variables and therefore satisfy the classical limit theorems of probability theory.

The dynamical systems which are amenable to current technology are in a large part [*But other approaches are possible. Let us quote the work* [98] *on partially hyperbolic systems, for instance.*] those that can be reduced to the following type:

Definition 1 Let $T: X \to X$ be a nonsingular map on a probability, metric space $(X, \mathcal{B}, \mu, d)$ with bounded diameter, preserving the probability measure μ. This map is said to be **Gibbs–Markov** if there exists a countable (measurable) partition α of X such that:

1. For all $a \in \alpha$, T is injective on a and $T(a)$ is a union of elements of α.
2. There exists $\lambda > 1$ such that, for all $a \in \alpha$, for all points $x, y \in a$, $d(Tx, Ty) \geq \lambda d(x,y)$.
3. Let Jac be the *inverse* of the Jacobian of T. There exists $C > 0$ such that, for all $a \in \alpha$, for all points $x, y \in a$, $|1 - Jac(x)/Jac(y)| \leq Cd(Tx, Ty)$.
4. The map T has the "big image property": $\inf_{a \in \alpha} \mu(Ta) > 0$.

Some piecewise expanding and C^2 maps are obviously Gibbs–Markov but the real point is that many dynamics can be reduced to that class by the use of inducing and tower constructions as in [262], in particular. This includes possibly piecewise uniformly hyperbolic diffeomorphisms, Collet–Eckmann maps of the interval [21] (typical chaotic maps in the quadratic family), billiards with convex scatterers [262], the stadium billiard [71], Hénon-like mappings [266].

Copyrighted material

diffeomorphisms for which the set of the entropies of ergodic invariant measures is always the interval $[0, h_{\text{top}}(f)]$ as a consequence of [147].

Local Complexity

Recall that the topological entropy can be computed as: $h_{\text{top}}(f) = \lim_{\epsilon\to 0} h_{\text{top}}(f,\epsilon)$ where:

$$h_{\text{top}}(f,\epsilon) := \lim_{n\to\infty} \frac{1}{n} \log s(\delta, n, X)$$

where $s(\delta, n, E)$ is the maximum cardinality of a subset S of E such that:

$$x \neq y \implies \exists 0 \le k < n \; d(f^k x, f^k y) \ge \epsilon$$

(see Bowen's formula of the topological entropy ▶ Entropy in Ergodic Theory). Likewise, the measure-theoretic entropy $h(T,\mu)$ of an ergodic invariant probability measure μ is $\lim_{\epsilon\to 0} h(f,\mu,\epsilon)$ where:

$$h(T,\epsilon) := \lim_{n\to\infty} \frac{1}{n} \log r(\delta, n, \mu)$$

where $r(\delta, n, \mu)$ is the minimum cardinality of $C \subset X$ such that

$$\mu\Big(\big\{x \in X : \exists y \in C \quad \text{such that} \quad \forall 0 \le k < n \; d(f^k x, f^k y) < \epsilon\big\}\Big) > 1/2\,.$$

One can ask *at which scales does entropy arise for a given dynamical system?*, i. e., how the above quantities $h(T,\epsilon)$, $h(T,\mu,\epsilon)$ converge when $\epsilon \to 0$.

An answer is provided by the *local entropy*. [*This quantity was introduced by Misiurewicz* [186] *under the name* conditional topological entropy *and is called* tail entropy *by Downarowicz* [100].] For a continuous map f of a compact metric space X, it is defined as:

$$\begin{aligned} h_{\text{loc}}(f) &:= \lim_{\epsilon\to 0} h_{\text{loc}}(f,\epsilon) \quad \text{with} \\ h_{\text{loc}}(f,\epsilon) &:= \sup_{x\in X} h_{\text{loc}}(f,\epsilon,x) \quad \text{and} \\ h_{\text{loc}}(f,\epsilon,x) &:= \lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log s\Big(\delta, n, \big\{y \in X : \\ &\qquad \forall k \ge 0 \; d(f^k y, f^k x) < \epsilon\big\}\Big) \end{aligned}$$

Clearly from the above formulas:

$$\begin{aligned} &h_{\text{top}}(f) \le h_{\text{top}}(f,\delta) + h_{\text{loc}}(f,\delta) \quad \text{and} \\ &h(f,\mu) \le h(f,\mu,\delta) + h_{\text{loc}}(f,\delta)\,. \end{aligned}$$

Thus the local entropy bounds the defect in uniformity with respect to the measure of the pointwise limit $h(f,\mu) = \lim_{\delta\to 0} h(f,\mu,\delta)$. An exercise in topology shows that the local entropy therefore also bounds the defect in upper semicontinuity of $\mu \mapsto h(f,\mu)$. In fact, by a result of Downarowicz [100] (extended by David Burguet to the non-invertible case), there is a *local variational principle*:

$$\begin{aligned} h_{\text{loc}}(f) &= \lim_{\delta\to 0} \left(\sup_{\mu} h(f,\mu) - h(f,\mu,\delta)\right) \\ &= \sup_{\mu} \limsup_{\nu\to\mu} h(f,\nu) - h(f,\mu) \end{aligned}$$

for any continuous self-map f of a compact metric space.

The local entropy is easily bounded for smooth maps using Yomdin's theory:

Theorem 10 ([64]) *For any C^r map f of a compact manifold, $h_{\text{loc}}(f) \le \frac{d}{r} \log \sup_x \|f'(x)\|$. In particular, $h_{\text{loc}}(f) = 0$ if $r = \infty$.*

Thus C^∞ smoothness implies the existence of a maximum entropy measure (this was proved first by Newhouse) and the existence of *symbolic extension*: a subshift over a finite alphabet $\sigma : \Sigma \to \Sigma$ and a continuous and onto map $\pi : \Sigma \to M$ such that $\pi \circ \sigma = f \circ \sigma$. More precisely,

Theorem 11 (Boyle, Fiebig, Fiebig [56]) *Given a homeomorphism f of a compact metric space X, there exists a principal symbolic extension $\sigma : \Sigma \to \Sigma$, i. e., a symbolic extension such that, for every σ-invariant probability measure ν, $h(\sigma,\nu) = h(f, \nu\circ\pi^{-1})$, if and only if $h_{\text{loc}}(f) = 0$.*

We refer to [55,101] for further results, including a realization theorem showing that the continuity properties of the measured entropy are responsible for the properties of symbolic extensions and also results in finite smoothness.

Global Simplicity

One can marvel at the power of mathematical analysis to analyze such complex evolutions. Of course another way to look at this is to remark that this analysis is possible *once this evolution has been fitted in a simple setting*: one had to move focus away from an individual, unpredictable orbit, of, say, the full tent map to the set of all the orbits of that map, which is essentially the set of all infinite sequences over two symbols: a very simple set indeed corresponding to full *combinatorial freedom* [[253] *describes a weakening of this which holds for all positive entropy symbolic dynamics.*]. The complete description of a given typical orbit requires an infinite amount of information, whereas the set of all orbits has a finite and very tractable definition. The complexity of the individual orbits is seen now as coming from purely random choices inside a simple structure.

Copyrighted material

The classical systems, namely uniformly expanding maps or hyperbolic diffeomorphisms of compact spaces, have a *simple* symbolic dynamics. It is not necessarily a full shift like for the tent map, but it is a subshift of finite type, i. e., a subshift obtained from a full shift by forbidding finitely many finite subwords. What happens outside of the uniform setting?

A fundamental example is provided by *piecewise monotone maps*, i. e., interval maps with finitely many critical points or discontinuities. The partition cut by these points defines a symbolic dynamics. This subshift is usually not of finite type. Indeed, the topological entropy taking arbitrary finite nonnegative values [*For instance, the topological entropy of the* β-transformation, $x \mapsto \beta x \bmod 1$, *is* $\log \beta$ *for* $\beta \geq 1$.], a representation that respects it has to use an uncountable class of models. In particular models defined by finite data, like the subshifts of finite type, cannot be generally adequate. However there are tractable "almost finite representations" in the following senses:

Most symbolic dynamics $\Sigma(T)$ of piecewise monotone maps T can be defined by finitely many infinite sequences, the *kneading invariants* of Milnor and Thurston: $\kappa_0^+, \kappa_1^-, \kappa_1^+, \ldots, \kappa_{d+1}^- \in \{0, \ldots, d\}^{\mathbb{N}}$ if d is the number of critical/discontinuity points. [*The kneading invariants are the suitably defined (left and right) itineraries of the critical/discontinuity points and endpoints.*] Namely,

$$\Sigma(T) = \left\{\alpha \in \{0, \ldots, d\}^{\mathbb{N}} : \forall n \geq 0 \quad \kappa_{\alpha_n}^+ \preceq \sigma^n \alpha \preceq \kappa_{\alpha_n+1}^-\right\}$$

where $\preceq$ is a total order on $\{0, \ldots, d\}^{\mathbb{N}}$ making the coding $x \mapsto \alpha$ non-decreasing. Observe how the kneading invariants determine $\Sigma(T)$ in an effective way: knowing their first n symbols is enough to know the sequences of length n which begin sequences of $\Sigma(T)$. We refer to [91] for the wealth of information that can be extracted from these kneading invariants following Milnor and Thurston [184].

This form of global simplicity can be extended to other classes of non-uniformly expanding maps, including those like Eq. (4) using the notions of subshifts and puzzles of *quasi-finite type* [68,69]. This leads to the notion and analysis of *entropy-expanding maps*, a new open class of non-uniformly expanding maps admitting critical hypersurfaces, defined purely in terms of entropies including the otherwise untractable examples of Eq. (4).

A generalization of the representation of uniform systems by subshifts of finite type is provided by *strongly positive recurrent countable state Markov shifts*, a subclass of Markov shifts (*see glossary.*) which shares many properties with the subshifts of finite type [57,123,124,224,229].

These "simple" systems admit a *classification result* which in particular identifies their measures with entropy close to the maximum [57]. Such a classification generalizes [164]. The "ideology" here is that complexity of individual orbits in a simple setting must come from randomness, but purely random systems are classified by their entropy according to Ornstein [197].

Stability

By definition, chaotic dynamical systems have orbits which are unstable and numerically unpredictable. It is all the more surprising that, once one accepts to consider their dynamics globally, they exhibit very good stability properties.

Structural Stability

A simple form of stability is *structural stability*: a system $f : M \to M$ is structurally C^r-stable if any system g sufficiently C^r-close to f is topologically the same as f, formally: g is topologically conjugate, i. e., there is some homeomorphism [*If* h *were* C^1, *the conjugacy would imply, among other things, that for every p-periodic point:* $\det((f^p)'(x)) = \det((g^p)'(h(p)))$, *a much too strong requirement.*] $h : M \to M$ mapping the orbits of f to those of g, i. e., $g \circ h = h \circ f$.

Andronov and Pontryaguin argued in the 1930's that only such structurally stable systems are physically relevant. Their idea was that the model of a physical system is always known only to some degree of approximation, hence mathematical model whose structure depends on arbitrarily small changes should be irrelevant.

A first question is: *What are these structurally stable systems?* The answer is quite striking:

Theorem 12 (Mañé [182]) *Let* $f : M \to M$ *be a* C^1 *diffeomorphism of a compact manifold.*

f is structurally stable among C^1-diffeomorphisms of M if and only if f is uniformly hyperbolic on its chain recurrent set. [A point x is chain recurrent if, for all $\epsilon > 0$, there exists a finite sequence $x_0, x_1, \ldots, x_n$ such that $x_0 = x_n = x$ and $d(f(x_k), x_{k+1}) < \epsilon$. The chain recurrent set is the set of all chain recurrent points.]

A basic idea in the proof of the theorem is that failure of uniform hyperbolicity gives the opportunity to make an arbitrarily small perturbation contradicting the structural stability. In higher smoothness the required perturbation lemmas (e. g., the closing lemma [14,148,180]) are not available.

We note that uniform hyperbolicity without invertibility does not imply C^1-stability [210].

Copyrighted material

A second question is: *are these stable systems dense?* (So that one could offer structurally stable models for all physical situations). A deep discovery around 1970 is that this is not the case:

Theorem 13 (Abraham–Smale, Simon [3,234]) *For any $r \geq 1$ and any compact manifold M of dimension ≥ 3, the set of uniformly hyperbolic diffeomorphisms is not dense in the space of C^r diffeomorphisms of M. [They use the phenomenon called "heterodimensional homoclinic intersections".]*

Theorem 14 (Newhouse [194]) *For any $r \geq 2$, for any compact manifold M of dimension ≥ 2, the set of uniformly hyperbolic diffeomorphisms is not dense in the space of C^r diffeomorphisms of M. More precisely, there exists a non-empty open subset in this space which contains a dense G_δ subset of diffeomorphisms with infinitely many periodic sinks. [So these diffeomorphisms have no finite statistical description.]*

Observe that it is possible that uniform hyperbolicity could be dense among *surface C^1-diffeomorphisms* (this is the case for C^1 circle maps by a theorem of Jakobson [145]).

In light of Mañé C^1-stability theorem this implies that *structurally stable systems are not dense*, thus one can robustly see behaviors that are topologically modified by arbitrarily small perturbations (at least in the C^1-topology)! So one needs to look beyond these and face that topological properties of relevant dynamical system are not determined from "finite data". It is natural to ask whether the dynamics is almost determined by "sufficient data".

Continuity Properties of the Topological Dynamics

Structural stability asks the topological dynamics to remain *unchanged* by a small perturbation. It is probably at least as interesting to ask it to *change continuously*. This raises the delicate question of which topology should be put on the rather wild set of topological conjugacy classes. It is perhaps more natural to associate to the system a topological invariant taking value in a more manageable set and ask whether the resulting map is continuous.

A first possibility is Zeeman's *Tolerance Stability Conjecture*. He associated to each diffeomorphism the set of all the closures of all of its orbits and he asked whether the resulting map is continuous on a dense G_δ subset of the class of C^r-diffeomorphisms for any $r \geq 0$. This conjecture remains open, we refer to [85] for a discussion and related progress.

A simpler possibility is to consider our favorite topological invariant, the topological entropy, and thus ask whether the dynamical complexity as measured by the entropy is a stable phenomenon. $f \mapsto h_{\text{top}}(f)$ is lower semicontinuous for f among C^0 maps of the interval [*On the set of interval maps with a bounded number of critical points, the entropy is continuous* [188]. *Also $t \mapsto h_{\text{top}}(Q_t)$ is non-decreasing by complex arguments* [91], *though it is a non-smooth function.*] [187] and for f among $C^{1+\epsilon}$-diffeomorphisms of a compact surface [147]. [*It is an important open question whether this actually holds for C^1-diffeomorphisms. It fails for homeomorphisms* [214].] In both cases, one shows the existence of structurally stable invariant uniformly expanding or hyperbolic subsets with topological entropy close to that of the whole dynamics. On the other hand $f \mapsto h_{\text{top}}(f)$ is upper semi-continuous for C^∞ maps [195,254].

Statistical Stability

Statistical stability is the property that *deterministic perturbations* of the dynamical system cause only small changes in the physical measures, usually with respect to the weak star topology on the space of measures. When the physical measure μ_g is uniquely defined for all systems g near f, statistic stability is the continuity of the map $g \mapsto \mu_g$ thus defined.

Statistical stability is known in the uniform setting and also in the piecewise uniform case, provided the expansion is strong enough [25,42,151] (otherwise there are counter-examples, even in the one-dimensional case).

It also holds in some non-uniform settings without critical behavior, in particular for maps with dominated splitting satisfying robustly a separation condition between the positive and negative Lyapunov exponents [247] (see also [7]).

However statistical unstability occurs in dynamics with critical behaviors [18]:

Theorem 15 *Consider the quadratic family $Q_t(x) = tx(1-x)$, $t \in [0,4]$. Let $I \subset [0,4]$ be the full measure subset of $[0,4]$ of good parameters t such that in particular Q_t admits a physical measure (necessarily unique). For $t \in I$, let μ_t be this physical measure.*

Lebesgue almost every $t \in I$ such that μ_t is not carried by a periodic orbit [At such parameters, statistical stability is easily proved.], is a discontinuity point of the map $t \mapsto \mu_t$ [$M([0,1])$ is equipped with the vague topology.]. However, for Lebesgue almost every t, there is a subset $I_t \subset I$ for which t is a Lebesgue density point [t is a Lebesgue density point if for all $r > 0$, the Lebesgue measure $m(I_t \cap [t-r, t+r]) > 0$ and $\lim_{\epsilon \to 0} m(I_t \cap [t-\epsilon, t+\epsilon])/2\epsilon = 1$.] and such that $\mu : I_t \to M([0,1])$ is continuous at t.

Copyrighted material

Stochastic Stability

A physically motivated and a technically easier approach is to study stability of the physical measure under *stochastic perturbations*. For simplicity let us consider a diffeomorphism of a compact subset of $\mathbb{R}^d$ allowing for a direct definition of *additive noise*. Let $\psi(x)\mathrm{d}x$ be an absolutely continuous probability law with compact support. [*Sometimes additional properties are required of the density, e. g., $\psi(x) = \phi(x)1_B$ where 1_B is the characteristic function of the unit ball and $C^{-1} \le \phi(x) \le C$ for some $C < \infty$.*] For $\epsilon > 0$, consider the Markov chain f_ϵ with state space M and transition probabilities:

$$P_\epsilon(x, A) = \int_A \psi\left(\frac{y - f(x)}{\epsilon}\right) \frac{\mathrm{d}y}{\epsilon^d} .$$

The evolution of measures is given by: $(f_\epsilon\mu)(A) = \int_M P_\epsilon(x, A)\, \mathrm{d}\mu$. Under rather weak irreducibility assumptions on f, f_ϵ has a *unique invariant measure* μ_ϵ (contrarily to f) and μ_ϵ is absolutely continuous. When f has a unique physical measure μ, it is said to be *stochastically stable* if $\lim_{\epsilon\to 0} \mu_\epsilon = \mu$ in the appropriate topology (the weak star topology unless otherwise specified).

It turns out that stochastic stability is a rather common property of Sinai–Ruelle–Bowen measures. It holds not only for uniformly expanding maps or hyperbolic diffeomorphisms [153,258], but also for most interval maps [25], for partially hyperbolic systems of the type $E^u \oplus E^{cs}$ or Hénon-like diffeomorphisms [6,13,33,40]. We refer to the monographs [21,41] for more background and results.

Untreated Topics

For reasons of space and time, many important topics have been left out of this article. Let us list some of them.

Other phenomena related to chaotic dynamics have been studied: entrance times [77,79], spectral properties and dynamical zeta functions [21], escape rates [104], dimension [204], differentiability of physical measures with respect to parameters [99,223], entropies and volume or homological growth rates [118,139,254].

As far as the structure of the setting is concerned, one can go beyond maps or diffeomorphisms or flows and study: more general group actions [128]; holomorphic and meromorphic structures [72] and the references therein; symplectic or volume-preserving [90,137] and in particular the Pugh–Shub program around stable ergodicity of partially hyperbolic systems [211]; random iterations [17, 154,155].

A number of important problems have motivated the study of special forms of chaotic dynamics: equidistribution in number theory [105,109,138] and geometry [236]; quantum chaos [10,125]; chaotic control [227]; analysis of algorithms [245].

We have also omitted the important problem of *applying the above results*. Perhaps because of the lack of a general theory, this can often be a challenge (see for instance [244] for the already complex problem of verifying *uniform hyperbolicity* for a singular flow). Liverani has shown how theoretical results can lead to precise and efficient estimates for the toy model of piecewise expanding interval maps [176]. Ergodic theory implies that, in some settings at least, *adding noise may make some estimates more precise* (see [157]). We refer to [106] and the reference therein.

Future Directions

We conclude this article by a (very partial) selection of open problems.

General Theory

In dimension 1, we have seen that the analogue of the Palis conjecture (see above) is established (Theorem 2). However the description of the typical dynamics in Kolmogorov sense is only known in the unimodal, non-degenerate case by Theorem 3. Indeed, results like [63] suggest that the multimodal picture could be more complex.

In higher dimensions, our understanding is much more limited. As far as a general theory is concerned, a deep problem is the paucity of results on the generic dynamics in C^r smoothness with $r > 1$. The remarkable current progress in generic dynamics (culminating in the proof of the weak Palis conjecture, see [84], the references therein and [51] for background) seems restricted to the C^1 topology because of the lack of fundamental tools (e. g., closing lemmas) in higher smoothness. But Pesin theory requires higher smoothness at least technically. This is not only a hard technical issue but generic properties of physical measures, when they have been analyzed are often completely different between the C^1 case and higher smoothness [45].

Physical Measures

In higher dimensions, Benedicks and Carleson analysis of the Hénon map has given rise to a rather general theory of Hénon-like maps and more generally of the dynamical phenomena associated to homoclinic tangencies. However, the proofs are extremely technical. Could they be simplified? Current attempts like [266] center on the introduction of a simpler notion of critical points, possibly a non-inductive one [212].

Copyrighted material

Can this Hénon theory be extended to the weakly dissipative situation? to the conservative situation (for which the *standard map* is a well-known example defying analysis)? In the strongly dissipative setting, what are the typical phenomena on the complement of the Benedicks–Carleson set of parameters?

From a global perspective one of the main questions is the following:

> *Can infinitely many sinks coexist for a large set of parameters in a typical family or is Newhouse phenomenon atypical in the Kolmogorov or prevalent sense?*

This seems rather unlikely (see however [11]).

Away from such "critical dynamics", there are many results about systems with dominated splitting satisfying additional conditions. Can these conditions be weakened so they would be satisfied by typical systems satisfying some natural conditions (like robust transitivity)? For instance:

> *Could one analyze the physical measures of volume-hyperbolic systems?*

A more specific question is whether Tsujii's striking analysis of surface maps with one uniformly expanding direction can be extended to higher dimensions? can one weaken the uniformity of the expansion? The same questions for the corresponding invertible situation is considered in [83].

Maximum Entropy Measures and Topological Complexity

As we explained, C^∞ smoothness, by a Newhouse theorem, ensures the existence of maximum entropy measures, making the situation a little simpler than with respect to physical measures. This existence results allow in particular an easy formulation of the problem of the typicality of hyperbolicity:

> *Are maximum entropy ergodic measures of systems with positive entropy hyperbolic for most systems?*

A more difficult problem is that of the finite multiplicity of the maximum entropy measures. For instance:

> *Do typical systems possess finitely many maximum entropy ergodic measures?*

More specifically, can one prove intrinsic ergodicity (ie, uniqueness of the measure of maximum entropy) for an isolated homoclinic class of some diffeomorphisms (perhaps C^1-generic)? Can a generic C^1-diffeomorphism carry an infinite number of homoclinic classes, each with topological entropy bounded away from zero?

A perhaps more tractable question, given the recent progress in this area: Is a C^1-generic partially hyperbolic diffeomorphisms, perhaps with central dimension 1 or 2, intrinsically ergodic?

We have seen how uniform systems have simple symbolic dynamics, i. e., subshifts of finite type, and how interval maps and more generally entropy-expanding maps keep some of this simplicity, defining subshifts or puzzles of quasi-finite type [68,69]. [264] have defined symbolic dynamics for topological Hénon-like map which seems close to that of that of a one-dimensional system.

> *Can one describe Wang and Young symbolic dynamics of Hénon-like attractors and fit it in a class in which uniqueness of the maximum entropy measure could be proved?*

More generally, can one define nice combinatorial descriptions, for surface diffeomorphisms? Can one formulate variants of the entropy-expansion condition [*For instance building on our "entropy-hyperbolicity".*] of [66,70], that would be satisfied by a large subset of the diffeomorphisms?

Another possible approach is illustrated by the pruning front conjecture of [87] (see also [88,144]). It is an attempt to build a combinatorial description by trying to generalize the way that, for interval maps, kneading invariants determine the symbolic dynamics by considering the bifurcations from a trivial dynamics to an arbitrary one.

We hope that our reader has shared in our fascination with this subject, the many surprising and even paradoxical discoveries that have been made and the exciting current progress, despite the very real difficulties both in the analysis of such non-uniform systems as the Henon map and in the attemps to establish a general (and practical) ergodic theory of chaotic dynamical systems.

Acknowledgments

I am grateful for the advice and/or comments of the following colleagues: P. Collet, J.-R. Chazottes, and especially S. Ruette. I am also indebted to the anonymous referee.

Bibliography

Primary Literature

1. Aaronson J, Denker M (2001) Local limit theorems for partial sums of stationary sequences generated by Gibbs–Markov maps. Stoch Dyn 1:193–237

Copyrighted material

tion for Gibbs measures and flows under a function. Ergod Theory Dyn Syst 4:541–552
93. Detmers MF, Liverani C (2007) Stability of Statistical Properties in Two-dimensional Piecewise Hyperbolic Maps. Trans Amer Math Soc 360:4777–4814
94. Devaney RL (1989) An introduction to chaotic dynamical systems, 2nd ed. Addison–Wesley, Redwood
95. Diaz L, Rocha J, Viana M (1996) Strange attractors in saddle-node cycles: prevalence and globality. Invent Math 125: 37–74
96. Dolgopyat D (2000) On dynamics of mostly contracting diffeomorphisms. Comm Math Phys 213:181–201
97. Dolgopyat D (2002) On mixing properties of compact group extensions of hyperbolic systems. Isr J Math 130:157–205
98. Dolgopyat D (2004) Limit theorems for partially hyperbolic systems. Trans Amer Math Soc 356:1637–1689
99. Dolgopyat D (2004) On differentiability of SRB states for partially hyperbolic systems. Invent Math 155:389–449
100. Downarowicz T (2005) Entropy structure. J Anal Math 96: 57–116
101. Downarowicz T, Newhouse S (2005) Symbolic extensions and smooth dynamical systems. Invent Math 160:453–499
102. Downarowicz T, Serafin J (2003) Possible entropy functions. Isr J Math 135:221–250
103. Duhem P (1906) La théorie physique, son objet et sa structure. Vrin, Paris 1981
104. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617–656
105. Eskin A, McMullen C (1993) Mixing, counting, and equidistribution in Lie groups. Duke Math J 71:181–209
106. Fiedler B (ed) (2001) Ergodic theory, analysis, and efficient simulation of dynamical systems. Springer, Berlin
107. Fiedler B (ed) (2002) Handbook of dynamical systems, vol 2. North-Holland, Amsterdam
108. Fisher A, Lopes AO (2001) Exact bounds for the polynomial decay of correlation, 1/*f* noise and the CLT for the equilibrium state of a non-Hölder potential. Nonlinearity 14:1071–1104
109. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. In: MB Porter Lectures. Princeton University Press, Princeton, NJ
110. Gallavotti G (1999) Statistical mechanics. In: A short treatise. Texts and Monographs in Physics. Springer, Berlin
111. Gatzouras D, Peres Y (1997) Invariant measures of full dimension for some expanding maps. Ergod Theory Dynam Syst 17:147–167
112. Glasner E, Weiss B (1993) Sensitive dependence on initial conditions. Nonlinearity 6:1067–1075
113. Gordin MI (1969) The central limit theorem for stationary processes. (Russian) Dokl Akad Nauk SSSR 188:739–741
114. Gouëzel S (2004) Central limit theorem and stable laws for intermittent maps. Probab Theory Relat Fields 128:82–122
115. Gouëzel S (2005) Berry–Esseen theorem and local limit theorem for non uniformly expanding maps. Ann Inst H Poincaré 41:997–1024
116. Gouëzel S, Liverani C (2006) Banach spaces adapted to Anosov systems. Ergod Theory Dyn Syst 26:189–217
117. Graczyk J, Świątek G (1997) Generic hyperbolicity in the logistic family. Ann of Math 146(2):1–52
118. Gromov M (1987) Entropy, homology and semialgebraic geometry. Séminaire Bourbaki, vol 1985/86. Astérisque No 145–146:225–240
119. Guivarc'h Y, Hardy J (1998) Théorèmes limites pour une classe de chaînes de Markov et applications aux difféomorphismes d'Anosov. Ann Inst H Poincaré 24:73:98
120. Guckenheimer J (1979) Sensitive dependence on initial conditions for one-dimensional maps. Commun Math Phys 70:133–160
121. Guckenheimer J, Holmes P (1990) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Revised and corrected reprint of the 1983 original. In: Applied Mathematical Sciences, 42. Springer, New York
122. Guckenheimer J, Wechselberger M, Young LS (2006) Chaotic attractors of relaxation oscillators. Nonlinearity 19:701–720
123. Gurevic BM (1996) Stably recurrent nonnegative matrices. (Russian) Uspekhi Mat Nauk 51(3)(309):195–196; translation in: Russ Math Surv 51(3):551–552
124. Gurevic BM, Savchenko SV (1998) Thermodynamic formalism for symbolic Markov chains with a countable number of states. (Russian) Uspekhi Mat Nauk 53(2)(320):3–106; translation in: Russ Math Surv 53(2):245–344
125. Gutzwiller M (1990) Chaos in classical and quantum mechanics. In: Interdisciplinary Applied Mathematics, 1. Springer, New York
126. Hadamard J (1898) Les surfaces à courbures opposees et leurs lignes geodesiques. J Math Pures Appl 4:27–73
127. Hasselblatt B, Katok A (2003) A first course in dynamics. In: With a panorama of recent developments. Cambridge University Press, New York
128. Hasselblatt B, Katok A (eds) (2002,2006) Handbook of dynamical systems, vol 1A and 1B. Elsevier, Amsterdam
129. Hennion H (1993) Sur un théorème spectral et son application aux noyaux lipchitziens. Proc Amer Math Soc 118:627–634
130. Hénon M (1976) A two-dimensional mapping with a strange attractor. Comm Math Phys 50:69–77
131. Hesiod (1987) Theogony. Focus/R. Pullins, Newburyport
132. Hochman M (2006) Upcrossing inequalities for stationary sequences and applications. arXiv:math/0608311
133. Hofbauer F (1979) On intrinsic ergodicity of piecewise monotonic transformations with positive entropy. Isr J Math 34(3):213–237
134. Hofbauer F (1985) Periodic points for piecewise monotonic transformations. Ergod Theory Dynam Syst 5:237–256
135. Hofbauer F, Keller G (1982) Ergodic properties of invariant measures for piecewise monotonic transformations. Math Z 180:119–140
136. Hofbauer F, Keller G (1990) Quadratic maps without asymptotic measure. Comm Math Phys 127:319–337
137. Hofer H, Zehnder E (1994) Symplectic invariants and Hamiltonian dynamics. Birkhäuser, Basel
138. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. Ann of Math 161(2):397–488
139. Hua Y, Saghin R, Xia Z (2006) Topological Entropy and Partially Hyperbolic Diffeomorphisms. arXiv:math/0608720
140. Hunt FY (1998) Unique ergodicity and the approximation of attractors and their invariant measures using Ulam's method. (English summary) Nonlinearity 11:307–317
141. Hunt BR, Sauer T, Yorke JA (1992) Prevalence: a translation-invariant "almost every" on infinite-dimensional spaces. Bull Amer Math Soc (NS) 27:217–238
142. Ibragimov I, Linnik Yu, Kingman JFC (ed) (trans) (1971) In-

Copyrighted material

dependent and stationary sequences of random variables. Wolters-Noordhoff, Groningen
143. Ionescu Tulcea CT, Marinescu G (1950) Théorie ergodique pour des classes d'opérations non complètement continues. (French) Ann of Math 52(2):140–147
144. Ishii Y (1997) Towards a kneading theory for Lozi mappings. I. A solution of the pruning front conjecture and the first tangency problem. Nonlinearity 10:731–747
145. Jakobson MV (1981) Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. Comm Math Phys 81:39–88
146. Jenkinson O (2007) Optimization and majorization of invariant measures. Electron Res Announc Amer Math Soc 13:1–12
147. Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. Inst Hautes Etudes Sci Publ Math No 51:137–173
148. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. In: Encyclopedia of Mathematics and its Applications, 54. With a supplementary chapter by: Katok A, Mendoza L. Cambridge University Press, Cambridge
149. Keller G (1984) On the rate of convergence to equilibrium in one-dimensional systems. Comm Math Phys 96:181–193
150. Keller G (1990) Exponents, attractors and Hopf decompositions for interval maps. Ergod Theory Dynam Syst 10:717–744
151. Keller G, Liverani C (1999) Stability of the spectrum for transfer operators. Ann Scuola Norm Sup Pisa Cl Sci 28(4):141–152
152. Keller G, Nowicki T (1992) Spectral theory, zeta functions and the distribution of periodic points for Collet-Eckmann maps. Comm Math Phys 149:31–69
153. Kifer Yu (1977) Small random perturbations of hyperbolic limit sets. (Russian) Uspehi Mat Nauk 32(1)(193):193–194
154. Kifer Y (1986) Ergodic theory of random transformations. In: Progress in Probability and Statistics, 10. Birkhäuser Boston, Inc, Boston, MA
155. Kifer Y (1988) Random perturbations of dynamical systems. In: Progress in Probability and Statistics, 16. Birkhäuser Boston, Inc, Boston, MA
156. Kifer Yu (1990) Large deviations in dynamical systems and stochastic processes. Trans Amer Math Soc 321:505–524
157. Kifer Y (1997) Computations in dynamical systems via random perturbations. (English summary) Discret Contin Dynam Syst 3:457–476
158. Kolyada SF (2004) LI-Yorke sensitivity and other concepts of chaos. Ukr Math J 56:1242–1257
159. Kozlovski OS (2003) Axiom A maps are dense in the space of unimodal maps in the C^k topology. Ann of Math 157(2):1–43
160. Kozlovski O, Shen W, van Strien S (2007) Density of Axiom A in dimension one. Ann Math 166:145–182
161. Krengel U (1983) Ergodic Theorems. De Gruyter, Berlin
162. Krzyżewski K, Szlenk W (1969) On invariant measures for expanding differentiable mappings. Studia Math 33:83–92
163. Lacroix Y (2002) Possible limit laws for entrance times of an ergodic aperiodic dynamical system. Isr J Math 132:253–263
164. Ladler RL, Marcus B (1979) Topological entropy and equivalence of dynamical systems. Mem Amer Math Soc 20(219)
165. Lastoa A, Yorke J (1973) On the existence of invariant measures for piecewise monotonic transformations. Trans Amer Math Soc 186:481–488
166. Ledrappier F (1984) Proprietes ergodiques des mesures de Sinaï. Inst Hautes Etudes Sci Publ Math 59:163–188
167. Ledrappier F, Young LS (1985) The metric entropy of diffeomorphisms. I Characterization of measures satisfying Pesin's entropy formula. II Relations between entropy, exponents and dimension. Ann of Math 122(2):509–539, 540–574
168. Leplaideur R (2004) Existence of SRB-measures for some topologically hyperbolic diffeomorphisms. Ergod Theory Dynam Syst 24:1199–1225
169. Li TY, Yorke JA (1975) Period three implies chaos. Amer Math Monthly 82:985–992
170. Li M, Vitanyi P (1997) An introduction to Kolmogorov complexity and its applications. In: Graduate Texts in Computer Science, 2nd edn. Springer, New York
171. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
172. Liu PD, Qian M, Zhao Y (2003) Large deviations in Axiom A endomorphisms. Proc Roy Soc Edinb Sect A 133:1379–1388
173. Liverani C (1995) Decay of Correlations. Ann Math 142: 239–301
174. Liverani C (1995) Decay of Correlations in Piecewise Expanding maps. J Stat Phys 78:1111–1129
175. Liverani C (1996) Central limit theorem for deterministic systems. In: Ledrappier F, Lewowicz J, Newhouse S (eds) International conference on dynamical systems, Montevideo, 1995. Pitman Research Notes. In: Math 362:56–75
176. Liverani C (2001) Rigorous numerical investigation of the statistical properties of piecewise expanding maps – A feasibility study. Nonlinearity 14:463–490
177. Liverani C, Tsujii M (2006) Zeta functions and dynamical systems. (English summary) Nonlinearity 19:2467–2473
178. Lyubich M (1997) Dynamics of quadratic polynomials. I, II Acta Math 178:185–247, 247–297
179. Margulis G (2004) On some aspects of the theory of Anosov systems. In: With a survey by Richard Sharp: Periodic orbits of hyperbolic flows. Springer Monographs in Mathematics. Springer, Berlin
180. Mañé R (1982) An ergodic closing lemma. Ann of Math 116(2):503–540
181. Mañé R (1985) Hyperbolicity, sinks and measures in one-dimensional dynamics. Commun Math Phys 100:495–524
182. Mañé R (1988) A proof of the C^1 stability conjecture. Publ Math IHES 66:161–210
183. Melbourne I, Nicol M (2005) Almost sure invariance principle for nonuniformly hyperbolic systems. Comm Math Phys 260(1):131–146
184. Milnor J, Thurston W (1988) On iterated maps of the interval. Dynamical systems, College Park, MD, 1986–87, pp 465–563. Lecture Notes in Math, 1342. Springer, Berlin
185. Misiurewicz M (1976) A short proof of the variational principle for a Z_+^n action on a compact space. Bull Acad Polon Sci Sér Sci Math Astronom Phys 24(12)1069–1075
186. Misiurewicz M (1976) Topological conditional entropy. Studia Math 55:175–200
187. Misiurewicz M (1979) Horseshoes for mappings of the interval. Bull Acad Polon Sci Sér Sci Math 27:167–169
188. Misiurewicz M (1995) Continuity of entropy revisited. In: Dynamical systems and applications, pp 495–503. World Sci Ser Appl Anal 4. World Sci Publ, River Edge, NJ
189. Misiurewicz M, Smítal J (1988) Smooth chaotic maps with zero topological entropy. Ergod Theory Dynam Syst 8:421–424
190. Mora L, Viana M (1993) Abundance of strange attractors. Acta Math 171:1–71

Copyrighted material

191. Moreira CG, Palis J, Viana M (2001) Homoclinic tangencies and fractal invariants in arbitrary dimension. CRAS 333:475–480
192. Murray JD (2002,2003) Mathematical biology. I and II An introduction, 3rd edn. In: Interdisciplinary Applied Mathematics, 17 and 18. Springer, New York
193. Nagaev SV (1957) Some limit theorems for stationary Markov chains. Theor Probab Appl 2:378–406
194. Newhouse SE (1974) Diffeomorphisms with infinitely many sinks. Topology 13:9–18
195. Newhouse SE (1989) Continuity properties of entropy. Ann of Math 129(2):215–235; Erratum: Ann of Math 131(2):409–410
196. Nussbaum RD (1970) The radius of the essential spectrum. Duke Math J 37:473–478
197. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
198. Ornstein D, Weiss B (1988) On the Bernoulli nature of systems with some hyperbolic structure. Ergod Theory Dynam Syst 18:441–456
199. Ovid (2005) Metamorphosis. W.W. Norton, New York
200. Palis J (2000) A global view of dynamics and a conjecture on the denseness of tinitude of attractors. Asterisque 261: 335–347
201. Palis J, Yoccoz JC (2001) Fers a cheval non-uniformement hyperboliques engendres par une bifurcation homocline et densite nulle des attracteurs. CRAS 333:867–871
202. Pesin YB (1976) Families of invariant manifolds corresponding to non-zero characteristic exponents. Math USSR Izv 10:1261–1302
203. Pesin YB (1977) Characteristic exponents and smooth ergodic theory. Russ Math Surv 324:55–114
204. Pesin Ya (1997) Dimension theory in dynamical systems. In: Contemporary views and applications. Chicago Lectures in Mathematics. University of Chicago Press, Chicago
205. Pesin Ya, Sinaï Ya (1982) Gibbs measures for partially hyperbolic attractors. Ergod Theory Dynam Syst 2:417–438
206. Piorek J (1985) On the generic chaos in dynamical systems. Univ lagel Acta Math 25:293–298
207. Plykin RV (2002) On the problem of the topological classification of strange attractors of dynamical systems. Uspekhi Mat Nauk 57:123–166. Translation in: Russ Math Surv 57: 1163–1205
208. Poincare H (1892) Les methodes nouvelles de la mecanique céleste. Paris, Gauthier–Villars
209. Pollicott M, Sharp R (2002) Invariance principles for interval maps with an indifferent fixed point. Comm Math Phys 229: 337–346
210. Przytycki F (1977) On *U*-stability and structural stability of endomorphisms satisfying. Axiom A Studia Math 60:61–77
211. Pugh C, Shub M (1999) Ergodic attractors. Trans Amer Math Soc 312:1–54
212. Pujals E, Rodriguez–Hertz F (2007) Critical points for surface diffeomorphisms. J Mod Dyn 1:615–648
213. Puu T (2000) Attractors, bifurcations, and chaos. In: Nonlinear phenomena in economics. Springer, Berlin
214. Rees M (1981) A minimal positive entropy homeomorphism of the 2-torus. J London Math Soc 23(2):537–550
215. Robinson RC (2004) An introduction to dynamical systems: continuous and discrete. Pearson Prentice Hall, Upper Saddle River
216. Rousseau–Egele J (1983) Un théorème de la limite locale pour une classe de transformations dilatantes et monotones par morceaux. Ann Probab 11:772–788
217. Ruelle D (1968) Statistical mechanics of a one-dimensional lattice gas. Comm Math Phys 9:267–278
218. Ruelle D (1976) A measure associated with axiom-A attractors. Amer J Math 98:619–654
219. Ruelle D (1978) An inequality for the entropy of differentiable maps. Bol Soc Brasil Mat 9:83–87
220. Ruelle D (1982) Repellers for real analytic maps. Ergod Theory Dyn Syst 2:99–107
221. Ruelle D (1989) The thermodynamic formalism for expanding maps. Comm Math Phys 125:239–262
222. Ruelle D (2004) Thermodynamic formalism. In: The mathematical structures of equilibrium statistical mechanics, 2nd edn. Cambridge Mathematical Library, Cambridge University Press, Cambridge
223. Ruelle D (2005) Differentiating the absolutely continuous invariant measure of an interval map *f* with respect to *f*. Comm Math Phys 258:445–453
224. Ruette S (2003) On the Vere-Jones classification and existence of maximal measures for countable topological Markov chains. Pacific J Math 209:366–380
225. Ruette S (2003) Chaos on the interval. http://www.math.u-psud.fr/~ruette
226. Saari DG (2005) Collisions, rings, and other Newtonian *N*-body problems. In: CBMS Regional Conference Series in Mathematics, 104. Published for the Conference Board of the Mathematical Sciences, Washington, DC. American Mathematical Society, Providence
227. Saperstone SH, Yorke JA (1971) Controllability of linear oscillatory systems using positive controls. SIAM J Control 9: 253–262
228. Sarig O (1999) Thermodynamic formalism for countable Markov shifts. Ergod Theory Dynam Syst 19:1565–1593
229. Sarig O (2001) Phase Transitions for Countable Topological Markov Shifts. Commun Math Phys 217:555–577
230. Sataev EA (1992) Invariant measures for hyperbolic mappings with singularities. Uspekhi Mat Nauk 47:147–202. Translation in: Russ Math Surv 47:191–251
231. Saussol B (2000) Absolutely continuous invariant measures for multidimensional expanding maps. Isr J Math 116:223–48
232. Saussol B (2006) Recurrence rate in rapidly mixing dynamical systems. Discret Contin Dyn Syst 15(1):259–267
233. Shub M (1987) Global stability of dynamical systems. Springer, New York
234. Simon R (1972) A 3-dimensional Abraham-Smale example. Proc Amer Math Soc 34:629–630
235. Sinai Ya (1972) Gibbs measures in ergodic theory. Uspehi Mat Nauk 27(166):21–64
236. Starkov AN (2000) Dynamical systems on homogeneous spaces. In: Translations of Mathematical Monographs, 190. American Mathematical Society, Providence, RI
237. Stewart P (1964) Jacobellis v Ohio. US Rep 378:184
238. Szász D (ed) (2000) Hard ball systems and the Lorentz gas. Encyclopedia of Mathematical Sciences, 101. In: Mathematical Physics, II. Springer, Berlin
239. Tsujii M (1992) A measure on the space of smooth mappings and dynamical system theory. J Math Soc Jpn 44:415–425
240. Tsujii M (2000) Absolutely continuous invariant measures for piecewise real-analytic expanding maps on the plane. Comm Math Phys 208:605–622

Copyrighted material

241. Tsujii M (2000) Piecewise expanding maps on the plane with singular ergodic properties. Ergod Theory Dynam Syst 20:1851–1857
242. Tsujii M (2001) Absolutely continuous invariant measures for expanding piecewise linear maps. Invent Math 143:349–373
243. Tsujii M (2005) Physical measures for partially hyperbolic surface endomorphisms. Acta Math 194:37–132
244. Tucker W (1999) The Lorenz attractor exists. C R Acad Sci Paris Ser I Math 328:1197–1202
245. Vallée B (2006) Euclidean dynamics. Discret Contin Dyn Syst 15:281–352
246. van Strien S, Vargas E (2004) Real bounds, ergodicity and negative Schwarzian for multimodal maps. J Amer Math Soc 17: 749–782
247. Vasquez CH (2007) Statistical stability for diffeomorphisms with dominated splitting. Ergod Theory Dynam Syst 27: 253–283
248. Viana M (1993) Strange attractors in higher dimensions. Bol Soc Bras Mat (NS) 24:13–62
249. Viana M (1997) Multidimensional nonhyperbolic attractors. Inst Hautes Etudes Sci Publ Math No 85:63–96
250. Viana M (1997) Stochastic dynamics of deterministic systems. In: Lecture Notes 21st Braz Math Colloq IMPA. Rio de Janeiro
251. Viana M (1998) Dynamics: a probabilistic and geometric perspective. In: Proceedings of the International Congress of Mathematicians, vol I, Berlin, 1998. Doc Math Extra I:557–578
252. Wang Q, Young LS (2003) Strange attractors in periodically-kicked limit cycles and Hopf bifurcations. Comm Math Phys 240:509–529
253. Weiss B (2002) Single orbit dynamics. Amer Math Soc, Providence
254. Yomdin Y (1987) Volume growth and entropy. Isr J Math 57:285–300
255. Yoshihara KI (2004) Weakly dependent stochastic sequences and their applications. In: Recent Topics on Weak and Strong Limit Theorems, vol XIV. Sanseido Co Ltd, Chiyoda
256. Young LS (1982) Dimension, entropy and Lyapunov exponents. Ergod Th Dynam Syst 6:311–319
257. Young LS (1985) Bowen–Ruelle measures for certain piecewise hyperbolic maps. Trans Amer Math Soc 287:41–48
258. Young LS (1986) Stochastic stability of hyperbolic attractors. Ergod Theory Dynam Syst 6:311–319
259. Young LS (1990) Large deviations in dynamical systems. Trans Amer Math Soc 318:525–543
260. Young LS (1992) Decay of correlations for certain quadratic maps. Comm Math Phys 146:123–138
261. Young LS (1995) Ergodic theory of differentiable dynamical systems. Real and complex dynamical systems. Hillerad, 1993, pp 293–336. NATO Adv Sci Inst Ser C Math Phys Sci 464. Kluwer, Dordrecht
262. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. Ann Math 585–650
263. Young LS (1999) Recurrence times and rates of mixing. Isr J Math 110:153–188
264. Young LS, Wang D (2001) Strange attractors with one direction of instability. Comm Math Phys 218:1–97
265. Young LS, Wang D (2006) Nonuniformly expanding 1D maps. Commun Math Phys vol 264:255–282
266. Young LS, Wang D (2008) Toward a theory of rank one attractors. Ann Math 167:349–480
267. Zhang Y (1997) Dynamical upper bounds for Hausdorff dimension of invariant sets. Ergod Theory Dynam Syst 17: 739–756

Books and Reviews

Baladi V (2000) The magnet and the butterfly: thermodynamic formalism and the ergodic theory of chaotic dynamics. (English summary). In: Development of mathematics 1950–2000, Birkhäuser, Basel, pp 97–133

Bonatti C (2003) Dynamiques generiques: hyperbolicite et transitivite. In: Seminaire Bourbaki vol 2001/2002. Asterisque No 290:225–242

Kuksin SB (2006) Randomly forced nonlinear PDEs and statistical hydrodynamics in 2 space dimensions. In: Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich

Liu PD, Qian M (1995) Smooth ergodic theory of random dynamical systems. In: Lecture Notes in Mathematics, 1606. Springer, Berlin

Ornstein D, Weiss B (1991) Statistical properties of chaotic systems. With an appendix by David Fried. Bull Amer Math Soc (NS) 24:11–116

Young LS (1999) Ergodic theory of chaotic dynamical systems. XIth International Congress of Mathematical Physics, ICMP '97, Brisbane, Int Press, Cambridge, MA, pp 131–143

Copyrighted material

Chronological Calculus in Systems and Control Theory

MATTHIAS KAWSKI
Department of Mathematics, Arizona State University, Tempe, USA

Article Outline

Glossary
Definition of the Subject
Introduction, History, and Background
Fundamental Notions of the Chronological Calculus
Systems That Are Affine in the Control
Future Directions
Acknowledgments
Bibliography

Glossary

Controllability A control system is controllable if for every pair of points (*states*) p and q there exists an admissible control such that the corresponding solution curve that starts at p ends at q. Local controllability about a point means that all states in some open neighborhood can be reached.

Pontryagin Maximum Principle of optimal control Optimality of a control-trajectory pair geometrically is a property *dual* to local controllability in the sense that an optimal trajectory (endpoint) lies on the boundary of the reachable sets (after possibly augmenting the state of the system by the running cost). The maximum principle is a necessary condition for optimality. Geometrically it is based on analyzing the effect of families of control variations on the endpoint map. The chronological calculus much facilitates this analysis.

$\mathcal{E}(M) = C^\infty(M)$ The algebra of smooth functions on a finite dimensional manifold M, endowed with the topology of uniform convergence of derivatives of all orders on compact sets.

$\Gamma^\infty(M)$ The space of smooth vector fields on the manifold M.

Chronological calculus An approach to systems theory based on a functional analytic operator calculus, that replaces nonlinear objects such as smooth manifolds by infinite dimensional linear ones, by commutative algebras of smooth functions.

Chronological algebra A linear space with a bilinear product $\star$ that satisfies the identity $a \star (b \star c) - b \star (a \star c) = (a \star b) \star c - (b \star a) \star c$. This structure arises naturally via the product $(f \star g)_t = \int_0^t [f_s, g'_s] ds$ of time-varying vector fields f and g in the chronological calculus. Here $[\ ,\ ]$ denotes the Lie bracket.

Zinbiel algebra A linear space with a bilinear product that satisfies the identity $a * (b * c) = (a * b) * c + (b * a) * c$. This structure arises naturally in the special case of affine control systems for the product $(U * V)(t) = \int_0^t U(s)V'(s)ds$ of absolutely continuous scalar valued functions U and V.

The name Zinbiel is Leibniz read backwards, reflecting the duality with Leibniz algebras, a form of noncommutative Lie algebras. There has been some confusion in the literature with Zinbiel algebras incorrectly been called chronological algebras.

$\mathcal{ITF}(\mathcal{U}^Z)$ For a suitable space $\mathcal{U}$ of *time-varying scalars*, e. g. the space of locally absolutely continuous real-valued functions defined on a fixed time interval, and an indexing set Z, $\mathcal{ITF}(\mathcal{U}^Z)$ denotes the space of iterated integral functionals from the space of Z-tuples with values in $\mathcal{U}$ to the space $\mathcal{U}$.

Definition of the Subject

The chronological calculus is a functional analytic operator calculus tool for nonlinear systems theory. The central idea is to replace nonlinear objects by linear ones, in particular, smooth manifolds by commutative algebras of smooth functions. Aside from its elegance, its main virtue is to provide tools for problems that otherwise would effectively be untractable, and to provide new avenues to investigate the underlying geometry. Originally conceived to investigate problems in optimization and control, specifically for extending Pontryagin's Maximum Principle, the chronological calculus continues to establish itself as the preferred language of geometric control theory, and it is spawning new sets of problems, including its own set of algebraic structures that are now studied in their own right.

Introduction, History, and Background

This section starts with a brief historical survey of some landmarks that locate the chronological calculus at the interface of systems and control theory with functional analysis. It is understood that such brief survey cannot possibly do justice to the many contributors. Selected references given are meant to merely serve as starting points for the interested reader.

Many problems in modern systems and control theory are inherently nonlinear and e. g., due to conserved quantities or symmetries, naturally *live* on manifolds rather than on Euclidean spaces. A simple example is the problem of stabilizing the attitude of a satellite via feedback con-

Copyrighted material

the desired properties. For the reverse direction, suppose $\Psi\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$ is an automorphism. Then for every $p \in M$ the map $\hat{p} \circ \Psi\colon \mathcal{E}(M) \mapsto \mathbb{R}$ is a nontrivial linear multiplicative functional, and hence equals $\hat{q}$ for some $q \in M$. It is easy to see that the map $\Phi\colon M \mapsto M$ is indeed a diffeomorphism, and $\Psi = \hat{\Phi}$.

In the sequel we shall omit the *hats*, and simply write, say, p for the linear functional $\hat{p}$. The role of each object is usually clear from the order. For example, for a point p, a smooth function φ, smooth vector fields f, g, h, with flow e^{th} and its tangent map $\left(e^{th}\right)_*$, what in traditional format might be expressed as $\left(\left(e^{th}\right)_* g\right)\varphi(e^{\tau f}(p))$ is simply written as $pe^{\tau f} g e^{th}\varphi$, not requiring any parentheses.

Frechet Space and Convergence

The usual topology on the space $\mathcal{E}(M)$ is the one of uniform convergence of all derivatives on compact sets, i. e., a sequence of functions $\{\varphi_k\}_{k=1}^{\infty} \subseteq \mathcal{E}(M)$ converges to $\phi \in \mathcal{E}(M)$ if for every finite sequence $f_1, f_2, \ldots, f_s$ of smooth vector fields on M and every compact set $K \subseteq M$ the sequence $\{f_s \ldots f_2 f_1 \varphi_k\}_{k=1}^{\infty}$ converges uniformly on K to $f_s \ldots f_2 f_1 \varphi$.

This topology is also obtained by a countable family of semi-norms $\|\cdot\|_{s,K}$ defined by

$$\|\varphi\|_{s,K} = \sup\{\, |p f_s \ldots f_2 f_1 \varphi| : \\ p \in K,\ f_i \in \Gamma^{\infty}(M),\ s \in \mathbb{Z}^+ \,\} \quad (4)$$

where K ranges over a countable collection of compact subsets whose union is all of M. In an analogous way define semi-norms of smooth vector fields $f \in \Gamma^{\infty}(M)$

$$\|f\|_{s,K} = \sup\{\, \|f\varphi\|_{s,K} : \ \|f\varphi\|_{s+1,K} = 1 \,\}. \quad (5)$$

Finally, for every smooth diffeomorphism Φ of M, $s \in \mathbb{Z}^+$ and $K \subseteq M$ compact there exist $C_{s,K,\Phi} \in \mathbb{R}$ such that for all $\varphi \in \mathcal{E}(M)$

$$\|\Phi\varphi\|_{s,K} \ \leq\ C_{s,K,\Phi} \|\varphi\|_{s,\varphi(K)}\,. \quad (6)$$

Regularity properties of one-parameter families of vector fields and diffeomorphisms (*"time-varying vector fields and diffeomorphisms"*) are understood in the weak sense. In particular, for a family of smooth vector fields $f_t \in \Gamma^{\infty}(M)$, its integral and derivative (if they exist) are defined as the operators that satisfy for every $\varphi \in \mathcal{E}(M)$

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} f_t\right)\varphi = \frac{\mathrm{d}}{\mathrm{d}t}\left(f_t\varphi\right) \quad \text{and}$$

$$\left(\int_a^b f_t \mathrm{d}t\right)\varphi = \int_a^b \left(f_t\varphi\right)\mathrm{d}t\,. \quad (7)$$

The convergence of series expansions of vector fields and of diffeomorphisms encountered in the sequel are to be interpreted in the analogous weak sense.

The Chronological Exponential

This section continues to closely follow [5] which contains full details and complete proofs. On a manifold M consider generally time-varying differential equations of the form $\frac{\mathrm{d}}{\mathrm{d}t} q_t = f_t(q_t)$. To assure existence and uniqueness of solutions to initial value problems, make the typical regularity assumptions, namely that in every coordinate chart $U \subseteq M$, $x\colon U \mapsto \mathbb{R}^n$ the vector field $(x_* f_t)$ is (i) measurable and locally bounded with respect to t for every fixed x and (ii) smooth with locally bounded partial derivatives with respect to x for every fixed t. For the purposes of this article and for clarity of exposition, also assume that vector fields are complete. This means that solutions to initial value problems are defined for all times $t \in \mathbb{R}$. This is guaranteed if, for example, all vector fields considered vanish identically outside a common compact subset of M.

As in the previous sections, for each fixed t interpret $q_t\colon \mathcal{E}(M) \mapsto \mathbb{R}$ as a linear functional on $\mathcal{E}(M)$, and note that this family satisfies the time-varying, but linear differential equation, suggestively written as

$$\dot{q}_t = q_t \circ f_t\,. \quad (8)$$

on the space of linear functionals on $\mathcal{E}(M)$. It may be shown that under the above assumptions it has a unique solution, called the *right chronological exponential* of the vector field f_t as the corresponding flow. Formally, it satisfies for almost all $t \in \mathbb{R}$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\overrightarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right) = \left(\overrightarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right) \circ f_t\,. \quad (9)$$

Analogously the *left chronological exponential* satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\overleftarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right) = f_t \circ \left(\overleftarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right). \quad (10)$$

Formally, one obtains a series expansion for the chronological exponentials by rewriting the differential equation as an integral equation and solving it by iteration

$$q_t = q_0 + \int_0^t q(\tau) \circ f_\tau \,\mathrm{d}\tau \quad (11)$$

$$= q_0 + \int_0^t \left(q_0 + \int_0^\tau q(\sigma) \circ f_\sigma \, d\sigma\right) \circ f_\tau \,\mathrm{d}\tau \quad (12)$$

Copyrighted material

to eventually, formally, obtain the expansion

$$\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau \sim \mathrm{Id} + \sum_{k=1}^{\infty}\int_0^t\int_0^{t_k}\cdots\int_0^{t_2} \cdot f_{\tau_k}\circ\cdots\circ f_{\tau_2}\circ f_{\tau_1}\,\mathrm{d}\tau_k\ldots\mathrm{d}\tau_2\,\mathrm{d}\tau_1 \quad (13)$$

and analogously for the left chronological exponential

$$\overleftarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau \sim \mathrm{Id} + \sum_{k=1}^{\infty}\int_0^t\int_0^{t_k}\cdots\int_0^{t_2} \cdot f_{\tau_1}\circ f_{\tau_2}\circ\cdots\circ f_{\tau_k}\,\mathrm{d}\tau_k\ldots\mathrm{d}\tau_2\,\mathrm{d}\tau_1\,. \quad (14)$$

While this series never converges, not even in a weak sense, it nonetheless has an interpretation as an asymptotic expansion. In particular, for any fixed function $\phi \in \mathcal{E}(M)$ and any semi-norm as in the previous section, on any compact set one obtains an error estimate for the remainder after truncating the series at any order N which establishes that the truncation error is of order $O(t^N)$ as $t \to 0$. When one considers the restrictions to any f_t-invariant normed linear subspace $L \subseteq \mathcal{E}(M)$, i. e. for all t, $f_t(L) \subseteq L$, on which f_t is bounded, then the asymptotic series converges to the chronological exponential and it satisfies for every $\varphi \in L$

$$\left\| \left(\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right)\varphi \right\| \le \mathrm{e}^{\int_0^t \|f_\tau\|\,\mathrm{d}\tau}\,\|\varphi\|\,. \quad (15)$$

In the case of analytic vector fields f_τ and analytic functions φ one obtains convergence of the series for sufficiently small t.

While in general there is no reason why the vector fields f_t should commute at different times, the chronological exponentials nonetheless still share some of the usual properties of exponential and flows, for example the composition of chronological exponentials satisfies for all $t_i \in \mathbb{R}$

$$\left(\overrightarrow{\exp}\int_{t_1}^{t_2} f_\tau\,\mathrm{d}\tau\right)\circ\left(\overrightarrow{\exp}\int_{t_2}^{t_3} f_\tau\,\mathrm{d}\tau\right) = \left(\overrightarrow{\exp}\int_{t_1}^{t_3} f_\tau\,\mathrm{d}\tau\right). \quad (16)$$

Moreover, the left and right chronological exponentials are inverses of each other in the sense that for all $t_0, t_1 \in \mathbb{R}$

$$\left(\overrightarrow{\exp}\int_{t_0}^{t_1} f_\tau\,\mathrm{d}\tau\right)^{-1} = \overrightarrow{\exp}\int_{t_1}^{t_0} f_\tau\,\mathrm{d}\tau = \overleftarrow{\exp}\int_{t_0}^{t_1}(-f_\tau)\,\mathrm{d}\tau\,. \quad (17)$$

Variation of Parameters and the Chronological Logarithm

In control one rarely considers only a single vector field, and, instead, for example, is interested in the interaction of a perturbation or control vector field with a reference or drift vector field. In particular, in the nonlinear setting, this is where the chronological calculus very much facilitates the analysis. For simplicity consider a differential equation defined by two, generally time varying, vector fields (with the usual regularity assumptions)

$$\dot{q} = f_t(q) + g_t(q)\,. \quad (18)$$

The objective is to obtain a formula for the flow of the field $(f_t + g_t)$ as a *perturbation* of the flow of f_t. Writing

$$\Phi_t = \overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau \quad \text{and} \quad \Theta_t = \overrightarrow{\exp}\int_0^t (f_\tau+g_\tau)\,\mathrm{d}\tau, \quad (19)$$

this means one is looking for a family of operators Ψ_t such that for all t

$$\Theta_t = \Psi_t\circ\Phi_t\,. \quad (20)$$

Differentiating and using the invertibility of the flows one obtains a differential equation in the standard form

$$\dot{\Psi}(t) = \Psi_t\circ\Phi_t\circ g_t\circ\Phi_t^{-1} = \Psi_t\circ(\mathrm{Ad}\Phi_t)\,g_t = \Psi_t\circ\left(\overrightarrow{\exp}\int_0^t \mathrm{ad}f_\tau\,\mathrm{d}\tau\right)g_t\,, \quad (21)$$

which has the unique solution

$$\Psi_t = \overrightarrow{\exp}\int_0^t\left(\overrightarrow{\exp}\int_0^\tau \mathrm{ad}f_\sigma\,\mathrm{d}\sigma\right)g_\tau\,\mathrm{d}\tau\,, \quad (22)$$

and consequently one has the *variations formula*

$$\overrightarrow{\exp}\int_0^t (f_\tau+g_\tau)\,\mathrm{d}\tau = \left(\overrightarrow{\exp}\int_0^t\left(\overrightarrow{\exp}\int_0^\tau \mathrm{ad}f_\sigma\,\mathrm{d}\sigma\right)\cdot g_\tau\,\mathrm{d}\tau\right)\circ\left(\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right). \quad (23)$$

This formula is of fundamental importance and used ubiquitously for analyzing controllability and optimality: One typically considers f_t as defining a reference dynamical system and then considers the effect of adding a control term g_t. A special application is in the theory of averaging where g_t is considered a perturbation of the field f_t, which in the most simple form may be assumed to be time-periodic. The monograph [81] by Sanders and Verhulst is the

Copyrighted material

classical reference for the theory and applications of averaging using traditional language. The chronological calculus, in particular the above variations formula, have been used successfully by Sarychev [83] to investigate in particular higher order averaging, and by Bullo [13] for averaging in the context of mechanical systems and its interplay with mechanical connections. Instrumental to the work [83] is the notion of the chronological logarithm, which is motivated by rewriting the defining Eqs. (9) and (10) [2,80] as

$$f_t = \left(\overrightarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right)^{-1} \circ \frac{\mathrm{d}}{\mathrm{d}t}\left(\overrightarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right)$$
$$= \frac{\mathrm{d}}{\mathrm{d}t}\left(\overleftarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right) \circ \left(\overleftarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right)^{-1} . \quad (24)$$

While in general the existence of the logarithm is a delicate question [62,83], under suitable hypotheses it is instrumental for the derivation of the nonlinear Floquet theorem in [83].

Theorem 1 (Sarychev [83]) *Suppose $f_t = f_{t+1}$ is a period-one smooth vector field generating the flow Φ_t and $\Lambda = \log \Phi_1$, then there exists a period-one family of diffeomorphisms $\Psi_t = \Psi_{t+1}$ such that for all t $\Phi_t = e^{t\Lambda} \circ \Psi_t$.*

The logarithm of the chronological exponential of a time-varying vector field εf_t may be expanded as a formal chronological series

$$\log \overrightarrow{\exp}\left(\int_0^1 \varepsilon f_\tau \,\mathrm{d}\tau\right) = \sum_{j=1}^{\infty} \varepsilon^j \Lambda^j . \quad (25)$$

The first three terms of which are given explicitly [83] as

$$\Lambda^1 = \int_0^1 f_\tau \,\mathrm{d}\tau\,, \quad \Lambda^2 = \frac{1}{2}\int_0^1 \left[\int_0^\tau f_\sigma \,\mathrm{d}\sigma\,,\, f_\tau\right] \mathrm{d}\tau,$$
$$\text{and} \quad \Lambda^3 = -\frac{1}{2}[\Lambda^1, \Lambda^2] + \frac{1}{3}\int_0^1 \mathrm{ad}^2\left(\int_0^\tau f_\sigma \,\mathrm{d}\sigma\right) f_\tau \,\mathrm{d}\tau\,. \quad (26)$$

Aside from such applications as averaging, the main use of the chronological calculus has been for studying the dual problems of optimality and controllability, in particular controllability of families of diffeomorphisms [4]. In particular, this approach circumvents the limitations encountered by more traditional approaches that use parametrized families of control variations that are based on a finite number of switching times. However, since [52] it is known that a general theory must allow also, e. g., for increasing numbers of switchings and other more general families of control variations. The chronological calculus is instrumental to developing such general theory [4].

Chronological Algebra

When using the chronological calculus for studying controllability, especially when differentiating with respect to a parameter, and also in the aforementioned averaging theory, the following *chronological product* appears almost everywhere: For two time-varying vector fields f_t and g_t that are absolutely continuous with respect to t, define their chronological product as

$$(f \star g)_t = \int_0^t [f_s, g_s']\mathrm{d}s \quad (27)$$

where $[\cdot,\cdot]$ denotes the Lie bracket. It is easily verified that this product is generally neither associative nor satisfies the Jacobi identity, but instead it satisfies the chronological identity

$$x \star (y \star z) - y \star (x \star z) = (x \star y) \star z - (y \star x) \star z\,. \quad (28)$$

One may define an abstract chronological algebra (over a field $\mathbf{k}$) as a linear space that is equipped with a bilinear product that satisfies the chronological identity (28). More compactly, this property may be written in terms of the left translation λ that associates with any element $x \in \mathcal{A}$ of a not-necessarily associative algebra $\mathcal{A}$ the map $\lambda_x\colon y \mapsto xy$. Using this λ, the chronological identity (28) is simply

$$\lambda_{[x,y]} = [\lambda_x, \lambda_y]\,, \quad (29)$$

i. e., the requirement that the map $x \mapsto \lambda_x$ is a homomorphism of the algebra $\mathcal{A}$ with the commutator product $(x, y) \mapsto xy - yx$ into the Lie algebra of linear maps from $\mathcal{A}$ into $\mathcal{A}$ under its commutator product.

Basic general properties of chronological algebras have been studied in [2], including a discussion of free chronological algebras (over a fixed generating set S). Of particular interest are bases of such free chronological algebras as they allow one to minimize the number of terms in series expansions by avoiding redundant terms. Using the graded structure, it is shown in [2] that an ordered basis of a free chronological algebra over a set S may be obtained recursively from products of the form $\lambda_{b_1}\lambda_{b_2}\cdots\lambda_{b_k}s$ where $s \in S$ is a generator and $b_1 \preccurlyeq b_2 \preccurlyeq \cdots \preccurlyeq b_k$ are previously constructed basis elements. According to [2], the first elements of such a basis over the single-element generating set $S = \{s\}$ may be chosen as (using juxtaposition for multiplication)

$$b_1 = s\,,\quad b_2 = \lambda_{b_1}s = s^2\,,\quad b_3 = \lambda_{b_2}s = s^2 s\,,$$
$$b_4 = \lambda_{b_1}\lambda_{b_1}s = ss^2\,,\quad b_5 = \lambda_{b_3}s = (s^2 s)s\,,$$
$$b_6 = \lambda_{b_4}s = (ss^2)s\,,\quad b_7 = \lambda_{b_1}\lambda_{b_2}s = s(s^2 s)\,,$$
$$b_8 = \lambda_{b_2}\lambda_{b_1}s = s(ss^2)\,,\ldots \quad (30)$$

Copyrighted material

Using only this algebraic and combinatorial structure, one may define exponentials (and logarithms) and analyze the group of formal flows [2]. Of particular interest systems and control is an explicit formula that allows one to express products (compositions) of two or more exponentials (flows) as a single exponential. This is tantamount to a formula for the logarithm of a product of exponentials of two noncommuting indeterminates X and Y. The classical version is known as the Campbell–Baker–Hausdorff formula [16], the first few terms being

$$e^X \cdot e^Y = e^{X+Y+\frac{1}{2}[X,Y]+\frac{1}{12}[X-Y,[X,Y]]-\frac{1}{24}[X,[Y,[X,Y]]]-\cdots} . \tag{31}$$

A well-known short-coming of the classical formula is that the higher-order iterated Lie brackets appearing in this formula are always linearly dependent and hence the coefficients are never well-defined. Hence one naturally looks for more compact, minimal formulas which avoid such redundancies. The chronological algebra provides one such alternative and has been developed and utilized in [2]. The interested reader is referred to the original reference for technical details.

Systems That Are Affine in the Control

Separating Time-Varying Control and Geometry

The chronological calculus was originally designed for work with nonstationary (time-varying) families of vector fields, and it arguably reaps the biggest benefits in that setting where there are few other powerful tools available. Nonetheless, the chronological calculus also much streamlines the analysis of autonomous vector fields. The best studied case is that of affine control systems in which the time-varying control can be separated from the geometry determined by autonomous vector fields. In classical notation such control systems are written in the form

$$\dot{x} = u_1(t) f_1(x) + \cdots + u_m(t) f_m(x) . \tag{32}$$

where f_i are vector fields on a manifold M and u is a measurable function of time taking values typically in a compact subset $U \subseteq \mathbb{R}^m$. This description allows one to also accommodate physical systems that have an *uncontrolled drift term* by simply fixing $u_1 \equiv 1$. For the sake of clarity, and for best illustrating the kind of results available, assume that the fields f_i are smooth and complete. Later we specialize further to real analytic vector fields. This set-up does not necessarily require an interpretation of the u_i as controls: they may equally well be disturbances, or it may simply be a dynamical system which splits in this convenient way.

As a starting point consider families of piecewise constant control functions $u\colon [0,T] \mapsto U \subseteq \mathbb{R}^m$. On each interval $[t_j, t_{j+1}]$ on which the control is constant $u_{[t_j,t_{j+1}]} = u^{(j)}$, the right hand side of (32) is a fixed vector field $g_j = u_1^{(j)} f_1 + \ldots u_m^{(j)} f_m$. The endpoint of the solution curve starting from $x(0) = p$ is computed as a directed product of ordinary exponentials (flows) of vector fields

$$p e^{(t_1-t_0)g_1} \circ e^{(t_2-t_1)g_2} \circ \cdots e^{(T-t_{s-1})g_s} . \tag{33}$$

But even in this case of autonomous vector fields the chronological calculus brings substantial simplifications. Basically this means that rather than considering the expression (33) as a point on M, consider this product as a functional on the space $\mathcal{E}(M)$ of smooth functions on M. To reiterate the benefit of this approach [5,57] consider the simple example of a tangent vector to a smooth curve $\gamma\colon (-\varepsilon,\varepsilon) \mapsto M$ which one might want to define as

$$\dot{\gamma}(0) = \lim_{t\to 0} \frac{1}{t}(\gamma(t) - \gamma(0)) . \tag{34}$$

Due to the lack of a linear structure on a general manifold, with a classical interpretation this does not make any sense at all. Nonetheless, when interpreting $\gamma(t)$, $\gamma(0)$, $\gamma'(0)$, and $\frac{1}{t}(\gamma(t)-\gamma(0))$ as linear functionals on the space $\mathcal{E}(M)$ of smooth functions on M this is perfectly meaningful. The meaning of the limit can be rigorously justified as indicated in the preceding sections, compare also [1].

A typical example to illustrate how this formalism almost trivializes important calculations involving Lie brackets is the following, compare [57]. Suppose f and g are smooth vector fields on the manifold M, $p \in M$, and $t \in R$ is sufficiently small in magnitude. Using that $\frac{d}{dt} p e^{tf} = p e^{tf} f$ and so on, one immediately calculates

$$\begin{aligned}\frac{d}{dt}&\left(p e^{tf} e^{tg} e^{-tf} e^{-tg}\right)\\ &= p e^{tf} f e^{tg} e^{-tf} e^{-tg} + p e^{tf} e^{tg} g e^{-tf} e^{-tg}\\ &\quad - p e^{tf} e^{tg} e^{-tf} f e^{-tg} - p e^{tf} e^{tg} e^{-tf} e^{-tg} g .\end{aligned} \tag{35}$$

In particular, at $t = 0$ this expression simplifies to $pf + pg - pf - pg = 0$. Analogously the second derivative is calculated as

$$\begin{aligned}\frac{d^2}{dt^2}&\left(p e^{tf} e^{tg} e^{-tf} e^{-tg}\right)\\ &= p e^{tf} f^2 e^{tg} e^{-tf} e^{-tg} + p e^{tf} f\\ &\cdot e^{tg} g e^{-tf} e^{-tg} - p e^{tf} f e^{tg} e^{-tf} f e^{-tg} - p e^{tf} f e^{tg} e^{-tf}\\ &\cdot e^{-tg} g + p e^{tf} f e^{tg} g e^{-tf} e^{-tg} + p e^{tf} e^{tg} g^2 e^{-tf} e^{-tg}\\ &- p e^{tf} e^{tg} g e^{-tf} f e^{-tg} - p e^{tf} e^{tg} g e^{-tf} e^{-tg} g - p e^{tf} f\end{aligned}$$

Copyrighted material

$$\cdot e^{tg}e^{-tf}fe^{-tg} - pe^{tf}e^{tg}ge^{-tf}fe^{-tg} + pe^{tf}e^{tg}e^{-tf} \cdot f^2e^{-tg} + pe^{tf}e^{tg}e^{-tf}fe^{-tg}g - pe^{tf}fe^{tg}e^{-tf}e^{-tg}g - pe^{tf}e^{tg}ge^{-tf}e^{-tg}g + pe^{tf}e^{tg}e^{-tf}fe^{-tg}g + pe^{tf} \cdot e^{tg}e^{-tf}e^{-tg}g^2 \quad (36)$$

which at $t = 0$ evaluates to

$$pf^2 + pfg - pf^2 - pfg + pfg + pg^2 - pgf - pg^2 - pf^2 - pgf + pf^2 + pfg - pfg - pg^2 + pfg + pg^2 = 2pfg - 2pgf = 2p[f, g] .$$

While this simple calculation may not make sense on a manifold with a classical interpretation in terms of points and vectors on a manifold, it does so in the context of the chronological calculus, and establishes the familiar formula for the Lie bracket as the limit of a composition of flows

$$pe^{tf}e^{tg}e^{-tf}e^{-tg} = p + t^2p[f, g] + O(t^2) \text{ as } t \to 0. \quad (37)$$

Similar examples may be found in [5] (e. g. p. 36), as well as analogous simplifications for the variations formula for autonomous vector fields ([5] p. 43).

Asymptotic Expansions for Affine Systems

Piecewise constant controls are a starting point. But a general theory requires being able to take appropriate limits and demands completeness. Typically, one considers measurable controls taking values in a compact subspace $U \subseteq \mathbb{R}^m$, and uses $L^1([0, t], U)$ as the space of admissible controls. Correspondingly the finite composition of exponentials as in (33) is replaced by continuous formulae. These shall still separate the effects of the time-varying controls from the underlying geometry determined by the autonomous vector fields. This objective is achieved by the Chen–Fliess series which may be interpreted in a number of different ways. Its origins go back to the 1950s when K. T. Chen [18], studying geometric invariants of curves in $\mathbb{R}^n$, associated to each smooth curve a formal noncommutative power series. In the early 1970s, Fliess [30,31] recognized the utility of this series for studying control systems. Using a set $X_1, X_2, \ldots, X_m$ of indeterminates, the Chen–Fliess series of a measurable control $u \in L^1([0, t], U)$ as above is the formal power series

$$S_{\mathrm{CF}}(T, u) = \sum_I \int_0^T u_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} u_{i_2}(t_2) \cdot \int_0^{t_2} u_{i_1}(t_1)\, dt_1 \ldots dt_s \cdot X_{i_1} \ldots X_{i_s} \quad (38)$$

where the sum ranges over all multi-indices $I = (i_1, \ldots, i_s), s \geq 0$ with $i_j \in \{1, 2 \ldots m\}$. This series gives rise to an asymptotic series for solution curves of the system (32). For example, in the analytic setting, following [91], one has:

Theorem 2 *Suppose f_i are analytic vector fields on $\mathbb{R}^n$ $\phi\colon \mathbb{R}^n \mapsto \mathbb{R}$ is analytic and $U \subset \mathbb{R}^m$ is compact. Then for every compact set $K \subseteq \mathbb{R}^n$, there exists $T > 0$ such that the series*

$$S_{CF,f}(T, u, p)(\varphi) = \sum_I \int_0^T u_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} u_{i_2}(t_2) \cdot \int_0^{t_2} u_{i_1}(t_1)\, dt_1 \ldots dt_s \cdot pf_{i_1} \ldots f_{i_s}\varphi \quad (39)$$

converges uniformly to the solution of (32) for initial conditions $x(0) = p \in K$ and $u\colon [0, T] \mapsto U$ measurable.

Here the series $S_{\mathrm{CF},f}(T, u, p)$ is, for every fixed triple (f, u, p), interpreted as an element of the space of linear functionals on $\mathcal{E}(M)$. But one may abstract further. In particular, for each fixed multi-index I as above, the iterated integral coefficient is itself a functional that takes as input a control function $u \in \mathcal{U}^m$ and maps it to the corresponding iterated integral. It is convenient to work with the primitives $U_j\colon t \mapsto \int_0^t u_j(s)\, ds$ of the control functions u_j rather than the controls themselves. More specifically, if e. g. $\mathcal{U} = AC([0, T], [-1, 1])$ then for every multi-index $I = (i_1, \ldots, i_s) \in \{1, \ldots, m\}^s$ as above one obtains the iterated integral functional $\Upsilon_I\colon \mathcal{U}^m \mapsto \mathcal{U}$ defined by

$$\Upsilon_I\colon U \mapsto \int_0^T U'_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} U'_{i_2}(t_2) \cdot \int_0^{t_2} U'_{i_1}(t_1)\, \mathrm{d}t_1 \mathrm{d}t_2 \ldots \mathrm{d}t_s . \quad (40)$$

Denoting by $\mathcal{II}\mathcal{F}(\mathcal{U}^Z)$ the linear space of iterated integral functionals spanned by the functionals of the above form, the map Υ maps multi-indices to $\mathcal{II}\mathcal{F}(\mathcal{U}^Z)$. The algebraic and combinatorial nature of these spaces and this map will be explored in the next section.

On the other side, there is the map $\mathcal{F}$ that maps a multi-index $(i_1, i_2, \ldots, i_s)$ to the partial differential operator $f_{i_1} \ldots f_{i_s}\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$, considered as a linear transformation of $\mathcal{E}(M)$. In the analytic case one may go further, since as a basic principle, the relations between the iterated Lie brackets of the vector fields f_j completely determine the geometry and dynamical behavior [90,91]. Instead of considering actions on $\mathcal{E}(M)$, it suffices to consider the action of these partial differential operators, and

Copyrighted material

(iii) Suppose $u, v, w, (u, v) \in \tilde{\mathcal{H}}$. Then $(t', (t''', t''')) \in \tilde{\mathcal{H}}$ iff $t''' \le t' \le (t''', t''')$ and $t' < (t', (t''', t'''))$.

There are natural mappings $\theta\colon \mathcal{T}(Z) \mapsto L(Z) \subseteq A(Z)$ and $\vartheta\colon \mathcal{T}(Z) \mapsto A(Z)$ (the *foliage map*) that map labeled binary trees to Lie polynomials and to words, and which are defined for $a \in Z$, $(t', t'') \in \mathcal{T}(Z)$ by $\theta(a) = \vartheta(a) = a$ and recursively $\theta((t', t'')) = [\theta(t'), \theta(t'')]$ and $\vartheta((t', t'')) = \vartheta(t')\vartheta(t'')$. The image of a Hall set under the map θ is a basis for the free Lie algebra $L(Z)$. Moreover, the restriction of the map ϑ to a Hall-set is one-to-one which leads to a fundamental unique factorization property of words into products of Hall-words, and a unique way of recovering Hall trees from the Hall words that are their images under ϑ [101]. This latter property is very convenient as it allows one to use Hall words rather than Hall trees for indexing various objects.

The construction of Hall sets, as well as the proof that they give rise to bases of free Lie algebras, rely in an essential way on the process of *Lazard elimination* [9] which is intimately related to the solution of differential Eqs. (32) or (46) by a successive variation of parameters [93]. As a result, using Hall sets it is possible to obtain an extremely simple and elegant recursive formula for the coefficients ξ_h in (51), whereas formulas for the ζ_h are less immediate, although they may be computed by straightforward formulas in terms of natural maps of the Hopf algebra structure on $A(Z)$ [33]. Up to a normalization factor μ in terms of multi-factorials, the formula for the ξ_h is extremely simple in terms of the Zinbiel product on $A(Z)$. For letters $a \in Z$ and Hall words $h, k, hk \in \mathcal{H}$ one has

$$\xi_a = a \quad \text{and} \quad \xi_{hk} = \mu_{hk} \cdot \xi_h * \xi_k \,. \tag{52}$$

Using completely different notation, this formula appeared originally in the work of Schützenberger [84], and was later rediscovered in various settings by Grayson and Grossman [38], Melancon and Reutenauer [69] and Sussmann [93]. In the context of control, it appeared first in [93].

To illustrate the simplicity of the recursive formula, which is based on the factorization of Hall words, consider the following *normal form* for a free nilpotent system (of rank $r = 5$) using a typical Hall set over the alphabet $Z = \{0, 1\}$. This is the closest nonlinear analogue to the Kalman controller normal form of a linear system, which is determined by Kronecker indices that classify the lengths of chains of integrators. For practical computations nilpotent systems are commonly used as approximations of general nonlinear systems – and every nilpotent system (that is, a system of form (32) whose vector fields f_a generate a nilpotent Lie algebra) is the image of a free nilpotent system as below under some projection map. For convenience, the example again uses the notation x_h instead of ξ_h.

$$\begin{aligned}
\dot{x}_0 &= u_0 \\
\dot{x}_1 &= u_1 \\
\dot{x}_{01} &= x_0 \cdot \dot{x}_1 = x_0\, u_1 \\
\dot{x}_{001} &= x_0 \cdot \dot{x}_{01} = x_0^2\, u_1 \\
&\quad \text{using } \vartheta^{-1}(001) = (0(01)) \\
\dot{x}_{101} &= x_1 \cdot \dot{x}_{01} = x_1 x_0\, u_1 \\
&\quad \text{using } \vartheta^{-1}(101) = (1(01)) \\
\dot{x}_{0001} &= x_0 \cdot \dot{x}_{001} = x_0^3\, u_1 \\
&\quad \text{using } \vartheta^{-1}(0001) = (0(0(01))) \\
\dot{x}_{1001} &= x_1 \cdot \dot{x}_{001} = x_1 x_0^2\, u_1 \\
&\quad \text{using } \vartheta^{-1}(1001) = (1(0(01))) \\
\dot{x}_{1101} &= x_1 \cdot \dot{x}_{101} = x_1^2 x_0\, u_1 \\
&\quad \text{using } \vartheta^{-1}(1101) = (1(1(01))) \\
\dot{x}_{00001} &= x_0 \cdot \dot{x}_{001} = x_0^4\, u_1 \\
&\quad \text{using } \vartheta^{-1}(00001) = (0(0(0(01)))) \\
\dot{x}_{10001} &= x_1 \cdot \dot{x}_{0001} = x_1 x_0^3\, u_1 \\
&\quad \text{using } \vartheta^{-1}(10001) = (1(0(0(01)))) \\
\dot{x}_{11001} &= x_1 \cdot \dot{x}_{1001} = x_1^2 x_0^2\, u_1 \\
&\quad \text{using } \vartheta^{-1}(11001) = (1(1(0(01)))) \\
\dot{x}_{01001} &= x_{01} \cdot \dot{x}_{001} = x_{01} x_0^3\, u_1 \\
&\quad \text{using } \vartheta^{-1}(01001) = ((01)(0(01))) \\
\dot{x}_{01101} &= x_{01} \cdot \dot{x}_{101} = x_{01} x_1^2 x_0\, u_1 \\
&\quad \text{using } \vartheta^{-1}(01101) = ((01)(1(01)))\,.
\end{aligned}$$

This example may be interpreted as a system of form (32) with the x_h being a set of coordinate functions on the manifold M, or as the truncation of the lifted system (46) with the x_h being a basis for a finite dimensional subspace of the linear space $\mathcal{E}(M)$. For more details and the technical background see [56,57].

Future Directions

The chronological calculus is still a young methodology. Its intrinsically infinite dimensional character demands a certain sophistication from its users, and thus it may take some time until it reaches its full potential. From a different point of view this also means that there are many opportunities to explore its advantages in ever new areas of applications. A relatively straightforward approach simply checks classical topics of systems and control for whether

Copyrighted material

they are amenable to this approach, and whether this will be superior to classical techniques. The example of [70] which connects the chronological calculus with discrete time systems is a model example for such efforts. There are many further opportunities, some more speculative than others.

Control The chronological calculus appears ideally suited for the investigation of fully nonlinear systems – but in this area there is still much less known than in the case of nonlinear control systems that are affine in the control, or the case of linear systems. Specific conditions for the existence of solutions, controllability, stabilizability, and normal forms are just a few subjects that, while studies have been initiated [1,2,4], deserve further attention.

Computation In recent years much effort has been devoted to understanding the foundations of numerical integration algorithms in nonlinear settings, and to eventually design more efficient algorithms and eventually prove their superiority. The underlying mathematical structures, such as the compositions of noncommuting flows are very similar to those studies in control, as observed in e. g. [23], one of the earlier publications in this area. Some more recent work in this direction is [17,46,72,73,74]. This is also very closely related to the studies of the underlying combinatorial and algebraic structures – arguably [27,28,29,76] are some of the ones most closely related to the subject of this article. There appears to be much potential for a two-way exchange of ideas and results.

Geometry Arguably one of the main thrusts will continue to be the use of the chronological calculus for studying the differential geometric underpinnings of systems and control theory. But it is conceivable that such studies may eventually abstract further – just like after the 1930s – to studies of generalizations of algebras that stem from systems on classical manifolds, much in the spirit of modern noncommutative geometry with [19] being the best-known proponent.

Algebra and combinatorics In general very little is known about possible ideal and subalgebra structures of chronological and Zinbiel algebras. This work was initiated in [2], but relatively little progress has been made since. (Non)existence of finite dimensional (nilpotent) Zinbiel algebras has been established, but only over a complex field [25,26]. Bases of free chronological algebras are discussed in [2], but many other aspects of this algebraic structure remain unexplored. This also intersects with the aforementioned efforts in foundations of computation – on one side there are Hopf-algebraic approaches to free Lie algebras [78] and nonlinear control [33,40,41], while the most recent breakthroughs involve dendriform algebras and related subjects [27,28,29]. This is an open field for uncovering the connections between combinatorial and algebraic objects on one side and geometric and systems objects on the other side. The chronological calculus may well serve as a vehicle to elucidate the correspondences.

Acknowledgments

This work was partially supported by the National Science Foundation through the grant DMS 05-09030

Bibliography

1. Agrachëv A, Gamkrelidze R (1978) Exponential representation of flows and chronological calculus. Math Sbornik USSR (Russian) 107(N4):487–532. Math USSR Sbornik (English translation) 35:727–786
2. Agrachëv A, Gamkrelidze R (1979) Chronological algebras and nonstationary vector fields. J Soviet Math 17:1650–1675
3. Agrachëv A, Gamkrelidze R, Sarychev A (1989) Local invariants of smooth control systems. Acta Appl Math 14:191–237
4. Agrachëv A, Sachkov YU (1993) Local controllability and semigroups of diffeomorphisms. Acta Appl Math 32:1–57
5. Agrachëv A, Sachkov YU (2004) Control Theory from a Geometric Viewpoint. Springer, Berlin
6. Agrachëv A, Sarychev A (2005) Navier Stokes equations: controllability by means of low modes forcing. J Math Fluid Mech 7:108–152
7. Agrachëv A, Vakhrameev S (1983) Chronological series and the Cauchy–Kowalevski theorem. J Math Sci 21:231–250
8. Boltyanski V, Gamkrelidze R, Pontryagin L (1956) On the theory of optimal processes (in Russian). Doklady Akad Nauk SSSR, vol.10, pp 7–10
9. Bourbaki N (1989) Lie Groups and Lie algebras. Springer, Berlin
10. Brockett R (1971) Differential geometric methods in system theory. In: Proc. 11th IEEE Conf. Dec. Cntrl., Berlin, pp 176–180
11. Brockett R (1976) Volterra series and geometric control theory. Autom 12:167–176
12. Bullo F (2001) Series expansions for the evolution of mechanical control systems. SIAM J Control Optim 40:166–190
13. Bullo F (2002) Averaging and vibrational control of mechanical systems. SIAM J Control Optim 41:542–562
14. Bullo F, Lewis A (2005) Geometric control of mechanical systems: Modeling, analysis, and design for simple mechanical control systems. Texts Appl Math 49 IEEE
15. Caiado MI, Sarychev AV () On stability and stabilization of elastic systems by time-variant feedback. ArXiv:math.AP/0507123
16. Campbell J (1897) Proc London Math Soc 28:381–390
17. Casas F, Iserles A (2006) Explicit Magnus expansions for nonlinear equations. J Phys A: Math General 39:5445–5461
18. Chen KT (1957) Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula. Ann Math 65:163–178

Copyrighted material

19. Connes A (1994) Noncommutative geometry. Academic Press, San Diego
20. Cortés J, Martinez S (2003) Motion control algorithms for simple mechanical systems with symmetry. Acta Appl Math 76:221–264
21. Cortés J, Martinez S, Bullo F (2002) On nonlinear controllability and series expansions for Lagrangian systems with dissipative forces. Trans IEEE Aut Control 47:1401–1405
22. Crouch P (1981) Dynamical realizations of finite Volterra series. SIAM J Control Optim 19:177–202
23. Crouch P, Grossman R (1993) Numerical integration of ordinary differential equations on manifolds. J Nonlinear Sci 3:1–33
24. Crouch P, Lamnabhi-Lagarrigue F (1989) Algebraic and multiple integral identities. Acta Appl Math 15:235–274
25. Dzhumadil'daev A (2007) Zinbiel algebras over a q-commutator. J Math Sci 144:3909–3925
26. Dzhumadil'daev A, Tulenbaev K (2005) Nilpotency of Zinbiel algebras. J Dyn Control Syst 11:195–213
27. Ebrahimi-Fard K, Guo L (2007) Rota–Baxter algebras and dendriform algebras. J Pure Appl Algebra 212:320–339
28. Ebrahimi-Fard K, Manchon D, Patras F (2007) A Magnus- and Fer-type formula in dendriform algebras. J Found Comput Math (to appear) http://springerlink.com/content/106038/
29. Ebrahimi-Fard K, Manchon D, Patras F (2008) New identities in dendriform algebras. J Algebr 320:708–727
30. Fliess M (1978) Développements fonctionelles en indéterminées non commutatives des solutions d'équations différentielles non linéaires forcées. CR Acad Sci France Ser A 287:1133–1135
31. Fliess M (1981) Fonctionelles causales nonlinéaires et indeterminées noncommutatives. Bull Soc Math France 109:3–40
32. Gamkrelidze RV, Agrachëv AA, Vakhrameev SA (1991) Ordinary differential equations on vector bundles and chronological calculus. J Sov Math 55:1777–1848
33. Gehrig E (2007) Hopf algebras, projections, and coordinates of the first kind in control theory. Ph D Dissertation, Arizona State University
34. Gelfand I (1938) Abstract functions and linear operators. Math Sbornik NS 4:235–284
35. Gelfand I, Raikov D, Shilov G (1964) Commutative normed rings. (Chelsea) New York (translated from the Russian, with a supplementary chapter), Chelsea Publishing, New York
36. Ginzburg V, Kapranov M (1994) Koszul duality for operads. Duke Math J 76:203–272
37. Gray W, Wang Y (2006) Noncausal fliess operators and their shuffle algebra. In: Proc MTNS 2006 (Mathematical Theory of Networks and Systems). MTNS, Kyoto, pp 2805–2813
38. Grayson M, Grossman R (1990) Models for free nilpotent algebras. J Algebra 135:177–191
39. Grayson M, Larson R (1991) The realization of input-output maps using bialgebras. Forum Math 4:109–121
40. Grossman R, Larson R (1989) Hopf-algebraic structure of combinatorial objects and differential operators. Israel J Math 72:109–117
41. Grossman R, Larson R (1989) Hopf-algebraic structure of families of trees. J Algebra 126:184–210
42. Hall M (1950) A basis for free Lie rings and higher commutators in free groups. Proc Amer Math Soc 1:575–581
43. Haynes G, Hermes H (1970) Nonlinear controllability via Lie theory. SIAM J Control 8:450–460
44. Herman R (1963) On the accessibility problem in control theory. In: Int. Symp. Nonlinear Diff. Eqns. Nonlinear Mechanics. Academic Press, New York, pp 325–332
45. Hermann R, Krener A (1977) Nonlinear controllability and observability. IEEE Trans Aut Control 22:728–740
46. Iserles A, Munthe-Kaas H, Nrsett S, Zanna A (2000) Lie-group methods. Acta numerica 9:215–365
47. Jacob G (1991) Lyndon discretization and exact motion planning. In: Proc. Europ. Control Conf., pp 1507–1512, ECC, Grenoble
48. Jacob G (1992) Motion planning by piecewise constant or polynomial inputs. In: Proc. IFAC NOLCOS. Int Fed Aut, Pergamon Press, Oxford
49. Jakubczyk B (1986) Local realizations of nonlinear causal operators. SIAM J Control Opt 24:231–242
50. Jurdjevic V, Sussmann H (1972) Controllability of nonlinear systems. J Diff Eqns 12:95–116
51. Kalman R (1960) A new approach to linear filtering and prediction problems. Trans ASME – J Basic Eng 82:35–45
52. Kawski M (1988) Control variations with an increasing number of switchings. Bull Amer Math Soc 18:149–152
53. Kawski M (1990) High-order small-time local controllability. In: Sussmann H (ed) Nonlinear Controllability and Optimal Control. Dekker, pp 441–477, New York
54. Kawski M (2000) Calculating the logarithm of the Chen Fliess series. In: Proc. MTNS 2000, CDROM. Perpignan, France
55. Kawski M (2000) Chronological algebras: combinatorics and control. Itogi Nauki i Techniki 68:144–178 (translation in J Math Sci)
56. Kawski M (2002) The combinatorics of nonlinear controllability and noncommuting flows. In: Abdus Salam ICTP Lect Notes 8. pp 223–312, Trieste
57. Kawski M, Sussmann HJ (1997) Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory. In: Helmke U, Prätzel–Wolters D, Zerz E (eds) Operators, Systems, and Linear Algebra. Teubner, pp 111–128 , Stuttgart
58. Kirov N, Krastanov M (2004) Higher order approximations of affinely controlled nonlinear systems. Lect Notes Comp Sci 2907:230–237
59. Kirov N, Krastanov M (2005) Volterra series and numerical approximation of ODEs. Lect Notes Comp Sci 2907:337–344. In: Li Z, Vulkov L, Was'niewski J (eds) Numerical Analysis and Its Applications. Springer, pp 337–344, Berlin
60. Komleva T, Plotnikov A (2000) On the completion of pursuit for a nonautonomous two-person game. Russ Neliniini Kolyvannya 3:469–473
61. Krener A, Lesiak C (1978) The existence and uniqueness of Volterra series for nonlinear systems. IEEE Trans Aut Control 23:1090–1095
62. Kriegl A, Michor P (1997) The convenient setting of global analysis. Math Surv Monogr 53. Amer Math Soc, Providence
63. Lafferiere G, Sussmann H (1991) Motion planning for controllable systems without drift. In: IEEE Conf. Robotics and Automation. pp 1148–1153, IEEE Publications, New York
64. Lafferiere G, Sussmann H (1993) A differential geometric approach to motion planning. In: Li Z, Canny J (eds) Nonholonomic Motion Planning. Kluwer, Boston, pp 235–270
65. Lobry C (1970) Controllabilit'e des systèmes non linéares. SIAM J Control 8:573–605

Copyrighted material

66. Loday JL (1993) Une version non commutative des algèbres de Lie: les algèbres de Leibniz. Enseign Math 39:269–293
67. Loday JL, Pirashvili T (1996) Leibniz representations of Lie algebras. J Algebra 181:414–425
68. Martinez S, Cortes J, Bullo F (2003) Analysis and design of oscillatory control systems. IEEE Trans Aut Control 48:1164–1177
69. Melançon G, Reutenauer C (1989) Lyndon words, free algebras and shuffles. Canadian J Math XLI:577–591
70. Monaco S, Normand-Cyrot D, Califano C (2007) From chronological calculus to exponential representations of continuous and discrete-time dynamics: a lie-algebraic approach. IEEE Trans Aut Control 52:2227–2241
71. Morgansen K, Vela P, Burdick J (2002) Trajectory stabilization for a planar carangiform robot fish. In: Proc. IEEE Conf. Robotics and Automation. pp 756–762, New York
72. Munthe-Kaas H, Owren B (1999) Computations in a free Lie algebra. Royal Soc Lond Philos Trans Ser A 357:957–981
73. Munthe-Kaas H, Wright W (2007) On the Hopf algebraic structure of lie group integrators. J Found Comput Math 8(2):227–257
74. Munthe-Kaas H, Zanna A (1997) Iterated commutators, lie's reduction method and ordinary differential equations on matrix lie groups. In: Cucker F (ed) Found. Computational Math. Springer, Berlin, pp 434–441
75. Murray R, Sastry S (1993) Nonholonomic path planning: steering with sinusoids. IEEE T Autom Control 38:700–716
76. Murua A (2006) The Hopf algebra of rooted trees, free Lie algebras, and Lie series. J Found Comput Math 6:387–426
77. Ree R (1958) Lie elements and an algebra associated with shuffles. Annals Math 68:210–220
78. Reutenauer C (1991) Free Lie Algebras. Oxford University Press, New York
79. Rocha E (2003) On computation of the logarithm of the Chen–Fliess series for nonlinear systems. In: Zinober I, Owens D (eds) Nonlinear and adaptive Control, Lect Notes Control Inf Sci 281:317–326, Sprtinger, Berlin
80. Rocha E (2004) An algebraic approach to nonlinear control theory. Ph D Dissertation, University of Aveiro, Portugal
81. Sanders J, Verhulst F (1985) Averaging methods in nonlinear dynamical systems. Appl Math Sci 59. Springer, New York
82. Sarychev A (2001) Lie- and chronologico-algebraic tools for studying stability of time-varying systems. Syst Control Lett 43:59–76
83. Sarychev A (2001) Stability criteria for time-periodic systems via high-order averaging techniques. In: Lect. Notes Control Inform. Sci. 259. Springer, London, pp 365–377
84. Schützenberger M (1958) Sur une propriété combinatoire des algèbres de Lie libres pouvant être utilisée dans un probléme de mathématiques appliquées. In: Dubreil S (ed) Algèbres et Théorie des Nombres. Faculté des Sciences de Paris vol 12 no 1 (1958-1959), Exposé no 1 pp 1–23
85. Serres U (2006) On the curvature of two-dimensional optimal control systems and zermelos navigation problem. J Math Sci 135:3224–3243
86. Sigalotti M (2005) Local regularity of optimal trajectories for control problems with general boundary conditions. J Dyn Control Syst 11:91–123
87. Sontag E, Wang Y (1992) Generating series and nonlinear systems: analytic aspects, local realizability and i/o representations. Forum Math 4:299–322
88. Stefani G (1985) Polynomial approximations to control systems and local controllability. In: Proc. 25th IEEE Conf. Dec. Cntrl., pp 33–38, New York
89. Stone M (1932) Linear Transformations in Hilbert Space. Amer Math Soc New York
90. Sussmann H (1974) An extension of a theorem of Nagano on transitive Lie algebras. Proc Amer Math Soc 45:349–356
91. Sussmann H (1983) Lie brackets and local controllability: a sufficient condition for scalar-input systems. SIAM J Cntrl Opt 21:686–713
92. Sussmann H (1983) Lie brackets, real analyticity, and geometric control. In: Brockett RW, Millman RS, Sussmann HJ (eds) Differential Geometric Control. pp 1–116, Birkhauser
93. Sussmann H (1986) A product expansion of the Chen series. In: Byrnes C, Lindquist A (eds) Theory and Applications of Nonlinear Control Systems. Elsevier, North-Holland, pp 323–335
94. Sussmann H (1987) A general theorem on local controllability. SIAM J Control Opt 25:158–194
95. Sussmann H (1992) New differential geometric methods in nonholonomic path finding. In: Isidori A, Tarn T (eds) Progr Systems Control Theory 12. Birkhäuser, Boston, pp 365–384
96. Tretyak A (1997) Sufficient conditions for local controllability and high-order necessary conditions for optimality. A differential-geometric approach. J Math Sci 85:1899–2001
97. Tretyak A (1998) Chronological calculus, high-order necessary conditions for optimality, and perturbation methods. J Dyn Control Syst 4:77–126
98. Tretyak A (1998) Higher-order local approximations of smooth control systems and pointwise higher-order optimality conditions. J Math Sci 90:2150–2191
99. Vakhrameev A (1997) A bang-bang theorem with a finite number of switchings for nonlinear smooth control systems. Dynamic systems 4. J Math Sci 85:2002–2016
100. Vela P, Burdick J (2003) Control of biomimetic locomotion via averaging theory. In: Proc. IEEE Conf. Robotics and Automation. pp 1482–1489, IEEE Publications, New York
101. Viennot G (1978) Algèbres de Lie Libres et Monoïdes Libres. Lecture Notes in Mathematics, vol 692. Springer, Berlin
102. Visik M, Kolmogorov A, Fomin S, Shilov G (1964) Israil Moiseevich Gelfand, On his fiftieth birthday. Russ Math Surv 19:163–180
103. Volterra V (1887) Sopra le funzioni che dipendono de altre funzioni. In: Rend. R Academia dei Lincei. pp 97–105, 141–146, 153–158
104. von Neumann J (1932) Mathematische Grundlagen der Quantenmechanik. Grundlehren Math. Wissenschaften 38. Springer, Berlin
105. Zelenko I (2006) On variational approach to differential invariants of rank two distributions. Diff Geom Appl 24:235–259

Copyrighted material

Control of Non-linear Partial Differential Equations

FATIHA ALABAU-BOUSSOUIRA[1],
PIERMARCO CANNARSA[2]
[1] L.M.A.M., Université de Metz, Metz, France
[2] Dipartimento di Matematica, Università di Roma "Tor Vergata", Rome, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Controllability
Stabilization
Optimal Control
Future Directions
Bibliography

Glossary

$\mathbb{R}$ denotes the **real line**, $\mathbb{R}^n$ the n-dimensional Euclidean space, $x \cdot y$ stands for the Euclidean scalar product of $x, y \in \mathbb{R}^n$, and $|x|$ for the norm of x.

State variables quantities describing the state of a system; in this note they will be denoted by u; in the present setting, u will be either a function defined on a subset of $\mathbb{R} \times \mathbb{R}^n$, or a function of time taking its values in an Hilbert space H.

Space domain the subset of $\mathbb{R}^n$ on which state variables are defined.

Partial differential equation a differential equation containing the unknown function as well as its partial derivatives.

State equation a differential equation describing the evolution of the system of interest.

Control function an external action on the state equation aimed at achieving a specific purpose; in this note, control functions they will be denoted by f; f will be used to denote either a function defined on a subset of $\mathbb{R} \times \mathbb{R}^n$, or a function of time taking its values in an Hilbert space F. If the state equation is a partial differential equation of evolution, then a control function can be:

1. *distributed* if it acts on the whole space domain;
2. *locally distributed* if it acts on a subset of the space domain;
3. *boundary* if it acts on the boundary of the space domain;
4. *optimal* if it minimizes (together with the corresponding trajectory) a given cost;
5. *feedback* if it depends, in turn, on the state of the system.

Trajectory the solution of the state equation u_f that corresponds to a given control function f.

Distributed parameter system a system modeled by an evolution equation on an infinite dimensional space, such as a partial differential equation or a partial integro-differential equation, or a delay equation; unlike systems described by finitely many state variables, such as the ones modeled by ordinary differential equations, the information concerning these systems is "distributed" among infinitely many parameters.

$\mathbb{1}_A$ denotes the **characteristic function** of a set $A \subset \mathbb{R}^n$, that is,

$$\mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \in \mathbb{R}^n \setminus A \end{cases}$$

∂_t, ∂_{x_i} denote **partial derivatives** with respect to t and x_i, respectively.

$L^2(\Omega)$ denotes the **Lebesgue space** of all real-valued square integrable functions, where functions that differ on sets of zero Lebesgue measure are identified.

$H_0^1(\Omega)$ denotes the **Sobolev space** of all real-valued functions which are square integrable together with their *first order* partial derivatives in the sense of distributions in Ω, and vanish on the boundary of Ω; similarly $H^2(\Omega)$ denotes the space of all functions which are square integrable together with their *second order* partial derivatives.

$H^{-1}(\Omega)$ denotes the dual of $H_0^1(\Omega)$.

$\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional **Hausdorff measure**.

H denotes a **normed space**s over $\mathbb{R}$ with norm $\|\cdot\|$, as well as an **Hilbert space** with the scalar product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$.

$L^2(0, T; H)$ is the space of all square integrable functions $f\colon [0, T] \to H$; $C([0, T]; H)$ (continuous functions) and $H^1(0, T; H)$ (Sobolev functions) are similarly defined .

Given Hilbert spaces F and H, $\mathcal{L}(F, H)$ denotes the (Banach) space of all bounded linear operators $\Lambda\colon F \to H$ with norm $\|\Lambda\| = \sup_{\|x\|=1} \|\Lambda x\|$ (when $F = H$, we use the abbreviated notation $\mathcal{L}(H)$); $\Lambda^*\colon H \to F$ denotes the adjoint of Λ given by $\langle \Lambda^* u, \phi \rangle = \langle u, \Lambda \phi \rangle$ for all $u \in H, \phi \in F$.

Definition of the Subject

Control theory (abbreviated, CT) is concerned with several ways of influencing the evolution of a given system by an external action. As such, it originated in the nine-

Copyrighted material

has a unique *mild solution* $u_f \in C([0,T];H)$ given by

$$u_f(t) = \mathrm{e}^{tA}u_0 + \int_0^t \mathrm{e}^{(t-s)A}Bf(s) \quad \forall t \geq 0 \tag{13}$$

Note 1 Boundary control problems can be reduced to the same abstract form as above. In this case, however, B in (12) turns out to be an unbounded operator related to suitable fractional powers of $-A$, see, e. g., [22].

For any $t \geq 0$ let us denote by $\Lambda_t \colon L^2(0,t;F) \to H$ the bounded linear operator

$$\Lambda_t f = \int_0^t \mathrm{e}^{(t-s)A}Bf(s)\,\mathrm{d}s \quad \forall f \in L^2(0,t;F)\,. \tag{14}$$

The *attainable* (or *reachable*) set from u_0 at time t, $\mathcal{A}(u_0,t)$ is the set of all points in H of the form $u_f(t)$ for some control function f, that is

$$\mathcal{A}(u_0,t) \doteq \mathrm{e}^{tA}u_0 + \Lambda_t L^2(0,t;F)\,.$$

We introduce below the main notions of controllability for (7). Let $T > 0$.

Definition 1 System (7) is said to be:

- **exactly controllable** in time T if $\mathcal{A}(u_0,T) = H$ for all $u_0 \in H$, that is, if for all $u_0, u_1 \in H$ there is a control function $f \in L^2(0,T;F)$ such that $u_f(T) = u_1$;
- **null controllable** in time T if $0 \in \mathcal{A}(u_0,T)$ for all $u_0 \in H$, that is, if for all $u_0 \in H$ there is a control function $f \in L^2(0,T;F)$ such that $u_f(T) = 0$;
- **approximately controllable** in time T if $\mathcal{A}(u_0,T)$ is dense in H for all $u_0 \in H$, that is, if for all $u_0, u_1 \in H$ and for any $\varepsilon > 0$ there is a control function $f \in L^2(0,T;F)$ such that $\|u_f(T) - u_1\| < \varepsilon$.

Clearly, if a system is exactly controllable in time T, then it is also null and approximately controllable in time T. Although these last two notions of controllability are strictly weaker than strong controllability, for specific problems—like when A generates a strongly continuous group—some of them may coincide.

Since controllability properties concern, ultimately, the range of the linear operator Λ_T defined in (14), it is not surprising that they can be characterized in terms of the adjoint operator $\Lambda_T^* \colon H \to L^2(0,T;F)$, which is defined by

$$\int_0^T \langle \Lambda_T^* u(s), f(s)\rangle \mathrm{d}s = \langle u_0, \Lambda_T f\rangle \qquad \forall u \in H\,,\ \forall f \in L^2(0,T;F)\,.$$

Such a characterization is the object of the following theorem. Notice that the above identity and (14) yield

$$\Lambda_T^* u(s) = B^* \mathrm{e}^{(T-s)A^*}u \quad \forall s \in [0,T]\,.$$

Theorem 1 *System (7) is:*

- ***exactly controllable*** *in time T if and only if there is a constant $C > 0$ such that*

$$\int_0^T \left\| B^* e^{tA^*}u \right\|^2 dt \geq C\|u\|^2 \quad \forall u \in H; \tag{15}$$

- ***null controllable*** *in time T if and only if there is a constant $C > 0$ such that*

$$\int_0^T \left\| B^* e^{tA^*}u \right\|^2 dt \geq C\left\| e^{TA^*}u \right\|^2 \quad \forall u \in H; \tag{16}$$

- ***approximately controllable*** *in time T if and only if, for every $u \in H$,*

$$B^* e^{tA^*}u = 0 \quad t \in [0,T] \ \ a.e. \implies u = 0\,. \tag{17}$$

To benefit the reader who is more familiar with optimization theory than abstract functional analysis, let us explain, by a variational argument, why estimate (16) implies null controllability. Consider, for every $\varepsilon > 0$, the penalized problem

$$\min\left\{J_\varepsilon(f) \,:\, f \in L^2(0,T;H)\right\},$$

where

$$J_\varepsilon(f) = \frac{1}{2}\int_0^T \|f(t)\|^2\,\mathrm{d}t + \frac{1}{2\varepsilon}\|u_f(T)\|^2 \qquad \forall f \in L^2(0,T;H)\,.$$

Since J_ε is strictly convex, it admits a unique minimum point f_ε. Set $u_\varepsilon = u_{f_\varepsilon}$. Recalling (13) we have, By Fermat's rule,

$$0 = J_\varepsilon'(f_\varepsilon)g = \int_0^T \langle f_\varepsilon(t), g(t)\rangle\,\mathrm{d}t + \frac{1}{\varepsilon}\langle u_\varepsilon(T), \Lambda_T g\rangle \quad \forall g \in L^2(0,T;H)\,. \tag{18}$$

Therefore, passing to the adjoint of Λ_T,

$$\int_0^T \left\langle f_\varepsilon(t) + \frac{1}{\varepsilon}\left(\Lambda_T^* u_\varepsilon(T)\right)(t), g(t)\right\rangle \mathrm{d}t = 0 \qquad \forall g \in L^2(0,T;H)\,,$$

Copyrighted material

whence, owing to (14),

$$f_\varepsilon(t) = -\frac{1}{\varepsilon}\big(\Lambda_T^* u_\varepsilon(T)\big)(t) = -B^* v_\varepsilon(t) \qquad \forall\, t \in [0,T]\,, \quad (19)$$

where $v_\varepsilon(t) \doteq \frac{1}{\varepsilon} e^{(T-t)A^*} u_\varepsilon(T)$ is the solution of the *dual problem*

$$\begin{cases} v' + A^* v = 0 & t \in [0,T] \\ v(T) = \frac{1}{\varepsilon} u_\varepsilon(T)\,. \end{cases}$$

It turns out that

$$\frac{1}{2}\int_0^T \|f_\varepsilon(t)\|^2\,dt + \frac{1}{\varepsilon}\|u_\varepsilon(T)\|^2 \le C\|u_0\|^2 \qquad \forall\, \varepsilon > 0 \quad (20)$$

for some positive constant C. Indeed, observe that, in view of (19),

$$\begin{cases} \langle u_\varepsilon' - Au_\varepsilon + BB^* v_\varepsilon\,,\, v_\varepsilon\rangle = 0\,, & u_\varepsilon(0) = u_0 \\ \langle v_\varepsilon' + A^* v_\varepsilon\,,\, u_\varepsilon\rangle = 0\,, & v_\varepsilon(T) = \frac{1}{\varepsilon} u_\varepsilon(T)\,. \end{cases}$$

So,

$$\int_0^T \left[\frac{\mathrm{d}}{\mathrm{d}t}\langle u_\varepsilon\,,\, v_\varepsilon\rangle + \left\|B^* v_\varepsilon\right\|^2\right] dt = 0\,,$$

hence

$$\frac{1}{\varepsilon}\|u_\varepsilon(T)\|^2 + \int_0^T \left\|B^* v_\varepsilon\right\|^2 \mathrm{d}t = \langle u_0\,,\, v_\varepsilon(0)\rangle\,. \quad (21)$$

Now, apply estimate (16) with $u = \frac{u_\varepsilon(T)}{\varepsilon}$ and note that $v_\varepsilon(T-t) = e^{tA^*}\frac{u_\varepsilon(T)}{\varepsilon}$ to obtain

$$\int_0^T \left\|B^* v_\varepsilon(t)\right\|^2 \mathrm{d}t \ge C\,\|v_\varepsilon(0)\|^2$$

for some positive constant C. Hence, (20) follows from (21) and (19).

Finally, from (20) one deduces the existence of a weakly convergent subsequence f_{ε_j} in $L^2(0,T;F)$. Then, called f_0 the weak limit of f_{ε_j}, $u_{\varepsilon_j}(t) \to u_{f_0}(t)$ for all $t \in [0,T]$. So, owing to (20), $u_{f_0}(T) = 0$.

Heat Equation

It is not hard to see that the heat equation (9) with Dirichlet boundary conditions (2) fails to be exactly controllable. On the other hand, one can show that it is null controllable in any time $T > 0$, hence approximately controllable. Let ω be an open subset of Ω such that $\overline{\omega} \subset \Omega$.

Taking

$$H = L^2(\Omega) = F\,, \quad Bf = \mathbb{1}_\omega f \quad \forall f \in L^2(\Omega)$$

and A as in (8), one obtains that, for any $u_0 \in L^2(\Omega)$ and $f \in L^2(Q_T)$, the initial-boundary value problem

$$\begin{cases} \partial_t u = \Delta u + \mathbb{1}_\omega f & \text{in } Q_T \\ u = 0 & \text{on } \Sigma_T \\ u(0,x) = u_0(x) & x \in \Omega \end{cases} \quad (22)$$

has a unique mild solution $u_f \in C([0,T];L^2(\Omega))$. Moreover, multiplying both sides of equation (9) by u and integrating by parts, it is easy to see that

$$\partial_{x_i} u \in L^2(Q_T) \quad \forall\, i = 1,\dots,n\,. \quad (23)$$

Notice that the above property already suffices to explain why the heat equation cannot be exactly controllable: it is impossible to attain a state $u_1 \in L^2(\Omega)$ which is not compatible with (23).

On the other hand, null controllability holds true in any positive time.

Theorem 2 *Let $T > 0$ and let ω be an open subset of Ω such that $\overline{\omega} \subset \Omega$. Then the heat equation (9) with homogeneous Dirichlet boundary conditions is null controllable in time T, i. e., for every initial condition $u_0 \in L^2(\Omega)$ there is a control function $f \in L^2(Q_T)$ such that the solution u_f of (22) satisfies $u_f(T,\cdot) \equiv 0$. Moreover,*

$$\iint_{Q_T} |f|^2 dx dt \le C_T \int_\Omega |u_0|^2 dx$$

for some positive constant C_T.

The above property is a consequence of the abstract result in Theorem 1 and of concrete estimates for solutions of parabolic equations. Indeed, in order to apply Theorem 1 one has to translate (16) into an estimate for the heat operator. Now, observing that both A and B are self-adjoint, one promptly realizes that (16) reduces to

$$\int_0^T \int_\omega |v(t,x)|^2 \mathrm{d}x\mathrm{d}t \ge C \int_\Omega |v(T,x)|^2 \mathrm{d}x \quad (24)$$

for every solution v of the problem

$$\begin{cases} \partial_t v = \Delta v & \text{in } Q_T \\ v = 0 & \text{on } \Sigma_T\,. \end{cases} \quad (25)$$

Estimate (24) is called an *observability inequality* for the heat operator for obvious reasons: problem (25) is not

Copyrighted material

well-posed since the initial condition is missing. Nevertheless, if, "observing" a solution v of such a problem on the "small" cylinder $(0, T) \times \omega$, you find that it vanishes, then you can conclude that $v(T, \cdot) \equiv 0$ in the whole domain Ω. Thus, $v(0, \cdot) \equiv 0$ by backward uniqueness.

In conclusion, as elegant as the abstract approach to null controllability may be, one is confronted by the difficult task of proving observability estimates. In fact, for the heat operator there are several ways to prove inequality (24). One of the most powerful, basically due to Fursikov and Imanuvilov [65], relies on global *Carleman estimates*. Compared to other methods that can be used to derive observability, such a technique has the advantage of applying to second order parabolic operators with variable coefficients, as well as to more general operators.

Global Carleman estimates are a priori estimates in weighted norms for solutions of the problem

$$\begin{cases} \partial_t v = \Delta v + f & \text{in } Q_T \\ v = 0 & \text{on } \Sigma_T . \end{cases} \tag{26}$$

regardless of initial conditions. The weight function is usually of the form

$$\psi_r(t, x) \doteq \theta(t)\big(e^{2r\|\phi\|_{\infty,\Omega}} - e^{r\phi(x)}\big) \quad (t, x) \in Q_T , \tag{27}$$

where r is a positive constant, ϕ is a given function in $C^2(\overline{\Omega})$ such that

$$\nabla\phi(x) \neq 0 \quad \forall x \in \overline{\Omega} , \tag{28}$$

and

$$\theta(t) \doteq \frac{1}{t(T-t)} \quad 0 < t < T .$$

Note that

$$\theta > 0 , \quad \theta(t) \to \infty \quad t \to 0 , T$$

$$\psi_r > 0 , \quad \psi_r(t, x) \to \infty \quad t \downarrow 0 , t \uparrow T .$$

Using the above notations, a typical global Carleman estimate for the heat operator is the following result obtained in [65]. Let us denote by $\nu(x)$ the outword unit normal to Γ at a point $x \in \Gamma$, and by

$$\frac{\partial\phi}{\partial\nu}(x) = \nabla\phi(x) \cdot \nu(x)$$

the normal derivative of ϕ at x.

Theorem 3 *Let Ω be a bounded domain of $\mathbb{R}^n$ with boundary of class C^2, let $f \in L^2(Q_T)$, and let ϕ be a function satisfying (28). Let v be a solution of (26). Then there are positive constants r, s_0 and C such that, for any $s > s_0$,*

$$\begin{aligned} s^3 \iint_{Q_T} \theta^3(t)|v(t,x)|^2 e^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \\ \leq C \iint_{Q_T} |f(t,x)|^2 e^{-2s\psi_r} \mathrm{d}x\mathrm{d}t + Cs \int_0^T \theta(t)\mathrm{d}t \\ \times \int_\Gamma \frac{\partial\phi}{\partial\nu}(x) \left|\frac{\partial v}{\partial\nu}(t,x)\right|^2 e^{-2s\psi_r} \mathrm{d}\mathcal{H}^{n-1}(x) \end{aligned} \tag{29}$$

It is worth underlying that, thanks to the singular behavior of θ near 0 and T, the above result is independent of the initial value of v. Therefore, it can be applied, indifferently, to any solution of (26) as well as to any solution of the *backward problem*

$$\begin{cases} \partial_t v + \Delta v = f & \text{in } Q_T \\ v = 0 & \text{on } \Sigma_T . \end{cases}$$

Moreover, inequality (29) can be completed adding first and second order terms to its left-hand side, each with its own adapted power of s and θ.

Instead of trying to sketch the proof of Theorem 3, which would go beyond the scopes of this note, it is interesting to explain how it can be used to recover the observability inequality (24), which is what is needed to show that the heat equation is null controllable. The reasoning—not completely straightforward—is based on the following topological lemma, proved in [65].

Lemma 1 *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with boundary Γ of class C^k, for some $k \geq 2$, and let $\omega \subset \Omega$ be an open set such that $\overline{\omega} \subset \Omega$.*

Then there is function $\phi \in C^k(\overline{\Omega})$ such that

$$\begin{cases} (i) & \phi(x) = 0 \quad \text{and} \quad \frac{\partial\phi}{\partial\nu}(x) < 0 \quad \forall x \in \Gamma \\ (ii) & \{x \in \Omega | \nabla\phi(x) = 0\} \subset \omega . \end{cases} \tag{30}$$

Now, given a solution v of (25) and an open set ω such that $\overline{\omega} \subset \Omega$, let $\omega' \subset\subset \omega'' \subset\subset \omega$ be subdomains with smooth boundary. Then the above lemma ensures the existence of a function ϕ such that

$$\{x \in \Omega | \nabla\phi(x) = 0\} \subset \omega' .$$

"Localizing" problem (25) onto $\Omega' \doteq \Omega \setminus \omega'$ by a cutoff function $\eta \in C^\infty(\mathbb{R}^n)$ such that

$$0 \leq \eta \leq 1 , \quad \eta \equiv 1 \quad \text{on } \mathbb{R}^n \setminus \omega'' , \quad \eta \equiv 0 \quad \text{on } \omega' ,$$

that is, taking $w = \eta v$, gives

$$\begin{cases} \partial_t w = \Delta w + h & \text{in } Q'_T \doteq (0, T) \times \Omega' \\ w(t, \cdot) = 0 & \text{on } \partial\Omega' = \partial\Omega \cup \partial\omega' , \end{cases} \tag{31}$$

Copyrighted material

with $h := -\nu\Delta\eta + 2\nabla\eta\cdot\nabla u$. Since $\nabla\phi \neq 0$ on Ω', Theorem 3 can be applied to w on Q_T' to obtain

$$
\begin{aligned}
&s^3\iint_{Q_T'}\theta^3|w|^2\mathrm{e}^{-2s\psi_r}\mathrm{d}x\mathrm{d}t\\
&\leq C\iint_{Q_T'}|h|^2\mathrm{e}^{-2s\psi_r}\mathrm{d}x\mathrm{d}t\\
&+ Cs\int_0^T\theta\mathrm{d}t\int_\Gamma\frac{\partial\phi}{\partial\nu}\left|\frac{\partial w}{\partial\nu}\right|^2\mathrm{e}^{-2s\psi_r}d\mathcal{H}^{n-1}\\
&+ Cs\int_0^T\theta\mathrm{d}t\int_{\partial\omega'}\frac{\partial\phi}{\partial\nu}\left|\frac{\partial w}{\partial\nu}\right|^2\mathrm{e}^{-2s\psi_r}d\mathcal{H}^{n-1}\\
&\leq C\iint_{Q_T'}|h|^2\mathrm{e}^{-2s\psi_r}\mathrm{d}x\mathrm{d}t
\end{aligned}
$$

for s sufficiently large. On the other hand, for any $0 < T_0 < T_1 < T$,

$$
\begin{aligned}
&s^3\iint_{Q_T'}\theta^3|w|^2\mathrm{e}^{-2s\psi_r}\mathrm{d}x\mathrm{d}t\\
&\geq s^3\int_{T_0}^{T_1}\mathrm{d}t\int_{\Omega\setminus\omega}\theta^3|w|^2\mathrm{e}^{-2s\psi_r}\mathrm{d}x\mathrm{d}t\\
&\geq\int_{T_0}^{T_1}\mathrm{d}t\int_{\Omega\setminus\omega}|v|^2\mathrm{d}x
\end{aligned}
$$

Therefore, recalling the definition of h,

$$
\begin{aligned}
&\int_{T_0}^{T_1}\mathrm{d}t\int_{\Omega\setminus\omega}|v|^2\mathrm{d}x\leq C\iint_{Q_T'}|h|^2\mathrm{e}^{-2s\psi_r}\mathrm{d}x\mathrm{d}t\\
&\leq C\int_0^T\mathrm{d}t\int_{\omega''\setminus\omega'}\left[|\nabla^2\eta|^2\nu^2+|\nabla\eta|^2|\nabla v|^2\right]\mathrm{e}^{-2s\psi_r}\mathrm{d}x\\
&\leq C\int_0^T\mathrm{d}t\int_\omega|v|^2\mathrm{d}x+C\int_0^T\mathrm{d}t\\
&\qquad\times\int_{\omega''\setminus\omega'}|\nabla v|^2\mathrm{e}^{-2s\psi_r}\mathrm{d}x\,.
\end{aligned}
$$

Now, fix $T_0 = T/3$, $T_1 = 2T/3$ and use Caccioppoli's inequality (a well-known estimate for solution of elliptic and parabolic PDE's)

$$
\begin{aligned}
&\int_0^T\mathrm{d}t\int_{\omega''\setminus\omega'}|\nabla v|^2\mathrm{e}^{-2s\psi_r}\,\mathrm{d}x\\
&\qquad\leq C\int_0^T\mathrm{d}t\int_\omega|v|^2\mathrm{e}^{-2s\psi_r}\,\mathrm{d}x\,,
\end{aligned}
$$

to conclude that

$$
\int_{T/3}^{2T/3}\mathrm{d}t\int_{\Omega\setminus\omega}|v|^2\,\mathrm{d}x\leq C\int_0^T\mathrm{d}t\int_\omega|v|^2\,\mathrm{d}x
$$

or

$$
\int_{T/3}^{2T/3}\mathrm{d}t\int_\Omega|v|^2\,\mathrm{d}x\leq(1+C)\int_0^T\mathrm{d}t\int_\omega|v|^2\,\mathrm{d}x
$$

for some constant C. Then, the dissipativity of the heat operator (that is, the fact that $\int_\Omega|v(t,x)|^2\mathrm{d}x$ is decreasing with respect to t) implies that

$$
\begin{aligned}
\int_\Omega v^2(T,x)\,\mathrm{d}x&\leq\frac{3}{T}\int_{T/3}^{2T/3}\mathrm{d}t\int_\Omega v^2(t,x)\,\mathrm{d}x\\
&\leq(1+C)\frac{3}{T}\int_0^T\mathrm{d}t\int_\omega v^2(t,x)\,\mathrm{d}x\,,
\end{aligned}
$$

which is exactly (24).

Wave Equation

Compared to the heat equation, the wave equation (5) exhibits a quite different behavior from the point of view of exact controllability. Indeed, on the one hand, there is no obstruction to exact controllability since no regularizing effect is connected with wave propagation. On the other hand, due to the finite speed of propagation, exact controllability cannot be expected to hold true in arbitrary time, as null controllability does for the heat equation.

In fact, a typical result that holds true for the wave equation is the following, where a boundary control of Dirichlet type acts on a part $\Gamma_1\subset\Gamma$, while homogeneous boundary conditions are imposed on $\Gamma_0=\Gamma\setminus\Gamma_1$:

$$
\begin{cases}
\partial_t^2u=\Delta u & \text{in } Q_T\\
u=f\mathbb{1}_{\Gamma_1} & \text{on } \Sigma_T\\
u(0,x)=u_0(x)\,,\ \partial_tu(0,x)=u_1(x) & x\in\Omega
\end{cases}
\tag{32}
$$

Observe that problem (32) is well-posed taking

$$
\begin{aligned}
&u_0\in L^2(\Omega)\,,\quad u_1\in H^{-1}(\Omega)\\
&f\in L^2(0,T;L^2(\Gamma))\\
&u\in C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))\,.
\end{aligned}
$$

Theorem 4 *Let Ω be a bounded domain of $\mathbb{R}^n$ with boundary of class C^2 and suppose that, for some point $x_0\in\mathbb{R}^n$,*

$$
\begin{cases}
(x-x_0)\cdot\nu(x)>0 & \forall x\in\Gamma_1\\
(x-x_0)\cdot\nu(x)\leq 0 & \forall x\in\Gamma_0\,.
\end{cases}
$$

Let

$$
R=\sup_{x\in\Omega}|x-x_0|\,.
$$

Copyrighted material

If $T > 2R$, then, for all $(u_0, u_1), (v_0, v_1) \in L^2(\Omega) \times H^{-1}(\Omega)$ there is a control function $f \in L^2(0, T; L^2(\Gamma))$ such that the solution u_f of (32) satisfies

$$u_f(T, x) = v_0(x), \quad \partial_t u_f(T, x) = v_1(x).$$

As we saw for abstract evolution equations, the above exact controllability property is proved to be equivalent to an observability estimate for the dual homogeneous problem using, for instance, the Hilbert Uniqueness Method (HUM) by J.-L. Lions [86].

Bibliographical Comments

The literature on controllability of parabolic equations and related topics is so huge, that no attempt to provide a comprehensive account of it would fit within the scopes of this note. So, the following comments have to be taken as a first hint for the interested reader to pursue further bibliographical research.

The theory of exact controllability for parabolic equations was initiated by the seminal paper [58] by Fattorini and Russell. Since then, it has experienced an enormous development. Similarly, the multiplier method to obtain observability inequalities for the wave equation was developed in [17,73,74,77,86]. Some fundamental early contributions were surveyed by Russell [108]. The next essential progress was made in the work by Lebeau and Robbiano [83] and then by Fursikov and Imanuvilov in a series of papers. In [65] one can find an introduction to global Carleman estimates, as well as applications to the controllability of several ODE's. In particular, the presentation of this paper as for observability inequalities and Carleman estimates for the heat operator is inspired by the last monograph. General perspectives for the understanding of global Carleman estimates and their applications to unique continuation and control problems for PDE's can be found in the works by Tataru [113,114,115,116]. Usually, the above approach requires coefficients to be sufficiently smooth. Recently, however, interesting adaptations of Carleman estimates to parabolic operators with discontinuous coefficients have been obtained in [21,82].

More recently, interest has focussed on control problems for nonlinear parabolic equations. Different approaches to controllability problems have been proposed in [57] and [44]. Then, null and approximate controllability results have been improved by Fernandez–Cara and Zuazua [61,62]. Techniques to produce insensitizing controls have been developed in [117]. These techniques have been successfully applied to the study of Navier–Stokes equations by several authors, see e. g. [63].

Fortunately, several excellent monographs are now available to help introduce the reader to this subject. For instance, the monograph by Zabczyk [121] could serve as a clean introduction to control and stabilization for finite- and infinite-dimensional systems. Moreover, [22,50,51], as well as [80,81] develop all the basic concepts of control and system theory for distributed parameter systems with special emphasis on abstract formulation. Specific references for the controllability of the wave equation by HUM can be found in [86] and [74]. More recent results related to series expansion and Ingham type methods can be found in [75]. For the control of Navier–Stokes equations the reader is referred to [64], as well as to the book by Coron [43], which contains an extremely rich collection of classical results and modern developments.

Stabilization

Stabilization of flexible structures such as beams, plates, up to antennas of satellites, or of fluids as, for instance, in aeronautics, is an important part of CT. In this approach, one wants either to derive feedback laws that will allow the system to autoregulate once they are implemented, or study the asymptotic behavior of the stabilized system i. e. determine whether convergence toward equilibrium states as times goes to infinity holds, determine its speed of convergence if necessary or study how many feedback controls are required in case of coupled systems.

Different mathematical tools have been introduced to handle such questions in the context of ODE's and then of PDE's. Stabilization of ODE's goes back to the work of Lyapunov and Lasalle. The important property is that trajectories decay along Lyapunov functions. If trajectories are relatively compact in appropriate spaces and the system is autonomous, then one can prove that trajectories converge to equilibria asymptotically. However, the construction of Lyapunov functions is not easy, in general.

This section will be concerned with some aspects of the stabilization of second order hyperbolic equations, our model problem being represented by the wave equation with distributed damping

$$\begin{cases} \partial_{tt} u - \Delta u + a(x) u_t = 0 & \text{in } \Omega \times \mathbb{R}, \\ u = 0 & \text{on } \Sigma = (0, \infty) \times \Gamma \\ (u, \partial_t u)(0) = (u^0, u^1) & \text{on } \Omega, \end{cases} \tag{33}$$

in a bounded domain $\Omega \subset \mathbb{R}^n$ with a smooth boundary Γ. For $n = 2$, $u(t, x)$ represents the displacement of point x of the membrane at time t. Therefore, equation (33) describes an elastic system. The energy of such

Copyrighted material

Using once again (38) with $t = 0$ in this last inequality and the definition of ϕ, one has

$$\int_t^\infty E(s)\,\mathrm{d}s \le cE(0)\exp(-t/c) \quad \forall t \ge 0\,.$$

Since E is a nonnegative and nonincreasing function

$$cE(t) \le \int_{t-c}^{t} E(s)\,\mathrm{d}s \le \int_{t-c}^{\infty} E(s)\,\mathrm{d}s \le cE(0)\exp(-(t-c)/c)\,,$$

so that (39) is proved.

An alternative method is to introduce a modified (or perturbed) energy E_ε which is equivalent to the natural one for small values of the parameter ε as in Komornik and Zuazua [76]. Then one shows that this modified energy satisfies a differential Gronwall inequality so that it decays exponentially at infinity. The exponential decay of the natural energy follows then at once. In this case, the modified energy is indeed a Lyapunov function for the PDE. The natural energy cannot be in general such a Lyapunov function due to the finite speed of propagation (consider initial data which have compact support compactly embedded in $\Omega\backslash\omega$).

There are also very interesting approaches using the frequency domain approach, or spectral analysis such as developed by K. Liu [87] Z. Liu and S. Zheng [88]. In the sequel, we concentrate on the integral inequality method. This method has been generalized in several directions and we present in this note some results concerning extensions to

- nonlinear feedback
- indirect or single feedback for coupled system
- memory type feedbacks

Generalizations to Nonlinear Feedbacks Assume now that the feedback term $a(x)u_t$ in (33) is replaced by a nonlinear feedback $a(x)\rho(u_t)$ where ρ is a smooth, increasing function satisfying $v\rho(v) \ge 0$ for $v \in \mathbb{R}$, linear at ∞ and with polynomial growth close to zero, that is: $\rho(v) = |v|^p$ for $|v| \le 1$ where $p \in (1, \infty)$.

Assume moreover that ω satisfies Zuazua's multiplier geometric condition (MGC) or Liu's piecewise multiplier method (PWMGC). Then using multipliers of the space and time variables defined as $E(s)^{(p-1)/2}\,Mu(x)$ where $Mu(x)$ are multipliers of the form described in section 5.1 and integrating by parts expressions of the form

$$\int_t^T E(s)^{(p-1)/2} \times \int_\Omega \big(\partial_t^2 u - \Delta u + a(x)\rho(u_t)\big)Mu(x)\,\mathrm{d}x\,\mathrm{d}s = 0\,,$$

one can prove that the energy E of solutions satisfies the following inequality for all $0 \le t \le T$

$$\int_t^T E^{(p+1)/2}(s)\,\mathrm{d}t \le cE^{(p+1)/2}(t) + c\int_t^T E^{(p-1)/2}(s) \times \left(\int_\Omega \rho(u_t)^2 + \int_\omega |u_t|^2\right).$$

One can remark than an additional multiplicative weight in time depending on the energy has to be taken. This weight is $E^{(p-1)/2}$. Then as in the linear case, but in a more involved way, thanks to the dissipation relation

$$E'(t) = -\int_\Omega a(x)u_t\rho(u_t)\,, \tag{40}$$

one can prove that E satisfies the following nonlinear integral inequality

$$\int_t^T E^{(p+1)/2}(s)\,\mathrm{d}s \le cE(t)\,, \quad \forall 0 \le t \le T\,.$$

Thanks to the fact that E is nonincreasing, a well-known result by Komornik [74] shows that E is polynomially decaying, as $t^{-2/(p-1)}$ at infinity. The above type results have been obtained by many authors under weaker form (see also [40,41,71,98,122]).

Extensions to nonlinear feedbacks without growth conditions close to zero have been studied by Lasiecka and Tataru [78], Martinez [93], W. Liu and Zuazua [89], Eller Lagnese and Nicaise [56] and Alabau–Boussouira [5]. We present the results obtained in this last reference since they provide *optimal* decay rates.

The method is as follows. Define respectively the linear and nonlinear kinetic energies

$$\begin{cases} \int_\omega |u_t|^2\,\mathrm{d}x\,, \\ \int_\Omega |a(x)\rho(u_t)|^2\,\mathrm{d}x\,, \end{cases}$$

and use a weight function in time $f(E(s))$ which is to be determined later on in an optimal way. Integrating by parts expressions of the form

$$\int_t^T f(E(s))\int_\Omega \big(\partial_t^2 u - \Delta u + a(x)\rho(u_t)\big)Mu(x)\,\mathrm{d}x\,\mathrm{d}s = 0\,,$$

one can prove that the energy E of solutions satisfies the

Copyrighted material

following inequality for all $0 \le t \le T$

$$\int_t^T E(s) f(E(s))\,\mathrm{d}s \le c f(E(t)) + c \int_t^T f(E(s)) \times \left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2 \right). \quad (41)$$

The difficulty is to determine the optimal weight under general growth conditions on the feedback close to 0, in particular for cases for which the feedback decays to 0 faster than polynomials.

Assume now that the feedback satisfies

$$g(|v|) \le |\rho(v)| \le C g^{-1}(|v|), \quad \forall |v| \le 1, \quad (42)$$

where g is continuously differentiable on $\mathbb{R}$ strictly increasing with $g(0) = 0$ and

$$\begin{cases} g \in C^2([0, r_0]), r_0 \text{ sufficiently small}, \\ H(.) = \sqrt{.}\, g(\sqrt{.}) \text{ is strictly convex on } \left[0, r_0^2\right], \\ g \text{ is odd}. \end{cases}$$

Moreover, ρ is assumed to have a linear growth at infinity. We define the *optimal* weight function f as follows.

We first extend H to a function $\hat{H}$ define on all $\mathbb{R}$

$$\hat{H}(x) = \begin{cases} H(x) \text{ if } x \in \left[0, r_0^2\right], \\ +\infty \text{ otherwise}, \end{cases}$$

then, define a function F as follows:

$$F(y) = \begin{cases} \dfrac{\hat{H}^*(y)}{y} & \text{if } y \in (0, +\infty), \\ 0 & \text{if } y = 0, \end{cases}$$

where $\hat{H}^*$ stands for the convex conjugate of $\hat{H}$, that is

$$\hat{H}^*(y) = \sup_{x \in \mathbb{R}} \{x\, y - \hat{H}(x)\}.$$

Then the *optimal* weight function f is determined in the following way

$$f(s) = F^{-1}(s/2\beta) \quad s \in \left[0, 2\beta r_0^2\right),$$

where β is of the form $\max(\eta_1, \eta_2 E(0))$, η_1 and η_2 being explicit positive constants.

One can prove that the above formulas make sense, and in particular that F is invertible and smooth. More precisely, F is twice continuously differentiable strictly increasing, one-to-one function from $[0, +\infty)$ onto $[0, r_0^2)$. Note that since the feedback is supposed to be linear at infinity, if one wants to obtain results for general growth types of the feedback, one can assume convexity of H only in a neighborhood of 0.

One can prove from (41) that there exists an (explicit) $T_0 > 0$ such that for all initial data, E satisfies the following nonlinear integral inequality

$$\int_t^T E(s) f(E(s))\,\mathrm{d}s \le T_0 E(t) \quad \forall 0 \le t \le T. \quad (43)$$

This inequality is proved thanks to convexity arguments as follows. Thanks to the convexity of $\hat{H}$, one can use Jensen's inequality and (42), so that

$$\int_{\Omega_t} |a(x)\rho(u_t)|^2\,\mathrm{d}x \le \gamma_1(t) \hat{H}^{-1} \times \left(\frac{1}{\gamma_1(t)} \int_\Omega a(x) u_t \rho(u_t)\,\mathrm{d}x \right)$$

In a similar way, one proves that

$$\int_{\omega_t} |u_t|^2\,\mathrm{d}x \le \gamma_2(t) \hat{H}^{-1} \left(\frac{1}{\gamma_2(t)} \int_\Omega a(x) u_t \rho(u_t)\,\mathrm{d}x \right)$$

where Ω_t and ω_t are time-dependent sets of respective Lebesgue measures $\gamma_1(t)$ and $\gamma_2(t)$ on which the velocity $u_t(t, x)$ is sufficiently small. Using the above two estimates, together with the linear growth of ρ at infinity, one proves

$$\int_t^T f(E(s)) \left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2 \right) \le \int_t^T f(E(s)) \hat{H}^{-1} \left(\frac{1}{c} \int_\Omega a(x) u_t \rho(u_t)\,\mathrm{d}x \right)$$

Using then Young's inequality, together with the dissipation relation (40) in the above inequality, one obtains

$$\int_t^T f(E(s)) \left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2 \right) \le C_1 \int_t^T \hat{H}^\star \left(f(E(s)) \right) \mathrm{d}s + C_2 E(t), \quad (44)$$

where $C_i > 0$ $i = 1, 2$ is a constant independent of the initial data. Using the dissipation relation (40) in the above inequality, this gives for all $0 \le t \le T$

Combining this last inequality with (41) gives

$$\int_t^T E(s) f(E(s))\,\mathrm{d}s \le \beta \int_t^T (\hat{H})^\star \left(f(E(s)) \right) \mathrm{d}s + C_2 E(t)$$

where β is chosen of the form $\max(\eta_1, \eta_2 E(0))$, η_1 and η_2 being explicit positive constants to guarantee that the argument E of f stays in the domain of definition of f.

Copyrighted material

Thus (43) is proved, thanks to the fact that the weight function has been chosen so that

$$\beta \hat{H}^{\star}(f(E(s)) = \frac{1}{2}E(s)f(E(s)) \quad \forall\, 0 \le s\,.$$

Therefore E satisfies a nonlinear integral inequality with a weight function $f(E)$ which is defined in a semi-explicit way in general cases of feedback growth.

The last step is to prove that a nonincreasing and nonnegative absolutely continuous function E satisfying a nonlinear integral inequality of the form (43) is decaying at infinity, and to establish at which rate this holds. For this, one proceeds as in [5].

Let $\eta > 0$ and $T_0 > 0$ be fixed given real numbers and F be a strictly increasing function from $[0, +\infty)$ on $[0, \eta)$, with $F(0) = 0$ and $\lim_{y\to+\infty} F(y) = \eta$.

For any $r \in (0, \eta)$, we define a function K_r from $(0, r]$ on $[0, +\infty)$ by

$$K_r(\tau) = \int_{\tau}^{r} \frac{\mathrm{d}y}{yF^{-1}(y)}\,, \tag{45}$$

and a function ψ_r which is a strictly increasing function from $[\frac{1}{F^{-1}(r)}, +\infty)$ onto $[\frac{1}{F^{-1}(r)}, +\infty)$ by

$$\psi_r(z) = z + K_r(F(\frac{1}{z})) \ge z, \quad \forall z \ge \frac{1}{F^{-1}(r)}\,, \tag{46}$$

One can prove that if E is a nonincreasing, absolutely continuous function from $[0, +\infty)$ on $[0, +\infty)$, satisfying $0 < E(0) < \eta$ and the inequality

$$\int_{t}^{T} E(s)F^{-1}(E(s))\,\mathrm{d}s \le T_0E(S)\,, \quad \forall\, 0 \le t \le T\,, \tag{47}$$

then E satisfies the following estimate:

$$E(t) \le F\left(\frac{1}{\psi_r^{-1}(\frac{t}{T_0})}\right), \quad \forall t \ge \frac{T_0}{F^{-1}(r)}\,, \tag{48}$$

where r is any real such that

$$\frac{1}{T_0}\int_{0}^{+\infty} E(\tau)F^{-1}(E(\tau))\,\mathrm{d}\tau \le r \le \eta\,.$$

Thus, one can apply the above result to E with $\eta = r_0^2$ and show that $\lim t \to +\infty E(t) = 0$, the decay rate being given by estimate (48).

If g is polynomial close to zero, one gets back that the energy $E(t)$ decays as $t^{\frac{-2}{p-1}}$ at infinity. If $g(v)$ behaves as $\exp(-1/|v|)$ close to zero, then $E(t)$ decays as $1/(\ln(t))^2$ at infinity.

The usefulness of convexity arguments has been first pointed out by Lasiecka and Tataru [78] using Jensen's inequality and then in different ways by Martinez [93] (the weight function does not depend on the energy) and W. Liu and Zuazua [89] and Eller Lagnese and Nicaise [56]. Optimal decay rates have been obtained by Alabau–Boussouira [5,6] using a weight function determined through the theory of convex conjugate functions and Young's (named also as Fenchel–Moreau's) inequality. This argument was also used by W. Liu and Zuazua [89] in a slightly different way and combined to a Lyapunov technique. Optimality of estimates in [5] is proved in one-dimensional situation and for boundary dampings applying optimality results of Vancostenoble [119] (see also Martinez and Vancostenoble [118]).

Indirect Damping for Coupled Systems

Many complex phenomena are modeled through coupled systems. In stabilizing (or controlling) energies of the vector state, one has very often access only to some components of this vector either due to physical constraints or to cost considerations. In this case, the situation is to stabilize a full system of coupled equation through a reduced number of feedbacks. This is called indirect damping. This notion has been introduced by Russell [109] in 1993.

As an example, we consider the following system:

$$\begin{cases} \partial_t^2 u - \Delta u + \partial_t u + \alpha v = 0 \\ \partial_t^2 v - \Delta v + \alpha u = 0 \end{cases}$$
$$\text{in } \Omega \times \mathbb{R}\,, \quad u = 0 = v \quad \text{on } \partial\Omega \times \mathbb{R}\,. \tag{49}$$

Here, the first equation is damped through a linear distributed feedback, while no feedback is applied to the second equation. The question is to determine if this coupled system inherits any kind of stability for nonzero values of the coupling parameter α from the stabilization of the first equation only.

In the finite dimensional case, stabilization (or control) of coupled ODE's can be analyzed thanks to a powerful rank type condition named Kalman's condition. The situation is much more involved in the case of coupled PDE's.

One can show first show that the above system fails to be exponentially stable (see also [66] for related results). More generally, one can study the stability of the system

$$\begin{cases} u'' + A_1 u + Bu' + \alpha v = 0 \\ v'' + A_2 v + \alpha u = 0 \end{cases} \tag{50}$$

in a separable Hilbert space H with norm $|\cdot|$, where A_1, A_2 and B are self-adjoint positive linear operators in H. Moreover, B is assumed to be a bounded operator. So, our analysis applies to systems with internal damping supported

Copyrighted material

in the whole domain Ω such as (49); the reader is referred to [1,2] for related results concerning boundary stabilization problems (see also Beyrath [23,24] for localized indirect dampings).

In light of the above observations, system (50) fails to be exponentially stable, at least when H is infinite dimensional and A_1 has a compact resolvent as in (49). Indeed it is shown in Alabau, Cannarsa and Komornik [8] that the total energy of sufficiently smooth solutions of (50) decays polynomially at infinity whenever $|\alpha|$ is small enough but nonzero. From this result we can also deduce that any solution of (50) is strongly stable regardless of its smoothness: this fact follows by a standard density argument since the semigroup associated with (50) is a contraction semigroup.

A brief description of the key ideas of the approach developed in [2,8] is as follows. Essentially, one uses a finite iteration scheme and suitable multipliers to obtain an estimate of the form

$$\int_0^T E(u(t),v(t))\mathrm{d}t \le c\sum_{k=0}^{j} E(u^{(k)}(0),v^{(k)}(0)) \qquad \forall\, T\ge 0\,, \quad (51)$$

where j is a positive integer and E denotes the total energy of the system

$$E(u,v)=\frac{1}{2}\left(|A_1^{1/2}u|^2+|u'|^2\right)+\frac{1}{2}\left(|A_2^{1/2}v|^2+|v'|^2\right)+\alpha\langle u,v\rangle\,.$$

Once (51) is proved, an abstract lemma due to Alabau [1,2] shows that $E(u(t),v(t))$ decays polynomially at ∞. This abstract lemma can be stated as follows.

Let A be the infinitesimal generator of a continuous semi-group $\exp(tA)$ on an Hilbert space $\mathcal{H}$, and $D(A)$ its domain. For U^0 in $\mathcal{H}$ we set in all the sequel $U(t)=\exp(tA)U^0$ and assume that there exists a functional E defined on $C([0,+\infty),\mathcal{H})$ such that for every U^0 in $\mathcal{H}$, $E(\exp(.A))$ is a non-increasing, locally absolutely continuous function from $[0,+\infty)$ on $[0,+\infty)$. Assume moreover that there exist an integer $k\in\mathbb{N}^\star$ and nonnegative constants c_p for $p=0,\dots k$ such that

$$\int_S^T E(U(t))\mathrm{d}t \le \sum_{p=0}^{k} c_p E(U^{(p)}(S)) \qquad \forall 0\le S\le T\,,\forall U^0\in D(A^k)\,. \quad (52)$$

Then the following inequalities hold for every U^0 in $D(A^{kn})$ and all $0\le S\le T$ where n is any positive integer:

$$\int_S^T E(U(\tau))\frac{(\tau-S)^{n-1}}{(n-1)!}\mathrm{d}\tau \le c\sum_{p=0}^{kn} E\big(U^{(p)}(S)\big)\,, \quad (53)$$

and

$$E(U(t))\le c\sum_{p=0}^{kn} E\big(U^{(p)}(0)\big)t^{-n} \qquad \forall t>0\,,\quad \forall U^0\in D\big(A^{kn}\big)\,,$$

where c is a constant which depends on n.

First (53) is proved by induction on n. For $n=1$, it reduces to the hypothesis (52). Assume now that (53) holds for n and let U^0 be given in $D(A^{k(n+1)})$. Then we have

$$\int_S^T\int_t^T E(U(\tau))\frac{(\tau-t)^{n-1}}{(n-1)!}\mathrm{d}\tau\mathrm{d}t \le c\sum_{p=0}^{kn}\int_S^T E\big(U^{(p)}(t)\big)\mathrm{d}t \qquad \forall\, 0\le S\le T\,,\forall\, U^0\in D\big(A^{kn}\big).$$

Since U^0 is in $D(A^{k(n+1)})$ we deduce that $U^{(p)}(0)=A^pU^0$ is in $D(A^k)$ for $p\in\{0,\dots kn\}$. Hence we can apply the assumption (52) to the initial data $U^{(p)}(0)$. This together with Fubini's Theorem applied on the left hand side of the above inequality give (53) for $n+1$. Using the property that $E(U(t))$ is non increasing in (53) we easily obtain the last desired inequality.

Applications to wave-wave, wave-Petrowsky equations and various concrete examples hold.

The above results have been studied later on by Batkai, Engel, Prüss and Schnaubelt [18] using very interesting resolvent and spectral criteria for polynomial stability of abstract semigroups. The above abstract lemma in [2] has also been generalized using interpolation theory. One should note that this integral inequality involving higher order energies of solutions is not of differential nature, unlike Haraux's and Komornik's integral inequalities. Another approach based on decoupling techniques and for slightly different abstract systems have been introduced by Ammar Khodja Bader and Ben Abdallah [12].

Spectral conditions have also been studied by Z. Liu [88] and later on by Z. Liu and Rao [90], Loreti and Rao [92] for peculiar abstract systems and in general for coupled equations only of the *same* nature (wave-wave for instance), so that a dispersion relation for the eigenvalues of the coupled system can be derived. Also these last results are given for internal stabilization only. Because of the above limitations, Z. Liu–Rao and Loreti–Rao's results

Copyrighted material

are less powerful in generality than the ones given by Alabau, Cannarsa and Komornik [8] and Alabau [2]. Moreover results through energy type estimates and integral inequalities can be generalized to include nonlinear indirect dampings as shown in [7]. On the other side spectral methods are very useful to obtain optimal decay rates provided that one can determine at which speed the eigenvalues approach the imaginary axis for high frequencies.

Memory Dampings

We consider the following model problem

$$\begin{cases} u_{tt}(t,x) - \Delta u(t,x) + \int_0^t \beta(t-s)\,\Delta u(s,x)\,\mathrm{d}s = \\ \qquad\qquad\qquad\qquad |u(t,x)|^\gamma u(t,x) \\ u(t,\cdot)_{|\partial\Omega} = 0 \\ (u(0,\cdot), u_t(0,\cdot)) = (u_0, u_1) \end{cases} \tag{54}$$

where $0 < \gamma \leq \frac{2}{N-2}$ holds. The second member is a source term. The damping

$$\int_0^t \beta(t-s)\Delta u(s,x)\,\mathrm{d}s$$

is of memory type.

The energy is defined by

$$\begin{aligned} E_u(t) = {} & \frac{1}{2}\|u_t(t)\|^2_{L^2(\Omega)}\,\mathrm{d}x \\ & + \frac{1}{2}\left(1 - \int_0^t \beta(s)\,\mathrm{d}s\right)\|\nabla u(t)\|^2_{L^2(\Omega)} \\ & - \frac{1}{\gamma+2}\|u(t)\|^{\gamma+2}_{L^{\gamma+2}(\Omega)} \\ & + \frac{1}{2}\int_0^t \beta(t-s)\|\nabla u(t) - \nabla u(s)\|^2_{L^2(\Omega)}\,\mathrm{d}s \end{aligned}$$

The damping term produces dissipation of the energy, that is (for strong solutions)

$$\begin{aligned} E'_u(t) = {} & -\frac{1}{2}\beta(t)\|\nabla u(t)\|^2 \\ & + \frac{1}{2}\int_0^t \beta'(t)\|\nabla u(s) - \nabla u(t)\|^2\,\mathrm{d}s \leq 0 \end{aligned}$$

One can consider more general abstract equations of the form

$$u''(t) + Au(t) - \int_0^t \beta(t-s)Au(s)\,\mathrm{d}s = \nabla F(u(t)) \qquad t \in (0,\infty) \tag{55}$$

in a Hilbert space X, where $A\colon D(A) \subset X \to X$ is an accretive self-adjoint linear operator with dense domain, and ∇F denotes the gradient of a Gâteaux differentiable functional $F\colon D(A^{1/2}) \to \mathbb{R}$. In particular, equation (54) fits into this framework as well as several other classical equations of mathematical physics such as the linear elasticity system.

We consider the following assumptions.

Assumptions (H1)

1. A is a self-adjoint linear operator on X with dense domain $D(A)$, satisfying
$$\langle Ax, x\rangle \geq M\|x\|^2 \quad \forall x \in D(A) \tag{56}$$
for some $M > 0$.
2. $\beta\colon [0,\infty) \to [0,\infty)$ is a locally absolutely continuous function such that
$$\int_0^\infty \beta(t)\mathrm{d}t < 1\ \beta(0) > 0 \quad \beta'(t) \leq 0 \qquad \text{for a.e. } t \geq 0\,.$$
3. $F\colon D(A^{1/2}) \to \mathbb{R}$ is a functional such that
   1. F is Gâteaux differentiable at any point $x \in D(A^{1/2})$;
   2. for any $x \in D(A^{1/2})$ there exists a constant $c(x) > 0$ such that
   $$|DF(x)(y)| \leq c(x)\|y\|, \qquad \text{for any } y \in D(A^{1/2}),$$
   where $DF(x)$ denotes the Gâteaux derivative of F in x; consequently, $DF(x)$ can be extended to the whole space X (and we will denote by $\nabla F(x)$ the unique vector representing $DF(x)$ in the Riesz isomorphism, that is, $\langle \nabla F(x), y\rangle = DF(x)(y)$, for any $y \in X$);
   3. for any $R > 0$ there exists a constant $C_R > 0$ such that
   $$\|\nabla F(x) - \nabla F(y)\| \leq C_R\|A^{1/2}x - A^{1/2}y\|$$
   for all $x, y \in D(A^{1/2})$ satisfying $\|A^{1/2}x\|\,,\ \|A^{1/2}y\| \leq R$.

Assumptions (H2)

1. There exist $p \in (2,\infty]$ and $k > 0$ such that
$$\beta'(t) \leq -k\beta^{1+\frac{1}{p}}(t) \quad \text{for a.e. } t \geq 0$$
(here we have set $\frac{1}{p} = 0$ for $p = \infty$).

Copyrighted material

and h. For further reading, one may consult the monographs [22,85], and [79], for problems where the running and final costs are given by quadratic forms (the so-called Linear Quadratic problem), or [84] and [59] for more general optimal control problems.

Necessary Conditions

Once the existence of a solution to problem (60)–(61) has been established, the next important step is to provide conditions to detect a candidate solution, possibly showing that it is, in fact, optimal. By and large the optimality conditions of most common use are the ones known as Pontryagin's Maximum Principle named after the Russian mathematician L.S. Pontryagin who greatly contributed to the development of control theory, see [100,101].

So, suppose $\{u_*, f_*\}$, where $u_* = u_{f_*}$ is a candidate optimal pair and consider the so-called adjoint system

$$\begin{cases} -p'(t) = A^* p(t) + \partial_u L(t, u_*(t), f_*(t)) = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad t \in [0, T] \ \text{ a.e.} \\ p(T) = \partial \ell(u_*(T)) \,, \end{cases}$$

where $\partial_u L(t, u, \phi)$ and $\partial \ell(u)$ denote the Fréchet gradients of the maps $L(t, \cdot, \phi)$ and ℓ at u, respectively. Observe that the above is a backward linear Cauchy problem with terminal condition, which can obviously be reduced to a forward one by the change of variable $t \to T - t$. So, it admits a unique mild solution, labeled p_*, which is called the *adjoint state* associated with $\{u_*, f_*\}$.

Pontryagin's Maximum Principle states that, if $\{u_*, f_*\}$ is optimal, then

$$\langle p_*(t), B f_*(t) \rangle + L(t, u_*(t), f_*(t)) = \min_{\phi \in F} \left[\langle p_*(t), B\phi \rangle + L(t, u_*(t), \phi) \right] \qquad t \in [0, T] \ \text{ a.e.} \quad (65)$$

The name Maximum Principle rather than Minimum Principle, as it would be more appropriate, is due to the fact that, traditionally, attention was focussed on the *maximization*—instead of minimization—of the functional in (61). Even today, in most models from economics, one is interested in maximizing payoffs, such as revenues, utility, capital and so on. In that case, (65) would still be true, with a "max" instead of a "min".

At first glance, it might be hard to understand the revelance of (65) to problem (61). To explain this, introduce the function, called the *Hamiltonian*,

$$\mathcal{H}(t, u, p) = \min_{\phi \in F} \left[\langle p, B\phi \rangle + L(t, u, \phi) \right] \qquad (t, u, p) \in [0, T] \times H \times H \,. \quad (66)$$

Then, Fermat's rule yields $B^* p + \partial_\phi L(t, u, \phi) = 0$ at every $\phi \in F$ at which the minimum in (66) is attained. Therefore, from (65) it follows that

$$B^* p_*(t) + \partial_\phi L(t, u_*(t), f_*(t)) = 0 \quad t \in [0, T] \ \text{ a.e.} \quad (67)$$

which provides a much-easier-to-use optimality condition.

There is a vast literature on necessary conditions for optimality for distributed parameter systems. The set-up that was considered above can be generalized in several ways: one can consider nonlinear state equations as in (62), nonsmooth running and finals costs, constraints on both state and control, problems with infinite horizon or exit times. Further reading and useful references on most of these extensions can be found in the aforementioned monographs [22,79,84,85], and in [59] which is mainly concerned with time optimal control problems.

Dynamic Programming

Though useful as it may be, Pontryagin's Maximum Principle remains a necessary condition. So, without further information, it does not suffice to prove the optimality of a give trajectory/control pair. Moreover, even when the map $\phi \mapsto \partial_\phi L(t, u, \phi)$ turns out to be invertible, the best result identity (67) can provide, is a representation of $f_*(t)$ in terms of $u_*(t)$ and $p_*(t)$: not enough to determine $f_*(t)$, in general.

This is why other methods to construct optimal controls have been proposed over the years. One of the most interesting ones is the so-called *dynamic programming* method (abbreviated, DP), initiated by the work of R. Bellman [20]. Such a method will be briefly described below in the set-up of distributed parameter systems.

Fix $T > 0$, s such that $0 \le s \le T$, and consider the optimal control problem

to minimize

$$J^{s,\nu}(f) = \int_s^T L\left(t, u_f^{s,\nu}(t), f(t)\right) \mathrm{d}t + \ell\left(u_f^{s,\nu}(T)\right) \quad (68)$$

over all control functions $f \in L^2(s, T; F)$, where $u_f^{s,\nu}(t)$ is the solution of the controlled system

$$\begin{cases} u'(t) = Au(t) + Bf(t) \quad t \in [s, T] \\ u(s) = \nu \,. \end{cases} \quad (69)$$

The *value function* U associated to (68)-(69) is the real-valued function defined by

$$U(s, \nu) = \inf_{f \in L^2(s,T;F)} J^{s,\nu}(f) \quad \forall (s, \nu) \in [0, T] \times H \,. \quad (70)$$

Copyrighted material

A fundamental step of DP is the following result, known as Bellman's *optimality principle*.

Theorem 5 *For any* $(s, v) \in [0, T] \times H$ *and any* $f \in L^2(s, T; F)$

$$U(s, v) \leq \int_s^r L\left(t, u_f^{s,v}(t), f(t)\right) dt + U\left(r, u_f^{s,v}(r)\right) \qquad \forall r \in [s, T]\,.$$

Moreover, $f^*(\cdot)$ *is optimal if and only if*

$$U(s, v) = \int_s^r L\left(t, u_f^{s,v}(t), f(t)\right) dt + U\left(r, u_f^{s,v}(r)\right) \qquad \forall r \in [s, T]\,.$$

The connection between DP and optimal control is based on the properties of the value function. Indeed, applying Bellman's optimality principle, one can show that, if U is Fréchet differentiable, then

$$\begin{cases} \partial_s U(s, v) + \langle Av, \partial_v U(s, v)\rangle + \mathcal{H}\left(s, v, \partial_v U(s, v)\right) = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad (s, v) \in (0, T) \times D(A) \\ U(T, v) = \ell(v) \quad v \in H \end{cases}$$

where $\mathcal{H}$ is the Hamiltonian defined in (66). The above equation is the celebrated *Hamilton–Jacobi equation* of DP. To illustrate its connections with the original optimal control problem, a useful formal argument—that can, however, be made rigorous—is the following. Consider a sufficiently smooth solution W of the above problem and let $(s, v) \in (0, T) \times D(A)$. Then, for any trajectory/control pair $\{u, f\}$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} W(t, u(t)) &= \partial_s W(t, u(t)) + \langle \partial_v W(t, u(t)), Au(t) \\ &\quad + Bf(t)\rangle \\ &= \langle \partial_v W(t, u(t)), Bf(t)\rangle \\ &\quad - \mathcal{H}\left(t, u(t), \partial_v W(t, u(t))\right) \\ &\geq -L(t, u(t), f(t)) \end{aligned} \tag{71}$$

by the definition of $\mathcal{H}$. Therefore, integrating from s to T,

$$\ell(u(T)) - W(s, v) \geq -\int_s^T L(t, u(t), f(t))\mathrm{d}t,$$

whence $J^{s,v}(f) \geq W(s, v)$. Thus, taking the infimum over all $f \in L^2(s, T; F)$,

$$W(s, v) \leq U(s, v) \quad \forall (s, v) \in (0, T) \times D(A)\,. \tag{72}$$

Now, suppose there is a control function $f_* \in L^2(s, T; F)$ such that, for all $t \in [s, T]$,

$$\begin{aligned} &\langle \partial_v W(t, u_*(t)), Bf_*(t)\rangle + L(t, u_*(t), f_*(t)) \\ &\qquad = \mathcal{H}\left(t, u_*(t), \partial_v W(t, u_*(t))\right), \end{aligned} \tag{73}$$

where $u_*(\cdot) = u_{f_*}^{s,v}(\cdot)$. Then, from (71) and (73) it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t} W(t, u_*(t)) = -L(t, u(t), f(t))\,,$$

whence

$$W(s, v) = J^{s,v}(f_*) \geq U(s, v)\,.$$

From the above inequality and (72) it follows that $W(s, v) = U(s, v)$ for all $(s, v) \in (0, T) \times D(A)$, hence for all $(s, v) \in (0, T) \times H$ since $D(A)$ is dense in H. So, f_* is an **optimal control**.

Note 2 The above considerations lead to the following procedure to obtain optimal an optimal trajectory:

- find a smooth solution of the Hamilton–Jacobi equation;
- for every $(t, v) \in (0, T) \times D(A)$ provide a feedback $f(t, v)$ such that

$$\begin{aligned} &\langle \partial_v W(t, v), Bf(t, v)\rangle + L(t, v, f(t, v)) \\ &\qquad = \mathcal{H}(t, v, \partial_v W(t, v)) \end{aligned}$$

- solve the so-called *closed loop equation*

$$\begin{cases} u'(t) = Au(t) + Bf(t, u(t)) \quad t \in [s, T] \\ u(s) = v \end{cases}$$

Notice that not only is trajectory u optimal, but the corresponding control f is given in feedback form as well.

Linear Quadratic Optimal Control

One of the most successful applications of DP is the so-called Linear Quadratic optimal control problem. Consider problem (68)–(69) with costs L and ℓ given by

$$L(t, u, \phi) = \langle M(t)u, u\rangle + \langle N(t)\phi, \phi\rangle \qquad \forall (T, u, \phi) \in [0, T] \times H \times F$$

and

$$\ell(u) = \langle Du, u\rangle \quad \forall u \in H\,,$$

where

- $M\colon [0, T] \to \mathcal{L}(H)$ is continuous, $M(t)$ is symmetric and $\langle M(t)u, u\rangle \geq 0$ for every $(t, u) \in [0, T] \times H$;

Copyrighted material

- $N\colon [0,T] \to F$ is continuous, $N(t)$ is symmetric and $\langle N(t)\phi,\phi\rangle \geq c_0|\phi|^2$ for every $(t,\phi) \in [0,T]\times F$ and some constant $c_0 > 0$;
- $D \in \mathcal{L}(H)$ is symmetric and $\langle Du,u\rangle \geq 0$ for every $u \in H$.

Then, assumptions (63) and (64) are satisfied. So, a solution to (68)–(69) does exist. Moreover, it is unique because of the strict convexity of functional $J^{s,v}$.

In order to apply DP, one computes the Hamiltonian

$$\mathcal{H}(t,u,p) = \min_{\phi\in F}\Big[\langle p,B\phi\rangle + \langle M(t)u,u\rangle + \langle N(t)\phi,\phi\rangle\Big]$$
$$= \langle M(t)u,u\rangle - \frac{1}{4}\langle BN^{-1}(t)B^*p,p\rangle\,,$$

where the above minimum is attained at

$$\phi_*(t,p) = -\frac{1}{2}N^{-1}(t)B^*p\,. \tag{74}$$

Therefore, the Hamilton–Jacobi equation associated to the problem is

$$\begin{cases} \partial_s W(s,v) + \langle Av,\partial_v W(s,v)\rangle + \langle M(s)v,v\rangle \\ \quad -\frac{1}{4}\langle BN^{-1}(s)B^*\partial_v W(s,v),\partial_v W(s,v)\rangle = 0 \\ \qquad\qquad \forall (s,v)\in(0,T)\times D(A) \\ w(T,v) = \langle Dv,v\rangle \quad \forall v\in H \end{cases}$$

It is quite natural to search a solution of the above problem in the form

$$W(s,v) = \langle P(s)v,v\rangle \quad \forall (s,v)\in[0,T]\times H\,,$$

with $P\colon [0,T]\to\mathcal{L}(H)$ continuous, symmetric and such that $\langle P(t)u,u\rangle \geq 0$. Substituting into the Hamilton–Jacobi equation yields

$$\begin{cases} \langle P'(s)v,v\rangle + \langle [A^*P(s)+P(s)A]v,v\rangle + \langle M(s)v,v\rangle \\ \quad -\langle BN^{-1}(s)B^*P(s)v,P(s)v\rangle = 0 \\ \qquad\qquad \forall (s,v)\in(0,T)\times D(A) \\ \langle P(T)v,v\rangle = \langle Dv,v\rangle \quad \forall v\in H \end{cases}$$

Therefore, P must be a solution of the so-called *Riccati equation*

$$\begin{cases} P'(s) + A^*P(s) + P(s)A + M(s) \\ \quad -P(s)BN^{-1}(s)B^*P(s) = 0 \qquad \forall s\in(0,T) \\ P(T) = D \end{cases}$$

Once a solution $P(\cdot)$ of the Riccati equation is known, the procedure described in Note 2 can be applied. Indeed, recalling (74) and the fact that $\partial_v W(t,v) = 2P(t)v$, one concludes that $f(t,v) = -N^{-1}(t)B^*P(t)v$ is a feedback law. So, solving the closed loop equation

$$\begin{cases} u'(t) = [A - BN^{-1}(t)B^*P(t)]u(t) \quad t\in(s,T) \\ u(s) = v \end{cases}$$

one obtains the unique optimal trajectory of problem (68)–(69).

In sum, by DP one reduces the original Linear Quadratic optimal control problem to the problem of finding the solution of the Riccati equation, which is easier to solve than the Hamilton–Jacobi equation.

Bibliographical Comments

Different variants of the Riccati equation have been successfully studied by several authors in connection with different state equations and cost functionals, including boundary control problems and problems for other functional equations, see [22,79] and the references therein. Sometimes, the solution of the Riccati equation related to a linearized model provides feedback stabilization for nonlinear problems as in [104].

Unfortunately, the DP method is hard to implement for general optimal control problems, because of several obstructions: nonsmoothness of solutions to Hamilton–Jacobi equations, selection problems that introduce discontinuities, unboundedness of the coefficients, numerical complexity. Besides the Linear Quadratic case, the so-called Linear Convex case is the other example that can be studied by DP under fairly general conditions, see [14]. For nonlinear optimal control problems some of the above difficulties have been overcome extending the notion of viscosity solutions to infinite dimensional spaces, see [45,46,47,48,49], see also [28,29,30,31,32,33] and [112]. Nevertheless, finding additional ideas to make a generalized use of DP for distributed parameter systems possible, remains a challenging problem for the next generations.

Future Directions

In addition to all considerations spread all over this article on promising developments of recent—as well as established—research lines, a few additional topics deserve to be mentioned.

The one subject that has received the highest attention, recently, is that of numerical approximation of control problems, from the point of view of both controllability and optimal control. Here the problem is that, due to high frequency spurious numerical solutions, stable algorithms for solving initial-boundary value problems do not necessarily yield convergent algorithms for computing

Copyrighted material

controls. This difficulty is closely related to the existence of concentrated numerical solutions that escape the observation mechanisms. Nevertheless, some interesting results have been obtained so far, see, e. g., [124,125].

Several interesting results for nonlinear control problems have been obtained by the *return method*, developed initially by Coron [42] for a stabilization problem. This and other techniques have then been applied to fluid models ([68,69]), the Korteweg–de Vries equation ([105,106,107]), and Schrödinger type equations ([19]), see also [43] and the references therein. It seems likely that these ideas, possibly combined with other techniques like Carleman estimates as in [70], will lead to new exiting results in the years to come.

A final comment on null controllability for *degenerate parabolic equations* is in order. Indeed, many problems that are relevant for applications are described by parabolic equation equations in divergence form

$$\partial_t u = \nabla\cdot(A(x)\nabla u)+b(x)\cdot\nabla u+c(t,x)u+f \quad \text{in} \quad Q_T\,,$$

or in the general form

$$\partial_t u = \mathrm{Tr}\,[A(x)\nabla^2 u]+b(x)\cdot\nabla u+c(t,x)u+f \quad \text{in} \quad Q_T\,,$$

where $A(x)$ is a symmetric matrix, positive definite in Ω but possibly singular on Γ. For instance, degenerate parabolic equations arise in fluid dynamics as suitable transformations of the Prandtl equations, see, e. g., [94]. They can also be obtained as Kolmogorov equations of diffusions processes on domains that are invariant for stochastic flows, see, e. g., [52]. The latter interpretation explains why they have been applied to biological problems, such as gene frequency models for population genetics (see, e. g., the Wright–Fischer model studied in [111]).

So far, null controllability properties of degenerate parabolic equations have been fully understood only in dimension one: for some kind of degeneracy, null controllability holds true (see [36] and [9]), but, in general, one can only expect regional null controllability (see [35]). Since very little is known on null controllability for degenerate parabolic equations in higher space dimensions, it is conceivable that such a topic will provide interesting problems for future developments.

Bibliography

1. Alabau F (1999) Stabilisation frontière indirecte de systèmes faiblement couplés. C R Acad Sci Paris Sér I 328:1015–1020
2. Alabau F (2002) Indirect boundary stabilization of weakly coupled systems. SIAM J Control Optim 41(2):511–541
3. Alabau-Boussouira F (2003) A two-level energy method for indirect boundary observability and controllability of weakly coupled hyperbolic systems. SIAM J Control Optim 42(3):871–906
4. Alabau-Boussouira F (2004) A general formula for decay rates of nonlinear dissipative systems. C R Math Acad Sci Paris 338:35–40
5. Alabau-Boussouira F (2005) Convexity and weighted integral inequalities for energy decay rates of nonlinear dissipative hyperbolic systems. Appl Math Optim 51(1):61–105
6. Alabau-Boussouira F (2006) Piecewise multiplier method and nonlinear integral inequalities for Petrowsky equations with nonlinear dissipation. J Evol Equ 6(1):95–112
7. Alabau-Boussouira F (2007) Asymptotic behavior for Timoshenko beams subject to a single nonlinear feedback control. NoDEA 14(5–6):643–669
8. Alabau F, Cannarsa P, Komornik V (2002) Indirect internal damping of coupled systems. J Evol Equ 2:127–150
9. Alabau-Boussouira F, Cannarsa P, Fragnelli G (2006) Carleman estimates for degenerate parabolic operators with applications to null controllability. J Evol Equ 6:161–204
10. Alabau-Boussouira F, Cannarsa P, Sforza D (2008) Decay estimates for second order evolution equations with memory. J Funct Anal 254(5):1342–1372
11. Ammari K, Tucsnak M (2001) Stabilization of second order evolution equations by a class of unbounded feedbacks. ESAIM Control Optim Calc Var 6:361–386
12. Ammar-Khodja F, Bader A, Benabdallah A (1999) Dynamic stabilization of systems via decoupling techniques. ESAIM Control Optim Calc Var 4:577–593
13. Barbu V (2003) Feedback stabilization of Navier–Stokes equations. ESAIM Control Optim Calc Var 9:197–206
14. Barbu V Da Prato G (1983) Hamilton Jacobi equations in Hilbert spaces. Pitman, London
15. Barbu V, Lasiecka I, Triggiani R (2006) Tangential boundary stabilization of Navier–Stokes equations. Mem Amer Math Soc 181(852):128
16. Barbu V, Triggiani R (2004) Internal stabilization of Navier–Stokes equations with finite-dimensional controllers. Indiana Univ Math J 53(5):1443–1494
17. Bardos C, Lebeau G, Rauch R (1992) Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. SIAM J Control Optim 30:1024–1065
18. Bátkai A, Engel KJ, Prüss J, Schnaubelt R (2006) Polynomial stability of operator semigroups. Math Nachr 279(13–14):1425–1440
19. Beauchard K (2005) Local controllability of a 1-D Schödinger equation. J Math Pures Appl 84(7):851–956
20. Bellman R (1957) Dynamic Programming. Princeton University Press, Princeton
21. Benabdallah A, Dermenjian Y, Le Rousseau J (2007) Carleman estimates for the one-dimensional heat equation with a discontinuous coefficient and applications to controllability and an inverse problem. J Math Anal Appl 336(2):865–887
22. Bensoussan A, Da Prato G, Delfour MC, Mitter SK (1993) Representation and Control of Infinite Dimensional Systems. Systems and Control: Foundations and applications, Birkhäuser, Boston
23. Beyrath A (2001) Stabilisation indirecte localement distribué de systèmes faiblement couplés. C R Acad Sci Paris Sér I Math 333(5):451–456

Copyrighted material

24. Beyrath A (2004) Indirect linear locally distributed damping of coupled systems. Bol Soc Parana Mat 22(2):17–34
25. Burq N, Gérard P (1997) Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes. C R Acad Sci Paris Sér I Math 325(7):749–752
26. Burq N, Hitrik M (2007) Energy decay for damped wave equations on partially rectangular domains. Math Res Lett 14(1):35–47
27. Burq N, Lebeau G (2001) Mesures de défaut de compacité, application au système de Lamé. Ann Sci École Norm Sup (4) 34(6):817–870
28. Cannarsa P (1989) Regularity properties of solutions to Hamilton–Jacobi equations in infinite dimensions and nonlinear optimal control. Diff Integral Equ 2:479–493
29. Cannarsa P, Da Prato G (1990) Some results on non-linear optimal control problems and Hamilton–Jacobi equations in infinite dimensions. J Funct Anal 90:27–47
30. Cannarsa P, Gozzi F, Soner HM (1991) A boundary value problem for Hamilton–Jacobi equations in Hilbert spaces. Applied Math Optim 24:197–220
31. Cannarsa P, Gozzi F, Soner HM (1993) A dynamic programming approach to nonlinear boundary control problems of parabolic type. J Funct Anal 117:25–61
32. Cannarsa P, Di Blasio G (1995) A direct approach to infinite dimensional Hamilton–Jacobi equations and applications to convex control with state constraints. Differ Integral Equ 8:225–246
33. Cannarsa P, Tessitore ME (1996) Infinite dimensional Hamilton–Jacobi equations and Dirichlet boundary control problems of parabolic type. SIAM J Control Optim 34:1831–1847
34. Cannarsa P, Komornik V, Loreti P (2002) One-sided and internal controllability of semilinear wave equations with infinitely iterated logarithms. Discret Contin Dyn Syst 8(3):745–756
35. Cannarsa P, Martinez P, Vancostenoble J (2004) Persistent regional null controllability for a class of degenerate parabolic equations. Commun Pure Appl Anal 3(4):607–635
36. Cannarsa P, Martinez P, Vancostenoble (2005) Null Controllability of degenerate heat equations. Adv Differ Equ 10(2):153–190
37. Cavalcanti MM, Oquendo HP (2003) Frictional versus viscoelastic damping in a semilinear wave equation. SIAM J Control Optim 42:1310–1324
38. Chen G (1981) A note on boundary stabilization of the wave equation. SIAM J Control Optim 19:106–113
39. Chueshov I, Lasiecka I, Toundykov D (2008) Long-term dynamics of semilinear wave equation with nonlinear localized interior damping and a source term of critical exponent. Discret Contin Dyn Syst 20(3):459–509
40. Conrad F, Rao B (1993) Decay of solutions of the wave equation in a star-shaped domain with nonlinear boundary feedback. Asymptot Anal 7:159–177
41. Conrad F, Pierre (1994) Stabilization of second order evolution equations by unbounded nonlinear feedback. Ann Inst H Poincaré Anal Non Linéaire 11(5):485–515, Asymptotic Anal 7:159–177
42. Coron JM (1992) Global asymptotic stabilization for controllable systems without drift. Math Control Signals Syst 5(3):295–312
43. Coron JM (2007) Control and nonlinearity. Mathematical surveys and monographs vol 136, Providence, RI: xiv+426
44. Coron JM, Trélat E (2004) Global steady-state controllability of one-dimensional semilinear heat equations. SIAM J Control Optim 43(2):549–569
45. Crandall MG, Lions PL (1985) Hamilton Jacobi equation in infinite dimensions I: Uniqueness of viscosity solutions. J Funct Anal 62:379–396
46. Crandall MG, Lions PL (1986) Hamilton Jacobi equation in infinite dimensions II: Existence of viscosity solutions. J Funct Anal 65:368–425
47. Crandall MG, Lions PL (1986) Hamilton Jacobi equation in infinite dimensions III. J Funct Anal 68:214–247
48. Crandall MG, Lions PL (1990) Hamilton Jacobi equation in infinite dimensions IV: Hamiltonians with unbounded linear terms. J Funct Anal 90:237–283
49. Crandall MG, Lions PL (1991) Hamilton Jacobi equation in infinite dimensions V: Unbounded linear terms and *B*-continuous solutions. J Funct Anal 97:417–465
50. Curtain RF, Weiss G (1989) Well posedness of triples of operators (in the sense of linear systems theory). Control and Estimation of Distributed Parameter Systems (Vorau, 1988). Internat. Ser. Numer. Math., vol. 91, Birkhäuser, Basel
51. Curtain RF, Zwart H (1995) An introduction to infinite-dimensional linear systems theory. Texts in Applied Mathematics, vol 21. Springer, New York
52. Da Prato G, Frankowska H (2007) Stochastic viability of convex sets. J Math Anal Appl 333(1):151–163
53. Dafermos CM(1970) Asymptotic stability inviscoelasticity. Arch Ration Mech Anal 37:297–308
54. Dafermos CM (1970) An abstract Volterra equation with applications to linear viscoelasticity. J Differ Equ 7:554–569
55. Engel KJ, R Nagel R (2000) One-parameter semigroups for linear evolution equations. Springer, New York
56. Eller M, Lagnese JE, Nicaise S (2002) Decay rates for solutions of a Maxwell system with nonlinear boundary damping. Comp Appl Math 21:135–165
57. Fabre C, Puel J-P, Zuazua E (1995) Approxiamte controllability of the semilinear heat equation. Proc Roy Soc Edinburgh Sect A 125(1):31–61
58. Fattorini HO, Russell DL (1971) Exact controllability theorems for linear parabolic equations in one space dimension. Arch Rat Mech Anal 4:272–292
59. Fattorini HO (1998) Infinite Dimensional Optimization and Control theory. Encyclopedia of Mathematics and its Applications, vol 62. Cambridge University Press, Cambridge
60. Fattorini HO (2005) Infinite dimensional linear control systems. North-Holland Mathematics Studies, vol 201. Elsevier Science B V, Amsterdam
61. Fernández-Cara E, Zuazua E (2000) Null and approximate controllability for weakly blowing up semilinear heat equations. Ann Inst H Poincaré Anal Non Linéaire 17(5):583–616
62. Fernández-Cara E, Zuazua E (2000) The cost approximate controllability for heat equations: The linear case. Adv Differ Equ 5:465–514
63. Fernández-Cara E, Guerrero S, Imanuvilov OY, Puel J-P (2004) Local exact controllability of the Navier–Stokes system. J Math Pures Appl 83(9–12):1501–1542
64. Fursikov A (2000) Optimal control of distributed systems. Theory and applications. Translation of Mathematical Monographs, vol 187. American Mathematical Society, Providence
65. Fursikov A, Imanuvilov OY (1996) Controllability of evolution

Copyrighted material

Diagrammatic Methods in Classical Perturbation Theory

GUIDO GENTILE
Dipartimento di Matematica, Università di Roma Tre, Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Examples
Trees and Graphical Representation
Small Divisors
Multiscale Analysis
Resummation
Generalizations
Conclusions and Future Directions
Bibliography

Glossary

Dynamical system Let $W \subseteq \mathbb{R}^N$ be an open set and $f\colon W \times \mathbb{R} \to \mathbb{R}^N$ be a smooth function. The ordinary differential equation $\dot{x} = f(x,t)$ on W defines a continuous dynamical system. A discrete dynamical system on W is defined by a map $x \to x' = F(x)$, with F depending smoothly on x.

Hamiltonian system Let $\mathcal{A} \subseteq \mathbb{R}^d$ be an open set and $\mathcal{H}\colon \mathcal{A} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ be a smooth function ($\mathcal{A} \times \mathbb{R}^d$ is called the phase space). Consider the system of ordinary differential equations $\dot{q}_k = \partial \mathcal{H}(q,p,t)/\partial p_k$, $\dot{p}_k = -\partial \mathcal{H}(q,p,t)/\partial q_k$, for $k = 1,\dots,d$. The equations are called Hamilton equations, and $\mathcal{H}$ is called a Hamiltonian function. A dynamical system described by Hamilton equations is called a Hamiltonian system.

Integrable system A Hamiltonian system is called integrable if there exists a system of coordinates $(\alpha, A) \in \mathbb{T}^d \times \mathbb{R}^d$, called angle-action variables, such that in these coordinates the motion is $(\alpha, A) \to (\alpha + \omega(A)t, A)$, for some smooth function $\omega(A)$. Hence in these coordinates the Hamiltonian function $\mathcal{H}$ depends only on the action variables, $\mathcal{H} = \mathcal{H}_0(A)$.

Invariant torus Given a continuous dynamical system we say that the motion occurs on an invariant d-torus if it takes place on a d-dimensional manifold and its position on the manifold is identified through a coordinate in $\mathbb{T}^d$. In an integrable Hamiltonian system all phase space is filled by invariant tori. In a quasi-integrable system the KAM theorem states that most of the invariant tori persist under perturbation, in the sense that the relative Lebesgue measure of the fraction of phase space filled by invariant tori tends to 1 as the perturbation tends to disappear. The persisting invariant tori are slight deformations of the unperturbed invariant tori.

Quasi-integrable system A quasi-integrable system is a Hamiltonian system described by a Hamiltonian function of the form $\mathcal{H} = \mathcal{H}_0(A) + \varepsilon f(\alpha, A)$, with (α, A) angle-action variables, ε a small real parameter and f periodic in its arguments α.

Quasi-periodic motion Consider the motion $\alpha \to \alpha + \omega t$ on $\mathbb{T}^2$, with $\omega = (\omega_1, \omega_2)$. If ω_1/ω_2 is rational, the motion is periodic, that is there exists $T > 0$ such that $\omega_1 T = \omega_2 T = 0 \bmod 2\pi$. If ω_1/ω_2 is irrational, the motion never returns to its initial value. On the other hand it densely fills $\mathbb{T}^2$, in the sense that it comes arbitrarily close to any point of $\mathbb{T}^2$. We say in that case that the motion is quasi-periodic. The definition extends to $\mathbb{T}^d$, $d > 2$: a linear motion $\alpha \to \alpha + \omega t$ on $\mathbb{T}^d$ is quasi-periodic if the components of ω are rationally independent, that is if $\omega \cdot \nu = \omega_1 \nu_1 + \cdots + \omega_d \nu_d = 0$ for $\nu \in \mathbb{Z}^d$ if and only if $\nu = 0$ ($a \cdot b$ is the standard scalar product between the two vectors a, b). More generally we say that a motion on a manifold is quasi-periodic if, in suitable coordinates, it can be described as a linear quasi-periodic motion. The vector ω is usually called the frequency or rotation vector.

Renormalization group By renormalization group one denotes the set of techniques and concepts used to study problems where there are some scale invariance properties. The basic mechanism consists in considering equations depending on some parameters and defining some transformations on the equations, including a suitable rescaling, such that after the transformation the equations can be expressed, up to irrelevant corrections, in the same form as before but with new values for the parameters.

Torus The 1-torus $\mathbb{T}$ is defined as $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, that is the set of real numbers defined modulo 2π (this means that x is identified with y if $x - y$ is a multiple of 2π). So it is the natural domain of an angle. One defines the d-torus $\mathbb{T}^d$ as a product of d 1-tori, that is $\mathbb{T}^d = \mathbb{T} \times \cdots \times \mathbb{T}$. For instance one can imagine $\mathbb{T}^2$ as a square with the opposite sides glued together.

Tree A graph is a collection of points, called nodes, and of lines which connect the nodes. A walk on the graph is a sequence of lines such that any two successive lines in the sequence share a node; a walk is nontrivial if it

Copyrighted material

contains at least one line. A tree is a planar graph with no closed loops, that is, such that there is no nontrivial walk connecting any node to itself. An oriented tree is a tree with a special node such that all lines of the tree are oriented toward that node. If we add a further oriented line connecting the special node to another point, called the root, we obtain a rooted tree (see Fig. 1 in Sect. "Trees and Graphical Representation").

Definition of the Subject

Recursive equations naturally arise whenever a dynamical system is considered in the regime of perturbation theory; for an introductory article on perturbation theory see ► Perturbation Theory. A classical example is provided by Celestial Mechanics, where perturbation series, known as Lindstedt series, are widely used; see Gallavotti [21] and ► Perturbation Theory in Celestial Mechanics.

A typical problem in Celestial Mechanics is to study formal solutions of given ordinary differential equations in the form of expansions in a suitable small parameter, the perturbation parameter. In the case of quasi-periodic solutions, the study of the series, in particular of its convergence, is made difficult by the presence of the small divisors – which will be defined later on. Under some non-resonance condition on the frequency vector, one can show that the series are well-defined to any order. The first proof of such a property was given by Poincaré [53], even if the convergence of the series remained an open problem up to the advent of KAM theory – an account can be found in Gallavotti [17] and in Arnold et al. [2]; see also ► Kolmogorov–Arnold–Moser (KAM) Theory. KAM is an acronym standing for Kolmogorov [47], Arnold [1] and Moser [51], who proved in the middle of last century the persistence of most of invariant tori for quasi-integrable systems.

Kolmogorov and Arnold proofs apply to analytic Hamiltonian systems, while Moser's approach deals also with the differentiable case; the smoothness condition on the Hamiltonian function was thereafter improved by Pöschel [54]. In the analytic case, the persisting tori turn out to be analytic in the perturbation parameter, as explicitly showed by Moser [52]. In particular, this means that the perturbation series not only are well-defined, but also converge. However, a systematic analysis with diagrammatic techniques started only recently after the pioneering, fundamental works by Eliasson [16] and Gallavotti [18], and were subsequently extended to many other problems with small divisors, including dynamical systems with infinitely many degrees of freedom, such as nonlinear partial differential equations, and non-Hamiltonian systems. Some of these extensions will be discussed in Sect. "Generalizations".

From a technical point of view, the diagrammatic techniques used in classical perturbation theory are strongly reminiscent of the Feynman diagrams used in quantum field theory: this was first pointed out by Gallavotti [18]. Also the multiscale analysis used to control the small divisors is typical of renormalization group techniques, which have been successfully used in problems of quantum field theory, statistical mechanics and classical mechanics; see Gallavotti [20] and Gentile & Mastropietro [38] for some reviews.

Note that there exist other renormalization group approaches to the study of dynamical systems, and of KAM-like problems in particular, different from that outlined in this article. By confining ourselves to the framework of problems of KAM-type, we can mention the paper by Bricmont et al. [11], which also stressed the similarity of the technique with quantum field theory, and the so called dynamical renormalization group method – see MacKay [50] – which recently produced rigorous proofs of persistence of quasi-periodic solutions; see for instance Koch [46] and Khanin et al. [45].

Introduction

Consider the ordinary differential equation on $\mathbb{R}^d$

$$Du = G(u) + \varepsilon F(u)\,, \tag{1}$$

where D is a pseudo-differential operator and G, F are real analytic functions. Assume that (1) admits a solution $u^{(0)}(t)$ for $\varepsilon = 0$, that is $Du^{(0)} = G(u^{(0)})$. The problem we are interested in is to investigate whether there exists a solution of (1) which reduces to $u^{(0)}$ as $\varepsilon \to 0$. For simplicity assume $G = 0$ in the following.

The first attempt one can try is to look for solutions in the form of power series in ε,

$$u(t) = \sum_{k=0}^{\infty} \varepsilon^k u^{(k)}(t)\,, \tag{2}$$

which, inserted into (1), when equating the left and right hand sides order by order, gives the list of recursive equations $Du^{(0)} = 0$, $Du^{(1)} = F(u^{(0)})$, $Du^{(2)} = \partial_u F(u^{(0)})u^{(1)}$, and so on. In general to order $k \geq 1$ one has

$$Du^{(k)} = \sum_{s=0}^{k-1} \frac{1}{s!} \partial_u^s F(u^{(0)}) \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} u^{(k_1)} \dots u^{(k_s)}\,, \tag{3}$$

where $\partial_u^s F$, the sth derivative of F, is a tensor with $s+1$ indices (s must be contracted with the vectors

Copyrighted material

$u^{(k_1)}, \dots, u^{(k_s)}$), and the term with $s = 0$ in the sum has to be interpreted as $F(u^{(0)})$ and appears only for $k = 1$.

For instance for $F(u) = u^3$ the first orders give

$$\begin{aligned} Du^{(1)} &= u^{(0)3} , \\ Du^{(2)} &= 3u^{(0)2}u^{(1)} , \\ Du^{(3)} &= 3u^{(0)2}u^{(2)} + 3u^{(0)}u^{(1)2} , \\ Du^{(4)} &= 3u^{(0)2}u^{(3)} + 6u^{(0)}u^{(1)}u^{(2)} + u^{(1)3} , \end{aligned} \tag{4}$$

as is easy to check.

If the operator D can be inverted then the recursions (3) provide an algorithm to compute the functions $u^{(k)}(t)$. In that case we say that (2) defines a *formal power series*: by this we mean that the functions $u^{(k)}(t)$ are well-defined for all $k \geq 0$. Of course, even if this can be obtained, there is still the issue of the convergence of the series that must be dealt with.

Examples

In this section we consider a few paradigmatic examples of dynamical systems which can be described by equations of the form (1).

A Class of Quasi-integrable Hamiltonian Systems

Consider the Hamiltonian system described by the Hamiltonian function

$$\mathcal{H}(\alpha, A) = \tfrac{1}{2}A^2 + \varepsilon f(\alpha) , \tag{5}$$

where $(\alpha, A) \in \mathbb{T}^d \times \mathbb{R}^d$ are angle-action variables, with $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, f is a real analytic function, 2π-periodic in each of its arguments, and $A^2 = A \cdot A$, if (here and henceforth) $\cdot$ denotes the standard scalar product in $\mathbb{R}^d$, that is $a \cdot b = a_1 b_1 + \cdots + a_d b_d$. Assume also for simplicity that f is a trigonometric polynomial of degree N.

The corresponding Hamilton equations are (we shorten $\partial_x = \partial/\partial_x$)

$$\begin{aligned} \dot\alpha &= \partial_A \mathcal{H}(\alpha, A) = A , \\ \dot A &= -\partial_\alpha \mathcal{H}(\alpha, A) = -\varepsilon \partial_\alpha f(\alpha) , \end{aligned}$$

which can be written as an equation involving only the angle variables:

$$\ddot\alpha = -\varepsilon \partial_\alpha f(\alpha) , \tag{6}$$

which is of the form (1) with $u = \alpha$, $G = 0$, $F = -\partial_\alpha f$, and $D = \mathrm{d}^2/\mathrm{d}t^2$.

For $\varepsilon = 0$, $\alpha^{(0)}(t) = \alpha_0 + \omega t$ is a solution of (6) for any choice of $\alpha_0 \in \mathbb{T}^d$ and $\omega \in \mathbb{R}^d$. Take for simplicity $\alpha_0 = 0$: we shall see that this choice makes sense. We say that for $\varepsilon = 0$ the Hamiltonian function (5) describes a system of d *rotators*.

We call ω the *frequency vector*, and we say that ω is irrational if its components are rationally independent, that is if $\omega \cdot \nu = 0$ for $\nu \in \mathbb{Z}^d$ if and only if $\nu = 0$. For irrational ω the solution $\alpha^{(0)}(t)$ describes a *quasi-periodic motion* with frequency vector ω, and it densely fills $\mathbb{T}^d$.

Then (3) becomes

$$\begin{aligned} \ddot\alpha^{(k)} &= -\big[\varepsilon f(\alpha)\big]^{(k)} \\ &:= -\sum_{s=0}^{k-1} \frac{1}{s!} \partial_\alpha^{s+1} f(\omega t) \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} \alpha^{(k_1)} \dots \alpha^{(k_s)} . \end{aligned} \tag{7}$$

We look for a quasi-periodic solution of (6), that is a solution of the form $\alpha(t) = \omega t + h(\omega t)$, with $h(\omega t) = O(\varepsilon)$. We call h the *conjugation function*, as it "conjugates" (that is, maps) the perturbed solution $\alpha(t)$ to the unperturbed solution ωt. In terms of the function h (6) becomes

$$\ddot h = -\varepsilon \partial_\alpha f(\omega t + h) , \tag{8}$$

where ∂_α denotes derivative with respect to the argument. Then (8) can be more conveniently written in Fourier space, where the operator D acts as a multiplication operator.

If we write

$$h(\omega t) = \sum_{\nu \in \mathbb{Z}^d} \mathrm{e}^{i\omega\cdot\nu t} h_\nu , \qquad h_\nu = \sum_{k=1}^{\infty} \varepsilon^k h_\nu^{(k)} , \tag{9}$$

and insert (9) into (8) we obtain

$$\begin{aligned} (\omega \cdot \nu)^2 h_\nu^{(k)} &= \big[\varepsilon \partial_\alpha f(\alpha)\big]_\nu^{(k)} \\ &:= \sum_{s=0}^{k-1} \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{\nu_0+\nu_1+\cdots+\nu_s=\nu \\ \nu_i \in \mathbb{Z}^n}} \frac{1}{s!} \\ &\quad \cdot (i\nu_0)^{s+1} f_{\nu_0} h_{\nu_1}^{(k_1)} \dots h_{\nu_s}^{(k_s)} . \end{aligned} \tag{10}$$

These equations are well-defined to all orders provided $[\varepsilon \partial_\alpha f(\alpha)]_\nu^{(k)} = 0$ for all ν such that $\omega \cdot \nu = 0$. If ω is an irrational vector we need $[\varepsilon \partial_\alpha f(\alpha)]_0^{(k)} = 0$ for the equations to be well-defined. In that case the coefficients $h_0^{(k)}$ are left undetermined, and we can fix them arbitrarily to vanish (which is a convenient choice).

We shall see that under some condition on ω a quasi-periodic solution $\alpha(t)$ exists, and densely fills a d-dimensional manifold. The analysis carried out above for $\alpha_0 = 0$ can be repeated unchanged for all values of $\alpha_0 \in \mathbb{T}^d$: α_0

Copyrighted material

represents the *initial phase* of the solution, and by varying α_0 we cover all the manifold. Such a manifold can be parametrized in terms of α_0, so it represents an invariant torus for the perturbed system.

A Simplified Model with No Small Divisors

Consider the same equation as (8) with $D = \mathrm{d}^2/\mathrm{d}t^2$ replaced by -1, that is

$$h = \varepsilon \partial_\alpha f(\omega t + h) \,. \tag{11}$$

Of course in this case we no longer have a differential equation; still, we can look again for quasi-periodic solutions $h(\omega t) = O(\varepsilon)$ with frequency vector ω. In such a case in Fourier space we have

$$\begin{aligned} h_\nu^{(k)} &= \left[\varepsilon \partial_\alpha f(\alpha)\right]_\nu^{(k)} \\ &:= \sum_{s=0}^{k-1} \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{\nu_0+\nu_1+\cdots+\nu_s=\nu \\ \nu_i \in \mathbb{Z}^n}} \frac{1}{s!} (i\nu_0)^{s+1} \\ &\quad \cdot f_{\nu_0} h_{\nu_1}^{(k_1)} \dots h_{\nu_s}^{(k_s)} \,. \end{aligned} \tag{12}$$

For instance if $d = 1$ and $f(\alpha) = \cos\alpha$ the equation, which is known as the *Kepler equation*, can be explicitly solved by the Lagrange inversion theorem [55], and gives

$$h_\nu^{(k)} = \begin{cases} \dfrac{-i(-1)^{k+(\nu-k)/2}}{2^k((k-\nu)/2)!((k+\nu)/2)!} \,, & |\nu| \leq k, \\ & \nu + k \text{ even} \,, \\ 0, & \text{otherwise} \,. \end{cases} \tag{13}$$

We shall show in Sect. "Small Divisors" that a different derivation can be provided by using the forthcoming diagrammatic techniques.

The Standard Map

Consider the finite difference equation

$$D\alpha = -\varepsilon \sin\alpha \,, \tag{14}$$

on $\mathbb{T}$, where now D is defined by

$$D\alpha(\psi) := 2\alpha(\psi) - \alpha(\psi+\omega) - \alpha(\psi-\omega) \,. \tag{15}$$

By writing $\alpha = \psi + h(\psi)$, (14) becomes

$$Dh = -\varepsilon \sin(\psi + h) \,, \tag{16}$$

which is the functional equation that must be solved by the *conjugation function* of the *standard map*

$$\begin{cases} x' = x + y + \varepsilon \sin x \,, \\ y' = y + \varepsilon \sin x \,. \end{cases} \tag{17}$$

In other words, by writing $x = \psi + h(\psi)$ and $y = \omega + h(\psi) - h(\psi-\omega)$, with (ψ, ω) solving (17) for $\varepsilon = 0$, that is $(\psi', \omega') = (\psi+\omega, \omega)$, we obtain a closed-form equation for h, which is exactly (16).

In Fourier space the operator D acts as $D\colon \mathrm{e}^{i\nu\psi} \to 4\sin^2(\omega\nu/2)\mathrm{e}^{i\nu\psi}$, so that, by expanding h according to (9), we can write (16) as

$$\begin{aligned} h_\nu^{(k)} = \frac{1}{4\sin^2(\omega\nu/2)} \sum_{s=0}^{k-1} \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{\nu_0+\nu_1+\cdots+\nu_s=\nu \\ \nu_i \in \mathbb{Z}}} \frac{1}{s!} \\ \cdot (i\nu_0)^{s+1} f_{\nu_0} h_{\nu_1}^{(k_1)} \dots h_{\nu_s}^{(k_s)} \,, \end{aligned} \tag{18}$$

where $\nu_0 = \pm 1$ and $f_{\pm 1} = 1/2$.

Note that (17) is a discrete dynamical system. However, when passing to Fourier space, (18) acquires the same form as for the continuous dynamical systems previously considered, simply with a different kernel for D. In particular if we replace D with 1 we recover the Kepler equation.

The number ω is called the *rotation number*. We say that ω is irrational if the vector $(2\pi, \omega)$ is irrational according to the previous definition.

Trees and Graphical Representation

Take ω to be irrational. We study the recursive equations

$$\begin{cases} h_\nu^{(k)} = g(\omega\cdot\nu)\left[\varepsilon\partial_\alpha f(\alpha)\right]_\nu^{(k)} \,, & \nu \neq 0 \,, \\ \left[\varepsilon\partial_\alpha f(\alpha)\right]_0^{(k)} = 0 \,, & \nu = 0 \,, \end{cases} \tag{19}$$

where the form of g depends on the particular model we are investigating. Hence one has either $g(\omega\cdot\nu) = (\omega\cdot\nu)^{-2}$ or $g(\omega\cdot\nu) = 1$ or $g(\omega\cdot\nu) = (2\sin(\omega\nu/2))^{-2}$ according to models described in Sect. "Examples".

For $\nu \neq 0$ we have equations which express the coefficients $h_\nu^{(k)}$, $\nu \in \mathbb{Z}^d$, in terms of the coefficients $h_\nu^{(k')}$, $\nu \in \mathbb{Z}^d$, with $k' < k$, provided the equations for $\nu = 0$ are satisfied for all $k \geq 1$. Recursive equations, such as (19), naturally lead to a graphical representation in terms of trees.

Trees

A connected *graph* $\mathcal{G}$ is a collection of points (nodes) and lines connecting all of them. Denote with $N(\mathcal{G})$ and $L(\mathcal{G})$ the set of nodes and the set of lines, respectively. A path between two nodes is the minimal subset of $L(\mathcal{G})$ connecting the two nodes. A graph is planar if it can be drawn in a plane without graph lines crossing.

A *tree* is a planar graph $\mathcal{G}$ containing no closed loops. Consider a tree $\mathcal{G}$ with a single special node v_0: this introduces a natural partial ordering on the set of lines and nodes, and one can imagine that each line carries an arrow

Copyrighted material

pointing toward the node v_0. We add an extra oriented line ℓ_0 exiting the special node v_0; the added line will be called the *root line* and the point it enters (which is not a node) will be called the *root* of the tree. In this way we obtain a *tree* θ defined by $N(\theta) = N(\mathcal{G})$ and $L(\theta) = L(\mathcal{G}) \cup \ell_0$. A *labeled tree* is a rooted tree θ together with a label function defined on the sets $L(\theta)$ and $N(\theta)$.

We call *equivalent* two rooted trees which can be transformed into each other by continuously deforming the lines in the plane in such a way that the lines do not cross each other. We can extend the notion of equivalence also to labeled trees, by considering equivalent two labeled trees if they can be transformed into each other in such a way that the labels also match. In the following we shall deal mostly with nonequivalent labeled trees: for simplicity, where no confusion can arise, we call them just trees.

Given two nodes $v, w \in N(\theta)$, we say that $w \prec v$ if v is on the path connecting w to the root line. We can identify a line ℓ through the node v it exits by writing $\ell = \ell_v$.

We call *internal nodes* the nodes such that there is at least one line entering them, and *end-points* the nodes which have no entering line. We denote with $L(\theta)$, $V(\theta)$ and $E(\theta)$ the set of lines, internal nodes and end-points, respectively. Of course $N(\theta) = V(\theta) \cup E(\theta)$.

The number of unlabeled trees with k nodes (and hence with k lines) is bounded by 2^{2k}, which is a bound on the number of random walks with $2k$ steps [38].

For each node v denote by $S(v)$ the set of the lines entering v and set $s_v = |S(v)|$. Hence $s_v = 0$ if v is an end-node, and $s_v \geq 1$ if v is an internal node. One has

$$\sum_{v \in N(\theta)} s_v = \sum_{v \in V(\theta)} s_v = k - 1 \,; \tag{20}$$

this can be easily checked by induction on the order of the tree. An example of unlabeled tree is represented in Fig. 1.

For further details on graphs and trees we refer to the literature; cf. for instance Harary [43].

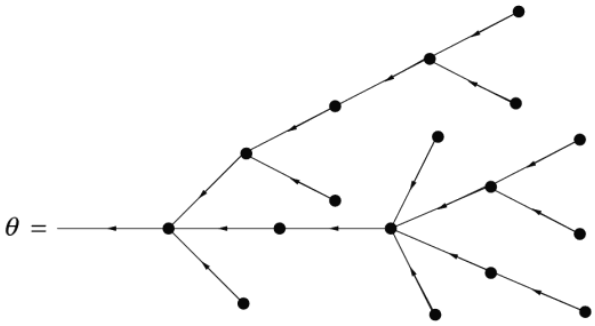
Labels and Diagrammatic Rules

We associate with each node $v \in N(\theta)$ a *mode* label $\nu_v \in \mathbb{Z}^d$, and with each line $\ell \in L(\theta)$ a *momentum* label $\nu_\ell \in \mathbb{Z}^d$, with the constraint

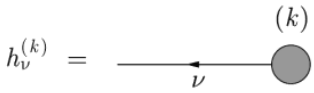
$$\nu_{\ell_v} = \sum_{\substack{w \in N(\theta) \\ w \preceq v}} \nu_w = \nu_v + \sum_{\ell \in S(v)} \nu_\ell \,, \tag{21}$$

which represents a *conservation rule* for each node.

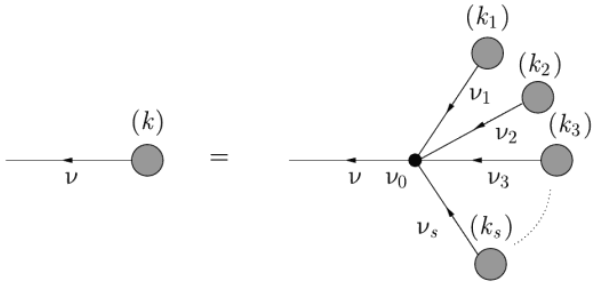
Call $\mathcal{T}_{k,\nu}$ the set of all trees θ with k nodes and momentum ν associated with the root line. We call k and ν the order and the momentum of θ, respectively.



Diagrammatic Methods in Classical Perturbation Theory, Figure 1
An unlabeled tree with 17 nodes



Diagrammatic Methods in Classical Perturbation Theory, Figure 2
Graph element



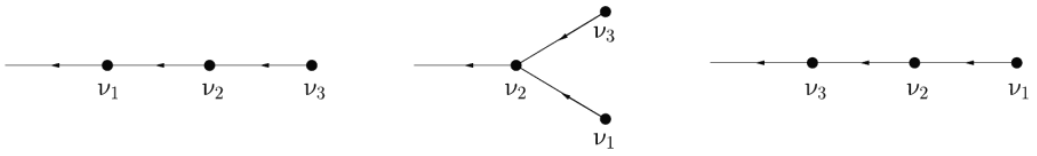
Diagrammatic Methods in Classical Perturbation Theory, Figure 3
Graphical representation of the recursive equations

We want to show that trees naturally arise when studying (19). Let $h_\nu^{(k)}$ be represented with the graph element in Fig. 2 as a line with label ν exiting from a ball with label (k).

Then we can represent (19) graphically as depicted in Fig. 3. Simply represent each factor $h_{\nu_i}^{(k_i)}$ on the right hand side as a graph element according to Fig. 2. The lines of all such graph elements enter the same node v_0. This is a graphical expedient to recall the conservation rule: the momentum ν of the root line is the sum of the mode label ν_0 of the node v_0 plus the sum of the momenta of the lines entering v_0.

The first few orders $k \leq 4$ are as depicted in Fig. 4. For each node the conservation rule (21) holds: for instance for $k = 2$ one has $\nu = \nu_1 + \nu_2$, for $k = 3$ one has $\nu = \nu_1 + \nu_{\ell_1}$ and $\nu_{\ell_1} = \nu_2 + \nu_3$ in the first tree and $\nu = \nu_1 + \nu_2 + \nu_3$ in

Copyrighted material



Diagrammatic Methods in Classical Perturbation Theory, Figure 6
Trees to be considered together to prove that $[\varepsilon\partial f(\alpha)]_0^3 = 0$

ily provide an exact expression for the coefficients $h_\nu^{(k)}$: all the computational difficulties reduce to a combinatorial check, which can be found in Gentile & van Erp [42], and the formula (13) is recovered.

However for the models where $g_\ell \neq 1$, the situation is much more involved: the propagators can be arbitrarily close to zero for ν large enough. This is the so-called *small divisor problem*. The series (25) is formally well–defined, assuming only an irrationality condition on ω. But to prove the convergence of the series, we need a stronger condition. For instance one can require the *standard Diophantine condition*

$$|\omega\cdot\nu| > \frac{\gamma}{|\nu|^\tau} \qquad \forall \nu \neq 0\,, \tag{29}$$

for suitable positive constants γ and τ. For fixed $\tau > d-1$, the sets of vectors which satisfy (29) for some constant $\gamma > 0$ has full Lebesgue measure in $\mathbb{R}^d$ [17]. We can also impose a weaker condition, known as the *Bryuno condition*, which can be expressed by requiring

$$\mathcal{B}(\omega) := \sum_{k=0}^{\infty} \frac{1}{2^k} \log \frac{1}{\min_{0<|\nu|\le 2^k} |\omega\cdot\nu|} < \infty\,. \tag{30}$$

We call *Diophantine vectors* and *Bryuno vectors* the vectors satisfying (29) and (30), respectively. Note that any Diophantine vector satisfies (30).

If we assume only analyticity on f – that is, we remove the assumption that f be a trigonometric polynomial, – then we need a Diophantine condition, such as (29) or (30), also to show that the series (25) are well-defined as formal power series: the condition is needed, as one can easily check, in order to sum over the mode labels.

In the case of the standard map $|\omega\cdot\nu|$ must be replaced with $\min_{p\in\mathbb{Z}} |\omega\nu - p|$, and the function $\mathcal{B}(\omega)$ can be expressed in terms of the *best approximants* of the number ω [15].

For the models with small divisors considered in Sect. "Examples" convergence of the series (25) can be proved for ω satisfying (29) or (30). But this requires more work, and, in particular, a detailed discussion of the product of propagators in (24). In the following we shall consider the case of Diophantine vectors, by referring to the bibliography for the Bryuno vectors (cf. Sect. "Generalizations").

Multiscale Analysis

Consider explicitly the case $g(\omega\cdot\nu) = (\omega\cdot\nu)^{-2}$ in (19) and ω satisfying (29). We can introduce a label characterizing the size of the propagator: we say that $\nu \in \mathbb{Z}^d$ is on *scale*

$$\begin{cases} n \ge 1\,, & \text{if } 2^{-n}\gamma \le |\omega\cdot\nu| < 2^{-(n-1)}\gamma\,, \\ n = 0\,, & \text{if } \gamma \le |\omega\cdot\nu|\,, \end{cases} \tag{31}$$

where γ is the constant appearing in (29), and we say that a line ℓ has a scale label $n_\ell = n$ if ν_ℓ is on scale n. If $\mathfrak{N}_n(\theta)$ denotes the number of lines $\ell \in L(\theta)$ with scale $n_\ell = n$, then we can bound in (23)

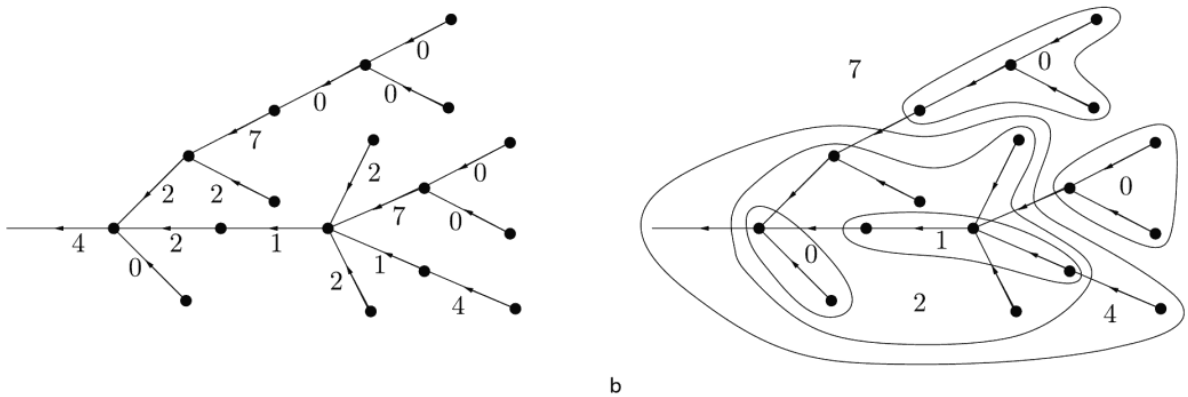
$$\Big|\prod_{\ell\in L(\theta)} g_\ell\Big| \le \gamma^{-2k} \prod_{n=0}^{\infty} 2^{2n\mathfrak{N}_n(\theta)}\,, \tag{32}$$

so that the problem is reduced to bounding $\mathfrak{N}_n(\theta)$.

The product of propagators gives problems when the small divisors "accumulate". To make more precise the idea of accumulation we introduce the notion of cluster. Once all lines of a tree θ have been given their scale labels, for any $n \ge 0$ we can identify the maximal connected sets of lines with scale not larger than n. If at least one among such lines has scale equal to n we say that the set is a *cluster* on scale n. For instance consider the tree in Fig. 1, and assign the mode labels to the nodes: this uniquely fixes the momenta, and hence the scale labels, of the lines. Suppose that we have found the scales as in Fig. 7a. Then a cluster decomposition as in Fig. 7b follows. Given a cluster T call $L(T)$ the set of lines of θ contained in T, and denote by $N(T)$ the set of nodes connected by such lines. We define $k_T = |N(T)|$ the order of the cluster T.

Any cluster has either one or no exiting line, and can have an arbitrary number of entering lines. We call *self-energy clusters* the clusters which have one exiting line and only one entering line and are such that both lines have the same momentum. This means that if T is a self-energy

Copyrighted material



Diagrammatic Methods in Classical Perturbation Theory, Figure 7
Example of clusters with their scales

cluster and ℓ_1 and ℓ_2 are the lines entering and exiting T, respectively, then $\nu_{\ell_1} = \nu_{\ell_2}$, so that

$$\sum_{\nu \in N(T)} \nu_\nu = 0 \,. \tag{33}$$

By definition of scale, both lines ℓ_1 and ℓ_2 have the same scale, say n, and that, by definition of cluster, one has $n_\ell < n$ for all $\ell \in L(T)$.

We define the *value* of the self-energy cluster T whose entering line has momentum ν as the matrix

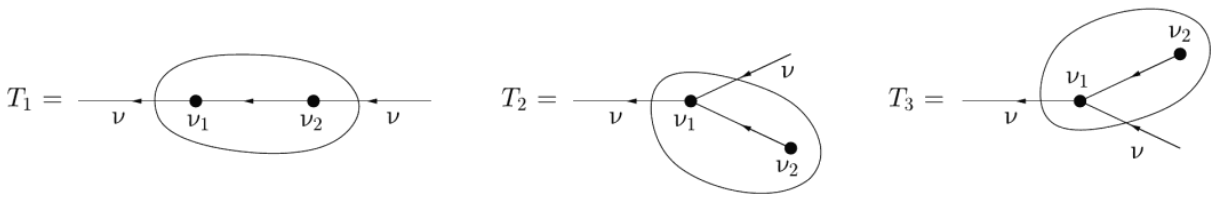
$$\mathcal{V}_T(\omega \cdot \nu) := \Big(\prod_{\nu \in N(T)} F_\nu \Big) \Big(\prod_{\ell \in L(T)} g_\ell \Big) \,. \tag{34}$$

where all the indices of the node factors must be contracted except those associated with the line ℓ_1 entering T and with the line ℓ_2 exiting T.

We can extend the notion of self-energy clusters also to a single node, by saying that ν is a self-energy cluster if $s_\nu = 1$ and the line entering ν has the same momentum as the exiting line. In that case (34) has to be interpreted as $\mathcal{V}_T(\omega \cdot \nu) = F_\nu$: in particular it is independent of $\omega \cdot \nu$.

The simplest self-energy cluster one can think of consists of only one node ν, but then (33) implies $\nu_\nu = 0$, so that $F_\nu = 0$, see (22), hence the corresponding value is zero. Thus the simplest non-trivial self-energy clusters contain at least two nodes, and are represented by the clusters T_1, T_2 and T_3 of Fig. 8. In all cases one has $\nu_1 + \nu_2 = 0$. By using the definition (34) one has $\mathcal{V}_{T_1}(x) = (i\nu_1)^2 f_{\nu_1} (i\nu_2)^2 f_{\nu_2} g(\omega \cdot \nu_2 + x)$ and $\mathcal{V}_{T_2}(x) = \mathcal{V}_{T_3}(x) = (i\nu_1)^3 f_{\nu_1} (i\nu_2) f_{\nu_2} g(\omega \cdot \nu_2)/2$, with $x = \omega \cdot \nu$. Hence for $x = 0$ the sum of the three values gives zero. It is not difficult to see that also the first derivatives of the values of all self-energy clusters of order 2 sum up to zero, that is the sum of the values of all possible self-energy clusters of order $k = 2$ gives zero up to order x^2. (The only *caveat* is that, if we want to derive $\mathcal{V}_T(x)$, the sharp multiscale decomposition in (31) can be a little annoying, hence it can be more convenient to replace it with a smooth decomposition through C^∞ compact support functions; we refer to the literature for details). This property generalizes to all orders k, and the underlying cancellation mechanism is essentially the same that ensures the validity of the second relation in (19).

The reason why it is important to introduce the self-energy clusters is that if we could neglect them then the product of small divisors would be controlled. Indeed, let us denote by $\mathfrak{N}_n(\theta)$ the number of lines on scale n which do exit a self-energy cluster, and set $\mathfrak{N}^*_n(\theta) = \mathfrak{N}_n(\theta) -$



Diagrammatic Methods in Classical Perturbation Theory, Figure 8
Self-energy clusters of order $k = 2$

Copyrighted material

$\mathfrak{R}_n(\theta)$. Then an important result, known as the *Siegel–Bryuno lemma*, is that

$$\mathfrak{R}_n^*(\theta) \le c2^{-n/\tau}k\,, \tag{35}$$

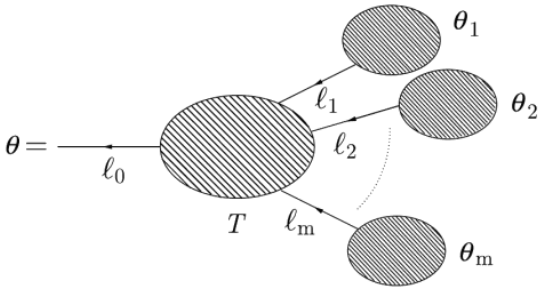
for some constant c, where k is the order of θ and τ is the Diophantine exponent in (29).

The bound (35) follows from the fact that if $\mathfrak{R}_n^*(\theta) \neq 0$ then $\mathfrak{R}_n^*(\theta) \le E(n,k) := 2Nk2^{-(n-2)/\tau} - 1$, which can be proved by induction on k as follows. Given a tree θ let ℓ_0 be its root line, let $\ell_1, \dots, \ell_m$, $m \ge 0$, be the lines on scales $\ge n$ which are the closest to ℓ_0, and let $\theta_1, \dots, \theta_m$ the trees with root lines $\ell_1, \dots, \ell_m$, respectively (cf. Fig. 9 – note that by construction all lines ℓ in the subgraph T have scales $n_\ell < n$, so that if $n_{\ell_0} \ge n$ then T is necessarily a cluster). If either ℓ_0 is not on scale n or it is on scale n but exits a self-energy cluster then $\mathfrak{R}_n^*(\theta) = \mathfrak{R}_n^*(\theta_1) + \cdots + \mathfrak{R}_n^*(\theta_m)$ and the bound $\mathfrak{R}_n^*(\theta) \le E(n,k)$ follows by the inductive hypothesis. If ℓ_0 does not exit a self-energy cluster and $n_{\ell_0} = n$ then $\mathfrak{R}_n^*(\theta) = 1 + \mathfrak{R}_n^*(\theta_1) + \cdots + \mathfrak{R}_n^*(\theta_m)$, and the lines $\ell_1, \dots, \ell_m$ enter a cluster T with $k_T = k - (k_1 + \cdots + k_m)$, where $k_1, \dots, k_m$ are the orders of $\theta_1, \dots, \theta_m$, respectively. If $m \ge 2$ the bound $\mathfrak{R}_n^*(\theta) \le E(n,k)$ follows once more by the inductive hypothesis. If $m = 0$ then $\mathfrak{R}_n^*(\theta) = 1$; on the other hand for ℓ_0 to be on scale $n_{\ell_0} = n$ one must have $|\omega \cdot \nu_{\ell_0}| < 2^{-n+1}\gamma$ (see (31)), which, by the Diophantine condition (29), implies $Nk \ge |\nu_{\ell_0}| > 2^{(n-1)/\tau}$, hence $E(n,k) > 1$. If $m = 1$ call ν_1 and ν_2 the momenta of the lines ℓ_0 and ℓ_1, respectively. By construction T cannot be a self-energy cluster, hence $\nu_1 \neq \nu_2$, so that, by the Diophantine condition (29),

$$2^{-n+2}\gamma \ge |\omega \cdot \nu_1| + |\omega \cdot \nu_2| \ge |\omega \cdot (\nu_1 - \nu_2)| > \frac{\gamma}{|\nu_1 - \nu_2|^\tau}\,, \tag{36}$$

because $n_{\ell_0} = n$ and $n_{\ell_1} \ge n$. Thus, one has

$$Nk_T \ge \sum_{\nu \in N(T)} |\nu_\nu| \ge |\nu_1 - \nu_2| > 2^{(n-2)/\tau}\,, \tag{37}$$



Diagrammatic Methods in Classical Perturbation Theory, Figure 9
Construction for the proof of the Siegel–Bryuno lemma

hence T must contain "many nodes". In particular, one finds also in this case $\mathfrak{R}_n^*(\theta) = 1 + \mathfrak{R}_n^*(\theta_1) \le 1 + E(n,k_1) \le 1 + E(n,k) - E(n,k_T) \le E(n,k)$, where we have used that $E(n,k_T) \ge 1$ by (37).

The argument above shows that small divisors can accumulate only by allowing self-energy clusters. That accumulation really occurs is shown by the example in Fig. 10, where a tree θ of order k containing a chain of p self-energy clusters is depicted. Assume for simplicity that $k/3$ is an integer: then if $p = k/3$ the subtree θ_1 with root line ℓ is of order $k/3$. If the line ℓ entering the rightmost self-energy cluster T_p has momentum ν, also the lines exiting the p self-energy clusters have the same momentum ν. Suppose that $|\nu| \approx Nk/3$ and $|\omega \cdot \nu| \approx \gamma/|\nu|^\tau$ (this is certainly possible for some ν). Then the value of the tree θ grows like $a_1^k(k!)^{a_2}$, for some constants a_1 and a_2: a bound of this kind prevents the convergence of the perturbation series (25).

If no self-energy clusters could occur (so that $\mathfrak{R}_n(\theta) = 0$) the Siegel–Bryuno lemma would allow us to bound in (32)

$$\prod_{n=0}^{\infty} 2^{2n\mathfrak{R}_n(\theta)} = \prod_{n=0}^{\infty} 2^{2n\mathfrak{R}_n^*(\theta)} \le \exp\left(C_1 k \sum_{n=0}^{\infty} n2^{-n/\tau}\right) \le C_2^k\,, \tag{38}$$

for suitable constants C_1 and C_2. In that case convergence of the series for $|\varepsilon| < \varepsilon_0$ would follow, with ε_0 defined as in (26) with $C_0 = \gamma^2/C_2$. However, there are self-energy clusters and they produce factorials, as the example in Fig. 10 shows, so that we have to deal with them.

Resummation

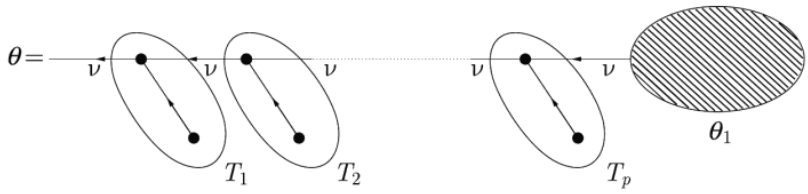
Let us come back to the Eq. (10). If we expand $g(\omega \cdot \nu)[\varepsilon\partial_\alpha f(\alpha)]_\nu^{(k)}$ in trees according to the diagrammatic rules described in Sect. "Trees and Graphical Representation", we can distinguish between contributions in which the root line exits a self-energy cluster T, that we can write as

$$g(\omega \cdot \nu) \sum_{T:\, k_T < k} \mathcal{V}_T(\omega \cdot \nu) h_\nu^{(k-k_T)}\,, \tag{39}$$

and all the other contributions, that we denote by $[\varepsilon\partial_\alpha f(\alpha)]_\nu^{(k)*}$. We can shift the contributions (39) to the left hand side of (10) and divide by $g(\omega \cdot \nu)$, so to obtain

$$D(\omega \cdot \nu)h_\nu^{(k)} - \sum_{T:\, k_T < k} \mathcal{V}_T(\omega \cdot \nu) h_\nu^{(k-k_T)} = [\varepsilon\partial_\alpha f(\omega t + h)]_\nu^{(k)*}. \tag{40}$$

Copyrighted material



Diagrammatic Methods in Classical Perturbation Theory, Figure 10
Example of accumulation of small divisors because of the self-energy clusters

where $D(\omega\cdot\nu) = 1/g(\omega\cdot\nu) = (\omega\cdot\nu)^2$. By summing over k and setting

$$M(\omega\cdot\nu;\varepsilon) = \sum_{k=1}^{\infty}\varepsilon^k \sum_{T:\, k_T=k} \mathcal{V}_T(\omega\cdot\nu)\,, \tag{41}$$

then (40) gives

$$\begin{aligned} &\mathcal{D}(\omega\cdot\nu)h_\nu = [\varepsilon\partial_\alpha f(\omega t+h)]^*_\nu,\\ &\mathcal{D}(\omega\cdot\nu) := D(\omega\cdot\nu) - M(\omega\cdot\nu;\varepsilon)\,. \end{aligned} \tag{42}$$

The motivation for proceeding in this way is that, at the price of changing the operator D into $\mathcal{D}$, hence of changing the propagators, lines exiting self-energy clusters no longer appear. Therefore, in the tree expansion of the right hand side of the equation, we have eliminated the self-energy clusters, that is the source of the problem of accumulation of small divisors.

Unfortunately the procedure described above has a problem: $M(\omega\cdot\nu;\varepsilon)$ itself is a sum of self-energy clusters, which can still contain some other self-energy clusters on lower scales. So finding a good bound for $M(\omega\cdot\nu;\varepsilon)$ could have the same problems as for the values of the trees.

To deal with such a difficulty we modify the prescription by proceeding recursively, in the following sense. Let us start from the momenta ν which are on scale $n = 0$. Since there are no self-energy clusters with exiting line on scale $n = 0$, for such ν one has $M(\omega\cdot\nu;\varepsilon) = 0$. Next, we consider the momenta ν which are on scale $n = 1$, and we can write (42), where now all self-energy clusters T whose values contribute to $M(\omega\cdot\nu;\varepsilon)$ cannot contain any self-energy clusters, because the lines $\ell \in T$ are on scale $n_\ell = 0$. Then, we consider the momenta ν which are on scale $n = 2$: again all the self-energy clusters contributing to $M(\omega\cdot\nu;\varepsilon)$ do not contain any self-energy clusters, because the lines on scale $n = 0, 1$ cannot exit self-energy clusters by the construction of the previous step, and so on. The conclusion is that we have obtained a different expansion for $h(\omega t)$, that we call a *resummed series*,

$$h(\omega t) = \sum_{\nu\in\mathbb{Z}^d} e^{i\omega\cdot\nu t}h_\nu\,, \qquad h_\nu = \sum_{k=1}^{\infty}\varepsilon^k h^{[k]}_\nu(\varepsilon)\,, \tag{43}$$

where the self-energy clusters do not appear any more in the tree expansion and the propagators must be defined recursively: the propagator g_ℓ of a line ℓ on scale $n_\ell = n$ and momentum $\nu_\ell = \nu$ is the matrix

$$g_\ell := g^{[n]}(\omega\cdot\nu;\varepsilon) = \big(D(\omega\cdot\nu) - \mathcal{M}^{[n]}(\omega\cdot\nu;\varepsilon)\big)^{-1}\,, \tag{44}$$

with

$$\mathcal{M}^{[n]}(\omega\cdot\nu;\varepsilon) := \sum_{T:\, n_T<n} \varepsilon^{k_T}\mathcal{V}_T(\omega\cdot\nu)\,, \tag{45}$$

where the value $\mathcal{V}_T(\omega\cdot\nu)$ is written in accord with (34), with all the lines $\ell' \in L(T)$ on scales $n_{\ell'} < n$ and the corresponding propagators $g_{\ell'}$ expressed in terms of matrices $\mathcal{M}^{[n_{\ell'}]}(\omega\cdot\nu_{\ell'};\varepsilon)$ as in (44). By construction, the new propagators depend on ε, so that the coefficients $h^{[k]}_\nu(\varepsilon)$ depend explicitly on ε: hence (43) is not a power series expansion.

The coefficients $h^{[k]}_\nu(\varepsilon)$ still admit a tree expansion

$$\begin{aligned} &h^{[k]}_\nu(\varepsilon) = \sum_{\theta\in\mathcal{T}^{\mathcal{R}}_{k,\nu}} \mathrm{Val}(\theta), \quad \nu\neq 0, k\geq 1\,,\\ &\mathrm{Val}(\theta) := \Big(\prod_{\nu\in N(\theta)} F_\nu\Big)\Big(\prod_{\ell\in L(\theta)} g^{[n_\ell]}(\omega\cdot\nu_\ell;\varepsilon)\Big)\,, \end{aligned} \tag{46}$$

which replaces (24). In particular $\mathcal{T}^{\mathcal{R}}_{k,\nu}$ is defined as the set of *renormalized trees* of order k and momentum ν, where "renormalized" means that the trees do not contain any self-energy clusters. Now the propagators are matrices: the contractions of the indices in (46) yields that if ℓ connects ν to ν' the indices of F_ν and $F_{\nu'}$ associated with ℓ must be equal to the column and row indices of g_ℓ, respectively.

Since for any tree $\theta \in \mathcal{T}^{\mathcal{R}}_{k,\nu}$ one has $\mathfrak{N}_n(\theta) = \mathfrak{N}^*_n(\theta)$, we can bound the product of propagators according to (36) *provided the propagators on scale n can still be bounded proportionally to* 2^{2n}. This is certainly not obvious, because of the extra term $\mathcal{M}^{[n]}(\omega\cdot\nu;\varepsilon)$ appearing in (44).

It is a remarkable cancellation that $\mathcal{M}^{[n]}(x;\varepsilon)$ vanishes up to second order, that is $\mathcal{M}^{[n]}(x;\varepsilon) = O(x^2)$, so that, by

Copyrighted material

taking into account also that $\mathcal{V}_T(x) \neq 0$ requires $k_T \geq 2$, we can write, for some constant C,

$$\mathcal{M}^{[n]}(x;\varepsilon) = \varepsilon^2 x^2 \overline{\mathcal{M}}^{[n]}(x;\varepsilon)\,, \qquad \left\| \overline{\mathcal{M}}^{[}n](x;\varepsilon) \right\| \leq C\,, \tag{47}$$

where $\|\cdot\|$ denotes – say – the uniform norm. The cancellation leading to (47) can be proved as discussed in Sect. "Multiscale Analysis" – where it has been explicitly proved for $k = 2$ in the absence of resummation. Now the propagators are matrices, but one can prove (by induction on n) that $\mathcal{M}^{[n]}(x;\varepsilon) = (\mathcal{M}^{[n]}(-x;\varepsilon))^T = (\mathcal{M}^{[n]}(x;\varepsilon))^\dagger$, with T and $\dagger$ denoting transposition and adjointness, respectively, and this is enough to see that the same cancellation mechanism applies [23].

Thus (47) implies that

$$\left\| g^{[n]}(x;\varepsilon) \right\| = \left\| \left(x^2 - \varepsilon^2 x^2 \overline{\mathcal{M}}^{[n]}(x;\varepsilon) \right)^{-1} \right\| \leq \frac{2}{x^2}\,, \tag{48}$$

and the same argument as used in Sect. "Small Divisors" implies that the series in (43) for h_ν converges for $|\varepsilon| < \varepsilon_0$, where ε_0 is defined as in (26) with $C_0 = 2\gamma^2/C_2$. Moreover (47) also implies that $h(\omega t)$ is analytic in ε (notwithstanding that the expansion is not a power expansion), so that we can say a posteriori that the original power series (25) also converges.

Generalizations

The case where the function $G(u)$ in (1) does not vanish is notationally more involved, and it is explicitly discussed in Gentile [28].

Extensions of the results described above to Hamiltonian functions more general than (5) require only some notational complications, and can be found in Gentile & Mastropietro [37] for anisochronous systems and in Bartuccelli & Gentile [3] for isochronous systems. The case (5) with f a real analytic function was first discussed by using trees in Chierchia & Falcolini [12].

Lower-Dimensional Tori

A tree formalism for hyperbolic lower-dimensional tori was introduced and studied in Gallavotti [19], Gentile [26] and Gallavotti, Gentile & Mastropietro [25], for systems consisting of a set of rotators interacting with a pendulum. In that case also the stable and unstable manifolds (*whiskers*) of the tori were studied with diagrammatic techniques.

For Hamiltonian functions of the form (5) solutions of the form (43) describe quasi-periodic motions with frequency vector ω on a d-dimensional manifold. If we fix t, say $t = 0$, and keep α_0 as a parameter, we obtain a parametrization of the manifold in terms of $\alpha_0 \in \mathbb{T}^d$, hence the manifold is an invariant torus. We say that the torus is a *maximal torus*.

We can study the problem of persistence of *lower-dimensional tori* by considering unperturbed solutions $\alpha(t) = \alpha_0 + \omega t$, where the components of ω are rationally dependent. For instance we can imagine that there exist s linearly independent vectors $\hat{\nu}_1, \ldots, \hat{\nu}_s \in \mathbb{Z}^d$ such that $\omega \cdot \hat{\nu}_k = 0$ for $k = 1, \ldots, s$. In that case, we say that the unperturbed torus is a *resonant* torus of order s. We can imagine performing a linear change of variables which transforms the frequency vector into a new vector, that we still denote with ω, such that $\omega = (\bar{\omega}, 0)$, where $\bar{\omega} \in \mathbb{R}^r$ and $0 \in \mathbb{R}^s$, with $r + s = d$, and $\bar{\omega}$ is irrational. This naturally suggests that we write $\alpha = (\bar{\alpha}, \beta)$, with $\bar{\alpha} \in \mathbb{T}^r$ and $\beta \in \mathbb{T}^s$. More generally, here and henceforth in this subsection for any vector $\nu \in \mathbb{R}^d$ we denote by $\bar{\nu}$ the vector in $\mathbb{R}^r$ whose components are the first r components of ν. For instance for the initial phase α_0 we write $\alpha_0 = (\bar{\alpha}_0, \beta_0)$.

In general the resonant torus is destroyed by the perturbation, and only some lower-dimensional tori persist under perturbation. We assume on $\bar{\omega}$ one of the Diophantine conditions of Sect. "Small Divisors" in $\mathbb{R}^r$, for instance $|\bar{\omega} \cdot \bar{\nu}| > \gamma/|\bar{\nu}|^\tau$ for all $\bar{\nu} \in \mathbb{Z}^r$, $\bar{\nu} \neq 0$.

To prove the existence of a maximal torus for the system with Hamiltonian function (5) we needed no condition on the perturbation f. On the contrary to prove existence of lower-dimensional tori, we need some *non-degeneracy condition*: by defining

$$f_0(\beta) = \int_{\mathbb{T}^r} \frac{\mathrm{d}\bar{\alpha}}{(2\pi)^r} f(\bar{\alpha}, \beta)\,, \tag{49}$$

if $\partial_\beta f_0(\beta_*) = 0$ for some β_* then we assume that the matrix $\partial^2_\beta f_0(\beta_*)$ is positive definite (more generally one could assume it to be non-singular, that is $\det \partial^2_\beta f_0(\beta_*) \neq 0$).

The formal analysis can be carried out as in Sect. "Trees and Graphical Representation", with the only difference being that now the compatibility conditions $[\varepsilon \partial_\alpha f(\alpha)]^{(k)}_\nu = 0$ have to be imposed for all ν such that $\bar{\nu} = 0$, because $\omega \cdot \nu = \bar{\omega} \cdot \bar{\nu}$. It turns out to be an identity only for the first r components (this is trivial for $k = 1$, whereas it requires some work for $k > 1$). For the last s components, for $k = 1$ it reads $\partial_\beta f_0(\beta_0) = 0$, hence it fixes β_0 to be a stationary point for $f_0(\beta)$, while for higher values of k it fixes the corrections of higher order of these values (to do this we need the non-degeneracy condition). Thus, we are free to choose only $\bar{\alpha}_0$ as a free parameter, since the last s components of α_0 have to be fixed.

Copyrighted material

Partial Differential Equations

Existence of quasi-periodic solutions in systems described by one-dimensional nonlinear partial differential equations (finite-dimensional tori in infinite-dimensional systems) was first studied by Kuksin [48], Craig and Wayne [14] and Bourgain [9]. In these systems, even the case of periodic solutions yields small divisors, and hence requires a multiscale analysis. The study of persistence of periodic solutions for nonlinear Schrödinger equations and nonlinear wave equations, with the techniques discussed here, can be found in Gentile & Mastropietro [39], Gentile et al. [40], and Gentile & Procesi [41].

The models are still described by (1), with $G(u) = 0$, but now D is given by $D = \partial_t^2 - \Delta + \mu$ in the case of the wave equation and by $D = i\partial_t - \Delta + \mu$ in the case of the Schrödinger equation, where Δ is the Laplacian and $\mu \in \mathbb{R}$. In dimension 1, one has $\Delta = \partial_x^2$. If we look for periodic solutions with frequency ω it can be convenient to pass to Fourier space, where the operator D acts as

$$D\colon e^{i\omega nt+imx} \to \left(-\omega^2 n^2 + m^2 + \mu\right) e^{i\omega nt+imx}\,, \quad (60)$$

for the wave equation; a similar expression holds for the Schrödinger equation. Therefore the kernel of D can be arbitrarily close to zero for n and m large enough.

Then one can consider, say, (1) for $x \in [0, \pi]$ and $F(u) = u^3$, with Dirichlet boundary conditions $u(0) = u(\pi) = 0$, and study the existence of periodic solutions with frequency ω close to some of the unperturbed frequencies. We refer to the cited bibliography for results and proofs.

Conclusions and Future Directions

The diagrammatic techniques described above have been applied also in cases where no small divisors appear; cf. Berretti & Gentile [4] and Gentile et al. [34]. Of course, such problems are much easier from a technical point of view, and can be considered as propaedeutic examples to become familiar with the tree formalism. Also the study of lower-dimensional tori becomes easy for $r = 1$ (periodic solutions): in that case one has $|\tilde{\omega}\cdot\tilde{\nu}| \geq |\tilde{\omega}|$ for all $\tilde{\nu} \neq 0$, so that the product of the propagators is bounded by $|\tilde{\omega}|^{-2k}$, and one can proceed as in Sect. "Small Divisors" to obtain analyticity of the solutions.

In the case of hyperbolic lower-dimensional tori, if ω is a two-dimensional Diophantine vector of constant type (that is, with $\tau = 1$) the conjugation function h can be proved to be Borel summable [13]. Analogous considerations hold for the one-dimensional systems in the presence of friction and of a quasiperiodic forcing term described in Sect. "Other Ordinary Differential Equations"; in that case one has Borel summability also for one-dimensional ω, that is for periodic forcing [33]. It would be interesting to investigate whether Borel summability could be obtained for higher values of τ.

Recently existence of finite-dimensional tori in the nonlinear Schrödinger equation in higher dimensions was proved by Bourgain [10]. It would be interesting to investigate how far the diagrammatic techniques extend to deal with such higher dimensional generalizations. The main problem is that (the analogues of) the second Melnikov conditions in (57) cannot be imposed.

In certain cases the tree formalism was extended to non-analytic systems, such as some quasi-integrable systems of the form (5) with f in a class of C^p functions for some finite p [7,8]. However, up to exceptional cases, the method described here seems to be intrinsically suited in cases in which the vector fields are analytic. The reason is that in order to exploit the expansion (3), we need that F be infinitely many times differentiable and we need a bound on the derivatives. It is a remarkable property that the perturbation series can be given a meaning also in cases where the solutions are not analytic in ε.

An advantage of the diagrammatic method is that it allows rather detailed information about the solutions, hence it could be more convenient than other techniques to study problems where the underlying structure is not known or too poor to exploit general abstract arguments.

Another advantage is the following. If one is interested not only in proving the existence of the solutions, but also in explicitly constructing them with any prefixed precision, this requires performing analytical or numerical computations with arbitrarily high accuracy. Then high perturbation orders have to be reached, and the easiest and most direct way to proceed is just through perturbation theory: so the approach illustrated here allows a unified treatment for both theoretical investigations and computational ones.

The resummation technique described in Sect. "Resummation" can also be used for computational purposes. With respect to the naive power series expansion it can reduce the computation time required to approximate the solution within a prefixed precision. It can also provide accurate information on the analyticity properties of the solution. For instance, for the Kepler equation, Levi-Civita at the beginning of the last century described a resummation rule (see [49]), which gives immediately the radius of convergence of the perturbation series. Of course, in the case of small divisor problems, everything becomes much more complicated.

Copyrighted material

Bibliography

Primary Literature

1. Arnold VI (1963) Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally periodic motions under a small perturbation of the Hamiltonian (Russian). Uspehi Mat Nauk 18(5):13–40
2. Arnold VI, Kozlov VV, Neĭshtadt AI (1988) Dynamical systems III, Encyclopaedia of Mathematical Sciences, vol 3. Springer, Berlin
3. Bartuccelli MV, Gentile G (2002) Lindstedt series for perturbations of isochronous systems: a review of the general theory. Rev Math Phys 14(2):121–171
4. Berretti A, Gentile G (1999) Scaling properties for the radius of convergence of a Lindstedt series: the standard map. J Math Pures Appl 78(2):159–176
5. Berretti A, Gentile G (2000) Scaling properties for the radius of convergence of a Lindstedt series: generalized standard maps. J Math Pures Appl 79(7):691–713
6. Berretti A, Gentile G (2001) Bryuno function and the standard map. Comm Math Phys 220(3):623–656
7. Bonetto F, Gallavotti G, Gentile G, Mastropietro V (1998) Quasi linear flows on tori: regularity of their linearization. Comm Math Phys 192(3):707–736
8. Bonetto F, Gallavotti G, Gentile G, Mastropietro V (1998) Lindstedt series, ultraviolet divergences and Moser's theorem. Ann Scuola Norm Sup Pisa Cl Sci 26(3):545–593
9. Bourgain J (1998) Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Schrödinger equations. Ann of Math 148(2):363–439
10. Bourgain J (2005) Green's function estimates for lattice Schrödinger operators and applications. Annals of Mathematics Studies 158. Princeton University Press, Princeton
11. Bricmont J, Gawędzki K, Kupiainen A (1999) KAM theorem and quantum field theory. Comm Math Phys 201(3):699–727
12. Chierchia L, Falcolini C (1994) A direct proof of a theorem by Kolmogorov in Hamiltonian systems. Ann Scuola Norm Sup Pisa Cl Sci 21(4):541–593
13. Costin O, Gallavotti G, Gentile G, Giuliani A (2007) Borel summability and Lindstedt series. Comm Math Phys 269(1):175–193
14. Craig W, Wayne CE (1993) Newton's method and periodic solutions of nonlinear wave equations. Comm Pure Appl Math 46(11):1409–1498
15. Davie AM (1994) The critical function for the semistandard map. Nonlinearity 7(1):219–229
16. Eliasson LH (1996) Absolutely convergent series expansions for quasi periodic motions. Math Phys Electron J 2(4):1–33
17. Gallavotti G (1983) The elements of mechanics. Texts and Monographs in Physics. Springer, New York
18. Gallavotti G (1994) Twistless KAM tori. Comm Math Phys 164(1):145–156
19. Gallavotti G (1994) Twistless KAM tori, quasi flat homoclinic intersections, and other cancellations in the perturbation series of certain completely integrable Hamiltonian systems. A review. Rev Math Phys 6(3):343–411
20. Gallavotti G (2001) Renormalization group in statistical mechanics and mechanics: gauge symmetries and vanishing beta functions, Renormalization group theory in the new millennium III. Phys Rep 352(4–6):251–272
21. Gallavotti G (2006) Classical mechanics. In: Françoise JP, Naber GL, Tsun TS (eds) Encyclopedia of Mathematical Physics, vol 1. Elsevier, Oxford
22. Gallavotti G, Gentile G (2002) Hyperbolic low-dimensional invariant tori and summations of divergent series. Comm Math Phys 227(3):421–460
23. Gallavotti G, Bonetto F, Gentile G (2004) Aspects of ergodic, qualitative and statistical theory of motion. Texts and Monographs in Physics. Springer, Berlin
24. Gallavotti G, Gentile G, Giuliani A (2006) Fractional Lindstedt series. J Math Phys 47(1):1–33
25. Gallavotti G, Gentile G, Mastropietro V (1999) A field theory approach to Lindstedt series for hyperbolic tori in three time scales problems. J Math Phys 40(12):6430–6472
26. Gentile G (1995) Whiskered tori with prefixed frequencies and Lyapunov spectrum. Dyn Stab Syst 10(3):269–308
27. Gentile G (2003) Quasi-periodic solutions for two-level systems. Comm Math Phys 242(1-2):221–250
28. Gentile G (2006) Brjuno numbers and dynamical systems. Frontiers in number theory, physics, and geometry. Springer, Berlin
29. Gentile G (2006) Diagrammatic techniques in perturbation theory. In: Françoise JP, Naber GL, Tsun TS (eds) Encyclopedia of Mathematical Physics, vol 2. Elsevier, Oxford
30. Gentile G (2006) Resummation of perturbation series and reducibility for Bryuno skew-product flows. J Stat Phys 125(2): 321–361
31. Gentile G (2007) Degenerate lower-dimensional tori under the Bryuno condition. Ergod Theory Dyn Syst 27(2):427–457
32. Gentile G, Bartuccelli MV, Deane JHB (2005) Summation of divergent series and Borel summability for strongly dissipative differential equations with periodic or quasiperiodic forcing terms. J Math Phys 46(6):1–21
33. Gentile G, Bartuccelli MV, Deane JHB (2006) Quasiperiodic attractors, Borel summability and the Bryuno condition for strongly dissipative systems. J Math Phys 47(7):1–10
34. Gentile G, Bartuccelli MV, Deane JHB (2007) Bifurcation curves of subharmonic solutions and Melnikov theory under degeneracies. Rev Math Phys 19(3):307–348
35. Gentile G, Cortez DA, Barata JCA (2005) Stability for quasi-periodically perturbed Hill's equations. Comm Math Phys 260(2):403–443
36. Gentile G, Gallavotti G (2005) Degenerate elliptic tori. Comm Math Phys 257(2):319–362
37. Gentile G, Mastropietro V (1996) Methods for the analysis of the Lindstedt series for KAM tori and renormalizability in classical mechanics. A review with some applications. Rev Math Phys 8(3):393–444
38. Gentile G, Mastropietro V (2001) Renormalization group for one-dimensional fermions. A review on mathematical results. Renormalization group theory in the new millennium III. Phys Rep 352(4–6):273–437
39. Gentile G, Mastropietro V (2004) Construction of periodic solutions of nonlinear wave equations with Dirichlet boundary conditions by the Lindstedt series method. J Math Pures Appl 83(8):1019–1065
40. Gentile G, Mastropietro V, Procesi M (2005) Periodic solutions for completely resonant nonlinear wave equations with Dirichlet boundary conditions. Comm Math Phys 256(2):437–490
41. Gentile G, Procesi M (2006) Conservation of resonant peri-

Copyrighted material

odic solutions for the one-dimensional nonlinear Schrödinger equation. Comm Math Phys 262(3):533–553
42. Gentile G, van Erp TS (2005) Breakdown of Lindstedt expansion for chaotic maps. J Math Phys 46(10):1–20
43. Harary F (1969) Graph theory. Addison-Wesley, Reading
44. Iserles A, Nørsett SP (1999) On the solution of linear differential equations in Lie groups. Royal Soc Lond Philos Trans Ser A Math Phys Eng Sci 357(1754):983–1019
45. Khanin K, Lopes Dias J, Marklof J (2007) Multidimensional continued fractions, dynamical renormalization and KAM theory. Comm Math Phys 270(1):197–231
46. Koch H (1999) A renormalization group for Hamiltonians, with applications to KAM tori. Ergod Theory Dyn Syst 19(2):475–521
47. Kolmogorov AN (1954) On conservation of conditionally periodic motions for a small change in Hamilton's function (Russian). Dokl Akad Nauk SSSR 98:527–530
48. Kuksin SB (1993) Nearly integrable infinite-dimensional Hamiltonian systems. Lecture Notes in Mathematics, vol 1556. Springer, Berlin
49. Levi-Civita T (1954) Opere matematiche. Memorie e note. Zanichelli, Bologna
50. MacKay RS (1993) Renormalisation in area-preserving maps. Advanced Series in Nonlinear Dynamics, 6. World Scientific Publishing, River Edge
51. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. Nachr Akad Wiss Göttingen Math-Phys Kl. II 1962:1–20
52. Moser J (1967) Convergent series expansions for quasi-periodic motions. Math Ann 169:136–176
53. Poincaré H (1892–1899) Les méthodes nouvelles de la mécanique céleste. Gauthier-Villars, Paris
54. Pöschel J (1982) Integrability of Hamiltonian systems on Cantor sets. Comm Pure Appl Math 35(5):653–696
55. Wintner A (1941) The analytic foundations of celestial mechanics. Princeton University Press, Princeton

Books and Reviews

Berretti A, Gentile G (2001) Renormalization group and field theoretic techniques for the analysis of the Lindstedt series. Regul Chaotic Dyn 6(4):389–420
Gentile G (1999) Diagrammatic techniques in perturbations theory, and applications. Symmetry and perturbation theory. World Science, River Edge

Copyrighted material

Discrete Control Systems

TAEYOUNG LEE[1], MELVIN LEOK[2], HARRIS MCCLAMROCH[1]
[1] Department of Aerospace Engineering, University of Michigan, Ann Arbor, USA
[2] Department of Mathematics, Purdue University, West Lafayette, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Discrete Lagrangian and Hamiltonian Mechanics
Optimal Control of Discrete Lagrangian and Hamiltonian Systems
Controlled Lagrangian Method for Discrete Lagrangian Systems
Future Directions
Acknowledgments
Bibliography

Glossary

Discrete variational mechanics A formulation of mechanics in discrete-time that is based on a discrete analogue of *Hamilton's principle*, which states that the system takes a trajectory for which the action integral is stationary.

Geometric integrator A numerical method for obtaining numerical solutions of differential equations that preserves geometric properties of the continuous flow, such as symplecticity, momentum preservation, and the structure of the configuration space.

Lie group A differentiable manifold with a group structure where the composition is differentiable. The corresponding *Lie algebra* is the tangent space to the Lie group based at the identity element.

Symplectic A map is said to be symplectic if given any initial volume in phase space, the sum of the signed projected volumes onto each position-momentum subspace is invariant under the map. One consequence of symplecticity is that the map is volume-preserving as well.

Definition of the Subject

Discrete control systems, as considered here, refer to the control theory of discrete-time Lagrangian or Hamiltonian systems. These discrete-time models are based on a discrete variational principle, and are part of the broader field of geometric integration. Geometric integrators are numerical integration methods that preserve geometric properties of continuous systems, such as conservation of the symplectic form, momentum, and energy. They also guarantee that the discrete flow remains on the manifold on which the continuous system evolves, an important property in the case of rigid-body dynamics.

In nonlinear control, one typically relies on differential geometric and dynamical systems techniques to prove properties such as stability, controllability, and optimality. More generally, the geometric structure of such systems plays a critical role in the nonlinear analysis of the corresponding control problems. Despite the critical role of geometry and mechanics in the analysis of nonlinear control systems, nonlinear control algorithms have typically been implemented using numerical schemes that ignore the underlying geometry.

The field of discrete control systems aims to address this deficiency by restricting the approximation to the choice of a discrete-time model, and developing an associated control theory that does not introduce any additional approximation. In particular, this involves the construction of a control theory for discrete-time models based on geometric integrators that yields numerical implementations of nonlinear and geometric control algorithms that preserve the crucial underlying geometric structure.

Introduction

The dynamics of Lagrangian and Hamiltonian systems have unique geometric properties; the Hamiltonian flow is symplectic, the total energy is conserved in the absence of non-conservative forces, and the momentum maps associated with symmetries of the system are preserved. Many interesting dynamics evolve on a non-Euclidean space. For example, the configuration space of a spherical pendulum is the two-sphere, and the configuration space of rigid body attitude dynamics has a Lie group structure, namely the special orthogonal group. These geometric features determine the qualitative behavior of the system, and serve as a basis for theoretical study.

Geometric numerical integrators are numerical integration algorithms that preserve structures of the continuous dynamics such as invariants, symplecticity, and the configuration manifold (see [14]). The exact geometric properties of the discrete flow not only generate improved qualitative behavior, but also provide accurate and efficient numerical techniques. In this article, we view a geometric integrator as an intrinsically discrete dynamical system, instead of concentrating on the numerical approximation of a continuous trajectory.

Copyrighted material

Numerical integration methods that preserve the simplecticity of a Hamiltonian system have been studied (see [28,36]). Coefficients of a Runge–Kutta method are carefully chosen to satisfy a simplecticity criterion and order conditions to obtain a symplectic Runge–Kutta method. However, it can be difficult to construct such integrators, and it is not guaranteed that other invariants of the system, such as a momentum map, are preserved. Alternatively, variational integrators are constructed by discretizing Hamilton's principle, rather than discretizing the continuous Euler–Lagrange equation (see [31,34]). The resulting integrators have the desirable property that they are symplectic and momentum preserving, and they exhibit good energy behavior for exponentially long times (see [2]). Lie group methods are numerical integrators that preserve the Lie group structure of the configuration space (see [18]). Recently, these two approaches have been unified to obtain Lie group variational integrators that preserve the geometric properties of the dynamics as well as the Lie group structure of the configuration space without the use of local charts, reprojection, or constraints (see [26,29,32]).

Optimal control problems involve finding a control input such that a certain optimality objective is achieved under prescribed constraints. An optimal control problem that minimizes a performance index is described by a set of differential equations, which can be derived using Pontryagin's maximum principle. Discrete optimal control problems involve finding a control input for a discrete dynamic system such that an optimality objective is achieved with prescribed constraints. Optimality conditions are derived from the discrete equations of motion, described by a set of discrete equations. This approach is in contrast to traditional techniques where a discretization appears at the last stage to solve the optimality condition numerically. Discrete mechanics and optimal control approaches determine optimal control inputs and trajectories more accurately with less computational load (see [19]). Combined with an indirect optimization technique, they are substantially more efficient (see [17,23,25]).

The geometric approach to mechanics can provide a theoretical basis for innovative control methodologies in geometric control theory. For example, these techniques allow the attitude of satellites to be controlled using changes in its shape, as opposed to chemical propulsion. While the geometric structure of mechanical systems plays a critical role in the construction of geometric control algorithms, these algorithms have typically been implemented using numerical schemes that ignore the underlying geometry. By applying geometric control algorithms to discrete mechanics that preserve geometric properties, we obtain an exact numerical implementation of the geometric control theory. In particular, the method of controlled Lagrangian systems is based on the idea of adopting a feedback control to realize a modification of either the potential energy or the kinetic energy, referred to as potential shaping or kinetic shaping, respectively. These ideas are applied to construct a real-time digital feedback controller that stabilizes the inverted equilibrium of the cart-pendulum (see [10,11]).

In this article, we will survey discrete Lagrangian and Hamiltonian mechanics, and their applications to optimal control and feedback control theory.

Discrete Lagrangian and Hamiltonian Mechanics

Mechanics studies the dynamics of physical bodies acting under forces and potential fields. In Lagrangian mechanics, the trajectory of the object is derived by finding the path that extremizes the integral of a Lagrangian over time, called the action integral. In many classical problems, the Lagrangian is chosen as the difference between kinetic energy and potential energy. The Legendre transformation provides an alternative description of mechanical systems, referred to as Hamiltonian mechanics.

Discrete Lagrangian and Hamiltonian mechanics has been developed by reformulating the theorems and the procedures of Lagrangian and Hamiltonian mechanics in a discrete time setting (see, for example, [31]). Therefore, discrete mechanics has a parallel structure with the mechanics described in continuous time, as summarized in Fig. 1 for Lagrangian mechanics. In this section, we describe discrete Lagrangian mechanics in more detail, and we derive discrete Euler–Lagrange equations for several mechanical systems.

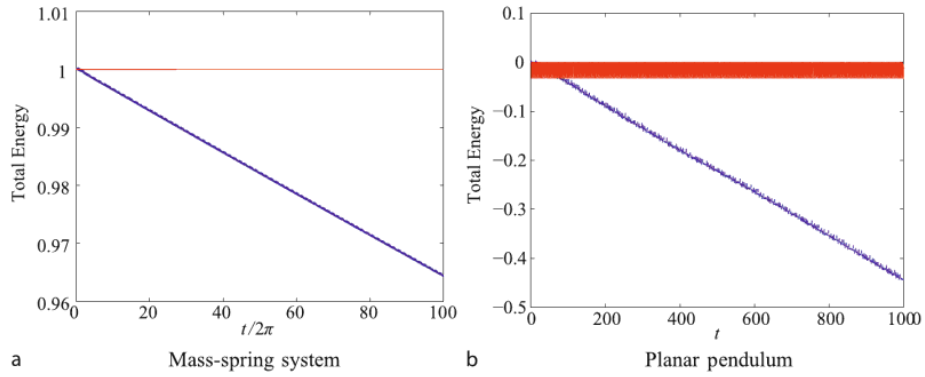
Consider a mechanical system on a configuration space Q, which is the space of possible positions. The Lagrangian depends on the position and velocity, which are elements of the tangent bundle to Q, denoted by TQ. Let $L\colon TQ \to \mathbb{R}$ be the Lagrangian of the system. The discrete Lagrangian, $L_{\mathrm{d}}\colon Q \times Q \to \mathbb{R}$ is an approximation to the exact discrete Lagrangian,

$$L_{\mathrm{d}}^{\mathrm{exact}}(q_0, q_1) = \int_0^h L(q_{01}(t), \dot{q}_{01}(t))\,\mathrm{d}t\,, \tag{1}$$

where $q_{01}(0) = q_0$, $q_{01}(h) = q_1$, and $q_{01}(t)$ satisfies the Euler–Lagrange equation in the time interval $(0, h)$. A discrete action sum $\mathfrak{G}_{\mathrm{d}}\colon Q^{N+1} \to \mathbb{R}$, analogous to the action integral, is given by

$$\mathfrak{G}_{\mathrm{d}}(q_0, q_1, \ldots, q_N) = \sum_{k=0}^{N-1} L_{\mathrm{d}}(q_k, q_{k+1})\,. \tag{2}$$

Copyrighted material



Discrete Control Systems, Figure 2
Computed total energy (RK45: *blue, dotted*, VI: *red, solid*)

sphere $\mathbb{S}^1 = \{q \in \mathbb{R}^2 \mid \|q\| = 1\}$. While it is common to parametrize the one-sphere by an angle, we develop parameter-free equations of motion in the special orthogonal group SO(2), which is a group of 2×2 orthogonal matrices with determinant 1, i. e. $\mathrm{SO}(2) = \{R \in \mathbb{R}^{2\times2} \mid R^{\mathrm{T}}R = I_{2\times2}, \det[R] = 1\}$. SO(2) is diffeomorphic to the one-sphere. It is also possible to develop global equations of motion on the one-sphere directly, as shown in the next example, but here we focus on the special orthogonal group in order to illustrate the key steps to develop a Lie group variational integrator.

We first exploit the basic structures of the Lie group SO(2). Define a hat map $\hat{\cdot}$, which maps a scalar Ω to a 2×2 skew-symmetric matrix $\hat{\Omega}$ as

$$\hat{\Omega} = \begin{bmatrix} 0 & -\Omega \\ \Omega & 0 \end{bmatrix} .$$

The set of 2×2 skew-symmetric matrices forms the Lie algebra $\mathfrak{so}(2)$. Using the hat map, we identify $\mathfrak{so}(2)$ with $\mathbb{R}$. An inner product on $\mathfrak{so}(2)$ can be induced from the inner product on $\mathbb{R}$ as $\langle \hat{\Omega}_1, \hat{\Omega}_2 \rangle = 1/2\mathrm{tr}[\hat{\Omega}_1^{\mathrm{T}}\hat{\Omega}_2] = \Omega_1\Omega_2$ for any $\Omega_1, \Omega_2 \in \mathbb{R}$. The matrix exponential is a local diffeomorphism from $\mathfrak{so}(2)$ to SO(2) given by

$$\exp\hat{\Omega} = \begin{bmatrix} \cos\Omega & -\sin\Omega \\ \sin\Omega & \cos\Omega \end{bmatrix} .$$

The kinematics equation for $R \in \mathrm{SO}(2)$ can be written in terms of a Lie algebra element as

$$\dot{R} = R\hat{\Omega} . \tag{12}$$

Continuous equations of motion: The Lagrangian for a planar pendulum $L\colon \mathrm{SO}(2)\times\mathfrak{so}(2) \to \mathbb{R}$ can be written as

$$\begin{aligned} L(R, \hat{\Omega}) &= \frac{1}{2}ml^2\Omega^2 + mgle_2^{\mathrm{T}}Re_2 \\ &= \frac{1}{2}ml^2\langle\hat{\Omega}, \hat{\Omega}\rangle + mgle_2^{\mathrm{T}}Re_2 , \end{aligned} \tag{13}$$

where the constant $g \in \mathbb{R}$ is the gravitational acceleration. The mass and the length of the pendulum are denoted by $m, l \in \mathbb{R}$, respectively. The second expression is used to define a discrete Lagrangian later. We choose the bases of the inertial frame and the body-fixed frame such that the unit vector along the gravity direction in the inertial frame, and the unit vector along the pendulum axis in the body-fixed frame are represented by the same vector $e_2 = [0; 1] \in \mathbb{R}^2$. Thus, for example, the hanging attitude is represented by $R = I_{2\times2}$. Here, the rotation matrix $R \in \mathrm{SO}(2)$ represents the linear transformation from a representation of a vector in the body-fixed frame to the inertial frame.

Since the special orthogonal group is not a linear vector space, the expression for the variation should be carefully chosen. The infinitesimal variation of a rotation matrix $R \in \mathrm{SO}(2)$ can be written in terms of its Lie algebra element as

$$\delta R = \left.\frac{\mathrm{d}}{\mathrm{d}\epsilon}\right|_{\epsilon=0} R\exp\epsilon\hat{\eta} = R\hat{\eta}\exp\epsilon\hat{\eta}\Big|_{\epsilon=0} = R\hat{\eta} , \tag{14}$$

where $\eta \in \mathbb{R}$ so that $\hat{\eta} \in \mathfrak{so}(2)$. The infinitesimal variation of the angular velocity is induced from this expression and (12) as

$$\begin{aligned} \delta\hat{\Omega} &= \delta R^{\mathrm{T}}\dot{R} + R^{\mathrm{T}}\delta\dot{R} \\ &= (R\hat{\eta})^{\mathrm{T}}\dot{R} + R^{\mathrm{T}}(\dot{R}\hat{\eta} + R\dot{\hat{\eta}}) \\ &= -\hat{\eta}\hat{\Omega} + \hat{\Omega}\hat{\eta} + \dot{\hat{\eta}} = \dot{\hat{\eta}} , \end{aligned} \tag{15}$$

where we used the equality of mixed partials to compute $\delta\dot{R}$ as $\mathrm{d}/\mathrm{d}t(\delta R)$.

Copyrighted material

Define the action integral to be $\mathfrak{G} = \int_0^T L(R, \hat{\Omega})\mathrm{d}t$. The infinitesimal variation of the action integral is obtained by using (14) and (15). Hamilton's principle yields the following continuous equations of motion.

$$\dot{\Omega} + \frac{g}{l} e_2^{\mathrm{T}} R e_1 = 0 \,, \tag{16}$$

$$\dot{R} = R\hat{\Omega} \,. \tag{17}$$

If we parametrize the rotation matrix as

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} ,$$

these equations are equivalent to

$$\ddot{\theta} + \frac{g}{l} \sin\theta = 0 \,. \tag{18}$$

Discrete equations of motion: We develop a Lie group variational integrator on SO(2). Similar to (4), define $F_k \in \mathrm{SO}(2)$ such that

$$R_{k+1} = R_k F_k \,. \tag{19}$$

Thus, $F_k = R_k^{\mathrm{T}} R_{k+1}$ represents the relative update between two integration steps. If we find $F_k \in \mathrm{SO}(2)$, the orthogonal structure is preserved through (19) since multiplication of orthogonal matrices is also orthogonal. This is a key idea of Lie group variational integrators.

Define the discrete Lagrangian $L_{\mathrm{d}} \colon \mathrm{SO}(2)\times\mathrm{SO}(2) \to \mathbb{R}$ to be

$$\begin{aligned} L_{\mathrm{d}}(R_k, F_k) &= \frac{1}{2h} m l^2 \langle F_k - I_{2\times 2}, F_k - I_{2\times 2} \rangle \\ &\quad + \frac{h}{2} mgl e_2^{\mathrm{T}} R_k e_2 + \frac{h}{2} mgl e_2^{\mathrm{T}} R_{k+1} e_2 \\ &= \frac{1}{2h} m l^2 \mathrm{tr}[I_{2\times 2} - F_k] + \frac{h}{2} mgl e_2^{\mathrm{T}} R_k e_2 \\ &\quad + \frac{h}{2} mgl e_2^{\mathrm{T}} R_{k+1} e_2 \,, \end{aligned} \tag{20}$$

which is obtained by an approximation $h\hat{\Omega}_k \simeq R_k^{\mathrm{T}}(R_{k+1} - R_k) = F_k - I_{2\times 2}$, applied to the continuous Lagrangian given by (13).

As for the continuous time case, expressions for the infinitesimal variations should be carefully chosen. The infinitesimal variation of a rotation matrix is the same as (14), namely

$$\delta R_k = R_k \hat{\eta}_k \,, \tag{21}$$

for $\eta_k \in \mathbb{R}$, and the constrained variation of F_k is obtained from (19) as

$$\begin{aligned} \delta F_k &= \delta R_k^{\mathrm{T}} R_{k+1} + R_k^{\mathrm{T}} \delta R_{k+1} \\ &= -\hat{\eta}_k F_k + F_k \hat{\eta}_{k+1} = F_k(\hat{\eta}_{k+1} - F_k^{\mathrm{T}} \hat{\eta}_k F_k) \\ &= F_k(\hat{\eta}_{k+1} - \hat{\eta}_k) \,, \end{aligned} \tag{22}$$

where we use the fact that $F\hat{\eta}F^{\mathrm{T}} = \hat{\eta}$ for any $F \in \mathrm{SO}(2)$ and $\hat{\eta} \in \mathfrak{so}(2)$.

Define an action sum $\mathfrak{G}_{\mathrm{d}} \colon \mathrm{SO}(2)^{N+1} \to \mathbb{R}$ as $\mathfrak{G}_{\mathrm{d}} = \sum_{k=0}^{N-1} L_{\mathrm{d}}(R_k, F_k)$. Using (21) and (22), the variation of the action sum is written as

$$\begin{aligned} \delta\mathfrak{G}_{\mathrm{d}} = \sum_{k=1}^{N-1} \Big\langle & \frac{1}{2h} m l^2 \left(F_{k-1} - F_{k-1}^{\mathrm{T}}\right) \\ & - \frac{1}{2h}\left(F_k - F_k^{\mathrm{T}}\right) - hmgl\widehat{e_2^{\mathrm{T}} R_k e_1}, \ \hat{\eta}_k \Big\rangle . \end{aligned}$$

From the discrete Hamilton's principle, $\delta\mathfrak{G}_{\mathrm{d}} = 0$ for any $\hat{\eta}_k$. Thus, we obtain the Lie group variational integrator on SO(2) as

$$\left(F_k - F_k^{\mathrm{T}}\right) - \left(F_{k+1} - F_{k+1}^{\mathrm{T}}\right) - \frac{2h^2 g}{l} \widehat{e_2^{\mathrm{T}} R_{k+1} e_1} = 0 \,, \tag{23}$$

$$R_{k+1} = R_k F_k \,. \tag{24}$$

For a given (R_k, F_k) and $R_{k+1} = R_k F_k$, (23) is solved to find F_{k+1}. This yields a discrete map $(R_k, F_k) \mapsto (R_{k+1}, F_{k+1})$. If we parametrize the rotation matrices R and F with θ and $\Delta\theta$ and if we assume that $\Delta\theta \ll 1$, these equations are equivalent to

$$\frac{1}{h}(\theta_{k+1} - 2\theta_k + \theta_k) + \frac{hg}{l} \sin\theta_k = 0 \,.$$

The discrete version of the Legendre transformation provides the discrete Hamiltonian map as follows.

$$F_k - F_k^{\mathrm{T}} = 2h\hat{\Omega} - \frac{h^2 g}{l} \widehat{e_2^{\mathrm{T}} R_k e_1} \,, \tag{25}$$

$$R_{k+1} = R_k F_k \,, \tag{26}$$

$$\Omega_{k+1} = \Omega_k - \frac{hg}{2l} e_2^{\mathrm{T}} R_k e_1 - \frac{hg}{2l} e_2^{\mathrm{T}} R_{k+1} e_1 \,. \tag{27}$$

For a given (R_k, Ω_k), we solve (25) to obtain F_k. Using this, (R_{k+1}, Ω_{k+1}) is obtained from (26) and (27). This yields a discrete map $(R_k, \Omega_k) \mapsto (R_{k+1}, \Omega_{k+1})$.

Numerical example: We compare the computational properties of the discrete equations of motion given by (25)–(27) with a 4(5)th order variable step size Runge–Kutta method. We choose $m = 1\,\mathrm{kg}$, $l = 9.81\,\mathrm{m}$. The initial conditions are $\theta_0 = \pi/2\,\mathrm{rad}$, $\Omega = 0$, and the total energy is $E = 0\,\mathrm{Nm}$. The simulation time is 1000 sec, and the step-size $h = 0.03$ of the discrete equations of motion is chosen such that the CPU times are identical. Figure 2b shows the computed total energy for both methods. The variational integrator preserves the total energy well. There is no drift in the computed total energy, and the

Copyrighted material

mean variation is 1.0835×10^{-2} Nm. But, there is a notable dissipation of the computed total energy for the Runge–Kutta method. Note that the computed total energy would further decrease as the simulation time increases.

Example 3 (Spherical Pendulum) A spherical pendulum is a mass particle connected to a frictionless, two degree-of-freedom pivot by a rigid massless link. The mass particle acts under a uniform gravitational potential. The configuration space is the two-sphere $\mathbb{S}^2 = \{q \in \mathbb{R}^3 \,|\, \|q\| = 1\}$. It is common to parametrize the two-sphere by two angles, but this description of the spherical pendulum has a singularity. Any trajectory near the singularity encounters numerical ill-conditioning. Furthermore, this leads to complicated expressions involving trigonometric functions.

Here we develop equations of motion for a spherical pendulum using the global structure of the two-sphere without parametrization. In the previous example, we develop equations of motion for a planar pendulum using the fact that the one-sphere $\mathbb{S}^1$ is diffeomorphic to the special orthogonal group SO(2). But, the two-sphere is not diffeomorphic to a Lie group. Instead, there exists a natural Lie group action on the two-sphere. That is the 3-dimensional special orthogonal group SO(3), a group of 3×3 orthogonal matrices with determinant 1, i. e. $\mathrm{SO}(3) = \{R \in \mathbb{R}^{3\times3} \,|\, R^{\mathrm{T}}R = I_{3\times3},\ \det[R] = 1\}$. The special orthogonal group SO(3) acts on the two-sphere in a transitive way; for any $q_1, q_2 \in \mathbb{S}^2$, there exists a $R \in \mathrm{SO}(3)$ such that $q_2 = Rq_1$. Thus, the discrete update for the two-sphere can be expressed in terms of a rotation matrix as (19). This is a key idea to develop a discrete equations of motion for a spherical pendulum.

Continuous equations of motion: Let $q \in \mathbb{S}^2$ be a unit vector from the pivot point to the point mass. The Lagrangian for a spherical pendulum can be written as

$$L(q, \dot{q}) = \frac{1}{2} m l^2 \dot{q} \cdot \dot{q} + mgle_3 \cdot q \,, \tag{28}$$

where the gravity direction is assumed to be $e_3 = [0; 0; 1] \in \mathbb{R}^3$. The mass and the length of the pendulum are denoted by $m, l \in \mathbb{R}$, respectively. The infinitesimal variation of the unit vector q can be written in terms of the vector cross product as

$$\delta q = \xi \times q \,, \tag{29}$$

where $\xi \in \mathbb{R}^3$ is constrained to be orthogonal to the unit vector, i. e. $\xi \cdot q = 0$. Using this expression for the infinitesimal variation, Hamilton's principle yields the following continuous equations of motion.

$$\ddot{q} + (\dot{q} \cdot \dot{q})q + \frac{g}{l}(q \times (q \times e_3)) = 0 \,. \tag{30}$$

Since $\dot{q} = \omega \times q$ for some angular velocity $\omega \in \mathbb{R}^3$ satisfying $\omega \cdot q = 0$, this can also be equivalently written as

$$\dot{\omega} - \frac{g}{l} q \times e_3 = 0 \,, \tag{31}$$

$$\dot{q} = \omega \times q \,. \tag{32}$$

These are global equations of motion for a spherical pendulum; these are much simpler than the equations expressed in term of angles, and they have no singularity.

Discrete equations of motion: We develop a variational integrator for the spherical pendulum defined on $\mathbb{S}^2$. Since the special orthogonal group acts on the two-sphere transitively, we can define the discrete update map for the unit vector as

$$q_{k+1} = F_k q_k \tag{33}$$

for $F_k \in \mathrm{SO}(3)$. The rotation matrix F_k is not uniquely defined by this condition, since $\exp(\lambda \hat{q}_k)$, which corresponds to a rotation about the q_k direction fixes the vector q_k. Consequently, if F_k satisfies (33), then $F_k \exp(\lambda \hat{q}_k)$ does as well. We avoid this ambiguity by requiring that F_k does not rotate about q_k, which can be achieved by letting $F_k = \exp(\hat{f}_k)$, where $f_k \cdot q_k = 0$.

Define a discrete Lagrangian $L_{\mathrm{d}} \colon \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}$ to be

$$L_{\mathrm{d}}(q_k, q_{k+1}) = \frac{1}{2h} m l^2 (q_{k+1} - q_k) \cdot (q_{k+1} - q_k) + \frac{h}{2} mgle_3 \cdot q_k + \frac{h}{2} mgle_3 \cdot q_{k+1} \,.$$

The variation of q_k is the same as (29), namely

$$\delta q_k = \xi_k \times q_k \tag{34}$$

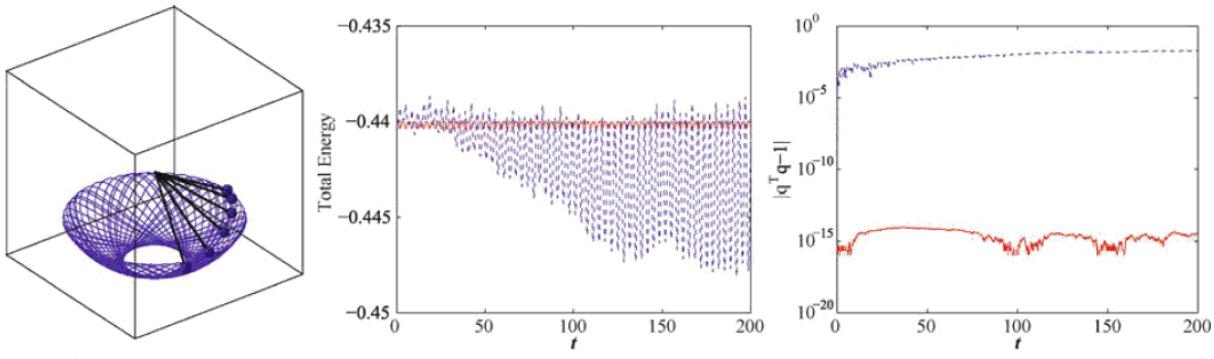
for $\xi_k \in \mathbb{R}^3$ with a constraint $\xi_k \cdot q_k = 0$. Using this discrete Lagrangian and the expression for the variation, discrete Hamilton's principle yields the following discrete equations of motion for a spherical pendulum.

$$q_{k+1} = \left(h\omega_k + \frac{h^2 g}{2l} q_k \times e_3 \right) \times q_k + \left(1 - \left\| h\omega_k + \frac{h^2 g}{2l} q_k \times e_3 \right\|^2 \right)^{1/2} q_k \,, \tag{35}$$

$$\omega_{k+1} = \omega_k + \frac{hg}{2l} q_k \times e_3 + \frac{hg}{2l} q_{k+1} \times e_3 \,. \tag{36}$$

Since an explicit solution for $F_k \in \mathrm{SO}(3)$ can be obtained in this case, the rotation matrix F_k does not appear in the equations of motion. This variational integrator on $\mathbb{S}^2$ exactly preserves the unit length of q_k, the constraint

Copyrighted material



Discrete Control Systems, Figure 3
Numerical simulation of a spherical pendulum (RK45: *blue, dotted*, VI: *red, solid*)

$q_k \cdot \omega_k = 0$, and the third component of the angular velocity $\omega_k \cdot e_3$ which is conserved since gravity exerts no moment along the gravity direction e_3.

Numerical example: We compare the computational properties of the discrete equations of motion given by (35) and (36) with a 4(5)th order variable step size Runge–Kutta method for (31) and (32). We choose $m = 1\,\mathrm{kg}$, $l = 9.81\,\mathrm{m}$. The initial conditions are $q_0 = [\sqrt{3}/2, 0, 1/2]$, $\omega_0 = 0.1[\sqrt{3}, 0, 3]\,\mathrm{rad/sec}$, and the total energy is $E = -0.44\,\mathrm{Nm}$. The simulation time is 200 sec, and the step-size $h = 0.05$ of the discrete equations of motion is chosen such that the CPU times are identical. Figure 3 shows the computed total energy and the unit length errors. The variational integrator preserves the total energy and the structure of the two-sphere well. The mean total energy deviation is $1.5460 \times 10^{-4}\,\mathrm{Nm}$, and the mean unit length error is 3.2476×10^{-15}. But, there is a notable dissipation of the computed total energy for the Runge–Kutta method. The Runge–Kutta method also fails to preserve the structure of the two-sphere. The mean unit length error is 1.0164×10^{-2}.

Example 4 (Rigid Body in a Potential Field) Consider a rigid body under a potential field that is dependent on the position and the attitude of the body. The configuration space is the special Euclidean group, which is a semi-direct product of the special orthogonal group and Euclidean space, i. e. $\mathrm{SE}(3) = \mathrm{SO}(3) \circledS \mathbb{R}^3$.

Continuous equations of motion: The equations of motion for a rigid body can be developed either from Hamilton's principle (see [26]) in a similar way as Example 2, or directly from the generalized discrete Euler–Poincaré equation given at (5). Here, we summarize the results. Let $m \in \mathbb{R}$ and $J \in \mathbb{R}^{3\times3}$ be the mass and the moment of inertia matrix of a rigid body. For $(R, x) \in \mathrm{SE}(3)$, the linear transformation from the body-fixed frame to the inertial frame is denoted by the rotation matrix $R \in \mathrm{SO}(3)$, and the position of the mass center in the inertial frame is denoted by a vector $x \in \mathbb{R}^3$. The vectors $\Omega, v \in \mathbb{R}^3$ are the angular velocity in the body-fixed frame, and the translational velocity in the inertial frame, respectively. Suppose that the rigid body acts under a configuration-dependent potential $U\colon \mathrm{SE}(3) \to \mathbb{R}$. The continuous equations of motion for the rigid body can be written as

$$\dot{R} = R\hat{\Omega}\,, \tag{37}$$

$$\dot{x} = v\,, \tag{38}$$

$$J\dot{\Omega} + \Omega \times J\Omega = M\,, \tag{39}$$

$$m\dot{v} = -\frac{\partial U}{\partial x}\,, \tag{40}$$

where the hat map $\hat{\cdot}$ is an isomorphism from $\mathbb{R}^3$ to 3×3 skew-symmetric matrices $\mathfrak{so}(3)$, defined such that $\hat{x}y = x \times y$ for any $x, y \in \mathbb{R}^3$. The moment due to the potential $M \in \mathbb{R}^3$ is obtained by the following relationship:

$$\hat{M} = \frac{\partial U}{\partial R}^{\mathrm{T}} R - R^{\mathrm{T}} \frac{\partial U}{\partial R}\,. \tag{41}$$

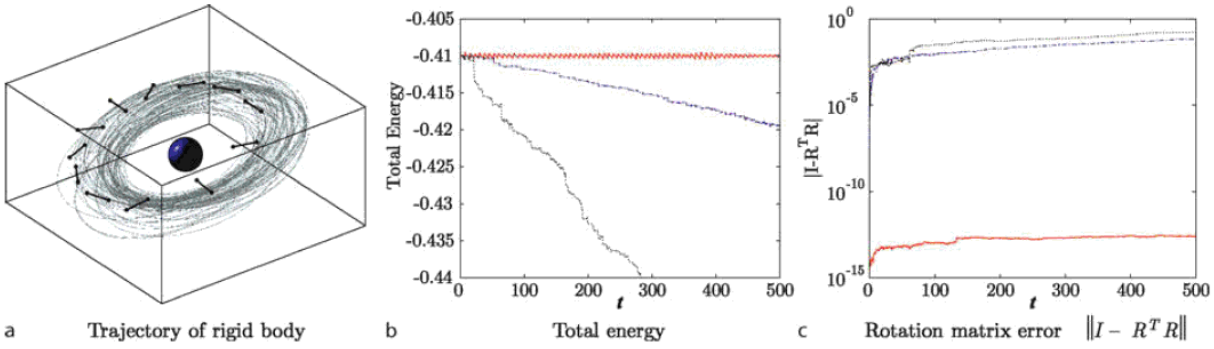
The matrix $\partial U/\partial R \in \mathbb{R}^{3\times3}$ is defined such that $[\partial U/\partial R]_{ij} = \partial U/\partial [R]_{ij}$ for $i, j \in \{1, 2, 3\}$, where the i,jth element of a matrix is denoted by $[\cdot]_{ij}$.

Discrete equations of motion: The corresponding discrete equations of motion are given by

$$h\widehat{J\Omega_k} + \frac{h^2}{2}\hat{M}_k = F_k J_{\mathrm{d}} - J_{\mathrm{d}} F_k^{\mathrm{T}}\,, \tag{42}$$

$$R_{k+1} = R_k F_k\,, \tag{43}$$

Copyrighted material



Discrete Control Systems, Figure 4

Numerical simulation of a dumbbell rigid body (LGVI: *red, solid*, RK45 with rotation matrices: *blue, dash-dotted*, RK45 with quaternions: *black, dotted*)

$$x_{k+1} = x_k + hv_k - \frac{h^2}{2m}\frac{\partial U_k}{\partial x_k}\,, \tag{44}$$

$$J\Omega_{k+1} = F_k^{\mathrm{T}} J\Omega_k + \frac{h}{2}F_k^{\mathrm{T}} M_k + \frac{h}{2}M_{k+1}\,, \tag{45}$$

$$mv_{k+1} = mv_k - \frac{h}{2}\frac{U_k}{\partial x_k} - \frac{h}{2}\frac{\partial U_{k+1}}{\partial x_{k+1}}\,, \tag{46}$$

where $J_{\mathrm{d}} \in \mathbb{R}^{3\times 3}$ is a non-standard moment of inertia matrix defined as $J_{\mathrm{d}} = 1/2\mathrm{tr}[J]I_{3\times 3} - J$. For a given $(R_k, x_k, \Omega_k, v_k)$, we solve the implicit Eq. (42) to find $F_k \in \mathrm{SO}(3)$. Then, the configuration at the next step R_{k+1}, x_{k+1} is obtained by (43) and (44), and the moment and force $M_{k+1}, (\partial U_{k+1})/(\partial x_{k+1})$ can be computed. Velocities Ω_{k+1}, v_{k+1} are obtained from (45) and (46). This defines a discrete flow map, $(R_k, x_k, \Omega_k, v_k) \mapsto (R_{k+1}, x_{k+1}, \Omega_{k+1}, v_{k+1})$, and this process can be repeated. This Lie group variational integrator on SE(3) can be generalized to multiple rigid bodies acting under their mutual gravitational potential (see [26]).

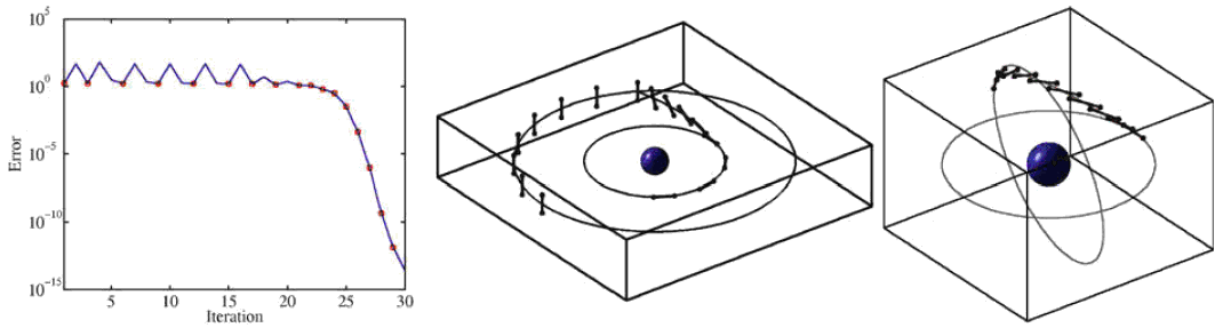
Numerical example: We compare the computational properties of the discrete equations of motion given by (42)–(46) with a 4(5)th order variable step size Runge–Kutta method for (37)–(40). In addition, we compute the attitude dynamics using quaternions on the unit three-sphere $\mathbb{S}^3$. The attitude kinematics Eq. (37) is rewritten in terms of quaternions, and the corresponding equations are integrated by the same Runge–Kutta method.

We choose a dumbbell spacecraft, that is two spheres connected by a rigid massless rod, acting under a central gravitational potential. The resulting system is a restricted full two body problem. The dumbbell spacecraft model has an analytic expression for the gravitational potential, resulting in a nontrivial coupling between the attitude dynamics and the orbital dynamics.

As shown in Fig. 4a, the initial conditions are chosen such that the resulting motion is a near-circular orbit combined with a rotational motion. Figure 4b and c show the computed total energy and the orthogonality errors of the rotation matrix. The Lie group variational integrator preserves the total energy and the Lie group structure of SO(3). The mean total energy deviation is 2.5983×10^{-4}, and the mean orthogonality error is 1.8553×10^{-13}. But, there is a notable dissipation of the computed total energy and the orthogonality error for the Runge–Kutta method. The mean orthogonality errors for the Runge–Kutta method are 0.0287 and 0.0753, respectively, using kinematics equation with rotation matrices, and using the kinematics equation with quaternions. Thus, the attitude of the rigid body is not accurately computed for Runge–Kutta methods. It is interesting to see that the Runge–Kutta method with quaternions, which is generally assumed to have better computational properties than the kinematics equation with rotation matrices, has larger total energy error and orthogonality error. Since the unit length of the quaternion vector is not preserved in the numerical computations, orthogonality errors arise when converted to a rotation matrix. This suggests that it is critical to preserve the structure of SO(3) in order to study the global characteristics of the rigid body dynamics.

The importance of simultaneously preserving the symplectic structure and the Lie group structure of the configuration space in rigid body dynamics can be observed numerically. Lie group variational integrators, which preserve both of these properties, are compared to methods that only preserve one, or neither, of these properties (see [27]). It is shown that the Lie group variational integrator exhibits greater numerical accuracy and efficiency. Due to these computational advantages, the Lie group variational integrator has been used to study the dynamics of the binary near-Earth asteroid 66391 (1999 KW_4) in joint work between the University of Michigan and the Jet Propulsion Laboratory, NASA (see [37]).

Copyrighted material



Discrete Control Systems, Figure 6
Optimal control of a rigid body

variation of the rotation matrix and the angular velocity given in (14) and (15), the infinitesimal variation can be written as

$$\delta \mathcal{J}_a = \sum_{k=1}^{N-1} h\delta u_k^{f,\mathrm{T}} \left\{ W_f u_k^f + \lambda_{k-1}^2 \right\} + h\delta u_k^{m,\mathrm{T}} \left\{ W_m u_k^m + \lambda_{k-1}^4 \right\} + z_k^{\mathrm{T}} \left\{ -\lambda_{k-1} + A_k^{\mathrm{T}} \lambda_k \right\} ,$$

where $\lambda_k = [\lambda_k^1; \lambda_k^2; \lambda_k^3; \lambda_k^4] \in \mathbb{R}^{12}$, and $z_k \in \mathbb{R}^{12}$ represents the infinitesimal variation of $(R_k, x_k, \Omega_k, v_k)$, given by $z_k = [\mathrm{logm}(R_k^{\mathrm{T}} \delta R_k)^\vee; \delta x_k, \delta \Omega_k, \delta v_k]$. The matrix $A_k \in \mathbb{R}^{12\times 12}$ is defined in terms of $(R_k, x_k, \Omega_k, v_k)$. Thus, necessary conditions for optimality are given by

$$u_{k+1}^f = -W_f^{-1} \lambda_k^2 , \tag{47}$$

$$u_{k+1}^m = -W_m^{-1} \lambda_k^4 , \tag{48}$$

$$\lambda_k = A_{k+1}^{\mathrm{T}} \lambda_{k+1} \tag{49}$$

together with the discrete equations of motion and the boundary conditions.

Computational approach: Necessary conditions for optimality are expressed in terms of a two point boundary problem. The problem is to find the optimal discrete flow, multipliers, and control inputs to satisfy the equations of motion, optimality conditions, multiplier equations, and boundary conditions simultaneously. We use a neighboring extremal method (see [12]). A nominal solution satisfying all of the necessary conditions except the boundary conditions is chosen. The unspecified initial multiplier is updated by successive linearization so as to satisfy the specified terminal boundary conditions in the limit. This is also referred to as the shooting method. The main advantage of the neighboring extremal method is that the number of iteration variables is small.

The difficulty is that the extremal solutions are sensitive to small changes in the unspecified initial multiplier values. The nonlinearities also make it hard to construct an accurate estimate of sensitivity, thereby resulting in numerical ill-conditioning. Therefore, it is important to compute the sensitivities accurately to apply the neighboring extremal method. Here the optimality conditions (47) and (48) are substituted into the equations of motion and the multiplier equations, which are linearized to obtain sensitivity derivatives of an optimal solution with respect to boundary conditions. Using this sensitivity, an initial guess of the unspecified initial conditions is iterated to satisfy the specified terminal conditions in the limit. Any type of Newton iteration can be applied. We use a line search with backtracking algorithm, referred to as Newton–Armijo iteration (see [22]).

Figure 6 shows optimized maneuvers, where a dumbbell spacecraft on a reference circular orbit is transferred to another circular orbit with a different orbital radius and inclination angle. Figure 6a shows the violation of the terminal boundary condition according to the number of iterations on a logarithmic scale. Red circles denote outer iterations in the Newton–Armijo iteration to compute the sensitivity derivatives. The error in satisfaction of the terminal boundary condition converges quickly to machine precision after the solution is close to the local minimum at around the 20th iteration. These convergence results are consistent with the quadratic convergence rates expected of Newton methods with accurately computed gradients.

The neighboring extremal method, also referred to as the shooting method, is numerically efficient in the sense that the number of optimization parameters is minimized. But, in general, this approach may be prone to numerical ill-conditioning (see [3]). A small change in the initial multiplier can cause highly nonlinear behavior of the terminal attitude and angular momentum. It is difficult

Copyrighted material

to compute the gradient for Newton iterations accurately, and the numerical error may not converge. However, the numerical examples presented in this article show excellent numerical convergence properties. The dynamics of a rigid body arises from Hamiltonian mechanics, which have neutral stability, and its adjoint system is also neutrally stable. The proposed Lie group variational integrator and the discrete multiplier equations, obtained from variations expressed in the Lie algebra, preserve the neutral stability property numerically. Therefore the sensitivity derivatives are computed accurately.

Controlled Lagrangian Method for Discrete Lagrangian Systems

The method of controlled Lagrangians is a procedure for constructing feedback controllers for the stabilization of relative equilibria. It relies on choosing a parametrized family of controlled Lagrangians whose corresponding Euler–Lagrange flows are equivalent to the closed loop behavior of a Lagrangian system with external control forces. The condition that these two systems are equivalent results in matching conditions. Since the controlled system can now be viewed as a Lagrangian system with a modified Lagrangian, the global stability of the controlled system can be determined directly using Lyapunov stability analysis.

This approach originated in Bloch et al. [5] and was then developed in Auckly et al. [1]; Bloch et al. [6,7,8,9]; Hamberg [15,16]. A similar approach for Hamiltonian controlled systems was introduced and further studied in the work of Blankenstein, Ortega, van der Schaft, Maschke, Spong, and their collaborators (see, for example, [33,35] and related references). The two methods were shown to be equivalent in Chang et al. [13] and a nonholonomic version was developed in Zenkov et al. [40,41], and Bloch [4].

Since the method of controlled Lagrangians relies on relating the closed-loop dynamics of a controlled system with the Euler–Lagrange flow associated with a modified Lagrangian, it is natural to discretize this approach through the use of variational integrators. In Bloch et al. [10,11], a discrete theory of controlled Lagrangians was developed for variational integrators, and applied to the feedback stabilization of the unstable inverted equilibrium of the pendulum on a cart.

The pendulum on a cart is an example of an underactuated control problem, which has two degrees of freedom, given by the pendulum angle and the cart position. Only the cart position has control forces acting on it, and the stabilization of the pendulum has to be achieved indirectly through the coupling between the pendulum and the cart. The controlled Lagrangian is obtained by modifying the kinetic energy term, a process referred to as kinetic shaping. Similarly, it is possible to modify the potential energy term using potential shaping.

Since the pendulum on a cart model involves both a planar pendulum, and a cart that translates in one-dimension, the configuration space is a cylinder, $\mathbb{S}^1 \times \mathbb{R}$.

Continuous kinetic shaping: The Lagrangian has the form kinetic minus potential energy

$$L(q,\dot{q}) = \frac{1}{2}\left[\alpha\dot{\theta}^2 + 2\beta(\theta)\dot{\theta}\dot{s} + \gamma\dot{s}^2\right] - V(q), \tag{50}$$

and the corresponding controlled Euler–Lagrange dynamics is

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}} - \frac{\partial L}{\partial \theta} = 0\,, \tag{51}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{s}} = u\,, \tag{52}$$

where u is the control input.

Since the potential energy is translation invariant, i. e., $V(q) = V(\theta)$, the *relative equilibria* $\theta = \theta_e$, $\dot{s} = \text{const}$ are unstable and given by non-degenerate critical points of $V(\theta)$. To stabilize the relative equilibria $\theta = \theta_e$, $\dot{s} = \text{const}$ with respect to θ, kinetic shaping is used. The controlled Lagrangian in this case is defined by

$$L^{\tau,\sigma}(q,\dot{q}) = L(\theta,\dot{\theta},\dot{s}+\tau(\theta)\dot{\theta}) + \frac{1}{2}\sigma\gamma(\tau(\theta)\dot{\theta})^2\,, \tag{53}$$

where $\tau(\theta) = \kappa\beta(\theta)$. This velocity shift corresponds to a new choice of the horizontal space (see [8] for details). The dynamics is just the Euler–Lagrange dynamics for controlled Lagrangian (53),

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L^{\tau,\sigma}}{\partial \dot{\theta}} - \frac{\partial L^{\tau,\sigma}}{\partial \theta} = 0\,, \tag{54}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L^{\tau,\sigma}}{\partial \dot{s}} = 0\,. \tag{55}$$

The Lagrangian (53) satisfies the simplified matching conditions of Bloch et al. [9] when the kinetic energy metric coefficient γ in (50) is constant.

Setting $u = -\mathrm{d}\big(\gamma\tau(\theta)\dot{\theta}\big)/\mathrm{d}t$ defines the control input, makes Eqs. (52) and (55) identical, and results in controlled momentum conservation by dynamics (51) and (52). Setting $\sigma = -1/\gamma\kappa$ makes Eqs. (51) and (54) identical when restricted to a level set of the controlled momentum.

Discrete kinetic shaping: Here, we adopt the following notation:

$$q_{k+1/2} = \frac{q_k + q_{k+1}}{2}\,,$$

$$\Delta q_k = q_{k+1} - q_k\,, \quad q_k = (\theta_k, s_k)\,.$$

Copyrighted material

Then, a second-order accurate discrete Lagrangian is given by,

$$L_{\mathrm{d}}(q_k, q_{k+1}) = hL(q_{k+1/2}, \Delta q_k/h)\,.$$

The discrete dynamics is governed by the equations

$$\frac{\partial L_{\mathrm{d}}(q_k, q_{k+1})}{\partial \theta_k} + \frac{\partial L_{\mathrm{d}}(q_{k-1}, q_k)}{\partial \theta_k} = 0\,, \tag{56}$$

$$\frac{\partial L_{\mathrm{d}}(q_k, q_{k+1})}{\partial s_k} + \frac{\partial L_{\mathrm{d}}(q_{k-1}, q_k)}{\partial s_k} = -u_k\,, \tag{57}$$

where u_k is the control input. Similarly, the discrete controlled Lagrangian is,

$$L_{\mathrm{d}}^{\tau,\sigma}(q_k, q_{k+1}) = hL^{\tau,\sigma}(q_{k+1/2}, \Delta q_k/h)\,,$$

with discrete dynamics given by,

$$\frac{\partial L_{\mathrm{d}}^{\tau,\sigma}(q_k, q_{k+1})}{\partial \theta_k} + \frac{\partial L_{\mathrm{d}}^{\tau,\sigma}(q_{k-1}, q_k)}{\partial \theta_k} = 0\,, \tag{58}$$

$$\frac{\partial L_{\mathrm{d}}^{\tau,\sigma}(q_k, q_{k+1})}{\partial s_k} + \frac{\partial L_{\mathrm{d}}^{\tau,\sigma}(q_{k-1}, q_k)}{\partial s_k} = 0\,. \tag{59}$$

Equation (59) is equivalent to the *discrete controlled momentum conservation:*

$$p_k = \mu\,,$$

where

$$\begin{aligned} p_k &= -\frac{\partial L_{\mathrm{d}}^{\tau,\sigma}(q_k, q_{k+1})}{\partial s_k} \\ &= \frac{(1+\gamma\kappa)\beta(\theta_{k+1/2})\Delta\theta_k + \gamma\Delta s_k}{h}\,. \end{aligned}$$

Setting

$$u_k = -\frac{\gamma\Delta\theta_k\tau(\theta_{k+1/2}) - \gamma\Delta\theta_{k-1}\tau(\theta_{k-1/2})}{h}$$

makes Eqs. (57) and (59) identical and allows one to represent the discrete momentum equation (57) as the discrete momentum conservation law $p_k = p$.

The condition that (56)–(57) are equivalent to (58)–(59) yield the discrete matching conditions. The dynamics determined by Eqs. (56)–(57) restricted to the momentum level $p_k = p$ is equivalent to the dynamics of Eqs. (58)–(59) restricted to the momentum level $p_k = \mu$ if and only if the matching conditions

$$\sigma = -\frac{1}{\gamma\kappa}, \qquad \mu = \frac{p}{1+\gamma\kappa},$$

hold.

Numerical example: Simulating the behavior of the discrete controlled Lagrangian system involves viewing Eqs. (58)–(59) as an implict update map $(q_{k-2}, q_{k-1}) \mapsto (q_{k-1}, q_k)$. This presupposes that the initial conditions are given in the form (q_0, q_1); however it is generally preferable to specify the initial conditions as $(q_0, \dot{q}_0)$. This is achieved by solving the boundary condition,

$$\frac{\partial L}{\partial \dot{q}}(q_0, \dot{q}_0) + D_1 L_{\mathrm{d}}(q_0, q_1) + F_{\mathrm{d}}^{-}(q_0, q_1) = 0\,,$$

for q_1. Once the initial conditions are expressed in the form (q_0, q_1), the discrete evolution can be obtained using the implicit update map.

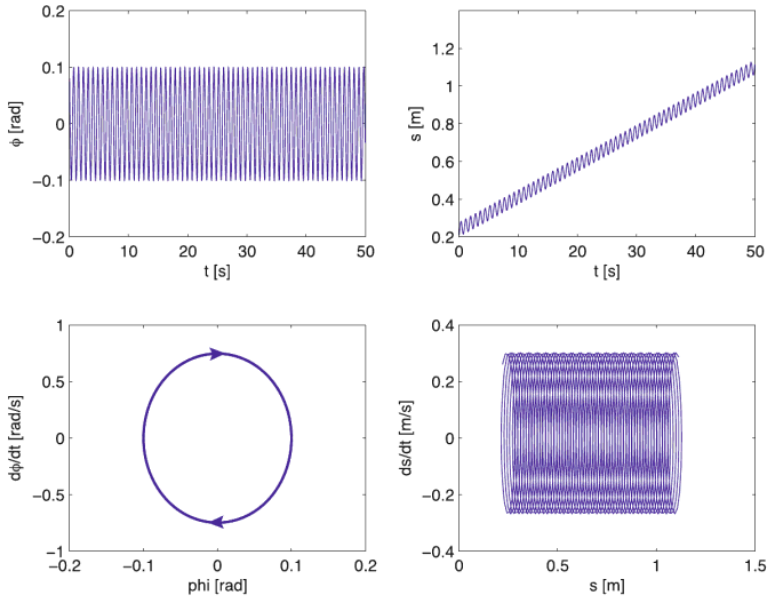
We first consider the case of kinetic shaping on a level surface, when κ is twice the critical value, and without dissipation. Here, $h = 0.05$ sec, $m = 0.14$ kg, $M = 0.44$ kg, and $l = 0.215$ m. As shown in Fig. 7, the θ dynamics is stabilized, but since there is no dissipation, the oscillations are sustained. The s dynamics exhibits both a drift and oscillations, as potential shaping is necessary to stabilize the translational dynamics.

Future Directions

Discrete Receding Horizon Optimal Control: The existing work on discrete optimal control has been primarily focused on constructing the optimal trajectory in an open loop sense. In practice, model uncertainty and actuation errors necessitate the use of feedback control, and it would be interesting to extend the existing work on optimal control of discrete systems to the feedback setting by adopting a receding horizon approach.

Discrete State Estimation: In feedback control, one typically assumes complete knowledge regarding the state of the system, an assumption that is often unrealistic in practice. The general problem of state estimation in the context of discrete mechanics would rely on good numerical methods for quantifying the propagation of uncertainty by solving the Liouville equation, which describes the evolution of a phase space distribution function advected by a prescribed vector field. In the setting of Hamiltonian systems, the solution of the Liouville equation can be solved by the method of characteristics (Scheeres et al. [38]). This implies that a collocational approach (Xiu [39]) combined with Lie group variational integrators, and interpolation based on noncommutative harmonic analysis on Lie groups could yield an efficient means of propagating uncertainty, and serve as the basis of a discrete state estimation algorithm.

Copyrighted material



Discrete Control Systems, Figure 7

Discrete controlled dynamics with kinetic shaping and without dissipation. The discrete controlled system stabilizes the θ motion about the equilibrium, but the s dynamics is not stabilized; since there is no dissipation, the oscillations are sustained

Forced Symplectic–Energy-Momentum Variational Integrators: One of the motivations for studying the control of Lagrangian systems using the method of controlled Lagrangians is that the method provides a natural candidate Lyapunov function to study the global stability properties of the controlled system. In the discrete theory, this approach is complicated by the fact that the energy of a discrete Lagrangian system is not exactly conserved, but rather oscillates in a bounded fashion.

This can be addressed by considering the symplectic-energy-momentum [20] analogue to the discrete Lagrange-d'Alembert principle,

$$\delta \sum_{k=0}^{N_1} L_{\mathrm{d}}(q_k, q_{k+1}, h_k) = \sum_{k=0}^{N-1} [F_{\mathrm{d}}^{-}(q_k, q_{k+1}, h_k) \cdot \delta q_k + F_{\mathrm{d}}^{+}(q_k, q_{k+1}, h_k) \cdot \delta q_{k+1}] \,,$$

where the timestep h_k is allowed to vary, and is chosen to satisfy the variational principle. The variations in h_k yield an Euler–Lagrange equation that reduces to the conservation of discrete energy in the absence of external forces. By developing a theory of controlled Lagrangians around a geometric integrator based on the symplectic-energy-momentum version of the Lagrange–d'Alembert principle, one would potentially be able to use Lyapunov techniques to study the global stability of the resulting numerical control algorithms.

Acknowledgments

TL and ML have been supported in part by NSF Grant DMS-0504747 and DMS-0726263. TL and NHM have been supported in part by NSF Grant ECS-0244977 and CMS-0555797.

Bibliography

Primary Literature

1. Auckly D, Kapitanski L, White W (2000) Control of nonlinear underactuated systems. Commun Pure Appl Math 53:354–369
2. Benettin G, Giorgilli A (1994) On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms. J Stat Phys 74:1117–1143
3. Betts JT (2001) Practical Methods for Optimal Control Using Nonlinear Programming. SIAM, Philadelphia, PA
4. Bloch AM (2003) Nonholonomic Mechanics and Control. In: Interdisciplinary Applied Mathematics, vol 24. Springer, New York
5. Bloch AM, Leonard N, Marsden JE (1997) Matching and stabilization using controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Hyatt Regency San Diego, San Diego, CA, 10–12 December 1997, pp 2356–2361
6. Bloch AM, Leonard N, Marsden JE (1998) Matching and stabilization by the method of controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Hyatt Regency Westshore, Tampa, FL, 16–18 December 1998, pp 1446–1451
7. Bloch AM, Leonard N, Marsden JE (1999) Potential shaping and the method of controlled Lagrangians. In: Proceedings of the

Copyrighted material

IEEE Conference on Decision and Control. Crowne Plaza Hotel and Resort, Phoenix, AZ, 7–10 December 1999, pp 1652–1657
8. Bloch AM, Leonard NE, Marsden JE (2000) Controlled Lagrangians and the stabilization of mechanical systems I: The first matching theorem. IEEE Trans Syst Control 45:2253–2270
9. Bloch AM, Chang DE, Leonard NE, Marsden JE (2001) Controlled Lagrangians and the stabilization of mechanical systems II: Potential shaping. IEEE Trans Autom Contr 46:1556–1571
10. Bloch AM, Leok M, Marsden JE, Zenkov DV (2005) Controlled Lagrangians and stabilization of the discrete cart-pendulum system. In: Proceedings of the IEEE Conference on Decision and Control. Melia Seville, Seville, Spain, 12–15 December 2005, pp 6579–6584
11. Bloch AM, Leok M, Marsden JE, Zenkov DV (2006) Controlled Lagrangians and potential shaping for stabilization of discrete mechanical systems. In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 3333–3338
12. Bryson AE, Ho Y (1975) Applied Optimal Control. Hemisphere, Washington, D.C.
13. Chang D-E, Bloch AM, Leonard NE, Marsden JE, Woolsey C (2002) The equivalence of controlled Lagrangian and controlled Hamiltonian systems. Control Calc Var (special issue dedicated to Lions JL) 8:393–422
14. Hairer E, Lubich C, Wanner G (2006) Geometric Numerical Integration, 2nd edn. Springer Series in Computational Mathematics, vol 31. Springer, Berlin
15. Hamberg J (1999) General matching conditions in the theory of controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Crowne Plaza Hotel and Resort, Phoenix, AZ, 7–10 December 1999, pp 2519–2523
16. Hamberg J (2000) Controlled Lagrangians, symmetries and conditions for strong matching. In: Lagrangian and Hamiltonian Methods for Nonlinear Control. Elsevier, Oxford
17. Hussein II, Leok M, Sanyal AK, Bloch AM (2006) A discrete variational integrator for optimal control problems in *SO*(3). In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 6636–6641
18. Iserles A, Munthe-Kaas H, Nørsett SP, Zanna A (2000) Lie-group methods. In: Acta Numerica, vol 9. Cambridge University Press, Cambridge, pp 215–365
19. Junge D, Marsden JE, Ober-Blöbaum S (2005) Discrete mechanics and optimal control. In: IFAC Congress, Praha, Prague, 3–8 July 2005
20. Kane C, Marsden JE, Ortiz M (1999) Symplectic-energy-momentum preserving variational integrators. J Math Phys 40(7):3353–3371
21. Kane C, Marsden JE, Ortiz M, West M (2000) Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems. Int J Numer Meth Eng 49(10):1295–1325
22. Kelley CT (1995) Iterative Methods for Linear and Nonlinear Equations. SIAM, Philadelphia, PA
23. Lee T, Leok M, McClamroch NH (2005) Attitude maneuvers of a rigid spacecraft in a circular orbit. In: Proceedings of the American Control Conference. Portland Hilton, Portland, OR, 8–10 June 2005, pp 1742–1747
24. Lee T, Leok M, McClamroch NH (2005) A Lie group variational integrator for the attitude dynamics of a rigid body with applications to the 3D pendulum. In: Proceedings of the IEEE Conference on Control Applications. Toronto, Canada, 28–31 August 2005, pp 962–967
25. Lee T, Leok M, McClamroch NH (2006) Optimal control of a rigid body using geometrically exact computations on SE(3). In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 2710–2715
26. Lee T, Leok M, McClamroch NH (2007) Lie group variational integrators for the full body problem. Comput Method Appl Mech Eng 196:2907–2924
27. Lee T, Leok M, McClamroch NH (2007) Lie group variational integrators for the full body problem in orbital mechanics. Celest Mech Dyn Astron 98(2):121–144
28. Leimkuhler B, Reich S (2004) Simulating Hamiltonian Dynamics. Cambridge Monographs on Applied and Computational Mathematics, vol 14. Cambridge University Press, Cambridge
29. Leok M (2004) Foundations of Computational Geometric Mechanics. Ph D thesis, California Instittute of Technology
30. Marsden JE, Ratiu TS (1999) Introduction to Mechanics and Symmetry, 2nd edn. Texts in Applied Mathematics, vol 17. Springer, New York
31. Marsden JE, West M (2001) Discrete mechanics and variational integrators. In: Acta Numerica, vol 10. Cambridge University Press, Cambridge, pp 317–514
32. Marsden JE, Pekarsky S, Shkoller S (1999) Discrete Euler–Poincarée and Lie–Poisson equations. Nonlinearity 12(6):1647–1662
33. Maschke B, Ortega R, van der Schaft A (2001) Energy-based Lyapunov functions for forced Hamiltonian systems with dissipation. IEEE Trans Autom Contr 45:1498–1502
34. Moser J, Veselov AP (1991) Discrete versions of some classical integrable systems and factorization of matrix polynomials. Commun Math Phys 139:217–243
35. Ortega R, Spong MW, Gómez-Estern F, Blankenstein G (2002) Stabilization of a class of underactuated mechanical systems via interconnection and damping assignment. IEEE Trans Autom Contr 47:1218–1233
36. Sanz-Serna JM (1992) Symplectic integrators for Hamiltonian problems: an overview. In: Acta Numerica, vol 1. Cambridge University Press, Cambridge, pp 243–286
37. Scheeres DJ, Fahnestock EG, Ostro SJ, Margot JL, Benner LAM, Broschart SB, Bellerose J, Giorgini JD, Nolan MC, Magri C, Pravec P, Scheirich P, Rose R, Jurgens RF, De Jong EM, Suzuki S (2006) Dynamical configuration of binary near-Earth asteroid (66391) 1999 KW4. Science 314:1280–1283
38. Scheeres DJ, Hsiao F-Y, Park RS, Villac BF, Maruskin JM (2006) Fundamental limits on spacecraft orbit uncertainty and distribution propagation. J Astronaut Sci 54:505–523
39. Xiu D (2007) Efficient collocational approach for parametric uncertainty analysis. Comm Comput Phys 2:293–309
40. Zenkov DV, Bloch AM, Leonard NE, Marsden JE (2000) Matching and stabilization of low-dimensional nonholonomic systems. In: Proceedings of the IEEE Conference on Decision and Control. Sydney Convention and Exhibition Centre, Sydney, NSW Australia; 12–15 December 2000, pp 1289–1295
41. Zenkov DV, Bloch AM, Leonard NE, Marsden JE (2002) Flat nonholonomic matching. In: Proceedings of the American Control Conference. Hilton Anchorage, Anchorage, AK, 8–10 May 2002, pp 2812–2817

Copyrighted material

results for nonlinear equations, existence of low regularity solutions, scattering, qualitative study of evolution equations on manifolds, and many others.

Introduction

To a large extent, mathematical physics is the study of qualitative and quantitative properties of the solutions to systems of differential equations. Phenomena which evolve in time are usually described by nonlinear evolution equations, and their natural setting is the Cauchy problem: the state of the system is assigned at an initial time, and the following evolution is, or should be, completely determined by the equations.

The most basic question concerns the local existence and uniqueness of the solution; we should require that, on one hand, the set of initial conditions is not too restrictive, so that a solution exists, and on the other hand it is not too lax, so that a unique solution is specified. For physically meaningful equations it is natural to require in addition that small errors in the initial state propagate slowly in time, so that two solutions corresponding to slightly different initial data remain close, at least for some time (continuous dependence on the data). This set of requirements is called *local well posedness* of the system. *Global* well posedness implies the possibility of extending these properties to arbitrary large times.

However, local well posedness is just the first step in understanding a nonlinear equation; basically, it amounts to a check that the equation and its Cauchy problem are meaningful from the mathematical and physical point of view. But the truly interesting questions concern global existence, asymptotic behavior and more generally the structure and shape of solutions. Indeed, if an equation describes a microscopic system, all quantities measured in an experiment correspond to asymptotic features of the solution, and local properties are basically useless in this context.

The classical tool to prove local well posedness is the energy method; it can be applied to many equations and leads to very efficient proofs. But the energy method is intrinsically unable to give answers concerning the global existence and behavior of solutions.

Let us illustrate this point more precisely. Most nonlinear evolution equations are of the form

$$Lu = N(u)$$

with a linear part L (e. g., the d'Alembert or Schrödinger operator) and a nonlinear term $N(u)$. The two terms are in competition: the linear flow is well behaved, usually with several conserved quantities bounded by norms of the initial data, while the nonlinear term tends to concentrate the peaks of u and make them higher, increasing the singularity of the solution. At the same time, a small function u is made even smaller by a power-like nonlinearity $N(u)$.

The energy method tries to use only the conserved quantities, which remain constant during the evolution of the linear flow. Essentially, this means to regard the equation as an ODE of the form $y'(t) = N(y(t))$. This approach is very useful if the full (nonlinear) equation has a positive conserved quantity. But in general the best one can prove using the energy method is local well posedness, while a blow up in finite time can not be excluded, even if the initial data are assumed to be small, as the trivial example $y' = y^2$ clearly shows.

To improve on this situation, the strength of the linear flow must be used to a full extent. This is where dispersive phenomena enter the picture. A finer study of the operator L shows that there are quantities, different from the standard L^2 type norms used in the energy estimates, which actually *decay* during the evolution. Hence the term L has an additional advantage in the competition with N which can lead to global existence of small solutions. See Sect. "The Nonlinear Wave Equation" for more details.

This circle of ideas was initiated at the beginning of the 1970s and rapidly developed during the 1980s in several papers, (see [26,27,31,32,33,34,38,40,41,42,50] to mention just a few of the fundamental contributions on the subject). Global existence of small solutions was proved for many nonlinear evolution equations, including the nonlinear wave, Klein–Gordon and Schrödinger equations. The proof of the stability of vacuum for the Einstein equations [15,35] can be regarded as an extreme development of this direction of research.

Another early indication of the importance of L^p estimates came from the separate direction of harmonic analysis. In 1970, at a time when general agreement in the field was that the L^2 framework was the only correct one for the study of the wave equation, R. Strichartz ([44,45], see also [10,11,46]) proved the estimate

$$u_{tt} - \Delta u = F\,, \quad u(0,x) = u_t(0,x) = 0 \qquad \Longrightarrow \qquad \|u\|_{L^p(\mathbb{R}^{n+1})} \lesssim \|F\|_{L^{p'}(\mathbb{R}^{n+1})}$$

for $p = 2(n+1)/(n-1)$. This was a seminal result, both for the emphasis on the L^p approach, and for the idea of regarding all variables including time on the same level, an idea which would play a central role in the later developments. At the same time, it hinted at a deep connection

Copyrighted material

between dispersive properties and the techniques of harmonic analysis.

The new ideas were rapidly incorporated in the mainstream theory of partial differential equations and became an important tool ([22,23,52]); the investigation of *Strichartz estimates* is still in progress. The freedom to work in an L^p setting was revealed as useful in a wide variety of contexts, including linear equations with variable coefficients, scattering, existence of low regularity solutions, and several others.

We might say that this point of view created a new important class of *dispersive equations*, complementing and overlapping with the standard classification into elliptic, hyperbolic, parabolic equations. Actually this class is very wide and contains most evolution equations of mathematical physics: the wave, Schrödinger and Klein–Gordon equations, Dirac and Maxwell systems, their nonlinear counterparts including wave maps, Yang–Mills and Einstein equations, Korteweg de Vries, Benjamin–Ono, and many others. In addition, many basic physical problems are described by systems obtained by coupling these equations; among the most important ones we mention the Dirac–Klein–Gordon, Maxwell–Klein–Gordon, Maxwell–Dirac, Maxwell–Schrödinger and Zakharov systems. Finally, a recent line of research pursues the extension of the dispersive techniques to equations with variable coefficients and equations on manifolds.

The main goal of this article is to give a quick overview of the basic dispersive techniques, and to show some important examples of their applications. It should be kept in mind however that the field is evolving at an impressive speed, mainly due to the contribution of ideas from Fourier and harmonic analysis, and it would be difficult to give a self-contained account of the recent developments, for which we point at the relevant bibliography. In Sect. "The Mechanism of Dispersion" we analyze the basic mechanism of dispersion, with special attention given to two basic examples (Schrödinger and wave equation). Section "Strichartz Estimates" is a coincise exposition of Strichartz estimates, while Sects. "The Nonlinear Wave Equation" and "The Nonlinear Schrödinger Equation" illustrate some typical applications of dispersive techniques to problems of global existence for nonlinear equations. In the last Sect. "Future Directions" we review some of the more promising lines of research in this field.

The Mechanism of Dispersion

The Fourier transform puts in duality oscillations and translations according to the elementary rule

$$\mathcal{F}(f(x+h)) = \mathrm{e}^{-ix\cdot h}\mathcal{F}f\,.$$

The same mechanism lies behind dispersive phenomena, and indeed two dual explanations are possible:

a) in physical space, dispersion is characterized by a finite speed of propagation of the signals; data concentrated on some region tend to spread over a larger volume following the evolution flow. Correspondingly, the L^∞ and other norms of the solution tend to decrease;
b) in Fourier space, dispersion is characterized by oscillations in the Fourier variable ξ, which increase for large $|\xi|$ and large t. This produces cancelation and decay in suitable norms.

We shall explore these dual points of view in two fundamental cases: the Schrödinger and the wave equation. These are the most important and hence most studied dispersive equations, and a fairly complete theory is available for both.

The Schrödinger Equation in Physical Space

The solution $u(t,x)$ of the Schrödinger equation

$$iu_t + \Delta u = 0\,, \quad u(0,x) = f(x)$$

admits several representations which are not completely equivalent. From spectral theory we have an abstract representation as a group of L^2 isometries

$$u(t,x) = \mathrm{e}^{it\Delta} f\,; \tag{1}$$

then we have an explicit representation using the fundamental solution

$$\begin{aligned} u(t,x) &= \frac{1}{(4\pi\, it)^{n/2}} \mathrm{e}^{-i\frac{|x|^2}{4t}} * f \\ &= \frac{1}{(4\pi\, it)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-i\frac{|x-y|^2}{4t}} f(y)\mathrm{d}y\,; \end{aligned} \tag{2}$$

finally, we can represent the solution via Fourier transform as

$$u(t,x) = (2\pi)^{-n} \int \mathrm{e}^{i(t|\xi|^2 + x\cdot\xi)} \widehat{f}(\xi)\mathrm{d}\xi\,. \tag{3}$$

Notice for instance that the spectral and Fourier representations are well defined for $f \in L^2$, while it is not immediately apparent how to give a meaning to (2) if $f \notin L^1$.

From (1) we deduce immediately the *dispersive estimate* for the Schrödinger equation, which has the form

$$|\mathrm{e}^{it\Delta} f| \le (4\pi)^{-n/2} \frac{1}{t^{n/2}} \|f\|_{L^1}\,. \tag{4}$$

Since the L^2 norm of u is constant in time

$$\|\mathrm{e}^{it\Delta} f\|_{L^2} = \|f\|_{L^2}\,, \tag{5}$$

Copyrighted material

this might suggest that the constant L^2 mass spreads on a region of volume $\sim t^n$ i. e. of diameter t. Apparently, this is in contrast with the notion that the Schrödinger equation has an infinite speed of propagation. Indeed, this is a correct but incomplete statement. To explain this point with an explicit example, consider the evolution of a *wave packet*

$$\pi_{x_0,\xi_0} = \mathrm{e}^{-\frac{1}{2}|x-x_0|^2}\mathrm{e}^{i(x-x_0)\cdot\xi_0}\,.$$

A wave packet is a mathematical representation of a particle localized near position x_0, with frequency localized near ξ_0; notice indeed that its Fourier transform is

$$\widehat{\pi}_{x_0,\xi_0} = c\mathrm{e}^{-\frac{1}{2}|\xi-\xi_0|^2}\mathrm{e}^{i\xi\cdot x_0}$$

for some constant c. Exact localization both in space and frequency is of course impossible by Heisenberg's principle; however the approximate localization of the wave packet is a very good substitute, since the gaussian tails decay to 0 extremely fast. The evolution of a wave packet is easy to compute explicitly:

$$\mathrm{e}^{it\Delta}\pi_{x_0,\xi_0} = (1+2it)^{-\frac{n}{2}}\,\exp\left(-\frac{1}{2}\frac{(x-x_0-2t\xi_0)^2}{(1+4t^2)}\right)\mathrm{e}^{i\Gamma}$$

where the real valued function Γ is given by

$$\Gamma = \frac{t(x-x_0)^2 + (x-x_0-t\xi_0)\cdot\xi_0}{(1+4t^2)}\,.$$

We notice that the particle "moves" with velocity $2\xi_0$ and the height of the gaussian spreads precisely at a rate $t^{n/2}$. The mass is essentially concentrated in a ball of radius t, as suspected. This simple remark is at the heart of one of the most powerful techniques to have appeared in harmonic analysis in recent years, the wave packet decomposition, which has led to breakthroughs in several classical problems (see e. g. [7,8,48,51]).

The Schrödinger Equation in Fourier Space

In view of the ease of proof of (4), it may seem unnecessary to examine the solution in Fourier space, however a quick look at this other representation is instructive and prepares the discussion for the wave equation. If we examine (3), we see that the integral is a typical oscillatory integral which can be estimated using the stationary phase method. The phase here is

$$\phi(t,x;\xi) = t|\xi|^2 + x\cdot\xi \quad\Longrightarrow\quad \nabla_\xi\phi = 2t\xi + x$$

so that for each (t,x) we have a unique stationary point $\xi = -x/2t$; the n curvatures at that point are different from zero, and this already suggests a decay of the order $t^{-n/2}$. This argument can be made precise and gives an alternative proof of (4).

We also notice that the Fourier transform in both space and time of $u(t,x) = \mathrm{e}^{it\Delta}f$ is a temperate distribution in $S'(\mathbb{R}^{n+1})$ which can be written, apart from a constant,

$$\tilde{u}(\tau,\xi) = \delta(\tau - |\xi|^2)$$

and is supported on the paraboloid $\tau = |\xi|^2$. This representation of the solution gives additional insight to its behavior in certain frequency regimes, and has proved to be of fundamental importance when studying the existence of low regularity solutions to the nonlinear Schrödinger equation [6,9].

The Wave Equation in Physical Space

For the wave equation, the situation is reversed: the most efficient proof of the dispersive estimate is obtained via the Fourier representation of the solution. This is mainly due to the complexity of the fundamental solution; we recall the relevant formulas briefly. The solution to the homogeneous wave equation

$$w_{tt} - \Delta w = 0\,,\quad w(0,x) = f(x)\,,\quad w_t(0,x) = g(x)$$

can be written as

$$u(t,x) = \cos(t|D|)f + \frac{\sin(t|D|)}{|D|}g$$

where we used the symbolic notations

$$\cos(t|D|)f = \mathcal{F}^{-1}\cos(t|\xi|)\mathcal{F}f\,,$$
$$\frac{\sin(t|D|)}{|D|}g = \mathcal{F}^{-1}\frac{\sin(t|\xi|)}{|\xi|}\mathcal{F}g$$

(recall that $\mathcal{F}$ denotes the Fourier transform in space variables). Notice that

$$\cos(t|D|)f = \frac{\partial}{\partial t}\frac{\sin(t|D|)}{|D|}f\,,$$

hence the properties of the first operator can be deduced from the second. Moreover, by Duhamel's formula, we can express the solution to the nonhomogeneous wave equation

$$w_{tt} - \Delta w = F(t,x)\,,\quad w(0,x) = 0\,,\quad w_t(0,x) = 0$$

as

$$u(t,x) = \int_0^t \frac{\sin((t-s)|D|)}{|D|}F(u(s,\cdot))\mathrm{d}s\,.$$

Copyrighted material

In other words, a study of the operator

$$S(t)g = \frac{\sin(t|D|)}{|D|}g$$

is sufficient to deduce the behavior of the complete solution to the nonhomogeneous wave equation with general data.

The operator $S(t)$ can be expressed explicitly using the fundamental solution of the wave equation. We skip the one-dimensional case since then from the expression of the general solution

$$u(t,x) = \phi(t+x) + \psi(t-x)$$

it is immediately apparent that no dispersion can occur. In odd space dimension $n \geq 3$ we have

$$\frac{\sin(t|D|)}{|D|}g = c_n \left(\frac{1}{t}\partial_t\right)^{\frac{n-3}{2}} \left(\frac{1}{t}\int_{|y|=t} g(x-y)\mathrm{d}\sigma_y\right),$$

and in even space dimension $n \geq 2$

$$\frac{\sin(t|D|)}{|D|}g = c_n \left(\frac{1}{t}\partial_t\right)^{\frac{n-2}{2}} \left(\frac{1}{t}\int_{|y|<t} \frac{g(x-y)}{\sqrt{t^2-|y|^2}}\mathrm{d}y\right)$$

for some constant c_n. These formulas can be expanded and the resulting integrals estimated in order to compute the explicit decay rate of the solution; this approach was followed in [50] (see also [43] for a more concise proof based on the fundamental solution). We shall prove the dispersive estimate using Fourier methods in the next section, but here it is instructive to consider the particularly simple case $n = 3$, when the above representation reduces to *Kirchhoff's formula*

$$\frac{\sin(t|D|)}{|D|}g = \frac{c}{t}\int_{|y|=t} g(x-y)\mathrm{d}\sigma_y\,.$$

Using the divergence theorem it is easy to deduce that

$$\left|\frac{\sin(t|D|)}{|D|}g\right| \lesssim \frac{1}{t}\|\nabla f\|_{L^1}\,.$$

Recall that the rate of decay for the Schrödinger equation in dimension 3 is $\sim t^{-3/2}$; for the wave equation the rate of decay is slower, in addition the estimate shows a loss of one derivative with respect to the initial data. Actually this estimate is optimal, as it is evident from the explicit representation of the solution. In dimension n the correct rate of decay is $t^{-\frac{n-1}{2}}$ and the loss of derivative is of order $(n-1)/2$, as we shall see in the next section.

This result is in apparent contrast with our physical space picture. Indeed the L^2 norm of the first derivatives of the solution remains constant, in view of the energy estimate

$$\|u_t(t,\cdot)\|_{L^2}^2 + \|\nabla u(t,\cdot)\|_{L^2}^2 = \|u_t(0,\cdot)\|_{L^2}^2 + \|\nabla u(0,\cdot)\|_{L^2}^2;$$

on the other hand, the velocity of propagation for the wave equation is exactly 1, so that the constant L^2 energy is spread on a volume which increases at a rate $\sim t^n$; as a consequence, we would expect the same rate of decay $\sim t^{-n/2}$ as for the Schrödinger equation. However, a more detailed study of the shape of the solution reveals that it tends to concentrate in a shell of constant thickness around the light cone. For instance, in odd dimension larger than 3 initial data with support in a ball $B(0,R)$ produce a solution which at time t is supported in the shell $B(0,t+R)\setminus B(0,t-R)$. Also in even dimension the solution tends to concentrate along the light cone $|t| = |x|$, with a faster decreasing tail far from the cone. Thus, the energy spreads over a volume of size t^{n-1}, in accordance with the rate of decay in the dispersive estimate.

The Wave Equation in Fourier Space

Since

$$2i\frac{\sin(t|D|)}{|D|}g = \frac{\mathrm{e}^{it|D|}}{|D|}g - \frac{\mathrm{e}^{-it|D|}}{|D|}g,$$
$$2\cos(t|D|)f = \mathrm{e}^{it|D|}f + \mathrm{e}^{-it|D|}f,$$

we can unify the study of the various components of the solution by considering the operator

$$\mathrm{e}^{it|D|}f = \mathcal{F}^{-1}\mathrm{e}^{it|\xi|}\mathcal{F}f = (2\pi)^{-n}\int \mathrm{e}^{i(x\cdot\xi+t|\xi|)}\widehat{f}\mathrm{d}\xi\,.$$

The phase of the last integral is

$$\phi(t,x;\xi) = t|\xi| + x\cdot\xi \quad\Longrightarrow\quad \nabla_\xi\phi = t\frac{\xi}{|\xi|} + x$$

and we notice immediately that the structure of stationary points is more complex than for Schrödinger: if $|t| = |x|$, i. e. if the point (t,x) is on the light cone, the phase has a half line of critical points

$$|t| = |x| \quad\Longrightarrow\quad \frac{\xi}{|\xi|} = -\frac{x}{t}$$

which of course are degenerate since only $n-1$ curvatures are different from zero; on the other hand, when (t,x) is not on the light cone, the phase is nondegenerate at all points.

In the nondegenerate case $|t| \neq |x|$, we notice that

$$\nabla_\xi\phi = t\frac{\xi}{|\xi|} + x \quad\Longrightarrow\quad |\nabla_\xi\phi| \geq ||t| - |x||$$

Copyrighted material

hence by direct integration by parts we obtain

$$\left| \int e^{i(x\cdot\xi + t|\xi|)} \widehat{f} d\xi \right| \le C(f) ||t| - |x||^{-(n-1)} . \tag{6}$$

Notice that the singularity $|\xi|$ of the phase makes it impossible to perform the integration by parts more than $n-1$ times, but this is largely sufficient for our purposes here.

In the degenerate case $|t| = |x|$, we can eliminate the degeneracy by passing to polar coordinates:

$$\int e^{i(x\cdot\xi + t|\xi|)} \widehat{f} d\xi = \int_0^\infty e^{it\rho} \rho^{n-1} \times \int_{|\omega|=1} e^{ix\cdot\rho\omega} \widehat{f}(\rho\omega) d\omega d\rho .$$

Now the phase in the inner integral has exactly two isolated critical points for each value of x; by a standard application of the stationary phase method we can obtain an estimate of the form

$$\left| \int e^{i(x\cdot\xi + t|\xi|)} \widehat{f} d\xi \right| \le C(f) |x|^{-\frac{n-1}{2}} . \tag{7}$$

The power $(n-1)/2$ is directly related to the dimension $n-1$ of the unit sphere.

From the above computations it is possible to obtain a very detailed description of the shape of the solution (see e. g. [20]). However, here we are only interested in the rate of decay of the L^∞ norm of the solution. Thus, using estimate (6) when $|x| < |t|/2$ and (7) when $|x| > |t|/2$ we obtain

$$|e^{it|D|} f| \le C(f) |t|^{-\frac{n-1}{2}} .$$

It is also possible to make precise the form of the constant $C(f)$ appearing in the estimate. This can be done in several ways; one of the most efficient exploits the scaling properties of the wave equation (see [10,23]). In order to state the optimal estimate we briefly recall the definition of some function spaces. The basic ingredient is the homogeneous *Paley–Littlewood partition of unity* which has the form

$$\sum_{j\in\mathbb{Z}} \phi_j(\xi) = 1 , \quad \phi_j(\xi) = \phi_0(2^{-j}\xi) , \quad \sup \phi_j = \{\xi : 2^{j-1} \le |\xi| \le 2^{j+1}\}$$

for a fixed $\phi_0 \in C_0^\infty(\mathbb{R}^n)$ (the following definitions are independent of the precise choice of ϕ_0). To each symbol $\phi_j(\xi)$ we associate by the standard calculus the operator $\phi_j(D) = \mathcal{F}^{-1}\phi_j(\xi)\mathcal{F}$. Now the *homogeneous Besov norm* $\dot{B}^s_{1,1}$ can be defined as follows:

$$\|f\|_{\dot{B}^s_{1,1}} = \sum_{j\in\mathbb{Z}} 2^{js} \|\phi_j(D) f\|_{L^1} \tag{8}$$

and the corresponding Besov space can be obtained by completing C_0^∞ (see [4] for a thorough description of these definitions).

Then the optimal dispersive estimate for the wave equation is

$$|e^{it|D|} f| \le C \|f\|_{\dot{B}^{\frac{n-1}{2}}_{1,1}} |t|^{-\frac{n-1}{2}} . \tag{9}$$

We might add that in odd space dimension a slightly better estimate can be proved involving only the standard Sobolev norm $\dot{W}^{\frac{n-1}{2}}$ instead of the Besov norm.

The singularity in t at the right hand side of (9) can be annoying in applications, thus it would be natural to try to replace $|t|$ by $(1+|t|)$ by some different proof. However, this is clearly impossible since both sides have the same scaling with respect to the transformation $u(t,x) \to u(\lambda t, \lambda x)$ which takes a solution of the homogeneous wave equation into another solution. Moreover, if such an estimate were true, taking the limit as $t \to 0$ we would obtain a false Sobolev embedding. Thus, the only way to get a nonsingular estimate is to replace the Besov norm of the data by some stronger norm.

Indeed, for $|t| < 1$ we can use the L^2 conservation which implies, for any $\epsilon > 0$,

$$|e^{it|D|} f| \le \|e^{it|D|} f\|_{H^{n/2+\epsilon}} = \|f\|_{H^{n/2+\epsilon}} .$$

Combining the two estimates, we arrive at

$$|e^{it|D|} f| \lesssim \left(\|f\|_{\dot{B}^{\frac{n-1}{2}}_{1,1}} + \|f\|_{H^{n/2+\epsilon}} \right) (1+|t|)^{-\frac{n-1}{2}} . \tag{10}$$

Other Dispersive Equations

Dispersive estimates can be proved for many other equations; see for instance [3] for general results concerning the equations of the form

$$iu_t = P(D)u , \quad P \text{ polynomial}$$

which correspond to integrals of the form

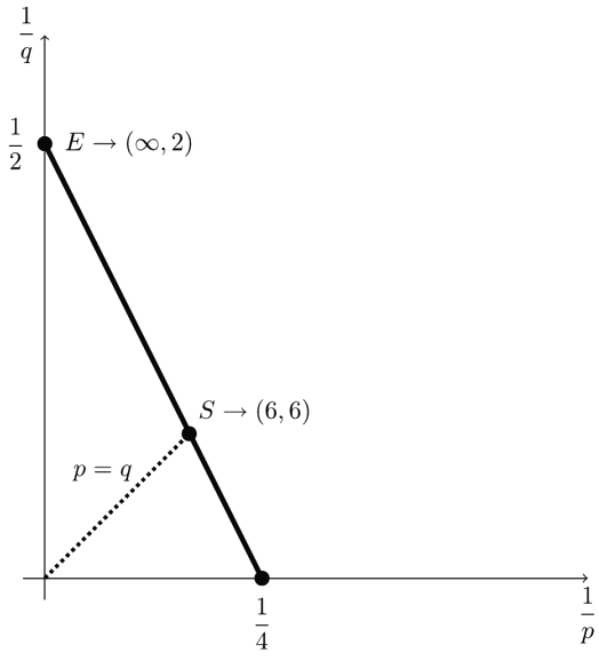
$$\int e^{i(x\cdot x i + P(\xi))} \widehat{f}(\xi) d\xi .$$

Here we would like to add some more details on two equations of particular importance for physics. The *Klein–Gordon* equation

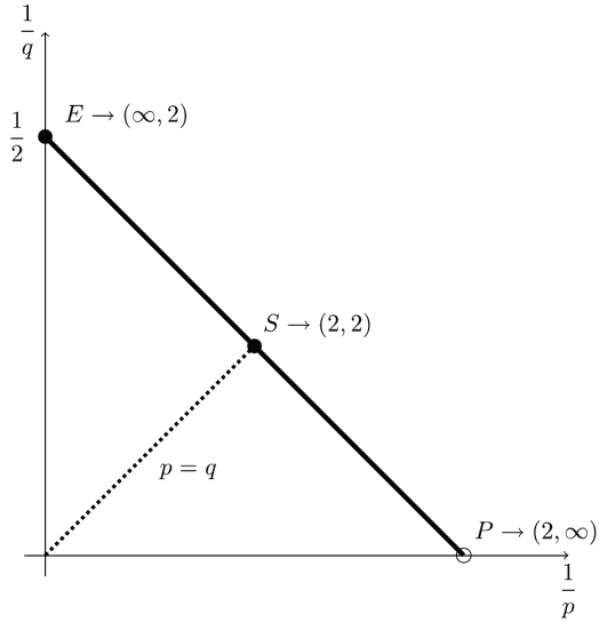
$$u_{tt} - \Delta u + u = 0 , \quad u(0,x) = f(x) , \quad u_t(0,x) = g(x)$$

shows the same decay rate as the Schrödinger equation, although for most other properties it is very close to the wave

Copyrighted material

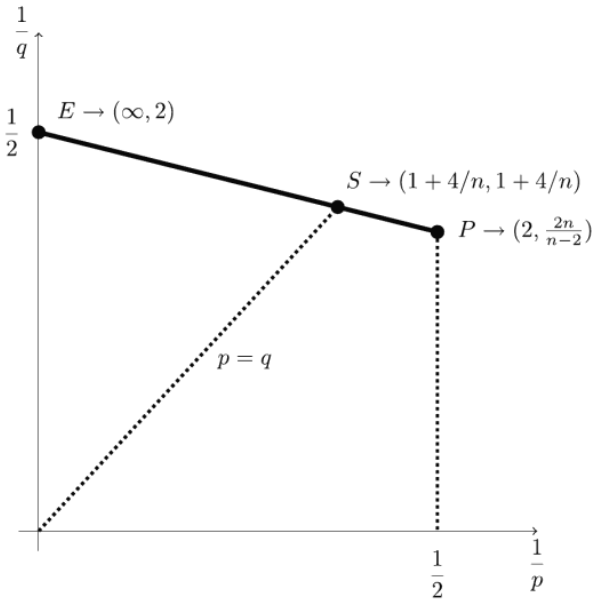


Dispersion Phenomena in Partial Differential Equations, Figure 1
Schrödinger admissible couples, $n = 1$



Dispersion Phenomena in Partial Differential Equations, Figure 2
Schrödinger admissible couples, $n = 2$

for all admissible couples (p, q) and $(\tilde{p}, \tilde{q})$, which can be chosen independently. Here p' denotes the conjugate exponent to p. These estimates continue to be true in their *local* version, with the norms at both sides restricted to a fixed time interval $I \subset \mathbb{R}$. Notice that the inhomogeneous global estimates are optimal since other combination of indices are excluded by scale invariance; however, the local Strichartz estimates can be slightly improved allowing for indices outside the admissible region (see [21]).



Dispersion Phenomena in Partial Differential Equations, Figure 3
Schrödinger admissible couples, $n \geq 3$

We also mention that a much larger set of estimates can be deduced from the above ones by combining them with Sobolev embeddings; in the above diagrams, all couples in the region bounded by the line of admissibility, the axes and the line $1/p = 2$ can be reached, while the external region is excluded by suitable counterexamples (see [30]).

Strichartz Estimates for the Wave Equation

Most of the preceding section has a precise analogue for the wave equation. Since the decay rate for the dispersive estimates is $t^{-(n-1)/2}$ for the wave equation instead of $t^{-n/2}$ as for the Schrödinger equation, this causes a shift $n \to n-1$ in the numerology of indices. A more serious difference is the loss of derivatives: this was already apparent in the original estimate by Strichartz [44], which can be written

$$\|\mathrm{e}^{it|D|} f\|_{L^{\frac{2(n+1)}{n-1}}} \lesssim \|f\|_{\dot{H}^{1/2}}, \quad n \geq 2 .$$

The general homogeneous Strichartz estimates for the wave equation take the form

$$\|\mathrm{e}^{it|D|} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{1}{p}\,\frac{n+1}{n-1}}} ;$$

Copyrighted material

now the *(wave) admissible* couples of indices must satisfy the scaling condition

$$\frac{2}{p} + \frac{n-1}{q} = \frac{n-1}{2} \tag{22}$$

and the constraints $2 \le p \le \infty$,

$$2 \le q \le \frac{2(n-1)}{n-3} \quad \text{if} \quad n \ge 3\,, \quad q \ne \infty \tag{23}$$

$$2 \le q \le \infty \quad \text{if} \quad n = 2\,, \tag{24}$$

while p varies accordingly between 2 and ∞ (between 4 and ∞ in dimension 2). Recall that in dimension 1 the solutions of the wave equation do not decay since they can be decomposed into traveling waves. We omit the graphic representation of these sets of conditions since it is completely analogous to the Schrödinger case (apart from the shift $n \to n-1$).

The inhomogeneous estimates may be written

$$\left\| \int_0^t \mathrm{e}^{i(t-s)\Delta} F(s)\mathrm{d}s \right\|_{L^p \dot{H}_q^{-\frac{1}{p}\frac{n+1}{n-1}}} \lesssim \|F\|_{L^{\tilde{p}'} \dot{H}_{q'}^{-\frac{1}{\tilde{p}}\frac{n+1}{n-1}}} \tag{25}$$

in terms of the homogeneous Sobolev norms

$$\|g\|_{\dot{H}_q^s} = \||D|^s g\|_{L^q}\,, \quad |D|^s = (1-\Delta)^{s/2}\,.$$

Other Equations

The general procedure of Keel and Tao can be applied to a large variety of equations. For the convenience of the reader we list here the precise form of the homogeneous Strichartz estimates in some important cases.

For the *Klein–Gordon* equation we have

$$\|\mathrm{e}^{it\langle D\rangle} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{1}{p}\frac{n+2}{n}}}$$

with (p, q) Schrödinger admissible (recall that the rate of dispersion is $t^{-n/2}$ for Klein–Gordon). For the *massless Dirac* system we have

$$\|\mathrm{e}^{it\mathcal{D}} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{2}{p}}}\,, \quad n = 3\,,$$

with (p, q) wave admissible in dimension 3, while for the *massive Dirac* system the estimate takes the form

$$\|\mathrm{e}^{it(\mathcal{D}+\beta)} f\|_{L^p L^q} \lesssim \|f\|_{H^{\frac{5}{3p}}}\,, \quad n = 3$$

with (p, q) Schrödinger admissible in dimension 3.

The Nonlinear Wave Equation

In this section we shall illustrate, using the semilinear wave equation as a basic example, what kind of improvements can be obtained over the classical energy method via the additional information given by dispersive techniques.

From the linear theory it is well known that the correct setting of the Cauchy problem for the wave equation

$$u_{tt} - \Delta u = F(t,x)\,, \quad u(t,x)\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{C}$$

requires two initial data at $t = 0$:

$$u(0,x) = u_0(x)\,, \quad u_t(0,x) = u_1(x)\,.$$

For this equation the energy method is particularly simple. If we multiply the equation by $\overline{u}_t$ we can recast it in the form

$$\partial_t \left(|u_t|^2 + |\nabla u|^2 \right) = 2\Re \nabla \cdot (\overline{u}_t \nabla u) + 2\Re(\overline{u}_t F)$$

and integrating this identity on different domains we can obtain several conservation laws. Integrating on a slab $0 \le t \le T$, $x \in \mathbb{R}^n$ we arrive at an estimate of the form

$$\begin{aligned} &\|u_t\|_{L_T^\infty L^2} + \|\nabla u\|_{L_T^\infty L^2} \\ &\qquad \lesssim \|u_1\|_{L^2} + \|\nabla u_0\|_{L^2} + \|F\|_{L_T^1 L^2} \end{aligned} \tag{26}$$

where we are using the shorthand notation

$$L_T^p L^q = L^p(0, T; L^q(\mathbb{R}^n))\,.$$

Estimate (26) is the basic *energy estimate*, which is enough to prove the existence of a global solution $u \in C(\mathbb{R}; H^1(\mathbb{R}^n)) \cap C^1(\mathbb{R}; L^2(\mathbb{R}^n))$ of the linear wave equation for data in $(u_0, u_1) \in H^1 \times L^2$ and a forcing term in $F \in L_T^1 L^2$ for all T. Uniqueness and continuous dependence on the data are simple consequences of the energy estimate. Moreover, if we differentiate the equation k times, the same computations lead to the higher order energy estimates

$$\begin{aligned} &\|u_t\|_{L_T^\infty H^k} + \|u\|_{L_T^\infty H^{k+1}} \\ &\lesssim \|u_1\|_{H^k} + \|u_0\|_{H^{k+1}} + \|F\|_{L_T^1 H^k}\,, \quad k = 0, 1, 2, \ldots \end{aligned} \tag{27}$$

Consider now the nonlinear wave equation

$$u_{tt} - \Delta u = F(u) \tag{28}$$

for some function $F(s)$ sufficiently smooth. The classical approach to solve it is to look for a fixed point of the mapping $u = \Phi(v)$, where $u(t,x)$ solves the linearized equation

$$u_{tt} - \Delta u = F(v)\,. \tag{29}$$

Copyrighted material

If we apply the energy estimate to the difference of two solutions $u_1 = \Phi(v_1)$ and $u_2 = \Phi(v_2)$ with the same initial data, we obtain

$$\|\partial_t(u_1 - u_2)\|_{L^\infty_T H^k} + \|u_1 - u_2\|_{L^\infty_T H^{k+1}} \lesssim \|F(v_1) - F(v_2)\|_{L^1_T H^k} .$$

Using a standard Moser-type estimate (see e. g. [25,49])

$$\|f(v_1) - f(v_2)\|_{H^k} \lesssim C(\|v_1\|_{L^\infty} + \|v_2\|_{L^\infty})\|v_1 - v_2\|_{H^k}$$

we arrive at the inequality

$$\|\partial_t(u_1 - u_2)\|_{L^\infty_T H^k} + \|u_1 - u_2\|_{L^\infty_T H^{k+1}} \lesssim C(\|v_1\|_{L^\infty_T L^\infty} + \|v_2\|_{L^\infty_T L^\infty}) \cdot \|v_1 - v_2\|_{L^1_T H^k} .$$

In particular, writing $X_T = L^\infty_T H^k$, and noticing that

$$\|G\|_{L^1_T H^k} \le \|G\|_{X_T} \cdot T ,$$

we obtain

$$\|u_1 - u_2\|_{X_T} \lesssim C(\|v_1\|_{L^\infty_T L^\infty} + \|v_2\|_{L^\infty_T L^\infty}) \cdot \|v_1 - v_2\|_{X_T} \cdot T .$$

We still have to handle the L^∞ norm of v_j, which can be controlled by the H^k norm provided $k > n/2$. In conclusion

$$k > \frac{n}{2} \quad \Longrightarrow \quad \|u_1 - u_2\|_{X_T} \lesssim C(\|v_1\|_{X_T} + \|v_2\|_{X_T}) \cdot \|v_1 - v_2\|_{X_T} \cdot T .$$

At this point the philosophy of the method should be clear: if we agree to differentiate $n/2$ times the equation, we can prove a nonlinear estimate, with a coefficient which can be made small for T small. Thus, for $T \ll 1$ the mapping $v \mapsto u$ is a contraction, and we obtain a local (in time) solution of the nonlinear equation.

The energy method is quite elementary, yet it has the remarkable property that it can be adapted to a huge number of equations: it can be applied to equations with variable coefficients, and even to fully nonlinear equations. An outstanding example of its flexibility is the proof of the local solvability of Einstein's equations ([14]).

The main drawbacks of the energy method are:

1) A high regularity of the data and of the solution is required;
2) The nonlinear nature of the estimate leads to a *local* existence result, leaving completely open the question of global existence.

Actually, the two difficulties are related to each other. Indeed, from the above argument one can check that the lifespan of the local solution is a function of the H^k norm of the initial data. Most equations with a physical origin possess conserved quantities, which are essentially H^k norms of the solution with k low, e. g. $k = 0$ or 1. Assume now that we can prove the local existence for data say in H^1, on some interval $[0, T_0]$ with T_0 depending on the H^1 norm of the data. Since this norm is conserved, we can take T_0 as a new initial time and construct a solution on $[T_0, 2T_0]$, and so on. This is the essence of the *continuation method*. As a consequence, the local existence of low regularity solution with the addition of suitable conservation laws can lead to global existence.

We shall now apply the dispersive information to improve on the preceding result and obtain the global existence of small solution, under suitable assumptions on the nonlinear term. Consider a free wave $w(t, x)$, solution of

$$w_{tt} - \Delta w = 0 , \quad w(0, x) = f(x) , \quad w_t(0, x) = g(x) .$$

Recalling the dispersive estimate (10), here the following simplified version

$$|w(t, x)| \le C(1 + t)^{-\frac{n-1}{2}} \|f\|_{W^{N,1}} , \quad N > n$$

will be sufficient. This suggests modification of the energy norm by adding a term that incorporates the decay information. To this end, we introduce the time-dependent quantity

$$M(t) = \sup_{0 \le \tau \le t} \left\{ \|u(\tau, \cdot)\|_{H^{k+1}} + (1 + \tau)^{\frac{n-1}{2}} \|u(\tau, \cdot)\|_{L^\infty} \right\} ,$$
$$k = \left[\frac{n}{2} \right] .$$

Consider now the nonlinear Eq. (28) with initial data u_0, u_1. We already know that a local solution exists, provided the initial data are smooth enough; we shall now try to extend this solution to a global one. Using Duhamel's principle we rewrite the equation as

$$u(t, x) = w(t, x) + \int_0^t \frac{\sin((t - s)|D|)}{|D|} F(u(s, \cdot))\mathrm{d}s$$

where we are using the symbolic notations of Sect. "Other Dispersive Equations". $\mathcal{F}$ being the space Fourier transform. Note that the above dispersive estimate can be applied in particular to the last term; we obtain

$$\left| \frac{\sin((t - s)|D|)}{|D|} F(s) \right| \le C(1 + t - s)^{-\frac{n-1}{2}} \|F\|_{W^{N,1}} .$$

Copyrighted material

By this inequality and the energy estimate already proved, in a few steps we arrive at the estimate

$$M(T) \lesssim C_0 + \|F\|_{L^1_T H^k} + (1+T)^{\frac{n-1}{2}} \int_0^T (1+T-s)^{-\frac{n-1}{2}} \|F(u(s))\|_{W^{N,1}} \, ds$$

where $C_0 = \|u_0\|_{W^{N,1}} + \|u_1\|_{W^{N,1}}$ and the mute constant is independent of T. By suitable nonlinear estimates of Moser type, we can bound the derivatives of the nonlinear term using only the L^∞ and the H^k norm of u, i. e., the quantity $M(t)$ itself. If we assume that $F(u) \sim |u|^\gamma$ for small u, we obtain

$$M(T) \lesssim C_0 + M(T)^\gamma (1+T)^{\frac{n-1}{2}} \int_0^T (1+T-s)^{-\frac{n-1}{2}} (1+s)^{-\frac{n-1}{2}(\gamma-2)} ds \, .$$

Now we have

$$\int_0^t (1+t-\tau)^{-\frac{n-1}{2}} \cdot (1+\tau)^{-\frac{n-1}{2}\cdot(\gamma-2)} d\tau \leq C(1+t)^{-\frac{n-1}{2}}$$

provided

$$\gamma > 2 + \frac{2}{n-1}$$

so that, for such values of γ, we obtain

$$M(T) \lesssim C_0 + M(T)^\gamma .$$

For small initial data, which means small values of the constant C_0, (30) implies that $M(T)$ is uniformly bounded as long as the smooth solution exists. By a simple continuation argument, this implies that the solution is in fact global, provided the initial data are small enough.

The Nonlinear Schrödinger Equation

Strichartz estimates for the Schrödinger equation are expressed entirely in terms of L^p norms, without loss of derivatives, in contrast with the case of the wave equations. For this reason they represent a perfect tool to handle semilinear perturbations. Using Strichartz estimates, several important results can be proved with minimal effort. As an example, we chose to illustrate in this Section the critical and subcritical well posedness in L^2 in the case of power nonlinearities (see [13] for a systematic treatment).

Consider the Cauchy problem

$$iu_t - \Delta u = |u|^\gamma , \quad u(0,x) = f(x) \tag{31}$$

for some $\gamma \geq 1$. We want to study the local and global solvability in L^2 of this Cauchy problem. Recall that this is a difficult question, still not completely solved; however, at least in the subcritical range, a combination of dispersive techniques and the contraction mapping principle gives the desired result easily.

Instead of Eq. (31), we are actually going to solve the corresponding integral equation

$$u(t,x) = e^{it\Delta} f + i \int_0^t e^{i(t-s)\Delta} |u(s)|^\gamma ds \, .$$

Our standard approach is to look for a fixed point of the mapping $u = \Phi(v)$ defined by

$$\Phi(v) = e^{it\Delta} f + i \int_0^t e^{i(t-s)\Delta} |v(s)|^\gamma ds \, ;$$

it is easy to check, using Strichartz estimates, that the mapping Φ is well defined, provided the function v is in the suitable $L^p L^q$. Indeed we can write, for all admissible couples (p,q) and $(\tilde{p},\tilde{q})$

$$\|\Phi(v)\|_{L^p L^q} \lesssim \|f\|_{L^2} + \||\Phi(v)|^\gamma\|_{L^{\tilde{p}'} L^{\tilde{q}'}}$$

and by Hölder's inequality this implies

$$\|\Phi(v)\|_{L^p L^q} \leq C_1 \|f\|_{L^2} + C_1 \|v\|^\gamma_{L^{\gamma\tilde{p}'} L^{\gamma\tilde{q}'}} \, . \tag{32}$$

Thus, we see that the mapping Φ operates on the space $L^p L^q$, provided we can satisfy the following set of conditions:

$$\begin{gathered} \tilde{p}'\gamma = p \, , \quad \tilde{q}'\gamma = q \, , \\ \frac{2}{p} + \frac{n}{q} = \frac{n}{2} \, , \quad \frac{2}{\tilde{p}} + \frac{n}{\tilde{q}} = \frac{n}{2} \, , \\ p, \tilde{p} \in [2, \infty] \, , \quad \tilde{q} \in \left[2, \frac{2n}{n-2}\right] \, . \end{gathered} \tag{30}$$

Notice that from the first two identities we have

$$\frac{2\gamma}{p} + \frac{n\gamma}{q} = \alpha - \frac{2}{\tilde{p}} + n - \frac{n}{\tilde{q}}$$

and by the admissibility conditions this implies

$$\frac{n}{2}\gamma = 2 + n - \frac{n}{2} \, ,$$

which forces the value of γ to be

$$\gamma = 1 + \frac{4}{n} \, . \tag{33}$$

This value of γ is called L^2 *critical*.

Copyrighted material

Let us assume for the moment that γ is exactly the critical exponent (33). Then it is easy to check that there exist choices of (p, q) and $(\tilde{p}, \tilde{q})$ which satisfy all the above constraints. To prove that Φ is a contraction on the space

$$X = L^p L^q$$

we take any $v_1, v_2 \in X$ and compute

$$\begin{aligned}\|\Phi(v_1) - \Phi(v_2)\|_X &\lesssim \left\| |v_1|^\gamma - |v_2|^\gamma \right\|_{L^{\tilde{p}'} L^{\tilde{q}'}} \\ &\lesssim \left\| |v_1 - v_2| (|v_1|^{\gamma-1} + |v_2|^{\gamma-1}) \right\|_{L^{\tilde{p}'} L^{\tilde{q}'}} ;\end{aligned}$$

the same computation as above gives

$$\|\Phi(v_1) - \Phi(v_2)\|_X \le C_2 \|v_1 - v_2\|_X \| |v_1| + |v_2| \|_X^{\gamma-1} \quad (34)$$

for some constant C_2. Now assume that

$$\|v_i\|_X < \varepsilon , \quad \|f\|_{L^2} < \delta ;$$

using (32) we can write

$$\|\Phi(v)\|_X \le C_1(\delta + \varepsilon^\gamma) < \varepsilon$$

provided ε, δ are so small that

$$C_1 \delta < \frac{\varepsilon}{2} , \quad C_1 \varepsilon^\gamma < \frac{\varepsilon}{2} .$$

On the other hand, by (34) we get

$$\|\Phi(v_1) - \Phi(v_2)\|_X \le C_2 \|v_1 - v_2\|_X (2\varepsilon)^{\gamma-1} < \frac{1}{2} \|v_1 - v_2\|_X$$

provided ε is so small that

$$C_2 (2\varepsilon)^{\gamma-1} < \frac{1}{2} .$$

In conclusion, if the initial data are small enough, the mapping Φ is a contraction on a ball $B(0, \varepsilon) \subset X$ and the unique fixed point is a global small solution of the critical L^2 equation

$$iu_t - \Delta u = |u|^{1+\frac{4}{n}} .$$

Notice that in estimate (32) the first couple can also be taken equal to $(\infty, 2)$, thus the solution we have constructed belongs to $L^\infty L^2$. A limit argument (dominated convergence) in the identity

$$u = \mathrm{e}^{it\Delta} f + i \int_0^t \mathrm{e}^{i(t-s)\Delta} |u(s)|^\gamma \mathrm{d}s ;$$

proves that we have actually $u \in C(\mathbb{R}; L^2(\mathbb{R}^n))$.

We briefly sketch a proof of the local existence in the subcritical case $1 < \gamma < 1 + \frac{4}{n}$. Consider the local space

$$X_T = L_T^p L^q = L^p(0, T; L^q(\mathbb{R}^n)) .$$

If we follow the steps of the preceding proof and choose the same indices, we have now

$$\tilde{p}' \gamma < p , \quad \tilde{q}' \gamma < q .$$

It is easy to check that we can shift the point $(\tilde{p}, \tilde{q})$ so that

$$\tilde{p}' \gamma < p , \quad \tilde{q}' \gamma = q$$

and proceeding as above and using Hölder inequality in time, we obtain the inequality. Indeed, choosing the indices as above and applying Hölder's inequality in time we have

$$\|\Phi(v)\|_{X_T} \le C_1 \|f\|_{L^2} + C_1 T^\lambda \|v\|_{X_T} \quad (35)$$

for some strictly positive $\lambda > 0$. Analogously, estimate (34) will be replaced by

$$\begin{aligned}&\|\Phi(v_1) - \Phi(v_2)\|_{X_T} \\ &\qquad \le C_2 T^\lambda \|v_1 - v_2\|_{X_T} \| |v_1| + |v_2| \|_{X_T}^{\gamma-1} . \quad (36)\end{aligned}$$

Now let M be so large and T so small that

$$\|f\|_{L^2} < \frac{M}{2C_1} , \quad C_1 T^\lambda < \frac{M}{2}$$

we obtain from (35) that Φ takes the ball $B(0, M) \subset X_T$ into itself; if in addition we have also, by possibly taking T smaller,

$$C_2 T^\lambda (2M)^{\gamma-1} < \frac{1}{2}$$

we see by (36) that Φ is a contraction, and in conclusion we obtain the existence of a unique local solution to the subcritical equation. By the same argument as above this solution is in $C([0, T]; L^2(\mathbb{R}^n))$.

We notice that in the above proof the quantity M is only determined by the L^2 norm of f, and hence also the lifespan of the local solution is a function of $\|f\|_{L^2}$. If for some reason we know a priori that the L^2 norm of the solution is conserved by the evolution, this remark implies immediately that the solution is in fact global, by an elementary continuation argument. This is the case for instance of the *gauge invariant* equation

$$iu_t - \Delta u = \pm |u|^{\gamma-1} u$$

for which any solution continuous with values in L^2 satisfies $\|u(t)\|_{L^2} = \|u(0)\|_{L^2}$. The above computations apply without modification and we obtain the global well posedness for all subcritical exponents γ.

Copyrighted material

Dynamics on Fractals

RAYMOND L. ORBACH
Department of Physics, University of California, Riverside, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Fractal and Spectral Dimensions
Nature of Dynamics on Fractals – Localization
Mapping of Physical Systems onto Fractal Structures
Relaxation Dynamics on Fractal Structures
Transport on Fractal Structures
Future Directions
Bibliography

Glossary

Fractal Fractal structures are of two types: deterministic fractals and random fractals. The former involves repeated applications of replacing a given structural element by the structure itself. The process proceeds indefinitely, leading to *dilation symmetry*: if we magnify part of the structure, the enlarged portion looks the same as the original. Examples are the Mandelbrot–Given fractal and the Sierpinski gasket. A random fractal obeys the same properties (e. g. dilation symmetry), but only in terms of an ensemble average. The stereotypical example is the percolating network. Fractals can be constructed in any dimension, d, with for example, $d = 6$ being the mean-field dimension for percolating networks.

Fractal dimension The fractal dimension represents the "mass" dependence upon length scale (measuring length). It is symbolized by D_f, with the number of sites on a fractal as a function of the measurement length L being proportional to L^{D_f}, in analogy to a homogeneous structure embedded in a dimension d having mass proportional to the volume spanned by a length L proportional to L^d.

Spectral (or fracton) dimension The spectral (or fracton) dimension refers to the dynamical properties of fractal networks. It is symbolized by $\tilde{d}_s$ and can be most easily thought of in terms of the density of states of a dynamical fractal structure (e. g. vibrations of a fractal network). Thus, if the excitation spectrum is measured as a function of energy ω, the density of states for excitations of a fractal network would be proportional to $\omega^{(\tilde{d}_s-1)}$, in analogy to a homogeneous structure embedded in a dimension d having a density of states proportional to $\omega^{(d-1)}$.

Localization exponent Excitations on a fractal network are in general *strongly localized* in the sense that wave functions fall off more rapidly than a simple exponential, $\Psi(r) \sim \exp[-\{[r/\Lambda(\omega)]^{d_\phi}\}]$ where $\Lambda(\omega)$ is an energy dependent localization length, and the exponent d_ϕ is in general greater than unity.

Definition of the Subject

The dynamical properties of fractal networks are very different from homogeneous structures, dependent upon strongly localized excitations. The thermal properties of fractals depend upon a "spectral dimension" $\tilde{d}_s$ less than the "Euclidean" or embedding dimension d. Fractal randomness introduces statistical properties into relaxation dynamics. Transport takes place via hopping processes, reminiscent of Mott's variable range-rate hopping transport for electronic impurity states in semi-conductors. Fractal dynamics serve as a guide for the behavior of random structures where the short length scale excitations are localized.

Introduction

The study of the dynamics on and of fractal networks [29,30,31] is not an arcane investigation, with little application to real physical systems. Rather, there are many examples of real materials that exhibit fractal dynamics. Examples are the vibrational properties of silica aerogels [21,45], and magnetic excitations in diluted antiferromagnets [18,19,43]. Beyond these explicit physical realizations, one can learn from the very nature of the excitations on fractal structures how localized excitations in homogeneous materials behave. That is, the dynamics of fractal networks serve as a calculable model for transport of localized excitations in random structures that are certainly not mass fractals. Examples are thermal transport in glasses above the so-called "plateau" temperature [2,20,26], and the lifetime of high-energy lattice vibrations in a-Si [32,33,36,37]. As we shall show in subsequent sections, the former is an example of vibrational hopping-type transport (similar in nature to Mott's variable-range hopping [22,23] for localized electronic states), while the latter exhibits counter-intuitive energy dependences for vibrational lifetimes.

The following Sect. "Fractal and Spectral Dimensions" describes in brief terms the concept of the fractal (or mass) dimension, D_f [7,8,9,10,15,40,41,46], and the spectral (or fracton) dimension $\tilde{d}_s$ [5,14,17,34,35]. The latter is re-

Copyrighted material

lated to the anomalous diffusion characteristics of fractals (the so-called "ant in a labyrinth" introduced by de Gennes [8]), and is at the heart of the thermal properties of fractal structures. A conjecture introduced by Alexander and Orbach [5] suggested that the mean-field value (exact at $d = 6$) for $\tilde{d}_s$ for percolating networks, $\tilde{d}_s = 4/3$, might be universal, independent of dimension. It is now generally regarded that this conjecture is only approximate, though a very good approximation for percolating networks in $2 = d < 6$. Were it to be exact, it would mean that the dynamical properties of percolating networks could be expressed in terms of their geometrical properties.

Section "Nature of Dynamics on Fractals–Localization" introduces the nature of the wave function for an excitation embedded in a fractal structure. In general, random self-similar structures generate "super-localized" wave functions, falling off faster than exponential. We will characterize the envelope of the wave function by the functional form [1]

$$\Psi(r) \sim \exp[-\{[r/\Lambda(\omega)]^{d_\phi}\}] ,$$

where $\Lambda(\omega)$ is an energy-dependent localization length, and the exponent d_ϕ is in general greater than unity. For excitations on fractal structures, the localization length $\Lambda(\omega)$ can be calculated analytically, allowing for explicit computation of physical phenomena. This allows analytic expressions for scattering lifetimes and excitation transport.

Mapping of the diffusion problem onto the secular equation for scalar lattice vibrations allows the extraction of the density of vibrational states [5]. It is given by the simple relationship $D(\omega) \sim \omega^{(\tilde{d}_s - 1)}$ displaying the utility of the spectral dimension $\tilde{d}_s$. A similar though slightly more complicated mapping can be applied to diluted Heisenberg antiferromagnets. The silica aerogels [13,16,42,44,45], and $Mn_xZn_{1-x}F_2$ [43] and $RbMn_xMg_{1-x}F_3$ [18,19] are realizations of such systems, respectively, and have been extensively investigated experimentally.

The random nature of fractal structures introduces statistical complexity into relaxation processes. The direct (single vibration) [38] and Raman (scattering of two vibrations) [24,39] spin-lattice relaxation processes are discussed for random fractal networks in Sect. "Relaxation Dynamics on Fractal Structures". The theoretical results are complex, related to the "Devil's staircase" for Raman relaxation processes. Experiments probing these rather bizarre behaviors would be welcome.

Given the "super-localization" of excitations on fractal networks [1,42], transport properties need to be expressed according to the same concepts as introduced by Mott [22,23] for the insulating phase of doped semiconductors—the so-called "variable range-rate hopping". This is developed in Sect. "Transport on Fractal Structures", and applied to the thermal transport of fractal structures [2,20,26]. The "lessons learned" from these applications suggest an interpretation of the thermal transport of glassy materials above the so-called "plateau" temperature range. This instance of taking the insights from fractal structures, and applying it to materials that are certainly not fractal yet possess localized vibrational states, is an example of the utility of the analysis of fractal dynamics.

Finally, some suggestions, hinted at above, for future experimental investigations are briefly discussed in Sect. "Future Directions".

Fractal and Spectral Dimensions

Fractals can be defined by the nature of the symmetry they exhibit: self-similar geometry or equivalently dilation symmetry [7,8,9,10,15,40,41,46]. This symmetry is most assuredly not translational, leading to a very different set of behaviors for physical realizations that shall be explored in subsequent Sections. Stated most simply, self-similar structures "look the same" on any length scale. Conversely, one cannot extract a length scale from observation of a self-similar structure. There are clearly limits to self-similarity for physical realizations. For example, at the atomic level, dilation symmetry must come to an end. And for finite structures, there will be a largest length scale. There are other length scales that bracket the self-similar regime. Thus, for percolating networks, an example of a random fractal [38,39], the percolation correlation length, $\xi(p)$, where p is the concentration of occupied sites or bonds, defines an upper limit to the regime of fractal geometry. In general, for percolating networks, the range of length scales ℓ over which the structure is fractal is sandwiched between the atomic length a, and the percolation correlation length $\xi(p)$: $a < \ell < \xi(p)$.

Fractal structures can be deterministic or random. The former is a result of applying a structural rule indefinitely, replacing an element of a design by the design itself. Examples are the Mandelbrot–Given fractal and the Sierpinski gasket. The latter arises from a set of probabilistic rules. A simple example is the percolating network where a site or bond (e. g. in $d = 2$, a square grid) is occupied randomly with probability p. The resulting clusters contain a random number of sites or bonds, where the distribution function for the number of finite clusters of size s is denoted by $n_s(p)$. There exists a critical value $p = p_c$ where a connected cluster extends over all space (the "in-

Copyrighted material

finite" cluster). For $p < p_c$ only finite clusters exist, with the largest cluster spanned by the length scale $\xi(p)$. $\xi(p)$ diverges at $p = p_c$ as $|p - p_c|^{-\nu}$. For $p > p_c$ finite clusters and the infinite cluster coexist, the number of finite clusters vanishing at p approaches unity. The probability that a site belongs to the infinite cluster varies with occupation probability p as $(p - p_c)^{\beta}$.

The characteristic length $\xi(p)$ for $p > p_c$ is a measure of the largest finite cluster size. Hence, for length scales $\ell > \xi(p)$ the systems looks homogeneous. Thus, the fractal regime is sandwiched between the lattice constant a and the percolation correlation length $\xi(p)$. The remarkable utility of the percolating network becomes clear from these properties: there is a crossover length scale between fractal and homogeneous structural behavior that can be chosen at will by choosing a value for the site or bond occupation probability $p > p_c$. Excitations of structures that map onto percolating networks can change their nature from fractal to continuous as a function of their length scale ℓ, the transition occurring when $\ell_c \sim \xi(p)$. This property makes the percolating network the "fruit fly" of random structures. The flexibility of adjusting the crossover length ℓ_c enables a fit between those properties calculated for a percolating network with the properties of those physical systems that map onto percolating networks.

The fractal dimension, D_f, is defined from the dependence of the "mass" on length scale. For fractals, the number of points or bonds on the infinite cluster within a length scale ℓ (hence, the "mass"), for $\ell \ll \xi(p)$, depends upon ℓ as $M(\ell) \sim \ell^{D_f}$. From above it is straight forward to show $D_f = d - (\beta/\nu)$. For a percolation network [29] in $d = 2, D_f = 91/48$. For $d = 3, D_f = 2.48 \pm 0.09$. In mean field, $d = 6, D_f = 4$.

The spectral dimension, $\tilde{d}_s$, follows from the analysis of diffusion on a fractal network, as first postulated by de Gennes [14]: "An ant parachutes down onto an occupied site of the infinite cluster of a percolating network. At every time unit, the ant makes one attempt to jump to one of its adjacent sites. If that site is occupied, it moves there. If it is empty, the ant stays at its original site. What is the ensemble-averaged square distance that the ant travels in time t?" Gefen et al. [17] found that $\langle r^2(t)\rangle \sim t^{2+\theta}$ where θ depends upon the scaling exponent μ (conductivity exponent), the probability that a site belongs to the infinite cluster scaling exponent β, and the scaling exponent ν for $\xi(p)$: $\theta = (\mu - \beta)/\nu$. Alexander and Orbach [5] defined the spectral (or fracton) dimension as $\tilde{d}_s = 2D_f/(2 + \theta)$. The importance of $\tilde{d}_s$can be seen from the calculation of the probability of finding a diffusing particle at the starting point at time t, $P_0(t)$, which for compact diffusion is the inverse of the number of visited sites in time t, $V(t) \sim t^{(\tilde{d}_s/2)}$. For scalar elasticity, the vibrational density of states

$$D(\omega) = -(2\omega/\pi)\mathrm{Im}\ \tilde{P}_0(-\omega^2 + i0^+) \sim \omega^{(\tilde{d}_s - 1)} ,$$

where $\tilde{P}_0(\omega)$ is the Laplace transform of $P_0(t)$. Noting that for homogeneous systems, $D(\omega) \sim \omega^{(d-1)}$, the name spectral dimension for $\tilde{d}_s$ becomes evident. Alexander and Orbach [5] named the vibrational excitations of fractal structures, fractons, in analogy with the vibrational excitations of homogeneous structures, phonons. Hence, they termed $\tilde{d}_s$the fracton dimension.

Alexander and Orbach noted that for percolation structures in the mean-field limit, $d = 6$, $\tilde{d}_s = 4/3$ precisely. At that time (1982), values for $\tilde{d}_s$ for $d < 6$ appeared close to that value, and they conjectured that $\tilde{d}_s = 4/3$ for $2 \le d \le 6$. We now know that this is only approximate (but remarkably close [24]): $\tilde{d}_s = 1.325 \pm 0.002$ in $d = 2$, and 1.317 ± 0.003 in $d = 3$. Were the conjecture exact, a numerical relationship between structure and transport would be exact $[\theta = (3D_f/2) - 2]$, dictated solely by fractal geometry.

For percolating networks, the structure appears fractal for length scales $\ell \le \xi(p)$ and homogeneous for $\ell \ge \xi(p)$. The vibrational excitations in the fractal regime are termed fractons, and in the homogeneous regime phonons. We define the crossover frequency ω_c as that for which the excitation length scale equals $\xi(p)$. Then $\omega_c \sim (p - p_c)^{(\nu D_f/\tilde{d}_s)}$. Continuity leads to a phonon velocity $v(p) \sim (p - p_c)^{\nu}[(D_f/\tilde{d}_s) - 1]$. This then leads to the dispersion relations:

$$[\ell \gg \xi(p), \omega \ll \omega_c\ ,\ \text{phonon regime}]\omega \sim v(p)k\ ,$$

and

$$[\ell \ll \xi(p), \omega \gg \omega_c\ ,\ \text{fracton regime}]\omega \sim k^{(D_f/\tilde{d}_s)}\ .$$

In the latter case, as we shall show in the next Section, vibrational excitations in the fractal regime are localized, so that k should not be thought of as a wave vector but rather the inverse of the localization length $\Lambda(\omega)$.

Nature of Dynamics on Fractals – Localization

The conductance of a d dimensional percolating network of size L, $G(L)$, can be shown to be proportional to L^{β} where the exponent $\beta = (D_f/\tilde{d}_s)(\tilde{d}_s - 2)$. In the Anderson sense [6], localization occurs when $\beta \le 0$ (marginal at $\beta = 0$). The Alexander–Orbach conjecture [5], $\tilde{d}_s \approx 4/3$, leads to β well less than zero for all embedding dimensions d. One can think of this form of localization as geometrical, as opposed to the scattering localization that traditionally is associated with Anderson localization. The lo-

Copyrighted material

calization is strong, the wave function falling off faster than simple exponential.

The wave function has the form [1]: $\langle\Psi(r)\rangle \sim \exp[-\{[r/\Lambda(\omega)]^{d_\phi}\}]$. The exponent d_ϕ is a geometrical exponent describing the fact that an exponential decay along the fractal will be distorted when viewed in real space. In general, $1 \leq d_\phi \leq d_{\min}$, where $d_{\min}$ is defined by $\ell \sim R^{d_{\min}}$, with ℓ being the shortest path along the network between two points separated by a Pythagorean distance R. Bunde and Roman [11] have found that for percolating networks $d_\phi = 1$. The decay length $\Lambda(\omega)$ is the localization length from Sect. "Introduction", $\Lambda(\omega) \sim \omega^{(-\tilde{d}_s/D_f)}$.

For a random system such as a percolating network, a given realization of $\Psi(r)$ is badly behaved, depending upon the particular choice of origin taken for the position coordinate r. The ensemble average takes the average over all realizations of $\Psi(r)$, and is denoted by $\langle\Psi(r)\rangle$. This has its dangers, as the calculation of physical properties should be performed using a specific realization for $\Psi(r)$, and then ensemble averaged. Calculation of physical properties utilizing an ensemble averaged wave function $\langle\Psi(r)\rangle$ will not yield the same result, and could be very misleading.

Specific realizations for $\Psi(r)$ for two-dimensional bond percolation vibrational networks have been calculated by Nakayama and Yakubo [25,28,47]. The fracton "core" (or largest amplitude) possesses very clear boundaries for the edges of the excitation, with an almost step-like character and a long tail. The tail extends over large distances and the amplitude oscillates in sign. This is required for orthogonality to the uniform translational mode with $\omega = 0$. Clearly these modes do not look like simple exponentials, hence the warning about the use of $\langle\Psi(r)\rangle$ as compared to $\Psi(r)$ for the calculation of matrix elements.

Mapping of Physical Systems onto Fractal Structures

The foregoing description of dynamics of fractal networks would be interesting in and of itself, but its applicability to physical systems makes it of practical importance. Further, the nature of localization, and computation of "hopping-type" transport on fractal structures, has application to materials that are not fractal. Thus, it is worthwhile to consider what systems are fractal, the nature of dynamics on such systems, and the lessons we can learn from them for structures that possess similar characteristics but are not fractal. Specifically, Mott's [22,23] variable range-rate hopping for localized electronic impurity states has direct application to transport of vibrational energy via fracton hopping on fractal networks [2], and we shall argue to thermal transport in glassy materials above the "plateau" temperature [20,26].

The fundamental equations for fractal dynamics spring from the equation for diffusion of a random walker on a network: $\mathrm{d}P_i/\mathrm{d}t = \sum_{j\neq I} w_{ij}(P_j - P_i)$ where P_i is the probability that the ith site is occupied, w_{ij} is the probability per unit time for hopping from site i to site j, and $w_{ij} = w_{ji}$. Introducing new quantities $W_{ij} = w_{ij}$ for $i \neq j$, and $W_{ii} = -\sum_{j\neq I} w_{ij}$, one can rewrite the diffusion equation as $\mathrm{d}P_i/\mathrm{d}t = \sum_j W_{ij}P_j$ with the defined relation $\sum_j W_{ij} = 0$.

For scalar elasticity, the equation of motion for atomic vibrations is $\mathrm{d}^2u_i/\mathrm{d}t^2 = \sum_j K_{ij}u_j$, with u_i the atomic displacement at position i, and $K_{ij}(i \neq j)$ is given by $K_{ij} = k_{ij}/m_i$ with k_{ij} the force constant connecting two sites i and j, and m_i the mass of the ith site. The diagonal element W_{ii} is obtained from the balancing of forces at the ith site: $\sum_j K_{ij} = 0$.

The only difference between these two equations is the order of the time derivatives, leading to the equivalence of the eigenvalue problems. We have used this equivalence to extract the density of vibrational states for scalar elasticity, and the definition of the spectral (or fracton) dimension in Sect. "Introduction".

The most direct physical example of a vibrating fractal network is the silica aerogel [13,16,42,44]. Different preparation conditions can change the density of these materials to very small values, indicative of the open porous structure of these materials. Two length scales are present: the fundamental building block of the structure (the silica), and the correlation length of the gel. In between, the clusters possess a fractal structure, while at length scales beyond the correlation length the gel behaves as a homogeneous porous glass. This crossover from fractal to homogeneous is the physical example of our previous discussion for percolating networks when the length scale of excitations passes through the percolation correlation length $\xi(p)$. The thermal, transport, and scattering properties of aerogels have been well studied, and are the classic example of fractal behavior.

There are other examples of structures that map onto the eigenvalue spectrum of the diffusion equation. Ferromagnetic and antiferromagnetic materials can be diluted by non-magnetic impurities. This leads to random site dilution, mapping directly onto the site-diluted percolating network for magnetic interactions. This can be seen from the Heisenberg Hamiltonian for spin systems: $H = 1/2\sum_{i,j} J_{ij}\mathbf{S}_i\mathbf{S}_j$, where $\mathbf{S}_i$ is the vector spin at the ith site, and J_{ij} the isotropic exchange interaction between sites i and j. The equation of motion for the spin operator $S_i^+ = S_i^x + \mathrm{i}S_i^x$ becomes for diluted ferromagnets

Copyrighted material

$i\hbar\partial S_i^+/\partial t = \sum_{j\neq i} J_{ij}(S_i^z S_j^+ - S_j^z S_i^+)$. For low levels of spin wave excitations, $S_i^z \approx S_j^z \approx S$, so that one obtains the linear equation of motion $i\hbar\partial S_i^+/\partial t = \sum_{j\neq I} J_{ij}(S_j^+ - S_i^+)$. This is the same form as that for diffusion of a random walker on a network, so that, in analogy with the scalar vibrational network, the eigenvalue spectrum is the same. This mapping allows the spin wave spectrum to exhibit magnons and fracton waves obeying the same properties that phonons and fractons exhibit for scalar vibrations. For antiferromagnets, the change in sign of S_i^z for the two sublattices complicates the linearized equations of motion. They become, $i\hbar\partial S_i^+/\partial t = \sigma_i \sum_{j\neq I} J_{ij}(S_j^+ + S_i^+)$, where $\sigma_i = -1$ for down spins and $\sigma_i = 1$ for up spins. This alters the dynamical equations for diluted antiferromagnets, and requires a separate calculation of the eigenvalue spectrum.

Neutron diffraction experiments exhibit both conventional magnon and fracton excitations. In the case of the site-diluted antiferromagnet $Mn_xZn_{1-x}F_2$ both excitations are observed simultaneously for length scales in the vicinity of $\xi(p)$ [43].

Relaxation Dynamics on Fractal Structures

Electron spin resonance and non-radiative decay experiments can be performed in fractal materials. It is of interest to compare these dynamical properties with those found in translationally ordered structures. At first sight, for vibrational relaxation, it might be thought that the only difference would originate with the differences in the densities of vibrational states $D(\omega)$. However, the localized nature of the vibrational states (fractons) and their random positions in the fractal network, introduces properties that are specific to fractal lattices. There are two cases of interest: one-fracton relaxation, and two-fracton inelastic scattering. The former is referred to as the "direct" relaxation process; the latter the "Raman" relaxation process. In both cases, the main effect of vibrational localization is on the relaxation time profile. Different spatial sites can have very different relaxation rates because of the randomness of their distance from the appropriate fracton sites. The result is a strongly non-exponential time decay.

For one-fracton relaxation [1], a straightforward generalization of the direct relaxation process rate, $W(\omega_0, L)$, for relaxation energy ω_0 arising from interaction with a fracton of the same energy centered a distance L away, is proportional to

$$W(\omega_0, L) \sim \omega_0^{2q-1}\{[\Lambda(\omega_0)]^{(-D_f)}\} \cdot [\coth(\beta\omega_0/2)](1/\delta_L)\exp\{-[L/\Lambda(\omega_0)]^{d_\phi}\}\,.$$

Here, $\beta = 1/k_B T$ and the factor δ_L represents the energy width of the fracton state. In the homogeneous limit, where phonons represent the extended vibrational states, an energy conserving delta function would replace δ_L. There are two limits for δ_L, case (a) when the fracton relaxation rate δ caused by anharmonicity dominates; and case (b) when the electronic relaxation rate caused by the electron-fracton interaction is greater than δ. In the former case, $\delta_L = \delta$; while in the latter, $\delta_L = W(\omega_0, L)$ itself, leading to a self-consistent determination of the relaxation rate. The latter case requires explicit consideration of the L dependence of δ_L, further complicating the calculation of $W(\omega_0, L)$. The calculation of the time profile of the electronic state population and the average relaxation rate are complex, and the reader is referred to [32] for details.

For case (a), the population of the initial electronic state is found to vary as

$$(\ln t)^{[(D_f - d_\phi)/2d_\phi]}\left[t^{\{-c_1(\ln t)^{[(D_f/d_\phi)-1]}\}}\right]\,.$$

Here, c_1 is a constant independent of time, but dependent upon ω_0 and δ. The population of the initial electronic state decays is thus found to be faster than a power law, but slower than exponential or stretched exponential.

For case (b), the population of the initial electronic state is found to vary as

$$(1/t)\left\{(\ln t)^{[(D_f/d_\phi)-1]}\right\}\,.$$

The decay is slower than in case (a), and is closer to a power law.

The average relaxation rates for cases (a) and (b) have the same functional form: $\langle W\rangle \sim D(\omega_0)[(\omega_0)^{(2q-1)}]\coth(\beta\omega_0/2)$ where $q = \tilde{d}_s(d_\phi/D_f)$. The temperature dependence is the usual one found for the direct process, but the energy dependence differs, depending on the values of the dimensions and parameters for fractal networks.

Two fracton relaxation is considerably more complex [3,4]. In summary, the time profile of the initial electronic state in the long time regime begins as a stretched exponential, then crosses over into a form similar to case (a) for the one fracton decay process. In the presence of rapid electronic cross relaxation, the time profile is exponential, with a low-temperature relaxation time $\langle 1/T_1\rangle$ proportional to $T^{\{2\tilde{d}_s[1+2(d_\phi/D_f)]-1\}}$ for Kramers transitions (between half-integer time-reversed spin states), and $T^{\{2\tilde{d}_s[1+2(d_\phi/D_f)]-3\}}$ for non-Kramers transitions (between integer time-reversed spin states).

The complexity of random systems adds an interesting consideration: the nature of the relaxation rate *at a specific site* as compared to the *average* relaxation rate calculated above. The decay profile of single-site spin-lattice

Copyrighted material

16. Freltoft T, Kjems J, Richter D (1987) Density of states in fractal silica smoke-particle aggregates. Phys Rev Lett 59:1212–1215
17. Gefen Y, Aharony A, Alexander S (1983) Anomalous diffusion on percolating clusters. Phys Rev Lett 50:77–80
18. Ikeda H, Fernandez-Baca JA, Nicklow RM, Takahashi M, Iwasa I (1994) Fracton excitations in a diluted Heisenberg antiferromagnet near the percolation threshold: $RbMn_{0.39}Mg_{0.61}F_3$. J Phys: Condens Matter 6:10543–10549
19. Ikeda H, Itoh S, Adams MA, Fernandez-Baca JA (1998) Crossover from Homogeneous to Fractal Excitations in the Near-Percolating Heisenberg Antiferromagnet $RbMn_{0.39}Mg_{0.61}F_3$. J Phys Soc Japan 67:3376–3379
20. Jagannathan A, Orbach R, Entin-Wohlman O (1989) Thermal conductivity of amorphous materials above the plateau. Phys Rev B 39:13465–13477
21. Kistler SS (1932) Coherent Expanded–Aerogels. J Phys Chem 36:52–64
22. Mott NF (1967) Electrons in disordered structures. Adv Phys 16:49–144
23. Mott NF (1969) Conduction in non-crystalline materials III. Localized states in a pseudogap and near extremities of conduction and valence bands. Philos Mag 19:835–852
24. Nakayama T (1992) Dynamics of random fractals: large-scale simulations. Physica A 191:386–393
25. Nakayama T (1995) Elastic vibrations of fractal networks. Japan J Appl Phys 34:2519–2524
26. Nakayama T, Orbach R (1999) Anharmonicity and thermal transport in network glasses. Europhys Lett 47:468–473
27. Nakayama T, Orbach RL (1999) On the increase of thermal conductivity in glasses above the plateau region. Physica B 263–264:261–263
28. Nakayama T, Yakubo K (2001) The forced oscillator method: eigenvalue analysis and computing linear response functions. Phys Rep 349:239–299
29. Nakayama T, Yakubo K (2003) Fractal Concepts in Condensed Matter Physics. Solid-State Sciences. Springer, Berlin
30. Nakayama T, Yakubo K, Orbach R (1994) Dynamical properties of fractal networks: Scaling, numerical simulations, and physical realizations. Rev Mod Phys 66:381–443
31. Orbach R (1986) Dynamics of Fractal Networks. Science 231:814–819
32. Orbach R (1996) Transport and vibrational lifetimes in amorphous structures. Physica B 219–220:231–234
33. Orbach R, Jagannathan A (1994) High energy vibrational lifetimes in a-Si:H. J Phys Chem 98:7411–7413
34. Rammal R (1984) Spectrum of harmonic excitations on fractals. J Phys (Paris) 45:191–206
35. Rammal R, Toulouse G (1983) Random walks on fractal structures and percolation clusters. J Phys Lett (Paris) 44:L13–L22
36. Scholten AJ, Dijkhuis JI (1996) Decay of high-frequency phonons in amorphous silicon. Phys Rev B 53:3837–3840
37. Scholten AJ, Verleg PAWE, Dijkjuis JI, Akimov AV, Meltzer RS, Orbach R (1995) The lifetimes of high-frequency phonons in amorphous silicon: evidence for phonon localization. Solid State Phenom 44–46:289–296
38. Stanley HE (1977) Cluster shapes at the percolation threshold: and effective cluster dimensionality and its connection with critical-point exponents. J Phys A 10:L211–L220
39. Stauffer D, Aharony A (1992) Introduction to Percolation Theory, 2nd edn. Taylor and Francis, London/Philadelphia
40. Stauffer D, Stanley HE (1995) From Newton to Mandelbrot: A primer in Theoretical Physics, 2nd edn. Springer, Berlin
41. Takayasu H (1990) Fractals in the Physical Sciences. Manchester University Press, Manchester
42. Tsujimi Y, Courtens E, Pelous J, Vacher R (1988) Raman-scattering measurements of acoustic superlocalization in silica aerogels. Phys Rev Lett 60:2757–2760
43. Uemura YJ, Birgeneau RJ (1987) Magnons and fractons in the diluted antiferromagnet $Mn_xZn_{1-x}F_2$. Phys Rev B 36:7024–7035
44. Vacher R, Courtens E, Coddens G, Pelous J, Woignier T (1989) Neutron-spectroscopy measurement of a fracton density of states. Phys Rev B 39:7384–7387
45. Vacher R, Woignier T, Pelous J, Courtens E (1988) Structure and self-similarity of silica aerogels. Phys Rev B 37:6500–6503
46. Vicsek T (1993) Fractal Growth Phenomena. World Scientific, Singapore
47. Yakubo K, Nakayama T (1989) Direct observation of localized fractons excited on percolating nets. J Phys Soc Japan 58:1504–1507
48. Yamaguchi M, Nakayama T, Yagi T (1999) Effects of high pressure on the Bose peak in a-GeS_2 studied by light scattering. Physica B 263–264:258–260
49. Yamaguchi M, Yagi T (1999) Anharmonicity of low-frequency vibrations in a-GeS_2 studied by light scattering. Europhys Lett 47:462–467

Recommended Reading

While many of the original articles are somewhat difficult to access, the reader is nevertheless recommended to read them as there are subtle issues that tend to disappear as references are quoted and re-quoted over time. For a very helpful introduction to percolation theory, Ref. 39 is highly recommended. A thorough review of dynamics of/on fractal structures can be found in Ref. 30. Finally, a comprehensive treatment of localization and multifractals, and their relevance to dynamics of/on fractal networks, can be found in Ref. 29.

Copyrighted material

Dynamics of Parametric Excitation

ALAN CHAMPNEYS
Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

Article Outline

Glossary
Definition of the Subject
Introduction
Linear Resonance or Nonlinear Instability?
Multibody Systems
Continuous Systems
Future Directions
Bibliography

Glossary

Parametric excitation Explicit time-dependent variation of a parameter of a dynamical system.

Parametric resonance An instability that is caused by a rational relationship between the frequency of parametric excitation and the natural frequency of free oscillation in the absence of the excitation. If ω is the excitation frequency, and ω_0 the natural frequency, then parametric resonance occurs when $\omega = (n/2)\omega_0$ for any positive integer n. The case $n = 1$ is usually the most prominent form of parametric resonance, and is sometimes called the principle subharmonic resonance.

Autoparametric resonance A virtual parametric resonance that occurs due to the coupling between two independent degrees of freedom within a system. The output of one degree of freedom acts like the parametric excitation of the other.

Ince–Strutt diagram A two-parameter bifurcation diagram indicating stable and unstable regions, specifically plotting the instability boundaries as the required amplitude of parametric excitation against the square of the ratio of natural to excitation frequency.

Floquet theory The determination of the eigenvalue spectrum that governs the stability of periodic solutions to systems of ordinary differential equations.

Bifurcation A qualitative change in a system's dynamics as a parameter is varied. One-parameter bifurcation diagrams often depict invariant sets of the dynamics against a single parameter, indicating stability and any bifurcation points. Two-parameter bifurcation diagrams depict curves in a parameter plane on which one-parameter bifurcations occur.

Modal analysis The study of continuum models by decomposing their spatial parts into eigenmodes of the dynamic operator. The projection of the full system onto each mode gives an infinite system of differential equations in time, one for each mode.

Monodromy matrix The matrix used in Floquet theory, whose eigenvalues (also known as Floquet multipliers) determine stability of a periodic solution.

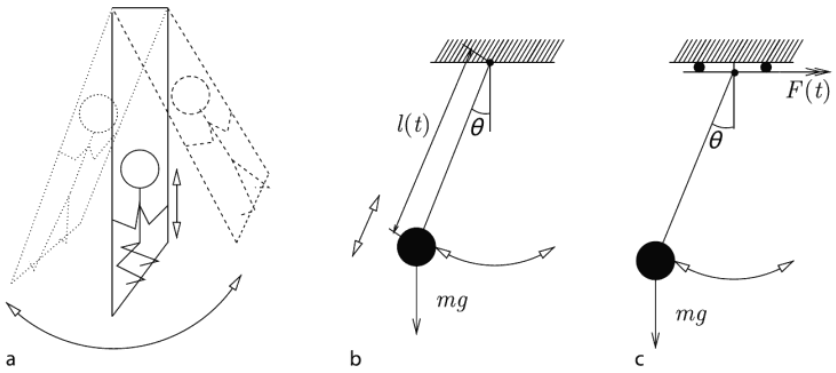
Definition of the Subject

Parametric excitation of a system differs from direct forcing in that fluctuations appear as temporal modulation of a parameter rather than as a direct additive term. A common paradigm is that of a pendulum hanging under gravity whose support is subjected to a vertical sinusoidal displacement. In the absence of any dissipative effects, instabilities occur to the trivial equilibrium whenever the natural frequency is a multiple of *half* the excitation frequency. At amplitude levels beyond the instability, further *bifurcations* (dynamical transitions) can lead to more complex quasi-periodic or chaotic dynamics. In multibody mechanical systems, one mode of vibration can effectively act as the parametric excitation of another mode through the presence of multiplicative nonlinearity. Thus *autoparametric* resonance occurs when one mode's frequency is a multiple of half of the other. Other effects include combination resonance, where the excitation is at a sum or difference of two modal frequencies. Parametric excitation is important in continuous systems too and can lead to counterintuitive effects such as stabilization "upside-down" of slender structures, complex wave patterns on a fluid free surface, stable pulses of light that overcome optical loss, and fluid-induced motion from parallel rather than transverse flow.

Introduction

Have you observed a child on a playground swing? In either a standing or a sitting position, generations of children have learned how to "pump" with their legs to set the swing in motion without making contact with the ground, and without external forcing. Note that pumping occurs at *twice* the natural frequency of swing of the pendulum, since the child extends her legs maximally at the two extremities of the swing's cycle, that is, two times per period. What is happening here? The simplest case to analyze is where the child is standing. Then, it has been argued by Curry [17] that the child plus swing is effectively a simple pendulum system, where the child's pumping has the effect of periodic variation of the position of the center of the mass along the arm of the pendulum (see Fig. 1a,b).

Copyrighted material



Dynamics of Parametric Excitation, Figure 1
a A child standing on a swing modeled as b a parametrically excited pendulum c a directly forced pendulum

We shall return to the self-propelled swing problem in Sect. "Future Directions" below, where we shall find out that all is not what it seems with this motivating example, but for the time being lets make the simple hypothesis that the child's pumping causes the center of mass to move up and down.

Using the standard model of a pendulum that consists of a lumped mass suspended by a rigid massless rod of length l from a pivot, the equations of motion for the angular displacement $\theta(t)$ of a pendulum hanging under gravity can be written in the form

$$\ddot{\theta}(t) + \frac{g}{l}\sin[\theta(t)] = 0\,.$$

Here g is the acceleration due to gravity, a dot denotes differentiation with respect to time t and, for the time being, we have ignored any damping that must inevitably be present. Now, the moving center of mass means that the effective length of the rigid rod oscillates, and let us suppose for simplicity that this oscillation is sinusoidal about some mean length l_0:

$$l(t) = l_0[1 - \varepsilon\cos(\omega t)]\,.$$

Here $\varepsilon > 0$ assumed to be small and as yet we haven't specified the frequency ω. Then, upon Taylor expansion in ε, we find to leading order

$$\ddot{\theta} + \omega_0^2[1 + \varepsilon\cos(\omega t)]\sin\theta = 0\,,$$

where $\omega_0 = \sqrt{(g/l_0)}$ is the natural frequency of small amplitude swings in the absence of the pump. Finally, upon rescaling time $\tilde{t} = \omega t$ and dropping the tildes we find

$$\ddot{\theta} + [\alpha + \beta\cos t]\sin\theta = 0\,, \tag{1}$$

where

$$\alpha = \frac{\omega_0^2}{\omega^2} \quad \text{and} \quad \beta = \varepsilon\frac{\omega_0^2}{\omega^2} \tag{2}$$

are dimensionless parameters that respectively describe the square of the ratio of natural frequency to excitation frequency, and the amplitude of the excitation.

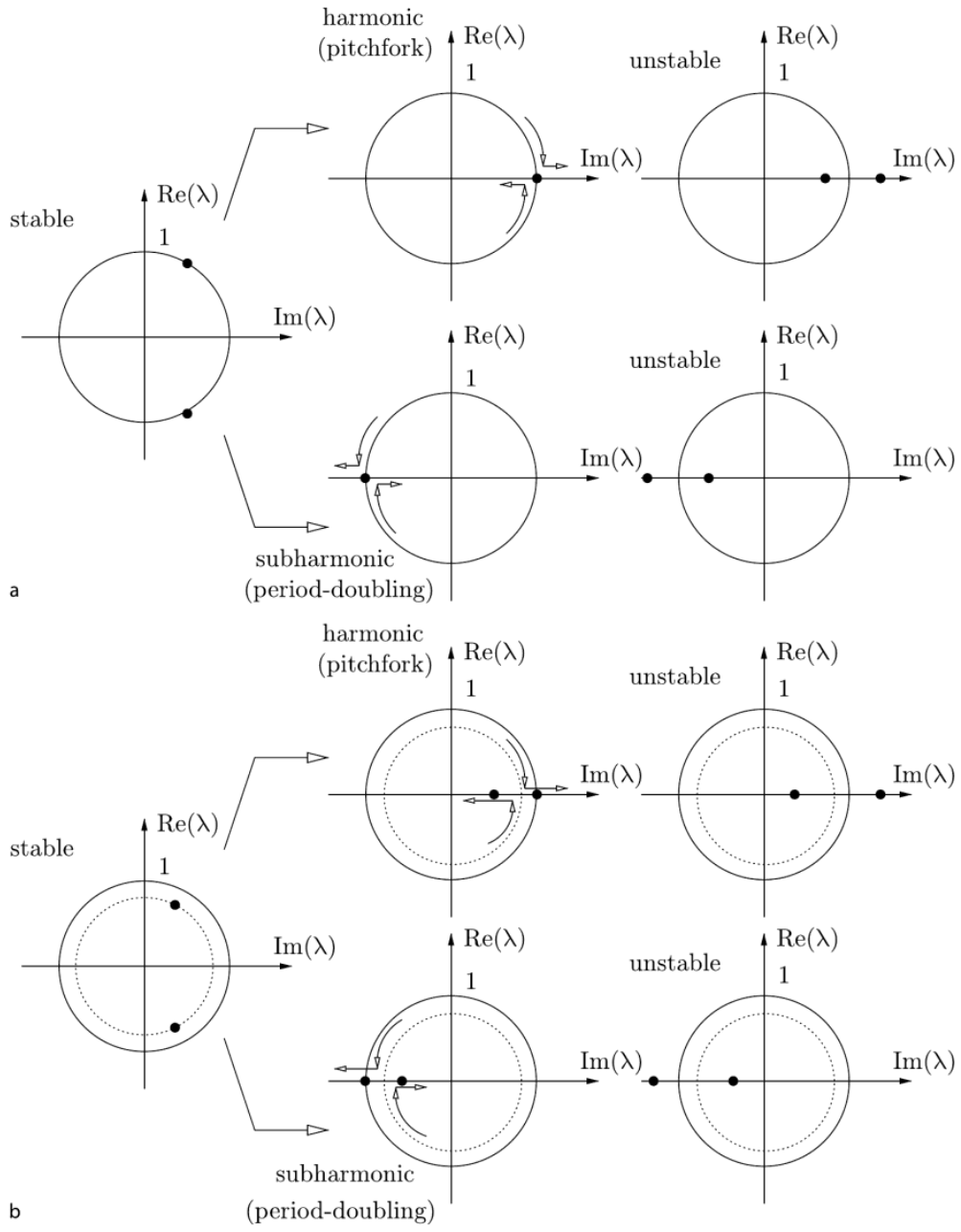
Equation (1) is a canonical example of a *parametrically excited* system. Note that (1) also describes, at least to leading order, the case of a pendulum hanging under gravity that is excited by a vertical force of size $\varepsilon\omega_0^2\cos t$ which we can think of as a periodic modulation of the gravity. This would be in contrast to a *directly forced* simple pendulum where the forcing would act in the direction of swing for small amplitude motion; see Fig. 1c. For the forced pendulum, the equivalent equations of motion would look more like

$$\ddot{\theta} + \alpha\sin\theta = \beta\cos t\,, \tag{3}$$

where the periodic forcing occurs as an additive input into the differential equation. In contrast, for (1) the periodic excitation occurs as a modulation of a *parameter* (which you can think of as gravity), hence the name 'parametric'. Another feature of (1) that makes it a good paradigm for this article is that it is nonlinear and, as we shall see, a full appreciation of the dynamics resulting from parametric excitation requires treatment of nonlinear terms.

An interesting contrast between parametrically excited and directly forced systems is in the nature of the trivial response to a non-zero input. In particular, for any α and β, Eq. (1) admits the trivial solution $\theta(t) \equiv 0$, whereas there is no such solution to (3) for non-zero β. The simplest "steady" solution to the directly forced system for small β is a small-amplitude periodic oscillation of period 2π (in our dimensionless time units). So any analysis of a parametrically excited system should start with an analysis of the steady state $\theta = 0$, which corresponds to the pendulum just hanging under gravity. To analyze stability of this equilibrium state let $\theta = 0 + x$ where x is small, then x sat-

Copyrighted material



Dynamics of Parametric Excitation, Figure 2
Possible behavior of Floquet multipliers λ (eigenvalues of the monodromy matrix M) for a a conservative Hill equation and b in the presence of small damping

isfies linearization (1)

$$\ddot{x} + (\alpha + \beta \cos t)x = 0\,. \tag{4}$$

The equilibrium position is stable if all solutions $x(t)$ of this equation, remain bounded as $t \to \infty$ and it is unstable if there are solutions $x(t) \to \infty$ as $t \to \infty$.

Equation (4) is known as the *Mathieu equation* [42], and is a specific example of Hill's equation [29,40]

$$\ddot{x} + [\alpha + V(t)]x = 0\,, \tag{5}$$

where $V(t)$ is any periodic function with period 2π. Hill's equation can be solved using classical methods for linear

Copyrighted material

time-dependent ordinary differential equations (ODEs), see e. g. [31,33]. In particular, by writing $y_1 = x$ and $y_2 = \dot{x}$, Hill's equation (5) is just a two-dimensional example of a linear first-order time-periodic equation

$$\dot{\mathbf{y}} = P(t)\mathbf{y}\,, \qquad \mathbf{y} \in \mathbb{R}^n\,. \tag{6}$$

The general solutions to (6) for any initial condition can be expressed in terms of a fundamental matrix $\Phi(t)$, with $\Phi(0) = I_n$ (the identity matrix of dimension n):

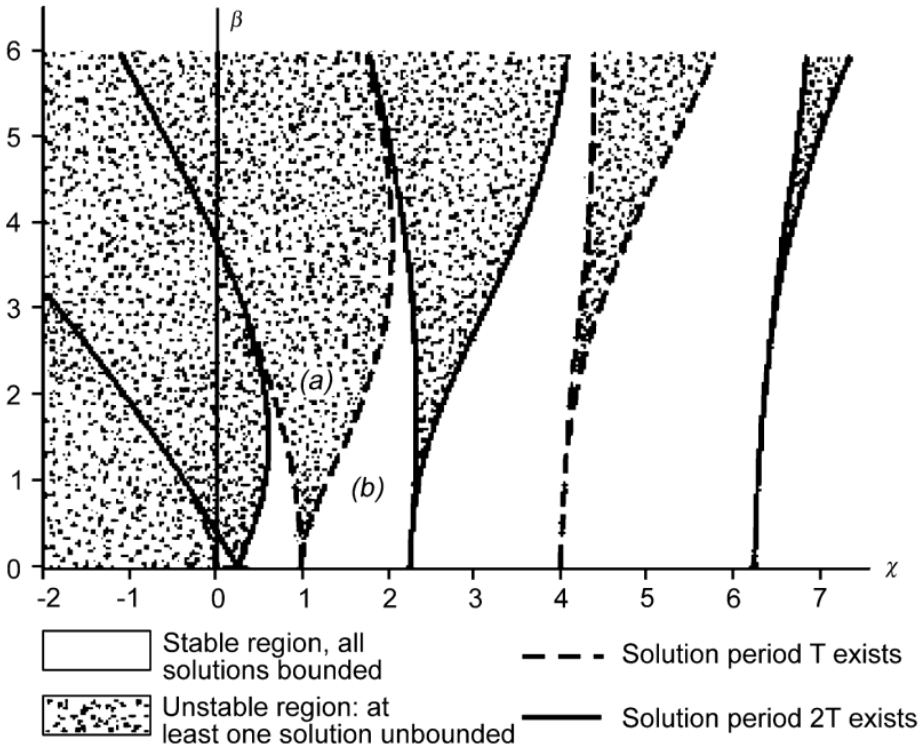
$$\mathbf{y}(t) = \Phi(t)\mathbf{y}(0)\,.$$

Note that just because $P(t)$ has period 2π it does not follow that $\Phi(t)$ does. Also $\Phi(t)$ cannot be computed exactly, except in highly specialized cases, and one has to rely on approximate methods as described in the next Sect. "Floquet Analysis" and in the companion article to this one ▶ Perturbation Analysis of Parametric Resonance. However, in order to study stability we don't actually need to construct the full solution $\Phi(t)$, it is sufficient to consider $M = \Phi(2\pi)$, which is called the *monodromy matrix* associated with Hill's equation. Stability is determined by studying the eigenvalues of M: solutions to (5) with small initial conditions remain small for all time if all eigenvalues of M lie on or within the unit circle in the complex plane; whereas a general initial condition will lead to unbounded motion if there are eigenvalues outside this circle. See Fig. 2a. Eigenvalues of the Monodromy matrix are also known as *Floquet multipliers* of the system, and the general study of the stability of periodic systems is called *Floquet theory*.

In the case of the Mathieu equation, where $V(t) = \beta\cos(t)$, we have

$$P(t) = \begin{bmatrix} 0 & 1 \\ 0 & \alpha + \beta\cos(t) \end{bmatrix}\,. \tag{7}$$

Figure 3 depicts graphically the behavior of Floquet multipliers for this case. Here, in what is known as an *Ince–Strutt diagram* [31,63], the shaded regions represent the values of parameters at which the origin $x = 0$ is stable (with Floquet multipliers on the unit circle) and the white regions are where the origin is unstable (at least one multiplier outside the unit circle). We describe how to calculate such a diagram in the next section. For the time be-



Dynamics of Parametric Excitation, Figure 3

Ince–Strutt diagram showing resonance tongues of the Mathieu equation (4). Regions of instability of the trivial solution $x = 0$ are *shaded* and a distinction is drawn between instability curves corresponding to a period-T ($T = 2\pi$) and to a period-$2T$ orbit. After (2nd edn. in [33]), reproduced with permission from Oxford University Press

Copyrighted material

where an asterisk represents complex conjugation. Note that the final condition implies that x is real. Clearly, for $\beta = 0$, a nontrivial solution can be found for any integer p which has $c_n = 0$ for all $n \neq p$, provided $\alpha^2 = p^2$. That is, solution curves bifurcate from $\beta = 0$ for each α-value that is the square of a whole integer. Furthermore, nonlinear analysis can be used to show that for small positive β, there are *two* solution branches that emerge from each such point (provided $p \neq 0$); one corresponding to c_p being real, which corresponds to motion $x(t) \sim \cos(pt)$ in phase with the excitation; and another corresponding to c_p imaginary, which corresponds to motion $x(t) \sim \sin(pt)$ out of phase with the excitation. Similarly, if we look for solutions which are 4π-periodic by making the substitution

$$x(t) = \sum_{m=-\infty}^{m=+\infty} d_m e^{imt/2}\,, \tag{11}$$

we arrive at an infinite system of equations

$$(1/2)\beta d_{m+2} + (\alpha - (1/4)m^2)d_m + (1/2)\beta d_{m-1} = 0\,,$$
$$m = -\infty, \ldots, \infty\,, \quad \text{with } d_{-m} = d_m^*\,. \tag{12}$$

Note that the equations for m even and m odd in (12) are decoupled. The equations for even m become, on setting $m = 2n$, precisely the system (10) with $d_{2m} = c_n$. For $m = 2n + 1$, however, we arrive at new points of bifurcation, by exactly the same logic as we used to analyze (10). Namely whenever $\alpha^2 = (p/2)^2$ for any odd integer p we have the bifurcation from $\beta = 0$ of two 4π-periodic solutions (one like $\cos(pt/2)$ and one like $\sin(pt/2)$).

Perturbation analysis can be used to obtain asymptotic expressions for the transition curves that bifurcate from $\beta = 0$ at these special α-values; see the companion article ▶ Perturbation Analysis of Parametric Resonance for the details. In particular it can be shown that the width of the stability tongue (separation in α-values from the sine and the cosine instability curves) originating at $\alpha = (p/2)^2$ (for p being odd or even) scales like

$$\text{tongue width}_p \sim \beta^p + \text{h.o.t}\,.$$

Hence the principal subharmonic instability arising from $\alpha = 1/4$ has width $O(\beta)$ (actually the transition curves are given by $\alpha = 1/4 \pm (1/2)\beta + O(\beta^2)$), and the fundamental harmonic instability arising from $\alpha = 1$ has width $O(\beta^2)$, etc. Hence the principal subharmonic instability is by far the most prevalent as it occupies the greatest frequency width for small β, and is often thought to be the hallmark of parametric resonance in applications.

Note that we have not as yet mentioned the transition curve emerging from $\alpha = \beta = 0$, which corresponds to $n = 0$ in the above analysis. This case can also be analyzed asymptotically, and the single transition curve can be shown to scale like $\alpha = -(1/2)\beta^2 + O(\beta^3)$ for small β. It is this bending back into $\alpha < 0$ that leads to the Kapitza pendulum effect, which we shall return to in Sect. "Upside-Down Stability" below.

Finally, the addition of small damping can be shown to narrow the tongues and lift them off from the $\beta = 0$ axis as shown in Fig. 4. The amount of lift is in proportion to the width of the tongue. This property further enhances the pre-eminence of the principal subharmonic instability for parametrically excited systems in practice, as they thus require the least amplitude of vibration in order to overcome the damping forces.

Bifurcation Theory

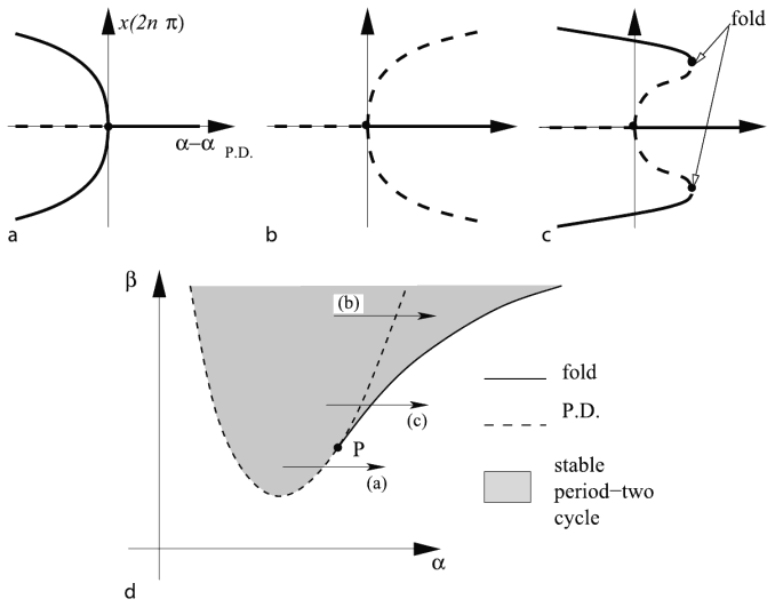
We have just analyzed the Mathieu equation, which is purely linear. But, for example as shown in the Introduction, such an equation is often derived in the course of analyzing the stability of a trivial solution to a nonlinear parametrically excited system. It is worthwhile therefore to consider the nonlinear implications of the harmonic and subharmonic instabilities. The time-periodic Mathieu equation can be written as an example of a three-dimensional autonomous system of ODEs

$$\dot{x} = y\,, \quad \dot{y} = f(x, y, s)\,, \quad \dot{s} = 1 \pmod{2\pi}\,,$$

for $(x, y, s) \in \mathbb{R}^2 \times S^1$, where $f(x, y, s) = -\delta y - (\alpha + \beta\cos(s))x$. As such, the trivial solution $x = 0$ should actually be seen as the periodic solution $x = y = 0$, $s = t \pmod{2\pi}$. Hence we can understand instability of the trivial solution as points of bifurcation of periodic solutions, for which there is a mature theory, see for example [27,37,65,67]. In particular, the subharmonic instability with Floquet multiplier -1 is a *period doubling bifurcation*, and the harmonic bifurcation corresponding to multiplier $+1$ is a *pitchfork* bifurcation. Note that both these bifurcations come in super- and sub-critical forms. Hence, depending on the nature of the nonlinearity we may find either stable bounded motion close to the instability inside the resonant tongue close to the boundary edge, or motion that is not of small amplitude. See Fig. 5.

The case of the Mathieu or Hill equations without damping have received special attention in the literature. This is because of the extra mathematical structure they posses, namely that they can be expressed as *reversible* and *Hamiltonian* systems. Specifically, Hill's equation of the form (5) can be shown to conserve the energy function

Copyrighted material



Dynamics of Parametric Excitation, Figure 5

Contrasting a super-critical and b sub-critical bifurcations close to the resonance tongue boundary of a subharmonic (period-doubling) instability, here represented by the abbreviation P.D. *Solid lines* represent stable solutions, *dashed lines* represent unstable solutions. c A one-parameter bifurcation diagram close to the super/sub-critical boundary. d A two parameter unfolding of the codimension-two point *P* where the super/sub-critical transition occurs. *Shading* is used to represent where a stable period-two cycle exists. Note in this case that for higher β-values the right-hand boundary of this existence region is therefore the fold curve and not the period-doubling point. Between the period-doubling and the fold there is hysteresis (coexistence) between the stable trivial solution and the stable period-two cycle

$$H(x, y, t) = (1/2)y^2 + (1/2)(\alpha + V(t))x^2 \,,$$

$y = \dot{x}$, and to be expressible in Hamiltonian form

$$\dot{x} = \frac{\partial H}{\partial y} \,, \qquad \dot{y} = -\frac{\partial H}{\partial x} \,.$$

Reversibility of Hill's equation is expressed in the fact that the model is invariant under the transformation $t \to -t$, $y \to -y$. Broer and Levi [7] used Hamiltonian and reversible theory to look at Hill's equations for general periodic functions $V(t)$. They showed that under certain conditions, the resonance tongues may become of zero width for nonzero β. Effectively, the left and right-hand boundaries of the tongue pass through each other. This is a generic phenomenon, in that a small perturbation within the class of Hill's equations will lead not destroy such an intersection. However, the Mathieu equation for which $V(t) = \cos(t)$ does not posses this property.

Many nonlinear models that undergo parametric resonance also preserve the properties of being Hamiltonian and reversible. Special tools from reversible or Hamiltonian systems theory can sometimes be used to analyze the bifurcations that occur under parametric resonance in nonlinear systems, see for example [8,9]. Points outside the resonance tongues correspond to where the trivial solution is an *elliptic* (weakly stable) fixed point of the time-2π map, and those inside the resonance tongue to where it is a *hyperbolic* (unstable) fixed point. One therefore does not generally see asymptotically stable non-trivial solutions being born at a resonance tongue boundary, rather one sees elliptic nontrivial dynamics. In fact, for a general periodic solution in a Hamiltonian system, pitchfork and period-doubling bifurcations are not the only kind of bifurcations that produce non-trivial elliptic periodic orbits. If the Floquet multipliers pass through any nth root of unity, there is a *period multiplying* (subharmonic) bifurcation at which a periodic orbit of period $2n\pi$ is born (see, e. g. [62]). A period-doubling is just the case $n = 2$. Such bifurcations can also occur for the non-trivial elliptic periodic orbits within a resonance tongue, but they are all destroyed by the presence of even a small amount of damping.

Beyond the Initial Instability

For a nonlinear system, the pitchfork or period-doubling bifurcation that occurs upon crossing a resonance tongue might just be the first in a sequence of bifurcations that is

Copyrighted material

encountered as a parameter is varied. As a general rule, the more β is increased for fixed α, the more irregular the stable dynamics becomes. For example, the period-doubling bifurcation that is encountered upon increasing β for α close to 1/4 is typically just the first in a sequence of period-doublings, that accumulate in the creation of a chaotic attractor, obeying the famous Feigenbaum scaling of this route to chaos (see e. g. [18]).

Van Noort [64] produced a comprehensive numerical and analytical study into the nonlinear dynamics of the parametrically excited pendulum (1) without damping. Of interest in that study was to complete the nonlinear dynamics of one of the simplest non-integrable Hamiltonian and reversible dynamical systems. The parameter β acts like a perturbation to integrability. Certain structures that can be argued to exist in the integrable system, like closed curves of quasiperiodic motion in a Poincaré section around an elliptic fixed point, serve as approximate organizers of the dynamics for non-zero β. However, KAM theory (see e. g. [55]) shows that these closed curves become invariant tori that typically breakup as β is increased. At the heart of the regions of chaotic dynamics lie the resonance tongues associated with a period-multiplying bifurcation. Inside the resonance tongues are nontrivial fixed points, whose surrounding invariant tori break up into heteroclinic chains, which necessarily imply chaos.

Acheson [2] studied the dynamics of the parametrically forced pendulum in the presence of damping (i. e. Eq. (1) with an additional term $\Delta\dot{x}$ on the right-hand side) via computer simulation. He found parameter regions inside the principal resonance tongue of the pendulum where it undergoes what he referred to as *multiple nodding* oscillations. The subharmonic motion inside the tongue should be such that during one cycle of the excitation, the pendulum swings to the left and back, and during the next cycle it swings to the right and back. That is, one complete oscillation per two cycles. Multiple nodding motion by contrast is asymmetric and involves two swings to the left, say, (one swing per cycle of the excitation), followed by one swing to the right, with the motion repeating not every two but every three cycles. He also found similar four-cycle and five-cycle asymmetric motion. This behavior can be explained as pockets of stable behavior that survives from the period-multiplying bifurcations that are only present in the case of zero damping [21].

Clearly, beyond the initial instability the universality of the Ince–Strutt diagram for the Mathieu equation, or for any other Hill equation for that matter, disappears and the precise features of the nonlinear dynamics depends crucially on the details of the particular nonlinear system being investigated.

Multibody Systems

Of course, there is little in the real world that can be accurately described by deterministic, parametrically excited single degree-of-freedom ODEs. In this section we shall consider slightly more realistic situations that are governed by higher-dimensional systems, albeit still deterministic, and still described by finitely many dynamic variables. For simplicity, we shall restrict attention to rigid-body mechanisms that can be described by a system of (finitely many) coupled modes of vibration. Furthermore, we shall adopt a Lagrangian framework where the dynamics of mode i can be represented by a generalized coordinate q_i and its associated generalized velocity $\dot{q}_i$, such that we arrive at a system of (generally nonlinear) coupled second-order ODEs; one equation for each acceleration $\ddot{q}_i$.

Autoparametric Resonance

Autoparametric resonance occurs when there is asymmetric nonlinear multiplicative coupling between modes of a multibody system. Specifically, suppose that a mode $(q_1, \dot{q}_1)$ is coupled to a mode $(q_2, \dot{q}_2)$ through a term proportional to $q_1 q_2$ in the equation of motion for the 2nd mode (i. e. the $\ddot{q}_2$ equation), yet there is no corresponding $q_2 q_1$ term in the $\ddot{q}_1$ equation. For example, consider

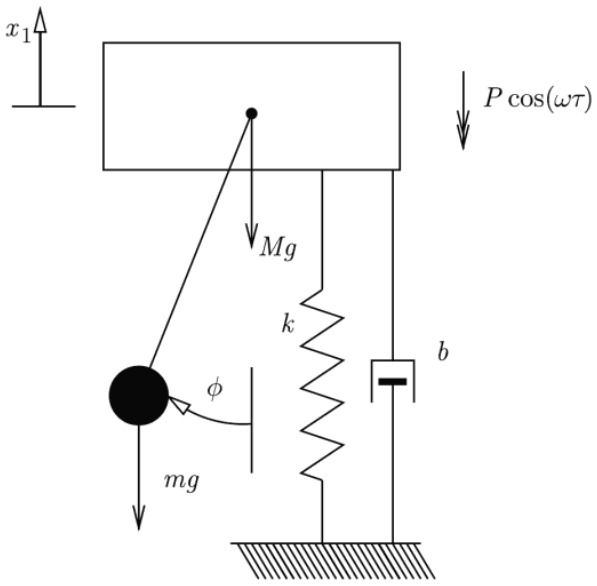
$$\ddot{q}_1 + \delta_1 \dot{q}_1 + \alpha_1 q_1 = f_1(q_1, t) \tag{13}$$

$$\ddot{q}_2 + \delta_2 \dot{q}_2 + \alpha_2 q_2 + \beta q_1 q_2 = f_2(q_2)\,, \tag{14}$$

where f_1 and f_2 are nonlinear functions of their arguments and the dependence of f_1 on t is periodic. Furthermore suppose that mode 1 undergoes a periodic motion (either as a result of direct forcing, its own parametric resonance, or self-excitation). For example, if $f_2 = 0$ and f_1 were simply $\cos(\omega t)$ for some ω^2 sufficiently detuned from α_1, then the solution $q_1(t)$ to (13) would, after allowing sufficient time for transients to decay, be simply proportional to $\cos(\omega t + \phi)$ for some phase ϕ. Then, the $q_1 q_2$ term in (14) would be effectively act like $\beta \cos(\omega t + \phi) q_2$ and so (14) would be precisely the Mathieu equation. Values of $\omega^2 = 4\alpha_2/n$ for any positive integer n correspond to the root points of the resonance tongues of the Ince–Strutt diagram. Such points are referred to as *internal* or *autoparametric* resonances.

A canonical example of an autoparametrically resonant system is the simple spring-mass-pendulum device depicted in Fig. 6, which we shall use to discuss the some key generic features of autoparametric resonance; see the book by Tondl et al. [58] for more details. In the figure, a mass M is mounted on a linear spring with stiffness k,

Copyrighted material



Dynamics of Parametric Excitation, Figure 6
A mass-spring-pendulum system

is forced by an applied displacement $P\cos(\omega\tau)$ and is attached to the pivot of a pendulum of length l and lumped mass m that hangs under gravity. The motion of the mass and the pendulum are assumed to be subject to proportional viscous damping with coefficients b and c respectively. Using standard techniques, the equations of motion can of such a device can be written in the dimensionless form

$$\ddot{x} + \delta_1\dot{x} + \alpha_1 + \mu(\ddot{\phi}\sin(\phi) + \dot{\phi}^2\cos\phi) = A\cos t \quad (15)$$

$$\ddot{\phi} + \delta_2\dot{\phi} + \alpha_2\sin\phi + \ddot{x}\sin\phi = 0\,, \quad (16)$$

where $x = y/l$, time t has been scaled so that $t = \omega\tau$, and the dimensionless parameters are

$$\delta_1 = \frac{b}{\omega(M+m)}\,,\quad \alpha_1 = \frac{k}{\omega^2(M+m)}\,,\quad \mu = \frac{m}{M+m}\,,$$
$$A = \frac{P}{l\omega^2(M+m)}\,,\quad \delta_2 = \frac{c}{\omega m l^2}\,,\quad \alpha_2 = \frac{g}{l\omega^2}\,.$$

The simplest form of output of such a model is the so-called *semi-trivial* solution in which the pendulum is at rest, $\dot{\phi} = \phi \equiv 0$, and the mass-spring system alone satisfies the simple equation

$$\ddot{x} + \delta_1\dot{x} + \kappa_1 x = A\cos t\,.$$

After transients have died out, this single equation has the attracting solution

$$x(t) = R_0\cos(t+\psi_0)\,,$$
$$\text{where}\quad R_0 = \frac{A}{\sqrt{\delta_1^2 + (\alpha_1-1)^2}}$$
$$\text{and}\quad \tan\psi_0 = \frac{\delta_1}{\alpha_1 - 1}\,. \quad (17)$$

Note that the solution (17) exists for all values of the physical parameters, in particular for all positive α_1 provided $\delta_1 > 0$.

Substitution of the semi-trivial solution (17) into the pendulum Eq. (16), leads to

$$\ddot{\phi} + \delta_2\dot{\phi} + [\alpha_2 + R_0\cos(t+\psi_0)]\sin\phi = 0\,,$$

which, after a time shift $t \to t - \psi_0$ is precisely the Eq. (1) for a parametrically excited pendulum in the presence of damping. Thus, the Ince–Strutt diagram of the damped Mathieu equation, with $\beta = A/(\sqrt{\delta_1^2 + (\alpha_1-1)^2})$, $\delta = \delta_2$ and $\alpha = \alpha_2$ tell us precisely the parameter values for which the semi-trivial solution is stable. Inside the resonance tongues we have an instability of the semi-trivial solution, which leads to stable non-trivial motion with the pendulum swinging. To determine precisely what happens away from the boundaries of the stability tongue, one again has to perform fully nonlinear analysis, as outlined in Sect. "Linear Resonance or Nonlinear Instability?". Note now though that for nonzero pendulum motion, $\phi \not\equiv 0$, the Eqs. (15) and (16) become a fully coupled two-degree-of-freedom system, whose dynamics can be even more complex than that of the one-degree-of-freedom parametrically excited pendulum. Indeed in [58], a further destabilizing secondary-Hopf (Neimark–Sacker) bifurcation is identified within the principal ($\alpha_2 = 1/4$) subharmonic resonance tongue, and numerical evidence is found for several different kinds of quasiperiodic motion.

There is one difference though between autoparametric resonance and pure parametric resonance. The paradigm for parametric resonance is that the external parametric excitation is held constant (its amplitude and frequency are *parameters* of the system). Here, however, for nonzero ϕ, the system is fully coupled, so that the parametric term (17) can no longer be thought of as being independent of ϕ, effectively the excitation becomes a *dynamic variable* itself. Thus, motion of the pendulum can affect the motion of the mass-spring component of the system. In particular if there is an autoparametric instability where the pendulum is set in motion, then energy must be transferred from the mass-spring component to the pendulum, a phenomenon known as *quenching*. But the pendulum is only set in motion close to certain resonant frequency ratios α_2. Thus, we have a *tuned damper* that is designed to

Copyrighted material

take energy out of the motion of the mass at certain input frequencies ω frequencies.

To see how such a damper might work in practice, suppose we build a civil engineering structure such as a bridge or a building, that has a natural frequency ω_1 in some mode, whose amplitude we call $x(t)$. Suppose the system is subjected to an external dynamic loading (perhaps from traffic, wind or earthquakes) that is rich in a frequency ω which is close to ω_1. Then, if the damping is small (as is typical in large structures) using (15) with $\mu = 0$, we find that the response given by (17) is large since α_1 is close to 1 and δ_1 is small. This is the fundamental *resonant response* of the structure close to its natural frequency. Now suppose we add the pendulum as a tuned damper. In particular we design the pendulum so that $\alpha_2 \approx \alpha_1$. The pendulum now sees a large amplitude parametric excitation within its main subharmonic resonance tongue. It becomes violently unstable, thus sucking energy from the structure.

To see just how effective this quenching effect can be, consider the following asymptotic calculation [58]. Suppose the external forcing is very small, $A = \varepsilon^2 \hat{A}$, for some small parameter ε, and that the damping and frequency detunings scale with ε: $\delta_1 = \varepsilon\hat{\delta}_1$, $\delta_2 = \varepsilon\hat{\delta}_2$, $\alpha_1 = 1 + \varepsilon\hat{\sigma}_1$, $\alpha_2 = (1/4) + \varepsilon\hat{\sigma}_2$, where the hatted quantities are all assumed to be $O(1)$ as $\varepsilon \to 0$. Upon making these substitutions, and dropping all hats, Eqs. (15) and (16) become

$$\ddot{x} + x + \varepsilon[\delta_1\dot{x} + \sigma_1 - \frac{1}{4}\mu\phi^2 + \mu\dot{\phi}^2] = \varepsilon A\cos(t) + O(\varepsilon^2)\,, \quad (18)$$

$$\ddot{\phi} + \frac{1}{4}\phi + \varepsilon[\delta_2\dot{\phi} + \frac{1}{2}\sigma_2\phi - x\phi] = O(\varepsilon^2)\,. \quad (19)$$

We then look for solutions of the form

$$x = R_1(t)\cos\left[t + \psi_1(t)\right],$$
$$\phi = R_2(t)\cos\left[\frac{t}{2} + \psi_2(t)\right].$$

Making this substitution into (18), (19), and keeping only $O(\varepsilon)$ terms we find

$$2\dot{R}_1 = \varepsilon\left[-\delta_1 R_1 - \frac{1}{4}R_2^2\sin(\psi_1 - 2\psi_2) - A\sin(\psi_1)\right],$$
$$2R_1\dot{\psi}_1 = \varepsilon\left[\sigma_1 R_1 - \frac{1}{4}\mu R_2^2\cos(\psi_1 - 2\psi_2) - A\cos\psi_1\right],$$
$$2\dot{R}_2 = \varepsilon\left[-\delta_2 R_2 + R_1 R_2\sin(\psi_1 - 2\psi_2)\right],$$
$$2R_2\dot{\psi}_2 = \varepsilon\left[\sigma_2 R_2 - R_1 R_2\cos(\psi_1 - 2\psi_2)\right].$$

Equilibrium solutions of these equations correspond to 4π-periodic solutions that are valid close to the instability tongue. Upon setting the left-hand sides to zero, a straightforward combination of the final two equations yields

$$R_1 = \sqrt{\sigma_2^2 + \delta_2^2}\,, \quad (20)$$

and a slightly longer calculation reveals that this leads to a unique (stable) solution for R_2 provided $A^2 > (\delta_1^2 + \sigma_1^2)(\delta_2^2 + \sigma_2^2)$.

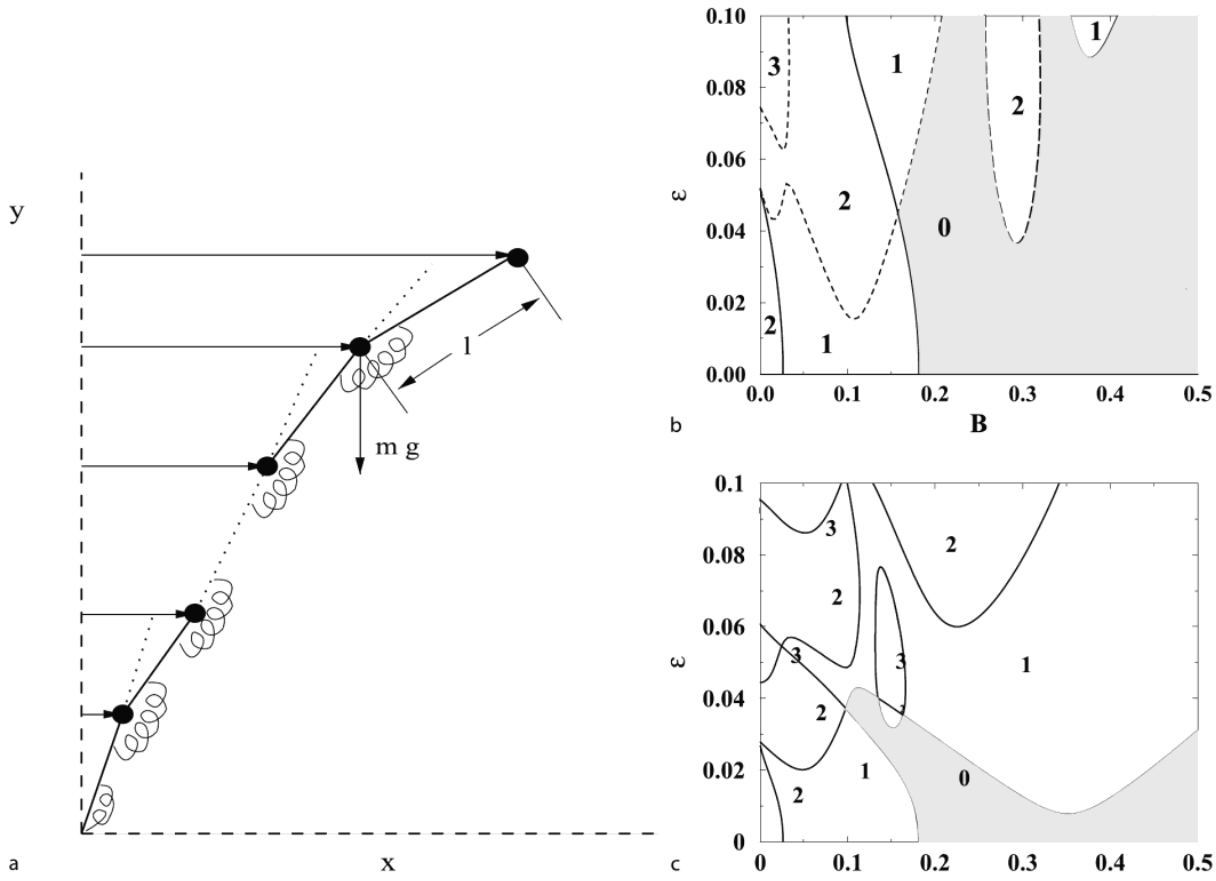
Equation (20) gives the amplitude of the quenched solution for the structure. Compare this with the amplitude R_0 of the semi-trivial solution (17), which is the forced response of the structure without the pendulum-tuned-damper added. Written in the rescaled coordinates, this solution is

$$R_0 = \frac{A^2}{\sigma_1^2 + \delta_1^2}\,.$$

Here we make the important observation that, unlike the simple forced response R_0, the quenched amplitude R_1 is independent of the forcing amplitude A, and does not blow up at the fundamental resonance $\sigma_1 \to 0$. In fact, at least to leading order in this asymptotic approximation, the amplitude of the quenched solution is *independent of the frequency detuning and damping constant of the structure, and depends only on the frequency detuning and proportional damping of the pendulum*!

Tondl et al. [58] give a number of applications of autoparametric resonance. First, the above quenching effect means that analogues of the mass-spring-pendulum system can be used as tuned dampers, that are capable of suppressing certain undesirable response frequencies from structures or machines. Another important example of parametric resonance they cite is in flow-induced oscillation, where periodic flow features such as vortex shedding can cause a parametric response of a structure. The galloping of cables in strong winds is an example of just such an effect. Cable-stayed bridges are engineering structures that are particularly susceptible to autoparametric effects. Not only is there potential for resonant interaction between the cables and the wind, but, since the cables are all of different lengths, there is a good chance that there is at least one cable whose natural frequency is either half or twice that of a global vibration mode of the entire bridge. Models of deck-cable interaction have shown the propensity for undesirable effects, see e. g. [23,24]. For example, vibrations caused by traffic could cause particular cables to undergo large-amplitude vibration leading to potential problems with wear. Alternatively, large-amplitude cable vibrations (caused for example by fluid-structure interaction) could cause significant deck vibration leading to an uncomfortable ride for those crossing the bridge.

Copyrighted material



Dynamics of Parametric Excitation, Figure 8

a A multiple pendulum with stiff joints. b,c Ince–Strutt diagrams of dimensionless amplitude of vibration ε against scaled elastic stiffness B for eight identical jointed pendulums with weak damping. b For dimensionless applied frequency $\omega = 10$; c $\omega = 20$. Shading represents the stability region of the upright equilibrium and digits in other regions give the number of unstable Floquet multipliers. After [22], to which the reader is referred to for the precise details; reproduced with permission from Elsevier

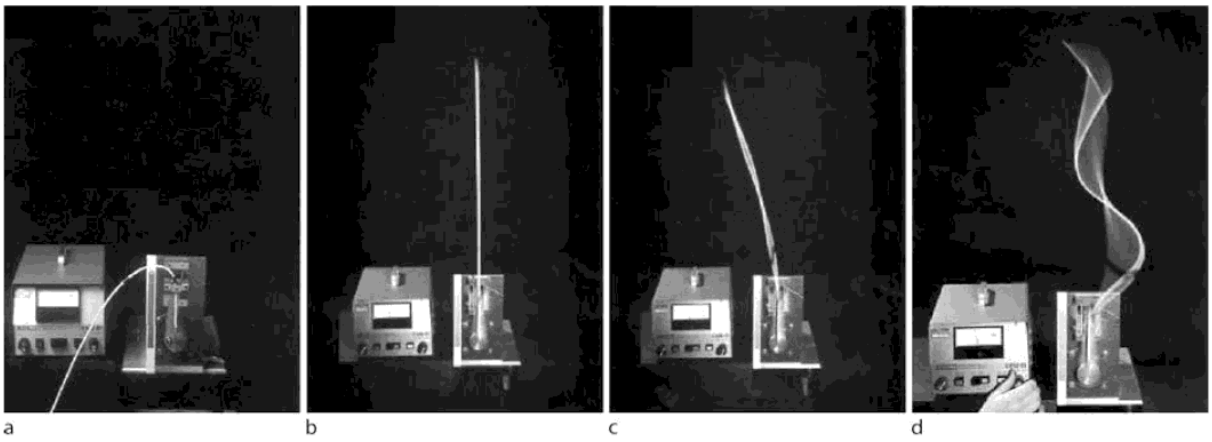
It has been suggested that the limit of such a multiple pendulum, in which the total length and mass of the system stays constant while the number of pendulums in the chain becomes infinite, could be used as a possible explanation of the so called '*Indian rope trick*', a classic conjuring trick in which a rope is 'charmed' by a magician to appear vertically out of a basket like a snake. Unfortunately, as noted [1], in this limit, which is that of a piece of string, the stability region becomes vanishingly small and so this potential scientific explanation of the magic trick fails.

However, a further experiment by Mullin announced in [4], demonstrated that a piece of 'bendy curtain wire', clamped at the bottom and free at the top, that is just too long to support its own weight can be stabilized by parametric excitation (see Fig. 9). In order to explain this result, Galán et al. [22] introduced the model depicted in Fig. 8 and studied the stability of the vertical equilibrium position in the presence of both damping and stiffness in each of the joints. The normal modes of this problem lead to fully coupled equations and so Acheson's argument does not apply. Using numerical bifurcation analysis Galán et al. were able to find stability boundaries and also to find an appropriate continuum limit; providing a good motivation for us to turn our attention to parametric resonance in continuous systems.

Continuous Systems

Parametric resonance is also of importance in continuum systems. Canonical effects we shall discuss here include the wire stiffening we have just introduced, the excitation of patterns at fluid interfaces, more general theories of flow-induced oscillations of structures, and the stabilizing effects of periodic modulation of optical fibers.

Copyrighted material



Dynamics of Parametric Excitation, Figure 9

Experimental results on a piece of 'bendy curtain wire' which is just long enough to buckle under its own weight (a). For tiny vertical oscillations between about 15 and 35 Hz, the wire can be made to stand upright (b), and to be stable against even quite large perturbations (c). Beyond the upper frequency threshold, a violent dynamic stability sets in which involves the third vibration mode excited harmonically (d). After [44], reproduced with permission from the Royal Society

Structural Stiffening; the 'Indian Rod Trick'

Thomsen [57] has shown that parametric excitation generally speaking has a stiffening effect on a structure. This is essentially because of the same effect that causes the upside down stability of the simple pendulum for high-enough frequencies and small enough amplitude. However, detailed experiments by Mullin and co-workers [44], reproduced in Fig. 9, showed that there can also be a destabilizing effect of parametric resonance. In particular, a piece of curtain wire (a thin helical spring surrounded by plastic) that was just too long (about 50 cm) to support its own weight, was clamped vertically upright and subjected to rapid, small-amplitude oscillations of about 1 cm peak-to-peak. At around 15 Hz the rod was found to be stabilized, such that when released it remains erect (see Fig. 9b). This effect continues for higher frequencies, until about 30 Hz at which point a violent dynamic instability sets in that is commensurate with the driving frequency (harmonic, rather than subharmonic). The experimentally determined parameter values in which the upside-down stabilization can be achieved are represented in terms of dimensionless parameters in Fig. 10b.

Champneys and Fraser [14,15,20], introduced a continuum theory to explain these experimental results, which produced the shaded region in Fig. 10b, that we now explain. Consider an intrinsically straight column of length ℓ, with a uniform circular cross-section of radius $a \ll \ell$ and mass linear density m per unit length (see Fig. 10a). The column is assumed to be inextensible, unshearable and linearly elastic with bending stiffness $\hat{B}$ and material damping coefficient $\hat{\gamma}$ (where hatted variables are dimensional). The lower end is clamped upright and is to vertical displacement $\Delta \cos \omega \hat{t}$, whereas the upper end is free. In [15], it is shown that the stability of the vertical equilibrium position can be studied by analyzing solutions to the PDE

$$\begin{aligned} &Bu_{ssss} + [(1-s)u_s]_s \\ &\quad + \eta\,(u_{tt} + \gamma u_{sssst} - \varepsilon[(1-s)u_s]_s \cos t) = 0\,, \end{aligned} \tag{27}$$

for the scaled lateral displacement $u(s,t)$ assumed without loss of generality to be in a fixed coordinate direction, where the scaled arc length $s \in (0,1)$, the scaled time variable $t = \omega \hat{t}$, and the boundary conditions are

$$\begin{aligned} &u = u_s = 0\,, \quad \text{at } s = 0\,; \\ &u_{ss} = u_{ssss} = 0\,, \text{ at } s = 1\,. \end{aligned} \tag{28}$$

The dimensionless parameters appearing in (27) are

$$\gamma = \frac{\Gamma \ell \omega}{mg}\,, \quad B = \frac{\hat{B}}{mg\ell^3}\,, \quad \eta = \frac{\omega^2 \ell}{g}\,, \quad \varepsilon = \frac{\Delta}{\ell}\,, \tag{29}$$

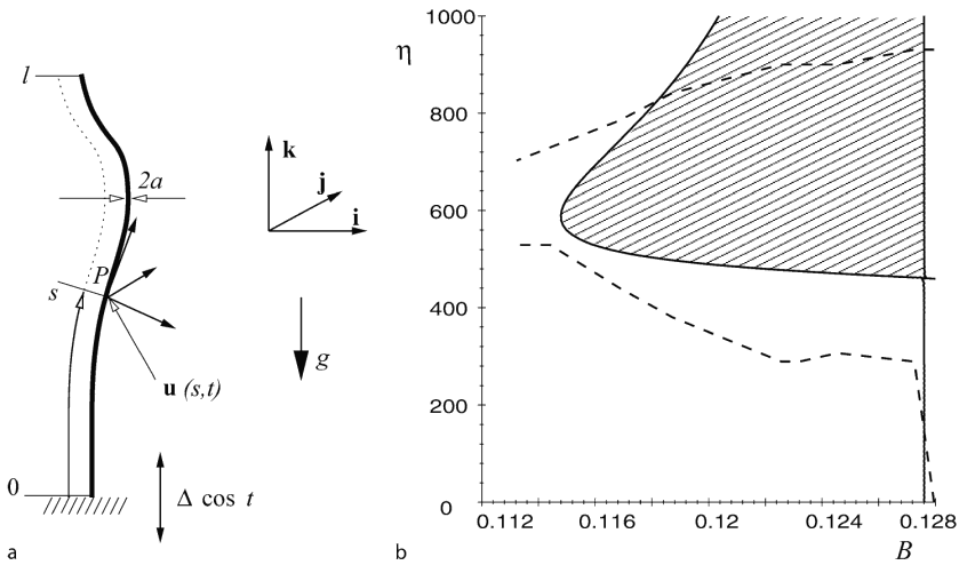
which represent respectively material damping, bending stiffness, the ratio of forcing frequency to the equivalent pendulum natural frequency, and forcing amplitude.

If we momentarily consider the problem with no forcing or damping, $\varepsilon = \gamma = 0$, then the eigenmodes ϕ_n and corresponding natural frequencies $\sqrt{\lambda_n/\eta}$ of free vibration of the rod satisfy

$$B(\phi_n)_{ssss} + [(1-s)(\phi_n)_s]_s \phi_n - \lambda_n \phi_n = 0\,,$$

and form a complete basis of functions that satisfy the boundary conditions (28). Now the eigenmodes $\phi_n(s;B)$

Copyrighted material



Dynamics of Parametric Excitation, Figure 10

a Definition sketch of the parametrically excited flexible rod model. b Comparison between theory (*solid line* and *shading*) and the experimental results of [44] (*dashed line*), for the *stability* region as a function of the dimensionless parameters B and η using the experimental value $\varepsilon = 0.02$, and the representative damping value $\gamma_1 = 0.01$ (which was not measured in the experiment) The theory is based on calculations close to the resonance tongue interaction between when $B_{0,1} = B_{1,3}$ when $\eta = \eta_c = 460.7$. After [15], reproduced with permission from SIAM

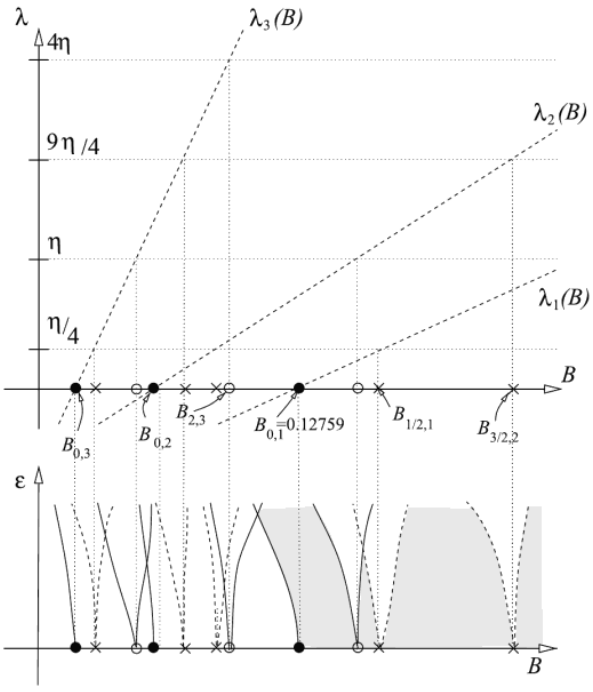
are in general not expressible in closed form except at the special values of B at which $\lambda_n(B) = 0$, in which case [20] there is a solution in terms of the Bessel function $J_{-1/3}$. The same analysis shows that there are infinitely many such B-values, $B_{0,n}$, for $n = 1, 2, 3, 4, \ldots$, accumulating at $B = 0$, the first few values of which are $B_{0,1} = 0.127594$, $B_{0,2} = 0.017864$, $B_{0,3} = 0.0067336$ and $B_{0,4} = 0.0003503$. These correspond respectively to where each eigenvalue locus $\lambda_n(B)$, for $n = 1, 2, 3, 4$, crosses the B-axis, with the corresponding eigenmodes there having $n - 1$ internal zeros. The lowering of B through each $B_{0,n}$-value implies that the nth mode becomes linearly unstable. Hence for $B > B_c := B_{0,1}$ the unforced rod is stable to self-weight buckling (a result known to Greenhill [26]). Moreover, it can be shown [14,15] that each eigenmode $\phi_n(s)$ retains a qualitatively similar mode shape for $B > B_{0,n}$ with the nth mode having $n = 1$ internal zeros, and that the corresponding loci $\lambda_n(B)$ are approximately straight lines as shown in the upper part of Fig. 11.

The lower part of Fig. 11 shows the results of a parametric analysis of the Eq. (27) for $\gamma = 0$. Basically, each time one of the eigenvalue loci passes through η times the square of a half-integer $(j/2)^2$ we reach a B-value for which there is the root point of an instability tongue. Essentially we have a Mathieu-equation-like Ince–Strutt diagram in the (B, ε) plane for *each* mode n. These diagrams are overlaid on top of each other to produce the overall stability plot. However, note from Fig. 11 that the slope of each eigenvalue locus varies with the frequency ratio η. Hence as η increases, the values of all the $B_{j/2,n}$, $j = 0, \ldots, \infty$ for a given n increase with the same fixed speed. But this speed varies with the mode number n. Hence the instability tongues slide over each other as the frequency varies.

Note the schematic diagram in the lower panel of Fig. 11 indicates a region of the stability for $B < B_{0,1}$ for small positive ε. Such a region would indicate the existence of rod of (marginally) lower bending stiffness than that required for static stability, that can be stabilized by the introduction of the parametric excitation. Essentially, the critical bending stiffness for the buckling instability is reduced (or, equivalently since $B \propto \ell^{-3}$, the critical length for buckling is increased). This effect has been dubbed the '*Indian rod trick*' because the need for bending stiffness in the wire means that mathematically this is a rod, not a rope (or string); see, e. g. the account in [4].

To determine in theory the precise parameter values at which the rod trick works turns out to be an involved process. That the instability boundary emerging from $B_{0,1}$ in the lower panel of Fig. 11 should *usually* bend back to the left can be argued geometrically [7]. However, for η-values at which $B_{0,1} \approx B_{1,n}$ for some n, it has been shown that there is actually a singularity in the coefficient of the quadratic part of this buckling curve causing the instability boundary to bend to the right [20]. A complete un-

Copyrighted material



Dynamics of Parametric Excitation, Figure 11

Summarizing schematically the results of the parametric resonance analysis of the parametric excited vertical column in the absence of damping. The *upper part* shows eigenvalue loci $\lambda_n(B)$ and the definition of the points $B_{j/2,n}$, $j = 0, 1, 2, 3, \ldots$ In the *lower part*, instability tongues in the (B, ε)-plane are shown to occur with root points $B_{j/2,n}$ and to have width ε^j. The *shaded regions* correspond to the where the vertical solution is stable. *Solid lines* represent neutral stability curves with Floquet multiplier + 1 (where α is an integer) and *dashed lines* to multipliers − 1 (where α is half an odd integer). Reproduced from [14] with permission from the Royal Society

folding of this so-called *resonance tongue interaction* was performed in [15], also allowing for the presence of material damping. The resulting asymptotic analysis needs to be carried out in the presence of *four* small parameters; ε, γ, $B - B_{0,1}$ and $\eta - \eta_c$ where η_c is the frequency-ratio value at which the resonance-tongue interaction occurs. The results produce the shaded region shown in Fig. 10b, which can be seen to agree well with the experimental results.

Faraday Waves and Pattern Formation

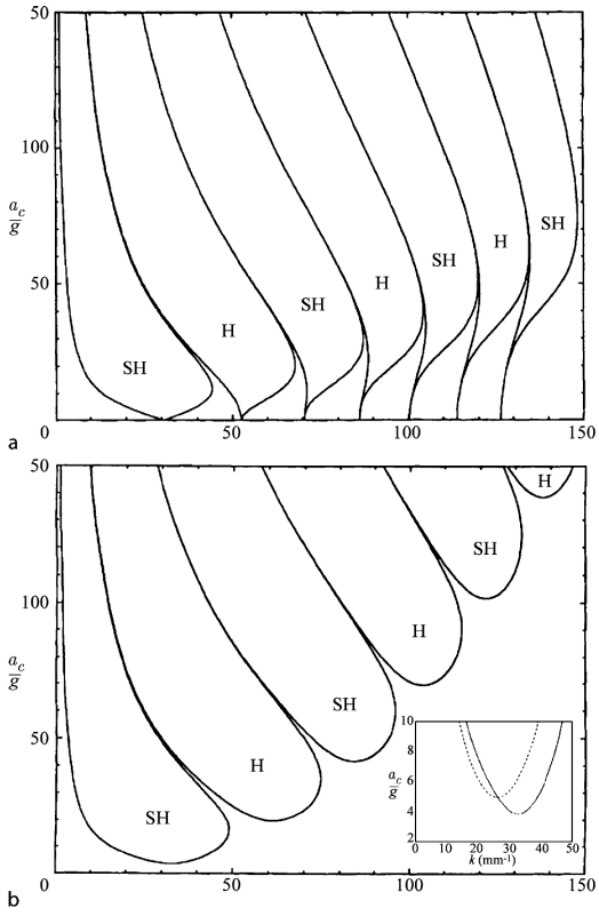
When a vessel containing a liquid is vibrated vertically, it is well known that certain standing wave patterns form on the free surface. It was Faraday in 1831 [19] who first noted that the frequency of the liquid motion is half that of the vessel. These results were confirmed by Rayleigh [50] who suggested that what we now know as parametric resonance was the cause. This explanation was shown to be correct by Benjamin and Ursell [6] who derived a system of uncoupled Mathieu equations from the Navier–Stokes equations describing the motion of an ideal (non-viscous) fluid. There is one Mathieu equation (4), for each wave number i that fits into the domain, each with its own parameters α_i and β_i. They argue that viscous effects would effectively introduce a linear damping term, as in (8) and they found good agreement between the broad principal (subharmonic) instability tongue and the results of a simple experiment. Later, Kumer and Tuckerman [36] produced a more accurate Ince–Strutt diagram (also known as a dispersion relation in this context) for the instabilities of free surfaces between two viscous fluids. Through numerical solution of appropriate *coupled* Mathieu equations that correctly take into account the boundary conditions, they showed that viscosity acts in a more complicated way than the simple velocity proportional damping term in (8). They produced the diagram shown in Fig. 12b, which showed a good match with the experimental results. Note that, compared with the idealized theory, Fig. 12a, the correct treatment of viscosity causes a small shift in the critical wavelength of each instability, a broadening of the higher-harmonic resonance tongues and a lowering of the amplitude required for the onset of the principal subharmonic resonance if damping is introduced to the idealized theory (see the inset to Fig. 12b).

More interest comes when Faraday waves are excited in a vessel where two different wave modes have their subharmonic resonance at nearly the same applied frequency, or when the excitation contains two (or more) frequencies each being close to twice the natural frequency of a standing wave. See for example the book by Hoyle [30] and references therein, for a glimpse at the amazing complexity of the patterns of vibration that can result and for their explanation in terms of dynamical systems with symmetry. Recently a new phenomenon has also been established, namely that of localized spots of pattern that can form under parametric resonance, so-called *oscillons* see e. g. [61] for a review and [39] for the first steps at an explanation.

Fluid-Structure Interaction

Faraday waves are parametric effects induced by the gross mechanical excitation of a fluid. In Subsect. "Autoparametric Resonance" we briefly mentioned that fluid-structure interaction can go the other way, namely that external fluid flow effects such as vortex shedding can excite (auto)parametric resonances in simple mechanical systems. Such flow-induced parametric resonance can also occur for continuous structures, in which waves can be excited in the structure due to external parametric loading of

Copyrighted material



Dynamics of Parametric Excitation, Figure 12

Stability boundaries for Faraday waves from: a Benjamin and Ursell's theory for ideal fluids; b the results of full hydrodynamic stability analysis by Kumar and Tuckerman. The horizontal axis is the frequency and the vertical scale the amplitude of the excitation. The *inset* to panel b shows the difference between the true instability boundary for the principal subharmonic resonance (*lower curve*) and the idealized boundary in the presence of simple proportional damping. After [36], to which the reader is referred for the relevant parameter values; reproduced with permission from Cambridge University Press

the fluid; see for example the extensive review by Païdousis and Li [47]. They also treat cases where the flow is internal to the structure: we are all familiar with the noisy ducts and pipes, for example in central heating or rumbling digestive systems, and it could well be that many of these effects are a result of parametric resonance.

Semler and Païdousis [52] consider in detail parametric resonance effects in the simple paradigm of a cantilevered (free at one end, clamped at the other) flexible pipe conveying fluid. Think of this as a hosepipe resting on a garden lawn. Many authors have considered such an apparatus as a paradigm for internal flow-induced oscillation problems (see the extensive references in [47,52]), starting from Benjamin's [5] pioneering derivation of equations of motion from the limit of a segmented rigid pipe. The equations of motion for such a situation are similar to those of the rod (27), but with significant nonlinear terms arising from inertial and fluid-structure interaction effects. In the absence of periodic fluctuations in the flow, then it is well known that there is a critical flow rate beyond which the pipe will become oscillatory, through a Hopf bifurcation (flutter instability). This effect can be seen in the way that hosepipes writhe around on a lawn, and indeed forms the basis for a popular garden toy. Another form of instability can occur when the mean flow rate is beneath the critical value for flutter, but the flow rate is subject to a periodic variation, see [52] and references therein. There, a parametric instability can occur when the pulse rate is twice the natural frequency of small amplitude vibrations of the pipe. In [52] this effect was observed in theory, in numerical simulation and in a careful experimental study, with good agreement between the three. Parametric resonance could also be found for super-critical flow rates. In this case, the effect was that the periodic oscillation born at the Hopf bifurcation was excited into a quasi-periodic response.

Another form of fluid structure interaction can occur when the properties of an elastic body immersed in a fluid are subject to a temporal periodic variation. With a view to potential biological applications such as active hearing processes and heart muscles immersed in blood, Cortez et al. [16] consider a two-dimensional patch of viscous fluid containing a ring filament that is excited via periodic variation of its elastic modulus. Using a mixture of immersed boundary numerical methods and analytic reductions, they are able to find an effective Ince–Strutt diagram for this system and demonstrate the existence of parametric instability in which the excitation of the structure causes oscillations in the fluid. Taking a fixed wave number p of the ring in space, they find a sequence of harmonic and subharmonic instabilities as a parameter effectively like α in the Mathieu equation is increased. Interestingly though, even for low viscosity, the principal subharmonic resonance tongue (corresponding to $\alpha = 1/4$) does not occur for low amplitude forcing, and the most easily excited instability is actually the fundamental, harmonic resonance (corresponding to $\alpha = 1$).

Dispersion Managed Optical Solitons

The stabilizing effect of high-frequency parametric excitation has in recent years seen a completely different application, in an area of high technological importance, namely

Copyrighted material

face of a liquid in vertical periodic motion. Proc Roy Soc Lond A 255:505–515
7. Broer H, Levi M (1995) Geometrical aspects of stability theory for Hills equations. Arch Ration Mech Anal 131:225–240
8. Broer H, Vegter G (1992) Bifurcation aspects of parametric resonance. Dynamics Reported (new series) 1:1–53
9. Broer H, Hoveijn I, van Noort M (1998) A reversible bifurcation analysis of the inverted pendulum. Physica D 112:50–63
10. Case W (1996) The pumping of a swing from the standing position. Am J Phys 64:215–220
11. Case W, Swanson M (1990) The pumping of a swing from the seated position. Am J Phys 58:463–467
12. Champneys A (1991) Homoclinic orbits in the dynamics of articulated pipes conveying fluid. Nonlinearity 4:747–774
13. Champneys A (1993) Homoclinic tangencies in the dynamics of articulated pipes conveying fluid. Physica D 62:347–359
14. Champneys A, Fraser W (2000) The 'Indian rope trick' for a continuously flexible rod; linearized analysis. Proc Roy Soc Lond A 456:553–570
15. Champneys A, Fraser W (2004) Resonance tongue interaction in the parametrically excited column. SIAM J Appl Math 65:267–298
16. Cortez R, Peskin C, Stockie J, Varela D (2004) Parametric resonance in immersed elastic boundaries. SIAM J Appl Math 65:494–520
17. Curry S (1976) How children swing. Am J Phys 44:924–926
18. Cvitanovic P (1984) Universality in Chaos. Adam Hilger, Bristol
19. Faraday M (1831) On the forms of states of fluids on vibrating elastic surfaces. Phil Trans Roy Soc Lond 52:319–340
20. Fraser W, Champneys A (2002) The 'Indian rope trick' for a parametrically excited flexible rod: nonlinear and subharmonic analysis. Proc Roy Soc Lond A 458:1353–1373
21. Galan J (2002) Personal communication
22. Galán J, Fraser W, Acheson D, Champneys A (2005) The parametrically excited upside-down rod: an elastic jointed pendulum model. J Sound Vibration 280:359–377
23. Gattulli V, Lepidi M (2003) Nonlinear interactions in the planar dynamics of cable-stayed beam. Int J Solids Struct 40:4729–4748
24. Gattulli V, Lepidi M, Macdonald J, Taylor C (2005) One-to-two global-local interaction in a cable-stayed beam observed through analytical, finite element and experimental models. Int J Nonlin Mech 40:571–588
25. Gordon J, Haus H (1986) Random walk of coherently amplified solitons in optical fiber transmission. Opt Lett 11:665–666
26. Greenhill A (1881) Determination of the greatest height consistent with stability that a pole or mast can be made. Proceedings of the Cambridge Philosophical Society IV, Oct 1880 – May 1883, pp 65–73
27. Guckenheimer J, Holmes P (1983) Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer, New York
28. Hasegawa A, Tappert F (1991) Transmission of stationary nonlinear optical pulses in dispersive dielectric fibres. i. anomolous dispersion. Appl Phys Lett 23:142–144
29. Hill G (1886) On the part of the motion of lunar perigee which is a function of the mean motions of the sun and moon. Acta Math 8:1–36
30. Hoyle R (2005) Pattern Formation; An Introduction to Methods. CUP, Cambridge
31. Ince E (1956) Ordinary Differential Equations. Dover, New York
32. Jones C (2003) Creating stability from instability. In: Nonlinear dynamics and chaos; where do we go from here? IOP Publishing, Bristol, chap 4, pp 73–90
33. Jordan DW, Smith P (1998) Nonlinear Ordinary Differential Equations, 3rd edn. Oxford University Press, Oxford
34. Kapitza P (1951) Dinamicheskaya ustoichivost mayatnika pri koleblyushcheisya tochke podvesa. Z Eksperimentalnoi I Teoreticheskoi Fiziki 21:588–597, in Russian
35. Knoz F, Forysiak W, Doran N (1995) 10 gbit/s soliton communication systems over standard fiber at 1.55μm and the use of dispersion compensation. IEEE J Lightwave Technol 13:1960–1995
36. Kumar K, Tuckerman L (1994) Parametric instability of the interface between two fluids. J Fluid Mech 279:49–68
37. Kuznetsov YA (2004) Elements of Applied Bifurcation Theory, 3rd edn. Springer, New York
38. Lin C, Kogelnik H, Cohen L (1980) Optical-pulse equilization of low-dispersion transmission in single fibers in the 1.3–1.7μm spectral region. Opt Lett 5:476–480
39. Lloyd D, Sandstede B, Avitabilie D, Champneys A (2008) Localized hexagon patterns in the planar swift-hohenberg equation. To appear in SIAM Appl Dyn Sys
40. Magnus W, Winkler S (1979) Hill's Equation. Dover, New York
41. Mailybayev A, Seyranian AP (2001) Parametric resonance in systems with small dissipation. PMM J App Math Mech 65:755–767
42. Mathieu E (1868) Mémoire sur le mouvement vibratoire d'une membrane de forme elliptique. J Math Pure Appl 13:137–203
43. Mollenauer L, Stolen R, Gordon J (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45:1095–1097
44. Mullin A, Champneys T, Fraser W, Galan J, Acheson D (2003) The 'Indian wire trick' via parametric excitation: a comparison between theory and experiment. Proc Roy Soc Lond A 459:539–546
45. Nayfeh A (2000) Nonlinear Interactions: Analytical, Computational and Experimental Methods. Wiley Interscience, New York
46. Otterbein S (1982) Stabilization of the N-pendulum and the Indian link trick. Arch Rat Mech Anal 78:381–393
47. Païdoussis MP, Li G (1993) Pipes conveying fluid: a model dynamical problem. J Fluids Structures 7:137–204
48. Piccoli B, Kulkarni J (2005) Pumping a swing by standing and squatting: Do children pump time optimally? IEEE Control Systems Magazine 25:48–56
49. Post A, de Groot G, Daffertshofer A, Beek P (2007) Pumping a playground swing. Motor Control 11:136–150
50. Rayliegh L (1883) On the crispations of fluid resting upon a vibrating support. Phil Mag 16:50
51. Romeiras F, Romeiras F, Bondeson A, Ott E, Antonsen TM, Grebogi C (1989) Quasiperiodic forcing and the observability of strange nonchaotic attractors. Phys Scr 40:442–444
52. Semler C, Païdoussis M (1996) Nonlinear analysis of the parametric resonances of a planar fluid-conveying cantilevered pipe. J Fluids Structures 10:787–825
53. Stephenson A (1908) On a new type of dynamical stability. Mem Proc Manch Lit Phil Soc 52:1–10
54. Suzuki M, Morita N, Edagawa I, Yamamoto S, Taga H, Akiba S (1995) Reduction of gordon-haas timing jitter by periodic

Copyrighted material

dispersion compensation in soliton transmission. Electron Lett 31:2027–2035
55. Tabor M (1989) Chaos and Integrability in Nonlinear Dynamics: An Introduction. Wiley, New York
56. Thomsen J (1995) Chaotic dynamics of the partially follower-loaded elastic double pendulum. J Sound Vibr 188:385–405
57. Thomsen J (2003) Vibrations and stability: advanced theory, analysis, and tools, 2nd edn. Springer, New York
58. Tondl A, Ruijgrok T, Verhulst F, Nabergoj R (2000) Autoparametric Resonance in Mechanical Systems. Cambridge University Press, Cambridge
59. Truman M, Galán J, Champneys A (2008) An example of difference combination resonance. In preparation
60. Turytsyn S, Shapiro E, Medvedev S, Fedoruk M, Mezentsev V (2003) Physics and mathematics of dispersion-managed optical solitons. CR Physique 4:145–161
61. Umbanhowar PB, Melo F, Swinney HL (1996) Localized excitations in a vertically vibrated granular layer. Nature 382:793–796
62. Vanderbauwhede A (1990) Subharmonic branching in reversible-systems. SIAM J Math Anal 21:954–979
63. van der Pol B, Strutt M (1928) On the stability of the solutions of mathieu's equation. Phil Mag 5:18–38, sp. Iss. 7th Series
64. van Noort M (2001) The parametrically forced pendulum. a case study in 1 1/2 degree of freedom. Ph D thesis, RU Groningen
65. Verhulst F (2000) Nonlinear Differential Equations and Dynamical Systems. Springer, New York
66. Wiesenfeld K, Moss F (1995) Stochastic resonance and the benefits of noise: from ice ages to crayfish and squids. Nature pp 33–36
67. Wiggins S (2003) Introduction to Applied Nonlinear Dynamical Systems and Chaos, 2nd edn. Springer, New York
68. Yakubovich VA, Starzhinskii VM (1987) Pammetric Resonance in Linear Systems. Nauka, Moscow, in Russian
69. Zakharov V, Shabat A (1971) Exact theory of two-dimensional self focussing and one-dimensional modulation of waves in nonlinear media. Sov Phys JETP 33:77–83

Copyrighted material

Entropy in Ergodic Theory

JONATHAN L. F. KING
University of Florida,
Gainesville, USA

Article Outline

Glossary
Definition of the Subject
Entropy example: *How many questions?*
Distribution Entropy
A Gander at Shannon's Noisy Channel Theorem
The Information Function
Entropy of a Process
Entropy of a Transformation
Determinism and Zero-Entropy
The Pinsker–Field and K-Automorphisms
Ornstein Theory
Topological Entropy
Three Recent Results
Exodos
Bibliography

Glossary

Some of the following definitions refer to the "Notation" paragraph immediately below. Use ***mpt*** for '*measure-preserving transformation*'.

Measure space A ***measure space*** $(X, \mathcal{X}, \mu)$ is a set X, a ***field*** (that is, a σ-algebra) $\mathcal{X}$ of subsets of X, and a countably-additive measure $\mu\colon \mathcal{X} \to [0, \infty]$. (We often just write (X, μ), with the field implicit.) For a collection $\mathcal{C} \subset \mathcal{X}$, use $\mathrm{Fld}(\mathcal{C}) \supset \mathcal{C}$ for the smallest field including $\mathcal{C}$. The number $\mu(B)$ is the "***μ-mass of*** B".

Measure-preserving map A ***measure-preserving map*** $\psi\colon (X, \mathcal{X}, \mu) \to (Y, \mathcal{Y}, \nu)$ is a map $\psi\colon X \to Y$ such that the inverse image of each $B \in \mathcal{Y}$ is in $\mathcal{X}$, and $\mu(\psi^{-1}(B)) = \nu(B)$. A ***(measure-preserving)*** *transformation* is a measure-preserving map $T\colon (X, \mathcal{X}, \mu) \to (X, \mathcal{X}, \mu)$. Condense this notation to $(T\colon X, \mathcal{X}, \mu)$ or $(T\colon X, \mu)$.

Probability space A ***probability space*** is a measure space (X, μ) with $\mu(X) = 1$; this μ is a ***probability measure***. All our maps/transformations in this article are on probability spaces.

Factor map A ***factor map***

$$\psi\colon (T\colon X, \mathcal{X}, \mu) \to (S\colon Y, \mathcal{Y}, \nu)$$

is a measure-preserving map $\psi\colon X \to Y$ which intertwines the transformations, $\psi \circ T = S \circ \psi$. And ψ is an ***isomorphism*** if – after deleting a nullset (a mass-zero set) in each space – this ψ is a bijection and ψ^{-1} is also a factor map.

Almost everywhere (a.e.) A measure-theoretic statement holds ***almost everywhere***, abbreviated ***a.e.***, if it holds off of a nullset. (Eugene Gutkin once remarked to me that the problem with Measure Theory is … that you have to say "almost everywhere", almost everywhere.) For example, $B \overset{a.e.}{\supset} A$ means that $\mu(B \smallsetminus A)$ is zero. The *a.e.* will usually be implicit.

Probability vector A ***probability vector*** $\vec{v} = (v_1, v_2, \dots)$ is a list of non-negative reals whose sum is 1. We generally assume that probability vectors and partitions (see below) have *finitely* many components. Write "*countable probability vector/partition*", when finitely or denumerably many components are considered.

Partition A ***partition*** $\mathsf{P} = (A_1, A_2, \dots)$ splits X into pairwise disjoint subsets $A_i \in \mathcal{X}$ so that the disjoint union $\bigsqcup_i A_i$ is all of X. Each A_i is an ***atom*** of P. Use $|\mathsf{P}|$ or $^{\#}\mathsf{P}$ for the number of atoms. When P partitions a probability space, then it yields a probability vector $\vec{v}$, where $v_j := \mu(A_j)$. Lastly, use $\mathsf{P}\langle x\rangle$ to denote the P-atom that owns x.

Fonts We use the font $\mathcal{H}, \mathcal{I}, \mathcal{I}$ for *distribution-entropy*, *entropy* and the *information function*. In contrast, the script font $\mathcal{ABC}\dots$ will be used for collections of sets; usually subfields of $\mathcal{X}$. Use $\mathbb{E}(\cdot)$ for the (conditional) expectation operator.

Notation $\mathbb{Z}$ = integers, $\mathbb{Z}_+$ = positive integers, and $\mathbb{N}$ = natural numbers = $0, 1, 2, \dots$. (Some well-meaning folk use $\mathbb{N}$ for $\mathbb{Z}_+$, saying '*Nothing could be more natural than the positive integers*'. And this is why $0 \in \mathbb{N}$. Use $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ for the ***ceiling*** and ***floor*** functions; $\lfloor\cdot\rfloor$ is also called the **"greatest-integer function"**. For an interval $J := [a, b) \subset [-\infty, +\infty]$, let $[a \dots b)$ denote the ***interval of integers*** $J \cap \mathbb{Z}$ (with a similar convention for closed and open intervals). E. g., $(\mathrm{e} \dots \pi] = (\mathrm{e} \dots \pi) = \{3\}$.

For subsets A and B of the same space, Ω, use $A \subset B$ for inclusion and $A \subsetneq B$ for *proper* inclusion. The difference set $B \smallsetminus A$ is $\{\omega \in B \mid \omega \notin A\}$. Employ A^c for the complement $\Omega \smallsetminus A$. Since we work in a probability space, if we let $x := \mu(A)$, then a convenient convention is to have

$$x^c \quad \text{denote} \quad 1 - x\,,$$

since then $\mu(A^c)$ equals x^c.

Use $A \triangle B$ for the ***symmetric difference*** $[A \smallsetminus B] \cup [B \smallsetminus A]$. For a collection $\mathcal{C} = \{E_j\}_j$ of sets in Ω, let the ***disjoint union*** $\bigsqcup_j E_j$ or $\bigsqcup(\mathcal{C})$ represent the union $\bigcup_j E_j$ *and also assert* that the sets are pairwise disjoint.

Copyrighted material

Use "$\forall_{\text{large}} n$" to mean: "$\exists n_0$ such that $\forall n > n_0$". To refer to left hand side of an Eq. (20), use LhS(20); do analogously for RhS(20), the right hand side.

Definition of the Subject

The word '*entropy*' (originally German, *Entropie*) was coined by Rudolf Julius Emanuel Clausius circa 1865 [2,3], taken from the Greek $\epsilon\nu\tau\rho o\pi\iota\alpha$, '*a turning towards*'. This article thus begins (*Prolegomenon*, "introduction") and ends (*Exodos*[1], "the path out") in Greek.

Clausius, in his coinage, was referring to the thermodynamic notion in physics. Our focus in this article, however, will be the concept in measurable and topological dynamics. (Entropy in differentiable dynamics[2] would require an article by itself.) Shannon's 1948 paper [6] on Information Theory, then Kolmogorov's [4] and Sinai's [7] generalization to dynamical systems, will be our starting point. Our discussion will be of the one-dimensional case, where the acting-group is $\mathbb{Z}$.

> "My greatest concern was what to call it. I thought of calling it '*information*', but the word was overly used, so I decided to call it '*uncertainty*'. John von Neumann had a better idea, he told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function goes by that name in statistical mechanics. In the second place, and more important, **nobody knows what entropy really is**, so in a debate you will always have the advantage.'" (Shannon as quoted in [59])

Entropy example: *How many questions?*

Imagine a dartboard, Fig. 1, split in five regions $A, \ldots, E$ with known probabilities. Blindfolded, you throw a dart at the board. What is the expected number V of Yes/No questions needed to ascertain the region in which the dart landed?

Solve this by always dividing the remaining probability in half. '*Is it A?*' – if Yes, then $V = 1$. Else: '*Is it B or C?*' – if Yes, then '*Is it B?*' – if No, then the dart landed in C, and $V = 3$ was the number of questions. Evidently $V = 3$ also for regions B, D, E. Using "log" to denote base-2 logarithm[3], the expected number of questions[4] is thus

$$\begin{aligned}\mathbb{E}(V) &= \frac{1}{2}\cdot 1 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 \\ &= \sum_{j=0}^{4} p_j \log\left(\frac{1}{p_j}\right) \overset{\text{note}}{=} 2\,. \end{aligned} \tag{1}$$

Letting $\vec{v} := \left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$ be the probability vector, we can write this expectation as

$$\mathbb{E}(V) = \sum_{x\in\vec{v}} \eta(x)\,.$$

Here, $\eta : [0,1] \to [0,\infty)$ is the important function[5]

$$\eta(x) := x\log(1/x)\,, \quad \text{so extending by continuity gives } \eta(0) = 0\,. \tag{2}$$

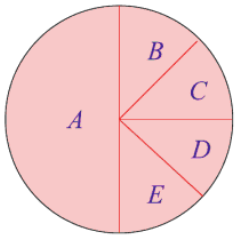
An interpretation of "$\eta(x)$" is the number of questions needed to winnow down to an event of probability x.

Distribution Entropy

Given a probability vector $\vec{v}$, define its ***distribution entropy*** as

$$\mathcal{H}(\vec{v}) := \sum_{x\in\vec{v}} \eta(x)\,. \tag{3}$$

This article will use the term ***distropy*** for '*distribution entropy*', reserving the word ***entropy*** for the corresponding



Entropy in Ergodic Theory, Figure 1
This dartboard is a probability space with a 5-set partition. The atoms have probabilities $\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}$. This probability distribution will be used later in Meshalkin's example

[1] This is the Greek spelling.

[2] For instance, see [15,18,24,25].

[3] In this paper, unmarked logs will be to base-2. In entropy theory, it does not matter much what base is used, but base-2 is convenient for computing entropy for messages described in bits. When using the natural logarithm, some people refer to the unit of information as a ***nat***. In this paper, I have picked bits rather than nats.

[4] This is holds when each probability p is a reciprocal power of two. For general probabilities, the "expected number of questions" interpretation holds in a weaker sense: Throw N darts independently at N copies of the dartboard. Efficiently ask Yes/No questions to determine where *all* N darts landed. Dividing by N, then sending $N \to \infty$, will be the $p\log\left(\frac{1}{p}\right)$ sum of Eq. (1).

[5] There does not seem to be a standard name for this function. I use η, since an uppercase η looks like an H, which is the letter that Shannon used to denote what I am calling distribution-entropy.

Copyrighted material

dynamical concept, when there is a notion of *time* involved. Getting ahead of ourselves, the *entropy* of a stationary process is the asymptotic average value that its distropy decays to, as we look at larger and larger finite portions of the process.

An equi-probable vector $\vec{v} := \left(\frac{1}{K}, \overset{K}{..}, \frac{1}{K}\right)$ evidently has $\mathcal{H}(\vec{v}) = \log(K)$. On a probability space, the "**distropy of partition** P", written $\mathcal{H}(\mathsf{P})$ or $\mathcal{H}(A_1, A_2, \ldots)$ shall mean the distropy of probability vector $j \mapsto \mu(A_j)$.

A (finite) partition necessarily has finite distropy. A *countable* partition can have finite distropy, e. g. $\mathcal{H}(1/2, 1/4, 1/8, \ldots) = 2$. One could also have infinite distropy: Consider a piece $B \subset X$ of mass $1/2^N$. Splitting B into 2^k many equal-mass atoms gives an η-sum of $2^k(k+N)/(2^k 2^N)$. Setting $k = k_N := 2^N - N$ makes this η-sum equal 1; so splitting the pieces of $X = \bigsqcup_{N=1}^{\infty} B_N$, with $\mu(B_N) = \frac{1}{2^N}$, yields an ∞-distropy partition.

Function η

The $\eta(x) = x\log(1/x)$ function[6] has vertical tangent at $x = 0$, maximum at 1/e and, when graphed in nats slope -1 at $x = 1$.

Consider partitions P and Q on the same space (X, μ). Their ***join***, written $\mathsf{P} \vee \mathsf{Q}$, has atoms $A \cap B$, for each pair $A \in \mathsf{P}$ and $B \in \mathsf{Q}$. They are ***independent***, written $\mathsf{P} \perp \mathsf{Q}$ if $\mu(A \cap B) = \mu(A)\mu(B)$ for each A, B pair. We write $\mathsf{P} \succcurlyeq \mathsf{Q}$, and say that "P ***refines*** Q", if each P-atom is a subset of some Q-atom. Consequently, each Q-atom is a union of P-atoms.

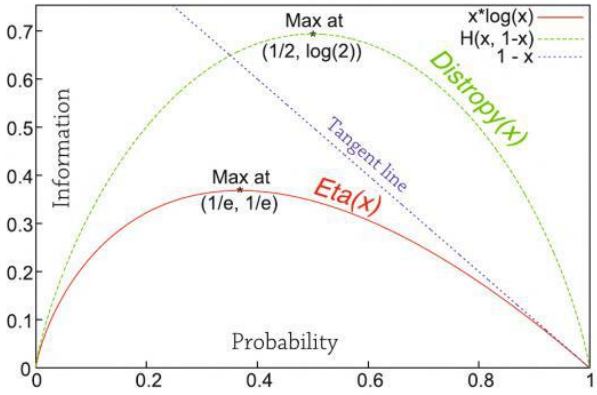
Recall, for δ a real number, our convention that δ^c means $1 - \delta$, in analogy with $\mu(B^c)$ equaling $1 - \mu(B)$ on a probability space.

Distropy Fact

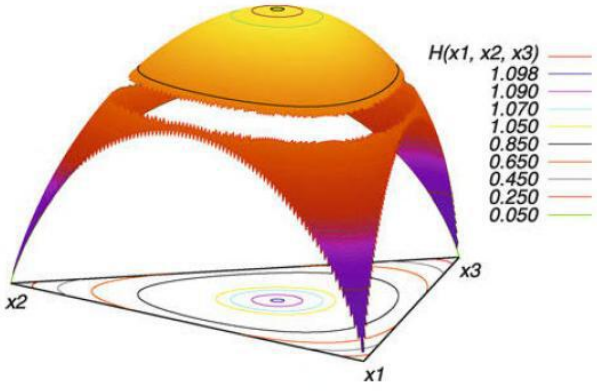
For partitions P, Q, R on probability space (X, μ):

(a) $\mathcal{H}(\mathsf{P}) \leq \log(^{\#}\mathsf{P})$, with equality IFF P is an equi-mass partition.
(b) $\mathcal{H}(\mathsf{Q} \vee \mathsf{R}) \leq \mathcal{H}(\mathsf{Q}) + \mathcal{H}(\mathsf{R})$, with equality IFF $\mathsf{Q} \perp \mathsf{R}$.
(c) For $\delta \in \left[0, \frac{1}{2}\right]$, the function $\delta \mapsto \mathcal{H}(\delta, \delta^c)$ is strictly increasing.
(d) $\mathsf{R} \preccurlyeq \mathsf{P}$ implies $\mathcal{H}(\mathsf{R}) \leq \mathcal{H}(\mathsf{P})$, with equality IFF $\mathsf{R} \overset{a.e.}{=} \mathsf{P}$.

[6] Curiosity: Just in this paragraph we compute distropy in ***nats***, that is, using natural logarithm. Given a small probability $p \in [0, 1]$ and setting $x := 1/p$, note that $\eta(p) = \log(x)/x \approx 1/\pi(x)$, where $\pi(x)$ denotes the number of prime numbers less-equal x. (This approximation is a weak form of the Prime Number Theorem.) Is there any actual connection between the '*approximate distropy*' function $\mathcal{H}_\pi(\vec{\mathsf{p}}) := \sum_{p \in \vec{\mathsf{p}}} 1/\pi(1/p)$ and Number Theory, other than a coincidence of growth rate?



Entropy in Ergodic Theory, Figure 2
Using natural log, here are the graphs of: $\eta(x)$ in *solid red*, $\mathcal{H}(x, x^c)$ in *dashed green*, $1 - x$ in *dotted blue*. Both $\eta(x)$ and $\mathcal{H}(x, x^c)$ are strictly convex-down. The $1 - x$ line is tangent to $\eta(x)$ at $x = 1$



Entropy in Ergodic Theory, Figure 3
Using natural log: The graph of $\mathcal{H}(x_1, x_2, x_3)$ in barycentric coordinates; a slice has been removed, between $z = 0.745$ and $z = 0.821$. The three arches are copies of the distropy curve from Fig. 2.

Proof Use the strict concavity of $\eta(\cdot)$, together with Jensen's inequality. □

Remark 1 Although we will not discuss it in this paper, most distropy statements remain true with '*partition*' replaced by '*countable partition of finite distropy*'.

Binomial Coefficients

The dartboard gave an example where distropy arises in a natural way. Here is a second example.

For a small $\delta > 0$, one might guess that the binomial coefficient $\binom{n}{\delta n}$ grows asymptotically (as $n \to \infty$) like $2^{\varepsilon n}$, for some small ε. But what is the correct relation between ε and δ?

Copyrighted material

Entropy of a Process

Consider a transformation $(T\colon X, \mu)$ and partition $\mathsf{P} = (A_1, A_2, \ldots)$. Each "time" n determines a partition $\mathsf{P}_n := T^n\mathsf{P}$, whose jth-atom is $T^{-n}(A_j)$. The ***process*** T, P refers to how T acts on the subfield $\bigvee_0^\infty \mathsf{P}_n \subset \mathcal{X}$. (An alternative view of a process is as a stationary sequence $V_0, V_1, \ldots$ of random variables $V_n\colon X \to \mathbb{Z}_+$, where $V_n(x) := j$ because x is in the jth-atom of P_n.)

Write $\mathcal{E}(T, \mathsf{P})$ or $\mathcal{E}^T(\mathsf{P})$ for the "***entropy***" of the T, P process". It is the limit of the ***conditional-distropy-numbers***

$$c_n := \mathcal{H}(\mathsf{P}_0|\mathsf{P}_1 \vee \mathsf{P}_2 \vee \cdots \vee \mathsf{P}_{n-1})\,.$$

This limit exists since $\mathcal{H}(\mathsf{P}) = c_1 \geq c_2 \geq \cdots \geq 0$.

Define the ***average-distropy-number*** $\frac{1}{n}h_n$, where

$$h_n := \mathcal{H}(\mathsf{P}_0 \vee \mathsf{P}_1 \vee \cdots \vee \mathsf{P}_{n-1})\,.$$

Certainly $h_n = c_n + \mathcal{H}(\mathsf{P}_1 \vee \cdots \vee \mathsf{P}_{n-1}) = c_n + h_{n-1}$, since T is measure preserving. Induction gives $h_n = \sum_{j=1}^n c_j$. So the Cesàro averages $\frac{1}{n}h_n$ converge to the entropy.

Theorem 5 *The entropy of process* $(T, P\colon X, \mathcal{X}, \mu)$ *equals*

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}\mathcal{H}(P_0 \vee \cdots \vee P_{n-1})\\ &\quad= \lim_{n\to\infty}\mathcal{H}\left(P_0\middle|\bigvee_1^n P_j\right)\\ &\quad= \mathcal{H}\left(P_0\middle|\bigvee_1^\infty P_j\right).\end{aligned}$$

Both limits are non-increasing. The entropy $\mathcal{E}^T(P) \geq 0$, *with equality* IFF $P \subset \bigvee_1^\infty P_j$. *And* $\mathcal{E}^T(P) \leq \mathcal{H}(P)$, *with equality* IFF T, P *is an independent process.*

Generators

We henceforth only discuss *invertible* mpts, that is, when T^{-1} is itself an mpt. Viewing the atoms of P as "letters", then, each $x \in X$ has a T, P-***name***

$$\ldots \mathrm{x}_{-2}\mathrm{x}_{-1}\mathrm{x}_0\mathrm{x}_1\mathrm{x}_2\mathrm{x}_3 \ldots,$$

where x_n is $\mathsf{P}\langle T^n(x)\rangle$, the P-letter owning $T^n(x)$.

A partition P ***generates*** (the whole field) under $(T\colon X, \mu)$, if $\bigvee_{-\infty}^\infty T^n\mathsf{P} =_\mu \mathcal{X}$. It turns out[8] that P generates IFF P ***separates points***. That is, after deleting a (T-invariant) nullset, distinct points of X have distinct T, P-names.

A finite set $[1 \ldots L]$ of integers, interpreted as an ***alphabet,*** yields the ***shift space*** $X := [1 \ldots L]^{\mathbb{Z}}$ of doubly-infinite sequences $x = (\ldots \mathrm{x}_{-1}\mathrm{x}_0\mathrm{x}_1 \ldots)$. The ***shift*** $T\colon X \to X$ acts on X by

$$T(x) := [n \mapsto \mathrm{x}_{n+1}]\,.$$

The time-zero partition P separates points, under the action of the shift. This L-atom ***time-zero partition*** has $\mathsf{P}\langle x\rangle = \mathsf{P}\langle y\rangle$ IFF $\mathrm{x}_0 = \mathrm{y}_0$. So no matter what shift-invariant measure is put on X, the time-zero partition will generate under the action of T.

Time Reversibility

A transformation need not be isomorphic to its inverse. Nonetheless, the average-distropy-numbers show that $\mathcal{E}(T^{-1}, \mathsf{P}) = \mathcal{E}(T, \mathsf{P})$; although this is not obvious from the conditioning-definition of entropy. Alternatively,

$$\begin{aligned}&\mathcal{H}\left(\mathsf{P}_0\middle|\bigvee_1^n \mathsf{P}_j\right)\\ &\quad= \mathcal{H}(\mathsf{P}_0 \vee \cdots \vee \mathsf{P}_n) - \mathcal{H}(\mathsf{P}_1 \vee \cdots \vee \mathsf{P}_n)\\ &\quad= \mathcal{H}(\mathsf{P}_{-n} \vee \cdots \vee \mathsf{P}_0) - \mathcal{H}(\mathsf{P}_{-n} \vee \cdots \vee \mathsf{P}_{-1})\\ &\quad= \mathcal{H}\left(\mathsf{P}_0\middle|\bigvee_{-1}^{-n} \mathsf{P}_j\right).\end{aligned} \tag{12}$$

Bernoulli Processes

A probability vector $\vec{\nu} := (\nu_1, \ldots, \nu_L)$ can be viewed as a measure on alphabet $[1 \ldots L]$. Let $\mu_{\vec{\nu}}$ be the resulting product measure on $X := [1 \ldots L]^{\mathbb{Z}}$, with T the shift on X and P the time-zero partition. The independent process $(T, \mathsf{P}\colon X, \mu_{\vec{\nu}})$ is called, by ergodic theorists, a ***Bernoulli process***. Not necessarily consistently, we tend to refer to the underlying *transformation* as a ***Bernoulli shift***.

The $\left(\frac{1}{2}, \frac{1}{2}\right)$-Bernoulli and the $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$-Bernoulli have different process-entropies, but perhaps their underlying transformations are isomorphic? Prior to the Kolmogorov–Sinai definition of entropy[9] of a *transformation*, this question remained unanswered.

[8] I am now at liberty to reveal that our X has always been a ***Lebesgue space***, that is, measure-isomorphic to an interval of $\mathbb{R}$ together with countably many point-atoms (points with positive mass). The equivalence of *generating* and *separating* is a technical theorem, due to Rokhlin.

Assuming μ to be Lebesgue is not much of a limitation. For instance, if μ is a finite measure on any Polish space, then μ extends to a Lebesgue measure on the μ-completion of the Borel sets. To not mince words: All spaces are Lebesgue spaces unless you are actively *looking* for trouble.

[9] This is sometimes called ***measure(-theoretic)*** *entropy* or (perhaps unfortunately) ***metric entropy***, to distinguish it from topological

Copyrighted material

Entropy of a Transformation

The Kolmogorov–Sinai definition of the entropy of an mpt is

$$\mathcal{E}(T) := \sup\{\mathcal{E}^T(\mathsf{Q}) \mid \mathsf{Q} \text{ a partition on } X\} .$$

Certainly entropy is an isomorphism invariant – but is it useful? After all, the supremum of *distropies* of partitions is always infinite (on non-atomic spaces) and one might fear that the same holds for entropies. The key observation (restated in Lemma 8c and proved below) was this, from [4] and [7].

Theorem 6 (Kolmogorov–Sinai Theorem) *If P generates under T, then $\mathcal{E}(T) = \mathcal{E}(T, P)$.*

Thereupon the $\left(\frac{1}{2}, \frac{1}{2}\right)$ and $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ Bernoulli-shifts are *not* isomorphic, since their respective entropies are $\log(2) \neq \log(3)$.

Wolfgang Krieger later proved a converse to the Kolmogorov–Sinai theorem.

Theorem 7 (Krieger Generator Theorem, 1970) *Suppose T ergodic. If $\mathcal{E}(T) < \infty$, then T has a generating partition. Indeed, letting K be the smallest integer $K > \mathcal{E}(T)$, there is a K-atom generator.*[10]

Proof See Rudolph [21], or § 5.1 in ▶ Joinings in Ergodic Theory, where Krieger's theorem is stated in terms of joinings. □

Entropy Is Continuous

Given *ordered* partitions $\mathsf{Q} = (\mathbf{B_1}, \ldots)$ and $\mathsf{Q}' = (\mathbf{B'_1}, \ldots)$, extend the shorter by null-atoms until $|\mathsf{Q}| = |\mathsf{Q}'|$. Let $Fat := \bigsqcup_j [B_j \cap B'_j]$; this set should have mass close to 1 if Q and Q' are almost the same partition. Define a new partition

$$\mathsf{Q}\triangle\mathsf{Q}' := \{Fat\} \sqcup \{B_i \cap B'_j \mid \text{with } i \neq j\} .$$

(In other words, take $\mathsf{Q} \vee \mathsf{Q}'$ and coalesce, into a single atom, all the $B_k \cap B'_k$ sets.) Topologize the space of partitions by saying[11] that $\mathsf{Q}^{(L)} \to \mathsf{Q}$ when $\mathcal{H}\left(\mathsf{Q}\triangle\mathsf{Q}^{(L)}\right) \to 0$.

Then Lemma 8b says that process-entropy varies continuously with varying the partition.

Lemma 8 *Fix a mpt $(T\colon X, \mu)$. For partitions P, Q, Q', define $R := Q\triangle Q'$ and let $\delta := \mathcal{H}(R)$. Then*

(a) $|\mathcal{H}(Q) - \mathcal{H}(Q')| \leq \delta$. (Distropy varies continuously with the partition.)
(b) $|\mathcal{E}^T(Q) - \mathcal{E}^T(Q')| \leq \delta$. (Process-entropy varies continuously with the partition.)
(c) For all partitions $Q \subset \mathrm{Fld}(T, P)$: $\mathcal{E}^T(Q) \leq \mathcal{E}^T(P)$.

Proof (of (a)) Evidently $\mathsf{Q}' \vee \mathsf{R} = \mathsf{Q}' \vee \mathsf{Q} = \mathsf{Q} \vee \mathsf{R}$. So $\mathcal{H}(\mathsf{Q}') \leq \mathcal{H}(\mathsf{Q} \vee \mathsf{R}) \leq \mathcal{H}(\mathsf{Q}) + \delta$. □

Proof (of (b)) As above,

$$\mathcal{H}\left(\bigvee_1^N \mathsf{Q}'_j\right) \leq \mathcal{H}\left(\bigvee_1^N \mathsf{Q}_j\right) + \mathcal{H}\left(\bigvee_1^N \mathsf{R}_j\right) .$$

Sending $N \to \infty$ gives $\mathcal{E}^T(\mathsf{Q}') \leq \mathcal{E}^T(\mathsf{Q}) + \mathcal{E}^T(\mathsf{R})$. Finally, $\mathcal{E}^T(\mathsf{R}) \leq \mathcal{H}(\mathsf{R})$ and so $\mathcal{E}^T(\mathsf{Q}') \leq \mathcal{E}^T(\mathsf{Q}) + \delta$. □

Proof (of (c)) Let $K := |\mathsf{Q}|$. Then there is a sequence of K-set partitions $\mathsf{Q}^{(L)} \to \mathsf{Q}$ with $\mathsf{Q}^{(L)} \preccurlyeq \bigvee_{-L}^{L} \mathsf{P}_\ell$. By above, $\mathcal{E}^T(\mathsf{Q}^{(L)}) \to \mathcal{E}^T(\mathsf{Q})$, so showing that

$$\mathcal{E}^T\left(\bigvee_{-L}^{L} \mathsf{P}_\ell\right) \stackrel{?}{\leq} \mathcal{E}^T(\mathsf{P})$$

will suffice. Note that

$$h_N := \mathcal{H}\left(\bigvee_{n=0}^{N-1} T^n \left(\bigvee_{-L}^{L} \mathsf{P}_\ell\right)\right) = \mathcal{H}\left(\bigvee_{j=-L}^{N-1+L} \mathsf{P}_j\right) .$$

So $\frac{1}{N} h_N \leq \frac{1}{N}\mathcal{H}\left(\bigvee_0^{N-1} \mathsf{P}_j\right) + \frac{1}{N} 2L\mathcal{H}(\mathsf{P})$. Now send $N \to \infty$. □

Entropy Is Not Continuous

The most common topology placed on the space Ω of mpts is the *coarse topology*[12] that Halmos discusses in his "little red book" [14].

The Rokhlin lemma (see p. 33 in [21]) implies that the isomorphism-class of *each* ergodic mpt is *dense* in Ω, (e. g., see p. 77 in [14]) disclosing that the $S \mapsto \mathcal{E}(S)$ map is exorbitantly discontinuous.

entropy. Tools known prior to entropy, such as *spectral* properties, did *not* distinguish the two Bernoulli-shifts; see ▶ Spectral Theory of Dynamical Systems for the definitions.

[10] It is an easier result, undoubtedly known much earlier, that every ergodic T has a *countable* generating partition – possibly of ∞-distropy.

[11] On the set of ordered K-set partitions (with K fixed) this convergence is the same as: $\mathsf{Q}^{(L)} \to \mathsf{Q}$ when $\mu\left(Fat\left(\mathsf{Q}^{(L)}, \mathsf{Q}\right)\right) \to 1$.

An alternative approach is the ***Rokhlin metric***, $\mathrm{Dist}(\mathsf{P}, \mathsf{Q}) := \mathcal{H}(\mathsf{P}|\mathsf{Q}) + \mathcal{H}(\mathsf{Q}|\mathsf{P})$, which has the advantage of working for *unordered* partitions.

[12] i. e, $S_n \to T$ IFF $\forall A \in \mathcal{X} : \mu(S_n{}^{-1}(A) \triangle T^{-1}(A)) \to 0$; this is a metric-topology, since our probability space is countably generated. This can be restated in terms of the unitary operator U_T on $\mathbb{L}^2(\mu)$, where $U_T(f) := f \circ T$. Namely, $S_n \to T$ in the coarse topology IFF $U_{S_n} \to U_T$ in the strong operator topology.

Copyrighted material

Indeed, the failure happens already for process-entropy with respect to a fixed partition. A Bernoulli process T, P has positive entropy. Take mpts $S_n \to T$, each isomorphic to an irrational rotation. Then each $\mathcal{E}(S_n, \mathsf{P})$ is zero, as shown in the later section "Determinism and Zero-Entropy".

Further Results When $\mathcal{F}$ is a T-invariant subfield, agree to use $T \restriction_{\mathcal{F}}$ for "T restricted to $\mathcal{F}$", which is a factor (see Glossary) of T. Transformations T and S are ***weakly isomorphic*** if each is isomorphic to a factor of the other.

The foregoing entropy tools make short shrift of the following.

Lemma 9 (Entropy Lemma) *Consider T-invariant subfields $\mathcal{G}_j$ and $\mathcal{F}$.*

(a) Suppose $\mathcal{G}_j \nearrow \mathcal{F}$. Then $\mathcal{E}\, T \restriction_{\mathcal{G}_j} \nearrow \mathcal{E}(T \restriction_{\mathcal{F}})$. In particular, $\mathcal{G} \subset \mathcal{F}$ implies that $\mathcal{E}(T \restriction_{\mathcal{G}}) \le \mathcal{E}(T \restriction_{\mathcal{F}})$, so entropy is an invariant of weak-isomorphism.

(b) $\mathcal{E}(T \restriction_{\mathcal{G}_1 \vee \mathcal{G}_2 \vee \ldots}) \le \sum_j \mathcal{E}(T \restriction_{\mathcal{G}_j})$. And $\mathcal{E}(T, Q_1 \vee Q_2 \vee \ldots) \le \sum_j \mathcal{E}(T, Q_j)$.

(c) For mpts $(S_j\colon Y_j, \nu_j)$: $\mathcal{E}(S_1 \times S_2 \times \cdots) = \sum_j \mathcal{E}(S_j)$.

(d) $\mathcal{E}(T^{-1}) = \mathcal{E}(T)$. More generally, $\mathcal{E}(T^n) = |n| \cdot \mathcal{E}(T)$.

Meshalkin's Map

In the wake of Kolmogorov's 1958 entropy paper, for two Bernoulli-shifts to be isomorphic one now knew that they had to have equal entropies. Meshalkin provided the first non-trivial example in 1959 [45].

Let $S : Y \to Y$ be the Bernoulli-shift over the "letter" alphabet $\{\mathtt{E}, \mathtt{D}, \mathtt{P}, \mathtt{N}\}$, with probability distribution $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. The letters $\mathtt{E}, \mathtt{D}, \mathtt{P}, \mathtt{N}$ stand for *Even, oDd, Positive, Negative*, and will be used to describe the code (isomorphism) between the processes.

Use $T : X \to X$ for the Bernoulli-shift over "digit" alphabet $\{0, +1, -1, +2, -2\}$, with probability distribution $\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$. Both distributions $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ and $\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$ have distropy log(4).

After deleting invariant nullsets from X and Y, the construction will produce a measure-preserving isomorphism $\psi\colon X \to Y$ so that $T \circ \psi = \psi \circ S$.

The Code In X, consider this point x:

$$\ldots \; \mathbf{0}\; \mathbf{0}\; \mathbf{0}\; -1\; \mathbf{0}\; \mathbf{0}\; +1\; +2\; -1\; +1\; \mathbf{0}\; \ldots$$

Regard each **0** as a left-parenthesis, and each non-zero as a right-parenthesis. Link them according to the legal way of matching parentheses, as shown in the top row, below:

$$\begin{array}{cccccccccccc} \mathbf{0} & \mathbf{0} & \mathbf{0} & -1 & \mathbf{0} & \mathbf{0} & +1 & +2 & -1 & +1 & \mathbf{0} \\ \mathtt{P} & \mathtt{N} & \mathtt{N} & \mathtt{D} & \mathtt{P} & \mathtt{P} & \mathtt{D} & \mathtt{E} & \mathtt{D} & \mathtt{D} & ? \end{array}$$

The leftmost **0** is linked to the rightmost +1, as indicated by the longest-overbar. The left/right-parentheses form a $\left(\frac{1}{2}, \frac{1}{2}\right)$-random-walk. Since this random walk is recurrent, every position in x will be linked (except for a nullset of points x).

Below each **0**, write "$\mathtt{P}$" or "$\mathtt{N}$" as the **0** is linked to a *positive* or *negative* digit. And below the other digits, write "$\mathtt{E}$" or "$\mathtt{D}$" as the digit is *even* or *odd*. So the upper name in X is mapped to the lower name, a point $y \in Y$.

This map $\psi\colon X \to Y$ carries the upstairs $\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$ distribution to $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$, downstairs. It takes some arguing to show that independence is preserved.

The inverse map ψ^{-1} views $\mathtt{D}$ and $\mathtt{E}$ as right-parentheses, and $\mathtt{P}$ and $\mathtt{N}$ as left. Above $\mathtt{D}$, write the odd digit $+\mathbf{1}$ or $-\mathbf{1}$, as this $\mathtt{D}$ is linked to Positive or Negative.

Markov Shifts

A Bernoulli process T, P has independence $\mathsf{P}_{(-\infty \ldots 0]} \perp \mathsf{P}_1$ whereas a ***Markov process*** is a bit less aloof:

> The infinite Past $\mathsf{P}_{(-\infty \ldots 0]}$ doesn't provide any more information about Tomorrow than Today did.

That is, the conditional distribution $\mathsf{P}_1 | \mathsf{P}_{(-\infty \ldots 0]}$ equals $\mathsf{P}_1 | \mathsf{P}_0$. Equivalently,

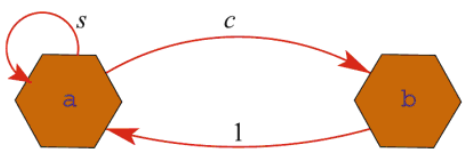
$$\mathcal{H}(\mathsf{P}_1 | \mathsf{P}_0) = \mathcal{H}\left(\mathsf{P}_1 | \mathsf{P}_{(-\infty \ldots 0]}\right) \overset{\text{note}}{=} \mathcal{E}(T, \mathsf{P})\,. \tag{13}$$

The simplest non-trivial Markov process $(T, \mathsf{P}\colon X, \mu)$ is over a two-letter alphabet $\{\mathtt{a}, \mathtt{b}\}$, and has transition graph Fig. 4, for some choice of transition probabilities s and c. The graph's Markov matrix is

$$\mathbf{M} = [m_{i,j}]_{i,j} = \begin{bmatrix} s & c \\ 1 & 0 \end{bmatrix},$$

where $c = 1 - s$, and $m_{i,j}$ denotes the probability of going from state i to state j.

If Today's distribution on the two states is the probability-vector $\vec{\nu} := [p_{\mathtt{a}} \quad p_{\mathtt{b}}]$, then Tomorrow's is the product $\vec{\nu}\mathbf{M}$. So a *stationary* process needs $\vec{\nu}\mathbf{M} = \vec{\nu}$. This equation has the unique solution $p_{\mathtt{a}} = \frac{1}{1+c}$ and $p_{\mathtt{b}} = \frac{c}{1+c}$.



Entropy in Ergodic Theory, Figure 4
Call the transition probabilities s := Prob(a→a) for stay, and c := Prob a→b for change. These are non-negative reals, and $s + c = 1$

Copyrighted material

An example of computing the probability of a word or cylinder set; (see Sect. "The Carathéodory Construction" in ▶ Measure Preserving Systems) in the process, is

$$\mu_s(\mathtt{baaaba}) = p_{\mathtt{b}} m_{\mathtt{ba}} m_{\mathtt{aa}} m_{\mathtt{aa}} m_{\mathtt{ab}} m_{\mathtt{ba}} = \frac{c}{1+c} \cdot 1 \cdot s \cdot s \cdot c \cdot 1 \,.$$

The subscript on μ_s indicates the dependence on the transition probabilities; let's also mark the mpt and call it T_s. Using Eq. (13), the entropy of our Markov map is

$$\mathcal{E}(T_s) = p_{\mathtt{a}} \cdot \mathcal{H}(s,c) + p_{\mathtt{b}} \cdot \underbrace{\mathcal{H}(1,0)}_{=0} = \frac{-1}{1+c} \cdot [s\log(s) + c\log(c)] \,. \tag{14}$$

Determinism and Zero-Entropy

Irrational rotations have zero-entropy; let's reveal this in two different ways.

Equip $X := [0,1)$ with "length" (Lebesgue) measure and wrap it into a circle. With "$\oplus$" denoting addition mod-1, have $T\colon X \to X$ be the rotation $T(x) := x \oplus \boldsymbol{\alpha}$, where the rotation number $\boldsymbol{\alpha}$ is irrational. Pick distinct points $y_0, z_0 \in X$, and let P be the partition whose two atoms are the intervals $[y_0, z_0)$ and $[z_0, y_0)$, wrapping around the circle.

The T-orbit of each point x is dense[13] in X. In particular, y_0 has dense orbit, so P separates points – hence generates – under T. Our goal, thus, is $\mathcal{E}(T,\mathsf{P}) \stackrel{?}{=} 0$.

Rotations are Deterministic

The forward T-orbit of each point is dense. This is true for y_0, and so the *backward* T,P-name of each x actually tells us which point x is. I. e., $\mathsf{P} \subset \bigvee_{-\infty}^{-1} T^n\mathsf{P}$, which is our definition of "process T,P is ***deterministic***". Our P being finite, this determinism implies that $\mathcal{E}(T,\mathsf{P})$ is zero, by Theorem 5.

Counting Names in a Rotation The $\mathsf{P}_0 \vee \cdots \vee \mathsf{P}_{n-1}$ partition places n translates of points y_0 and of z_0, cutting the circle into at most $2n$ intervals. Thus $\mathcal{H}(\mathsf{P}_0 \vee \ldots \vee \mathsf{P}_{n-1}) \le \log(2n)$. And $\frac{1}{n}\log(2n) \to 0$.

Alternatively, the below SMB-theorem implies, for an ergodic process T,P, that the number of length-n names is approximately $2^{\mathcal{E}(T,\mathsf{P})\cdot n}$; this, after discarding small mass from the space. But the growth of $n \mapsto 2n$ is sub-exponential and so, for our rotation, $\mathcal{E}(T,\mathsf{P})$ must be zero.

Theorem 10 (Shannon–McMillan–Breiman Theorem (SMB-Theorem)) *Set* $\mathbf{E} := \mathcal{E}(T,P)$, *where the tuple* $(T,P\colon X,\mu)$ *is an ergodic process. Then the average information function*

$$\frac{1}{n}\mathcal{I}_{\mathsf{P}_{[0\ldots n)}}(x) \stackrel{n\to\infty}{\longrightarrow} \mathbf{E} \,, \quad \textit{for a.e.} \quad x \in X \,. \tag{15}$$

The functions $f_n := \mathcal{I}_{\mathsf{P}_{[0\ldots n)}}$ *converge to the constant function* $\boldsymbol{E}$ *both in the* $\mathbb{L}^1$*-norm and in probability.* [14]

Proof See the texts of Karl Petersen (p. 261 in [20]), or Dan Rudolph (p. 77 in [21]). □

Consequences Recall that $\mathsf{P}_{[0\ldots n)}$ means $\mathsf{P}_0 \vee \mathsf{P}_1 \vee \cdots \vee \mathsf{P}_{n-1}$, where $\mathsf{P}_j := T^j\mathsf{P}$. As usual, $\mathsf{P}_{[0\ldots n)}\langle x\rangle$ denotes the $\mathsf{P}_{[0\ldots n)}$-atom owning x.

Having deleted a nullset, we can restate Eq. (15) to now say that $\forall\varepsilon, \forall x, \forall_{\text{large}} n$:

$$1/2^{[\mathbf{E}+\varepsilon]n} \le \mu\big(\mathsf{P}_{[0\ldots n)}\langle x\rangle\big) \le 1/2^{[\mathbf{E}-\varepsilon]n} \,. \tag{16}$$

This has the following consequence. Fixing a number $\delta > 0$, we consider any set with $\mu(B) \ge \delta$ and count the number of n-names of points in B. The SMB-Thm implies

$$\forall\varepsilon, \forall_{\text{large}} n, \forall B \stackrel{\mu}{\ge} \delta\colon \quad \big|\{n\text{-names in } B\}\big| \ge 2^{[\mathbf{E}-\varepsilon]n} \,. \tag{17}$$

Rank-1 Has Zero-Entropy

There are several equivalent definitions for "rank-1 transformation", several of which are discussed in the introduction of [28]. (See Chap. 6 in [13] as well as [51] and [27] for examples of stacking constructions.)

A **rank-1 transformation** $(T\colon X,\mu)$ admits a generating partition P and a sequence of Rokhlin stacks $S_n \subset X$, with heights going to ∞, and with $\mu(S_n) \to 1$. Moreover, each of these Rokhlin stacks is P-monochromatic, that is, each level of the stack lies entirely in some atom of P.

Taking a stack of some height $2n$, let $B = B_n$ be the union of the bottom n levels of the stack. There are at most n many length-n names starting in B_n, by monochromaticity. Finally, $\mu(B_n)$ is almost $\frac{1}{2}$, so is certainly larger than $\delta := \frac{1}{3}$. Thus Eq. (17) shows that our rank-1 T has zero entropy.

[13] Fix an $\varepsilon > 0$ and an $N > 1/\varepsilon$. Points $x, T(x), \ldots, T^N(x)$ have some two at distance less than $\frac{1}{N}$; say, $\mathrm{Dist}(T^i(x), T^j(x)) < \varepsilon$, for some $0 \le i < j \le N$. Since T is an isometry, $\varepsilon > \mathrm{Dist}(x, T^k(x)) > 0$, where $k := j - i$. So the T^k-orbit of x is ε-dense.

[14] In engineering circles, this is called the Almost-everywhere equipartition theorem.

Copyrighted material

Cautions on Determinism's Relation to Zero-Entropy

A finite-valued process T, P has zero-entropy iff $\mathsf{P} \subset \bigvee_{-\infty}^{-1} \mathsf{P}_j$. Iterating gives

$$\bigvee_{0}^{\infty} \mathsf{P}_j \subset \bigvee_{-\infty}^{-1} \mathsf{P}_j \,,$$

i. e., the future is measurable with respect to the past.

This was the case with the rotation, where a point's past uniquely identified the point, thus telling us its future.

While determinism and zero-entropy mean the same thing for finite-valued processes, this fails catastrophically for real-valued (i. e., continuum-valued) processes, as shown by an example of the author's. A stationary real-valued process $\mathbf{V} = \ldots V_{-1}\ V_0\ V_1\ V_2\ \ldots$ is constructed in [40] which is simultaneously

> ***strongly deterministic***: The two values V_0, V_1 determine all of $\mathbf{V}$, future and past.

and **non-consecutively independent**. This latter means that for each bi-infinite increasing integer sequence $\{n_j\}_{j=-\infty}^{\infty}$ with no consecutive pair (always $1 + n_j < n_{j+1}$), then the list of random variables $\ldots V_{n_{-1}}\ V_{n_0}\ V_{n_1}\ V_{n_2}\ \ldots$ is an independent process.

Restricting the random variables to be *countably*-valued, how much of the example survives? Joint work with Kalikow [39] produced a countably-valued stationary $\mathbf{V}$ which is non-consecutively independent as well as deterministic. (Strong determinism is ruled out, due to cardinality considerations.) A side-effect of the construction is that $\mathbf{V}$'s time-reversal $n \mapsto V_{-n}$ is **not** deterministic.

The Pinsker–Field and K-Automorphisms

Consider the collection of ***zero-entropy sets***,

$$\mathcal{Z} = \mathcal{Z}_T := \{A \in \mathcal{X} \mid \mathcal{E}(T, (A, A^c)) = 0\} \,. \tag{18}$$

Courtesy of Lemma 9b, $\mathcal{Z}$ is a T-invariant *field*, and

$$\forall \mathsf{Q} \subset \mathcal{Z} \colon \mathcal{E}(T, \mathsf{Q}) = 0 \,. \tag{19}$$

The ***Pinsker field***[15] of T is this $\mathcal{Z}$. It is maximal with respect to Eq. (19). Unsurprisingly, the ***Pinsker factor*** $T \restriction_{\mathcal{Z}}$ has zero entropy, that is, $\mathcal{E}(T \restriction_{\mathcal{Z}}) = 0$. A transformation T is said to have ***completely-positive entropy*** if it has no (non-trivial) zero-entropy factors. That is, its Pinsker field $\mathcal{Z}_T$ is the trivial field, $\emptyset := \{\varnothing, X\}$.

[15] Traditionally, this called the *Pinsker algebra* where, in this context, "algebra" is understood to mean "σ-algebra".

K-Processes

Kolmogorov introduced the notion of a ***K-process*** or ***Kolmogorov process***, in which the present becomes asymptotically independent of the distant past. The asymptotic past of the T, P process is called its ***tail field***, where

$$\mathrm{Tail}(T, \mathsf{P}) := \bigcap_{M=1}^{\infty} \bigvee_{j=-\infty}^{-M} \mathsf{P}_j \,.$$

This T, P is a K-process if $\mathrm{Tail}(T, \mathsf{P}) = \emptyset$. This turns out to be equivalent to what we might a call a strong form of *"sensitive dependence on initial conditions"*: For each fixed length L, the distant future

$$\bigvee_{j \in (G \ldots G+L]} \mathsf{P}_j$$

becomes more and more independent of

$$\bigvee_{j \in (-\infty \ldots 0]} \mathsf{P}_j \,,$$

as the gap $G \to \infty$.

A transformation T is a ***Kolmogorov automorphism*** if it possesses a generating partition P for which T, P is a K-process.

Here is a theorem that relates the "asymptotic forgetfulness" of $\mathrm{Tail}(T, \mathsf{P}) = \emptyset$, to the lack of determinism implied by having no zero-entropy factors (See Walters, [23], p. 113. Related results appear in Berg [44]).

Theorem 11 (Pinsker-Algebra Theorem) *Suppose P is a generating partition for an ergodic T. Then* $\mathrm{Tail}(T, P)$ *equals* $\mathcal{Z}_T$.

Since $\mathcal{Z}_T$ does not depend on P, this means that all generating partitions have the same tail field, and therefore K-ness of T can be detected from any generator.

Another non-evident fact follows from the above. The ***future field*** of T, P is defined to be $\mathrm{Tail}\,(T^{-1}\mathsf{P})$. It is not obvious that if the present is independent of the distant past, then it is automatically independent of the distant future. (Indeed, the precise definitions are important; witness the *Cautions on determinism* section.) But since the entropy of a process, $\mathcal{E}(T, (A, A^c))$, equals the entropy of the time-reversed process $\mathcal{E}(T^{-1}, (A, A^c))$, it follows that $\mathcal{Z}_T$ equals $\mathcal{Z}_{T^{-1}}$.

Ornstein Theory

In 1970, Don Ornstein [46] solved the long-standing problem of showing that entropy was a complete isomorphism-

Copyrighted material

This $\mathcal{D} := \{D_z\}_z$ is a cover, and it is d_n-2ε-small. Thus $\mathrm{Cov}(n, 2\varepsilon) \le |\mathcal{D}| = |F|$.

For any metric, a *maximal* ε-separated-set is automatically ε-spanning; adjoin a putative unspanned point to get a larger separated set.

Let A be a max-cardinality d_n-ε-separated set. Take $\mathcal{C}$, a min-cardinality subcover of $\mathcal{W}_{[0\ldots n)}$. For each $z \in A$, pick a $\mathcal{C}$-patch $C_z \ni z$. Could some pair $x, y \in A$ pick the same C? Well, write $C = \bigcap_{j=0}^{n-1} T^{-j}(W_j)$, with each $W_j \in \mathcal{W}$. For every $j \in [0 \ldots n)$, then,

$$d(T^j(x), T^j(y)) \le \mathrm{Diam}(W_j) < \varepsilon \,.$$

Hence $d_n(x, y) < \varepsilon$; so $x = y$. Accordingly, the $z \mapsto C_z$ map is injective, whence $|A| \le |\mathcal{C}|$. □

Proof (of (ii)) Choose a min-cardinality d_n-δ-small cover $\mathcal{C}$. For each $C \in \mathcal{C}$ and $j \in [0 \ldots n)$, the d-$\mathrm{Diam}(T^j C) < \delta$. So there is a $\mathcal{V}$-patch $V_{C,j} \supset T^j(C)$. Hence

$$\mathcal{V}_{[0\ldots n)} \overset{\text{note}}{\ni} \bigcap_{j=0}^{n-1} T^{-j}\left(V_{C,j}\right) \supset C \,.$$

Thus $\mathcal{V}_{[0\ldots n)} \preccurlyeq \mathcal{C}$. So

$$\mathrm{Card}\left(\mathcal{V}_{[0\ldots n)}\right) \le \mathrm{Card}(\mathcal{C}) \le |\mathcal{C}| = \mathrm{Cov}(n, \delta) \,. \qquad \square$$

Proof (of (iii)) To upper-bound $\mathrm{Cov}(k + l, \varepsilon)$ let $\mathcal{V}$ and $\mathcal{W}$ be min-cardinality ε-small covers, respectively, for metrics d_k and d_l. Then $\mathcal{V} \cap T^l(\mathcal{W})$ is a ε-small for d_{k+l}. Consequently $\mathrm{Cov}(k+l, \varepsilon) \le \mathrm{Cov}(k, \varepsilon) \cdot \mathrm{Cov}(l, \varepsilon)$. Thus $n \mapsto \log(\mathrm{Cov}(n, \varepsilon))$ is subadditive. □

Proof (of (iv)) Pick a $\mathcal{V}$ from the list in Sect. "Metric Preliminaries", choose some $2\varepsilon < \mathrm{LEB}(\mathcal{V})$ followed by an ε-small $\mathcal{W}$ from Sect. "Metric Preliminaries". Pushing $n \to \infty$ gives

$$\mathsf{L}^{\mathrm{Card}}(\mathcal{V}) \le \mathsf{L}^{\mathrm{Cov}}(2\varepsilon) \le \begin{matrix} \widehat{\mathsf{L}}^{\mathrm{Spn}}(\varepsilon) \le \widehat{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon) \\ \underline{\mathsf{L}}^{\mathrm{Spn}}(\varepsilon) \le \underline{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon) \end{matrix} \le \mathsf{L}^{\mathrm{Card}}(\mathcal{W}) \,. \tag{23}$$

Now send $\mathcal{V}$ and $\mathcal{W}$ along the list in Sect. "Metric Preliminaries". □

Pretension Topological entropy takes its values in $[0, \infty]$. A useful corollary of Eq. (23) can be stated in terms of any Distance$(\cdot, \cdot)$ which topologizes $[0, \infty]$ as a compact interval.

> For each continuous $T\colon X \to X$ on a compact metric-space, the Distance$\left(\widehat{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon), \underline{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon)\right)$ goes to zero as $\varepsilon \searrow 0$. Consequently, we can pretend that the
>
> $$\mathsf{L}^{\mathrm{Sep}}(\varepsilon) = \lim_{n\to\infty} \frac{1}{n} \log(\mathrm{Sep}(n, \varepsilon)) \tag{24}$$
>
> limit exists, in arguments that subsequently send $\varepsilon \searrow 0$. Ditto for $\mathsf{L}^{\mathrm{Spn}}(\varepsilon)$.

This will be used during the proof of the **Variational Principle**. But first, here are two entropy computations which illustrate the efficacy in having several characterizations of topological entropy.

$\mathcal{E}_{\mathrm{top}}$(Isometry) = 0 Suppose $(T\colon X, d)$ is a distance-preserving map of a compact metric-space. Fixing ε, a set is d_n-ε-separated IFF it is d-ε-separated. Thus $\mathrm{Sep}(n, \varepsilon)$ does *not* grow with n. So each $\widehat{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon)$ is zero.

Topological Markov Shifts Imagine ourselves back in the days when computer data is stored on large reels of fast-moving magnetic tape. One strategy to maximize the density of binary data stored is to *not* put timing-marks (which take up space) on the tape. This has the defect that when the tape-writer writes, say, 577 consecutive 1-bits, then the tape-reader may erroneously count 578 copies of 1. We sidestep this flaw by first encoding our data so as to avoid the $11^{577}1$ word, then writing to tape.

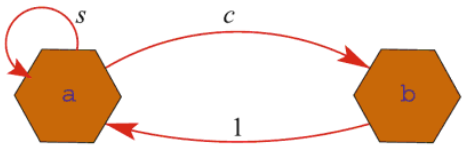
Generalize this to a finite alphabet Q and a finite list $\mathcal{F}$ of disallowed Q-words. Extend each word to a common length $K + 1$; now $\mathcal{F} \subset \mathsf{Q}^{K+1}$. The resulting "$K$-step **TMS** (***topological Markov shift***)" is the shift on the set of doubly-∞ Q-names having no substring in $\mathcal{F}$. In the above magnetic-tape example, $K = 576$. Making it more realistic, suppose that some string of zeros, say $0^{574}0$, is also forbidden[18]. Extending to length 577, we get 2^3=8 new disallowed words of form $0^{574}0b_1b_2b_3$.

We ***recode*** to a 1-step **TMS** (just called a **TMS** or a ***subshift of finite type***) over the alphabet $\mathsf{P} := \mathsf{Q}^K$. Each outlawed Q-word $w_0w_1\cdots w_K$ engenders a length-2 forbidden P-word $(w_0, \ldots, w_{K-1})(w_1, \ldots, w_K)$. The resulting **TMS** is topologically conjugate to the original K-step. The *allowed* length-2 words can be viewed as the edges in a directed-graph and the set of points $x \in X$ is the set of doubly-∞ paths through the graph. Once trivialities removed, this X is a Cantor set and the shift $T\colon X \to X$ is a homeomorphism.

The Golden Shift As the simplest example, suppose our magnetic-tape is constrained by the Markov graph, Fig. 5 that we studied measure-theoretically in Fig. 4.

[18] Perhaps the Ø-bad-length, 574, is shorter than the 1-bad-length because, say, Øs take less tape-space than 1s and so – being written more densely – cause ambiguity sooner.

Copyrighted material



Entropy in Ergodic Theory, Figure 5
Ignoring the labels on the edges, for the moment, the *Golden shift*, T, acts on the space of doubly-infinite paths through this graph. The space can be represented as a subset $X_{\text{Gold}} \subset \{\mathtt{a},\mathtt{b}\}^{\mathbb{Z}}$, namely, the set of sequences with no two consecutive $\mathtt{b}$ letters.

We want to store the text of *The Declaration of Independence* on our magnetic tape. Imagining that English is a stationary process, we'd like to encode English into this Golden **TMS** as efficiently as possible. We seek a shift-invariant measure μ on X_{Gold} of *maximum entropy*, should such exist.

View $\mathsf{P}=\{\mathtt{a},\mathtt{b}\}$ as the time-zero partition on X_{Gold}; that is, name $x=\ldots \mathtt{x}_{-1}\mathtt{x}_0\mathtt{x}_1\mathtt{x}_2\ldots$, is in atom $\mathtt{b}$ IFF letter $\mathtt{x}_0$ is "b". Any measure μ gives conditional probabilities

$$\mu(\mathtt{a}|\mathtt{a}) =: s\,, \qquad \mu(\mathtt{b}|\mathtt{a}) =: c\,,$$
$$\mu(\mathtt{a}|\mathtt{b}) \stackrel{\text{note}}{=} 1\,, \qquad \mu(\mathtt{b}|\mathtt{b}) \stackrel{\text{note}}{=} 0\,.$$

But recall, $\mathcal{E}(T) = \mathcal{H}(\mathsf{P}_1|\mathsf{P}_{[-\infty\ldots 0)}) \leq \mathcal{H}(\mathsf{P}_1|\mathsf{P}_0)$. So among all measures that make the conditional distribution $\mathsf{P}|\mathtt{a}$ equal (s, c), the *unique* one maximizing entropy is the (s, c)-Markov-process. Its entropy, derived in Eq. (14), is

$$f(s) := \frac{1}{2-s}\mathcal{H}(s, 1-s) = \frac{-1}{2-s}\big[s\log(s) + (1-s)\log(1-s)\big]\,. \tag{25}$$

Certainly $f(0) = f(1) = 0$, so f's maximum occurs at the (it turns out) *unique* point $\widehat{s}$ where the derivative $f'(\widehat{s})$ equals zero. This $\widehat{s} = (-1+\sqrt{5})/2$. Plugging in, the maximum entropy supportable by the Golden Shift is

$$\text{MaxEnt} = \frac{2}{5-\sqrt{5}}\left[\frac{-1+\sqrt{5}}{2}\log\left(\frac{2}{-1+\sqrt{5}}\right) + \frac{3-\sqrt{5}}{2}\log\left(\frac{2}{3-\sqrt{5}}\right)\right]\,. \tag{26}$$

Exponentiating, the number of μ-typical n-names grows like G^n, where

$$\mathrm{G} := \left[\frac{2}{-1+\sqrt{5}}^{\frac{-1+\sqrt{5}}{5-\sqrt{5}}}\right]\left[\frac{2}{3-\sqrt{5}}^{\frac{3-\sqrt{5}}{5-\sqrt{5}}}\right]\,. \tag{27}$$

This expression[19] looks unpleasant to simplify – it isn't even obviously an algebraic number – and yet topological entropy will reveal its familiar nature. This, because the Variational Principle (proved in the next section) says that the top-ent of a system is the supremum of measure-entropies supportable by the system.

Top-ent of the Golden Shift For a moment, let's work more generally on an arbitrary subshift (a closed, shift-invariant subset) $X \subset \mathsf{Q}^{\mathbb{Z}}$, where Q is a finite alphabet. Here, the transformation is always the shift – but the *space* is varying – so agree to refer to the top-ent as $\mathcal{E}_{\text{top}}(X)$. Let $\text{Names}_X(n)$ be the number of distinct words in the set $\{x\restriction_{[0\ldots n)} \mid x \in X\}$. Note that a metric inducing the product-topology on $\mathsf{Q}^{\mathbb{Z}}$ is

$$\mathrm{d}(x, x') := \frac{1}{1+|m|}\,, \tag{28}$$

for the smallest $|m|$ with $x_m \neq x'_m$.

Lemma 15 *Consider a subshift X. Then the*

$$\lim_{n\to\infty}\frac{1}{n}\log(\text{Names}_X(n))$$

exists in $[0,\infty]$, and equals $\mathcal{E}_{top}(X)$.

Proof With $\varepsilon \in (0,1)$ fixed, two n-names are d_n-ε-separated IFF they are not the same name. Hence $\text{Sep}(n,\varepsilon) = \text{Names}_X(n)$. □

To compute $\mathcal{E}_{\text{top}}(X_{\text{Gold}})$, declare that a word is "golden" if it appears in some $x \in X_{\text{Gold}}$. Each $[n+1]$-golden word ending in $\mathtt{a}$ has form $w\mathtt{a}$, where w is n-golden. An $[n+1]$-golden word ending in $\mathtt{b}$, must end in $\mathtt{ab}$ and so has form $w\mathtt{ab}$, where w is $[n-1]$-golden. Summing up,

$$\text{Names}_{X_{\text{Gold}}}(n+1) = \text{Names}_{X_{\text{Gold}}}(n) + \text{Names}_{X_{\text{Gold}}}(n-1)\,.$$

This is the Fibonacci recurrence, and indeed, these are *the* Fibonacci numbers, since $\text{Names}_{X_{\text{Gold}}}(0) = 1$ and $\text{Names}_{X_{\text{Gold}}}(1) = 2$. Consequently, we have that

$$\text{Names}_{X_{\text{Gold}}}(n) \sim \text{Const}\cdot\boldsymbol{\lambda}^n\,,$$

where $\boldsymbol{\lambda} = \frac{1+\sqrt{5}}{2}$ is the Golden Ratio. So the sesquipedalian number G from Eq. (27) is simply $\boldsymbol{\lambda}$, and $\mathcal{E}_{\text{top}}(X_{\text{Gold}}) = \log(\boldsymbol{\lambda})$.

Since $\log(\boldsymbol{\lambda}) \approx 0.694$, each thousand bits written on tape (subject to the "no $\mathtt{bb}$ substrings" constraint) can carry at most 694 bits of information.

[19] A popular computer-algebra-system was not, at least under my inexpert tutelage, able to simplify this. However, once top-ent gave the correct answer, the software *was* able to detect the *equality*.

Copyrighted material

*Top-ent of a General **TMS*** A (finite) digraph G engenders a **TMS** $T\colon X_G \to X_G$, as well as a $\{0,1\}$-valued adjacency matrix $\mathbf{A}{=}\mathbf{A}_G$, where $a_{i,j}$ is the number of directed-edges from state i to j. (Here, each $a_{i,j}$ is 0 or 1.) The (i,j)-entry in power $\mathbf{A}^n$ is automatically the number of length-n paths from i to j. Employing the matrix-norm $\|\mathbf{M}\| := \sum_{i,j} |m_{i,j}|$, then,

$$\|\mathbf{A}\|^n = \text{Names}_X(n)\,.$$

Happily Gelfand's formula (see 10.13 in [58] or Spectral_radius in [60]) applies: *For an arbitrary (square) complex matrix,*

$$\lim_{n\to\infty} \|\mathbf{A}^n\|^{\frac{1}{n}} = \text{SpecRad}(\mathbf{A})\,. \tag{29}$$

This right hand side, the ***spectral radius*** of **A**, means the maximum of the absolute values of **A**'s eigenvalues. So the top-ent of a **TMS** is thus the

$$\begin{aligned}\mathcal{E}_{\text{top}}(X_G) &= \text{SpecRad}(\mathbf{A}_G)\\ &:= \text{Max}\left\{|e| \;\middle|\; e \text{ is an eigenvalue of } \mathbf{A}_G\right\}\,.\end{aligned} \tag{30}$$

The $(\mathtt{a},\mathtt{b})$-adjacency matrix of Fig. 5 is

$$\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix},$$

whose eigenvalues are $\boldsymbol{\lambda}$ and $\frac{-1}{\boldsymbol{\lambda}}$.

Labeling Edges Interpret $(s,c,1)$ simply as edge-labels in Fig. 5. The set of doubly-∞ paths can also be viewed as a subset $Y_{\text{Gold}} \subset \{s,c,1\}^{\mathbb{Z}}$, and it too is a **TMS**. The shift on Y_{Gold} is conjugate (topologically isomorphic) to the shift on X_{Gold}, so they *a fortiori* have the same top-ent, $\log(\boldsymbol{\lambda})$. The $(s,c,1)$-adjacency matrix is

$$\begin{bmatrix}1 & 1 & 0\\ 0 & 0 & 1\\ 1 & 1 & 0\end{bmatrix}\,.$$

Its $|\cdot|$-largest eigenvalue is still $\boldsymbol{\lambda}$, as it must.

Now we make a new graph. We modify Fig. 5 by manufacturing a total of two s-edges, seven c-edges, and three edges $1_1, 1_2, 1_3$. Give these $2+7+3$ edges twelve *distinct* labels. We *could* compute the resulting **TMS**-entropy from the corresponding 12×12 adjacency matrix. Alternatively, look at the $(\mathtt{a},\mathtt{b})$-adjacency matrix

$$\mathbf{A} := \begin{bmatrix}2 & 7\\ 3 & 0\end{bmatrix}.$$

The roots of its characteristic polynomial are $1 \pm \sqrt{22}$. Hence $\mathcal{E}_{\text{top}}$ of this 12-symbol **TMS** is $\log(1+\sqrt{22})$.

The Variational Principle

Let $\mathcal{M} := \mathcal{M}(X,\mathrm{d})$ be the set of Borel probability measures, and $\mathcal{M}(T) := \mathcal{M}(T\colon X,\mathrm{d})$ the set of T-invariant $\mu \in \mathcal{M}$. Assign

$$\text{EntSup}(T) := \sup\{\mathcal{E}_\mu(T) \mid \mu \in \mathcal{M}(T)\}\,.$$

Theorem 16 (Variational principle (Goodson)) $\text{EntSup}(T) = \mathcal{E}_{\text{top}}(T)$.

This says that top-ent is the top entropy – *if* there is a measure μ which realizes the supremum. There doesn't have to be. Choose a sequence of metric-systems $(S_k\colon Y_k, \mathrm{m}_k)$ whose entropies *strictly* increase $\mathcal{E}_{\text{top}}(S_k) \nearrow L$ to some limit in $(0,\infty]$. Let $(S_\infty\colon Y_\infty, \mathrm{m}_\infty)$ be the identity-map on a 1-point space. Define a new system $(T\colon X,\mathrm{d})$, where $X := \bigsqcup_{k\in[1\ldots\infty]} Y_k$. Have $T(x) := S_k(x)$, for the unique k with $Y_k \ni x$. As for the metric, on Y_k let d be a scaled version of m_k, so that the d-Diam(Y_k) is less than $1/2^k$. Finally, for points in *distinct* components, $x \in Y_k$ and $z \in Y_\ell$, decree that $\mathrm{d}(x,z) := \left|2^{-k} - 2^{-\ell}\right|$. Our T is continuous, and is a homeomorphism if each of the S_k is. Certainly $\mathcal{E}_{\text{top}}(T) = L > \mathcal{E}_{\text{top}}(S_k)$, for every $k \in [1\ldots\infty]$.

If L is *finite* then there is *no* measure μ of maximal entropy; for μ must give mass to *some* Y_k; this pulls the entropy below L, since there are no compensatory components with entropy exceeding L.

In contrast, when $L = \infty$ then there *is* a maximal-entropy measure (put mass $1/2^j$ on some component Y_{k_j}, where $k_j \nearrow \infty$ swiftly); indeed, there are continuum-many maximal-entropy measures. *But* there is no [20] *ergodic* measure of maximal entropy.

For a concrete $L = \infty$ example, let S_k be the shift on $[1\ldots k]^{\mathbb{Z}}$.

Topology on $\mathcal{M}$ Let's arrange our tools for establishing the Variational Principle. The argument will follow Misiurewicz's proof, adapted from the presentations in [23] and [11].

Equip $\mathcal{M}$ with the ***weak-**** topology. [21] An $A \subset X$ is ***μ-nice*** if its topological boundary $\partial(A)$ is μ-null. And a *partition* is ***μ-nice*** if each atom is.

[20] The ergodic measures are the extreme points of $\mathcal{M}(T)$; call them $\mathcal{M}_{\text{Erg}}(T)$. This $\mathcal{M}(T)$ is the set of barycenters obtained from Borel probability measures on $\mathcal{M}_{\text{Erg}}(T)$ (see Krein-Milman_theorem, Choquet_theory in [60]). In this instance, what explains the failure to have an *ergodic* maximal-entropy measure? Let μ_k be an invariant ergodic measure on Y_k. These measures *do* converge to the one-point (ergodic) probability measure μ_∞ on Y_∞. But the map $\mu \mapsto \mathcal{E}_\mu(T)$ is not continuous at μ_∞.

[21] Measures $\alpha_L \to \mu$ IFF $\int f\mathrm{d}\alpha_L \to \int f\mathrm{d}\mu$, for each continuous $f\colon X \to \mathbb{R}$. This metrizable topology makes $\mathcal{M}$ compact. Always, $\mathcal{M}(T)$ is a non-void compact subset (see ▶ Measure Preserving Systems.)

Copyrighted material

Proposition 17 *If $\alpha_L \to \mu$ and $A \subset X$ is μ-nice, then $\alpha_L(A) \to \mu(A)$.*

Proof Define operator $\mathcal{U}(D) := \limsup_L \alpha_L(D)$. It suffices to show that $\mathcal{U}(A) \leq \mu(A)$. For since A^c is μ-nice too, then $\mathcal{U}(A^c) \leq \mu(A^c)$. Thus $\lim_L \alpha_L(A)$ exists, and equals $\mu(A)$.

Because $C := \overline{A}$ is closed, the continuous functions $f_N \searrow \mathbf{1}_C$ pointwise, where

$$f_N(x) := 1 - \mathrm{Min}(N \cdot \mathrm{d}(x, C), 1) .$$

By the Monotone Convergence theorem, then,

$$\int f_N \mathrm{d}\mu \xrightarrow{N} \mu(C) .$$

And $\mu(C) = \mu(A)$, since A is nice. Fixing N, then, it suffices to establish $\mathcal{U}(A) \leq \int f_N \mathrm{d}\mu$. But f_N is continuous, so

$$\int f_N \mathrm{d}\mu = \limsup_{L\to\infty} \int f_N \mathrm{d}\alpha_L \geq \limsup_{L\to\infty} \int \mathbf{1}_A \mathrm{d}\alpha_L = \mathcal{U}(A) . \qquad \square$$

Corollary 18 *Suppose $\alpha_L \to \mu$, and partition P is μ-nice. Then $\mathcal{H}_{\alpha_L}(\mathsf{P}) \to \mathcal{H}_\mu(\mathsf{P})$.*

The ***diameter*** of partition P is $\mathrm{Max}_{A\in\mathsf{P}}\mathrm{Diam}(A)$.

Proposition 19 *Take $\mu \in \mathcal{M}$ and $\varepsilon > 0$. Then there exists a μ-nice partition with $\mathrm{Diam}(\mathsf{P}) < \varepsilon$.*

Proof Centered at an x, the uncountably many balls $\{\mathrm{Bal}(x, r) \mid r \in (0, \varepsilon)\}$ have disjoint boundaries. So all but countably many are μ-nice; pick one and call it B_x. Compactness gives a finite nice cover, say, $\{B_1, \ldots, B_7\}$, at different centers. Then the partition $\mathsf{P} := (A_1, \ldots, A_7)$ is nice, [22] where $A_k := B_k \smallsetminus \bigcup_{j=1}^{k-1} B_j$. $\square$

Here is a consequence of Jensen's inequality.

Lemma 20 (Distropy-Averaging Lemma) *For $\mu, \nu \in \mathcal{M}$, a partition R, and a number $t \in [0, 1]$,*

$$t\mathcal{H}_\mu(\mathsf{R}) + t^c\mathcal{H}_\nu(\mathsf{R}) \leq \mathcal{H}_{t\mu+t^c\nu}(\mathsf{R}) .$$

Strategy for $\mathrm{EntSup}(T) \geq \mathcal{E}_{\mathrm{top}}(T)$. Choose an $\varepsilon > 0$. For $L = 1, 2, 3, \ldots$, take a maximal (L, ε)–separated–set $F_L \subset X$, then define

$$\mathbf{F} = \mathbf{F}_\varepsilon := \limsup_{L\to\infty} \frac{1}{L}\log(|F_L|) .$$

[22] For any two sets $B, B' \subset X$, the union $\partial B \cup \partial B'$ is a superset of the three boundaries $\partial(B \cup B')$, $\partial(B \cap B')$, $\partial(B \smallsetminus B')$.

Let $\varphi_L()$ be the equi-probable measure on F_L; each point has weight $1/|F_L|$. The desired invariant measure μ will come from the Cesàro averages

$$\alpha_L := \frac{1}{L}\sum_{\ell\in[0\ldots L)} T^\ell\varphi_L ,$$

which get more and more invariant.

Lemma 21 *Let μ be any weak-$*$ accumulation point of the above $\{\alpha_L\}_1^\infty$. (Automatically, μ is T-invariant.) Then $\mathcal{E}_\mu(T) \geq \mathbf{F}$. Indeed, if Q is any μ-nice partition with $\mathrm{Diam}(\mathsf{Q}) < \varepsilon$, then $\mathcal{E}_\mu(T, \mathsf{Q}) \geq \mathbf{F}$.*

Tactics As usual, $\mathsf{Q}_{[0\ldots N)}$ means $\mathsf{Q}_0 \vee \mathsf{Q}_1 \vee \ldots \vee \mathsf{Q}_{N-1}$. Our goal is

$$\forall N\colon \quad \mathbf{F} \stackrel{?}{\leq} \frac{1}{N}\mathcal{H}_\mu\big(\mathsf{Q}_{[0\ldots N)}\big) . \tag{31}$$

Fix N and $\mathsf{P} := \mathsf{Q}_{[0\ldots N)}$, and a $\delta > 0$. It suffices to verify: $\forall_{\mathrm{large}} L \gg N$,

$$\frac{1}{L}\log(|F_L|) \stackrel{?}{\leq} \delta + \frac{1}{N}\mathcal{H}_{\alpha_L}(\mathsf{P}) , \tag{32}$$

since this and Corollary 17 will prove Eq. (31): Pushing $L \to \infty$ along the sequence that produced μ essentially sends LhS(32) to $\mathbf{F}$, courtesy Eq. (24). And RhS(32) goes to $\delta + \frac{1}{N}\mathcal{H}_\mu(\mathsf{P})$, by Corollary 17, since P is μ-nice. Descending $\delta \searrow 0$, hands us the needed Eq. (31).

Remark 22 The idea in the following proof is to mostly fill interval $[0..L)$ with N-blocks, starting with a offset $K \in [0..N)$. Averaging over the offset will create a Cesàro average over each N-block. Averaging over the N-blocks will allow us to compute distropy with respect to the averaged measure, α_L.

Proof (of Eq. (32)) Since L is frozen, agree to use φ for the φ_L probability measure.

Our d_L-ε-separated set F_L has at most *one* point in any given atom of $\mathsf{Q}_{[0\ldots L)}$, thereupon

$$\log(|F_L|) = \mathcal{H}_\varphi\big(\mathsf{Q}_{[0\ldots L)}\big) .$$

Regardless of the "offset" $K \in [0 \ldots N)$, we can always fit $C := \lfloor\frac{L-N}{N}\rfloor$ many N-blocks into $[0 \ldots L)$. Denote by $\mathcal{G}(K) := [K \ldots K + CN)$, this union of N-blocks, the ***good*** set of indices. Unsurprisingly, $\mathcal{B}(K) := [0 \ldots L) \smallsetminus \mathcal{G}(K)$ is the ***bad*** index-set. Therefore,

$$\mathcal{H}_\varphi\big(\mathsf{Q}_{[0\ldots L)}\big) \leq \overbrace{\mathcal{H}_\varphi\Big(\bigvee_{j\in\mathcal{B}(K)} \mathsf{Q}_j\Big)}^{\mathrm{Bad}(K)} + \overbrace{\mathcal{H}_\varphi\Big(\bigvee_{j\in\mathcal{G}(K)} \mathsf{Q}_j\Big)}^{\mathrm{Good}(K)} . \tag{33}$$

Copyrighted material

Certainly $\text{Bad}(K) \le 3N\log(|\mathsf{Q}|)$. So

$$\frac{1}{NL}\sum_{K\in[0\dots N)} \text{Bad}(K) \le \frac{3N}{L}\log(|\mathsf{Q}|)\,.$$

This is less than δ, since L is large. Applying $\frac{1}{NL}\sum_{K\in[0\dots N)}$ to Eq. (28) now produces

$$\frac{1}{L}\log(|F_L|) \le \delta + \frac{1}{NL}\sum_K \text{Good}(K)\,. \tag{34}$$

Note

$$\bigvee_{j\in\mathcal{G}(K)} T^j(\mathsf{Q}) = \bigvee_{c\in[0\dots C)T^{K+cN}(\mathsf{P})}\,.$$

So

$$\text{Good}(K) \le \sum_c \mathcal{H}_\varphi(T^{K+cN}\mathsf{P})\,.$$

This latter, by definition, equals $\sum_c \mathcal{H}_{T^{K+cN}(\varphi)}(\mathsf{P})$. We conclude that

$$\begin{aligned}\frac{1}{NL}\sum_K \text{Good}(K) &\le \frac{1}{NL}\sum_K\sum_c \mathcal{H}_{T^{K+cN}\varphi}(\mathsf{P})\\ &\le \frac{1}{NL}\sum_{\ell\in[0\dots L)} \mathcal{H}_{T^\ell\varphi}(\mathsf{P})\,,\\ &\qquad\text{by adjoining a few translates of } \mathsf{P},\\ &\le \frac{1}{N}\mathcal{H}_{\alpha_L}(\mathsf{P})\,,\\ &\qquad\text{by the Distropy-averaging lemma 18,}\end{aligned}$$

since α_L is the average $\frac{1}{L}\sum_\ell T^\ell\varphi$. Thus Eq. (34) implies Eq. (32), our goal. □

Proof (of $\text{EntSup}(T) \le \mathcal{E}_{\text{top}}(T)$*)* Fix a T-invariant μ. For partition $\mathsf{Q} = (B_1,\dots,B_K)$, choose a compact set $A_k \subset B_k$ with $\mu(B_k \smallsetminus A_k)$ small. (This can be done, since μ is automatically a regular measure [58].) Letting $D := \left[\bigsqcup_i A_i\right]^c$ and $\mathsf{P} := (D, A_1,\dots,A_K)$, we can have made $\mathcal{H}(\mathsf{P}|\mathsf{Q})$ as small as desired. Courtesy of Lemma 8b, then, we only need consider partitions of the form that P has.

Open-cover $\mathcal{V} = (U_1,\dots,U_K)$ has patches $U_k := D\cup A_k$. What atoms of, say, $\mathsf{P}_{[0\dots3)}$, can the intersection $U_9 \cap T^{-1}(U_2)\cap T^{-2}(U_5)$ touch? Only the eight atoms

$$(D \text{ or } A_9)\cap T^{-1}(D \text{ or } A_2)\cap T^{-2}(D \text{ or } A_5)\,.$$

Thus $\sharp\mathsf{P}_{[0\dots n)} \le 2^n\cdot\sharp\mathcal{V}_{[0\dots n)}$. (Here, $\sharp()$ counts the number of non-void atoms/patches.) So

$$\begin{aligned}\frac{1}{n}\mathcal{H}_\mu\big(\mathsf{P}_{[0\dots n)}\big) &\le 1 + \frac{1}{n}\log\left(\sharp\mathcal{V}_{[0\dots n)}\right)\\ &\le 1 + 1 + \mathcal{E}_{\text{top}}(T)\,;\end{aligned}$$

this last inequality, when n is large. The upshot: $\mathcal{E}_\mu(T) \le 2 + \mathcal{E}_{\text{top}}(T)$.

Applied to a power T^ℓ, this asserts that $\mathcal{E}_\mu(T^\ell) \le 2 + \mathcal{E}_{\text{top}}(T^\ell)$. Thus

$$\mathcal{E}_\mu(T) \le \frac{2}{\ell} + \mathcal{E}_{\text{top}}(T)\,,$$

using Lemma 9d and using Lemma 13g. Now coax $\ell\to\infty$. □

Three Recent Results

Having given an survey of older results in measure-theoretic entropy and in topological entropy, let us end this survey with a brief discussion of a few recent results, chosen from many.

Ornstein–Weiss: Finitely-Observable Invariant

In a landmark paper [10], Ornstein and Weiss show that all "finitely observable" properties of ergodic processes are secretly entropy; indeed, they are continuous functions of entropy. This was generalized by Gutman and Hochman [9]; some of the notation below is from their paper.

Here is the setting. Consider an ergodic process, on a non-atomic space, taking on only finitely many values in $\mathbb{N}$; let $\mathcal{C}$ be some family of such processes. An ***observation scheme*** is a metric space (Ω, d) and a sequence of functions $\mathbf{S} = (S_n)_1^\infty$, where S_n maps $\mathbb{N}\times\overset{n}{\cdot}\times\mathbb{N}$ into Ω. On a point $\vec{x}\in\mathbb{N}^\infty$, the scheme ***converges*** if

$$n \mapsto S_n(x_1, x_2, \dots x_n) \tag{35}$$

converges in Ω. And on a particular process $\mathcal{X}$, say that $\mathbf{S}$ ***converges***, if $\mathbf{S}$ converges on a.e. $\vec{x}$ in $\mathcal{X}$.

A function $J\colon \mathcal{C}\to\Omega$ is isomorphism invariant if, whenever the underlying transformations of two processes $\mathcal{X},\mathcal{X}'\in\mathcal{C}$ are isomorphic, then $J(\mathcal{X}) = J(\mathcal{X}')$. Lastly, say that $\mathbf{S}$ "converges to J", if for each $\mathcal{X}\in\mathcal{C}$, scheme $\mathbf{S}$ converges to the value $J(\mathcal{X})$.

The work of David Bailey [38], a student of Ornstein, produced an observation scheme for entropy. The Lempel-Ziv algorithm [43] was another entropy observer, with practical application.

Ornstein and Weiss provided entropy schemes in [41] and [42]. Their recent paper *"Entropy is the only finitely-observable invariant"* [10], gives a converse, a uniqueness result.

Theorem 23 (Ornstein, Weiss) *Suppose J is a finitely observable function, defined on all ergodic finite-valued processes. If J is an isomorphism invariant, then J is a continuous function of the entropy.*

Copyrighted material

Abramov Formula

55. Ward T, Zhang Q (1992) The Abramov–Rohlin entropy addition formula for amenable group actions. Monatshefte Math 114:317–329

Miscellaneous

56. Newhouse SE (1989) Continuity properties of entropy. Ann Math 129:215–235
57. http://www.isib.cnr.it/control/entropy/
58. Rudin W (1973) Functional analysis. McGraw-Hill, New York
59. Tribus M, McIrvine EC (1971) Energy and information. Sci Am 224:178–184
60. Wikipedia, http://en.wikipedia.org/wiki/. pages: http://en.wikipedia.org/wiki/Spectral_radius, http://en.wikipedia.org/wiki/Information_entropy

Books and Reviews

Boyle M, Downarowicz T (2004) The entropy theory of symbolic extensions. Invent Math 156(1):119–161

Downarowicz T, Serafin J (2003) Possible entropy functions. Isr J Math 135:221–250

Hassner M (1980) A non-probabilistic source and channel coding theory. Ph D Dissertation, UCLA

Katok A, Sinai YG, Stepin AM (1977) Theory of dynamical systems and general transformation groups with an invariant measure. J Sov Math 7(6):974–1065

Copyrighted material

Ergodicity and Mixing Properties

ANTHONY QUAS
Department of Mathematics and Statistics,
University of Victoria, Victoria, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Basics and Examples
Ergodicity
Ergodic Decomposition
Mixing
Hyperbolicity and Decay of Correlations
Future Directions
Bibliography

Glossary

Bernoulli shift Mathematical abstraction of the scenario in statistics or probability in which one performs repeated independent identical experiments.

Markov chain A probability model describing a sequence of observations made at regularly spaced time intervals such that at each time, the probability distribution of the subsequent observation depends only on the current observation and not on prior observations.

Measure-preserving transformation A map from a measure space to itself such that for each measurable subset of the space, it has the same measure as its inverse image under the map.

Measure-theoretic entropy A non-negative (possibly infinite) real number describing the complexity of a measure-preserving transformation.

Product transformation Given a pair of measure-preserving transformations: T of X and S of Y, the product transformation is the map of $X \times Y$ given by $(T \times S)(x, y) = (T(x), S(y))$.

Definition of the Subject

Many physical phenomena in equilibrium can be modeled as measure-preserving transformations. Ergodic theory is the abstract study of these transformations, dealing in particular with their long term average behavior.

One of the basic steps in analyzing a measure-preserving transformation is to break it down into its simplest possible components. These simplest components are its ergodic components, and on each of these components, the system enjoys the ergodic property: the long-term time average of any measurement as the system evolves is equal to the average over the component. Ergodic decomposition gives a precise description of the manner in which a system can be split into ergodic components.

A related (stronger) property of a measure-preserving transformation is mixing. Here one is investigating the correlation between the state of the system at different times. The system is mixing if the states are asymptotically independent: as the times between the measurements increase to infinity, the observed values of the measurements at those times become independent.

Introduction

The term ergodic was introduced by Boltzmann [8,9] in his work on statistical mechanics, where he was studying Hamiltonian systems with large numbers of particles. The system is described at any time by a point of *phase space*, a subset of $\mathbb{R}^{6N}$ where N is the number of particles. The configuration describes the 3-dimensional position and velocity of each of the N particles. It has long been known that the Hamiltonian (i. e. the overall energy of the system) is invariant over time in these systems. Thus, given a starting configuration, all future configurations as the system evolves lie on the same *energy surface* as the initial one.

Boltzmann's *ergodic hypothesis* was that the trajectory of the configuration in phase space would fill out the entire energy surface. The term ergodic is thus an amalgamation of the Greek words for work and path. This hypothesis then allowed Boltzmann to conclude that the long-term average of a quantity as the system evolves would be equal to its average value over the phase space.

Subsequently, it was realized that this hypothesis is rarely satisfied. The ergodic hypothesis was replaced in 1911 by the *quasi-ergodic hypothesis* of the Ehrenfests [17] which stated instead that each trajectory is dense in the energy surface, rather than filling out the entire energy surface. The modern notion of ergodicity (to be defined below) is due to Birkhoff and Smith [7]. Koopman [44] suggested studying a measure-preserving transformation by means of the associated isometry on Hilbert space, $U_T : L^2(X) \to L^2(X)$ defined by $U_T(f) = f \circ T$. This point of view was used by von Neumann [91] in his proof of the mean ergodic theorem. This was followed closely by Birkhoff [6] proving the pointwise ergodic theorem. An ergodic measure-preserving transformation enjoys the property that Boltzmann first intended to deduce from his hypothesis: that long-term averages of an observable quantity coincide with the integral of that quantity over the phase space.

Copyrighted material

These theorems allow one to deduce a form of independence on the average: given two sets of configurations A and B, one can consider the volume of the phase space consisting of points that are in A at time 0 and in B at time t. In an ergodic measure-preserving transformation, if one computes the average of the volumes of these regions over time, the ergodic theorems mentioned above allow one to deduce that the limit is simply the product of the volume of A and the volume of B. This is the weakest mixing-type property. In this article, we will outline a rather full range of mixing properties with ergodicity at the weakest end and the Bernoulli property at the strongest end.

We will set out in some detail the various mixing properties, basing our study on a number of concrete examples sitting at various points of this hierarchy. Many of the mixing properties may be characterized in terms of the Koopman operators operators mentioned above (i. e. they are *spectral properties*), but we will see that the strongest mixing properties are not spectral in nature.

We shall also see that there are connections between the range of mixing properties that we discuss and measure-theoretic entropy. In measure-preserving transformations that arise in practice, there is a correlation between strong mixing properties and positive entropy, although many of these properties are logically independent.

One important issue for which many questions remain open is that of higher-order mixing. Here, the issue is if instead of asking that the observations at two times separated by a large time T be approximately independent, one asks whether if one makes observations at more times, each pair suitably separated, the results can be expected to be approximately independent. This issue has an analogue in probability theory, where it is well-known that it is possible to have a collection of random variables that are pairwise independent, but not mutually independent.

Basics and Examples

In this article, except where otherwise stated, the measure-preserving transformations that we consider will be defined on probability spaces.

More specifically, given a measurable space $(X, \mathcal{B})$ and a probability measure μ defined on $\mathcal{B}$, a *measure-preserving transformation* of $(X, \mathcal{B}, \mu)$ is a $\mathcal{B}$-measurable map $T\colon X \to X$ such that $\mu(T^{-1}B) = \mu(B)$ for all $B \in \mathcal{B}$.

While this definition makes sense for arbitrary measures, not simply probability measures, most of the results and definitions below only make sense in the probability measure case. Sometimes it will be helpful to make the assumption that the underlying probability space is a Lebesgue space (that is, the space together with its completed σ-algebra agrees up to a measure-preserving bijection with the unit interval with Lebesgue measure and the usual σ-algebra of Lebesgue measurable sets). Although this sounds like a strong restriction, in practice it is barely a restriction at all, as almost all of the spaces that appear in the theory (and all of those that appear in this article) turn out to be Lebesgue spaces. For a detailed treatment of the theory of Lebesgue spaces, the reader is referred to Rudolph's book [76]. The reader is referred also to the chapter on ▶ Measure Preserving Systems.

While many of the definitions that we shall present are valid for both invertible and non-invertible measure-preserving transformations, the strongest mixing conditions are most useful in the case of invertible transformations.

It will be helpful to present a selection of simple examples, relative to which we will be able to explore ergodicity and the various notions of mixing. These examples and the lemmas necessary to show that they are measure-preserving transformations as claimed may be found in the books of Petersen [64], Rudolph [76] and Walters [92]. More details on these examples can also be found in the chapter on ▶ Ergodic Theory: Basic Examples and Constructions.

Example 1 (Rotation on the circle) Let $\alpha \in \mathbb{R}$. Let $R_\alpha\colon [0,1) \to [0,1)$ be defined by $R_\alpha(x) = x + \alpha \bmod 1$. It is straightforward to verify that R_α preserves the restriction of Lebesgue measure λ to $[0,1)$ (it is sufficient to check that $\lambda(R_\alpha^{-1}(J)) = \lambda(J)$ for an interval J).

Example 2 (Doubling Map) Let $M_2\colon [0,1) \to [0,1)$ be defined by $M_2(x) = 2x \bmod 1$. Again, Lebesgue measure is invariant under M_2 (to see this, one observes that for an interval J, $M_2^{-1}(J)$ consists of two intervals, each of half the length of J). This may be generalized in the obvious way to a map M_k for any integer $k \geq 2$.

Example 3 (Interval Exchange Transformation) The class of interval exchange transformations was introduced by Sinai [85]. An interval exchange transformation is the map obtained by cutting the interval into a finite number of pieces and permuting them in such a way that the resulting map is invertible, and restricted to each interval is an order-preserving isometry.

More formally, one takes a sequence of positive lengths $\ell_1, \ell_2, \ldots, \ell_k$ summing to 1 and a permutation π of $\{1, \ldots, k\}$ and defines $a_i = \sum_{j<i} \ell_j$ and $b_i = \sum_{\pi(j)<\pi(i)} \ell_j$ (again with $b_0 = 0$). The interval exchange transformation defined by $(\ell_1, \ldots, \ell_k)$ and π is the map $T\colon [0,1) \to [0,1)$ defined by $T|_{[a_i, a_{i+1})}(x) = x + (b_i - a_i)$. It is straightforward to check that any such interval exchange transformation preserves Lebesgue measure on the unit interval.

Copyrighted material

Example 4 (Bernoulli Shift) Let A be a finite set and fix a vector $(p_i)_{i\in A}$ of positive numbers that sum to 1. Let $A^{\mathbb{N}}$ denote the set of sequences of the form $x_0x_1x_2\ldots$, where $x_n \in A$ for each $n \in \mathbb{N}$ and let $A^{\mathbb{Z}}$ denote the set of bi-infinite sequences of the form $\ldots x_{-2}x_{-1}\cdot x_0x_1x_2\ldots$ (the $\cdot$ is a placeholder that allows us to distinguish (for example) between the sequences $\ldots 01010\cdot 10101\ldots$ and $\ldots 10101\cdot 01010\ldots$).

We define a map (the *shift map*) S on $A^{\mathbb{N}}$ by $(S(x))_n = x_{n+1}$ and define S on $A^{\mathbb{Z}}$ by the same formula. Note that S is invertible as a transformation on $A^{\mathbb{Z}}$ but non-invertible as a transformation on $A^{\mathbb{N}}$.

We need to equip $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ with measures. This is done by defining the measure of a preferred class of sets, checking certain consistency conditions and appealing to the Kolmogorov extension theorem. Here the preferred sets are the *cylinder sets*. Given $m \le n$ in the invertible case and a sequence $a_m \ldots a_n$, we let $[a_m \ldots a_n]_m^n$ denote $\{x \in A^{\mathbb{Z}} : x_m = a_m, \ldots, x_n = a_n\}$ and define $\mu([a_m \ldots a_n]_m^n) = p_{a_m}p_{a_{m+1}}\ldots p_{a_n}$. This is then shown to uniquely define a measure μ on the σ-algebra of $A^{\mathbb{Z}}$ generated by the cylinder sets. It is immediate to see that for any cylinder set C, $\mu(S^{-1}C) = \mu(C)$, and it follows that S is a measure-preserving transformation of $(A^{\mathbb{Z}}, \mathcal{B}, \mu)$. The construction is exactly analogous in the non-invertible case. See the chapter on ▶ Measure Preserving Systems or the books of Walters [92] or Rudolph [76] for more details of defining measures in these systems.

The class of Bernoulli shifts will play a distinguished role in what follows.

Example 5 (Markov Shift) The spaces $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ are exactly as above, as is the shift map. All that changes is the measure.

To define a Markov shift, we need a stochastic matrix P (i. e. a matrix with non-negative entries whose rows sum to 1) with rows and columns indexed by A and a left eigenvector π for P with eigenvalue 1 with the property that the entries of π are non-negative and sum to 1. The existence of such an eigenvector is a consequence of the Perron–Frobenius theory of positive matrices. Provided that the matrix P is irreducible (for each a and a' in A, there is an $n > 0$ such that $P^n_{a,a'} > 0$), the eigenvector π is unique.

Given the pair (P, π), one defines the measure of a cylinder set by $\mu([a_m \ldots a_n]_m^n) = \pi_{a_m}P_{a_m a_{m+1}}\ldots P_{a_{n-1}a_n}$ and extends μ as before to a probability measure on $A^{\mathbb{N}}$ or $A^{\mathbb{Z}}$.

Example 6 (Hard Sphere Gases and Billiards) We wish to model the behavior of a gas in a bounded region. We make the assumption that the gas consists of a large number N of identical balls which move at constant velocity until two balls collide, whereupon they elastically swap momentum along the direction of contact. The phase space for this system is a region of $\mathbb{R}^{6N}$ (with N 3-dimensional position vectors and N 3-dimensional velocity vectors). More abstractly, the system is equivalent to the motion of a single point particle in a region of $\mathbb{R}^M \times \mathbb{R}^M$ (with the first M-vector representing position and the second representing velocity). The system is constrained in that its position is required to lie in a bounded region S of $\mathbb{R}^M$ with a piecewise smooth boundary. The system evolves by moving the position at a constant rate in the direction of the velocity vector until the point reaches ∂S, at which time the component of the velocity parallel to the normal to ∂S is reversed. This then defines a flow (i. e. a family of maps $(T_t)_{t\in\mathbb{R}}$ satisfying $T_{t+s} = T_t \circ T_s$) on the phase space. Since the magnitude of the velocity is conserved, it is convenient to restrict to flows with speed 1. This system is clearly the closest of the examples that we consider to the situation envisaged by Boltzmann. Perhaps not surprisingly, proofs of even the most basic properties for this system are much harder than those for the other examples that we consider.

We will need to make use of the concept of measure-theoretic isomorphism. Two measure-preserving transformations T of $(X, \mathcal{B}, \mu)$ and S of $(Y, \mathcal{F}, \nu)$ are *measure-theoretically isomorphic* (or just *isomorphic*) if there exist measurable maps $g\colon X \to Y$ and $h\colon Y \to X$ such that

1. $g \circ h$ and $h \circ g$ agree with the respective identity maps almost everywhere;
2. $\mu(g^{-1}F) = \nu(F)$ and $\nu(h^{-1}B) = \mu(B)$ for all $F \in \mathcal{F}$ and $B \in \mathcal{B}$; and
3. $S\circ g(x) = g\circ T(x)$ for μ-almost every x (or equivalently $T \circ h(y) = h \circ S(y)$ for ν-almost every y).

Measure-theoretic isomorphism is the basic notion of 'sameness' in ergodic theory. It is in some sense quite weak, so that systems may be isomorphic that feel very different (for example, as we discuss later, the time one map of a geodesic flow is isomorphic to a Bernoulli shift). For comparison, the notion of sameness in topological dynamical systems (topological conjugacy) is far stronger.

As an example of measure-theoretic isomorphism, it may be seen that the doubling map is isomorphic to the one-sided Bernoulli shift on $\{0, 1\}$ with $p_0 = p_1 = 1/2$ (the map g takes an $x \in [0, 1)$ to the sequence of 0's and 1's in its binary expansion (choosing the sequence ending with 0's, for example, if x is of the form $p/2^n$) and the inverse map h takes a sequence of 0's and 1's to the point in $[0, 1)$ with that binary expansion.)

Copyrighted material

Given a measure-preserving transformation T of a probability space $(X, \mathcal{B}, \mu)$, T is associated to an isometry of $L^2(X, \mathcal{B}, \mu)$ by $U_T(f) = f \circ T$. This operator is known as the *Koopman Operator*. In the case where T is invertible, the operator U_T is unitary. Two measure-preserving transformations T and S of $(X, \mathcal{B}, \mu)$ and $(Y, \mathcal{F}, \nu)$ are *spectrally isomorphic* if there is a Hilbert space isomorphism Θ from $L^2(X, \mathcal{B}, \mu)$ to $L^2(Y, \mathcal{F}, \nu)$ such that $\Theta \circ U_T = U_S \circ \Theta$. As we shall see below, spectral isomorphism is a strictly weaker property than measure-theoretic isomorphism.

Since in ergodic theory, measure-theoretic isomorphism is the basic notion of sameness, all properties that are used to describe measure-preserving systems are required to be invariant under measure-theoretic isomorphism (i. e. if two measure-preserving transformations are measure-theoretically isomorphic, the first has a given property if and only if the second does). On the other hand, we shall see that some mixing-type properties are invariant under spectral isomorphism, while others are not. If a property is invariant under spectral isomorphism, we say that it is a *spectral property*.

There are a number of mixing type properties that occur in the probability literature (α-mixing, β-mixing, ϕ-mixing, ψ-mixing etc.) (see Bradley's survey [12] for a description of these conditions). Many of these are stronger than the Bernoulli property, and are therefore not preserved under measure-theoretic isomorphism. For this reason, these properties are not widely used in ergodic theory, although β-mixing turns out to be equivalent to the so-called *weak Bernoulli* property (which turns out to be stronger than the Bernoulli property that we discuss in this article – see Smorodinsky's paper [87]) and α-mixing is equivalent to strong-mixing.

A basic construction (see the article on ▶ Ergodic Theory: Basic Examples and Constructions) that we shall require in what follows is the product of a pair of measure-preserving transformations: given transformations T of $(X, \mathcal{B}, \mu)$ and S of $(Y, \mathcal{F}, \nu)$, we define the *product transformation* $T \times S$: $(X \times Y, \mathcal{B} \otimes \mathcal{F}, \mu \times \nu)$ by $(T \times S)(x, y) = (Tx, Sy)$.

One issue that we face on occasion is that it is sometimes convenient to deal with invertible measure-preserving transformations. It turns out that given a non-invertible measure-preserving transformation, there is a natural way to uniquely associate an invertible measure-preserving transformation transformation sharing almost all of the ergodic properties of the original transformation. Specifically, given a non-invertible measure-preserving transformation T of $(X, \mathcal{B}, \mu)$, one lets $\overline{X} = \{(x_0, x_1, \dots)\colon x_n \in X \text{ and } T(x_n) = x_{n-1} \text{ for all } n\}$, $\overline{\mathcal{B}}$ be the σ-algebra generated by sets of the form $\bar{A}_n = \{\bar{x} \in \overline{X}\colon x_n \in A\}$, $\bar{\mu}(\bar{A}_n) = \mu(A)$ and $\overline{T}(x_0, x_1, \dots) = (T(x_0), x_0, x_1, \dots)$. The transformation $\overline{T}$ of $(\overline{X}, \overline{\mathcal{B}}, \overline{\mu})$ is called the *natural extension* of the transformation T of $(X, \mathcal{B}, \mu)$ (see the chapter on ▶ Ergodic Theory: Basic Examples and Constructions for more details). In situations where one wants to use invertibility, it is often possible to pass to the natural extension, work there and then derive conclusions about the original non-invertible transformation.

Ergodicity

Given a measure-preserving transformation $T\colon X \to X$, if $T^{-1}A = A$, then $T^{-1}A^c = A^c$ also. This allows us to decompose the transformation X into two pieces A and A^c and study the transformation T separately on each. In fact the same situation holds if $T^{-1}A$ and A agree up to a set of measure 0. For this reason, we call a set A *invariant* if $\mu(T^{-1}A \Delta A) = 0$.

Returning to Boltzmann's ergodic hypothesis, existence of an invariant set of measure between 0 and 1 would be a bad situation as his essential idea was that the orbit of a single point would 'see' all of X, whereas if X were decomposed in this way, the most that a point in A could see would be all of A, and similarly the most that a point in A^c could see would be all of A^c.

A measure-preserving transformation will be called *ergodic* if it has no non-trivial decomposition of this form. More formally, let T be a measure-preserving transformation of a probability space $(X, \mathcal{B}, \mu)$. The transformation T is said to be ergodic if for all invariant sets, either the set or its complement has measure 0.

Unlike the remaining concepts that we discuss in this article, this definition of ergodicity applies also to infinite measure-preserving transformations and even to certain non-measure-preserving transformations. See Aaronson's book [1] for more information.

The following lemma is often useful:

Lemma 1 *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $T\colon X \to X$ be a measure-preserving transformation. Then T is ergodic if and only if the only measurable functions f satisfying $f \circ T = f$ (up to sets of measure 0) are constant almost everywhere.*

For the straightforward proof, we notice that if the condition in the lemma holds and A is an invariant set, then $\mathbf{1}_A \circ T = \mathbf{1}_A$ almost everywhere, so that $\mathbf{1}_A$ is an a. e. constant function and so A or A^c is of measure 0. Conversely, if f is an invariant function, we see that for each α, $\{x\colon f(x) < \alpha\}$ is an invariant set and hence of measure 0

Copyrighted material

of the ergodic and quasi-ergodic hypotheses, which can be seen as demanding that μ_x is equal to μ for (almost) all x.

Proof The proof will be divided into 3 main steps: defining the measures μ_x, proving measurability with respect to x and proving ergodicity of the measures.

Step 1: Definition of μ_x

Given a function $f \in L^1(X, \mathcal{B}, \mu)$, Birkhoff's theorem states that for μ-almost every $x \in X$, $(A_N f(x))$ is convergent. It will be convenient to denote the limit by $\tilde{f}(x)$. Let $f_1, f_2, \ldots$ be a sequence of continuous functions that is dense in $C(X)$. For each k, there is a set B_k of x's measure 1 for which $(A_n f_k(x))_{n=1}^{\infty}$ is a convergent sequence. Intersecting these gives a set B of full measure such that for $x \in B$, for each $k \geq 1$, $A_n f_k(x)$ is convergent. A simple approximation argument shows that for $x \in B$ and f an arbitrary continuous function, $A_n f(x)$ is convergent. Given $x \in B$, define a map $L_x \colon C(X) \to \mathbb{R}$ by $L_x(f) = \tilde{f}(x)$. This is a continuous linear functional on $C(X)$, and hence by the Riesz Representation Theorem there exists a Borel measure μ_x such that $\tilde{f}(x) = \int f \, \mathrm{d}\mu_x$ for each $f \in C(X)$ and $x \in B$. Since $L_x(f) \geq 0$ when f is a non-negative function and $L_x(1) = 1$, the measure μ_x is a probability measure. Since $L_x(f \circ T) = L_x(f)$ for $f \in C(X)$, one can check that μ_x must be an invariant probability measure. For $x \notin B$, simply define $\mu_x = \mu$. Since B^c is a set of measure 0, this will not affect any of the statements that we are trying to prove.

Now for f continuous, we have $A_N f$ is a bounded sequence of functions with $A_N f(x)$ converging to $\int f \, \mathrm{d}\mu_x$ almost everywhere and $\int A_N f \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu$ since T is measure-preserving. It follows from the bounded convergence theorem that for $f \in C(X)$,

$$\int f \, \mathrm{d}\mu = \int \left(\int f \, \mathrm{d}\mu_x \right) \mathrm{d}\mu(x) \, . \tag{1}$$

Step 2: Measurability of $x \mapsto \mu_x(A)$

Lemma 7 *Let $C \in \mathcal{B}$ satisfy $\mu(C) = 0$. Then $\mu_x(C) = 0$ for μ-almost every $x \in X$.*

Proof Using regularity of Borel probability measures (see Rudin's book [75] for details), there exist open sets $U_1 \supset U_2 \supset \cdots \supset C$ with $\mu(U_k) < 1/k$. There exist continuous functions $g_{k,m}$ with $(g_{k,m}(x))_{m=1}^{\infty}$ increasing to $\mathbf{1}_{U_k}$ everywhere (e. g. $g_{k,m}(x) = \min(1, m \cdot d(x, U_k^c))$). By (1), we have $\int (\int g_{k,m} \, \mathrm{d}\mu_x) \, \mathrm{d}\mu(x) < 1/k$ for all k, m. Note that $\int g_{k,m} \, \mathrm{d}\mu_x = \lim_{n\to\infty} A_n g_{k,m}(x)$ is a measurable function of x, so that using the monotone convergence theorem (taking the limit in m), $x \mapsto \int \mathbf{1}_{U_k} \, \mathrm{d}\mu_x = \mu_x(U_k)$ is measurable and $\int (\int \mathbf{1}_{U_k} \, \mathrm{d}\mu_x) \, \mathrm{d}\mu(x) \leq 1/k$. We now see that $x \mapsto \lim_{k\to\infty} \mu_x(U_k) = \mu_x(\bigcap U_k)$ is also measurable, and by monotone convergence we see $\int \mu_x(\bigcap U_k) \, \mathrm{d}\mu(x) = 0$. It follows that $\mu_x(\bigcap U_k) = 0$ for μ-almost every x. Since $\bigcap U_k \supset C$, the lemma follows. □

Given a set $A \in \mathcal{B}$, let f_k be a sequence of continuous functions (uniformly bounded by 1) satisfying $\|f_k - \mathbf{1}_A\|_{L^1(\mu)} < 2^{-n}$, so that in particular $f_k(x) \to \mathbf{1}_A(x)$ for μ-almost every x. For each k, $x \mapsto \int f_k \, \mathrm{d}\mu_x = \lim_{n\to\infty} A_n f_k(x)$ is a measurable function. By Lemma 7, for μ-almost every x, $f_k \to 1_A$ μ_x-almost everywhere, so that by the bounded convergence theorem $\lim_{k\to\infty} \int f_k \, \mathrm{d}\mu_x = \mu_x(A)$ for μ-almost every x. Since the limit of measurable functions is measurable, it follows that $x \mapsto \mu_x(A)$ is measurable for any measurable set $A \in \mathcal{B}$.

This allows us to define a measure ν by $\nu(A) = \int \mu_x(A) \, \mathrm{d}\mu(x)$. For a bounded measurable function f, we have $\int f \, \mathrm{d}\nu = \int (\int f \, \mathrm{d}\mu_x) \, \mathrm{d}\mu(x)$. Since this agrees with $\int f \, \mathrm{d}\mu$ for continuous functions by (1), it follows that $\mu = \nu$. Conclusion (1) of the theorem now follows easily.

Given $f \in L^1(X)$, we let (f_k) be a sequence of continuous functions such that $\|f_k - f\|_{L^1(\mu)}$ is summable. This implies that $\|f_k - f\|_{L^1(\mu_x)}$ is summable for μ-almost every x and in particular, $\int f_k \, \mathrm{d}\mu_x \to \int f \, \mathrm{d}\mu_x$ for almost every x. On the other hand, by the remark following the statement of Birkhoff's theorem, we have $\tilde{f}_k = \mathbb{E}(f_k|\mathcal{I})$ so that $\|\tilde{f} - \tilde{f}_k\|_{L^1(\mu)}$ is summable and $\tilde{f}_k(x) \to \tilde{f}(x)$ for μ-almost every x. Combining these two statements, we see that for μ-almost every x, we have

$$\tilde{f}(x) = \lim_{k\to\infty} \tilde{f}_k(x) = \lim_{k\to\infty} \int f_k \, \mathrm{d}\mu_x = \int f \, \mathrm{d}\mu_x \, .$$

This establishes conclusion (2) of the theorem.

Step 3: Ergodicity of μ_x

We have shown how to disintegrate the invariant measure μ as an integral combination of μ_x's, and we have interpreted the μ_x's as describing the average behavior starting from x. It remains to show that the μ_x's are ergodic measures.

Fix for now a continuous function f and a number $0 < \epsilon < 1$. Since $A_n f(x) \to \tilde{f}(x)$ μ-almost everywhere, there exists an N such that $\mu\{x \colon |A_N f(x) - \tilde{f}(x)| > \epsilon/2\} < \epsilon^3/8$.

We now claim the following:

$$\mu\left\{x \colon \mu_x\{y \colon |\tilde{f}(y) - \textstyle\int f \, \mathrm{d}\mu_x| > \epsilon\} > \epsilon\right\} < \epsilon \, . \tag{2}$$

To see this, note that $\{y \colon |\tilde{f}(y) - \int f \, \mathrm{d}\mu_x| > \epsilon\} \subset \{y \colon |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} \cup \{y \colon |A_N f(y) - \tilde{f}(x)| > \epsilon/2\}$, so that if $\mu_x\{y \colon |\tilde{f}(y) - \int f \, \mathrm{d}\mu_x| > \epsilon\} > \epsilon$, then either

Copyrighted material

$\mu_x\{y\colon |\tilde{f}(y)-A_Nf(y)| > \epsilon/2\} > \epsilon/2$ or $\mu_x\{y\colon |A_Nf(y)-\tilde{f}(x)| > \epsilon/2\} > \epsilon/2$. We show that the set of x's satisfying each condition is small.

Firstly, we have $\epsilon^3/8 > \mu\{y\colon |\tilde{f}(y) - A_Nf(y)| > \epsilon/2\} = \int \mu_x\{y\colon |\tilde{f}(y) - A_Nf(y)| > \epsilon/2\}\, d\mu(x)$, so that $\mu\{x\colon \mu_x\{y\colon |\tilde{f}(y) - A_Nf(y)| > \epsilon/2\} > \epsilon/2\} < \epsilon^2/4 < \epsilon/2$.

For the second term, given $c \in \mathbb{R}$, let $F_c(x) = |A_Nf(x) - c|$ and $G(x) = F_{\tilde{f}(x)}(x)$. Note that

$$\int F_{\tilde{f}(x)}(y)\, d\mu_x(y) = \lim_{n\to\infty} A_n F_{\tilde{f}(x)}(x) = \lim_{n\to\infty} A_n G(x)$$

(using the facts that $y \mapsto F_{\tilde{f}(x)}(y)$ is a continuous function and that since $\tilde{f}(x)$ is an invariant function, $F_{\tilde{f}(x)}(T^k x) = G(T^k x)$). Since $\int G(x)\, d\mu(x) < \epsilon^3/8$, it follows that $\int F_{\tilde{f}(x)}(y)\, d\mu_x(y) \le \epsilon^2/4$ except on a set of x's of measure less than $\epsilon/2$. Outside this bad set, we have $\mu_x\{y\colon |A_Nf(y) - \tilde{f}(x)| > \epsilon/2\} < \epsilon/2$ so that $\mu\{x\colon \mu_x\{y\colon |A_Nf(y) - \tilde{f}(x)| > \epsilon/2\} > \epsilon/2\} < \epsilon/2$ as required.

This establishes our claim (2) above. Since $\epsilon > 0$ is arbitrary, it follows that for each $f \in C(X)$, for μ-almost every x, μ_x-almost every y satisfies $\tilde{f}(y) = \int f\, d\mu_x$. As usual, taking a countable dense sequence (f_k) in $C(X)$, it is the case that for all k and μ-almost every x, $\tilde{f}_k(y) = \int f_k\, d\mu_x$ μ_x-almost everywhere. Let the set of x's with this property be D. We claim that for $x \in D$, μ_x is ergodic. Suppose not. Then let $x \in D$ and let J be an invariant set of μ_x measure between δ and $1-\delta$ for some $\delta > 0$. Then by density of $C(X)$ in $L^1(\mu_x)$, there exists an f_k with $\|f_k - \mathbf{1}_J\|_{L^1(\mu_x)} < \delta$. Since $\mathbf{1}_J$ is an invariant function, we have $\tilde{\mathbf{1}}_J = \mathbf{1}_J$. On the other hand, $\tilde{f}_k$ is a constant function. It follows that $\|\tilde{f}_k - \tilde{\mathbf{1}}_J\|_{L^1(\mu_x)} \ge \delta > \|f_k - \mathbf{1}_J\|_{L_1(\mu_x)}$. This contradicts the identification of the limit as a conditional expectation and concludes the proof of the theorem. □

Mixing

As mentioned above, ergodicity may be seen as an independence on average property. More specifically, one wants to know whether in some sense $\mu(A \cap T^{-n}B)$ converges to $\mu(A)\mu(B)$ as $n \to \infty$. Ergodicity is the property that there is convergence in the Césaro sense. Weak-mixing is the property that there is convergence in the strong Césaro sense. That is, a measure-preserving transformation T is *weak-mixing* if

$$\frac{1}{N}\sum_{n=0}^{N-1} |\mu(A\cap T^{-n}B) - \mu(A)\mu(B)| \to 0 \quad \text{as} \quad N \to \infty.$$

In order for T to be *strong-mixing*, we require simply $\mu(A \cap T^{-N}B) \to \mu(A)\mu(B)$ as $N \to \infty$. It is clear that strong-mixing implies weak-mixing and weak-mixing implies ergodicity.

If T^d is not ergodic (so that $T^{-d}A = A$ for some A of measure strictly between 0 and 1), then $|\mu(T^{-nd}A \cap A) - \mu(A)^2| = \mu(A)(1-\mu(A))$, so that T is not weak-mixing.

An alternative characterization of weak-mixing is as follows:

Lemma 8 *The measure-preserving transformation T is weak-mixing if and only if for every pair of measurable sets A and B, there exists a subset J of $\mathbb{N}$ of density 1 (i. e. $\#(J \cap \{1,\dots,N\})/N \to 1$) such that*

$$\lim_{n\to\infty,\, n\notin J} \mu(A \cap T^{-n}B) = \mu(A)\mu(B)\,. \tag{3}$$

By taking a countable family of measurable sets that are dense (with respect to the metric $d(A,B) = \mu(A\Delta B)$) and taking a suitable intersection of the corresponding J sets, one shows that for a given weak-mixing measure-preserving transformation, there is a single set $J \subset \mathbb{N}$ such that (3) holds for *all* measurable sets A and B (see Petersen [64] or Walters [92] for a proof).

We show that an irrational rotation of the circle is not weak-mixing as follows: let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and let A be the interval $[\frac{1}{4}, \frac{3}{4})$. There is a positive proportion of n's in the natural numbers (in fact proportion 1/3) with the property that $|T^n(\frac{1}{2}) - \frac{1}{2}| < \frac{1}{6}$. For these n's $\mu(A \cap T^{-n}A) > \frac{1}{3}$, so that in particular $|\mu(A \cap T^{-n}A) - \mu(A)\mu(A)| > \frac{1}{12}$. Clearly this precludes the required convergence to 0 in the definition of weak-mixing, so that an irrational rotation is ergodic but not weak-mixing. Since $R_\alpha^n = R_{n\alpha}$, the earlier argument shows that R_α^n is ergodic, so that R_α is totally ergodic.

On the other hand, we show that any Bernoulli shift is strong-mixing. To see this, let A and B be arbitrary measurable sets. By standard measure-theoretic arguments, A and B may each be approximated arbitrarily closely by a finite union of cylinder sets. Since if A' and B' are finite unions of cylinder sets, we have that $\mu(A' \cap T^{-n}B')$ is *equal* to $\mu(A')\mu(B')$ for large n, it is easy to deduce that $\mu(A \cap T^{-n}B) \to \mu(A)\mu(B)$ as required. Since the doubling map is measure-theoretically isomorphic to a one-sided Bernoulli shift, it follows that the doubling map is also strong-mixing.

Similarly, if a Markov Chain is irreducible (i. e. for any states i and j, there exists an $n \ge 0$ such that $P_{ij}^n > 0$) and aperiodic (there is a state i such that $\gcd\{n\colon P_{ii}^n > 0\} = 1$), then given any pair of cylinder sets A' and B' we have by standard theorems of Markov chains $\mu(A' \cap T^{-n}B') \to$

Copyrighted material

$\mu(A')\mu(B')$. The same argument as above then shows that an aperiodic irreducible Markov Chain is strong-mixing. On the other hand, if a Markov chain is periodic ($d = \gcd\{n\colon P_{ii}^n > 0\} > 0$), then letting $A = B = \{x\colon x_0 = i\}$, we have that $\mu(A \cap T^{-n}B) = 0$ whenever $d \nmid n$. It follows T^d is not ergodic, so that T is not weak-mixing.

Both weak- and strong-mixing have formulations in terms of functions:

Lemma 9 *Let T be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$.*

1. *T is weak-mixing if and only if for every $f, g \in L^2$ one has*

$$\frac{1}{N}\sum_{n=0}^{N-1} |\langle f, g \circ T^n\rangle - \langle f, 1\rangle\langle 1, g\rangle| \to 0 \text{ as } N \to \infty\,.$$

2. *T is strong-mixing if and only if for every $f, g \in L^2$, one has*

$$\langle f, g \circ T^N\rangle \to \langle f, 1\rangle\langle 1, g\rangle \text{ as } N \to \infty\,.$$

Using this, one can see that both mixing conditions are spectral properties.

Lemma 10 *Weak- and strong-mixing are spectral properties.*

Proof Suppose S is a weak-mixing transformation of $(Y, \mathcal{F}, \nu)$ and the transformation T of $(X, \mathcal{B}, \mu)$ is spectrally isomorphic to S by the Hilbert space isomorphism Θ. Then for $f, g \in L^2(X, \mathcal{B}, \mu)$, $\langle f, g \circ T^n\rangle_X - \langle f, 1\rangle_X\langle 1, g\rangle_X = \langle \Theta(f), \Theta(g) \circ S^n\rangle_Y - \langle \Theta(f), \Theta(1)\rangle_Y\langle \Theta(1), \Theta(g)\rangle_Y$. Since 1 is an eigenfunction of U_T with eigenvalue 1, $\Theta(1)$ is an eigenfunction of U_S with an eigenvalue 1, so since S is ergodic, $\Theta(1)$ must be a constant function. Since Θ preserves norms, $\Theta(1)$ must have a constant value of absolute value 1 and hence $\langle f, g \circ T^n\rangle_X - \langle f, 1\rangle_X\langle 1, g\rangle_X = \langle \Theta(f), \Theta(g) \circ S^n\rangle_Y - \langle \Theta(f), 1\rangle_Y\langle 1, \Theta(g)\rangle_Y$. It follows from Lemma 9 that T is weak-mixing.

A similar proof shows that strong-mixing is a spectral property. □

Both weak- and strong-mixing properties are preserved by taking natural extensions.

Recent work of Avila and Forni [4] shows that for interval exchange transformations of $k \geq 3$ intervals with the underlying permutation satisfying the non-degeneracy condition above, almost all divisions of the interval (with respect to Lebesgue measure on the $k-1$-dimensional simplex) lead to weak-mixing transformations. On the other hand, work of Katok [36] shows that no interval exchange transformation is strong-mixing.

It is of interest to understand the behavior of the 'typical' measure-preserving transformation. There are a number of Baire category results addressing this. In order to state them, one needs a set of measure-preserving transformations and a topology on them. As mentioned earlier, it is effectively no restriction to assume that a transformation is a Lebesgue-measurable map on the unit interval preserving Lebesgue measure. The classical category results are then on the collection of invertible Lebesgue-measure preserving transformations of the unit interval. One topology on these is the 'weak' topology, where a sub-base is given by sets of the form $N(T, A, \epsilon) = \{S\colon \lambda(S(A)\Delta T(A)) < \epsilon\}$. With respect to this topology, Halmos [26] showed that a residual set (i. e. a dense G_δ set) of invertible measure-preserving transformations is weak-mixing (see also work of Alpern [3]), while Rokhlin [72] showed that the set of strong-mixing transformations is meagre (i. e. a nowhere dense F_σ set), allowing one to conclude that with respect to this topology, the typical transformation is weak- but not strong-mixing.

As often happens in these cases, even when a certain kind of behavior is typical, it may not be simple to exhibit concrete examples. In this case, a well-known example of a transformation that is weak-mixing but not strong-mixing was given by Chacon [13].

While on the face of it the formulation of weak-mixing is considerably less natural than that of strong-mixing, the notion of weak-mixing turns out to be extremely natural from a spectral point of view. Given a measure-preserving transformation T, let U_T be the Koopman operator described above. Since this operator is an isometry, any eigenvalue must lie on the unit circle. The constant function 1 is always an eigenfunction with eigenvalue 1. If T is ergodic and g and h are eigenfunctions of U_T with eigenvalue λ, then $g\bar{h}$ is an eigenfunction with eigenvalue 1, hence invariant, so that $g = Kh$ for some constant K. We see that for ergodic transformations, up to rescaling, there is at most one eigenfunction with any given eigenvalue.

If U_T has a non-constant eigenfunction f, then one has $|\langle U_T^n f, f\rangle| = \|f\|^2$ for each n, whereas by Cauchy–Schwartz, $|\langle f, 1\rangle|^2 < \|f\|^2$. It follows that $|\langle U_T^n f, f\rangle - \langle f, 1\rangle\langle 1, f\rangle| \geq c$ for some positive constant c, so that using Lemma 9, T is not weak-mixing.

Using the spectral theorem, the converse is shown to hold.

Theorem 11 *The measure-preserving transformation T is weak-mixing if and only U_T has no non-constant eigenfunctions.*

Copyrighted material

Of course this also shows that weak-mixing is a spectral property. Equivalently, this says that the transformation T is weak-mixing if and only if the apart from the constant eigenfunction, the operator U_T has only continuous spectrum (that is, the operator has no other eigenfunctions). For a very nice and concise development of the part of spectral theory relevant to ergodic theory, the reader is referred to the Appendix in Parry's book [63].

Using this theory, one can establish the following:

Theorem 12

1. *T is weak-mixing if and only if $T \times T$ is ergodic;*
2. *If T and S are ergodic, then $T \times S$ is ergodic if and only if U_S and U_T have no common eigenvalues other than 1.*

Proof The main factor in the proof is that the eigenvalues of $U_{T\times S}$ are precisely the set of $\alpha\beta$, where α is an eigenvalue of U_T and β is an eigenvalue of U_S. Further, the eigenfunctions of $U_{T\times S}$ with eigenvalue γ are spanned by eigenfunctions of the form $f \otimes g$, where f is an eigenfunction of U_T, g is an eigenfunction of U_S, and the product of the eigenvalues is γ.

Suppose that T is weak-mixing. Then the only eigenfunction is the constant function, so that the only eigenfunction of $U_{T\times T}$ is the constant function, proving that $T \times T$ is ergodic. Conversely, if U_T has an eigenvalue (so that $f \circ T = \alpha T$ for some non-constant f) then $f \otimes \bar{f}$ is a non-constant invariant function of $T \times T$ so that $T \times T$ is not ergodic.

For the second part, if U_S and U_T have a common eigenvalue other than 1 (say $f \circ T = \alpha f$ and $g \circ T = \alpha g$), then $f \otimes \bar{g}$ is a non-constant invariant function. Conversely, if $T \times S$ has a non-constant invariant function h, then h can be decomposed into functions of the form $f \otimes g$, where f and g are eigenfunctions of U_T and U_S respectively with eigenvalues α and β satisfying $\alpha\beta = 1$. Since the eigenvalues of S are closed under complex conjugation, we see that U_T and U_S have a common eigenvalue other than 1 as required. □

For a measure-preserving transformation T, we let K be the subspace of L^2 spanned by the eigenfunctions of U_T. It is a remarkable fact that K may be identified as $L^2(X, \mathcal{B}', \mu)$ where $\mathcal{B}'$ is a sub-σ-algebra of $\mathcal{B}$. The space K is called the *Kronecker factor* of T. The terminology comes from the fact that any sub-σ-algebra $\mathcal{F}$ of $\mathcal{B}$ gives rise to a factor mapping $\pi : (X, \mathcal{B}, \mu) \to (X, \mathcal{F}, \mu)$ with $\pi(x) = x$. By construction $L^2(X, \mathcal{B}', \mu)$ is the closed linear span of the eigenfunctions of T considered as a measure-preserving transformation of $(X, \mathcal{B}', \mu)$. By the Discrete Spectrum Theorem of Halmos and von Neumann [27], T acting on $(X, \mathcal{B}', \mu)$ is measure-theoretically isomorphic to a rotation on a compact group. This allows one to split $L^2(X, \mathcal{B}, \mu)$ as $L^2(X, \mathcal{B}', \mu) \oplus L^2_c(X, \mathcal{B}, \mu)$, where, as mentioned above the first part is the discrete spectrum part, spanned by eigenfunctions, and the second part is the continuous spectrum part, consisting of functions whose spectral measure is continuous. Since we have split L^2 into a discrete part and a continuous part, it is natural to ask whether the underlying transformation T can be split up in some way into a weak-mixing part and a discrete spectrum (compact group rotation) part, somewhat analogously to the ergodic decomposition. Unfortunately, there is no-such decomposition available. However for some applications, for example to multiple recurrence (starting with the work of Furstenberg [20,21]), the decomposition of L^2 (possibly into more complicated parts) plays a crucial role (see the chapters on ► Ergodic Theory: Recurrence and ► Ergodic Theory: Interactions with Combinatorics and Number Theory).

For non-invertible measure-preserving transformations, the transformation is weak- or strong-mixing if and only if its natural extension has that property.

The understanding of weak-mixing in terms of the discrete part of the spectrum of the operator also extends to total ergodicity. T^n is ergodic if and only if T has no eigenvalues of the form $e^{2\pi ip/n}$ other than 1. From this it follows that an ergodic measure-preserving transformation T is totally ergodic if and only if it has no *rational spectrum* (i. e. no eigenvalues of the form $e^{2\pi ip/q}$ other than the simple eigenvalue 1).

An intermediate mixing condition between strong- and weak- mixing is that a measure-preserving transformation is *mild-mixing* if whenever $f \circ T^{n_i} \to f$ for an L^2 function f and a sequence $n_i \to \infty$, then f is a.e. constant. Clearly mild-mixing is a spectral property. If a transformation has an eigenfunction f, then it is straightforward to find a sequence n_i such that $f \circ T^{n_i} \to f$, so we see that mild-mixing implies weak-mixing. To see that strong-mixing implies mild-mixing, suppose that T is strong-mixing and that $f \circ T^{n_i} \to f$. Then we have $\int f \circ T^{n_i} \bar{f} \to \|f\|^2$. On the other hand, the strong mixing property implies that $\int f \circ T^{n_i} \bar{f} \to |\langle f, 1\rangle|^2$. The equality of these implies that f is a. e. constant. Mild-mixing has a useful reformulation in terms of ergodicity of general (not necessarily probability) measure-preserving transformations: A transformation T is mild-mixing if and only if for every conservative ergodic measure-preserving transformation S, $T \times S$ is ergodic. See Furstenberg and Weiss' article [22] for further information on mild-mixing.

The strongest spectral property that we consider is that of having countable Lebesgue spectrum. While we

Copyrighted material

will avoid a detailed discussion of spectral theory in this article, this is a special case that can be described simply. Specifically, let T be an invertible measure-preserving transformation. Then T has countable Lebesgue spectrum if there is a sequence of functions $f_1, f_2, \ldots$ such that $\{1\} \cup \{U_T^n f_j : n \in \mathbb{Z}, j \in \mathbb{N}\}$ forms an orthonormal basis for $L^2(X)$.

To see that this property is stronger than strong-mixing, we simply observe that it implies that $\langle U_T^t U_T^n f_j, U_T^m f_k\rangle \to 0$ as $t \to \infty$. Then by approximating f and g by their expansions with respect to a finite part of the basis, we deduce that $\langle U_T^n f, g\rangle \to \langle f, 1\rangle\langle 1, g\rangle$ as required. Since already strong-mixing is atypical from the topological point of view, it follows that countable Lebesgue spectrum has to be atypical. In fact, Yuzvinskii [96] showed that the typical invertible measure-preserving transformation has simple singular spectrum.

The property of countable Lebesgue spectrum is by definition a spectral property. Since it completely describes the transformation up to spectral isomorphism, there can be no stronger spectral properties. The remaining properties that we shall examine are invariant under measure-theoretic isomorphisms only.

An invertible measure-preserving transformation T of $(X, \mathcal{B}, \mu)$ is said to be K (for Kolmogorov) if there is a sub-σ-algebra $\mathcal{F}$ of $\mathcal{B}$ such that

1. $\bigcap_{n=1}^{\infty} T^{-n}\mathcal{F}$ is the trivial σ-algebra up to sets of measure 0 (i. e. the intersection consists only of null sets and sets of full measure).
2. $\bigvee_{n=1}^{\infty} T^n\mathcal{F} = \mathcal{B}$ (i. e. the smallest σ-algebra containing $T^n\mathcal{F}$ for all $n > 0$ is $\mathcal{B}$).

The K property has a useful reformulation in terms of entropy as follows: T is K if and only if for every non-trivial partition $\mathcal{P}$ of X, the entropy of T with respect to the partition $\mathcal{P}$ is positive: T has *completely positive entropy*. See the chapter on ▶ Entropy in Ergodic Theory for the relevant definitions. The equivalence of the K property and completely positive entropy was shown by Rokhlin and Sinai [74]. For a general transformation T, one can consider the collection of all subsets B of X such that with respect to the partition $\mathcal{P}_B = \{B, B^c\}$, $h(\mathcal{P}_B) = 0$. One can show that this is a σ-algebra. This σ-algebra is known as the *Pinsker* σ-algebra. The above reformulation allows us to say that a transformation is K if and only if it has a trivial Pinsker σ-*algebra*.

The K property implies countable Lebesgue spectrum (see Parry's book [63] for a proof). To see that K is not implied by countable Lebesgue spectrum, we point out that certain measure-preserving transformations derived from Gaussian systems (see for example the paper of Parry and Newton [52]) have countable Lebesgue spectrum but zero entropy.

The fact that (two-sided) Bernoulli shifts have the K property follows from Kolmogorov's 0–1 law by taking $\mathcal{F} = \bigvee_{n=0}^{\infty} T^{-n}\mathcal{P}$, where $\mathcal{P}$ is the partition into cylinder sets (see Williams's book [93] for details of the 0–1 law).

Although the K property is explicitly an invertible property, it has a non-invertible counterpart, namely exactness. A transformation T of $(X, \mathcal{B}, \mu)$ is *exact* if $\bigcap_{n=0}^{\infty} T^{-n}\mathcal{B}$ consists entirely of null sets and sets of measure 1. It is not hard to see that a non-invertible transformation is exact if and only if its natural extension is K.

The final and strongest property in our list is that of being measure-theoretically isomorphic to a Bernoulli shift. If T is measure-theoretically isomorphic to a Bernoulli shift, we say that T has the *Bernoulli property*. While in principle this could apply to both invertible and non-invertible transformations, in practice the definition applies to a large class of invertible transformations, but occurs comparatively seldom for non-invertible transformations. For this reason, we will restrict ourselves to a discussion of the Bernoulli property for invertible transformations (see however work of Hoffman and Rudolph [29] and Heicklen and Hoffman [28] for work on the one-sided Bernoulli property).

In the case of invertible Bernoulli shifts, Ornstein [53,58] developed in the early 1970s a powerful isomorphism theory, showing that two Bernoulli shifts are measure-theoretically isomorphic if and only if they have the same entropy. Entropy had already been identified as an invariant by Kolmogorov and Sinai [43,82], so this established that it was a complete invariant for Bernoulli shifts. Keane and Smorodinsky [41] gave a proof which showed that two Bernoulli shifts of the same entropy are isomorphic using a conjugating map that is continuous almost everywhere. With other authors, this theory was extended to show that the property of being isomorphic to a Bernoulli shift applied to a surprisingly large class of measure-preserving transformations (e. g. geodesic flows on manifolds of constant negative curvature (Ornstein and Weiss [60]), aperiodic irreducible Markov chains (Friedman and Ornstein [19]), toral automorphisms (Katznelson [39]) and more generally many Gibbs measures for hyperbolic dynamical systems (see the book of Bowen [11])).

Initially, it was conjectured that the properties of being K and Bernoulli were the same, but since then a number of measure-preserving transformations that are K but not Bernoulli have been identified. The earliest was due to Ornstein [55]. Ornstein and Shields [59] then provided an uncountable family of non-isomorphic K automorphisms.

Copyrighted material

In a recent paper of McCutcheon and Quas [50], the analogous question was addressed in the case where T is a mild-mixing transformation.

Problem 3 (Pascal adic transformation) Vershik [89,90] introduced a family of transformations known as the adic transformations. The underlying spaces for these transformations are certain spaces of paths on infinite graphs, and the transformations act by taking a path to its lexicographic neighbor. Amongst the adic transformations, the so-called Pascal adic transformation (so-called because the underlying graph resembles Pascal's triangle) has been singled out for attention in work of Petersen and others [2,10,51,65]. In particular, it is unresolved whether this transformation is weak-mixing with respect to any of its ergodic measures. Weak-mixing has been shown by Petersen and Schmidt to follow from a number-theoretic condition on the binomial coefficients [2,10].

Problem 4 (Weak Pinsker Conjecture) Pinsker [67] conjectured that in a measure-preserving transformation with positive entropy, one could express the transformation as a product of a Bernoulli shift with a system with zero entropy. This conjecture (now known as the *Strong Pinsker Conjecture*) was shown to be false by Ornstein [56,57]. Shields and Thouvenot [78] showed that the collection of transformations that can be written as a product of a zero entropy transformation with a Bernoulli shift is closed in the so-called $\bar{d}$-metric that lies at the heart of Ornstein's theory.

It is, however, the case that if $T\colon X \to X$ has entropy $h > 0$, then for all $h' \leq h$, T has a factor S with entropy h' (this was originally proved by Sinai [83] and reproved using the Ornstein machinery by Ornstein and Weiss in [61]). The *Weak Pinsker Conjecture* states that if a measure-preserving transformation T has entropy $h > 0$, then for all $\epsilon > 0$, T may be expressed as a product of a Bernoulli shift and a measure-preserving transformation with entropy less than ϵ.

Bibliography

1. Aaronson J (1997) An Introduction to Infinite Ergodic Theory. American Mathematical Society, Providence
2. Adams TM, Petersen K (1998) Binomial coefficient multiples of irrationals. Monatsh Math 125:269–278
3. Alpern S (1976) New proofs that weak mixing is generic. Invent Math 32:263–278
4. Avila A, Forni G (2007) Weak-mixing for interval exchange transformations and translation flows. Ann Math 165:637–664
5. Bergelson V (1987) Weakly mixing pet. Ergodic Theory Dynam Systems 7:337–349
6. Birkhoff GD (1931) Proof of the ergodic theorem. Proc Nat Acad Sci 17:656–660
7. Birkhoff GD, Smith PA (1924) Structural analysis of surface transformations. J Math 7:345–379
8. Boltzmann L (1871) Einige allgemeine Sätze über Wärmegleichgewicht. Wiener Berichte 63:679–711
9. Boltzmann L (1909) Wissenschaftliche Abhandlungen. Akademie der Wissenschaften, Berlin
10. Boshernitzan M, Berend D, Kolesnik G (2001) Irrational dilations of Pascal's triangle. Mathematika 48:159–168
11. Bowen R (1975) Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms. Springer, Berlin
12. Bradley RC (2005) Basic properties of strong mixing conditions. A survey and some open questions. Probab Surv 2:107–144
13. Chacon RV (1969) Weakly mixing transformations which are not strongly mixing. Proc Amer Math Soc 22:559–562
14. Choquet G (1956) Existence des représentations intégrales au moyen des points extrémaux dans les cônes convexes. C R Acad Sci Paris 243:699–702
15. Choquet G (1956) Unicité des représentations intégrales au moyen de points extrémaux dans les cônes convexes réticulés. C R Acad Sci Paris 243:555–557
16. de la Rue T (2006) 2-fold and 3-fold mixing: why 3-dot-type counterexamples are impossible in one dimension. Bull Braz Math Soc (NS) 37(4):503–521
17. Ehrenfest P, Ehrenfest T (1911) Begriffliche Grundlage der statistischen Auffassung in der Mechanik. No. 4. In: Encyclopädie der mathematischen Wissenschaften. Teubner, Leipzig
18. Feller W (1950) An Introduction to Probability and its Applications. Wiley, New York
19. Friedman NA, Ornstein DS (1970) On isomorphism of weak Bernoulli transformations. Adv Math 5:365–394
20. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. J Analyse Math 31:204–256
21. Furstenberg H (1981) Recurrence in Ergodic Theory and Combinatorial Number Theory. Princeton University Press, Princeton
22. Furstenberg H, Weiss B (1978) The finite multipliers of infinite ergodic transformations. In: The Structure of Attractors in Dynamical Systems. Proc Conf, North Dakota State Univ, Fargo, N.D., 1977. Springer, Berlin
23. Garsia AM (1965) A simple proof of E. Hopf's maximal ergodic theory. J Math Mech 14:381–382
24. Girsanov IV (1958) Spectra of dynamical systems generated by stationary Gaussian processes. Dokl Akad Nauk SSSR 119:851–853
25. Góra P (1994) Properties of invariant measures for piecewise expanding one-dimensional transformations with summable oscillations of derivative. Ergodic Theory Dynam Syst 14:475–492
26. Halmos PA (1944) In general a measure-preserving transformation is mixing. Ann Math 45:786–792
27. Halmos P (1956) Lectures on Ergodic Theory. Chelsea, New York
28. Hoffman C, Heicklen D (2002) Rational maps are *d*-adic bernoulli. Ann Math 156:103–114
29. Hoffman C, Rudolph DJ (2002) Uniform endomorphisms which are isomorphic to a Bernoulli shift. Ann Math 76:79–101
30. Host B (1991) Mixing of all orders and pairwise independent joinings of systems with singular spectrum. Israel J Math 76:289–298

Copyrighted material

31. Hu H (2004) Decay of correlations for piecewise smooth maps with indifferent fixed points. Ergodic Theory Dynam Syst 24:495–524
32. Kalikow S () Outline of ergodic theory. Notes freely available for download. See http://www.math.uvic.ca/faculty/aquas/kalikow/kalikow.html
33. Kalikow S (1982) T, T^{-1} transformation is not loosely Bernoulli. Ann Math 115:393–409
34. Kalikow S (1984) Twofold mixing implies threefold mixing for rank one transformations. Ergodic Theory Dynam Syst 2:237–259
35. Kamae T (1982) A simple proof of the ergodic theorem using non-standard analysis. Israel J Math 42:284–290
36. Katok A (1980) Interval exchange transformations and some special flows are not mixing. Israel J Math 35:301–310
37. Katok A (1980) Smooth non-Bernoulli K-automorphisms. Invent Math 61:291–299
38. Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge, Cambridge
39. Katznelson Y (1971) Ergodic automorphisms of $\mathbb{T}^n$ are Bernoulli shifts. Israel J Math 10:186–195
40. Katznelson Y, Weiss B (1982) A simple proof of some ergodic theorems. Israel J Math 42:291–296
41. Keane M, Smorodinsky M (1979) Bernoulli schemes of the same entropy are finitarily isomorphic. Ann Math 109:397–406
42. Keane MS, Petersen KE (2006) Nearly simultaneous proofs of the ergodic theorem and maximal ergodic theorem. In: Dynamics and Stochastics: Festschrift in Honor of M.S. Keane. Institute of Mathematical Statistics, pp 248–251, Bethesda MD
43. Kolmogorov AN (1958) New metric invariant of transitive dynamical systems and endomorphisms of Lebesgue spaces. Dokl Russ Acad Sci 119:861–864
44. Koopman BO (1931) Hamiltonian systems and Hilbert space. Proc Nat Acad Sci 17:315–218
45. Krámli A, Simányi N, Szász D (1991) The *K*-property of three billiard balls. Ann Math 133:37–72
46. Ledrappier F (1978) Un champ Markovien peut être d'entropie nulle et mélangeant. C R Acad Sci Paris, Sér A-B 287:561–563
47. Liverani C (2004) Decay of correlations. Ann Math 159:1275–1312
48. Masser DW (2004) Mixing and linear equations over groups in positive characteristic. Israel J Math 142:189–204
49. Masur H (1982) Interval exchange transformations and measured foliations. Ann Math 115:169–200
50. McCutcheon R, Quas A (2007) Generalized polynomials and mild mixing systems. Canad J Math (to appear)
51. Méla X, Petersen K (2005) Dynamical properties of the Pascal adic transformation. Ergodic Theory Dynam Syst 25:227–256
52. Newton D, Parry W (1966) On a factor automorphism of a normal dynamical system. Ann Math Statist 37:1528–1533
53. Ornstein DS (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
54. Ornstein DS (1972) On the root problem in ergodic theory. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), vol II: Probability theory. Univ. California Press, pp 347–356
55. Ornstein DS (1973) An example of a Kolmogorov automorphism that is not a Bernoulli shift. Adv Math 10:49–62
56. Ornstein DS (1973) A K-automorphism with no square root and Pinsker's conjecture. Adv Math 10:89–102
57. Ornstein DS (1973) A mixing transformation for which Pinsker's conjecture fails. Adv Math 10:103–123
58. Ornstein DS (1974) Ergodic Theory, Randomness, and Dynamical Systems. Yale University Press, Newhaven
59. Ornstein DS, Shields PC (1973) An uncountable family of K-automorphisms. Adv Math 10:89–102
60. Ornstein DS, Weiss B (1973) Geodesic flows are Bernoullian. Israel J Math 14:184–198
61. Ornstein DS, Weiss B (1975) Unilateral codings of Bernoulli systems. Israel J Math 21:159–166
62. Oxtoby JC (1952) Ergodic sets. Bull Amer Math Soc 58:116–136
63. Parry W (1981) Topics in Ergodic Theory. Cambridge, Cambridge
64. Petersen K (1983) Ergodic Theory. Cambridge, Cambridge
65. Petersen K, Schmidt K (1997) Symmetric Gibbs measures. Trans Amer Math Soc 349:2775–2811
66. Phelps R (1966) Lectures on Choquet's Theorem. Van Nostrand, New York
67. Pinsker MS (1960) Dynamical systems with completely positive or zero entropy. Soviet Math Dokl 1:937–938
68. Quas A (1996) A C^1 expanding map of the circle which is not weak-mixing. Israel J Math 93:359–372
69. Quas A (1996) Non-ergodicity for C^1 expanding maps and *g*-measures. Ergodic Theory Dynam Systems 16:531–543
70. Rényi A (1957) Representations for real numbers and their ergodic properties. Acta Math Acad Sci Hungar 8:477–493
71. Riesz F (1938) Some mean ergodic theorems. J Lond Math Soc 13:274–278
72. Rokhlin VA (1948) A 'general' measure-preserving transformation is not mixing. Dokl Akad Nauk SSSR Ser Mat 60:349–351
73. Rokhlin VA (1949) On endomorphisms of compact commutative groups. Izvestiya Akad Nauk SSSR Ser Mat 13:329–340
74. Rokhlin VA, Sinai Y (1961) Construction and properties of invariant measurable partitions. Dokl Akad Nauk SSSR 141:1038–1041
75. Rudin W (1966) Real and Complex Analysis. McGraw Hill, New York
76. Rudolph DJ (1990) Fundamentals of Measurable Dynamics. Oxford, Oxford University Press
77. Ryzhikov VV (1993) Joinings and multiple mixing of the actions of finite rank. Funct Anal Appl 27:128–140
78. Shields P, Thouvenot J-P (1975) Entropy zero × Bernoulli processes are closed in the $\bar{d}$-metric. Ann Probab 3:732–736
79. Simányi N (2003) Proof of the Boltzmann–Sinai ergodic hypothesis for typical hard disk systems. Invent Math 154:123–178
80. Simányi N (2004) Proof of the ergodic hypothesis for typical hard ball systems. Ann Henri Poincaré 5:203–233
81. Simányi N, Szász D (1999) Hard ball systems are completely hyperbolic. Ann Math 149:35–96
82. Sinai YG (1959) On the notion of entropy of a dynamical system. Dokl Russ Acad Sci 124:768–771
83. Sinai YG (1964) On a weak isomorphism of transformations with invariant measure. Mat Sb (NS) 63:23–42
84. Sinai YG (1970) Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. Uspehi Mat Nauk 25:141–192
85. Sinai Y (1976) Introduction to Ergodic Theory. Princeton, Princeton, translation of the 1973 Russian original

Copyrighted material

86. Sinai YG, Chernov NI (1987) Ergodic properties of some systems of two-dimensional disks and three-dimensional balls. Uspekhi Mat Nauk 42:153–174, 256
87. Smorodinsky M (1971) A partition on a Bernoulli shift which is not weakly Bernoulli. Math Systems Th 5:201–203
88. Veech W (1982) Gauss measures for transformations on the space on interval exchange maps. Ann Math 115:201–242
89. Vershik A (1974) A description of invariant measures for actions of certain infinite-dimensional groups. Soviet Math Dokl 15:1396–1400
90. Vershik A (1981) Uniform algebraic approximation of shift and multiplication operators. Soviet Math Dokl 24:97–100
91. von Neumann J (1932) Proof of the quasi-ergodic hypothesis. Proc Nat Acad Sci USA 18:70–82
92. Walters P (1982) An Introduction to Ergodic Theory. Springer, Berlin
93. Williams D (1991) Probability with Martingales. Cambridge, Cambridge
94. Young LS (1995) Ergodic theory of differentiable dynamical systems. In: Real and Complex Dynamical Systems (Hillerød, 1993). Kluwer, Dordrecht, pp 293–336
95. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. Ann Math 147:585–650
96. Yuzvinskii SA (1967) Metric automorphisms with a simple spectrum. Soviet Math Dokl 8:243–245

Copyrighted material

Ergodic Theorems

ANDRÉS DEL JUNCO
Department of Mathematics,
University of Toronto, Toronto, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Ergodic Theorems for Measure-Preserving Maps
Generalizations to Continuous Time
and Higher–Dimensional Time
Pointwise Ergodic Theorems for Operators
Subadditive and Multiplicative Ergodic Theorems
Entropy and the Shannon–McMillan–Breiman Theorem
Amenable Groups
Subsequence and Weighted Theorems
Ergodic Theorems and Multiple Recurrence
Rates of Convergence
Ergodic Theorems for Non-amenable Groups
Future Directions
Bibliography

Glossary

Dynamical system in its broadest sense, any set X, with a map $T\colon X \to X$. The classical example is: X is a set whose points are the states of some physical system and the state x is succeeded by the state Tx after one unit of time.

Iteration repeated applications of the map T above to arrive at the state of the system after n units of time.

Orbit of x the forward images $x, Tx, T^2X \ldots$ of $x \in X$ under iteration of T. When T is invertible one may consider the forward, backward or two-sided orbit of x.

Automorphism a dynamical system $T\colon X \to X$, where X is a measure space and T is an invertible map preserving measure.

Ergodic average if f is a function on X let $A_n f(x) = n^{-1}\sum_{i=0}^{n-1} f(T^i x)$; the average of the values of f over the first n points in the orbit of x.

Ergodic theorem an assertion that ergodic averages converge in some sense.

Mean ergodic theorem an assertion that ergodic averages converge with respect to some norm on a space of functions.

Pointwise ergodic theorem an assertion that ergodic averages $A_n f(x)$ converge for some or all $x \in X$, usually for a.e. x.

Stationary process a sequence $(X_1, X_2, \ldots)$ of random variables (real or complex-valued measurable functions) on a probability space whose joint distributions are invariant under shifting $(X_1, X_2, \ldots)$ to $(X_2, X_3, \ldots)$.

Uniform distribution a sequence $\{x_n\}$ in $[0,1]$ is uniformly distributed if for each interval $I \subset [0,1]$, the time it spends in I is asymptotically proportional to the length of I.

Maximal inequality an inequality which allows one to bound the pointwise oscillation of a sequence of functions. An essential tool for proving pointwise ergodic theorems.

Operator any linear operator U on a vector space of functions on X, for example one arising from a dynamical system T by setting $Uf(x) = f(Tx)$. More generally any linear transformation on a real or complex vector space.

Positive contraction an operator T on a space of functions endowed with a norm $\|\cdot\|$ such that T maps positive functions to positive functions and $\|Tf\| \le |f|$.

Definition of the Subject

Ergodic theorems are assertions about the long-term statistical behavior of a dynamical system. The subject arose out of Boltzmann's ergodic hypothesis which sought to equate the spatial average of a function over the set of states in a physical system having a fixed energy with the time average of the function observed by starting with a particular state and following its evolution over a long time period.

Introduction

Suppose that $(X, \mathcal{B}, \mu)$ is a measure space and $T\colon (X, \mathcal{B}, \mu) \to (X, \mathcal{B}, \mu)$ is a measurable and measure-preserving transformation, that is $\mu(T^{-1}E) = \mu(E)$ for all $E \in \mathcal{B}$. One important motivation for studying such maps is that a Hamiltonian physical system (see the article by Petersen in this collection) gives rise to a one-parameter group $\{T_t : t \in \mathbb{R}\}$ of maps in the phase space of the system which preserve Lebesgue measure. The ergodic theorem of Birkhoff asserts that for $f \in L_1(\mu)$

$$\frac{1}{n}\sum_{i=0}^{n-1} f(T^i x) \tag{1}$$

converges a.e. and that if T is *ergodic* (to be defined shortly) then the limit is $\int f \, d\mu$. This may be viewed

Copyrighted material

as a justification for Boltzmann's ergodic hypothesis that "space averages equal time averages". See Zund [214] for some history of the ergodic hypothesis. For physicists, then, the problem is reduced to showing that a given physical system is ergodic, which can be very difficult. However ergodic systems arise in many natural ways in mathematics. One example is rotation $z \mapsto \lambda z$ of the unit circle if λ is a complex number of modulus one which is not a root of unity. Another is the shift transformation on a sequence of i.i.d. random variables, for example a coin-tossing sequence. Another is an automorphism of a compact Abelian group. Often a transformation possesses an invariant measure which is not obvious at first sight. Knowledge of such a measure can be a very useful tool. See Petersen's article for more examples.

If $(X, \mathcal{B}, \mu)$ is a probability space and T is ergodic then Birkhoff's ergodic theorem implies that if A is a measurable subset of X then for almost every x, the frequency with which x visits A is asymptotically equal to $\mu(A)$, a very satisfying justification of intuition. For example, applying this to the coin-tossing sequence one obtains the *strong law of large numbers* which asserts that almost every infinite sequence of coin tosses has tails occurring with asymptotic frequency $\frac{1}{2}$. One also obtains Borel's theorem on normal numbers which asserts that for almost all $x \in [0,1]$ each digit $0, 1, 2, \ldots, 9$ occurs with limiting frequency $\frac{1}{10}$. The so-called *continued fraction transformation* $x \mapsto x^{-1} \mod 1$ on $(0,1)$ has a finite invariant measure $\frac{dx}{1+x}$. (Throughout this article $x \mod 1$ denotes the fractional part of x.) Applying Birkhoff's theorem then gives precise information about the frequency of occurrence of any $n \in \mathbb{N}$ in the continued fraction expansion of x, for a.e. x. See for example Billingsley [34].

These are the classical roots of the subject of ergodic theorems. The subject has evolved from these simple origins into a vast field in its own right, quite independent of physics or probability theory. Nonetheless it still has close ties to both these areas and has also forged new links with many other areas of mathematics.

Our purpose here is to give a broad overview of the subject in a historical perspective. There are several excellent references, notably the books of Krengel [135] and Tempelman [194] which give a good picture of the state of the subject at the time they appeared. There has been tremendous progress since then. The time is ripe for a much more comprehensive survey of the field than is possible here.

Many topics are necessarily absent and many are only glimpsed. For example this article will not touch on random ergodic theorems. See the articles [75,143] for some references on this topic.

I thank Mustafa Akcoglu, Ulrich Krengel, Michael Lin, Dan Rudolph and particularly Joe Rosenblatt and Vitaly Bergelson for many helpful comments and suggestions. I would like to dedicate this article to Mustafa Akcoglu who has been such an important contributor to the development of ergodic theorems over the past 40 years. He has also played a vital role in my mathematical development as well as of many other mathematicians. He remains a source of inspiration to me, and a valued friend.

Ergodic Theorems for Measure-Preserving Maps

Suppose $(X, \mathcal{B}, \mu)$ is a measure space. A (measure-preserving) *endomorphism* of $(X, \mathcal{B}, \mu)$ is a measurable mapping $T: X \to X$ such that $\mu(T^{-1}E) = \mu(E)$ for any measurable subset $E \subset X$. If T has a measurable inverse then one says that it is an *automorphism*. In this article the unqualified terms "endomorphism" or "automorphism" will always mean measure-preserving. T is called *ergodic* if for all measurable E one has

$$T^{-1}E = E \Rightarrow \mu(E) = 0 \quad \text{or} \quad \mu(E^c) = 0\,.$$

A very general class of examples comes from the notion of a *stationary stochastic process* in probability theory. A stochastic process is a sequence of measurable functions $f_1, f_2, \ldots$ on a probability space $(X, \mathcal{B}, \mu)$ taking values in a measurable space (Y, C). The *distribution* of the process, a measure ν on $Y^{\mathbb{N}}$, is defined as the image of μ under the map $(f_1, f_2, \ldots) : X \to Y^{\mathbb{N}}$. ν captures all the essential information about the process $\{f_i\}$. In effect one may view any process $\{f_i\}$ as a probability measure on $Y^{\mathbb{N}}$. The process is said to be *stationary* if ν is invariant under the left shift transformation S on $Y^{\mathbb{N}}$, $S(y)(i) = y(i+1)$, that is S is an endomorphism of the probability space $(Y^{\mathbb{N}}, \nu)$.

From a probabilistic point of view the most natural examples of stationary stochastic processes are independent identically distributed processes, namely the case when ν is a product measure $\lambda^{\mathbb{N}}$ for some measure λ on (Y, C). More generally one can consider a stationary Markov process defined by transition probabilities on the state space Y and an invariant probability on Y. See for example Chap. 7 in [50], also Sect. "Pointwise Ergodic Theorems for Operators" below.

The first and most fundamental result about endomorphisms is the celebrated recurrence theorem of Poincaré [173]).

Theorem 1 *Suppose μ is finite, $A \in \mathcal{B}$ and $\mu(A) > 0$. Then for a.e. $x \in A$ there is an $n > 0$ such that $T^n x \in A$, in fact there are infinitely many such n.*

It may be viewed as an ergodic theorem, in that it is a qualitative statement about how x behaves under iteration of T.

Copyrighted material

$C(\lambda) \to 0$ as $\lambda \to \infty$ such that for all $x \in B$ and $\lambda > 0$ one has

$$\mu\{Mx \geq \lambda\|x\|\} \leq C(\lambda)\,. \tag{9}$$

Moreover the set of x for which $T_n x$ converges a.e. is closed in B.

The first chapter of Garsia's book [89] contains a nice introduction to the Banach principle.

It should be noted that, for general integrable f, Mf need not be in L^1. However if $f \in L^p$ for $p > 1$ then Mf does belong to L^p and one has the estimate

$$\|Mf\|_p \leq \frac{p}{p-1}\|f\|_p\,. \tag{10}$$

This can be derived from (6) using also the (obvious) fact that $\|Mf\|_\infty \leq \|f\|_\infty$. Any such estimate on the norm of a maximal function will be called a *strong L_p* inequality. It also follows from (6) that if $\mu(X)$ is finite and $f \in L\log L$, that is $\int |f|\log^+|f|\mathrm{d}\mu < \infty$, then $Mf \in L^1$. In fact Ornstein [160] has shown that the converse of this last statement holds provided T is ergodic.

There is a special setting where one has uniform convergence in the ergodic theorem. Suppose T is a homeomorphism of a compact metric space X. By a theorem of Krylov and Bogoliouboff [138] there is at least one probability measure on the Borel σ-algebra of X which is invariant under T. T is said to be *uniquely ergodic* if there is only one Borel probability measure, say μ, invariant under T. It is easy to see that when this is the case then T is an ergodic automorphism of $(X, \mathcal{B}, \mu)$. As an example, if α is an irrational number then the rotation $z \mapsto \mathrm{e}^{2\pi i\alpha}z$ is a uniquely ergodic transformation of the circle $\{|z| = 1\}$. Equivalently $x \mapsto x + \alpha \mod 1$ is a uniquely ergodic map on $[0, 1]$. A quick way to see this is to show that the Fourier co-efficients $\hat{\mu}(n)$ of any invariant probability μ are zero for $n \neq 0$. The Jewett–Krieger theorem (see Jewett [109] and Krieger [137]) guarantees that unique ergodicity is ubiquitous in the sense that any automorphism of a probability space is measure-theoretically isomorphic to a uniquely ergodic homeomorphism. The following important result is due to Oxtoby [165]).

Theorem 7 *If T is uniquely ergodic, μ is its unique invariant probability measure and $f \in C(X)$ then the ergodic averages $A_n(f)$ converge uniformly to $\int f\,d\mu$.*

This result can be proved along the same lines as the proof given above of von Neumann's theorem. In a nutshell, one uses the fact that the dual of $C_{\mathbb{R}}(X)$ is the space of finite signed measures on X and the unique ergodicity to show that functions of the form $f - f \circ T$ together with the invariant functions (which are just constant functions) span a dense subspace of $C(X)$.

A sequence $\{x_n\}$ in the interval $[0, 1]$ is said to be *uniformly distributed* if $\frac{1}{n}\sum_{i=1}^{n} f(x_n) \to \int_0^1 f(x)\mathrm{d}x$ for any $f \in C[0, 1]$ (equivalently for any Riemann integrable f or for any $f = 1_I$ where I is any subinterval of $[0, 1]$). As a simple application of Theorem 7 one obtains the linear case of the following result of Weyl [202].

Theorem 8 *If α is irrational and $p(x)$ is any non-constant polynomial with integer coefficients then the sequence $\{p(n)\alpha \mathrm{mod} 1\}$ is uniformly distributed.*

Furstenberg [84], see also [86], has shown that Weyl's result, in full generality, can be deduced from the unique ergodicity of certain affine transformations of higher dimensional tori. For polynomials of degree $k > 1$ Weyl's result is usually proved by inductively reducing to the case $k = 1$, using the following important lemma of van der Corput (see for example [139]).

Theorem 9 *Suppose that for each fixed $h > 0$ the sequence*

$$\{x_{n+h} - x_n \mod 1\}_n$$

is uniformly distributed. Then $\{x_n\}$ is uniformly distributed.

When μ is infinite and T is ergodic the limiting function in Birkhoff's theorem is 0 a.e. In 1937 Hopf [100] proved a generalization of Birkhoff's theorem which is more meaningful in the case of an infinite invariant measure. It is a special case of a later theorem of Hurewicz [105], which we will discuss first.

Suppose that $(X, \mathcal{B}, \mu)$ is a σ-finite measure space. If ν is another σ-finite measure on $\mathcal{B}$ write $\nu \ll \mu$ (ν is *absolutely continuous* relative to μ) if $\mu(E) = 0$ implies $\nu(E) = 0$ and one writes $\nu \sim \mu$ if $\nu \ll \mu$ and $\mu \ll \nu$. Consider a *non-singular automorphism* $\tau : X \to X$, meaning that τ is measurable with a measurable inverse and that $\mu(E) = 0$ if and only if $\mu(\tau E) = 0$. In other words $\nu = \mu \circ \tau \sim \mu$. By the Radon–Nikodym theorem there is a function $\rho \in L_1(\mu)$ such that $\rho > 0$ a.e. and $\nu(E) = \int_E \rho\,\mathrm{d}\mu$ for all measurable E. In order to obtain an associated operator T on L_1 which is an (invertible) isometry one defines

$$Tf(x) = \rho(x)f(\tau x)\,. \tag{11}$$

The dual operator on L_∞ is then given by $T^*g = g \circ \tau^{-1}$.

If ν is a σ-finite measure equivalent to μ which is invariant under τ then the ergodic theory of τ can be reduced to the measure-preserving case using σ. The interesting case is when there is no such ν. It was an open

Copyrighted material

problem for some time whether there is always an equivalent invariant measure. In 1960 Ornstein [163] gave an example of a τ which does not have an equivalent invariant measure. It is curious that, with hindsight, such examples were already known in the fifties to people studying von Neumann algebras.

For $f \in L_1$ let $S_n f = \sum_{i=0}^{n} T^i f$. τ is said to be *conservative* if there is no set E with $\mu(E) > 0$ such that $\tau^{-i}E, i = 0, 1, \ldots$ are pairwise disjoint, that is if the Poincaré recurrence theorem remains valid. For example, the shift on $\mathbb{Z}$ is not conservative. Hurewicz [105] proved the following ratio ergodic theorem.

Theorem 10 *Suppose τ is conservative, $f, g \in L_1$ and $g(x) > 0$ a.e. Then $S_n f / S_n g$ converges a.e. to a τ-invariant limit h. If τ is ergodic then $h = \frac{\int f \, d\mu}{\int g \, d\mu}$.*

In the case when μ is τ-invariant one has $Tf = f \circ \tau$. If μ is invariant and finite, taking $g = 1$ one recovers Birkhoff's theorem. If μ is invariant and σ-finite then Hurewicz's theorem becomes the theorem of Hopf alluded to earlier.

Wiener and Wintner [204] proved the following variant of Birkhoff's theorem.

Theorem 11 (Wiener–Wintner) *Suppose T is an automorphism of a probability space $(X, \mathcal{B}, \mu)$. Then for any $f \in L_1$ there is a subset $X' \subset X$ of measure one such that for each $x \in X'$ and $\lambda \in \mathbb{C}$ of modulus 1 the sequence $\frac{1}{n} \sum_{i=0}^{n-1} \lambda^i T^i f(x)$ converges.*

It is an easy consequence of the ergodic theorem that one has a.e. convergence for a given f and λ but the point here is that the set on which the convergence occurs is independent of λ.

Generalizations to Continuous Time and Higher-Dimensional Time

A (measure-preserving) flow is a one-parameter group $\{T_t, t \in \mathbb{R}\}$ of automorphisms of $(X, \mathcal{B}, \mu)$, that is $T_{t+s} = T_t T_s$, such that $T_t x$ is measurable as a function of (t, x). It will always be implicitly assumed that the map $(t, x) \mapsto T_t x$ from $\mathbb{R} \times X$ to X is measurable. Theorem 4 generalizes to flows by replacing sums with integrals and this generalization follows without difficulty from Theorem 4. (As already observed this observation reverses the historical record.) Theorem 4 may be viewed as a theorem about the "discrete flow" $\{T_n = T^n : n \in \mathbb{Z}\}$. Wiener was the first to generalize Birkhoff's theorem to families of automorphisms $\{T_g\}$ indexed by groups more general than $\mathbb{R}$ or $\mathbb{Z}$.

A measure-preserving flow is an *action* of $\mathbb{R}$ while a single automorphism corresponds to an action of $\mathbb{Z}$. A (measure-preserving) action of a group G is a homomorphism $T\colon g \mapsto T_g$ from G into the group of automorphisms of a measure space $(X, \mathcal{B}, \mu)$ (satisfying the appropriate joint measurability condition in case G is not discrete). Suppose now that $G = \mathbb{R}^k$ or $\mathbb{Z}^k$ and T is an action of G on $(X, \mathcal{B}, \mu)$. In the case of $\mathbb{Z}^k$ an action amounts to an arbitrary choice of commuting maps $T_1, \ldots T_k$, $T_i = T(e_i)$ where e_i is the standard basis of $\mathbb{Z}^k$. In the case of $\mathbb{R}^k$ one must specify k commuting flows.

Let m denote counting measure on G in case $G = \mathbb{Z}^k$ and Lebesgue measure in case $G = \mathbb{R}^k$. For any subset E of G with $m(E) < \infty$ let

$$A_E f(x) = \frac{1}{m(E)} \int_E f(T_g x) \mathrm{d}m(g) \,. \tag{12}$$

One may then ask whether $A_E f$ converges to a limit, either in the mean or pointwise, as E varies through some sequence of sets which "grow large" or, in case $G = \mathbb{R}^k$, "shrink to 0". The second case is referred to as a *local ergodic theorem*. In the case of ergodic theorems at infinity the continuous and discrete theories are rather similar and often the continuous analogue of a discrete result can be deduced from the discrete result.

In Wiener [203] proved the following result for actions of $G = \mathbb{R}^k$ and ergodic averages over Euclidean balls $B_r = \{x \in \mathbb{R}^k : \|x\|_2 \le r\}$.

Theorem 12 *Suppose T is an action of $\mathbb{R}^k$ on $(X, \mathcal{B}, \mu)$ and $f \in L_1(\mu)$.*

(a) *For $f \in L_1$ $\lim_{r\to\infty} A_{B_r} f = g$ exists a.e. If μ is finite the convergence also holds with respect to the L_1-norm, g is T-invariant and $\int_I g d\mu = \int_I f \, d\mu$ for every T-invariant set I.*

(b) $\lim_{r\to 0} A_{B_r} f = f$ *a.e.*

The local aspect of Wiener's theorem is closely related to the Lebesgue differentiation theorem, see for example Proposition 3.5.4 of [132], which, in its simplest form, states that for $f \in L^1(\mathbb{R}^k, m)$ one has a.e. convergence of $\frac{1}{m(B_r)} \int_{B_r} f(x + t) \mathrm{d}t$ to $f(x)$ as $r \to 0$. The local ergodic theorem implies Lebesgue's theorem, simply by considering the action of $\mathbb{R}^k$ on itself by translation. In fact the local ergodic theorem can also be deduced from Lebesgue's theorem by a simple application of Fubini's theorem (see for example [135], Chap. 1, Theorem 2.4 in the case $k = 1$).

The key point in Wiener's proof is the following weak L_1 maximal inequality, similar to (6).

Theorem 13 *Let $Mf = \sup_{r>0} |A_{B_r} f|$. Then one has*

$$\mu\{Mf > \alpha\} \le \frac{C}{\alpha} \|f\|_1 \,, \tag{13}$$

Copyrighted material

where C is a constant depending only on the dimension d. (In fact one may take $C = 3^d$.)

In the case when T is the action of R^k on itself by translation (13) is the well-known maximal inequality for the Hardy–Littlewood maximal function ([132], Lemma 3.5.3). Wiener proves (13) by way of the following *covering lemma*. If B is a ball in $\mathbb{R}^k$ let B' denote the concentric ball with three times the radius.

Theorem 14 *Suppose a compact subset K of $\mathbb{R}^k$ is covered by a (finite) collection $\mathcal{U}$ of open balls. Then there exist pairwise disjoint $B_i \in \mathcal{U}$, $i = 1, \ldots k$ such that $\bigcup_i B_i'$ covers K.*

To find the B_i it suffices to let B_1 be the largest ball in $\mathcal{U}$, then B_2 the largest ball which does not intersect B_1 and in general B_n the largest ball which does not intersect $\bigcup_{i=1}^{n-1} B_i$. Then it is not hard to argue that the B_i' cover K.

In general, a covering lemma is, roughly speaking, an assertion that, given a cover $\mathcal{U}$ of a set K in some space (often a group), one may find a subcollection $\mathcal{U}' \subset \mathcal{U}$ which covers a substantial part of K and is in some sense efficient in that $\sum_{U \in \mathcal{U}'} 1_U \leq C$, C some absolute constant. Covering lemmas play an important role in the proofs of many maximal inequalities. See Sect. 3.5 of [132] for a discussion of several of the best-known classical covering lemmas.

As Wiener was likely aware, the same kind of covering argument easily leads to maximal inequalities and ergodic theorems for averages over "sufficiently regular" sets. For example, in the case of $G = \mathbb{Z}^k$ use the standard total order $\prec$ on G and for $n \in \mathbb{N}^k$ let $S_n = \{m \in Z^k : 0 \prec m \prec n\}$. Let $e(n)$ denote the maximum value of n_i/n_j. Then one has the following result.

Theorem 15 *For any $e > 0$ and any integrable f*

$$\lim_{n \to \infty, e(n) < e} A_{S_n} f$$

exists a.e. and the limit is characterized by the same properties as in Theorem 12. Moreover there is a maximal inequality

$$\mu \left\{ \sup_{e(n) < e} |A_{S_n} f| > \alpha \right\} \leq \frac{C}{\alpha} \|f\|_1 , \tag{14}$$

where C is a constant depending only on e and the dimension k.

If one does not impose a bound on $e(n)$ the a.e. convergence of $A_{S_n} f$ may fail for $f \in L_1$, (although it does hold for any increasing sequence of n's by Theorem 30 below). Nonetheless Dunford [72] and independently Zygmund [215] showed that one does have unrestricted convergence if $f \in L_p$ for some $p > 1$, provided μ is finite. Let $T_1, \ldots, T_k$ be any k (possibly non-commuting!) automorphisms of a finite measure space $(X, \mathcal{B}, \mu)$. For $n = (n_1, n_2, \ldots, n_k)$ let $T_n = T_1^{n_1} \ldots T_k^{n_k}$. (This is not an action of $\mathbb{Z}^k$ unless the T_i commute with each other.) As before when $F \subset \mathbb{Z}^k$ is a finite subset write $A_F f = \frac{1}{|F|} \sum_{n \in F} T_n f$. Finally let $P_j f = \lim_n A_n(T_j) f$.

Theorem 16 *For $f \in L_p$*

$$\lim_{n \to \infty} A_{S_n} f = P_1 \ldots P_k f \tag{15}$$

both a.e. and in L_p.

The proof of Theorem 16 uses repeated applications of Birkhoff's theorem and hinges on (10).

Somewhat surprisingly Hurewicz's theorem was not generalized to higher dimensions until the very recent work of Feldman [79]. In fact the theorem fails if one considers averages over the cube $[0, n-1]^d$ in $\mathbb{Z}^d$. However Feldman was able to prove a suitably formulated generalization of Hurewicz's theorem for averages over symmetric cubes.

For $f \in L^1(\mathbb{R})$ the classical *Hilbert transform* is defined for a.e. t by

$$Hf(t) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \int_{|s| > \epsilon} \frac{f(t-s)}{s} ds \tag{16}$$

Let $H^* f(t) = \frac{1}{\pi} \sup_{\epsilon > 0} | \int_{|s| > \epsilon} \frac{f(t-s)}{s} ds |$, the corresponding maximal function. The proof that the limit (16) exists a.e. is based on the maximal inequality

$$m\{Hf > \lambda\} \leq \frac{C}{\lambda} \|f\|_1 , \tag{17}$$

where C is an absolute constant. See Loomis [146] (1946) for a proof of (17) using real-variable methods. Cotlar [65] proved the existence of the following ergodic analogue of the Hilbert transform (actually for the n-dimensional Hilbert transform).

Theorem 17 *Suppose $\{T_t\}$ is a measure-preserving flow on a probability space $(X, \mathcal{B}, \mu)$ and $f \in L^1$. Then*

$$\lim_{\epsilon \to 0} \int_{\epsilon < t < \epsilon^{-1}} \frac{f(T_t x)}{t} dt \tag{18}$$

exists for a.a. x.

Motivated by Cotlar's result Calderón [54]proved a general *transfer principle* which allows one to transfer maximal inequalities and convergence theorems for functions of a real or integer variable to the ergodic setting. Although stated for functions on $\mathbb{R}$ it applies equally well to $\mathbb{R}^k$

Copyrighted material

or $\mathbb{Z}^k$. This principle subsumes Birkhoff's theorem, Wiener's theorem and Cotlar's result. For simplicity only a special case of Calderón's result, in the discrete case, will be stated here.

Let T be an automorphism of a probability space $(X, \mathcal{B}, \mu)$, let m denote counting measure on $\mathbb{Z}$ and suppose σ is a probability measure on $\mathbb{Z}$. For $g \in L_1(m)$ define $\sigma(g)(n) = \sum_{i\in\mathbb{Z}} \sigma(i)g(n+i)$. For $f \in L_1(\mu)$ let $\sigma(T)f = \sum_{i\in\mathbb{Z}} \sigma(i)T^i f$, which is easily seen to converge a.e. and in L_1-norm. Given a fixed sequence σ_n of probabilities define $Mg = \sup_n \sigma_n g$ and $M(T)f = \sup_n \sigma_n(T)f$.

Theorem 18 (Calderón) *Suppose there is a constant C such that*

$$m\{Mg > \lambda\} < \frac{C}{\lambda}\|g\|_1 \quad \textit{for all } g \in L_1(m)\,. \tag{19}$$

Then one also has

$$\mu\{M(T)f > \lambda\} < \frac{C}{\lambda}\|f\|_1 \quad \textit{for all } T \quad \textit{and } g \in L_1(\mu)\,. \tag{20}$$

In other words, in order to prove a maximal inequality for general T it suffices to prove it in case T is the shift map on the integers. The idea of transference could already be seen in Wiener's proof of the $\mathbb{R}^n$ ergodic theorem. Transfer principles in various forms of have become an important tool in the study of ergodic theorems. See Bellow [18] for a very readable overview.

Pointwise Ergodic Theorems for Operators

Early in the history of ergodic theorems there were attempts to generalize the ergodic theorem to more general linear operators on L_p spaces, that is, operators which do not arise by composition with a mapping of X . In the case $p = 1$ the main motivation for this comes from the theory of Markov processes.

If $(X, \mathcal{B})$ is a measurable space a *sub-stochastic kernel* on X is a non-negative function P on $X \times \mathcal{B}$ such that

(a) for each $x \in XP(x, \cdot) = P_x$ is a measure on $\mathcal{B}$ such that $P_x(X) \leq 1$ and
(b) $P(\cdot, A)$ is a measurable function for each $A \in \mathcal{B}$.

It is most intuitive to think about the *stochastic* case, namely when each P_x is a probability measure. One then views $P(x, A)$ as the probability that the point x moves into the set A in one unit of time, so one has stochastic dynamics as opposed to deterministic dynamics, namely the case when $P_x = \delta_{Tx}$ for a map T. In this case the measures P_x are called *transition probabilities*.

If μ is a σ-finite measure on X one may define the measure $P\mu = \int P_x \mathrm{d}\mu(x)$. $P\nu$ is also meaningful if ν is a finite signed measure. The case when $P\mu = \mu$ is the stochastic analogue of measure-preserving dynamics and the case when $P\mu \ll \mu$ is the analogue of non-singular dynamics. It is easy to see that given any σ-finite measure λ there is always a μ such that $\lambda \ll \mu$ and $P\mu \ll \mu$. Let $\tilde{L}_1$ denote the space of finite signed measures ν such that $\nu \ll \mu$, which is identified with $L_1 = L_1(\mu, \mathbb{R})$ via the Radon–Nikodym theorem. If $P\mu \ll \mu$ then P maps $\tilde{L}_1(\mu)$ into itself so the restriction of P is an operator T on $L_1(\mu)$. T is a positive contraction, that is $\|T\| \leq 1$ and T maps non-negative functions to non-negative functions. As proved in, for example, [153], every positive contraction arises in this way from a substochastic kernel under the assumption that $(X, \mathcal{B})$ is standard. This simply means that there is some complete metric on X for which $\mathcal{B}$ is the σ-algebra of Borel sets. Virtually all measurable spaces encountered in analysis are standard so this should be viewed as a technicality only. See [153], [80] and [135] for more about the relation between kernels and positive contractions.

The case when X is finite, P is stochastic and μ is a a probability measure is classical in probability theory. P and μ determine a probability measure ν on $X^{\mathbb{N}}$ characterized by its values on cylinder sets, namely for all $x_1, \ldots, x_n \in X$

$$\begin{aligned}\nu\{x \in X^{\mathbb{N}} : x(i) = x_i, 1 \leq i \leq n-1\} \\ = \mu(x_1)\prod_{i=1}^{n-1} P_{x_i}(x_{i+1})\,.\end{aligned} \tag{21}$$

The co-ordinate functions $X_i(x) = x(i)$ on the space $X^{\mathbb{N}}$ endowed with the probability ν form a *Markov process*, which will be stationary if and only if μ is P-invariant. For a general X the analogous construction is possible provided X is standard.

Hopf [101] initiated the systematic study of positive L_1 contractions and proved the following ergodic theorem.

Theorem 19 *Suppose $(X, \mathcal{B}, \mu)$ is a probability space and T is a positive contraction on $L^1(\mu)$ satisfying $T1 = 1$ and $T^*1 = 1$. Then $\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} T^k f$ exists a.e.*

The importance of Hopf's article lies less in this convergence result than in the methods he developed. He proved that the maximal inequality (5) generalizes to all positive L_1 contractions T and used this to obtain the decomposition of X into its conservative and dissipative parts C and D, characterized by the fact that for any $p \in L_1$ such that $p > 0$ a.e. $\sum_{k=0}^{\infty} T^k p$ is infinite a.e. on C and finite a.e. on D. These results are the cornerstone of much of the subsequent work on L_1 contractions.

Copyrighted material

Theorem 19 contains Birkhoff's theorem for an automorphism τ of $(X, \mathcal{B}, \mu)$, simply by defining $Tf = f \circ \tau$. In fact one can also deduce Theorem 19 from Birkhoff's theorem (if one assumes only that X is standard). Indeed the hypotheses of Hopf's theorem imply that the kernel P associated to T is stochastic ($T^*1 = 1$) and and that μ is P-invariant ($T1 = 1$). Hopf's theorem then follows by applying Birkhoff's theorem to the shift on the stationary Markov process associated to P and μ. In fact Kakutani [123] (see also Doob [71]) had already made essentially the same observation, except that his result assumes the stationary Markov process to be already given.

In 1955 Dunford and Schwartz [73] made essential use of Hopf's work to prove the following result.

Theorem 20 *Suppose $\mu(X) = 1$ and T is a (not necessarily positive) contraction with respect to both the L_1 and L_∞ norms. Then the conclusion of Hopf's theorem remains valid.*

Note that the assumption that T contracts the L_∞-norm is meaningful, as $L_\infty \subset L_1$. The proof of the result is reduced to the positive case by defining a positive contraction $|T|$ analogous to the total variation of a complex measure.

Then in 1960 Chacón and Ornstein [61] proved a definitive ratio ergodic theorem for all positive contractions of L_1 which generalizes both Hopf's theorem and the Hurewicz theorem.

Theorem 21 *Suppose T is a positive contraction of $L_1(\mu)$, where μ is σ-finite, $f, g \in L_1$ and $g \geq 0$. Then*

$$\frac{\sum_{i=0}^{n} T^i f}{\sum_{i=0}^{n} T^i g} \tag{22}$$

converges a.e. on the set $\{\sum_{i=0}^{\infty} T^i g > 0\}$.

In 1963 Chacón [58] proved a very general theorem for non-positive operators which includes the Chacón–Ornstein theorem as well as the Dunford–Schwartz theorem.

Theorem 22 *Suppose T is a contraction of L_1 and $p_n \geq 0$ is a sequence of measurable functions with the property that*

$$g \in L_1, |g| \leq p_n \Rightarrow |Tg| \leq p_{n+1} \,. \tag{23}$$

Then

$$\frac{\sum_{i=0}^{n} T^i f}{\sum_{i=0}^{n} p_i} \tag{24}$$

converges a.e. to a finite limit on the set $\{\sum_{i=0}^{\infty} p_i > 0\}$.

If T is an L_1, L_∞-contraction and $p_n = 1$ for all n then the hypotheses of this theorem are satisfied, so Theorem 22 reduces to the result of Dunford and Schwartz. See [59] for a concise overview of all of the above theorems in this section and the relations between them.

The identification of the limit in the Chacón–Ornstein theorem on the conservative part C of X is a difficult problem. It was solved by Neveu [152] in case $C = X$, and in general by Chacón [57]. Chacón [60] has shown that there is a non-singular automorphism τ of $(X, \mathcal{B}, \mu)$ such that for the associated invertible isometry T of $L_1(\mu)$ given by (11) there is an $f \in L_1(\mu)$ such that $\limsup A_n f = \infty$ and $\liminf A_n f = -\infty$.

In 1975 Akcoglu [3] solved a major open problem when he proved the following celebrated theorem.

Theorem 23 *Suppose $T\colon L_p \mapsto L_p$ is a positive contraction. Then $A_n f = \frac{1}{n}\sum_{i=0}^{n-1} T^i f$ converges a.e. Moreover one has the strong L_p inequality*

$$\left\| \sup_n |A_n f| \right\|_p \leq \frac{p}{p-1} \|f\|_p \,. \tag{25}$$

As usual, the maximal inequality (25) is the key to the convergence. Note that it is identical in form to (10) the classical strong L_p inequality for automorphisms. (25) was proved by A. Ionesco–Tulcea (now Bellow) [106] in the case of positive invertible isometries of L_p. It is a result of Banach [16], see also [140], that in this case T arises from a non-singular automorphism τ of $(X, \mathcal{B}, \mu)$ in the form $Tf = \rho^{1/p} f \circ \tau$. By a series of reductions Bellow was able to show that in this case (25) can be deduced from (10).

Akcoglu's brilliant idea was to consider a *dilation* S of T which is a positive invertible isometry on a larger L_p space $\tilde{L}_p = L_p(Y, C, \nu)$. What this means is that there is a positive isometric injection $D\colon L_p \to \tilde{L}_p$ and a positive projection P on $\tilde{L}_p$ whose range is $D(L_p)$ such that $DT^n = PS^n D$ for all $n \geq 0$. Given the existence of such an S it is not hard to deduce (25) for T from (25) for S. In fact Akcoglu constructs a dilation only in the case when L_p is finite dimensional and shows how to reduce the proof of (25) to this case. In the finite dimensional case the construction is very concrete and P is a conditional expectation operator. Later Akcoglu and Kopp [7] gave a construction in the general case. It is noteworthy that the proof of Akcoglu's theorem consists ultimately of a long string of reductions to the classical strong L_p inequality (10), which in turn is a consequence of (5).

Subadditive and Multiplicative Ergodic Theorems

Consider a family $\{X_{n,m}\}$ of real-valued random variables on a probability space indexed by the set of pairs $(n, m) \in$

Copyrighted material

In other words, the actual information obtained per second by observing x over time converges to the constant $h(P, T)$, namely the limiting expected information per second. Shannon [185] formulated Theorem 28 and proved convergence in probability. McMillan [149] proved L_1 convergence and Breiman [49] obtained the a.e. convergence.

The original proofs of a.e. convergence used the martingale convergence theorem, were not very intuitive and did not generalize to $\mathbb{Z}^n$-actions, where the martingale theorem is not available. Ornstein and Weiss [161] (1983) found a beautiful and more natural argument which bypasses the martingale theorem and allows generalization to a class of groups which includes $\mathbb{Z}^n$.

Amenable Groups

Let G be any countable group and $T = \{T_g\}$ an action of G by automorphisms of a probability space. Suppose σ is a complex measure on G, that is, $\{\sigma(g)\}_{g\in G}$ is an absolutely summable sequence. Let T_g act on functions via $T_g f = f \circ T_g$, so T_g is an isometry of L_p for every $1 \le p \le \infty$. Let $\sigma(T) = \sum_{g\in G} \sigma(g) T_g$. A very general framework for formulating ergodic theorems is to consider a sequence $\{\sigma_n\}$ and ask whether $\sigma_n(T)f$ converges, a.e. or in mean, for f in some L_p space. When $\sigma_n(T)f$ converges for all actions T and all f in L_p in p-norm or a.e. then one says that σ_n is *mean* or *pointwise good* in L_p. When the σ_n are probability measures it is natural to call such results weighted ergodic theorems and this terminology is retained for complex σ as well.

Birkhoff's theorem says that if $G = \mathbb{Z}$ and σ_n is the normalized counting measure on $\{0, 1, \dots n-1\}$ then $\{\sigma_n\}$ is pointwise good in L_1. This section will be concerned only with sequences $\{\sigma_n\}$ such that σ_n is normalized counting measure on a finite subset $F_n \subset G$ so one speaks of mean or pointwise good sequences $\{F_n\}$. A natural condition to require of $\{F_n\}$, which will ensure that the limiting function is invariant, is that it be asymptotically (left) invariant in the sense that

$$\frac{|gF_n \triangle F_n|}{|F_n|} \to 0 \quad \forall g \in G\,. \tag{30}$$

Such a sequence is called a *Følner sequence* and a group G is *amenable* if it has a Følner sequence. As in most of this article G is restricted to be a discrete countable group for simplicity but most of the results to be seen actually hold for a general locally compact group.

Amenability of G is equivalent to the existence of a finitely additive left invariant probability measure on G. It is not hard to see that any Abelian, and more generally any solvable, group is amenable. On the other hand the free group F_2 on two generators is not amenable. See Paterson [167] for more information on amenable groups. The Følner property by itself is enough to give a mean ergodic theorem.

Theorem 29 *Any Følner sequence is mean good in L_p for $1 \le p < \infty$.*

The proof of this result is rather similar to the proof of Theorem 3. In fact Theorem 29 is only a special case of quite general results concerning amenable semi-groups acting on abstract Banach spaces. See the book of Paterson [167] for more on this.

Turning to pointwise theorems, the Følner condition alone does not yield a pointwise theorem, even when $G = \mathbb{Z}$ and the F_n are intervals. For example Akcoglu and del Junco [6] have shown that when $G = \mathbb{Z}$ and $F_n = [n, n + \sqrt{n}] \cap \mathbb{Z}$ the pointwise ergodic theorem fails for any aperiodic T and for some characteristic function f. See also del Junco and Rosenblatt [119].

The following pointwise result of Tempelman [196] is often quoted. A Følner sequence $\{F_n\}$ is called *regular* if there is a constant C such that $|F_n^{-1}F_n| \le C|F_n|$ and there is an increasing sequence F_n' such that $F_n \subset F_n'$ and $|F_n'| \le C|F_n|$.

Theorem 30 *Any regular Følner sequence is pointwise good in L_1.*

In case the F_n are intervals in $\mathbb{Z}^n$ this result can be proved by a variant of Wiener's covering argument and in the general case by an abstraction thereof. The condition $|F_n^{-1}F_n| \le C|F_n|$ captures the property of rectangles which is needed for the covering argument. Emerson [78] independently proved a very similar result.

The work on ergodic theorems for abstract locally compact groups was pioneered by Calderón [53] who built on Wiener's methods. The main result in this paper is somewhat technical but it already contains the germ of Tempelman's theorem. Other ergodic theorems for amenable groups, whose main interest lies in the case of continuous groups, include Tempelman [195], Renaud [175], Greenleaf [93] and Greenleaf and Emerson[94]. The discrete versions of these results are all rather close to Tempelman's theorem.

Among pointwise theorems for discrete groups Tempelman's result was essentially the best available for a long time. It was not known whether every amenable group had a Følner sequence which is pointwise good for some L_p. In 1988 Shulman [187] introduced the notion of a *tempered*

Copyrighted material

Følner sequence $\{F_n\}$, namely one for which

$$\left|\bigcup_{i<n} F_i^{-1} F_n\right| < C|F_n|\,, \tag{31}$$

for some constant C. The advantage of the tempered condition is that any Følner sequence has a tempered subsequence, and in particular any amenable group has a tempered Følner sequence.

Shulman proved a maximal inequality in L_2 for such F_n which implies that $\{F_n\}$ is pointwise good in L_2. An account of this work may be found in Section 5.6 of Tempelman's book [194].

Lindenstrauss [145] was able to extend the result to L_1.

Theorem 31 *Any tempered Følner sequence is pointwise good in L_1.*

The key new idea in his proof is to use a probabilistic argument to establish a covering lemma. In the discrete case Ornstein and Weiss [201] have given a non-probabilistic proof of Lindenstrauss's covering lemma. Lindenstrauss also generalizes the a.e. convergence in the Shannon–McMillan–Breiman theorem to this setting. L_1 convergence was already established by Kieffer [128] in 1975.

Subsequence and Weighted Theorems

In this section G and T are as in the previous section and σ_n is a sequence of complex measures on G.

This section will be concerned with conditions on σ_n that ensure that it is mean or pointwise good. For the most part G will be $\mathbb{Z}$.

Hopf's ergodic theorem gives a class of examples for free. Choose any probability measure λ on G, define the operator $T_\lambda = \lambda(T)$ and observe that $(T_\lambda)^n = T_{\lambda^{*n}}$, where λ^{*n} denotes the convolution power. Since T_λ satisfies the hypotheses of Hopf's theorem it follows that the sequence $\sigma_n = \frac{1}{n}\sum_{i=0}^{n-1} \lambda^{*i}$ is pointwise good in L_1.

Another sort of example is given by choosing a sequence g_n in G and letting $\sigma_n = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{g_n}$. If one has convergence for such a sequence one speaks of a *subsequence ergodic theorem*. This section will mainly focus on the case $G = \mathbb{Z}$ and sequences which are the increasing enumeration of a subset $S \subset N$. Given $S \subset \mathbb{N}$ write $\sigma_{S,n}$ for the corresponding probabilities σ_n and say S is good if $\sigma_{S,n}$ is good.

Perhaps the first subsequence ergodic theorem is due to Blum and Hanson [38] who proved that an automorphism T of a probability space is *strongly mixing* if and only if $\frac{1}{n}\sum_{i=0}^{n-1} T^{m_i} f$ converges in L_2 norm for every $f \in L_2$ and every increasing $\{m_i\}$. Strong mixing means that $\mu(T^{-n}A \cap B) \to \mu(A)\mu(B)$. In 1969 Brunel and Keane [51] proved the first pointwise subsequence ergodic theorem

Theorem 32 *Suppose that T is a translation on a compact Abelian group G with Haar measure λ, $g \in G$, and $E \subset G$ is any Borel set with $\lambda(E) > 0$ and $\lambda(\partial E) = 0$. Let $S = \{i > 0 : T^i g \in E\}$. Then S is pointwise good in L_1.*

Krengel [133] constructed the first example of a sequence $S \subset \mathbb{N}$ which is pointwise universally bad, in the strong sense that for any aperiodic T the a.e. convergence of $\sigma_{S,n}(T)f$ fails for some characteristic function f. Bellow [17] proved that any lacunary sequence (meaning $a_{n+1} > ca_n$ for some $c > 1$) is pointwise universally bad in L_1.

Later Akcoglu at al. [1] were able to show that for lacunary sequences $\{\sigma_{S,n}\}$ is even *strongly sweeping out*. A sequence $\{\sigma_n\}$ of probability measures on $\mathbb{Z}$ is said to be strongly sweeping out if for any ergodic T and for all $\delta > 0$ there is a characteristic function f with $\int f\,d\mu < \delta$ such that $\limsup \sigma_n(T)f = 1$ a.e. It is not difficult to show that if $\{\sigma_n\}$ is strongly sweeping out then there are characteristic functions f such that $\liminf \sigma_n(T)f = 0$ and $\limsup \sigma_n(T)f = 1$. Thus for lacunary sequences the ergodic theorem fails in the worst possible way.

Bellow and Losert [21] gave the first example of a sequence $S \subset \mathbb{Z}$ of density 0 which is universally good for pointwise convergence, answering a question posed by Furstenberg. They construct an S which is pointwise good in L_1. This paper also contains a good overview of the progress on weighted and subsequence ergodic theorems at that time.

Weyl's theorem on uniform distribution (Theorem 9) suggests the possibility of an ergodic theorem for the sequence $\{n^2\}$. It is not hard to see that $\{n^2\}$ is mean good in L_2. In fact the spectral theorem and the dominated convergence theorem show that it is enough to prove that the L_∞-bounded sequence of functions $\frac{1}{n}\sum_{i=0}^{n-1} z^{n^2}$ on the unit circle converges at each point z of the unit circle. When z is not a root of unity the sequence converges to 0 by Weyl's result and when z is a root of unity the convergence is trivial because $\{z^{n^2}\}$ is periodic. In 1987 Bourgain [39,43] proved his celebrated pointwise ergodic theorem for polynomial subsequences.

Theorem 33 *If p is any polynomial with rational coefficients taking integer values on the integers then $S = \{p(n)\}$ is pointwise good in L_2.*

The first step in Bourgain's argument is to reduce the problem of proving a maximal inequality to the case of the shift map on the integers, via Calderón's transfer princi-

Copyrighted material

ple. Then the problem is transferred to the circle by using Fourier transforms. At this point the problem becomes a very delicate question about exponential sums and a whole arsenal of tools is brought to bear. See Rosenblatt and Wierdl [176] and Quas and Wierdl [28](Appendix B) for nice expositions of Bourgain's methods.

Bourgain subsequently improved this to all L_p, $p > 1$ and also extended it to sequences $\{[q(n)]\}$ where now q is an arbitrary real polynomial and $[\cdot]$ denotes the greatest integer function. He also announced that his methods can be used to show that the sequence of primes is pointwise good in L_p for any $p > \frac{1+\sqrt{3}}{2}$. Wierdl [205] (1988) soon extended the result for primes to all $p > 1$.

Theorem 34 *The primes are pointwise good in L_p for $p > 1$.*

It has remained a major open question for quite some time whether any of these results hold for $p = 1$. In 2005 there appeared a preprint of Mauldin and Buczolich [148], which remains unpublished, showing that polynomial sequences are L_1-universally bad.

Another major result of Bourgain's is the so-called return times theorem [44]. A simplification of Bourgain's original proof was published jointly with Furstenberg, Katznelson and Ornstein as an appendix to an article [47] of Bourgain. To state it let us agree to say that a a sequence of complex numbers $\{a(n)\}_{n\geq 0}$ has property P if the sequence of complex measures $\sigma_n = \frac{1}{n}\sum_{i=0}^{n-1} a(i)\delta_i$ has property P, where δ_i denotes the point mass at i.

Theorem 35 (Bourgain) *Suppose T is an automorphism of a probability space $(X, \mathcal{B}, \mu)$, $1 \leq p, q \leq \infty$ are conjugate exponents and $f \in L_p(\mu)$. Then for almost all x the sequence $\{f(T^n x)\}$ is pointwise good in L_q.*

Applying this to characteristic functions $f = 1_E$ one sees that the return time sequence $\{i > 0 : S^i x \in E\}$ is good for pointwise convergence in L_1. Theorem 32 is a very special case. It is also easy to see that Theorem 35 contains the Wiener–Wintner theorem.

In 1998 Rudolph [179] proved a far-reaching generalization of the return times theorem using the technique of *joinings*. For an introduction to joinings see the article by de la Rue in this collection and also Thouvenot [198], Glasner [90] (2003) and Rudolph's book [177]. Rudolph's result concerns the convergence of multiple averages

$$\frac{1}{N}\sum_{n=0}^{N-1}\prod_{j=1}^{k} f_j(T_j^n x) \tag{32}$$

where each T_j is an automorphism of a probability space $(X_j, \mathcal{B}_j, \mu_j)$ and the f_j are L_∞ functions. The point is that the convergence occurs whenever each $x_j \in X_j'$, sets of measure one which may be chosen sequentially for $j = 1, \ldots, k$ without knowing what T_i or f_i are for any $i > j$. He actually proves something stronger, namely he identifies an intrinsic property of a sequence $\{a_i\}$, which he calls *fully generic*, such that the following hold.

(a) The constant sequence $\{1\}$ is fully generic.

(b) If $\{a_i\}$ is fully generic then for any T and $f \in L_\infty$ the sequence $a_i f(T^i x)$ is fully generic for almost all x.

(c) Fully generic implies pointwise good in L_∞.

The definition of fully generic will not be quoted here as it is somewhat technical.

For a proof of the basic return times theorem using joinings see Rudolph [178]. Assani, Lesigne and Rudolph [13] took a first step towards the multiple theorem, a Wiener–Wintner version of the return times theorem. Also Assani [11] independently gave a proof of Rudolph's result in the case when all the T_j are weakly mixing.

Ornstein and Weiss [162] have proved the following version of the return times theorem for abstract discrete groups. As with $\mathbb{Z}$, let us say that a sequence $\{a_g\}_{g\in G}$ of complex numbers has property P for $\{F_n\}$ if the sequence $\sigma_n = \frac{1}{|F_n|}\sum_{g\in F_n} a(g)\delta_g$ of complex measures has property P.

Theorem 36 *Suppose that the increasing Følner sequence $\{F_n\}$ satisfies the Tempelman condition $\sup_n |F_n^{-1}F_n|/|F_n| < \infty$ and $\bigcup F_n = G$. If $b \in L_\infty$ then for a.a. x the sequence $\{b(T_g x)\}$ is pointwise good in L_1 for $\{F_n\}$.*

Recently Demeter, Lacey, Tao and Thiele [67] have proved that the return times theorem remains valid for any $1 < p \leq \infty$ and $q \geq 2$. On the other hand Assani, Buczolich and Mauldin [14] (2005) showed that it fails for $p = q = 1$.

Bellow, Jones and Rosenblatt have a series of papers [22,23,24,25] studying general weighted averages associated to a sequence σ_n of probability measures on $\mathbb{Z}$, and, in some cases, more general groups. The following are a few of their results. [23] is concerned with $\mathbb{Z}$-actions and *moving block* averages given by $\sigma_n = m_{I_n}$, where the I_n are finite intervals and m_I denotes normalized counting measure on I. They resolve the problem completely, obtaining a checkable necessary and sufficient condition for such a sequence to be pointwise good in L_1.

[24] gives sufficient conditions on a sequence σ_n for it to be pointwise good in L_p, $p > 1$, via properties of the Fourier transforms $\hat{\sigma}_n$. A particular consequence is that if $\lim_{n\to\infty}\sum_{k\in\mathbb{Z}} |\sigma_n(k) - \sigma_n(k-1)| = 0$ then $\{\sigma_n\}$ has a subsequence which is pointwise good in L_p,

Copyrighted material

$p > 1$. In [25] they obtain convergence results for sequences $\sigma_n = \sigma^n$, the convolution powers of a probability measure σ. A consequence of one of their main results is that if the expectation $\sum_{k\in\mathbb{Z}} k\sigma(k)$ is zero, the second moment $\sum_{k\in\mathbb{Z}} k^2\sigma(k)$ is finite and σ is *aperiodic* (its support is not contained in any proper coset in $\mathbb{Z}$) then σ^n is pointwise good in L_p for $p > 1$.

Bellow and Calderón [19] later showed that this last result is valid also for $p = 1$. This is a consequence of the following sufficient condition for a sequence T to satisfy a weak L_1 inequality. Given an automorphism of a probability space $(X, \mathcal{B}, \mu)$ let $Mf = \sup|\sigma_n(T)f(x)|$ be the maximal operator associated to $\{\sigma_n\}$.

Theorem 37 (Bellow and Calderón) *Suppose there is an $\alpha \in (0, 1]$ and $C > 0$ such that for each $n > 1$ one has*

$$|\sigma_n(x+y) - \sigma_n(x)| \le C\frac{|y|^\alpha}{|x|^{1+\alpha}} \quad \text{for all } x, y \in \mathbb{Z} \text{ such that } 0 < 2|y| \le |x|$$

Then there is a constant D such that

$$\mu\{Mf > \lambda\} \le \frac{D}{\lambda}\|f\|_1 \quad \text{for all } T, \quad f \in L_1(\mu) \text{ and } \lambda > 0\,.$$

Ergodic Theorems and Multiple Recurrence

Suppose $S \subset \mathbb{N}$. The *upper density* of S is

$$\bar{d}(S) = \limsup_n \frac{|S \cap [1, n]|}{n}\,. \tag{33}$$

and the *density* $d(S)$ is the limit of the same quantity, if it exists. In 1975 Szemerédi [190] proved the following celebrated theorem, answering an old question of Erdős and Turán.

Theorem 38 *Any subset of $\mathbb{N}$ with positive upper density contains an arithmetic progression of length k for each $k \ge 1$.*

This result has a distinctly ergodic-theoretic flavor. Letting T denote the shift map on $\mathbb{Z}$, it says that for each k there is an n such that $S' = \bigcap_{i=1}^k T^{-in}S$ is non-empty. In fact the result gives more: there is an n for which $\bar{d}(S') > 0$. In this light Szemerédi's theorem becomes a multiple recurrence theorem for the shift map on $\mathbb{N}$, equipped with the invariant "measure-like" quantity $\bar{d}$. Of course $\bar{d}$ is not even finitely additive so it is *not* a measure. d, however, is at least finitely additive, when defined, and $d(\mathbb{N}) = 1$.

This point of view suggests the following multiple recurrence theorem.

Theorem 39 *Suppose T is an automorphism of a probability space $(X, \mathcal{B}, \mu)$, $\mu(B) > 0$ and $k \ge 1$. Then there is an $n > 0$ such that $\mu(\bigcap_{i=1}^k T^{in}B) > 0$.*

In 1977, Furstenberg [85] proved the following ergodic theorem which implies the multiple recurrence theorem. He also established a general *correspondence principle* which puts the shaky analogy between the multiple recurrence theorem and Szemerédi's theorem on a firm footing and allows each to be deduced from the other. Thus he obtained an ergodic theoretic proof of Szemerédi's combinatorial result.

Theorem 40 *Suppose T is an automorphism of a probability space $(X, \mathcal{B}, \mu)$, $f \in L_\infty$, $f \ge 0$, $\int f d\mu > 0$ and $k \ge 1$. Then*

$$\liminf_N \frac{1}{N}\sum_{n=0}^{N-1}\int\prod_{i=1}^k T^{in}f\,d\mu > 0\,. \tag{34}$$

Furstenberg's result opened the door to the study of so-called ergodic Ramsey theory which has yielded a vast array of deep results in combinatorics, many of which have no non-ergodic proof as yet. The focus of this article is not on this direction but the reader is referred to Furstenberg's book [86] for an excellent introduction and to Bergelson [27,28] for surveys of later developments. There is also the article by Frantzikinakis and McCutcheon in this collection.

Furstenberg's proof relies on a deep structure theorem for a general automorphism which was also developed independently by Zimmer [213], [212] in a more general context. A *factor* of T is any sub-σ-algebra $\mathcal{F} \subset \mathcal{B}$ such that $T(\mathcal{F}) = \mathcal{F}$. (It is more accurate to think of the factor as the action of T on the measure space $(X, \mathcal{F}, \mu|\mathcal{F})$.) The structure theorem asserts that there is a transfinite increasing sequence of factors $\{\mathcal{F}_\alpha\}$ of T such that the following conditions hold.

(a) $\mathcal{F}_{\alpha+1}$ is *compact* relative to $\mathcal{F}_\alpha$.
(b) $\mathcal{F}_\alpha = \bigvee_{\beta<\alpha}\mathcal{F}_\beta$ whenever α is a limit ordinal.
(c) $\mathcal{B}$ is *weakly mixing* relative to $\bigvee\mathcal{F}_\alpha$.

($\bigvee$ denotes the σ-algebra generated by a collection of σ-algebras.) $\bigvee_\alpha\mathcal{F}_\alpha$ is called the *maximal distal factor* of T and if $\bigvee\mathcal{F}_\alpha = \mathcal{B}$ then T is called *distal*. The definitions of the relative properties in (a) and (c) above are somewhat technical so only their absolute versions (i. e. relative to the trivial σ-algebra $\{\emptyset, X\}$) will be described here.

T is said to be compact if for each $f \in L_2$ the orbit $\{T^i f : i \in \mathbb{Z}\}$ is pre-compact in the norm topology of L_2. This turns out to be equivalent to the statement that T is

Copyrighted material

a translation on a compact Abelian group endowed with its Haar measure. The property of weak mixing is a fundamental notion in ergodic theory which has many equivalent definitions. The most appropriate for our purposes is that T is weakly mixing if it has no compact factors. This turns out to be equivalent to the ergodicity of $T \times T$ acting on the product measure space $(X, \mathcal{B}, \mu) \times (X, \mathcal{B}, \mu)$.

The verification of 37 in the case of compact T is rather easy. In this case it is not hard to prove that for any $f \in L_\infty$ and $\epsilon > 0$ the set $\{n \in \mathbb{Z} : \|T^n f - f\|_2 < \epsilon\}$ has bounded gaps and (34) follows easily. In the case when T is weakly mixing (34) is a consequence of the following theorem which Furstenberg proves in [85] (as a warm-up for it's much harder relative version).

Theorem 41 *If T is weakly mixing and $f_1, f_2, \ldots, f_k$ are L_∞ functions then*

$$\lim_N \frac{1}{N} \sum_{n=0}^{N-1} \int \prod_{i=1}^{k} T^{in} f_i \, d\mu = \prod_{i=1}^{k} \int f_i \tag{35}$$

Later Bergelson [26] showed that the result can be obtained easily by an induction argument using the following Hilbert space generalization of van der Corput's lemma.

Theorem 42 (Bergelson) *Suppose $\{x_n\}$ is a bounded sequence of vectors in Hilbert space such that for each $h > 0$ one has $\frac{1}{N} \sum_{n=0}^{N-1} \langle x_{n+h}, x_n \rangle \to 0$ as $N \to \infty$. Then $\| \frac{1}{N} \sum_{n=0}^{N-1} x_n \| \to 0$.*

Ryzhikov has also given a beautiful short proof of Theorem 41 using joinings ([182]). Bergelson's van der Corput lemma and variants of it have been a key tool in subsequent developments in ergodic Ramsey theory and in the convergence results to be be discussed in this section. Bergelson [26] used it to prove the following mean ergodic theorem for weakly mixing automorphisms.

Theorem 43 *Suppose T is weakly mixing, $f_1, \ldots, f_k$ are L_∞ functions and $p_1, \ldots, p_k$ are polynomials with rational coefficients taking integer values on the integers such that no $p_i - p_j$ is constant for $i \neq j$. Then*

$$\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} \prod_{i=1}^{k} T^{p_i(n)} f_i - \prod_{i=1}^{k} \int f_i \, d\mu \right\| = 0 \,. \tag{36}$$

Theorems 40 and 41 immediately raise the question of convergence of the multiple averages $\frac{1}{N} \sum_{n=0}^{N-1} \prod_{i=1}^{k} T^{in} f_i$ for a general T. Several authors obtained partial results on the question of mean convergence. It was finally resolved only recently by Host and Kra [104], who proved the following landmark theorem.

Theorem 44 *Suppose $f_1, f_2, \ldots, f_k \in L_\infty$. Then there is a $g \in L_\infty$ such that*

$$\lim \left\| \frac{1}{N} \sum_{n=0}^{N-1} \prod_{i=1}^{k} T^{in} f_i - g \right\|_2 = 0 \,. \tag{37}$$

Independently and somewhat later Ziegler [211]obtained the same result by somewhat different methods. Furstenberg had already established Theorem 44 for $k = 2$ in [85]. It was proved for $k = 3$ in the case of a totally ergodic T by Conze and Lesigne [63]and in general by Host and Kra [102]. It can also be obtained using the methods developed by Furstenberg and Weiss [88]. In this paper Furstenberg and Weiss proved a result for polynomial powers of T in the case $k = 2$. They also formalized the key notion of a *characteristic factor*. A factor C of T is said to be characteristic for the averages (37) if, roughly speaking, the L_2 limiting behavior of the averages is unchanged when any one of the f_i's is replaced by its conditional expectation on C. This means that the question of convergence of these averages may be reduced to the case when f_i are all C-measurable. So the problem is to find the right (smallest) characteristic factor and prove convergence for that factor.

The importance of characteristic factors was already apparent in Furstenberg's original paper [85], where he showed that the maximal distal factor is characteristic for the averages (37). In fact he showed that for a given k a k-step distal factor is characteristic. (An automorphism is k-step distal if it is the top rung in a k-step ladder of factors as in the Furstenberg–Zimmer structure-theorem.) It turns out, though, that the right characteristic factor for (37) is considerably smaller. In their seminal paper [63] Conze and Lesigne identified the characteristic factor for $k = 3$, now called the Conze–Lesigne factor. As shown in [104], and [211], the characteristic factor for a general k is (isomorphic to) an inverse limit of k-step nilflows. A k-step nilflow is a compact homogeneous space N/Γ of a k-step nilpotent Lie group N, endowed with its unique left-invariant probability measure, on which T acts via left translation by an element of N. Ergodic properties of nilflows have been studied for some time in ergodic theory, for example in Parry [166]. In this way the problem of L_2-convergence of (37) is reduced to the case when T is a nilflow. In this case one has more: the averages converge pointwise by a result of Leibman [141] (See also Ziegler [210]).

There have already been a good many generalizations of (37). Host and Kra [103], Frantzikinakis and Kra [82, 83], and Leibman [141] have proved results which replace

Copyrighted material

ways a sequence $\{F_n\}$ which is pointwise good and grows at most exponentially? Can one do better either in general or in particular groups?

Lindenstrauss's theorem at least guarantees the existence of Følner sequences which are pointwise good in L_1 but in particular groups there are often natural sequences which one hopes might be good. For example in $\bigoplus_{i=1}^{\infty} \mathbb{Z}$ one may take F_n to be a cube based at 0 of sidelength l_n and dimension d_n (that is, all but the first d_n co-ordinates are zero), where both sequences increase to ∞. What conditions on l_n and d_n will give a good sequence? Note that no such sequence is regular in Tempelman's sense. If $d_n = n$ then $\{l_n\}$ must be superexponential to ensure Shulman's condition. Can one do better? What about $l_n = d_n = n$?

Bibliography

1. Akcoglu M, Bellow A, Jones RL, Losert V, Reinhold–Larsson K, Wierdl M (1996) The strong sweeping out property for lacunary sequences, Riemann sums, convolution powers, and related matters. Ergodic Theory Dynam Systems 16(2):207–253
2. Akcoglu M, Jones RL, Rosenblatt JM (2000) The worst sums in ergodic theory. Michigan Math J 47(2):265–285
3. Akcoglu MA (1975) A pointwise ergodic theorem in L_p-spaces. Canad J Math 27(5):1075–1082
4. Akcoglu MA, Chacon RV (1965) A convexity theorem for positive operators. Z Wahrsch Verw Gebiete 3:328–332 (1965)
5. Akcoglu MA, Chacon RV (1970) A local ratio theorem. Canad J Math 22:545–552
6. Akcoglu MA, del Junco A (1975) Convergence of averages of point transformations. Proc Amer Math Soc 49:265–266
7. Akcoglu MA, Kopp PE (1977) Construction of dilations of positive L_p-contractions. Math Z 155(2):119–127
8. Akcoglu MA, Krengel U (1981) Ergodic theorems for superadditive processes. J Reine Angew Math 323:53–67
9. Akcoglu MA, Sucheston L (1978) A ratio ergodic theorem for superadditive processes. Z Wahrsch Verw Gebiete 44(4):269–278
10. Alaoglu L, Birkhoff G (1939) General ergodic theorems. Proc Nat Acad Sci USA 25:628–630
11. Assani I (2000) Multiple return times theorems for weakly mixing systems. Ann Inst H Poincaré Probab Statist 36(2):153–165
12. Assani I (2003) Wiener–Wintner ergodic theorems. World Scientific Publishing Co. Inc., River Edge, NJ
13. Assani I, Lesigne E, Rudolph D (1995) Wiener–Wintner return-times ergodic theorem. Israel J Math 92(1–3):375–395
14. Assani I, Buczolich Z, Mauldin RD (2005) An L^1 counting problem in ergodic theory. J Anal Math 95:221–241
15. Auslander L, Green L, Hahn F (1963) Flows on homogeneous spaces. With the assistance of Markus L, Massey W and an appendix by Greenberg L Annals of Mathematics Studies, No 53. Princeton University Press, Princeton, NJ
16. Banach S (1993) Théorie des opérations linéaires. Éditions Jacques Gabay, Sceaux, reprint of the 1932 original
17. Bellow A (1983) On "bad universal" sequences in ergodic theory II. In: Belley JM, Dubois J, Morales P (eds) Measure theory and its applications. Lecture Notes in Math, vol 1033. Springer, Berlin, pp 74–78
18. Bellow A (1999) Transference principles in ergodic theory. In: Christ M, Kenig CE, Sadowsky C (eds) Harmonic analysis and partial differential equations, Chicago Lectures in Math. Univ Chicago Press, Chicago, pp 27–39
19. Bellow A, Calderón A (1999) A weak-type inequality for convolution products. In: Christ M, Kenig CE, Sadowsky C (eds) Harmonic analysis and partial differential equations, Chicago Lectures in Math. Univ Chicago Press, Chicago, pp 41–48
20. Bellow A, Jones R (eds) (1991) Almost everywhere convergence, II. Academic Press, Boston
21. Bellow A, Losert V (1985) The weighted pointwise ergodic theorem and the individual ergodic theorem along subsequences. Trans Amer Math Soc 288(1):307–345
22. Bellow A, Jones R, Rosenblatt J (1989) Almost everywhere convergence of powers. In: Edgar GA, Sucheston L (eds) Almost everywhere convergence. Academic, Boston, pp 99–120
23. Bellow A, Jones R, Rosenblatt J (1990) Convergence for moving averages. Ergodic Theory Dynam Systems 10(1):43–62
24. Bellow A, Jones RL, Rosenblatt J (1992) Almost everywhere convergence of weighted averages. Math Ann 293(3):399–426
25. Bellow A, Jones R, Rosenblatt J (1994) Almost everywhere convergence of convolution powers. Ergodic Theory Dynam Systems 14(3):415–432
26. Bergelson V (1987) Weakly mixing PET. Ergodic Theory Dynam Systems 7(3):337–349
27. Bergelson V (1996) Ergodic Ramsey theory—an update. In: Pollicott M, Schmidt K (eds) Ergodic theory of Z^d actions. London Math Soc Lecture Note Ser, vol 228. Cambridge Univ Press, Cambridge, pp 1–61
28. Bergelson V (2006) Combinatorial and Diophantine applications of ergodic theory. In: Hasselblatt B, Katok A (eds) Handbook of dynamical systems, vol 1B, Appendix A by Leibman A, Appendix B by Quas A, Wierdl M. Elsevier, Amsterdam, pp 745–869
29. Bergelson V (2007) Some historical remarks and modern questions around the ergodic theorem. Internat Math Nachrichten 205:1–10
30. Bergelson V, Leibman A (2002) A nilpotent Roth theorem. Invent Math 147(2):429–470
31. Bergelson V, Leibman A (2004) Failure of the Roth theorem for solvable groups of exponential growth. Ergodic Theory Dynam Systems 24(1):45–53
32. Bergelson V, Leibman A (2007) Distribution of values of bounded generalized polynomials. Acta Math 198(2):155–230
33. Berkson E, Bourgain J, Gillespie TA (1991) On the almost everywhere convergence of ergodic averages for power-bounded operators on L^p-subspaces. Integral Equ Operator Theory 14(5):678–715
34. Billingsley P (1965) Ergodic theory and information. Wiley, New York
35. Birkhoff GD (1931) Proof of the ergodic theorem. Proc Nat Acad Sci USA 17:656–660
36. Bishop E (1966) An upcrossing inequality with applications. Michigan Math J 13:1–13
37. Bishop E (1967/1968) A constructive ergodic theorem. J Math Mech 17:631–639
38. Blum JR, Hanson DL (1960) On the mean ergodic theorem for subsequences. Bull Amer Math Soc 66:308–311

Copyrighted material

39. Bourgain J (1987) On pointwise ergodic theorems for arithmetic sets. C R Acad Sci Paris Sér I Math 305(10):397–402
40. Bourgain J (1988) Almost sure convergence and bounded entropy. Israel J Math 63(1):79–97
41. Bourgain J (1988) An approach to pointwise ergodic theorems. In: Lindenstrauss J, Milman VD (eds) Geometric aspects of functional analysis (1986/87). Lecture Notes in Math, vol 1317. Springer, Berlin, pp 204–223
42. Bourgain J (1988) On the maximal ergodic theorem for certain subsets of the integers. Israel J Math 61(1):39–72
43. Bourgain J (1988) On the pointwise ergodic theorem on L^p for arithmetic sets. Israel J Math 61(1):73–84
44. Bourgain J (1988) Temps de retour pour les systèmes dynamiques. C R Acad Sci Paris Sér I Math 306(12):483–485
45. Bourgain J (1988) Temps de retour pour les systèmes dynamiques. C R Acad Sci Paris Sér I Math 306(12):483–485
46. Bourgain J (1989) Almost sure convergence in ergodic theory. In: Edgar GA, Sucheston L (eds) Almost everywhere convergence. Academic, Boston, pp 145–151
47. Bourgain J (1989) Pointwise ergodic theorems for arithmetic sets. Inst Hautes Études Sci Publ Math (69):5–45, with an appendix by the author, Furstenberg H, Katznelson Y, Ornstein DS
48. Bourgain J (1990) Double recurrence and almost sure convergence. J Reine Angew Math 404:140–161
49. Breiman L (1957) The individual ergodic theorem of information theory. Ann Math Statist 28:809–811
50. Breiman L (1968) Probability. Addison–Wesley, Reading
51. Brunel A, Keane M (1969) Ergodic theorems for operator sequences. Z Wahrsch Verw Gebiete 12:231–240
52. Burkholder DL (1962) Semi–Gaussian subspaces. Trans Amer Math Soc 104:123–131
53. Calderon AP (1953) A general ergodic theorem. Ann Math (2) 58:182–191
54. Calderón AP (1968) Ergodic theory and translation-invariant operators. Proc Natl Acad Sci USA 59:349–353
55. Calderón AP (1968) Ergodic theory and translation-invariant operators. Proc Natl Acad Sci USA 59:349–353
56. Calderon AP, Zygmund A (1952) On the existence of certain singular integrals. Acta Math 88:85–139
57. Chacon RV (1962) Identification of the limit of operator averages. J Math Mech 11:961–968
58. Chacon RV (1963) Convergence of operator averages. In: Wright FB (ed) Ergodic theory. Academic, New York, pp 89–120
59. Chacon RV (1963) Linear operators in L_1. In: Wright FB (ed) Ergodic theory. Academic, New York, pp 75–87
60. Chacon RV (1964) A class of linear transformations. Proc Amer Math Soc 15:560–564
61. Chacon RV, Ornstein DS (1960) A general ergodic theorem. Illinois J Math 4:153–160
62. Conze JP, Lesigne E (1984) Théorèmes ergodiques pour des mesures diagonales. Bull Soc Math France 112(2):143–175
63. Conze JP, Lesigne E (1988) Sur un théorème ergodique pour des mesures diagonales. In: Probabilités, Publ Inst Rech Math Rennes, vol 1987. Univ Rennes I, Rennes, pp 1–31
64. Cotlar M (1955) On ergodic theorems. Math Notae 14:85–119 (1956)
65. Cotlar M (1955) A unified theory of Hilbert transforms and ergodic theorems. Rev Mat Cuyana 1:105–167 (1956)
66. Day M (1942) Ergodic theorems for Abelian semigroups. Trans Amer Math Soc 51:399–412
67. Demeter C, Lacey M, Tao T, Thiele C (2008) Breaking the duality in the return times theorem. Duke Math J 143(2):281–355
68. Derrien JM, Lesigne E (1996) Un théorème ergodique polynomial ponctuel pour les endomorphismes exacts et les K-systèmes. Ann Inst H Poincaré Probab Statist 32(6):765–778
69. Derriennic Y (2006) Some aspects of recent works on limit theorems in ergodic theory with special emphasis on the "central limit theorem". Discrete Contin Dyn Syst 15(1):143–158
70. Derriennic Y (2006) Some aspects of recent works on limit theorems in ergodic theory with special emphasis on the "central limit theorem". Discrete Contin Dyn Syst 15(1):143–158
71. Doob JL (1938) Stochastic processes with an integral-valued parameter. Trans Amer Math Soc 44(1):87–150
72. Dunford N (1951) An individual ergodic theorem for noncommutative transformations. Acta Sci Math Szeged 14:1–4
73. Dunford N, Schwartz J (1955) Convergence almost everywhere of operator averages. Proc Natl Acad Sci USA 41:229–231
74. Dunford N, Schwartz JT (1988) Linear operators, Part I. Wiley Classics Library, Wiley, New York. General theory, with the assistance of Bade WG, Bartle RG, Reprint of the 1958 original, A Wiley–Interscience Publication
75. Durand S, Schneider D (2003) Random ergodic theorems and regularizing random weights. Ergodic Theory Dynam Systems 23(4):1059–1092
76. Eberlein WF (1949) Abstract ergodic theorems and weak almost periodic functions. Trans Amer Math Soc 67:217–240
77. Edgar G, Sucheston L (eds) (1989) Almost everywhere convergence. Academic Press, Boston
78. Emerson WR (1974) The pointwise ergodic theorem for amenable groups. Amer J Math 96:472–487
79. Feldman J (2007) A ratio ergodic theorem for commuting, conservative, invertible transformations with quasi-invariant measure summed over symmetric hypercubes. Ergodic Theory Dynam Systems 27(4):1135–1142
80. Foguel SR (1969) The ergodic theory of Markov processes. In: Van Nostrand Mathematical Studies, No 21. Van Nostrand Reinhold, New York
81. Frantzikinakis N, Kra B (2005) Convergence of multiple ergodic averages for some commuting transformations. Ergodic Theory Dynam Systems 25(3):799–809
82. Frantzikinakis N, Kra B (2005) Polynomial averages converge to the product of integrals. Israel J Math 148:267–276
83. Frantzikinakis N, Kra B (2006) Ergodic averages for independent polynomials and applications. J London Math Soc (2) 74(1):131–142
84. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. Math Systems Theory 1:1–49
85. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. J Analyse Math 31:204–256
86. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. Princeton University Press, Princeton, NJ, m B Porter Lectures
87. Furstenberg H, Kesten H (1960) Products of random matrices. Ann Math Statist 31:457–469

Copyrighted material

88. Furstenberg H, Weiss B (1996) A mean ergodic theorem for $(1/N)\sum_{n=1}^{N} f(T^n x)g(T^{n^2}x)$. In: Bergelson V, March P, Rosenblatt J (eds) Convergence in ergodic theory and probability. Ohio State Univ Math Res Inst Publ, vol 5. de Gruyter, Berlin, pp 193–227
89. Garsia AM (1970) Topics in almost everywhere convergence. In: Lectures in Advanced Mathematics, vol 4. Markham, Chicago
90. Glasner E (2003) Ergodic theory via joinings, Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence
91. Gowers WT (2001) A new proof of Szemerédi's theorem. Geom Funct Anal 11(3):465–588
92. Green B, Tao T (2004) The primes contain arbitrarily large arithmetic progressions. http://arxivorg/abs/mathNT/0404188
93. Greenleaf FP (1973) Ergodic theorems and the construction of summing sequences in amenable locally compact groups. Comm Pure Appl Math 26:29–46
94. Greenleaf FP, Emerson WR (1974) Group structure and the pointwise ergodic theorem for connected amenable groups. Adv Math 14:153–172
95. Guivarc'h Y (1969) Généralisation d'un théorème de von Neumann. C R Acad Sci Paris Sér A-B 268:A1020–A1023
96. Halmos PR (1946) An ergodic theorem. Proc Natl Acad Sci USA 32:156–161
97. Halmos PR (1949) A nonhomogeneous ergodic theorem. Trans Amer Math Soc 66:284–288
98. Halmos PR (1960) Lectures on ergodic theory. Chelsea, New York
99. Hammersley JM, Welsh DJA (1965) First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory. In: Proc Internat Res Semin. Statist Lab, University of California, Berkeley. Springer, New York, pp 61–110
100. Hopf E (1937) Ergodentheorie. In: Ergebnisse der Mathematik und ihrer Grenzgebiete, vol 5. Springer, Berlin
101. Hopf E (1954) The general temporally discrete Markoff process. J Rational Mech Anal 3:13–45
102. Host B, Kra B (2001) Convergence of Conze–Lesigne averages. Ergodic Theory Dynam Systems 21(2):493–509
103. Host B, Kra B (2005) Convergence of polynomial ergodic averages. Israel J Math 149:1–19
104. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. Ann Math (2) 161(1):397–488
105. Hurewicz W (1944) Ergodic theorem without invariant measure. Ann Math (2) 45:192–206
106. Tulcea AI (1964) Ergodic preperties of positive isometries. Bull AMS 70:366–371
107. Ivanov VV (1996) Geometric properties of monotone functions and the probabilities of random oscillations. Sibirsk Mat Zh 37(1):117–15
108. Ivanov VV (1996) Oscillations of averages in the ergodic theorem. Dokl Akad Nauk 347(6):736–738
109. Jewett RI (1969/1970) The prevalence of uniquely ergodic systems. J Math Mech 19:717–729
110. Jones R, Rosenblatt J, Tempelman A (1994) Ergodic theorems for convolutions of a measure on a group. Illinois J Math 38(4):521–553
111. Jones RL (1987) Necessary and sufficient conditions for a maximal ergodic theorem along subsequences. Ergodic Theory Dynam Systems 7(2):203–210
112. Jones RL (1993) Ergodic averages on spheres. J Anal Math 61:29–45
113. Jones RL, Wierdl M (1994) Convergence and divergence of ergodic averages. Ergodic Theory Dynam Systems 14(3):515–535
114. Jones RL, Olsen J, Wierdl M (1992) Subsequence ergodic theorems for L^p contractions. Trans Amer Math Soc 331(2):837–850
115. Jones RL, Kaufman R, Rosenblatt JM, Wierdl M (1998) Oscillation in ergodic theory. Ergodic Theory Dynam Systems 18(4):889–935
116. Jones RL, Lacey M, Wierdl M (1999) Integer sequences with big gaps and the pointwise ergodic theorem. Ergodic Theory Dynam Systems 19(5):1295–1308
117. Jones RL, Rosenblatt JM, Wierdl M (2001) Oscillation inequalities for rectangles. Proc Amer Math Soc 129(5):1349–1358 (electronic)
118. Jones RL, Rosenblatt JM, Wierdl M (2003) Oscillation in ergodic theory: higher dimensional results. Israel J Math 135:1–27
119. del Junco A, Rosenblatt J (1979) Counterexamples in ergodic theory and number theory. Math Ann 245(3):185–197
120. Kac M (1947) On the notion of recurrence in discrete stochastic processes. Bull Amer Math Soc 53:1002–1010
121. Kachurovskii AG (1996) Rates of convergence in ergodic theorems. Uspekhi Mat Nauk 51(4(310)):73–124
122. Kachurovskii AG (1996) Spectral measures and convergence rates in the ergodic theorem. Dokl Akad Nauk 347(5):593–596
123. Kakutani S (1940) Ergodic theorems and the Markoff process with a stable distribution. Proc Imp Acad Tokyo 16:49–54
124. Kalikow S, Weiss B (1999) Fluctuations of ergodic averages. In: Proceedings of the Conference on Probability, Ergodic Theory, and Analysis, vol 43. pp 480–488
125. Kamae T (1982) A simple proof of the ergodic theorem using nonstandard analysis. Israel J Math 42(4):284–290
126. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems, Encyclopedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge, with a supplementary chapter by Katok A and Mendoza L
127. Katznelson Y, Weiss B (1982) A simple proof of some ergodic theorems. Israel J Math 42(4):291–296
128. Kieffer JC (1975) A generalized Shannon–McMillan theorem for the action of an amenable group on a probability space. Ann Probability 3(6):1031–1037
129. Kingman JFC (1968) The ergodic theory of subadditive stochastic processes. J Roy Statist Soc Ser B 30:499–510
130. Kingman JFC (1976) Subadditive processes. In: École d'Été de Probabilités de Saint–Flour, V–1975. Lecture Notes in Math, vol 539. Springer, Berlin, pp 167–223
131. Koopman B (1931) Hamiltonian systems and transformations in hilbert spaces. Proc Natl Acad Sci USA 17:315–318
132. Krantz SG, Parks HR (1999) The geometry of domains in space. Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Boston Inc, Boston, MA
133. Krengel U (1971) On the individual ergodic theorem for subsequences. Ann Math Statist 42:1091–1095

Copyrighted material

134. Krengel U (1978/79) On the speed of convergence in the ergodic theorem. Monatsh Math 86(1):3–6
135. Krengel U (1985) Ergodic theorems. In: de Gruyter studies in mathematics, vol 6. de Gruyter, Berlin, with a supplement by Antoine Brunel
136. Krengel U, Lin M, Wittmann R (1990) A limit theorem for order preserving nonexpansive operators in L_1. Israel J Math 71(2):181–191
137. Krieger W (1972) On unique ergodicity. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol II. Probability Theory, University of California Press, Berkeley, pp 327–346
138. Kryloff N, Bogoliouboff N (1937) La théorie générale de la mesure dans son application à l'étude des systèmes dynamiques de la mécanique non linéaire. Ann Math (2) 38(1):65–113
139. Kuipers L, Niederreiter H (1974) Uniform distribution of sequences. Wiley, New York, Pure and Applied Mathematics
140. Lamperti J (1958) On the isometries of certain function-spaces. Pacific J Math 8:459–466
141. Leibman A (2005) Convergence of multiple ergodic averages along polynomials of several variables. Israel J Math 146:303–315
142. Leibman A (2005) Pointwise convergence of ergodic averages for polynomial actions of $\mathbb{Z}^d$ by translations on a nil-manifold. Ergodic Theory Dynam Systems 25(1):215–225
143. Lemańczyk M, Lesigne E, Parreau F, Volný D, Wierdl M
144. Lesigne E (1989) Théorèmes ergodiques pour une translation sur un nilvariété. Ergodic Theory Dynam Systems 9(1):115–126
145. Lindenstrauss E (1999) Pointwise theorems for amenable groups. Electron Res Announc Amer Math Soc 5:82–90
146. Loomis LH (1946) A note on the Hilbert transform. Bull Amer Math Soc 52:1082–1086
147. Lorch ER (1939) Means of iterated transformations in reflexive vector spaces. Bull Amer Math Soc 45:945–947
148. Mauldin D, Buczolich Z (2005) Concepts behind divergent ergodic averages along the squares. In: Assani I (ed) Ergodic theory and related fields. Contemp Math, vol 430. Amer Math Soc, Providence, pp 41–56
149. McMillan B (1953) The basic theorems of information theory. Ann Math Statistics 24:196–219
150. Merlevède F, Peligrad M, Utev S (2006) Recent advances in invariance principles for stationary sequences. Probab Surv 3:1–36
151. von Neumann J (1932) Proof of the quasi-ergodic hypothesis. Proc Natl Acad Sci USA 18:70–82
152. Neveu J (1961) Sur le théorème ergodique ponctuel. C R Acad Sci Paris 252:1554–1556
153. Neveu J (1965) Mathematical foundations of the calculus of probability. Translated by Amiel Feinstein, Holden-Day, San Francisco
154. Nevo A (1994) Harmonic analysis and pointwise ergodic theorems for noncommuting transformations. J Amer Math Soc 7(4):875–902
155. Nevo A (2006) Pointwise ergodic theorems for actions of groups. In: Hasselblatt B, Katok A (eds) Handbook of dynamical systems vol 1B. Elsevier, Amsterdam, pp 871–982
156. Nevo A, Stein EM (1994) A generalization of Birkhoff's pointwise ergodic theorem. Acta Math 173(1):135–154
157. Nguyen XX (1979) Ergodic theorems for subadditive spatial processes. Z Wahrsch Verw Gebiete 48(2):159–176
158. Orey S (1971) Lecture notes on limit theorems for Markov chain transition probabilities. In: Van Nostrand Reinhold Mathematical Studies, no 34. Van Nostrand Reinhold Co, London
159. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352 (1970)
160. Ornstein D (1971) A remark on the Birkhoff ergodic theorem. Illinois J Math 15:77–79
161. Ornstein D, Weiss B (1983) The Shannon–McMillan–Breiman theorem for a class of amenable groups. Israel J Math 44(1):53–60
162. Ornstein D, Weiss B (1992) Subsequence ergodic theorems for amenable groups. Israel J Math 79(1):113–127
163. Ornstein DS (1960) On invariant measures. Bull Amer Math Soc 66:297–300
164. Oseledec VI (1968) A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems. Trudy Moskov Mat Obšč 19:179–210
165. Oxtoby JC (1952) Ergodic sets. Bull Amer Math Soc 58:116–136
166. Parry W (1969) Ergodic properties of affine transformations and flows on nilmanifolds. Amer J Math 91:757–771
167. Paterson A (1988) Amenability, Mathematical Surveys and Monographs, vol 29. American Mathematical Society, Providence, RI
168. Peck JEL (1951) An ergodic theorem for a noncommutative semigroup of linear operators. Proc Amer Math Soc 2:414–421
169. Pesin JB (1977) Characteristic Ljapunov exponents, and smooth ergodic theory. Uspehi Mat Nauk 32(4(196)):55–112,287
170. Petersen K (1983) Another proof of the existence of the ergodic Hilbert transform. Proc Amer Math Soc 88(1):39–43
171. Petersen K (1989) Ergodic theory, Cambridge Studies in Advanced Mathematics, vol 2. Cambridge University Press, Cambridge
172. Pitt HR (1942) Some generalizations of the ergodic theorem. Proc Cambridge Philos Soc 38:325–343
173. Poincaré H (1987) Les méthodes nouvelles de la mécanique céleste. Tome I, II, III. Les Grands Classiques Gauthier–Villars. Librairie Scientifique et Technique Albert Blanchard, Paris
174. Raghunathan MS (1979) A proof of Oseledec's multiplicative ergodic theorem. Israel J Math 32(4):356–362
175. Renaud PF (1971) General ergodic theorems for locally compact groups. Amer J Math 93:52–64
176. Rosenblatt JM, Wierdl M (1995) Pointwise ergodic theorems via harmonic analysis. In: Peterson KE, Salama IA (eds) Ergodic theory and its connections with harmonic analysis, London Math Soc Lecture Note Ser, vol 205. Cambridge University Press, Cambridge, pp 3–151
177. Rudolph DJ (1990) Fundamentals of measurable dynamics. In: Fundamentals of measurable dynamics: Ergodic theory on Lebesgue spaces. Oxford Science Publications, The Clarendon Press, Oxford University Press, New York
178. Rudolph DJ (1994) A joinings proof of Bourgain's return time theorem. Ergodic Theory Dynam Systems 14(1):197–203
179. Rudolph DJ (1998) Fully generic sequences and a multiple-term return-times theorem. Invent Math 131(1):199–228
180. Ruelle D (1982) Characteristic exponents and invariant manifolds in Hilbert space. Ann Math (2) 115(2):243–290

Copyrighted material