The Princeton Companion to Applied Mathematics

Copyrighted material

The Princeton Companion to Applied Mathematics

EDITOR

Nicholas J. Higham
The University of Manchester

ASSOCIATE EDITORS

Mark R. Dennis
University of Bristol

Paul Glendinning
The University of Manchester

Paul A. Martin
Colorado School of Mines

Fadil Santosa
University of Minnesota

Jared Tanner
University of Oxford

Princeton University Press
Princeton and Oxford

Copyrighted material

Copyright © 2015 by Princeton University Press

Published by Princeton University Press,
41 William Street, Princeton, New Jersey 08540

In the United Kingdom: Princeton University Press,
6 Oxford Street, Woodstock, Oxfordshire OX20 1TW

press.princeton.edu

Jacket image courtesy of iStock

All Rights Reserved

Library of Congress Cataloging-in-Publication Data

The Princeton companion to applied mathematics / editor,
Nicholas J. Higham, The University of Manchester ;
associate editors, Mark R. Dennis, University of Bristol
[and four others].
pages cm
Includes bibliographical references and index.
ISBN 978-0-691-15039-0 (hardcover : alk. paper)
1. Algebra. 2. Mathematics. 3. Mathematical models.
I. Higham, Nicholas J., 1961– editor. II. Dennis, Mark R.,
editor. III. Title: Companion to applied mathematics.
IV. Title: Applied mathematics.

QA155.P75 2015
510—dc23 2015013024

British Library Cataloging-in-Publication Data is available

This book has been composed in LucidaBright

Project management, composition and copyediting
by T&T Productions Ltd, London

Printed on acid-free paper ♾
Printed in the United States of America

1 2 3 4 5 6 7 8 9 10

Copyrighted material

Contents

Copyrighted material

Part IV Areas of Applied Mathematics

Part V Modeling

Part VI Example Problems

Copyrighted material

Part VII Application Areas

Part VIII Final Perspectives

Copyrighted material

Copyrighted material

Preface

1 What Is *The Companion*?

The Princeton Companion to Applied Mathematics describes what applied mathematics is about, why it is important, its connections with other disciplines, and some of the main areas of current research. It also explains what applied mathematicians do, which includes not only studying the subject itself but also writing about mathematics, teaching it, and influencing policy makers.

The Companion differs from an encyclopedia in that it is not an exhaustive treatment of the subject, and it differs from a handbook in that it does not cover all relevant methods and techniques. Instead, the aim is to offer a broad but selective coverage that conveys the excitement of modern applied mathematics while also giving an appreciation of its history and the outstanding challenges. *The Companion* focuses on topics felt by the editors to be of enduring interest, and so it should remain relevant for many years to come.

With online sources of information about mathematics growing ever more extensive, one might ask what role a printed volume such as this has. Certainly, one can use Google to search for almost any topic in the book and find relevant material, perhaps on Wikipedia. What distinguishes *The Companion* is that it is a self-contained, structured reference work giving a consistent treatment of the subject. The content has been curated by an editorial board of applied mathematicians with a wide range of interests and experience, the articles have been written by leading experts and have been rigorously edited and copyedited, and the whole volume is thoroughly cross-referenced and indexed.

Within each article, the authors and editors have tried hard to convey the motivation for each topic or concept and the basic ideas behind it, while avoiding unnecessary detail. It is hoped that *The Companion* will be seen as a friendly and inspiring reference, containing both standard material and more unusual, novel, or unexpected topics.

2 Scope

It is difficult to give a precise definition of applied mathematics, as discussed in WHAT IS APPLIED MATHEMATICS? [I.1] and, from a historical perspective, in THE HISTORY OF APPLIED MATHEMATICS [I.6]. *The Companion* treats applied mathematics in a broad sense, and it cannot cover all aspects in equal depth. Some parts of mathematical physics are included, though a full treatment of modern fundamental theories is not given. Statistics and probability are not explicitly included, although a number of articles make use of ideas from these subjects, and in particular the burgeoning area of UNCERTAINTY QUANTIFICATION [II.34] brings together many ideas from applied mathematics and statistics. Applied mathematics increasingly makes use of algorithms and computation, and a number of aspects at the interface with computer science are included. Some parts of discrete and combinatorial mathematics are also covered.

3 Audience

The target audience for *The Companion* is mathematicians at undergraduate level or above; students, researchers, and professionals in other subjects who use mathematics; and mathematically interested lay readers. Some articles will also be accessible to students studying mathematics at pre-university level.

Prospective research students might use the book to obtain some idea of the different areas of applied mathematics that they could work in. Researchers who regularly attend seminars in areas outside their own specialities should find that the articles provide a gentle introduction to some of these areas, making good pre- or post-seminar reading.

In soliciting and editing the articles the editors aimed to maximize accessibility by keeping discussions at the lowest practical level. A good question is how much of the book a reader should expect to understand. Of course “understanding” is an imprecisely defined

Copyrighted material

concept. It is one thing to read along with an argument and find it plausible, or even convincing, but another to reproduce it on a blank piece of paper, as every undergraduate discovers at exam time. The very wide range of topics covered means that it would take a reader with an unusually broad knowledge to understand everything, but every reader from undergraduate level upward should find a substantial portion of the book accessible.

4 Organization

The Companion is organized in eight parts, which are designed to cut across applied mathematics in different ways.

Part I, "Introduction to Applied Mathematics," begins by discussing what applied mathematics is and giving examples of the use of applied mathematics in everyday life. THE LANGUAGE OF APPLIED MATHEMATICS [I.2] then presents basic definitions, notation, and concepts that are needed frequently in other parts of the book, essentially giving a brief overview of some key parts of undergraduate mathematics. This article is not meant to be a complete survey, and many later articles provide other introductory material themselves. METHODS OF SOLUTION [I.3] describes some general solution techniques used in applied mathematics. ALGORITHMS [I.4] explains the concept of an algorithm, giving some important examples and discussing complexity issues. The presence of this article in part I reflects the increasing importance of algorithms in all areas of applied mathematics. GOALS OF APPLIED MATHEMATICAL RESEARCH [I.5] describes the kinds of questions and issues that research in applied mathematics addresses and discusses some strategic aspects of carrying out research. Finally, THE HISTORY OF APPLIED MATHEMATICS [I.6] describes the history of the subject from ancient times up until the late twentieth century.

Part II, "Concepts," comprises short articles that explain specific concepts and their significance. These are mainly concepts that cut across different models and areas and provide connections to other parts of the book. This part is not meant to be comprehensive, and many other concepts are well described in later articles (and discoverable via the index).

Part III, "Equations, Laws, and Functions of Applied Mathematics," treats important examples of what its title describes. The choice of what to include was based on a mix of importance, accessibility, and interest. Many equations, laws, and functions not contained in this part are included in other articles.

Part IV, "Areas of Applied Mathematics," contains longer articles giving an overview of the whole subject and how it is organized, arranged by research area. The aim of this part is to convey the breadth, depth, and diversity of applied mathematics research. The coverage is not comprehensive, but areas that do not appear as or in article titles may nevertheless be present in other articles. For example, there is no article on geoscience, yet EARTH SYSTEM DYNAMICS [IV.30], INVERSE PROBLEMS [IV.15], and IMAGING THE EARTH USING GREEN'S THEOREM [VII.16] all cover specific aspects of this area. Nor is there a part IV article on numerical analysis, but this area is represented by APPROXIMATION THEORY [IV.9], NUMERICAL LINEAR ALGEBRA AND MATRIX ANALYSIS [IV.10], CONTINUOUS OPTIMIZATION (NONLINEAR AND LINEAR PROGRAMMING) [IV.11], NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS [IV.12], and NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS [IV.13].

Part V, "Modeling," gives a selection of mathematical models, explaining how the models are derived and how they are solved.

Part VI, "Example Problems," contains short articles covering a variety of interesting applied mathematics problems.

Part VII, "Application Areas," comprises articles on connections between applied mathematics and other disciplines, including such diverse topics as integrated circuit (chip) design, medical imaging, and the screening of luggage in airports.

Part VIII, "Final Perspectives," contains essays on broader aspects, including reading, writing, and typesetting mathematics; teaching applied mathematics; and how to influence government as a mathematician.

The articles within a given part vary significantly in length. This should not be taken as an indication of the importance of the corresponding topic, as it is partly due to the number of pages that could be allocated to each article, as well as to how authors responded to their given page limit.

The ordering of articles within a part is alphabetical for parts II and III. For part IV some attempt was made to place related articles together and to place one article before another if there is a natural order in which to read the two articles. The ordering is nevertheless somewhat arbitrary, and the reader should feel free to read the articles in any order. The articles within parts V–VIII are arranged only loosely by theme.

Copyrighted material

Contributors

David Acheson, *Emeritus Fellow, Jesus College, University of Oxford*
INVERTED PENDULUMS [VI.4],
TEACHING APPLIED MATHEMATICS [VIII.7]

Miguel A. Alonso, *Associate Professor, The Institute of Optics at the University of Rochester*
MODERN OPTICS [V.14]

Douglas N. Arnold, *McKnight Presidential Professor of Mathematics, University of Minnesota*
THE FLIGHT OF A GOLF BALL [VI.6]

Karl Johan Åström, *Emeritus Professor, Department of Automatic Control, Lund Institute of Technology/University of Lund*
CONTROL THEORY [IV.34]

David H. Bailey, *Lawrence Berkeley National Laboratory (retired); Research Fellow, University of California, Davis*
EXPERIMENTAL APPLIED MATHEMATICS [VIII.6]

June Barrow-Green, *Senior Lecturer in the History of Mathematics, The Open University*
THE HISTORY OF APPLIED MATHEMATICS [I.6]

Peter Benner, *Director, Max Planck Institute for Dynamics of Complex Technical Systems*
MODEL REDUCTION [II.26]

Andrew J. Bernoff, *Kenneth and Diana Jonsson Professor of Mathematics, Harvey Mudd College*
THE THIN-FILM EQUATION [III.29]

Michael V. Berry, *Melville Wills Professor of Physics (Emeritus), University of Bristol*
DIVERGENT SERIES: TAMING THE TAILS [V.8]

Michael W. Berry, *Professor, Department of Electrical Engineering and Computer Science, University of Tennessee*
TEXT MINING [VII.24]

Brett Borden, *Professor of Physics, The Naval Postgraduate School, Monterey, California*
RADAR IMAGING [VII.17]

Jeffrey T. Borggaard, *Professor of Mathematics, Virginia Tech*
OPTIMAL SENSOR LOCATION IN THE CONTROL OF ENERGY-EFFICIENT BUILDINGS [VI.13]

Jonathan M. Borwein, *Laureate Professor, School of Mathematical and Physical Sciences, University of Newcastle, Australia*
EXPERIMENTAL APPLIED MATHEMATICS [VIII.6]

Fred Brauer, *Professor Emeritus of Mathematics, University of Wisconsin–Madison*
THE SPREAD OF INFECTIOUS DISEASES [V.16]

Thomas J. Brennan, *Professor of Law, Harvard Law School*
PORTFOLIO THEORY [V.10]

David S. Broomhead, *Professor of Applied Mathematics, The University of Manchester (deceased)*
APPLICATIONS OF MAX-PLUS ALGEBRA [VII.4]

Kurt Bryan, *Professor of Mathematics, Rose–Hulam Institute of Technology*
CLOAKING [VI.1]

Dorothy Buck, *Reader in BioMathematics, Imperial College London*
KNOTTING AND LINKING OF MACROMOLECULES [VI.8]

Chris Budd, *Professor of Applied Mathematics, University of Bath; Professor of Mathematics, Royal Institution of Great Britain*
SLIPPING, SLIDING, RATTLING, AND IMPACT: NONSMOOTH DYNAMICS AND ITS APPLICATIONS [VI.15]

John A. Burns, *Hatcher Professor of Mathematics and Technical Director for the Interdisciplinary Center for Applied Mathematics, Virginia Tech*
OPTIMAL SENSOR LOCATION IN THE CONTROL OF ENERGY-EFFICIENT BUILDINGS [VI.13]

Daniela Calvetti, *The James Wood Williamson Professor, Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University*
DIMENSIONAL ANALYSIS AND SCALING [II.9]

Eric Cancès, *Professor of Analysis, Ecole des Ponts and INRIA*
ELECTRONIC STRUCTURE CALCULATIONS (SOLID STATE PHYSICS) [VII.14]

René Carmona, *Paul M. Wythes '55 Professor of Engineering and Finance, Bendheim Center for Finance, ORFE, Princeton University*
FINANCIAL MATHEMATICS [V.9]

C. J. Chapman, *Professor of Applied Mathematics, University of Keele*
SHOCK WAVES [V.20], AIRCRAFT NOISE [VII.1]

S. Jonathan Chapman, *Professor of Mathematics and Its Applications, University of Oxford*
THE GINZBURG–LANDAU EQUATION [III.14]

Gui-Qiang G. Chen, *Statutory Professor in the Analysis of Partial Differential Equations and Professorial Fellow of Keble College, University of Oxford*
THE TRICOMI EQUATION [III.30]

Margaret Cheney, *Professor of Mathematics and Albert C. Yates Endowment Chair, Colorado State University*
RADAR IMAGING [VII.17]

Copyrighted material

Peter A. Clarkson, *Professor of Mathematics, University of Kent*
THE PAINLEVÉ EQUATIONS [III.24]

Eugene M. Cliff, *Professor Emeritus, Interdisciplinary Center for Applied Mathematics, Virginia Tech*
OPTIMAL SENSOR LOCATION IN THE CONTROL OF ENERGY-EFFICIENT BUILDINGS [VI.13]

Paul G. Constantine, *Ben L. Fryrear Assistant Professor of Applied Mathematics and Statistics, Colorado School of Mines*
RANKING WEB PAGES [VI.9]

William Cook, *Professor of Combinatorics and Optimization, University of Waterloo*
THE TRAVELING SALESMAN PROBLEM [VI.18]

Robert M. Corless, *Distinguished University Professor, Department of Applied Mathematics, The University of Western Ontario*
THE LAMBERT W FUNCTION [III.17]

Darren Crowdy, *Professor of Applied Mathematics, Imperial College London*
CONFORMAL MAPPING [II.5]

James M. Crowley, *Executive Director, Society for Industrial and Applied Mathematics*
MATHEMATICS AND POLICY [VIII.9]

Annie Cuyt, *Professor, Department of Mathematics & Computer Science, University of Antwerp*
APPROXIMATION THEORY [IV.9]

E. Brian Davies, *Emeritus Professor of Mathematics, King's College London*
SPECTRAL THEORY [IV.8]

Timothy A. Davis, *Professor, Department of Computer Science and Engineering, Texas A&M University*
GRAPH THEORY [II.16], SEARCHING A GRAPH [VI.10]

Florent de Dinechin, *Professor of Applied Sciences, INSA—Lyon*
EVALUATING ELEMENTARY FUNCTIONS [VI.11]

Mark R. Dennis, *Professor of Theoretical Physics, University of Bristol*
INVARIANTS AND CONSERVATION LAWS [II.21], TENSORS AND MANIFOLDS [II.33], THE DIRAC EQUATION [III.9], MAXWELL'S EQUATIONS [III.22], SCHRÖDINGER'S EQUATION [III.26]

Jack Dongarra, *Professor, University of Tennessee; Professor, Oak Ridge National Laboratory; Professor, The University of Manchester*
HIGH-PERFORMANCE COMPUTING [VII.12]

David L. Donoho, *Anne T. and Robert M. Bass Professor in the Humanities and Sciences, Stanford University*
REPRODUCIBLE RESEARCH IN THE MATHEMATICAL SCIENCES [VIII.5]

Ivar Ekeland, *Professor Emeritus, CEREMADE and Institut de Finance, Université Paris-Dauphine*
MATHEMATICAL ECONOMICS [VII.20]

Yonina C. Eldar, *Professor of Electrical Engineering, Technion—Israel Institute of Technology, Haifa*
COMPRESSED SENSING [VII.10]

George F. R. Ellis, *Professor Emeritus, Mathematics Department, University of Cape Town*
GENERAL RELATIVITY AND COSMOLOGY [IV.40]

Charles L. Epstein, *Thomas A. Scott Professor of Mathematics, University of Pennsylvania*
MEDICAL IMAGING [VII.9]

Bard Ermentrout, *Distinguished University Professor of Computational Biology and Professor of Mathematics, University of Pittsburgh*
MATHEMATICAL NEUROSCIENCE [VII.21]

Maria Esteban, *Director of Research, CNRS*
MATHEMATICS AND POLICY [VIII.9]

Lawrence C. Evans, *Professor, Department of Mathematics, University of California, Berkeley*
PARTIAL DIFFERENTIAL EQUATIONS [IV.3]

Hans G. Feichtinger, *Faculty of Mathematics, University of Vienna*
FUNCTION SPACES [II.15]

Martin Feinberg, *Morrow Professor of Chemical & Biomolecular Engineering and Professor of Mathematics, The Ohio State University*
CHEMICAL REACTIONS [V.7]

Alistair D. Fitt, *Vice-Chancellor, Oxford Brookes University*
MATHEMATICS AND POLICY [VIII.9]

Irene Fonseca, *Mellon College of Science University Professor of Mathematics and Director of Center for Nonlinear Analysis, Carnegie Mellon University*
CALCULUS OF VARIATIONS [IV.6]

L. B. Freund, *Adjunct Professor, Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign*
MECHANICS OF SOLIDS [IV.32]

David F. Gleich, *Assistant Professor of Computer Science, Purdue University*
RANKING WEB PAGES [VI.9]

Paul Glendinning, *Professor of Applied Mathematics, The University of Manchester*
CHAOS AND ERGODICITY [II.3], COMPLEX SYSTEMS [II.4], HYBRID SYSTEMS [II.18], THE EULER–LAGRANGE EQUATIONS [III.12], THE LOGISTIC EQUATION [III.19], THE LORENZ EQUATIONS [III.20], BIFURCATION THEORY [IV.21]

Joe D. Goddard, *Professor of Applied Mechanics and Engineering Science, University of California, San Diego*
GRANULAR FLOWS [V.13]

Kenneth M. Golden, *Professor of Mathematics/Adjunct Professor of Bioengineering, University of Utah*
THE MATHEMATICS OF SEA ICE [V.17]

Timothy Gowers, *Royal Society Research Professor, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge*
MATHEMATICAL WRITING [VIII.1]

Thomas A. Grandine, *Senior Technical Fellow, The Boeing Company*
A HYBRID SYMBOLIC–NUMERIC APPROACH TO GEOMETRY PROCESSING AND MODELING [VII.2]

Andreas Griewank, *Professor of Mathematics, Humboldt University of Berlin*
AUTOMATIC DIFFERENTIATION [VI.7]

David Griffiths, *Emeritus Professor of Physics, Reed College*
QUANTUM MECHANICS [IV.23]

Copyrighted material

Peter Grindrod, *Professor of Mathematics, University of Oxford*
EVOLVING SOCIAL NETWORKS, ATTITUDES, AND BELIEFS—AND COUNTERTERRORISM [VII.5]

Julio C. Gutiérrez-Vega, *Director of the Optics Center, Technológico de Monterrey*
MATHIEU FUNCTIONS [III.21]

Ernst Hairer, *Honorary Professor of Mathematics, University of Geneva*
NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS [IV.12]

Ian Hawke, *Associate Professor, Mathematical Sciences, University of Southampton*
NUMERICAL RELATIVITY [V.15]

Stephan Held, *Professor, Research Institute for Discrete Mathematics, Bonn University*
CHIP DESIGN [VII.6]

Didier Henrion, *Professor, LAAS-CNRS, University of Toulouse; Professor, Faculty of Electrical Engineering, Czech Technical University in Prague*
CONVEXITY [II.8]

Willy A. Hereman, *Professor of Applied Mathematics, Colorado School of Mines*
THE KORTEWEG–DE VRIES EQUATION [III.16]

Desmond J. Higham, *1966 Professor of Numerical Analysis, University of Strathclyde*
BAYESIAN INFERENCE IN APPLIED MATHEMATICS [V.11]

Nicholas J. Higham, *Richardson Professor of Applied Mathematics, The University of Manchester*
WHAT IS APPLIED MATHEMATICS? [I.1], THE LANGUAGE OF APPLIED MATHEMATICS [I.2], METHODS OF SOLUTION [I.3], ALGORITHMS [I.4], GOALS OF APPLIED MATHEMATICAL RESEARCH [I.5], CONTROL [II.7], FINITE DIFFERENCES [II.11], THE FINITE-ELEMENT METHOD [II.12], FLOATING-POINT ARITHMETIC [II.13], FUNCTIONS OF MATRICES [II.14], INTEGRAL TRANSFORMS AND CONVOLUTION [II.19], THE JORDAN CANONICAL FORM [II.22], ORTHOGONAL POLYNOMIALS [II.29], THE SINGULAR VALUE DECOMPOSITION [II.32], VARIATIONAL PRINCIPLE [II.35], THE BLACK–SCHOLES EQUATION [III.3], THE SYLVESTER AND LYAPUNOV EQUATIONS [III.28], NUMERICAL LINEAR ALGEBRA AND MATRIX ANALYSIS [IV.10], COLOR SPACES AND DIGITAL IMAGING [VII.7], PROGRAMMING LANGUAGES: AN APPLIED MATHEMATICS VIEW [VII.11], HOW TO READ AND UNDERSTAND A PAPER [VIII.2], WORKFLOW [VIII.4]

Theodore P. Hill, *Professor Emeritus of Mathematics, Georgia Institute of Technology*
BENFORD'S LAW [III.1]

Philip Holmes, *Eugene Higgins Professor of Mechanical and Aerospace Engineering and Professor of Applied and Computational Mathematics, Princeton University*
DYNAMICAL SYSTEMS [IV.20]

Stefan Hougardy, *Professor of Mathematics, University of Bonn*
CHIP DESIGN [VII.6]

Christopher J. Howls, *Professor of Mathematics, University of Southampton*
DIVERGENT SERIES: TAMING THE TAILS [V.8]

Yifan Hu, *Principal Research Scientist, Yahoo Labs*
GRAPH THEORY [II.16]

David W. Hughes, *Professor of Applied Mathematics, University of Leeds*
MAGNETOHYDRODYNAMICS [IV.29]

Julian C. R. Hunt, *Emeritus Professor of Climate Modelling and Honorary Professor of Mathematics, University College London*
TURBULENCE [V.21]

Stefan Hutzler, *Associate Professor, School of Physics, Trinity College Dublin*
FOAMS [VI.3]

Richard D. James, *Professor, Department of Aerospace Engineering and Mechanics, University of Minnesota*
CONTINUUM MECHANICS [IV.26]

David J. Jeffrey, *Professor, Department of Applied Mathematics, The University of Western Ontario*
THE LAMBERT W FUNCTION [III.17]

Oliver E. Jensen, *Sir Horace Lamb Professor, School of Mathematics, The University of Manchester*
MATHEMATICAL BIOMECHANICS [V.4]

Chris R. Johnson, *Director, Scientific Computing and Imaging Institute; Distinguished Professor, School of Computing, University of Utah*
VISUALIZATION [VII.13]

Chandrika Kamath, *Member of Technical Staff, Lawrence Livermore National Laboratory*
DATA MINING AND ANALYSIS [IV.17]

Randall D. Kamien, *Vicki and William Abrams Professor in the Natural Sciences, University of Pennsylvania*
SOFT MATTER [IV.33]

Jonathan Peter Keating, *Henry Overton Wills Professor of Mathematics, University of Bristol*
RANDOM-MATRIX THEORY [IV.24]

David E. Keyes, *Professor of Applied Mathematics and Computational Science and Director, Extreme Computing Research Center, King Abdullah University of Science and Technology; Professor of Applied Mathematics and Applied Physics, Columbia University*
COMPUTATIONAL SCIENCE [IV.16]

Barbara Lee Keyfitz, *Dr. Charles Saltzer Professor of Mathematics, The Ohio State University*
CONSERVATION LAWS [II.6], SHOCKS [II.30]

David Krakauer, *President and Professor, Santa Fe Institute*
THE MATHEMATICS OF ADAPTATION (OR THE TEN AVATARS OF VISHNU) [V.1]

Rainer Kress, *Professor Emeritus, Institut für Numerische und Angewandte Mathematik, University of Göttingen*
INTEGRAL EQUATIONS [IV.4]

Alan J. Laub, *Professor, Department of Electrical Engineering/ Mathematics, University of California, Los Angeles*
THE RICCATI EQUATION [III.25]

Anita T. Layton, *Robert R. and Katherine B. Penn Associate Professor, Department of Mathematics, Duke University*
MATHEMATICAL PHYSIOLOGY [V.5]

Tanya Leise, *Associate Professor of Mathematics, Amherst College*
CLOAKING [VI.1]

Giovanni Leoni, *Professor, Department of Mathematical Sciences, Carnegie Mellon University*
CALCULUS OF VARIATIONS [IV.6]

Randall J. LeVeque, *Professor, Department of Applied Mathematics, University of Washington*
TSUNAMI MODELING [V.19]

Copyrighted material

Rachel Levy, *Associate Professor of Mathematics, Harvey Mudd College*
TEACHING APPLIED MATHEMATICS [VIII.7]

W. R. B. Lionheart, *Professor of Applied Mathematics, The University of Manchester*
AIRPORT BAGGAGE SCREENING WITH X-RAY TOMOGRAPHY [VII.19]

Andrew W. Lo, *Charles E. and Susan T. Harris Professor, Massachusetts Institute of Technology*
PORTFOLIO THEORY [V.10]

Christian Lubich, *Professor, Mathematisches Institut, Universität Tübingen*
NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS [IV.12]

Peter Lynch, *Emeritus Professor, School of Mathematical Sciences, University College, Dublin*
NUMERICAL WEATHER PREDICTION [V.18]

Malcolm A. H. MacCallum, *Emeritus Professor of Applied Mathematics, Queen Mary University of London*
EINSTEIN'S FIELD EQUATIONS [III.10]

Dian I. Martin, *CEO, Senior Consultant, Small Bear Technologies, Inc.*
TEXT MINING [VII.24]

P. A. Martin, *Professor of Applied Mathematics, Colorado School of Mines*
ASYMPTOTICS [II.1], BOUNDARY LAYER [II.2], INTEGRAL TRANSFORMS AND CONVOLUTION [II.19], SINGULARITIES [II.31], WAVE PHENOMENA [II.36], BESSEL FUNCTIONS [III.2], THE BURGERS EQUATION [III.4], THE CAUCHY–RIEMANN EQUATIONS [III.6], THE DIFFUSION EQUATION [III.8], THE EULER EQUATIONS [III.11], THE GAMMA FUNCTION [III.13], HOOKE'S LAW [III.15], LAPLACE'S EQUATION [III.18], THE SHALLOW-WATER EQUATIONS [III.27], THE WAVE EQUATION [III.31], COMPLEX ANALYSIS [IV.1]

Youssef Marzouk, *Associate Professor, Department of Aeronautics and Astronautics and Center for Computational Engineering, Massachusetts Institute of Technology*
UNCERTAINTY QUANTIFICATION [II.34]

Moshe Matalon, *Caterpillar Distinguished Professor, Mechanical Science and Engineering, University of Illinois at Urbana-Champaign*
FLAME PROPAGATION [VII.15]

Sean McKee, *Research Professor of Mathematics and Statistics, University of Strathclyde, Glasgow*
MODELING A PREGNANCY TESTING KIT [VII.18]

Ross C. McPhedran, *Emeritus Professor in Physics, CUDOS, University of Sydney*
EFFECTIVE MEDIUM THEORIES [IV.31]

John G. McWhirter, *Distinguished Research Professor, School of Engineering, Cardiff University*
SIGNAL PROCESSING [IV.35]

Beatrice Meini, *Professor of Numerical Analysis, University of Pisa*
MARKOV CHAINS [II.25]

James D. Meiss, *Professor, Department of Applied Mathematics, University of Colorado at Boulder*
ORDINARY DIFFERENTIAL EQUATIONS [IV.2]

Heather Mendick, *Reader in Education, Brunel University*
MEDIATED MATHEMATICS: REPRESENTATIONS OF MATHEMATICS IN POPULAR CULTURE AND WHY THESE MATTER [VIII.8]

Peter D. Miller, *Professor of Mathematics, The University of Michigan, Ann Arbor*
PERTURBATION THEORY AND ASYMPTOTICS [IV.5]

H. K. Moffatt, *Emeritus Professor of Mathematical Physics, University of Cambridge*
THE NAVIER–STOKES EQUATIONS [III.23], FLUID DYNAMICS [IV.28]

Esteban Moro, *Associate Professor, Department of Mathematics, Universidad Carlos III de Madrid*
NETWORK ANALYSIS [IV.18]

Clément Mouhot, *Professor of Mathematical Sciences, University of Cambridge*
KINETIC THEORY [IV.25]

Jean-Michel Muller, *Directeur de Recherche, CNRS*
EVALUATING ELEMENTARY FUNCTIONS [VI.11]

Tri-Dung Nguyen, *Associate Professor in Operational Research and Management Sciences, University of Southampton*
PORTFOLIO THEORY [V.10]

Qing Nie, *Professor, Department of Mathematics, Center for Mathematical and Computational Biology, Center for Complex Biological Systems, University of California, Irvine*
SYSTEMS BIOLOGY [VII.22]

Harald Niederreiter, *Senior Scientist, RICAM, Austrian Academy of Sciences, Linz*
RANDOM NUMBER GENERATION [VI.12]

Amy Novick-Cohen, *Professor, Department of Mathematics, Technion—Israel Institute of Technology, Haifa*
THE CAHN–HILLIARD EQUATION [III.5]

Bernt Øksendal, *Professor, Department of Mathematics, University of Oslo*
APPLICATIONS OF STOCHASTIC ANALYSIS [IV.14]

Alexander V. Panfilov, *Professor, Department of Physics and Astronomy, Gent University*
CARDIAC MODELING [V.6]

Nicola Parolini, *Associate Professor of Numerical Analysis, Dipartimento di Matematica, MOX Politecnico di Milano*
SPORT [V.2]

Kristin Potter, *Scientific Software Consultant, University of Oregon*
VISUALIZATION [VII.13]

Andrea Prosperetti, *C. A. Miller Jr. Professor of Mechanical Engineering, Johns Hopkins University; G. Berkhoff Professor of Applied Physics, University of Twente*
BUBBLES [VI.2]

Ian Proudler, *Professor of Signal Processing, Loughborough University*
SIGNAL PROCESSING [IV.35]

Alfio Quarteroni, *Professor and Director, Chair of Modelling and Scientific Computing, Ecole Polytechnique Fédérale de Lausanne*
SPORT [V.2]

Anders Rantzer, *Professor, Automatic Control, LTH Lund University*
CONTROL THEORY [IV.34]

Marcos Raydan, *Professor, Departamento de Cómputo Científico y Estadística, Universidad Simón Bolívar*
NONLINEAR EQUATIONS AND NEWTON'S METHOD [II.28]

Daniel N. Rockmore, *William H. Neukom 1964 Professor of Computational Science, Dartmouth College*
THE FAST FOURIER TRANSFORM [II.10], THE MATHEMATICS OF ADAPTATION (OR THE TEN AVATARS OF VISHNU) [V.1]

Copyrighted material

Donald G. Saari, *Distinguished Professor and Director, Institute for Mathematical Behavioral Sciences, University of California, Irvine*
FROM THE *N*-BODY PROBLEM TO ASTRONOMY AND DARK MATTER [VI.16], VOTING SYSTEMS [VII.25]

Fadil Santosa, *Professor, School of Mathematics, University of Minnesota; Director, Institute for Mathematics and its Applications*
HOMOGENIZATION [II.17], THE LEVEL SET METHOD [II.24], MULTISCALE MODELING [II.27], INVERSE PROBLEMS [IV.15]

Guillermo Sapiro, *Edmund T. Pratt, Jr. School Professor of Electrical and Computer Engineering, Duke University*
MATHEMATICAL IMAGE PROCESSING [VII.8]

Arnd Scheel, *Professor, School of Mathematics, University of Minnesota*
PATTERN FORMATION [IV.27]

Emily Shuckburgh, *Head of Open Oceans, British Antarctic Survey*
EARTH SYSTEM DYNAMICS [IV.30]

Reinhard Siegmund-Schultze, *Faculty of Engineering and Science, University of Agder*
THE HISTORY OF APPLIED MATHEMATICS [I.6]

Valeria Simoncini, *Professor of Numerical Analysis, Alma Mater Studiorum Università di Bologna*
KRYLOV SUBSPACES [II.23]

Ronnie Sircar, *Professor, Operations Research & Financial Engineering Department, Princeton University*
FINANCIAL MATHEMATICS [V.9]

Malcolm C. Smith, *Professor, Department of Engineering, University of Cambridge*
INERTERS [V.3]

Roel Snieder, *W. M. Keck Distinguished Professor of Basic Exploration Science, Colorado School of Mines*
IMAGING THE EARTH USING GREEN'S THEOREM [VII.16]

Erkki Somersalo, *Professor, Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University*
DIMENSIONAL ANALYSIS AND SCALING [II.9]

Frank Sottile, *Professor of Mathematics, Texas A&M University*
ALGEBRAIC GEOMETRY [IV.39]

Ian Stewart, *Emeritus Professor of Mathematics, University of Warwick*
SYMMETRY IN APPLIED MATHEMATICS [IV.22], HOW TO WRITE A GENERAL INTEREST MATHEMATICS BOOK [VIII.3]

Victoria Stodden, *Associate Professor, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign*
REPRODUCIBLE RESEARCH IN THE MATHEMATICAL SCIENCES [VIII.5]

Gilbert Strang, *Professor of Mathematics, Massachusetts Institute of Technology*
A SYMMETRIC FRAMEWORK WITH MANY APPLICATIONS [V.12], TEACHING APPLIED MATHEMATICS [VIII.7]

Agnès Sulem, *Researcher, INRIA Paris–Rocquencourt*
APPLICATIONS OF STOCHASTIC ANALYSIS [IV.14]

Endre Süli, *Professor of Numerical Analysis, University of Oxford*
NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS [IV.13]

William W. Symes, *Noah G. Harding Professor in Computational and Applied Mathematics and Professor of Earth Science, Rice University*
INVERSE PROBLEMS [IV.15]

Nico M. Temme, *Emeritus Researcher, Centrum Wiskunde & Informatica, Amsterdam*
SPECIAL FUNCTIONS [IV.7]

David Tong, *Professor of Theoretical Physics, University of Cambridge*
CLASSICAL MECHANICS [IV.19]

Warwick Tucker, *Professor of Mathematics, Uppsala University*
INTERVAL ANALYSIS [II.20], COMPUTER-AIDED PROOFS VIA INTERVAL ANALYSIS [VII.3]

Peter R. Turner, *Dean of Arts and Sciences and Professor of Mathematics, Clarkson University*
TEACHING APPLIED MATHEMATICS [VIII.7]

P. J. Upton, *Lecturer, Department of Mathematics and Statistics, The Open University*
THE DELTA FUNCTION AND GENERALIZED FUNCTIONS [III.7]

P. van den Driessche, *Professor Emeritus of Mathematics and Statistics, University of Victoria*
THE SPREAD OF INFECTIOUS DISEASES [V.16]

Sergio Verdú, *Eugene Higgins Professor of Electrical Engineering, Princeton University*
INFORMATION THEORY [IV.36]

Cédric Villani, *Professor of Mathematics, University Claude Bernard Lyon I; Director, Institut Henri Poincaré (CNRS/UPMC)*
KINETIC THEORY [IV.25]

Jens Vygen, *Professor, Research Institute for Discrete Mathematics, University of Bonn*
COMBINATORIAL OPTIMIZATION [IV.38], CHIP DESIGN [VII.6]

Charles W. Wampler, *Technical Fellow, General Motors Global Research and Development*
ROBOTICS [VI.14]

Z. Jane Wang, *Professor, Department of Physics, Cornell University*
INSECT FLIGHT [VI.5]

Denis Weaire, *Emeritus Professor, School of Physics, Trinity College Dublin*
FOAMS [VI.3]

Karen Willcox, *Professor of Aeronautics and Astronautics, Massachusetts Institute of Technology*
UNCERTAINTY QUANTIFICATION [II.34]

Walter Willinger, *Chief Scientist, NIKSUN, Inc.*
COMMUNICATION NETWORKS [VII.23]

Peter Winkler, *William Morrill Professor of Mathematics and Computer Science, Dartmouth College*
APPLIED COMBINATORICS AND GRAPH THEORY [IV.37]

Stephen J. Wright, *Professor, Department of Computer Sciences, University of Wisconsin-Madison*
CONTINUOUS OPTIMIZATION (NONLINEAR AND LINEAR PROGRAMMING) [IV.11]

Lexing Ying, *Professor of Mathematics, Stanford University*
THE *N*-BODY PROBLEM AND THE FAST MULTIPOLE METHOD [VI.17]

Ya-xiang Yuan, *Professor, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences*
MATHEMATICS AND POLICY [VIII.9]

Copyrighted material

Copyrighted material

Part I
Introduction to Applied Mathematics

I.1 What Is Applied Mathematics?

Nicholas J. Higham

1 The Big Picture

Applied mathematics is a large subject that interfaces with many other fields. Trying to define it is problematic, as noted by William Prager and Richard Courant, who set up two of the first centers of applied mathematics in the United States in the first half of the twentieth century, at Brown University and New York University, respectively. They explained that:

> Precisely to define applied mathematics is next to impossible. It cannot be done in terms of subject matter: the borderline between theory and application is highly subjective and shifts with time. Nor can it be done in terms of motivation: to study a mathematical problem for its own sake is surely not the exclusive privilege of pure mathematicians. Perhaps the best I can do within the framework of this talk is to describe applied mathematics as the bridge connecting pure mathematics with science and technology.
>
> Prager (1972)

> Applied mathematics is not a definable scientific field but a human attitude. The attitude of the applied scientist is directed towards finding clear cut answers which can stand the test of empirical observation. To obtain the answers to theoretically often insuperably difficult problems, he must be willing to make compromises regarding rigorous mathematical completeness; he must supplement theoretical reasoning by numerical work, plausibility considerations and so on.
>
> Courant (1965)

Garrett Birkhoff offered the following view in 1977, with reference to the mathematician and physicist Lord Rayleigh (John William Strutt, 1842–1919):

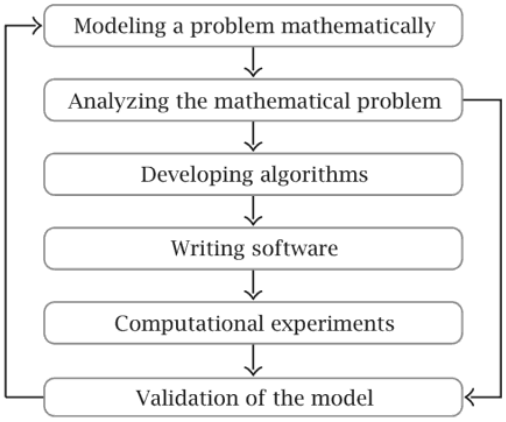


Figure 1 The main steps in solving a problem in applied mathematics.

> Essentially, mathematics becomes "applied" when it is used to solve real-world problems "neither seeking nor avoiding mathematical difficulties" (Rayleigh).

Rather than define what applied mathematics is, one can describe the methods used in it. Peter Lax stated of these methods, in 1989, that:

> Some of them are organic parts of pure mathematics: rigorous proofs of precisely stated theorems. But for the greatest part the applied mathematician must rely on other weapons: special solutions, asymptotic description, simplified equations, experimentation both in the laboratory and on the computer.

Here, instead of attempting to give our own definition of applied mathematics we describe the various facets of the subject, as organized around solving a problem. The main steps are described in figure 1. Let us go through each of these steps in turn.

Copyrighted material

Modeling a problem. Modeling is about taking a physical problem and developing equations—differential, difference, integral, or algebraic—that capture the essential features of the problem and so can be used to obtain qualitative or quantitative understanding of its behavior. Here, "physical problem" might refer to a vibrating string, the spread of an infectious disease, or the influence of people participating in a social network. Modeling is necessarily imperfect and requires simplifying assumptions. One needs to retain enough aspects of the system being studied that the model reproduces the most important behavior but not so many that the model is too hard to analyze. Different types of models might be feasible (continuous, discrete, stochastic), and for a given type there can be many possibilities. Not all applied mathematicians carry out modeling; in fact, most join the process at the next step.

Analyzing the mathematical problem. The equations formulated in the previous step are now analyzed and, ideally, solved. In practice, an explicit, easily evaluated solution usually cannot be obtained, so approximations may have to be made, e.g., by discretizing a differential equation, producing a reduced problem. The techniques necessary for the analysis of the equations or reduced problem may not exist, so this step may involve developing appropriate new techniques. If analytic or perturbation methods have been used then the process may jump from here directly to validation of the model.

Developing algorithms. It may be possible to solve the reduced problem using an existing algorithm—a sequence of steps that can be followed mechanically without the need for ingenuity. Even if a suitable algorithm exists it may not be fast or accurate enough, may not exploit available structure or other problem features, or may not fully exploit the architecture of the computer on which it is to be run. It is therefore often necessary to develop new or improved algorithms.

Writing software. In order to use algorithms on a computer it is necessary to implement them in software. Writing reliable, efficient software is not easy, and depending on the computer environment being targeted it can be a highly specialized task. The necessary software may already be available, perhaps in a package or program library. If it is not, software is ideally developed and documented to a high standard and made available to others. In many cases the software stage consists simply of writing short programs, scripts, or notebooks that carry out the necessary computations and summarize the results, perhaps graphically.

Computational experiments. The software is now run on problem instances and solutions obtained. The computations could be numeric or symbolic, or a mixture of the two.

Validation of the model. The final step is to take the results from the experiments (or from the analysis, if the previous three steps were not needed), interpret them (which may be a nontrivial task), and see if they agree with the observed behavior of the original system. If the agreement is not sufficiently good then the model can be modified and the loop through the steps repeated. The validation step may be impossible, as the system in question may not yet have been built (e.g., a bridge or a building).

Other important tasks for some problems, which are not explicitly shown in our outline, are to calibrate parameters in a model, to quantify the uncertainty in these parameters, and to analyze the effect of that uncertainty on the solution of the problem. These steps fall under the heading of UNCERTAINTY QUANTIFICATION [II.34].

Once all the steps have been successfully completed the mathematical model can be used to make predictions, compare competing hypotheses, and so on. A key aim is that the mathematical analysis gives new insights into the physical problem, even though the mathematical model may be a simplification of it.

A particular applied mathematician is most likely to work on just some of the steps; indeed, except for relatively simple problems it is rare for one person to have the skills to carry out the whole process from modeling to computer solution and validation.

In some cases the original problem may have been communicated by a scientist in a different field. A significant effort can be required to understand what the mathematical problem is and, when it is eventually solved, to translate the findings back into the language of the relevant field. Being able to talk to people outside mathematics is therefore a valuable skill for the applied mathematician.

It would be wrong to give the impression that all applied mathematics is done in the context of modeling. Frequently, a mathematical problem will be tackled because of its inherent interest (see the quote from Prager above) with the hope or expectation that a relevant application will be found. Indeed some applied

Copyrighted material

mathematicians spend their whole careers working in this way. There are many examples of mathematical results that provide the foundations for important practical applications but were developed without knowledge of those applications (sections 3.1 and 3.2 provide such examples).

Before the twentieth century, applied mathematics was driven by problems in astronomy and mechanics. In the twentieth century physics became the main driver, with other areas such as biology, chemistry, economics, engineering, and medicine also providing many challenging mathematical problems from the 1950s onward. With the massive and still-growing amounts of data available to us in today's digital society we can expect information, in its many guises, to be an increasingly important influence on applied mathematics in the twenty-first century.

For more on the definition and history of applied mathematics, including the development of the term "applied mathematics," see the article HISTORY OF APPLIED MATHEMATICS [I.6].

2 Applied Mathematics and Pure Mathematics

The question of how applied mathematics compares with pure mathematics is often raised and has been discussed by many authors, sometimes in controversial terms. We give a few highlights.

Paul Halmos wrote a 1981 paper provocatively titled "Applied mathematics is bad mathematics." However, much of what Halmos says would not be disputed by many applied mathematicians. For example:

> Pure mathematics can be practically useful and applied mathematics can be artistically elegant....
>
> Just as pure mathematics can be useful, applied mathematics can be more beautifully useless than is sometimes recognized....
>
> Applied mathematics is an intellectual discipline, not a part of industrial technology....
>
> Not only, as is universally admitted, does the applied need the pure, but, in order to keep from becoming inbred, sterile, meaningless, and dead, the pure needs the revitalization and the contact with reality that only the applied can provide.

G. H. Hardy's book *A Mathematician's Apology* (1940) is well known as a defense of mathematics as a subject that can be pursued for its own sake and beauty. As such it contains some criticism of applied mathematics:

> But is not the position of an ordinary applied mathematician in some ways a little pathetic? If he wants to be useful, he must work in a humdrum way, and he cannot give full play to his fancy even when he wishes to rise to the heights. "Imaginary" universes are so much more beautiful than this stupidly constructed "real" one; and most of the finest products of an applied mathematician's fancy must be rejected, as soon as they have been created, for the brutal but sufficient reason that they do not fit the facts.

Halmos and Hardy were pure mathematicians. Applied mathematicians C. C. Lin and L. A. Segel offer some insights in the introductory chapter of their classic 1974 book *Mathematics Applied to Deterministic Problems in the Natural Sciences*:

> The differences in motivation and objectives between pure and applied mathematics—and the consequent differences in emphasis and attitude—must be fully recognized. In pure mathematics, one is often dealing with such abstract concepts that logic remains the only tool permitting judgment of the correctness of a theory. In applied mathematics, empirical verification is a necessary and powerful judge. However . . . in some cases (e.g., celestial mechanics), rigorous theorems can be proved that are also valuable for practical purposes. On the other hand, there are many instances in which new mathematical ideas and new mathematical theories are stimulated by applied mathematicians or theoretical scientists.

They also opine that:

> Much second-rate pure mathematics is concealed beneath the trappings of applied mathematics (and vice versa). As always, knowledge and taste are needed if quality is to be assured.

The applied versus pure discussion is not always taken too seriously. Chandler Davis quotes the applied mathematician Joseph Keller as saying, "pure mathematics is a subfield of applied mathematics"!

The discussion can also focus on where in the spectrum a particular type of mathematics lies. An interesting story was told in 1988 by Clifford Truesdell of his cofounding in 1952 of the *Journal of Rational Mechanics and Analysis* (which later became *Archive for Rational Mechanics and Analysis*). He explained that

> In those days papers on the foundation of continuum mechanics were rejected by journals of mathematics as being applied, by journals of "applied" mathematics as being physics or pure mathematics, by journals of physics as being mathematics, and by all of them as too long, too expensive to print, and of interest to no one.

Copyrighted material

3 Applied Mathematics in Everyday Life

We now give three examples of applied mathematics in use in everyday life. These examples were chosen because they can be described without delving into too many technicalities and because they illustrate different characteristics. Some of the terms used in the descriptions are explained in THE LANGUAGE OF APPLIED MATHEMATICS [I.2].

3.1 Searching Web Pages

In the early to mid-1990s—the early days of the World Wide Web—search engines would find Web pages that matched a user's search query and would order the results by a simple criterion such as the number of times that the search query appears on a page. This approach became unsatisfactory as the Web grew in size and spammers learned how to influence the search results. From the late 1990s onward, more sophisticated criteria were developed, based on analysis of the links between Web pages. One of these is Google's PAGERANK ALGORITHM [VI.9]. Another is the hyperlink-induced topic search (HITS) algorithm of Kleinberg.

The HITS algorithm is based on the idea of determining hubs and authorities. *Authorities* are Web pages with many links to them and for which the linking pages point to many authorities. For example, the *New York Times* home page or a Wikipedia article on a popular topic might be an authority. *Hubs* are pages that point to many authorities. An example might be a page on a programming language that provides links to useful pages about that language but that does not necessarily contain much content itself. The authorities are the pages that we would like to rank higher among pages that match a search term. However, the definition of hubs and authorities is circular, as each depends on the other.

To resolve this circularity, associate an authority weight x_i and a hub weight y_i with page i, with both weights nonnegative. Let there be n pages to be considered (in practice this is a much smaller number than the total number of pages that match the search term). Define an $n \times n$ matrix $A = (a_{ij})$ by $a_{ij} = 1$ if there is a hyperlink from page i to page j and by $a_{ij} = 0$ otherwise. Let us make initial guesses $x_i^{(0)} = 1$ and $y_i^{(0)} = 1$, for $i = 1, 2, \ldots, n$. It is reasonable to update the authority weight x_i for page i by replacing it by the sum of the weights of the hubs that point to it. Similarly, the hub weight y_i for page i can be replaced by the sum of the weights of the authorities to which it points. In equations, these updates can be written as $x_i^{(1)} = \sum_{a_{ji}\neq 0} y_j^{(0)}$ and $y_i^{(1)} = \sum_{a_{ij}\neq 0} x_j^{(1)}$; note that in the latter equation we are using the updated authority weights, and the sums are over those j for which a_{ji} or a_{ij} is nonzero, respectively. This process can be iterated:

$$\left.\begin{aligned} x_i^{(k+1)} &= \sum_{a_{ji}\neq 0} y_j^{(k)} \\ y_i^{(k+1)} &= \sum_{a_{ij}\neq 0} x_j^{(k+1)} \end{aligned}\right\} \quad k = 0, 1, 2, \ldots.$$

The circular definition of hubs and authorities has been turned into an iteration. The iteration is best analyzed by rewriting it in matrix–vector form. Defining the n-vectors

$$x_k = [x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}]^{\mathrm{T}},$$
$$y_k = [y_1^{(k)}, y_2^{(k)}, \ldots, y_n^{(k)}]^{\mathrm{T}}$$

and recalling that the elements of A are 0 or 1, we can rewrite the iteration as

$$\left.\begin{aligned} x_{k+1} &= A^{\mathrm{T}} y_k \\ y_{k+1} &= A x_{k+1} \end{aligned}\right\} \quad k = 0, 1, 2, \ldots,$$

where $A^{\mathrm{T}} = (a_{ji})$ is the transpose of A. Combining the two formulas into one gives $x_{k+1} = A^{\mathrm{T}} y_k = A^{\mathrm{T}}(A x_k) = (A^{\mathrm{T}}A)x_k$. Hence the x_k are generated by repeatedly multiplying by the matrix $A^{\mathrm{T}}A$. Each element of $A^{\mathrm{T}}A$ is either zero or a positive integer, so the powers of $A^{\mathrm{T}}A$ will usually grow without bound. In practice we should therefore normalize the vectors x_k and y_k so that the largest element is 1; this avoids overflow and has no effect on the relative sizes of the components, which is all that matters. Our iteration is then

$$x_{k+1} = c_k^{-1} A^{\mathrm{T}} A x_k,$$

where c_k is the largest element of $A^{\mathrm{T}}Ax_k$. If the sequences x_k and c_k converge, say to x_* and c_*, respectively, then $A^{\mathrm{T}}Ax_* = c_* x_*$. This equation says that x_* is an eigenvector of $A^{\mathrm{T}}A$ with corresponding eigenvalue c_*. A similar argument shows that, if the normalized sequence of vectors y_k converges, then it must be to an eigenvector of AA^{T}.

This process of repeated multiplication by a matrix is known as the POWER METHOD [IV.10 §5.5]. The PERRON–FROBENIUS THEOREM [IV.10 §11.1] can be used to show that, provided the matrix $A^{\mathrm{T}}A$ has a property called irreducibility, it has a unique eigenvalue of largest magnitude and this eigenvalue is real and positive, with an associated eigenvector x having positive entries. Convergence theory for the power method then shows that

Copyrighted material

Figure 3 Close-up of part of a metal sign with a hot spot from a reflection. (a) Original image showing source and target regions. (b) The result of copying source to target. (c) The result of one application of Photoshop healing brush with same target area. In practice, multiple applications of the healing brush would be used with smaller, overlapping target areas.

between different computers and an improved ability for mathematicians to understand the way algorithms will behave when implemented in floating-point arithmetic.

In IEEE double-precision arithmetic, numbers are represented to a precision equivalent to about sixteen significant decimal digits. In many situations in life, results are needed to far fewer figures and a final result must be *rounded*. For example, a conversion from euros to dollars producing an answer \$110.89613 might be rounded up to \$110.90: the nearest amount in whole cents. A bank paying the dollars into a customer's account might prefer to round down to \$110.89 and keep the remainder. However, deciding on the rules for rounding was not so simple when the euro was founded in 1997. A twenty-nine-page document was needed to specify precisely how conversions among the fifteen currencies of the member states and the euro should be done. Its pronouncements included how many significant figures each individual conversion rate should have (six was the number that was chosen), how rounding should be done (round to the nearest six-digit number), and how ties should be handled (always round up).

Even when rounding should be straightforward it is often carried out incorrectly. In 1982 the Vancouver Stock Exchange established an index with an initial value of 1000. After twenty-two months the index had been hitting lows in the 520s, despite the exchange apparently performing well. The index was recorded to three decimal places and it was discovered that the computer program calculating the index always rounded down, hence always underestimating the index. Upon recalculation (presumably with round to nearest) the index almost doubled.

In 2006 athlete Justin Gatlin was credited with a new world record of 9.76 seconds for the 100 meters. Almost a week after the race the time was changed to 9.77 seconds, meaning that he had merely equaled the existing record held by Asafa Powell. The reason for the change was that his recorded time of 9.766 had incorrectly been rounded down to the nearest hundredth of a second instead of up as the International Association of Athletics Federations rules require.

4 What Do Applied Mathematicians Do?

Applied mathematicians can work in academia, industry, or government research laboratories. Their work may involve research, teaching, and (especially for more senior mathematicians) administrative tasks such as managing teams of people. They usually spend only part of their time doing mathematics in the traditional sense of sitting with pen and paper scribbling formulas on paper and trying to solve equations or prove theorems. Under the general heading of research, a lot of time is spent writing papers, books, grant proposals, reports, lecture notes, and talks; attending seminars, conferences, and workshops; writing and running computer programs; reading papers in the research literature; refereeing papers submitted to journals and grant proposals submitted to funding bodies; and commenting on draft papers and theses written by Ph.D. students.

Mathematics can be a lonely endeavor: one may be working on different problems from one's colleagues or may be the only mathematician in a company. Although some applied mathematicians prefer to work alone, many collaborate with others, often in faraway countries. Collaborations are frequently initiated through discussions at conferences, though sometimes papers are coauthored by people who have never met, thanks to the ease of email communication.

Applied mathematics societies provide an important source of identity and connectivity, as well as opportunities for networking and professional development. They mostly focus on particular countries or regions, an exception being the Society for Industrial and Applied Mathematics (SIAM), based in Philadelphia. SIAM is the largest applied mathematics organization

Copyrighted material

in the world and has a strong international outlook, with about one-third of its members residing outside the United States. A mathematician's activities are frequently connected with societies, whether it be through publishing in or editing their journals, attending their conferences, or keeping up with news through their magazines and newsletters. Most societies offer greatly reduced membership fees (sometimes free membership) for students.

Applied mathematicians can be part of multidisciplinary teams. Their skills in problem solving, thinking logically, modeling, and programming are sought after in other subjects, such as medical imaging, weather prediction, and financial engineering.

In the business world, applied mathematics can be invisible because it is called "analytics," "modeling," or simply generic "research." But whatever their job title, applied mathematicians play a crucial role in today's knowledge-based economy.

5 What Is the Impact of Applied Mathematics?

The impact of applied mathematics is illustrated in many articles in this volume, and in this section we provide just a brief overview, concentrating on the impact outside mathematics itself.

Applied mathematics provides the tools and algorithms to enable understanding and predictive modeling of many aspects of our planet, including WEATHER [V.18] (for which the accuracy of forecasts has improved greatly in recent decades), ATMOSPHERE AND THE OCEANS [IV.30], TSUNAMIS [V.19], and SEA ICE [V.17]. In many cases the models are used to inform policy makers.

At least two mathematical algorithms are used by most of us almost every day. The FAST FOURIER TRANSFORM [II.10] is found in any device that carries out signal processing, such as a smartphone. Photos that we take on our cameras or view on a computer screen are usually stored using JPEG COMPRESSION [VII.7 §5].

X-ray tomography devices, ranging from AIRPORT LUGGAGE SCANNERS [VII.19] to HUMAN BODY SCANNERS [VII.9], rely on the fast and accurate solution of INVERSE PROBLEMS [IV.15], which are problems in which we need to recover information about the internals of a system from (noisy) measurements taken outside the system.

Investments are routinely made on the basis of mathematical models, whether for individual options or collections of assets (portfolios): see FINANCIAL MATHEMATICS [V.9] and PORTFOLIO THEORY [V.10].

The clever use of mathematical modeling offers a competitive advantage in sports, such as YACHT RACING [V.2], SWIMMING [V.2], and FORMULA ONE RACING [V.3], where small improvements can be the difference between success and failure.

I.2 The Language of Applied Mathematics

Nicholas J. Higham

This article provides background on the notation, terminology, and basic results and concepts of applied mathematics. It therefore serves as a foundation for the later articles, many of which cross-reference it.

In view of the limited space, the material has been restricted to that common to many areas of applied mathematics. A number of later articles provide their own careful introduction to the language of their particular topic.

1 Notation

Table 1 lists the Greek alphabet, which is widely used to denote mathematical variables. Note that almost always δ and ε are used to denote small quantities, and π is used as a variable as well as for $\pi = 3.14159\ldots$.

Mathematics has a wealth of notation to express commonly occurring concepts. But notation is both a blessing and a curse. Used carefully, it can make mathematical arguments easier to read and understand. If overused it can have the opposite effect, and often it is better to express a statement in words than in symbols (see MATHEMATICAL WRITING [VIII.1]). Table 2 gives some notation that is common in informal contexts such as lectures and is occasionally encountered in this book. Table 3 summarizes basic notation used throughout the book.

2 Complex Numbers

Most applied mathematics takes place in the set of complex numbers, $\mathbb{C}$, or the set of real numbers, $\mathbb{R}$. A complex number $z = x + \mathrm{i}y$ has real and imaginary parts $x = \operatorname{Re} z$ and $y = \operatorname{Im} z$ belonging to $\mathbb{R}$, and the *imaginary unit* i denotes $\sqrt{-1}$. The imaginary unit is sometimes written as j, e.g., in electrical engineering and in the programming language PYTHON [VII.11].

We can represent complex numbers geometrically in the *complex plane*, in which a complex number $a + \mathrm{i}b$ is represented by the point with coordinates (a, b)

Copyrighted material

Table 1 The Greek alphabet. Where an uppercase Greek letter is the same as the Latin letter it is not shown.

α	alpha	ν	nu
β	beta	ξ, Ξ	xi
γ, Γ	gamma	o	omicron
δ, Δ	delta	π, ϖ, Π	pi
ϵ, ε	epsilon	ρ, ϱ	rho
ζ	zeta	$\sigma, \varsigma, \Sigma$	sigma
η	eta	τ	tau
$\theta, \vartheta, \Theta$	theta	υ, Υ	upsilon
ι	iota	ϕ, φ, Φ	phi
κ	kappa	χ	chi
λ, Λ	lambda	ψ, Ψ	psi
μ	mu	ω, Ω	omega

Table 2 Other notation.

$\Rightarrow$	Implies	$\exists$	There exists
$\Leftarrow$	Implied by	$\nexists$	There does not exist
$\Leftrightarrow$	If and only if	$\forall$	For all

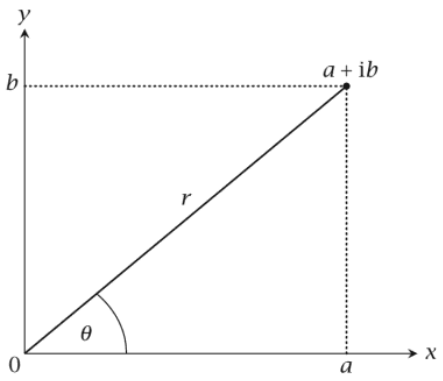


Figure 1 Complex plane with $z = a + \mathrm{i}b = r\mathrm{e}^{\mathrm{i}\theta}$.

(see figure 1). The corresponding diagram is called the *Argand diagram*. Important roles are played by the *right half-plane* $\{z\colon \operatorname{Re} z \geqslant 0\}$ and the *left half-plane* $\{z\colon \operatorname{Re} z \leqslant 0\}$. If we exclude the pure imaginary numbers ($\operatorname{Im} z = 0$) from these sets we obtain the *open* half-planes. Euler's formula, $\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$, is fundamental.

The *polar form* of a complex number is $z = r\mathrm{e}^{\mathrm{i}\theta}$, where $r \geqslant 0$ and the *argument* $\arg z = \theta$ are real, and θ can be restricted to any interval of length 2π, such as $[0, 2\pi)$ or $(-\pi, \pi]$. The *complex conjugate* of $z = x + \mathrm{i}y$ is $\bar{z} = x - \mathrm{i}y$, sometimes written z^*. The *modulus*, or *absolute value*, $|z| = (\bar{z}z)^{1/2} = (x^2 + y^2)^{1/2} = r$.

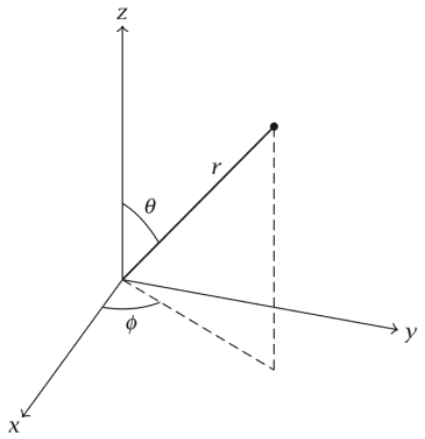


Figure 2 Spherical coordinates.

Complex arithmetic is defined in terms of real arithmetic according to the following rules, for $z_1 = x_1 + \mathrm{i}y_1$ and $z_2 = x_2 + \mathrm{i}y_2$:

$$z_1 \pm z_2 = x_1 \pm x_2 + \mathrm{i}(y_1 \pm y_2),$$
$$z_1 z_2 = x_1 x_2 - y_1 y_2 + \mathrm{i}(x_1 y_2 + x_2 y_1),$$
$$\frac{z_1}{z_2} = \frac{x_1 x_2 + y_1 y_2}{x_2^2 + y_2^2} + \mathrm{i}\frac{x_2 y_1 - x_1 y_2}{x_2^2 + y_2^2}.$$

In polar form multiplication and division become notationally simpler: if $z_1 = r_1\mathrm{e}^{\mathrm{i}\theta_1}$ and $z_2 = r_2\mathrm{e}^{\mathrm{i}\theta_2}$ then $z_1 z_2 = r_1 r_2 \mathrm{e}^{\mathrm{i}(\theta_1+\theta_2)}$ and $z_1/z_2 = (r_1/r_2)\mathrm{e}^{\mathrm{i}(\theta_1-\theta_2)}$.

3 Coordinate Systems

We are used to specifying a point in two dimensions by its x- and y-coordinates, and a point in three dimensions by its x-, y-, and z-coordinates. These are called *Cartesian coordinates*. In two dimensions we can also use *polar coordinates*, which are as described in the previous section if we identify (x, y) with $x + \mathrm{i}y$. *Spherical coordinates*, illustrated in figure 2, are an extension of polar coordinates to three dimensions. Here, (x, y, z) is represented by (r, θ, ϕ), where

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta,$$

with nonnegative radius r and angles θ and ϕ in the ranges $0 \leqslant \theta \leqslant \pi$ and $0 \leqslant \phi < 2\pi$.

Cylindrical coordinates provide another three-dimensional coordinate system. Here, polar coordinates are used in the xy-plane and z is retained, so (x, y, z) is represented by (r, θ, z).

Copyrighted material

Table 3 Notation frequently used in this book.

Notation	Meaning	Example
$\mathbb{R}$, $\mathbb{C}$	The real numbers, the complex numbers	
$\mathbb{R}^n$, $\mathbb{R}^{m\times n}$	The real n-vectors and real $m\times n$ matrices; similarly for $\mathbb{C}^n$ and $\mathbb{C}^{m\times n}$	
$\operatorname{Re} z$, $\operatorname{Im} z$	Real and imaginary parts of the complex number z	
$\mathbb{Z}$, $\mathbb{N}$	The integers, $\{0, \pm1, \pm2, \dots\}$, and the positive integers, $\{1, 2, \dots\}$	
$i = 1, 2, \dots, n$	The integer variable i takes on the values 1, 2, 3, and so on, up to n; also written $1 \leqslant i \leqslant n$ and $i = 1:n$	
$\approx$	Approximately equal; also written $\simeq$	$\pi \approx 3.14$
$\in$	Belongs to	$x \in \mathbb{R}$, $n \in \mathbb{Z}$
$\equiv$	Identically equal to $f \equiv 0$ means that f is the zero function, that is, f is zero for all values, not just some values, of its argument(s)	
$n!$	Factorial, $n! = n(n-1)\cdots 1$	
$\to$	Tends to, or converges to	$n \to \infty$
$\sum$	Summation	$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$
$\prod$	Product	$\prod_{i=1}^{3} x_i = x_1 x_2 x_3$
$\ll$, $\gg$	Much less than, much greater than	$n \gg 1$, $0 \leqslant \varepsilon \ll 1$
δ_{ij}	Kronecker delta: $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$	
$[a,b]$, (a,b), $[a,b)$	The closed interval $\{x\colon a \leqslant x \leqslant b\}$, the open interval $\{x\colon a < x < b\}$, and the half-closed, half-open interval $\{x\colon a \leqslant x < b\}$	
$f\colon P \to Q$	The function f maps the set P to the set Q, that is, $x \in P$ implies $f(x) \in Q$	
$f', f'', f''', f^{(k)}$	First, second, third, and kth derivatives of the function f	
$\dot{f}, \ddot{f}$	First and second derivatives of the function f	
$C[a,b]$	Real-valued continuous functions on $[a,b]$	$f \in C[a,b]$
$C^k[a,b]$	Real-valued functions with continuous derivatives of order $0, 1, \dots, k$ on $[a,b]$	$f \in C^2[a,b]$
$L^2[a,b]$	The functions $f\colon \mathbb{R} \to \mathbb{R}$ such that the Lebesgue integral $\int_a^b f(x)^2\,\mathrm{d}x$ exists	
$f \circ g$	Composition of functions: $(f \circ g)(x) = f(g(x))$	$\mathrm{e}^{x^2} = \mathrm{e}^x \circ x^2$
:=, =:	Definition of a variable or function, to distinguish from mathematical equality	$y' = 1 + y^4 =: f(y)$

4 Functions

A *function* f is a rule that assigns for each value of x a unique value $f(x)$. It can be thought of as a black box that takes an input x and produces an output $y = f(x)$. A function is sometimes called a *mapping*. If we write $y = f(x)$ then y is the *dependent variable* and x is the *independent variable*, also called the *argument* of f.

For some functions there is not a unique value of $f(x)$ for a given x, and these *multivalued functions* are not true functions unless restrictions are imposed. For example, consider $y = \log x$, which in general denotes any solution of the equation $\mathrm{e}^y = x$. There are infinitely many solutions, which can be written as $y = y_0 + 2\pi i k$ for $k \in \mathbb{Z}$, where y_0 is the *principal logarithm*, defined as the logarithm whose imaginary part lies in $(-\pi, \pi]$. The principal logarithm is often the one that is needed in practice and is usually the one computed by software. Multivalued functions of a complex variable can be elegantly handled using RIEMANN SURFACES [IV.1 §2] and BRANCH CUTS [IV.1 §2].

A function is *linear* if the independent variable appears only to the first power. Thus the function $f(x) = ax + b$, where a and b are constants, is linear in x. In some contexts, e.g., in convex optimization, $ax + b$ is called an *affine function* and the term linear is reserved for $f(x) = ax$, for which $f(tx) = tf(x)$ for all t.

A function f is *odd* if $f(x) = -f(-x)$ for all x and it is *even* if $f(x) = f(-x)$ for all x. For example, the sine function is odd, whereas x^2 and $|x|$ are even.

Copyrighted material

It is worth noting the distinction between the function f and its value $f(x)$ at a particular point x. Sometimes this distinction is blurred; for example, one might write "the function $f(u, v)$," in order to emphasize the symbols being used for the independent variables.

Functions with more than one independent variable are called *multivariate functions*. For ease of notation the independent variables can be collected into a vector. For example, the multivariate function $f(u, v) = \cos u \sin v$ can be written $f(x) = \cos x_1 \sin x_2$, where $x = [x_1, x_2]^{\mathrm{T}}$.

5 Limits and Continuity

The notion of a function converging to a limit as its argument approaches a certain value seems intuitively obvious. For example, the statement that $x^2 \to 4$ as $x \to 2$, where the symbol "$\to$" means tends to or converges to, is clearly true, as can be seen by considering the graph of x^2. However, we need to make the notion of convergence precise because a large number of definitions are built on it.

Let f be a real function of a real variable. We say that $f(x) \to \ell$ as $x \to a$, and we write $\lim_{x\to a} f(x) = \ell$, if for every $\varepsilon > 0$ there is a $\delta > 0$ such that $0 < |x-a| < \delta$ implies $|f(x) - \ell| < \varepsilon$. In other words, by choosing x close enough to a, $f(x)$ can be made as close as desired to ℓ. Showing that the definition holds in a particular case boils down to determining δ as a function of ε.

It is implicit in this definition that ℓ is finite. We say that $f(x) \to \infty$ as $x \to a$ if for every $\rho > 0$ there is a $\delta > 0$ such that $|x - a| < \delta$ implies $f(x) > \rho$.

In practice, mathematicians rarely prove existence of a limit by exhibiting the appropriate $\delta = \delta(\varepsilon)$ in these definitions. For example, one would argue that $\tan x \to \infty$ as $x \to \pi/2$ because $\sin x \to 1$ and $\cos x \to 0$ as $x \to \pi/2$. However, the definition might be used if f is an implicitly defined function whose behavior is not well understood.

We can also define one-sided limits, in which the limiting value of x is approached from the right or the left. For the right-sided limit $\lim_{x\to a^+} f(x) = \ell$, the definition of limit is modified so that $0 < |x - a| < \delta$ is replaced by $a < x < a + \delta$, and the left-sided limit $\lim_{x\to a^-} f(x)$ is defined analogously. The standard limit exists if and only if the right- and left-sided limits exist and are equal.

The function f is *continuous* at $x = a$ if $f(a)$ exists and $\lim_{x\to a} f(x) = f(a)$.

The definitions of limit and continuity apply equally well to functions of a complex variable. Here, the condition $|x - a| < \delta$ places x in a disk of radius less than δ in the complex plane instead of an open interval on the real axis.

The function f is continuous on $[a, b]$ if it is continuous at every point in that interval. A more restricted form of continuity is Lipschitz continuity. The function f is *Lipschitz continuous* on $[a, b]$ if

$$|f(x) - f(y)| \leqslant L|x - y| \quad \text{for all } x, y \in [a, b]$$

for some constant L, which is called the *Lipschitz constant*. This definition, which is quantitative as opposed to the purely qualitative usual definition of continuity, is useful in many settings in applied mathematics. A function may, however, be continuous without being Lipschitz continuous, as $f(x) = x^{1/2}$ on $[0, 1]$ illustrates.

A *sequence* $a_1, a_2, a_3, \ldots$ of real or complex numbers, written $\{a_n\}$, has limit c if for every $\varepsilon > 0$ there is a positive integer N such that $|a_n - c| < \varepsilon$ for all $n \geqslant N$. We write $c = \lim_{n\to\infty} a_n$. An *infinite series* $\sum_{i=1}^{\infty} a_i$ converges if the sequence of *partial sums* $\sum_{i=1}^{n} a_i$ converges.

6 Bounds

In applied mathematics we are often concerned with deriving bounds for quantities of interest. For example, we might wish to find a constant u such that $f(x) \leqslant u$ for all x on a given interval. Such a u, if it exists, is called an *upper bound*. Similarly, a lower bound is a constant ℓ such that $f(x) \geqslant \ell$ for all x on the interval. Of particular interest is the *least upper bound*, also called the *supremum* or sup, which is the smallest possible upper bound. The supremum might not actually be attained, as illustrated by the function $f(x) = x/(1+x)$ on $[0, \infty)$, which has supremum 1. The *infimum*, or inf, is the greatest possible lower bound.

A function that has an upper (or lower) bound is said to be *bounded above* (or *bounded below*). If the function is bounded both above and below it is said to be *bounded*. A function that is not bounded is *unbounded*.

Determining whether a certain function, perhaps a function of several variables or one defined in a FUNCTION SPACE [II.15], is bounded can be nontrivial and it is often a crucial step in proving the convergence of a process or determining the quality of an approximation.

Physical considerations sometimes imply that a function is bounded. For example, a function that represents energy must be nonnegative.

Copyrighted material

conditions can usually be derived if necessary. Sometimes, when deriving or using results, it is not possible to check smoothness conditions and one simply carries on anyway ("making compromises," as mentioned in the quote by Courant on page 1). It may be possible to verify by other means that an answer obtained in a nonrigorous way is valid.

For a function $f(x,y)$ of two variables, partial derivatives with respect to each of the two variables are defined by holding one variable constant and varying the other:

$$\frac{\partial f}{\partial x} = \lim_{\varepsilon\to 0} \frac{f(x+\varepsilon,y)-f(x,y)}{\varepsilon},$$
$$\frac{\partial f}{\partial y} = \lim_{\varepsilon\to 0} \frac{f(x,y+\varepsilon)-f(x,y)}{\varepsilon}.$$

Higher derivatives are defined recursively. For example,

$$\frac{\partial^2 f}{\partial x^2} = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial x}(x+\varepsilon,y)-\frac{\partial f}{\partial x}(x,y)}{\varepsilon},$$
$$\frac{\partial^2 f}{\partial x\partial y} = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial x}(x,y+\varepsilon)-\frac{\partial f}{\partial x}(x,y)}{\varepsilon},$$
$$\frac{\partial^2 f}{\partial y\partial x} = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial y}(x+\varepsilon,y)-\frac{\partial f}{\partial y}(x,y)}{\varepsilon}.$$

Common abbreviations are $f_x = \partial f/\partial x$, $f_{xy} = \partial^2 f/(\partial x\partial y)$, $f_{yy} = \partial^2 f/\partial y^2$, and so on. As long as they are continuous the two mixed second-order partial derivatives are equal: $f_{xy} = f_{yx}$.

For a function of n variables, $F\colon \mathbb{R}^n \to \mathbb{R}$, a Taylor series takes the form, for $x, a \in \mathbb{R}^n$,

$$F(x) = F(a) + \nabla F(a)^{\mathrm{T}}(x-a) + \tfrac{1}{2}(x-a)^{\mathrm{T}}\nabla^2 F(a)(x-a) + \cdots,$$

where $\nabla F(x) = (\partial F/\partial x_j) \in \mathbb{R}^n$ is the *gradient vector* and $\nabla^2 F(x) = (\partial^2 F/(\partial x_i \partial x_j)) \in \mathbb{R}^{n\times n}$ is the symmetric *Hessian matrix*, with x_j denoting the jth component of the vector x. The symbol ∇ is called nabla. Stationary points of F are zeros of the gradient and their nature (maximum, minimum, or saddle point) is determined by the eigenvalues of the Hessian (see CONTINUOUS OPTIMIZATION [IV.11 §2]).

Now we return to functions of a single (real) variable. The *indefinite integral* of $f(x)$ is $\int f(x)\,\mathrm{d}x$, while integrating between limits a and b gives the *definite integral* $\int_a^b f(x)\,\mathrm{d}x$. The definite integral can be interpreted as the area under the curve $f(x)$ between a and b. The inverse of differentiation is integration, as shown by the *fundamental theorem of calculus*, which states that, if f is continuous on $[a,b]$, then the function $g(x) = \int_a^x f(t)\,\mathrm{d}t$ is differentiable on (a,b) and $g'(x) = f(x)$. Generalizations of the fundamental theorem of calculus to functions of more than one variable are given in section 24.

For functions of two or more variables there are other kinds of integrals. When there are two variables, x and y, we can integrate over regions in the xy-plane (double integrals) or along curves in the plane (line integrals). For functions of three variables, x, y, and z, there are more possibilities. We can integrate over volumes (triple integrals) or over surfaces or along curves within xyz-space. As the number of variables increases, so does the number of different kinds of integrals. Multidimensional calculus shows how these different integrals can be calculated, used, and related. The number of variables can be very large (e.g., in mathematical finance) and the CURSE OF DIMENSIONALITY [I.3 §2] poses major challenges for numerical evaluation. Numerical integration in more than one dimension is an active area of research, and Monte Carlo methods and quasi-Monte Carlo methods are among the methods in use.

The *product rule* gives a formula for the derivative of a product of two functions:

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x)g(x) = f'(x)g(x) + f(x)g'(x).$$

Integrating this equation gives the rule for *integration by parts*:

$$\int f(x)g'(x)\,\mathrm{d}x = f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x.$$

In many problems functions are composed: the argument of a function is another function. Consider the example $f(x) = g(h(x))$. We would hope to be able to determine the derivative of f in terms of the derivatives of g and h. The *chain rule* provides the necessary formula: $f'(x) = h'(x)g'(h(x))$. An equivalent formulation is that, if f is a function of u, which is itself a function of x, then

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}u}\frac{\mathrm{d}u}{\mathrm{d}x}.$$

For example, if $f(x) = \sin x^2$ then with $f(x) = \sin u$ and $u = x^2$ we have $\mathrm{d}f/\mathrm{d}x = 2x\cos x^2$.

10 Ordinary Differential Equations

A differential equation is an equation containing one or more derivatives of an unknown function. It provides a relation among a function, its rate of change, and (possibly) higher-order rates of change. The independent

Copyrighted material

variable usually represents a spatial coordinate (x) or time (t). The differential equation may be accompanied by additional information about the function, called *boundary conditions* or *initial conditions*, that serve to uniquely determine the solution. A solution to a differential equation is a function that satisfies the equation for all values of the independent variables (perhaps in some region) and also satisfies the required boundary conditions or initial conditions. A differential equation can express a law of motion, a conservation law, or concentrations of constituents of a chemical reaction, for example.

Ordinary differential equations (ODEs) contain just one independent variable. The simplest nontrivial ODE is $\mathrm{d}y/\mathrm{d}t = ay$, where $y = y(t)$ is a function of t. This equation is linear in y and it is *first order* because only the first derivative of y appears. The general solution is $y(t) = ce^{at}$, where c is an arbitrary constant. To determine c, some value of y must be supplied, say $y(0) = y_0$, whence $c = y_0$.

A general first-order ODE has the form $y' = f(t, y)$ for some function f of two variables. The *initial-value problem* supplies an initial condition and asks for y at later times:

$$y' = f(t, y), \quad a \leqslant t \leqslant b, \quad y(a) = y_a.$$

A specific example is the Riccati equation

$$y' = t^2 + y^2, \quad 0 \leqslant t \leqslant 1, \quad y(0) = 0,$$

which is nonlinear because of the appearance of y^2.

For an example of a second-order ODE initial-value problem, that is, one involving y'', consider a mass m attached to a vertical spring and to a damper, as shown in figure 7. Let $y = y(t)$ denote how much the spring is stretched from its natural length at time t. Balancing forces using Newton's second law (force equals mass times acceleration) and HOOKE'S LAW [III.15] gives

$$my'' = mg - ky - cy',$$

where k is the spring constant, c is the damping constant, and g is the gravitational constant. With prescribed values for $y(0)$ and $y'(0)$ this is an initial-value problem. More generally, the spring might also be subjected to an external force $f(t)$, in which case the equation of motion becomes

$$my'' + cy' + ky = mg + f(t).$$

Second-order ODEs also arise in electrical networks. Consider the flow of electric current $I(t)$ in a simple RLC circuit composed of an inductor with inductance

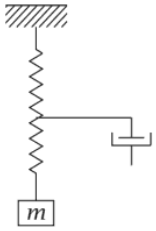


Figure 7 A spring system with damping.

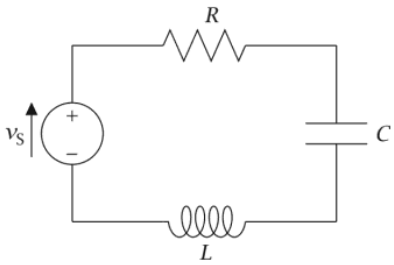


Figure 8 A simple RLC electric circuit.

L, a resistor with resistance R, a capacitor with capacitance C, and a source with voltage v_S, as illustrated in figure 8. The Kirchhoff voltage law states that the sum of the voltage drops around the circuit equals the input voltage, v_S. The voltage drops across the resistor, inductor, and capacitor are RI, $L\mathrm{d}I/\mathrm{d}t$, and Q/C, respectively, where $Q(t)$ is the charge on the capacitor, so

$$L\frac{\mathrm{d}I}{\mathrm{d}t} + RI + \frac{Q}{C} = v_S(t).$$

Since $I = \mathrm{d}Q/\mathrm{d}t$, this equation can be rewritten as the second-order ODE

$$L\frac{\mathrm{d}^2Q}{\mathrm{d}t^2} + R\frac{\mathrm{d}Q}{\mathrm{d}t} + \frac{1}{C}Q = v_S(t).$$

The unknown function y may have more than one component, as illustrated by the predator–prey model derived by Lotka and Volterra in the 1920s. In a population of rabbits (the prey) and foxes (the predators) let $r(t)$ be the number of rabbits at time t and $f(t)$ the number of foxes at time t. The model is

$$\frac{\mathrm{d}r}{\mathrm{d}t} = r - \alpha rf, \qquad r(0) = r_0,$$
$$\frac{\mathrm{d}f}{\mathrm{d}t} = -f + \alpha rf, \quad f(0) = f_0.$$

The rf term represents an interaction between the foxes and the rabbits (a fox eating a rabbit) and the parameter $\alpha \geqslant 0$ controls the amount of interaction. For $\alpha = 0$ there is no interaction and the solution is

Copyrighted material

$r(t) = r_0 e^t$, $f(t) = f_0 e^{-t}$: the foxes die from starvation and the rabbits go forth and multiply, unhindered. The aim is to investigate the behavior of the solutions for various parameters α and starting populations r_0 and f_0.

As we have described it, the predator–prey model has the apparent contradiction that r and f are integers by definition yet the solutions to the differential equation are real-valued. The way around this is to assume that r and f are large enough for the error in representing them by continuous variables to be small.

A *boundary-value problem* specifies the function at more than one value of the independent variable, as in the two-point boundary-value problem

$$y'' = f(t, y, y'), \quad a \leqslant t \leqslant b, \quad y(a) = y_a,\ y(b) = y_b.$$

An example is the Thomas–Fermi equation

$$y'' = t^{-1/2} y^{3/2}, \qquad y(0) = 1, \quad y(\infty) = 0,$$

which arises in a semiclassical description of the charge density in atoms of high atomic number. Another example, this time of third order, is the BLASIUS EQUATION [IV.28 §7.2]

$$2y''' + yy'' = 0, \qquad y(0) = y'(0) = 0, \quad y'(\infty) = 1,$$

which describes the boundary layer in a fluid flow.

A special type of ODE boundary-value problem is the *Sturm–Liouville problem*

$$-(p(x)y'(x))' + q(x)y(x) = \lambda r(x)y(x),$$
$$x \in [a, b], \quad y(a) = y(b) = 0.$$

This is an *eigenvalue problem*, meaning that the aim is to determine values of the parameter λ for which the boundary-value problem has a solution that is not identically zero.

11 Partial Differential Equations

Many important physical processes are modeled by partial differential equations (PDEs): differential equations containing more than one independent variable. We summarize a few key equations and basic concepts. We write the equations in forms where the unknown u has two space dimensions, $u = u(x, y)$, or one space dimension and one time dimension, $u = u(x, t)$. Where possible, the equations are given in parameter-free form, a form that is obtained by the process of NONDIMENSIONALIZATION [II.9]. Recall the abbreviations $u_t = \partial u/\partial t$, $u_{xx} = \partial^2 u/\partial x^2$, etc.

LAPLACE'S EQUATION [III.18] is

$$u_{xx} + u_{yy} = 0.$$

The left-hand side of the equation is the *Laplacian* of u, written Δu. This equation is encountered in electrostatics (for example), where u is the potential function. The equation $\Delta u = f$, for a given function $f(x, y)$, is known as *Poisson's equation*.

To define a problem with a unique solution it is necessary to augment the PDE with conditions on the solution: either boundary conditions for static problems or, for time-dependent problems, initial conditions. In the former class there are three main types of boundary conditions, with the problem being to determine u inside the boundary of a closed region.

- *Dirichlet conditions*, in which the function u is specified on the boundary.
- *Neumann conditions*, where the inner product (see section 19.1) of the gradient
$$\nabla u = [\partial u/\partial x, \partial u/\partial y]^{\mathrm{T}}$$
with the normal to the boundary is specified.
- *Cauchy conditions*, which comprise a combination of Dirichlet and Neumann conditions.

For time-dependent problems, which are known as evolution problems and represent equations of motion, initial conditions at the starting time, usually taken to be $t = 0$, are needed, the number of initial conditions depending on the highest order of time derivative in the PDE.

The WAVE EQUATION [III.31] is

$$u_{tt} = u_{xx}.$$

It describes linear, nondispersive propagation of a wave, represented by the wave function u, e.g., a vibrating string. Two initial conditions, prescribing $u(x, 0)$ and $u_t(x, 0)$, for example, are needed to determine u.

The HEAT EQUATION [III.8] (*diffusion equation*) is

$$u_t = u_{xx}, \tag{2}$$

which describes the diffusion of heat in a solid or the spread of a disease in a population. An initial condition prescribing u at $t = 0$ is usual. When a term $f(x, t, u)$ is added to the right-hand side of (2) the equation becomes a *reaction–diffusion equation*.

The *advection–diffusion equation* is

$$u_t + v u_x = u_{xx},$$

where v is a given function of x and t. Again, u is usually given at $t = 0$. For $v = 0$ this is just the heat equation. This PDE models the convection (or transport) of a quantity such as a pollutant in the atmosphere.

Copyrighted material

The general linear second-order PDE

$$au_{xx} + 2bu_{xt} + cu_{tt} = f(x, t, u, u_x, u_t) \qquad (3)$$

is classified into different types according to the (constant) coefficients of the second derivatives. Let $d = ac - b^2$, which is the determinant of the symmetric matrix $\left[\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right]$.

- If $d > 0$ the PDE is *elliptic*. These PDEs, of which the Laplace equation is a particular case, are associated with equilibrium or steady-state processes. The independent variables are denoted by x and y instead of x and t.
- If $d = 0$ the PDE is *parabolic*. This is an evolution problem governing a diffusion process. The heat equation is an example.
- If $d < 0$ the PDE is *hyperbolic*. This is an evolution problem, governing wave propagation. The wave equation is an example.

Some elliptic PDEs and parabolic PDEs have *maximum principles*, which say that the solution must take on its maximum value on the boundary of the domain over which it is defined.

In (3) we took a, b, and c to be constants, but they may also be specified as functions of x and t, in which case the nature of the PDE can change as x and t vary in the domain. For example, the TRICOMI EQUATION [III.30]

$$u_{xx} + xu_{yy} = 0$$

is hyperbolic for $x < 0$, elliptic for $x > 0$, and parabolic for $x = 0$.

The PDEs stated so far are all linear. Nonlinear PDEs, in which the unknown function appears nonlinearly, are of great practical importance. Examples are the KORTEWEG–DE VRIES EQUATION [III.16]

$$u_t + uu_x + u_{xxx} = 0,$$

the CAHN–HILLIARD EQUATION [III.5]

$$u_t = \Delta(-u + u^3 + \varepsilon^2 \Delta u),$$

and *Fisher's equation*

$$u_t = u_{xx} + u(1 - u),$$

a reaction–diffusion equation that describes PATTERN FORMATION [IV.27] and the propagation of genes in a population.

PDEs also occur in the form of eigenvalue problems. A famous example is the eigenvalue problem corresponding to the Laplace equation:

$$\Delta u + \lambda u = 0$$

on a membrane Ω, with boundary conditions that u vanishes on the boundary of Ω. A nonzero solution u is called an *eigenfunction* and λ is the corresponding *eigenvalue*. In a 1966 paper titled "Can one hear the shape of a drum?" Mark Kac asked the question of whether one can determine Ω given all the eigenvalues. In other words, do the frequencies at which a drum vibrates uniquely determine its shape? It was shown in a 1992 paper by Gordon, Webb, and Wolpert that the answer is no in general.

Higher-order PDEs also arise. For example, fluid dynamics problems involving surface tension forces are generally modeled by PDEs in space and time with fourth-order derivatives in space. The same is true of the *Euler–Bernoulli equation* for a beam, which has the form

$$\rho A \frac{\partial^2 u}{\partial t^2} + EI \frac{\partial^4 u}{\partial x^4} = f(x, t),$$

where $u(x, t)$ is the vertical displacement of the beam at time t and position x along the beam, ρ is the density of the beam, A its cross-sectional area, E is Young's modulus, I is the second moment of inertia, and $f(x, t)$ is an applied force.

12 Other Types of Differential Equations

Delay differential equations are differential equations in which the derivative of the unknown function y at time t (in general, a vector function) depends on past values of y and/or its derivatives. For example, $y'(t) = Ay(t - 1)$ is a delay differential equation analogue of the familiar $y'(t) = Ay(t)$. Looking for a solution of the form $y(t) = e^{wt}$ leads to the equation $we^w = A$, whose solutions are given by the LAMBERT W FUNCTION [III.17].

INTEGRAL EQUATIONS [IV.4] contain the unknown function inside an integral. Examples are *Fredholm equations*, which are of the form either

$$\int_0^1 K(x, y) f(y)\, dy = g(x),$$

where K and g are given and the task is to find f, or

$$\lambda \int_0^1 K(x, y) f(y)\, dy + g(x) = f(x),$$

where λ is an eigenvalue and again f is unknown. These two types of equations are analogous to a matrix linear system $Kf = g$ and an eigenvalue problem $(I - \lambda K)f = g$, respectively. *Integro-differential equations* involve both integrals and derivatives (see, for example, MODELING A PREGNANCY TESTING KIT [VII.18 §2]).

Copyrighted material

Fractional differential equations contain fractional derivatives. For example, $(\mathrm{d}/\mathrm{d}x)^{1/2}$ is defined to be an operator such that applying $(\mathrm{d}/\mathrm{d}x)^{1/2}$ twice in succession to a function $f(x)$ is the same as differentiating it once (that is, applying $\mathrm{d}/\mathrm{d}x$).

Differential–algebraic equations (DAEs) are systems of equations that contain both differential and algebraic equations. For example, the DAE

$$\begin{aligned} x'' &= -2\lambda x, \\ y'' &= -2\lambda y - g, \\ x^2 + y^2 &= L^2 \end{aligned}$$

describes the coordinates of an infinitesimal ball of mass 1 at the end of a pendulum of length L, where g is the gravitational constant and λ is the tension in the rod. DAEs often arise in the form $My' = f(t, y)$, where the matrix M is singular.

13 Recurrence Relations

Recurrence relations are the discrete counterpart of differential equations. They define a sequence $x_0, x_1, x_2, \ldots$ recursively, by specifying x_n in terms of earlier terms in the sequence. Such equations are also called *difference equations*, as they arise when derivatives in differential equations are replaced by FINITE DIFFERENCES [II.11].

A famous recurrence is the three-term recurrence that defines the *Fibonacci numbers*:

$$f_n = f_{n-1} + f_{n-2}, \quad n \geqslant 2, \quad f_0 = f_1 = 1.$$

This recurrence has the explicit solution $f_n = (\phi^n - (-\phi)^{-n})/\sqrt{5}$, where $\phi = (1 + \sqrt{5})/2$ is the *golden ratio*. An example of a two-term recurrence is $f(n) = nf(n-1)$, with $f(0) = 1$, which defines the factorial function $f(n) = n!$. Both the examples so far are linear recurrences, but in some recurrences the earlier terms appear nonlinearly, as in the LOGISTIC RECURRENCE [III.19] $x_{n+1} = \mu x_n(1 - x_n)$.

Although one can evaluate the terms in a recurrence one often needs an explicit formula for the general solution of the recurrence. Recurrence relations have a theory analogous to that of differential equations, though it is much less frequently encountered in courses and textbooks than it was fifty years ago.

The elements in a recurrence can be functions as well as numbers. Most transcendental functions that carry subscripts satisfy a recurrence. For example, the BESSEL FUNCTION [III.2] $J_n(x)$ of order n satisfies the three-term recurrence

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x).$$

An important source of three-term recurrences is ORTHOGONAL POLYNOMIALS [II.29].

14 Polynomials

Polynomials are one of the simplest and most familiar classes of functions and they find wide use in applied mathematics. A degree-n polynomial

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

is defined by its $n + 1$ coefficients $a_0, \ldots, a_n \in \mathbb{C}$ (with $a_n \neq 0$). Addition of two polynomials is carried out by adding the corresponding coefficients. Thus, if $q_n(x) = b_0 + b_1 x + \cdots + b_n x^n$ then $p_n(x) + q_n(x) = a_0 + b_0 + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n$. Multiplication is carried out by expanding the product term by term and collecting like powers of x:

$$\begin{aligned} p_n(x)q_n(x) = a_0 b_0 &+ (a_0 b_1 + a_1 b_0)x + \cdots \\ &+ (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0)x^n. \end{aligned}$$

The coefficient of x^n, $\sum_{i=0}^{n} a_i b_{n-i}$, is the *convolution* of the vectors $a = [a_0, a_1, \ldots, a_n]^{\mathrm{T}}$ and $b = [b_0, b_1, \ldots, b_n]^{\mathrm{T}}$. Polynomial division is also possible. Dividing p_n by q_m with $m \leqslant n$ results in

$$p_n(x) = q_m(x)g(x) + r(x), \tag{4}$$

where the quotient g and remainder r are polynomials and the degree of r is less than that of q_m.

The *fundamental theorem of algebra* says that a degree-n polynomial p_n has a *root* in $\mathbb{C}$; that is, there exists $z_1 \in \mathbb{C}$ such that $p_n(z_1) = 0$. If we take $q_m(x) = x - z_1$ in (4) then we have $p_n(x) = (x - z_1)g(x) + r(x)$, where $\deg r < 1$, so r is a constant. But setting $x = z_1$ we see that $0 = p_n(z_1) = r$, so $p_n(x) = (x - z_1)g(x)$ and g clearly has degree $n - 1$. Repeating this argument inductively on g, we end up with a factorization $p_n(x) = (x - z_1)(x - z_2)\cdots(x - z_n)$, which shows that p_n has n roots in $\mathbb{C}$ (not necessarily distinct). If the coefficients of p_n are real it does not follow that the roots are real, and indeed there may be no real roots at all, as the polynomial $x^2 + 1$ shows; however, nonreal roots must occur in complex conjugate pairs $x_j \pm \mathrm{i}y_j$.

Three basic problems associated with polynomials are as follows.

Evaluation: given the polynomial (specified by its coefficients), find its value at a given point. A standard way of doing this is HORNER'S METHOD [I.4 §6].

Interpolation: given the values of a degree-n polynomial at a set of $n + 1$ distinct points, find its coefficients. This can be done by various INTERPOLATION SCHEMES [I.3 §3.1].

Copyrighted material

$c_{ij} = a_{ij} + b_{ij}$ for all i and j. Multiplication by a scalar is defined in the natural way, so $C = \alpha A$ means that $c_{ij} = \alpha a_{ij}$ for all i and j. However, matrix multiplication is *not* defined element-wise. If A is $m \times r$ and B is $r \times n$ then the product $C = AB$ is $m \times n$ and is defined by

$$c_{ij} = \sum_{k=1}^{r} a_{ik} b_{kj}.$$

This formula can be obtained as follows. Write $B = [b^1, b^2, \ldots, b^n]$, where b^j is the jth column of B; this is a *partitioning* of B into its columns. Then $AB = A[b^1, b^2, \ldots, b^n] = [Ab^1, Ab^2, \ldots, Ab^n]$, where each Ab^j is a matrix–vector product. Matrix–vector products Ax with x an $r \times 1$ vector are in turn defined by

$$Ax = [a^1, a^2, \ldots, a^r] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} = x_1 a^1 + x_2 a^2 + \cdots + x_r a^r,$$

so that Ax is a *linear combination* of the columns of A.

Matrix multiplication is not commutative: $AB \neq BA$ in general, as is easily checked for 2×2 matrices. In some contexts the *commutator* (or *Lie bracket*) $[A, B] = AB - BA$ plays a role.

A linear system $Ax = b$ expresses the vector b as a linear combination of the columns of A. When A is square and of dimension n, this system provides n linear equations for the n components of x. The system has a unique solution when A is nonsingular, that is, when A has an inverse. An *inverse* of a square matrix A is a matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$, where I is the *identity matrix*, which has ones on the diagonal and zeros everywhere else. We can write $I = (\delta_{ij})$, where δ_{ij} is the *Kronecker delta* defined in table 3. The inverse is unique when it exists. If A is nonsingular then $x = A^{-1}b$ is the solution to $Ax = b$. While this formula is useful mathematically, in practice one almost never solves a linear system by inverting A and then multiplying the right-hand side by the inverse. Instead, GAUSSIAN ELIMINATION [IV.10 §2] with some form of pivoting is used.

Transposition turns an $m \times n$ matrix into an $n \times m$ one by interchanging the rows and columns: $C = A^{\mathrm{T}} \iff c_{ij} = a_{ji}$ for all i and j. *Conjugate transposition* also conjugates the elements: $C = A^* \iff c_{ij} = \overline{a_{ji}}$ for all i and j. The conjugate transpose of a product satisfies a useful reverse-order law: $(AB)^* = B^* A^*$.

Matrices can have a variety of different structures that can be exploited both in theory and in computation. A matrix $A \in \mathbb{R}^{n \times n}$ is *upper triangular* if $a_{ij} = 0$ for $i > j$, *lower triangular* if A^{T} is upper triangular, and *diagonal* if $a_{ij} = 0$ for $i \neq j$. For $n = 3$, such matrices have the forms

$$\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}, \quad \begin{bmatrix} \times & 0 & 0 \\ \times & \times & 0 \\ \times & \times & \times \end{bmatrix}, \quad \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix},$$

respectively, where $\times$ denotes a possibly nonzero entry; the third matrix is abbreviated $\mathrm{diag}(d_1, d_2, d_3)$. The matrix $A \in \mathbb{R}^{n \times n}$ is *symmetric* if $A^{\mathrm{T}} = A$, while $A \in \mathbb{C}^{n \times n}$ is *Hermitian* if $A^* = A$. If in addition the quadratic form $x^{\mathrm{T}} A x$ (or $x^* A x$) is always positive for nonzero vectors in $\mathbb{R}^n$ (or $\mathbb{C}^n$), then A is *positive-definite*. The term *self-adjoint* is sometimes used instead of symmetric or Hermitian. Also fundamental is the notion of orthogonality: $A \in \mathbb{R}^{n \times n}$ is *orthogonal* if $A^{\mathrm{T}} A = I$, and $A \in \mathbb{C}^{n \times n}$ is *unitary* if $A^* A = I$. These properties mean that the inverse of A is its (conjugate) transpose, but deeper properties of unitary matrices such as preservation of angles, norms, etc., under multiplication are what make them so important.

Structures can correspond to the pattern of the elements. A *Toeplitz matrix* has constant diagonals, made up from $2n - 1$ parameters a_i, $i = -(n-1), \ldots, n-1$. Thus a 5×5 Toeplitz matrix has the form

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 \\ a_{-1} & a_0 & a_1 & a_2 & a_3 \\ a_{-2} & a_{-1} & a_0 & a_1 & a_2 \\ a_{-3} & a_{-2} & a_{-1} & a_0 & a_1 \\ a_{-4} & a_{-3} & a_{-2} & a_{-1} & a_0 \end{bmatrix}.$$

Toeplitz matrices arise in SIGNAL PROCESSING [IV.35]. A *circulant matrix* is a special type of Toeplitz matrix in which each row is a cyclic permutation (one element to the right) of the row above. Circulant matrices have many special properties, including that an explicit formula exists for their inverses and their eigenvalues.

A *Hamiltonian matrix* is a $2n \times 2n$ matrix of the block form

$$\begin{bmatrix} A & F \\ G & -A^* \end{bmatrix},$$

where A, F, and G are $n \times n$ matrices and F and G are Hermitian. Hamiltonian matrices play an important role in CONTROL THEORY [III.25].

The *determinant* of an $n \times n$ matrix A is a scalar that can be defined inductively by

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij})$$

Copyrighted material

for any $i \in \{1, 2, \ldots, n\}$, where A_{ij} denotes the $(n-1) \times (n-1)$ matrix obtained from A by deleting row i and column j, and $\det(a) = a$ for a scalar a. This formula is called the expansion by minors because $\det(A_{kj})$ is a *minor* of A. The determinant is sometimes written with vertical bars, as $|A|$. Although determinants came before matrices historically, determinants have only a minor role in applied mathematics.

The quantity obtained by modifying the definition of determinant to remove the $(-1)^{i+j}$ term is the *permanent*, which is the sum of all possible products of n elements of A in which exactly one is taken from each row and each column. The permanent arises in combinatorics and in quantum mechanics.

19 Vector Spaces and Norms

A *vector space* is a mathematical structure in which a linear combination of elements can be taken, with the result remaining in the vector space. A vector space V has a binary operation, which we will write as addition, that is associative, is commutative, and has an identity (the "zero vector," written 0) and additive inverses. In other words, for any $a, b, c \in V$ we have $(a + b) + c = a + (b + c)$, $a + b = b + a$, $a + 0 = a$, and there is a d such that $a + d = 0$. There is also an underlying set of scalars, $\mathbb{R}$ or $\mathbb{C}$, such that V is closed under scalar multiplication. Moreover, for all $x, y \in V$ and scalars α, β we have $\alpha(x+y) = \alpha x + \alpha y$, $(\alpha+\beta)x = \alpha x + \beta x$, and $\alpha(\beta x) = (\alpha\beta)x$.

A vector space can take many possible forms. For example, the set of real-valued functions on an interval $[a, b]$ is a vector space over $\mathbb{R}$, and the set of polynomials of degree less than or equal to n with complex coefficients is a vector space over $\mathbb{C}$. Most importantly, the sets of n-vectors with real or complex coefficients are vector spaces over $\mathbb{R}$ and $\mathbb{C}$, respectively.

An important concept is that of linear independence. Vectors $v_1, v_2, \ldots, v_n$ in V are *linearly independent* if no nontrivial linear combination of them is zero, that is, if the equation $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$ holds only when the scalars α_i are all zero. If a collection of vectors is not linearly independent then it is *linearly dependent*.

Given vectors $v_1, v_2, \ldots, v_n$ in V we can form their *span*, which is the set of all possible linear combinations of them. A linearly independent collection of vectors whose span is V is a *basis* for V, and any vector in V can be written uniquely as a linear combination of these vectors.

The number of vectors in a basis for V is the *dimension* of V, written $\dim V$, and it can be finite or infinite. The vector space of functions mentioned above is infinite dimensional, while the vector space of polynomials of degree at most n has dimension $n + 1$, with a basis being $1, x, x^2, \ldots, x^n$ or any other sequence of polynomials of degrees $0, 1, 2, \ldots, n$.

A *subspace* of a vector space V is a subset of V that is itself a vector space under the same operations of addition and scalar multiplication.

19.1 Inner Products

Some vector spaces can be equipped with an *inner product*, which is a function $\langle x, y \rangle$ of two arguments that satisfies the conditions (i) $\langle x, x \rangle \geqslant 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$, (ii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$, (iii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$, and (iv) $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for all $x, y, z \in V$ and scalars α. The usual (Euclidean) inner product on $\mathbb{R}^n$ is $\langle x, y \rangle = x^{\mathrm{T}} y$; on $\mathbb{C}^n$ the conjugate transpose must be used: $\langle x, y \rangle = x^* y$. For the vector space $C[a, b]$ of real-valued continuous functions on $[a, b]$ an inner product is

$$\langle f, g \rangle = \int_a^b w(x) f(x) g(x) \, \mathrm{d}x, \tag{6}$$

where $w(x)$ is some given, positive weight function, while for the vector space of n-vectors of the form $[f(x_1), f(x_2), \ldots, f(x_n)]^{\mathrm{T}}$ for fixed points $x_i \in [a, b]$ and real-valued functions f an inner product is

$$\langle f, g \rangle = \sum_{i=1}^{n} w_i f(x_i) g(x_i), \tag{7}$$

where the w_i are positive weights. Note that (7) is not an inner product on the space of real-valued continuous functions because $\langle f, f \rangle = 0$ implies only that $f(x_i) = 0$ for all i and not that $f \equiv 0$.

The vector space $\mathbb{R}^n$ with the Euclidean inner product is known as *n-dimensional Euclidean space*.

19.2 Orthogonality

Two vectors u, v in an inner product space are *orthogonal* if $\langle u, v \rangle = 0$. For $\mathbb{R}^n$ and $\mathbb{C}^n$ this is just the usual notion of orthogonality: $u^{\mathrm{T}} v = 0$ and $u^* v = 0$, respectively. A set of vectors $\{u_i\}$ forms an *orthonormal set* if $\langle u_i, u_j \rangle = \delta_{ij}$ for all i and j.

For an inner product space with inner product (6) or (7), useful examples of orthogonal functions are ORTHOGONAL POLYNOMIALS [II.29], which have the important property that they satisfy a three-term recurrence relation. For example, the *Chebyshev polynomials*

Copyrighted material

T_k satisfy $T_0(x) = 1$, $T_1(x) = x$, and

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \quad k \geqslant 1, \tag{8}$$

and they are orthogonal on $[-1,1]$ with respect to the weight function $(1-x^2)^{-1/2}$:

$$\int_{-1}^{1} \frac{T_i(x)T_j(x)}{(1-x^2)^{1/2}}\,\mathrm{d}x = 0, \quad i \neq j.$$

Another commonly occurring class of orthogonal polynomials is the *Legendre polynomials* P_k, which are orthogonal with respect to $w(x) \equiv 1$ on $[-1,1]$ and satisfy the recurrence

$$P_{k+1}(x) = \frac{2k+1}{k+1}xP_k(x) - \frac{k}{k+1}P_{k-1}(x), \tag{9}$$

with $P_0(x) = 1$ and $P_1(x) = x$, when they are normalized so that $P_i(1) = 1$.

Figure 10 plots some Chebyshev polynomials and Legendre polynomials on $[-1,1]$. Both sets of polynomials are odd for odd degrees and even for even degrees. The values of the Chebyshev polynomials oscillate between -1 and 1, which is explained by the fact that $T_k(x) = \cos(k\theta)$, where $\theta = \cos^{-1} x$.

A beautiful theory surrounds orthogonal polynomials and their relations to various other areas of mathematics, including Padé approximation, spectral theory, and matrix eigenvalue problems.

If $\phi_1, \phi_2, \ldots$ is an orthogonal system, that is, $\langle \phi_i, \phi_j \rangle = 0$ for $i \neq j$, then the ϕ_i are necessarily linearly independent. Moreover, in an expansion

$$f(x) = \sum_{i=1}^{\infty} a_i \phi_i(x) \tag{10}$$

there is an explicit formula for the a_i. To determine it, we take the inner product of this equation with ϕ_j and use the orthogonality:

$$\langle f, \phi_j \rangle = \sum_{i=1}^{\infty} a_i \langle \phi_i, \phi_j \rangle = a_j \langle \phi_j, \phi_j \rangle,$$

so that $a_j = \langle f, \phi_j \rangle / \langle \phi_j, \phi_j \rangle$.

An important example of an orthogonal system of functions that are not polynomials is $1, \cos x, \sin x, \cos(2x), \sin(2x), \cos(3x), \ldots$, which are orthogonal with respect to the weight function $w(x) \equiv 1$ on $[-\pi, \pi]$, and for this basis (10) is a *Fourier series expansion.*

19.3 Norms

A common task is to approximate an element of a vector space V by the closest element in a subspace S. To define "closest" we need a way to measure the size of a vector. A norm provides such a measure.

A *norm* is a mapping $\|\cdot\|$ from V to the nonnegative real numbers such that $\|x\| = 0$ precisely when $x = 0$, $\|\alpha x\| = |\alpha|\,\|x\|$ for all scalars α and $x \in V$, and the *triangle inequality* $\|x + y\| \leqslant \|x\| + \|y\|$ holds for all $x, y \in V$. There are many possible norms, and on a finite-dimensional vector space all are *equivalent* in the sense that for any two norms $\|\cdot\|$ and $\|\cdot\|'$ there are positive constants c_1 and c_2 such that $c_1\|x\|' \leqslant \|x\| \leqslant c_2\|x\|'$ for all $x \in V$.

An example of a norm on $C[a,b]$ is

$$\|f\|_\infty = \max_{x \in [a,b]} |f(x)|, \tag{11}$$

known as the L_∞-norm, the supremum norm, the maximum norm, or the uniform norm. For $p \in [1, \infty)$,

$$\|f\|_p = \left(\int_a^b |f(x)|^p \,\mathrm{d}x \right)^{1/p}$$

is the L_p-norm on the space $L^p[a,b]$ of functions for which the (Lebesgue) integral is finite. Important special cases are the L_2-norm and the L_1-norm.

In an inner product space the natural norm is $\|x\| = \langle x, x \rangle^{1/2}$, and indeed the L_2-norm corresponds to the inner product (6) with unit weight function. A very useful inequality involving this norm is the *Cauchy–Schwarz inequality*:

$$|\langle x, y \rangle|^2 \leqslant \langle x, x \rangle \langle y, y \rangle = \|x\|^2 \|y\|^2$$

for all $x, y \in V$. This inequality shows that we can define the *angle* θ between two vectors x and y by $\cos\theta = \langle x, y \rangle / (\|x\|\,\|y\|) \in [-1,1]$. Thus orthogonality corresponds to an angle $\theta = \pm\pi/2$.

Several different norms are commonly used on the vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$. The vector p-norm is defined for real p by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 1 \leqslant p < \infty.$$

It includes the important special cases

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} = (x^* x)^{1/2},$$

$$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|.$$

The 2-norm is Euclidean length. The 1-norm is sometimes called the "Manhattan" or "taxi cab" norm, as when $x, y \in \mathbb{R}^2$ contain the coordinates of two locations in Manhattan (which has a regular grid of streets), $\|x - y\|_1$ measures the distance by taxi cab from x to y. Figure 11 shows the boundaries of the unit balls

Copyrighted material

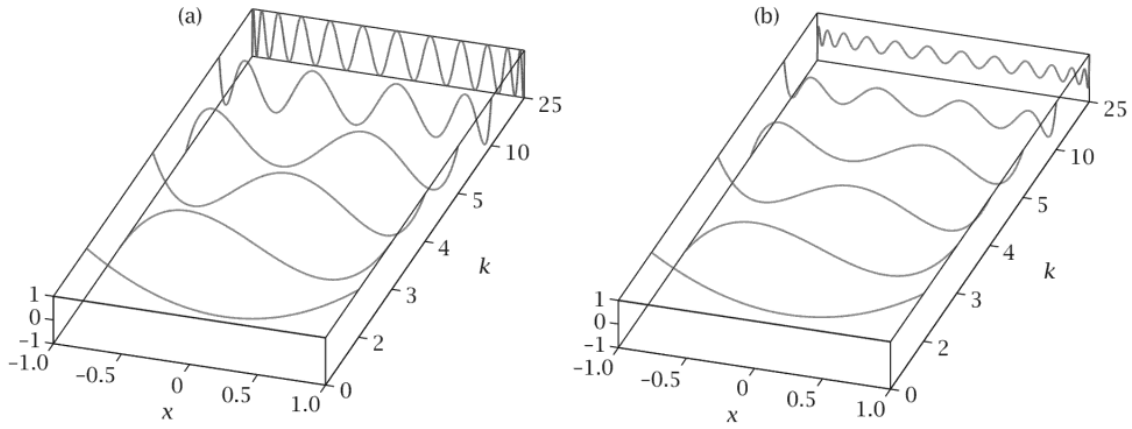


Figure 10 Selected (a) Chebyshev polynomials $T_k(x)$ and (b) Legendre polynomials $P_k(x)$ on $[-1,1]$.

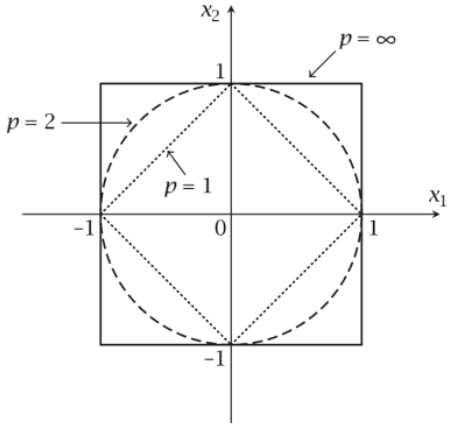


Figure 11 The boundary of the unit ball in $\mathbb{R}^2$ for the 1-, 2-, and ∞-norms.

$\{x \in \mathbb{R}^n : \|x\| = 1\}$ for the latter three p-norms. The very different shapes of the unit balls suggest that the appropriate choice of norm will depend on the problem, as is the case, for example, in DATA FITTING [IV.9 §3.2].

Related to norms is the notion of a *metric*, defined on a set M called a *metric space*. A metric on M is a nonnegative function d such that $d(x,y) = d(y,x)$ (symmetry), $d(x,z) \leqslant d(x,y) + d(y,z)$ (the *triangle inequality*), and for all $x, y, z \in M$, $d(x,y) = 0$ precisely when $x = y$. An example of a metric on the set of positive real numbers is $d(x,y) = |\log(x/y)|$. For a normed vector space, the function $d(x,y) = \|x - y\|$ is always a metric, so a normed vector space is always a metric space.

19.4 Convergence

We say that a sequence of points $x_1, x_2, \ldots$, each belonging to a normed vector space V, *converges* to a limit $x_* \in V$, written $\lim_{i\to\infty} x_i = x_*$ ($x_i \to x_*$ as $i \to \infty$), if for any $\varepsilon > 0$ there exists a positive integer N such that $\|x_* - x_i\| < \varepsilon$ for all $i \geqslant N$.

The sequence is a *Cauchy sequence* if for any $\varepsilon > 0$ there exists a positive integer N such that $\|x_i - x_j\| < \varepsilon$ for all $i, j \geqslant N$. A convergent sequence is a Cauchy sequence, but whether or not the converse is true depends on the space V.

A normed vector space is *complete* if every Cauchy sequence in V has a limit in V. A complete normed vector space is called a *Banach space*. In a Banach space we can therefore prove convergence of a sequence without knowing its limit by showing that it is a Cauchy sequence.

A complete inner product space is called a *Hilbert space*. The spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ with the Euclidean inner product are standard examples of Hilbert spaces.

20 Operators

An *operator* is a mapping from one vector space, U, to another, V (possibly the same one). A *linear operator* (or *linear transformation*) A is an operator such that $A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 A x_1 + \alpha_2 A x_2$ for all scalars α_1, α_2 and vectors $x_1, x_2 \in U$. For example, the differentiation operator is a linear operator that maps the vector space of polynomials of degree at most n to the vector space of polynomials of degree at most $n - 1$.

A natural measure of the size of a linear operator A mapping U to V is the *induced norm* (also called the

Copyrighted material

operator norm or subordinate norm),

$$\|A\| = \max\left\{\frac{\|Ax\|}{\|x\|} : x \in U,\ x \neq 0\right\},$$

where on the right-hand side $\|\cdot\|$ denotes both a norm on U (in the denominator) and a norm on V (in the numerator). For the rest of this section we assume that $U = V$ for simplicity. If $\|A\|$ is finite then A is said to be a *bounded* linear operator. On a finite-dimensional vector space all linear operators are bounded.

The definition of an operator norm yields the inequalities $\|Ax\| \leqslant \|A\|\,\|x\|$ (immediate) and $\|AB\| \leqslant \|A\|\,\|B\|$ (using the previous inequality), both of which are indispensable.

The operator A maps vectors in U to other vectors in U, and it may change the norm by as much as $\|A\|$. For some vectors, called *eigenvectors*, it is only the norm, and not the direction, that changes. A nonzero vector v is an eigenvector, with *eigenvalue* λ, if $Av = \lambda v$. Eigenvalues and eigenvectors play an important role in many areas of applied mathematics and appear in many places in this book. For example, SPECTRAL THEORY [IV.8] is about the eigenvalues and eigenvectors of linear operators on appropriate function spaces. The adjective *spectral* comes from *spectrum*, which is a set that contains the eigenvalues of an operator.

On taking norms in the relation $Av = \lambda v$ and using $\|v\| \neq 0$ we obtain $|\lambda| \leqslant \|A\|$. Thus all the eigenvalues of the operator A lie in a disk of radius $\|A\|$ centered at the origin. This is an example of a localization result.

An *invariant subspace* of an operator A that maps a vector space U to itself is a subspace X of U such that AX is a subset of X, so that $x \in X$ implies $Ax \in X$. An eigenvector is the special case of a one-dimensional invariant subspace.

For $n \times n$ matrices, the eigenvalue equation $Av = \lambda v$ says that $A - \lambda I$ is a singular matrix, which is equivalent to the condition $p(\lambda) = \det(\lambda I - A) = 0$. The polynomial p is the *characteristic polynomial* of A, and since it has degree n it follows from the fundamental theorem of algebra (section 14) that it has n roots in the complex plane, which are the eigenvalues of A. Whether there are n linearly independent eigenvectors associated with the eigenvalues depends on A and can be elegantly answered in terms of the JORDAN CANONICAL FORM [II.22]. For real symmetric and complex Hermitian matrices, the eigenvalues are all real and there is a set of n linearly independent eigenvectors, which can be taken to be orthonormal. If A is in addition positive-definite, then the eigenvalues are all positive.

For matrices on $\mathbb{C}^{m\times n}$ the operator matrix norms corresponding to the 1, 2, and ∞ vector norms have explicit formulas:

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{m} |a_{ij}|, \quad \text{“max column sum,”}$$

$$\|A\|_\infty = \max_{1\leqslant i\leqslant m} \sum_{j=1}^{n} |a_{ij}|, \quad \text{“max row sum,”}$$

$$\|A\|_2 = (\rho(A^*A))^{1/2}, \quad \text{spectral norm,}$$

where the *spectral radius*

$$\rho(B) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } B\}.$$

Another useful formula is $\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}(A)$ is the largest SINGULAR VALUE [II.32] of A. A further matrix norm that is commonly used is the Frobenius norm, given by

$$\|A\|_F = \Bigl(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\Bigr)^{1/2} = (\operatorname{trace}(A^*A))^{1/2},$$

where the *trace* of a square matrix is the sum of its diagonal elements. Note that $\|A\|_F$ is just the 2-norm of the vector obtained by stringing the columns of A out into one long vector. The Frobenius norm is not induced by any vector norm, as can be seen by taking A as the identity matrix.

21 Linear Algebra

Associated with a matrix $A \in \mathbb{C}^{m\times n}$ are four important subspaces, two in $\mathbb{C}^m$ and two in $\mathbb{C}^n$: the ranges and the nullspaces of A and A^*. The *range* of A is the set of all linear combinations of the columns: $\operatorname{range}(A) = \{Ax : x \in \mathbb{C}^n\}$. The *null space* of A is the set of vectors annihilated by A: $\operatorname{null}(A) = \{x \in \mathbb{C}^n : Ax = 0\}$.

The two most important laws of linear algebra are

$$\dim \operatorname{range}(A) = \dim \operatorname{range}(A^*),$$
$$\dim \operatorname{range}(A) + \dim \operatorname{null}(A) = n,$$

where dim denotes dimension. These equalities can be proved in various ways, one of which is via the SINGULAR VALUE DECOMPOSITION [II.32].

Suppose $x \in \operatorname{null}(A)$. Then x is orthogonal to every row of A and hence is orthogonal to the subspace spanned by the rows of A. Since the rows of A are the columns of A^*, it follows that $\operatorname{null}(A)$ is orthogonal to $\operatorname{range}(A^*)$, where two subspaces are said to be orthogonal if every vector in one of the subspaces is orthogonal to every vector in the other. In fact, it can be shown that $\operatorname{null}(A)$ and $\operatorname{range}(A^*)$ together span $\mathbb{C}^n$, and this

Copyrighted material

a "black box" that takes a vector x as input and returns the product Ax?

The problem of finding a minimum or maximum of a scalar function f of n variables provides a good example of a wide range of possible scenarios. At one extreme, f has derivatives of all orders and we can compute f and its first and second derivatives at any point (most methods do not use derivatives of higher than second order). At another extreme, f may be discontinuous and only function values may be available. It may even be that we are not able to evaluate f but only to test whether, for a given x and y, $f(x) < f(y)$ or vice versa. This is precisely the scenario for an optometrist formulating a prescription for a patient. The optometrist asks the patient to compare pairs of lenses and say which one gives the better vision. By suitably choosing the lenses the optometrist is able to home in on a prescription within a few minutes. In numerical optimization, DERIVATIVE-FREE METHODS [IV.11 §4.3] use only function values and many of them are based solely on comparisons of these values.

Another fundamental question is what is meant by a solution. If the solution is a function, would we accept its representation as an infinite series in some basis functions, or as an integral, or would we accept values of the function on a finite grid of points? If an inexact representation is allowed, how accurate must it be and what measure of error is appropriate?

2 Dimension Reduction

A common theme in many contexts is that of approximating a problem by one of smaller dimension and using the solution of the smaller problem to approximate the solution of the original problem. The motivation is that the large problem may be too expensive to solve, but of course this approach is viable only if the smaller problem can be constructed at low cost. In some situations the smaller problem is solved repeatedly, perhaps as some parameter varies, thereby amortizing the cost of producing it.

A ubiquitous example of this general approach concerns images displayed on Web pages. Modern digital cameras (even smartphones) produce images of 5 megapixels (million pixels) or more. Yet even a 27-inch monitor with a resolution of 2560×1440 pixels displays only about 3.7 megapixels. Since most images on Web pages are displayed at a small size within a page, it would be a great waste of storage and bandwidth to deal with them at their original size. They are therefore interpolated down to smaller dimensions appropriate for the intended usage (e.g., with longest side 400 pixels for an image on a news site). Here, dimension reduction is relatively straightforward and error is not an issue.

Often, though, an image is of intrinsic interest and we wish to keep it at its original dimensions and reduce the required storage, with minimal loss of quality. This is the more typical scenario for dimension reduction. The reason that dimension reduction is possible is that many images contain a high degree of redundancy. The SINGULAR VALUE DECOMPOSITION [II.32] (SVD) provides a way of capturing the important information in an image in a small number of vectors, at least for some images. A generally more effective reduction is produced by JPEG COMPRESSION [VII.7 §5], which uses two successive changes of basis in order to identify information that can be discarded.

A dynamical system may have many parameters but the behavior of interest may take place in a low-dimensional subspace. In this case we can try to identify that subspace and work within it, gaining a reduction in computation and storage. The general term for reducing dimension in a dynamical system is MODEL REDUCTION [II.26]. Model reduction has been an area of intensive research in the last thirty years, with applications ranging from the design of very large scale integration circuits to data assimilation in modeling the atmosphere.

Dimension reduction is fundamental to DATA ANALYSIS [IV.17 §4], where large data sets are transformed via a change of basis into lower-dimensional spaces that capture the behavior of the original data. Classic techniques are principal component analysis and application of the SVD. In the context of linear matrix equations such as the LYAPUNOV EQUATION [III.28], an approximation to a dominant invariant subspace of the solution (that is, an invariant subspace corresponding to the k eigenvalues of largest magnitude, for some k) can be as useful as an approximation to the whole solution, and such an approximation can often be computed at much lower cost.

A term often used in the context of dimension reduction is *curse of dimensionality*, which refers to the fact that many problems become much harder in higher dimensions and, more informally, that intuition gained from two and three dimensions does not necessarily translate to higher dimensions. A simple illustration is given by an n-sphere, or hypersphere, of radius r in $\mathbb{R}^n$, which comprises all n-vectors of 2-norm (Euclidean

Copyrighted material

norm) r. A hypersphere has volume

$$S_n = \frac{\pi^{n/2} r^n}{\Gamma(n/2+1)},$$

where Γ is the GAMMA FUNCTION [III.13]. Since $\Gamma(x) \sim \sqrt{2\pi/x}(x/e)^x$ (STIRLING'S APPROXIMATION [IV.7 §3]), for any fixed r we find that S_n tends to 0 as n tends to ∞, that is, the volume of the hypersphere tends to zero, which is perhaps surprising. For a means of comparison, consider the n-cube, or hypercube, with sides of length $2r$. It has volume

$$H_n = (2r)^n$$

and therefore $S_n/H_n \to 0$ as $n \to \infty$. In other words, most of the volume of a hypercube lies away from the enclosed hypersphere, and hence "in the corners." For $n = 2$, the ratio $S_2/H_2 = 0.785$ (see figure 11 in THE LANGUAGE OF APPLIED MATHEMATICS [I.2 §19.3]), which is already substantially less than 1. The sequence S_n/H_n continues 0.524, 0.308, 0.164, 0.081, ... This behavior is not too surprising when one realizes that any corner of the unit hypercube centered on the origin has coordinates $[\pm 1, \pm 1, \ldots, \pm 1]^T$, and so is at distance $\sqrt{n}$ from the origin, whereas any point on the unit hypersphere centered on the origin is at distance 1 from the origin, so the latter distance divided by the former tends to 0 as $n \to \infty$. The term curse of dimensionality was introduced by Richard Bellman in 1961, with reference to the fact that sampling a function of n variables on a grid with a fixed spacing requires a number of points that grows exponentially with n.

3 Approximation of Functions

We consider the problem of approximating a scalar function f, which may be given either as an explicit formula or implicitly, for example as the solution to an algebraic or differential equation. How the problem is solved depends on what is known about the function and what is required of the solution. We summarize some of the questions that must be answered before choosing a method.

- What form do we want the approximation to take: power series, polynomial, rational, Fourier series, ...?
- Do we want an approximation that has a desired accuracy within a certain region? If so, what measure of error should be used?
- Do we want an approximation that has certain qualitative features, such as convexity, monotonicity, or nonnegativity?

In this section we discuss a few examples of different types of approximation, touching on all the questions in this list. In the next three subsections f is assumed to be real (its argument being written x), whereas in the fourth subsection it can be complex (so its argument is written z). We consider first approximations based on polynomials.

3.1 Polynomials

Perhaps the simplest class of approximating functions is the polynomials, $p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$. Polynomials are easy to add, multiply, differentiate, and integrate, and their roots can be found by standard algorithms. Justification for the use of polynomials comes from *Weierstrass's theorem* of 1885, which states that for any $f \in C[a,b]$ and any $\varepsilon > 0$ there is a polynomial $p_n(x)$ such that $\|f - p_n\|_\infty < \varepsilon$, where the norm is the L_∞-NORM [I.2 §19.3] given by $\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$. Weierstrass's theorem assures us that any desired degree of accuracy in the maximum norm can be obtained by polynomials, though it does not bound the degree n, which may have to be high. Here are some of the ways in which polynomial approximations are constructed.

Truncated Taylor series. If f is sufficiently smooth that it has a Taylor series expansion and its derivatives can be evaluated, then a polynomial approximation can be obtained simply by truncating the Taylor series. The Taylor series with remainder tells us that we can write $f(x) = p_n(x) + E_n(x)$, where $p_n(x) = f(0) + f'(0)x + \cdots + f^{(n)}(0)x^n/n!$ is a degree-n polynomial and the remainder term has the form $E_n(x) = f^{(n+1)}(\xi)x^{n+1}/(n+1)!$ for some ξ on the interval with endpoints 0 and x. The value of n and the range of x for which the approximation $f(x) \approx p_n(x)$ is applied will depend on f and the desired accuracy. Figure 1 shows the degree-1, degree-3, and degree-5 Taylor approximants to the sine function.

Interpolation. We may require $p_n(x)$ to agree with $f(x)$ at certain specified points $x_i \in [a,b]$. Since p_n contains $n+1$ coefficients and each condition $p_n(x_i) = f(x_i)$ provides one equation, we need $n+1$ points in order to specify p_n. It can be shown that the $n+1$ interpolation equations in $n+1$ unknowns have a unique solution provided that the interpolation points $\{x_i\}_{i=0}^n$ are distinct, in which case there is a unique *interpolating polynomial*. There is a variety of ways of representing p_n (e.g., Lagrange form, barycentric form, and

Copyrighted material

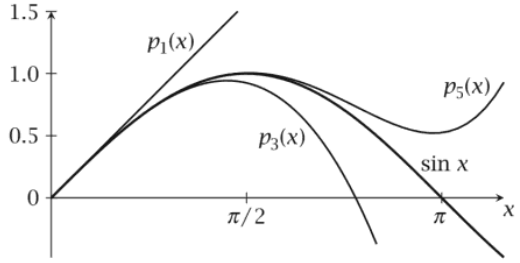


Figure 1 $\sin x$ and its Taylor approximants $p_1(x) = x$, $p_3(x) = x - x^3/3!$, and $p_5(x) = x - x^3/3! + x^5/5!$.

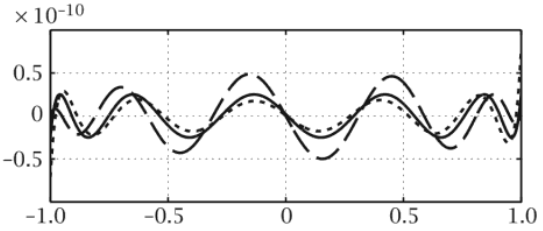


Figure 2 Error in polynomial approximations to e^x on $[-1, 1]$: solid line, L_∞ approximation; dashed line, Chebyshev interpolant; dotted line, least squares (L_2 approximation).

divided difference form). An explicit formula is available for the error: if f has $n+1$ continuous derivatives on $[a, b]$ then for any $x \in [a, b]$

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^{n} (x - x_i),$$

where ξ_x is some unknown point in the interval determined by $x_0, x_1, \ldots, x_n$, and x. This error formula can be used to obtain insight into how to choose the x_i. It turns out that equally spaced points are poor, whereas points derived by rescaling to $[a, b]$ the ZEROS OR EXTREMA OF THE CHEBYSHEV POLYNOMIAL [IV.9 §2.2] of degree $n+1$ or n, respectively, are good.

Least-squares approximation. In least-squares approximation we fix the degree n and then choose the polynomial p_n to minimize the L_2-norm

$$\left(\int_a^b |f(x) - p_n(x)|^2 \, dx \right)^{1/2},$$

where $[a, b]$ is the interval of interest. It turns out that there is a unique p_n minimizing the error, and its coefficients satisfy a linear system of equations called the *normal equations*. The normal equations tend to be ill-conditioned when p_n is represented in the *monomial basis*, $\{1, x, x^2, \ldots\}$, so in this context it is usual to write $p_n = \sum_{i=0}^{n} a_i \phi_i(x)$, where the ϕ_i are ORTHOGONAL POLYNOMIALS [II.29] on $[a, b]$. In this case the normal equations are diagonal and there is an explicit expression for the optimal coefficients: $a_i = \int_a^b \phi_i(x) f(x) \, dx / \int_a^b \phi_i(x)^2 \, dx$.

L_∞ approximation. Instead of using the L_2-norm we can use the L_∞-norm and so minimize $\|f - p_n\|_\infty$. A best L_∞ approximation always exists and is unique, and there is a beautiful theory that characterizes the solution in terms of *equioscillation*, whereby the error achieves its maximum magnitude at a certain number of points with alternating sign. An algorithm called the Remez algorithm is available for computing the best L_∞ approximation. One use of it is in EVALUATING ELEMENTARY FUNCTIONS [VI.11].

Figure 2 plots the absolute error $|f - p_n(x)|$ in three degree-10 polynomial approximations to e^x on $[-1, 1]$: the least-squares approximation; the L_∞ approximation; and a polynomial interpolant based on the Chebyshev points, $\cos(j\pi/n)$, $j = 0{:}n$. Note that the L_∞ approximation has equioscillating error with maximum error strictly less than that for the other two approximations, and that the error of the Chebyshev interpolant is zero at the eleven points where it interpolates, which include the endpoints. It is also clear that the Chebyshev approximation is not much worse than the L_∞ one—something that is true in general.

3.2 Piecewise Polynomials

High-degree polynomials have a tendency to wiggle. A degree-100 polynomial p has up to 100 points at which it crosses the x-axis on a plot of $y = p(x)$: the distinct real zeros of p. This can make high-degree polynomials unsatisfactory as approximating functions. Instead of using one polynomial of large degree it can be better to use many polynomials of low degree. This can be done by breaking the interval of interest into pieces and using a different low-degree polynomial on each piece, with the polynomials joined together smoothly to make up the complete approximating function. Such *piecewise polynomials* can produce functions with high approximating power while avoiding the oscillations possible with high-degree polynomials.

A trivial example of a piecewise polynomial is the absolute value function $|x|$, which is equal to $-x$ for $x \leqslant 0$ and x for $x \geqslant 0$ (see figure 4 on p. 13 in THE LANGUAGE OF APPLIED MATHEMATICS [I.2]). More generally, a piecewise polynomial g defined on an interval

Copyrighted material

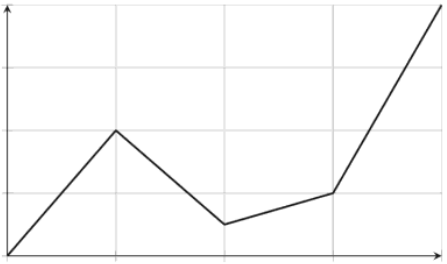


Figure 3 A piecewise-linear function (spline).

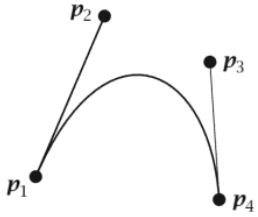


Figure 4 A cubic Bézier curve with four control points $\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4$.

$[a,b] =: [x_0, x_n]$ that is the union of n subintervals $[x_0,x_1], [x_1,x_2], \ldots, [x_{n-1},x_n]$ is defined by the property that $g(x) = p_i(x)$ for $x \in [x_i, x_{i+1}]$, where each p_i is a polynomial. Thus on each interval g is a polynomial, but each of these individual polynomials is in general different and possibly of different degree. Such a function is generally discontinuous, but we can ensure continuity by insisting that $p_{i-1}(x_i) = p_i(x_i)$, $i = 1:n-1$.

Important examples of piecewise polynomials are *splines*, which are piecewise polynomials g for which each individual polynomial has degree k or less and for which g has $k-1$ continuous derivatives on the interval. A spline therefore has the maximum possible smoothness. The most commonly used splines are linear splines and cubic splines, and an important application is in the FINITE-ELEMENT METHOD [II.12]. Figure 3 shows an example of a linear spline. Splines are commonly used in plotting data, where they provide a way of "joining up the dots," e.g., by straight lines in the case of a linear spline.

In computer-aided design the individual polynomials in a piecewise polynomial are often constructed as *Bézier curves*, which have the form

$$B_n(x) = \sum_{i=0}^{n} \binom{n}{i} \frac{(b-x)^{n-i}(x-a)^i}{(b-a)^n} \boldsymbol{p}_i$$

for an interval $[a,b]$. The $\boldsymbol{p}_i$ are control points in the plane that the user chooses via a graphical interface in order to achieve a desired form of curve. Figure 4 shows a cubic Bézier curve. The polynomials that multiply the $\boldsymbol{p}_i$ are called *Bernstein polynomials*, and they were originally introduced by Bernstein in 1912 in order to give a constructive proof of Weierstrass's theorem. The use of Bézier curves as a design tool to intuitively construct and manipulate complex shapes was initiated at the Citroën and Renault car companies in the 1960s. Today, cubic Bézier curves are widely used, e.g., in the design of fonts, in image manipulation programs such as Adobe Photoshop, and in the ISO standard for the Portable Document Format (PDF).

3.3 Wavelets

FOURIER ANALYSIS [I.2 §19.2] decomposes a function into a linear combination of trigonometric functions (sines and cosines) with different frequencies and so is a natural way to deal with periodic functions. Wavelet analysis, which was first developed in the 1980s, is designed to handle nonperiodic functions and does so by using basis functions that are rough and localized. Rather than varying the frequency as with the Fourier basis, a wavelet basis is constructed by translation ($f(x) \to f(x-1)$) and dilation ($f(x) \to f(2x)$). Given a mother wavelet $\psi(x)$, which has compact support (that is, it is zero outside a bounded interval), translations and dilations are created as $\psi(2^n x - k)$ with integer n and k. This leads to many different resolutions, and hence the term *multiresolution analysis* is used in this context. Larger n correspond to finer resolutions, and as k varies the support moves around.

The localized nature of the wavelet basis functions makes wavelet representations of many functions and data relatively sparse, which makes wavelets particularly suitable for data compression, detection of features in images (such as edges and other discontinuities), and noise reduction. These are some of the reasons for the success of wavelets in (for example) imaging, where they are used in the JPEG2000 STANDARD [VII.7 §5].

3.4 Series Solution

We now turn to the development of explicit series representations of a function. As an example we take the Airy function $w(z)$, which satisfies the differential equation

$$w'' - zw = 0.$$

Copyrighted material

We can look for a solution $w(z) = \sum_{k=0}^{\infty} a_k z^k$, where $a_0 = w(0)$ and $a_1 = w'(0)$ can be regarded as given. For simplicity we will take $a_0 = 1$ and $a_1 = 0$. Differentiating twice gives $w''(z) = \sum_{k=2}^{\infty} k(k-1)a_k z^{k-2}$. Substituting the power series for w and w'' into the differential equation we obtain $\sum_{k=2}^{\infty} k(k-1)a_k z^{k-2} - \sum_{k=0}^{\infty} a_k z^{k+1} = 0$. Since this equation must hold for all z we can equate coefficients of $z^0, z^1, z^2, \dots$ on both sides to obtain a sequence of equations that provide recurrence relations for the a_k, specifically $(k+1)(k+2)a_{k+2} = a_{k-1}$ along with $a_2 = 0$. We find that

$$w(z) = 1 + \frac{z^3}{6} + \frac{z^6}{180} + \frac{z^9}{12\,960} + \cdots.$$

The modulus of the ratio of successive nonzero terms tends to zero as the index of the terms tends to infinity, which ensures that the series is convergent for all z. Since a power series can be differentiated term by term within its radius of convergence, it follows that our series does indeed satisfy the Airy equation.

Constructing a series expansion does not always lead to a convergent series. Consider the exponential integral

$$E_1(z) = \int_z^{\infty} \frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t.$$

Integrating by parts repeatedly gives

$$\begin{aligned} E_1(z) &= \frac{\mathrm{e}^{-z}}{z} - \int_z^{\infty} \frac{\mathrm{e}^{-t}}{t^2}\,\mathrm{d}t \\ &= \frac{\mathrm{e}^{-z}}{z} - \frac{\mathrm{e}^{-z}}{z^2} + 2\int_z^{\infty} \frac{\mathrm{e}^{-t}}{t^3}\,\mathrm{d}t \\ &= \frac{\mathrm{e}^{-z}}{z}\left(1 - \frac{1}{z} + \frac{2!}{z^2} + \cdots + (-1)^{k-1}\frac{(k-1)!}{z^{k-1}}\right) + R_k. \end{aligned}$$

The remainder term, $R_k = (-1)^k k! \int_z^{\infty} (\mathrm{e}^{-t}/t^{k+1})\,\mathrm{d}t$, does not tend to zero as $k \to \infty$ for fixed z, so the series is not convergent. Nevertheless, $|R_k|$ does decrease with k before it increases, and a reasonable approximation to $E_1(z)$ can be obtained by choosing a suitable value of k. For example, with $z = 10$ the remainder starts increasing at $k = 11$, and taking $k = 10$ we obtain the approximation $E_1(10) \approx 4.156 \times 10^{-6}$, which is to be compared with $E_1(10) = 4.157 \times 10^{-6}$, where both results have been rounded to four significant figures. The series above is an example of an *asymptotic series*. In general, we say that the series $\sum_{k=0}^{\infty} a_k z^{-k}$ is an *asymptotic expansion* of f as $z \to \infty$ if

$$\lim_{z\to\infty} z^n \left(f(z) - \sum_{k=0}^{n} a_k z^{-k}\right) = 0$$

for every n, and we write $f(z) \sim \sum_{k=0}^{\infty} a_k z^{-k}$, where the symbol "$\sim$" is read as "is asymptotic to." This condition can also be written as

$$f(z) = \sum_{k=0}^{n-1} a_k z^{-k} + O(z^{-n}).$$

For the series for E_1 we have

$$\begin{aligned} |z^k R_k| &= |z|^k k! \left| \int_z^{\infty} \frac{\mathrm{e}^{-t}}{t^{k+1}}\,\mathrm{d}t \right| \\ &\leqslant \frac{k!}{|z|} \left| \int_z^{\infty} \mathrm{e}^{-t}\,\mathrm{d}t \right| = \frac{k!}{|z|}|\mathrm{e}^{-z}|, \end{aligned}$$

and the latter bound tends to zero as $|z| \to \infty$ if $\arg z \in (-\pi/2, \pi/2)$, so the series is asymptotic under this constraint on z.

By summing an appropriate number of terms, asymptotic series can deliver approximations of up to a certain, possibly good, accuracy, for large enough $|z|$, but beyond a certain point the accuracy worsens.

Suppose we have the quadratic $q_\varepsilon(x) = x^2 - x + \varepsilon = 0$, where ε is a small parameter and we wish to obtain a series expansion for x as a function of ε. This can be done by substituting $x(\varepsilon) = \sum_{k=0}^{\infty} a_k \varepsilon^k$ into the equation and setting the coefficients of each power of ε to zero. This produces a system of equations that can be used to express $a_1, a_2, \dots$ in terms of a_0. The two solutions of $q_\varepsilon(x) = 0$ for $\varepsilon = 0$ are 0 and 1, so we take $a_0 = 0, 1$ and obtain the series

$$x(\varepsilon) = \begin{cases} \varepsilon + \varepsilon^2 + 2\varepsilon^3 + \cdots, & a_0 = 0, \\ 1 - \varepsilon - \varepsilon^2 - 2\varepsilon^3 + \cdots, & a_0 = 1, \end{cases} \tag{1}$$

which describe how the roots 0 and 1 of $q_0(x)$ behave for small ε. Suppose now that it is the leading term that is small and that we have the quadratic $\tilde{q}_\varepsilon(x) = \varepsilon x^2 - x + 1 = 0$. If we repeat the process of looking for an expansion of $x(\varepsilon)$, we obtain $x(\varepsilon) = 1 + \varepsilon + 2\varepsilon^2 + 5\varepsilon^3 + \cdots$ describing the behavior of the root 1 of $\tilde{q}_0(x)$. But $\tilde{q}$ is a quadratic and so has two roots. What has happened to the other one? There is a change of degree as we go from $\varepsilon = 0$ to $\varepsilon \neq 0$, and this takes us into SINGULAR PERTURBATION THEORY [IV.5 §3.2]. In this simple case we can use the transformation $y = 1/x$ to write $\tilde{q}_\varepsilon(x) = q_\varepsilon(y)/y^2$, and so we obtain expansions for $x(\varepsilon)$ by inverting those in (1). Indeed, inverting the second expression in (1) and expanding in a power series recovers the expansion we just derived.

4 Symbolic Solution

Sometimes a useful representation of a solution can be obtained using a computer symbolic manipulation package. Such packages are, for example, very good at determining closed forms for indefinite integrals that

Copyrighted material

problem

$$y'(x) = f(x, y), \qquad y(0) = y_0.$$

Integrating between 0 and x leads to the equivalent problem

$$y(x) = y_0 + \int_0^x f(x, y(x))\,\mathrm{d}x,$$

which is a type of equation known as an INTEGRAL EQUATION [IV.4] because the unknown function occurs within an integral. Applying the fixed-point iteration idea we can make a guess ϕ_0 for y, plug it into the right-hand side of the integral equation, and call the result ϕ_1. The process can be iterated to produce a sequence of functions ϕ_k defined by

$$\phi_{k+1}(x) = y_0 + \int_0^x f(x, \phi_k(x))\,\mathrm{d}x, \quad k \geqslant 1.$$

In general, none of the ϕ_k will satisfy the differential equation, but we might hope that the sequence has a limit that does. Let us try out this idea on the problem

$$y' = 2x(1 + y), \qquad y(0) = 0$$

using first guess $\phi_0(x) = 0$. Then $\phi_1(x) = \int_0^x 2x\,\mathrm{d}x = x^2$ and $\phi_2(x) = \int_0^x 2x(1 + x^2)\,\mathrm{d}x = x^2 + x^4/2$. Continuing in this fashion yields $\phi_k(x) = x^2 + x^4/2! + x^6/3! + \cdots + x^{2k}/k!$. The limit as $k \to \infty$ exists and is $\mathrm{e}^{x^2} - 1$, which is the required solution.

The procedure we have just carried out is known as *Picard iteration*, or the method of successive approximation. Of course, in most cases it will not be possible to evaluate the integrals in closed form, and so Picard iteration is not a practical means for computing a solution. However, Picard iteration is the basis of the proof of the standard result on existence and uniqueness of solutions for ODEs. The result says that, if $f(x, y)$ is continuous for $x \in [a, b]$ and for all y and satisfies a *Lipschitz condition*

$$|f(x, u) - f(x, v)| \leqslant L|u - v| \quad \forall x \in [a, b],\ \forall u, v,$$

with Lipschitz constant L, then for any y_0 there is a unique continuously differentiable function $y(x)$ defined on $[a, b]$ that satisfies $y' = f(x, y)$ and $y(a) = y_0$.

7 Conversion to Another Problem

When we cannot solve a problem it can be useful to convert it to a different problem that is more amenable to attack. In this section we give several examples of such conversions.

We note first that it is not always obvious what is meant by a solution to a problem. Consider the ODE problem

$$\frac{\mathrm{d}y}{\mathrm{d}x} = 1 - 2xy, \qquad y(0) = 0.$$

The solution y can be written as

$$y(x) = \mathrm{e}^{-x^2} \int_0^x \mathrm{e}^{t^2}\,\mathrm{d}t,$$

which is known as *Dawson's integral* or *Dawson's function*. Which representation of y is better? If we need to obtain higher derivatives $\mathrm{d}^k y/\mathrm{d}x^k$, the differential equation is more convenient. To evaluate $y(x)$ for a given x, numerical methods can be applied to either representation. Both representations therefore have their uses.

7.1 Uncoupling

When we are solving equations, of whatever type, a particularly favorable circumstance is when the first equation involves only one unknown and each successive equation introduces only one new unknown. We can then solve the equations from first to last. The simplest example is a triangular system of linear equations, such as

$$\begin{aligned} a_{11}x_1 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3, \end{aligned}$$

which can be solved by finding x_1 from the first equation, then x_2 from the second, and finally x_3 from the third. This is the process known as *substitution*.

Most linear equation problems do not have this triangular structure, but the process of GAUSSIAN ELIMINATION [IV.10 §2] converts an arbitrary linear system into triangular form.

More generally we might have n nonlinear equations in n unknowns, and a natural way to solve them is to try to manipulate them into an analogous triangular form. In computer algebra a way of doing this for polynomial equations is provided by Buchberger's algorithm for computing a GRÖBNER BASIS [IV.39 §2.1].

A triangular problem is partially uncoupled. In a fully uncoupled system each equation contains only one unknown. A linear system of ODEs $y' = Ay$ with an $n \times n$ coefficient matrix A can be uncoupled if A is diagonalizable. Indeed, if $A = XDX^{-1}$ with X nonsingular and $D = \mathrm{diag}(\lambda_i)$, then the transformation $z = X^{-1}y$ gives $z' = Dz$, which represents n uncoupled scalar

Copyrighted material

equations $z_i' = \lambda_i z_i$, $i = 1:n$. The behavior of the vector y can now be understood by looking at the behavior of the n independent scalars z_i.

7.2 Polynomial Roots and Matrix Eigenvalues

Consider the problem of finding the roots (zeros) of a polynomial $p_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ with $a_n \neq 0$, that is, the values of x for which $p_n(x) = 0$. It is known from Galois theory that there is no explicit formula for the roots when $n \geqslant 5$. Many methods are available for computing polynomial roots, but not all are able to compute all n roots reliably and software might not be readily available. Consider the $n \times n$ matrix

$$C = \begin{bmatrix} -a_{n-1}/a_n & -a_{n-2}/a_n & \cdots & \cdots & -a_0/a_n \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & 0 \\ \vdots & & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \end{bmatrix}.$$

Let λ be a root of p_n. For the vector defined by $y = [\lambda^{n-1}\ \lambda^{n-2}\ \cdots\ 1]^T$ we have $Cy = \lambda y$, so λ is an eigenvalue of C with eigenvector y. In fact, the set of roots of p is the set of eigenvalues of C, so the polynomial root problem has been converted into an eigenvalue problem—albeit a specially structured one. The matrix C is called a *companion matrix*. Of course, one can go in the opposite direction: to find the eigenvalues of C one might look for solutions of $\det(C - \lambda I) = 0$, and the determinant is precisely $(-1)^n p_n(\lambda)/a_n$.

The eigenvector problem $Ax = \lambda x$ can be converted into a nonlinear system of equations $F(v) = 0$, where

$$F(v) = \begin{bmatrix} (A - \lambda I)x \\ e_s^T x - 1 \end{bmatrix}, \qquad v = \begin{bmatrix} x \\ \lambda \end{bmatrix}.$$

The last component of F serves to normalize the eigenvector and here s is some fixed integer, with e_s denoting the sth column of the identity matrix. By solving $F(v) = 0$ we obtain both an eigenvalue of A and the corresponding eigenvector.

7.3 Dubious Conversions

Converting one problem to an apparently simpler one is not always a good idea. The problem of solving the scalar nonlinear equation $f(x) = 0$ can be converted to the problem of minimizing the function $g(x) = f(x)^2$. Since the latter problem has a global minimum attained when $f(x) = 0$, the conversion might look attractive. However, it has a pitfall: since $g'(x) = 2f'(x)f(x)$, the derivative of g is zero whenever $f'(x) = 0$, and this means that methods for minimizing g might converge to points that are stationary points of g but not zeros of f.

For another example, consider the generalized eigenproblem in $n \times n$ matrices A and B, $Ax = \lambda Bx$, which arises in problems in engineering and physics. It is natural to attempt to convert it to the standard eigenproblem $B^{-1}Ax = \lambda x$ and then apply standard theory and algorithms. However, if B is singular this transformation is not possible, and when B is nonsingular but ILL-CONDITIONED [I.2 §22] the transformation is inadvisable in floating-point arithmetic as it will be numerically unstable. A further drawback is that if B is SPARSE [IV.10 §6] (has many zeros) then $B^{-1}A$ can have many more nonzeros than A or B.

7.4 High-Order Differential Equations

Methods of solution of differential equations have been more extensively developed for first-order equations than for higher-order ones, where order refers to the highest derivative in the equation. Fortunately, higher-order equations can always be converted to first-order ones. Consider the qth-order ODE

$$y^{(q)} = f(t, y, y', \ldots, y^{(q-1)})$$

with $y, y', \ldots, y^{(q-1)}$ given at $t = t_0$. Define new variables

$$z_1 = y, \quad z_2 = y', \quad \ldots, \quad z_q = y^{(q-1)}.$$

Then we have the first-order system of equations

$$\begin{aligned} z_1' &= z_2, \\ z_2' &= z_3, \\ &\vdots \\ z_{q-1}' &= z_q, \\ z_q' &= f(t, z_1, z_2, \ldots, z_q), \end{aligned}$$

with $z_1, z_2, \ldots, z_q$ given at $t = t_0$. We can write this system in vector form:

$$z' = f(t, z), \quad z = [z_1, z_2, \ldots, z_q]^T. \tag{4}$$

So we have traded high order for high dimension. Fortunately, the theory and the numerical methods developed for scalar first-order ODEs generally carry over straightforwardly to vector ODEs. We can go further and remove the explicit time dependence from (4) to put the system in *autonomous form*: with $w = [t, z^T]^T$, we have

$$w' = \begin{bmatrix} 1 \\ f(z) \end{bmatrix} = \begin{bmatrix} 1 \\ f(w_2, \ldots, w_n) \end{bmatrix} =: g(w).$$

Copyrighted material

7.5 Continuation

Suppose we have a hard problem "solve $f(x) = 0$" and another problem "solve $g(x) = 0$" that is trivial to solve. Consider the parametrized problem "solve $h(x,t) = tf(x) + (1-t)g(x) = 0$." We know the solution for $t = 0$ and wish to find it for $t = 1$. The idea of *continuation* (also called *homotopy*, or *incremental loading* in elasticity) is to traverse the interval from 0 to 1 in several steps: $0 < t_1 < t_2 < \cdots < t_n = 1$. On the kth step we use the solution x_{k-1} of the problem $h(x,t_{k-1}) = 0$ as the starting point for an iteration for solving $h(x,t_k) = 0$. We are therefore solving the original problem by approaching it gradually from a trivial problem. Continuation cannot be expected to work well in all cases. It is particularly well suited to cases where f already depends on a parameter and the problem is simpler for some value of that parameter.

Continuation is a very general technique and has close connections with BIFURCATION THEORY [IV.21]. A special case of it is the idea of SHRINKING [V.10 §2.2], whereby a convex combination is taken of a given object with another having more desirable properties.

8 Linearization

A huge body of mathematics is concerned with problems that are linear in the variables of interest, such as a system $Ax = b$ of n linear equations in n unknowns or a system of ODEs $\mathrm{d}y/\mathrm{d}t = A(t)y$. For linear problems it is usually easy to analyze the existence of solutions, to obtain an explicit formula for a solution, and to derive practical methods of solution that exploit the linearity. Unfortunately, many real-world processes are inherently nonlinear. This means, first of all, that it may not be easy to determine whether or not there is a solution at all or, if a solution exists, whether it is unique. Secondly, finding a solution is in general difficult. A general technique for solving nonlinear problems is to transform them into linear ones, thereby converting a problem that we cannot solve into one that we can. The transformation can rarely be done exactly, so what is usually done is to *approximate* the nonlinear problem by a linear one—the process of *linearization*—and carry out some sort of iteration or refinement process.

To illustrate the idea of linear approximations we consider the quadratic equation

$$x^2 - 10x + 1 = 0. \tag{5}$$

Because the coefficient of the linear term, 10, is large compared with that of the quadratic term, 1, we can think of (5) as a linear equation with a small quadratic perturbation:

$$x = \frac{1}{10} + \frac{x^2}{10}. \tag{6}$$

Indeed, if we solve the linear part we obtain $x = 1/10$, which leaves a residual of just $1/100$ when substituted into the left-hand side of (5). We can therefore say that $x \approx 1/10$ is a reasonable approximation to a root (in fact, to the smallest root, since the product of the roots must be 1). Note that this approximation is obtained by putting $x = 0$ in the right-hand side of (6). To obtain a better approximation we might try putting $x = 1/10$ into the right-hand side. Repeating this process leads to the fixed-point iteration

$$x_{k+1} = \frac{1 + x_k^2}{10}, \qquad x_0 = 0,$$

which yields 0, 0.10, 0.101, ... After ten iterations we have $x_{10} = 0.101020514433644$, which is correct to the fifteen significant digits shown. Of course we could have obtained this solution as $x = 5 - \sqrt{24}$ using the quadratic formula, but the linearization approach gives an instant approximation and provides insight. For the equation $x^7 - 10x + 1 = 0$, for which there is no explicit formula for the roots, $1/10$ is an even better approximation to the smallest root and the analogue of the iteration above converges even more quickly.

Linearization is the key concept underlying NEWTON'S METHOD [II.28], which we discussed in section 6. Suppose we wish to solve a nonlinear system $f(x) = 0$, where $f\colon \mathbb{R}^n \to \mathbb{R}^n$, and let x be an approximation to a solution x_*. Writing $x_* = x + h$, for sufficiently smooth f we have $0 = f(x_*) = f(x) + J(x)h + O(\|h\|^2)$, where $J(x) = (\partial f_i/\partial x_j) \in \mathbb{R}^{n\times n}$ is the *Jacobian matrix* and the big-oh term includes the second- and higher-order terms from a multidimensional Taylor series. Newton's method approximates f by the linear part of the series and so solves the linear system $J(x)h = -f(x)$ in order to produce a new approximation $x + h$. The process is iterated, yielding $x_{k+1} = x_k - J(x_k)^{-1}f(x_k)$. Theorems are available that guarantee when the linear approximations of the Newton method are good enough to ensure convergence to a solution. Indeed the Newton–Kantorovich theorem even uses Newton's method itself to prove the existence of a solution under certain conditions.

An equilibrium point (or critical point) of a nonlinear autonomous system of ODEs $y'(t) = f(y)$, where $f\colon \mathbb{R}^n \to \mathbb{R}^n$, is a vector y_0 such that $f(y_0) = 0$. For

Copyrighted material

such a point, $y(t) = y_0$ is a constant solution to the differential equations. Linear stability analysis determines the effect of small perturbations away from the equilibrium point. Let $y(t) = y_0 + h(t)$ with $h(0) = h_0$ small. We wish to determine the behavior of $h(t)$ as $t \to \infty$. A linear approximation to f at y_0 yields $h'(t) = y'(t) = f(y_0) + J(y_0)h = J(y_0)h$. The solution to this first-order system is $h(t) = \mathrm{e}^{J(y_0)t}h_0$, and so the behavior of h depends on the behavior of the MATRIX EXPONENTIAL [II.14] $\mathrm{e}^{J(y_0)t}$. In particular, whether or not $h(t)$ grows or decays as $t \to \infty$ depends on the real parts of the eigenvalues of $J(y_0)$. For the case where y has two components ($n = 2$), it is possible to give detailed classifications and plots (called phase-plane portraits) of the different qualitative behaviors that can occur. For more on the stability of ODEs see ORDINARY DIFFERENTIAL EQUATIONS [IV.2 §§8, 9].

An example of a nonlinear problem that can be linearized exactly, without any approximation, is the QUADRATIC EIGENVALUE PROBLEM [IV.10 §5.8].

Many other uses of linearization can be found throughout this book.

9 Recurrence Relations

A useful tactic for solving a problem whose solution is a number or function depending on a parameter is to try to derive a recurrence. For example, consider the integral

$$x_n = \int_0^1 \frac{t^n}{t+5}\,\mathrm{d}t.$$

It is easy to verify that x_n satisfies the recurrence $x_n + 5x_{n-1} = 1/n$ and $x_0 = \log(6/5)$, so values of x_n can easily be generated from the recurrence. However, when evaluating a recurrence numerically, one always needs to be aware of possible instability. Evaluating the recurrence in IEEE double-precision arithmetic (corresponding to about sixteen significant decimal digits) we find that $\hat{x}_{21} = -0.0159\ldots$, where the hat denotes the computed result. But

$$\frac{1}{6(n+1)} = \int_0^1 \frac{t^n}{6}\,\mathrm{d}t < x_n < \int_0^1 \frac{t^n}{5}\,\mathrm{d}t = \frac{1}{5(n+1)}$$

for all n, so this result is clearly not even of the right sign. The cause of the inaccuracy can be seen by considering the ideal case in which the only error, ε, say, occurs in evaluating x_0. That error is multiplied by -5 in computing x_1 and by a further factor of -5 on each step of the recurrence; overall, x_n will be contaminated by an error of $(-5)^n\varepsilon$. This is an example of *numerical instability* and it is something that recurrences are prone to. We can obtain a more accurate result by using the recurrence in the backward direction, which will result in errors being divided by -5, so that they are damped out. From the inequalities above we see that for large n, $x_n \approx 1/(5(n+1))$. Let us simply set $y_{20} = 1/105$. Then, using the recurrence backward in the form $x_{n-1} = (1/n - x_n)/5$, we find that x_0 is computed with a relative error of order 10^{-16}. For similar reasons, the recurrence relation in THE LANGUAGE OF APPLIED MATHEMATICS [I.2 §13] for the Bessel functions is also used in the backward direction for $x < n$.

10 Lagrange Multipliers

Optimization problems abound in applied mathematics because in many practical situations one wishes to maximize a desirable attribute (e.g., profit, or the strength of a structure) or minimize something that is desired to be small (such as cost or energy). More often than not, constraints impose limits on the variables and help to balance conflicting requirements. For example, in designing a tripod for cameras we may wish to minimize the weight of the tripod subject to it being able to support cameras up to a certain maximal weight, and a constraint might be a lower bound on the maximal height of the tripod.

Calculus enables us to characterize and find maxima and minima of functions. In the presence of constraints, though, the standard results are not so helpful. Consider the problem in three variables

$$\begin{aligned} &\text{minimize } f(x_1, x_2, x_3) \\ &\text{subject to } c(x_1, x_2, x_3) = 0, \end{aligned} \tag{7}$$

where the objective function f and constraint function c are scalars. We know that any minimizer of the unconstrained problem $\min f(x_1, x_2, x_3)$ has to have a zero gradient; that is, $\nabla f(x) = [\partial f/\partial x_1, \partial f/\partial x_2, \partial f/\partial x_3]^{\mathrm{T}}$ must be the zero vector. How can we take account of the constraint $c(x_1, x_2, x_3) = 0$?

Let $x_* \in \mathbb{R}^3$ be a *feasible point*, that is, a point satisfying the constraint $c(x_*) = 0$. Consider a smooth curve $z(t)$ with $z(0) = x_*$ that remains on the constraint, that is, $c(z(t)) = 0$ for all sufficiently small t. Differentiating the latter equation and using the chain rule gives $(\mathrm{d}z(t)/\mathrm{d}t)^{\mathrm{T}}\nabla c(z(t)) = 0$. Setting $t = 0$ and putting $p_* = \mathrm{d}z/\mathrm{d}t|_{t=0}$ gives

$$p_*^{\mathrm{T}}\nabla c(x_*) = 0. \tag{8}$$

Copyrighted material

For x_* to be optimal, the rate of change of f along z must be zero at x_*, so, using the chain rule again,

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} f(z(t))\Big|_{t=0} = \sum_{i=1}^{3} \frac{\partial f}{\partial z_i} \frac{\mathrm{d}z_i}{\mathrm{d}t}\Big|_{t=0}$$
$$= \nabla f(x_*)^{\mathrm{T}} p_*. \tag{9}$$

Now assume that $\nabla c(x_*) \neq 0$, which is known as a *constraint qualification*. This assumption ensures that every vector p_* satisfying (8) is the tangent at $t = 0$ to some curve $z(t)$. It then follows that since (8) and (9) hold for all p_*,

$$\nabla f(x_*) = \lambda_* \nabla c(x_*) \tag{10}$$

for some scalar λ_*. The scalar λ_* is called a *Lagrange multiplier*. The constraint equation $c(x) = 0$ and (10) together constitute four equations in four unknowns, x_1, x_2, x_3, and λ. We have therefore reduced the original constrained minimization problem to a nonlinear system of equations. The latter system can be solved by any means at our disposal, though being nonlinear it is not necessarily an easy problem.

Another way to express our findings is in terms of the *Lagrangian function* $L(x, \lambda) = f(x) - \lambda c(x)$. Since $\nabla_x L(x, \lambda) = \nabla f(x) - \lambda \nabla c(x)$, the Lagrange multiplier condition (10) says that the solution x_* is a stationary point of L with respect to x when $\lambda = \lambda_*$. Moreover, $\nabla_\lambda L(x, \lambda) = -c(x)$, so stationarity of L with respect to λ expresses the constraint $c(x) = 0$.

The development above was presented for a problem with three variables and one constraint, but it generalizes in a straightforward way to n variables and m constraints, with λ becoming an m-vector of Lagrange multipliers.

Let us see how Lagrange multipliers help us to solve the problem

$$\text{maximize } 8xyz \text{ subject to } \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

which defines the maximum rectangular block that fits inside the specified ellipsoid. Although our original problem (7) is a minimization problem, there is nothing in the development of (10) that is specific to minimization, and in fact the latter equation must be satisfied at any stationary point, so we can use it here. Setting $\tilde{x} = x/a$, $\tilde{y} = y/b$, $\tilde{z} = z/c$, the problem simplifies to

$$\text{maximize } 8abc\tilde{x}\tilde{y}\tilde{z} \text{ subject to } \tilde{x}^2 + \tilde{y}^2 + \tilde{z}^2 = 1.$$

The Lagrange multiplier condition is

$$8abc \begin{bmatrix} \tilde{y}\tilde{z} \\ \tilde{x}\tilde{z} \\ \tilde{x}\tilde{y} \end{bmatrix} = \lambda \begin{bmatrix} 2\tilde{x} \\ 2\tilde{y} \\ 2\tilde{z} \end{bmatrix}.$$

It is easily seen that these equations yield $\tilde{x} = \tilde{y} = \tilde{z} = 1/\sqrt{3}$ (and $\lambda_* = 4abc/\sqrt{3}$) and that the corresponding volume is $8abc/(3\sqrt{3})$. It is intuitively clear that this is a maximum, though in general checking for optimality requires further analysis involving inspection of second derivatives.

Lagrange multipliers and the Lagrangian function are widely used in applied mathematics in a variety of settings, including the CALCULUS OF VARIATIONS [IV.6] and LINEAR AND NONLINEAR OPTIMIZATION [IV.11]. One of the reasons for the importance of Lagrange multipliers is that they quantify the sensitivity of the optimal value to perturbations in the constraints. We can check this for our problem. If we perturb the constraint to $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1 + \varepsilon$, then it is easy to see that the solution is $V(\varepsilon) = 8abc((1+\varepsilon)/3)^{1/2}$, and hence $V'(0) = 4abc/\sqrt{3} = \lambda_*$.

11 Tricks and Techniques

As well as the general ideas and principles described in this article, applied mathematicians have at their disposal their own bags of tricks and techniques, which they bring into play when experience suggests they might be useful. Some will work only on very specific problems. Others might be nonrigorous but able to give useful insight. George Pólya is quoted as saying, "A trick used three times becomes a standard technique." Here are a few examples of tricks and techniques that prove useful on many different occasions, along with a very simple example in each case.

Use symmetry. When a problem has certain symmetries one can often argue that these must carry over into the solution. For example, the maximization problem at the end of the previous section is symmetric in $\tilde{x}$, $\tilde{y}$, and $\tilde{z}$, so one can argue that we must have $\tilde{x} = \tilde{y} = \tilde{z}$ at the solution.

Add and subtract a term, or multiply and divide by a term. As a very simple example, if A and B are $n \times n$ matrices with A nonsingular, then $AB = AB \cdot AA^{-1} = A(BA)A^{-1}$, which shows that AB and BA are similar and that they therefore have the same eigenvalues. A common scenario is that $\hat{x}$ is an approximation to x whose error cannot be directly estimated, but one can find another approximation $\tilde{x}$ whose relation to x and $\hat{x}$ is understood. One then writes $x - \hat{x} = (x - \tilde{x}) + (\tilde{x} - \hat{x})$ and thereby obtains, using the triangle inequality, the bound $\|x - \hat{x}\| \leqslant \|x - \tilde{x}\| + \|\tilde{x} - \hat{x}\|$. For example, x

Copyrighted material

problems. Of course, algorithm 1 does not cover all the possibilities. Another way to compute the sum is as $S_n = \log \prod_{i=1}^n e^{x_i}$. This formula has little to recommend it, but it is not so different from the expression $\exp(n^{-1} \log \sum_{i=1}^n x_i)$, which is a log-Euclidean mean of the x_i that has applications when the x_i are structured matrices or operators.

2 Bisection

The summation problem is unusual in that there is no difficulty in seeing the correctness of algorithm 1 or its computational cost. A slightly trickier algorithm is the bisection algorithm for finding a zero of a continuous function $f(x)$. The bisection algorithm takes as input an interval $[a, b]$ such that $f(a)f(b) < 0$; the intermediate-value theorem tells us that there must be a zero of f on this interval. The bisection algorithm repeatedly halves the interval and retains the half on which f has different signs at the endpoints, that is, the interval on which we can be sure there is a zero. To make the algorithm finite we need a stopping criterion. The following algorithm terminates once the interval is of length at most tol, a given tolerance.

Algorithm 2 (bisection algorithm). *This algorithm finds a zero of a continuous function $f(x)$ given an interval $[a, b]$ such that $f(a)f(b) < 0$ and an error tolerance* tol.

```
1  while b − a > tol
2        c = (a + b)/2
3        if f(c) = 0, quit, end
4        if f(c)f(b) < 0
5              a = c
6        else
7              b = c
8        end
9  end
10 x = (a + b)/2
```

To show the correctness of this algorithm note first that at the end of the while loop $f(a)f(b) < 0$ still holds; in other words, this inequality is an invariant of the loop. Therefore we have a sequence of intervals each of length half the previous interval and all containing a zero. This means that after k steps we have an interval of length $(b-a)/2^k$ containing a zero. The algorithm therefore terminates after $\lceil \log_2((|b-a|/\text{tol})) \rceil$ steps. Here, we are using the *ceiling function* $\lceil x \rceil$, which is the smallest integer greater than or equal to x. In the next section we will also need the *floor function* $\lfloor x \rfloor$, which is the largest integer less than or equal to x.

The algorithm returns as the approximate zero the midpoint of the final interval, which has length at most tol; since a zero lies in this interval, the absolute error is at most tol/2.

Algorithm 2 needs a number of refinements to make it more reliable and efficient for practical use. First, testing whether $f(c)$ and $f(b)$ have opposite signs should not be done by multiplying them, as the product could overflow or underflow in floating-point arithmetic. Instead, the signs should be directly compared. Second, $f(c)$ should not be computed twice, on lines 3 and 4, but rather computed once and its value reused. Finally, the convergence test is an absolute one, so is scale dependent. A better alternative is $|b - a| > \text{tol}(|a| + |b|)$, which is unaffected by scalings $a \leftarrow \theta a$, $b \leftarrow \theta b$.

Bisection is a widely applicable technique. For example, it can be used to search an ordered list to see if a given element is contained in the list; here it is known as *binary search*. It is also used for debugging. If the LaTeX source for this article fails to compile and I cannot spot the error, I will move the `\end{document}` command to the middle of the file and try again; I can thereby determine in which half of the file the error lies and can repeat the process to narrow the error down.

3 Divide and Conquer

The *divide and conquer* principle breaks a problem down into two (or more) equally sized subproblems and solves each subproblem recursively.

An example of how divide and conquer can be exploited is in the computation of a large integer power of a number. Computing x^n in the obvious way takes $n - 1$ multiplications. But x^{13}, for example, can be written x^8x^4x, which can be evaluated in just five multiplications instead of twelve by first forming x^2, $x^4 = (x^2)^2$, and $x^8 = (x^4)^2$. Notice that $13 = (1101)_2$ in base 2, and in general the base 2 representation of n tells us exactly how to break down the computation of x^n into products of terms x^{2^k}. However, by expressing the computation using divide and conquer we can avoid the need to compute the binary representation of n. The idea is to write $x^n = (x^{n/2})^2$ if n is even and $x^n = x(x^{\lfloor n/2 \rfloor})^2$ if n is odd. In either case the problem is reduced to one of half the size. The resulting algorithm is most elegantly expressed in recursive form, as an algorithm that calls itself.

Copyrighted material

Algorithm 3. *This algorithm computes x^n for a positive integer n.*

```
function y = power(x, n)
if n = 1, y = x, return
if n is odd
   y = x power(x, (n − 1)/2)^2    % Recursive call
else
   y = power(x, n/2)^2            % Recursive call
end
```

The number of multiplications required by algorithm 3 is bounded above by $2\lfloor \log_2 n \rfloor$.

Another example of how divide and conquer can be used is for computing the inverse of a nonsingular upper triangular matrix, $T \in \mathbb{C}^{n\times n}$. Write T in partitioned form as

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \tag{1}$$

where T_{11} has dimension $\lceil n/2 \rceil$. It is easy to check that

$$T^{-1} = \begin{bmatrix} T_{11}^{-1} & -T_{11}^{-1} T_{12} T_{22}^{-1} \\ 0 & T_{22}^{-1} \end{bmatrix}.$$

This formula reduces the problem to the computation of the inverses of two smaller matrices, namely, the diagonal blocks T_{11} and T_{22}, and their inverses can be expressed in the same way. The process can be repeated until scalars are reached and the inversion is trivial.

Algorithm 4. *This algorithm computes the inverse of a nonsingular upper triangular matrix T by divide and conquer.*

```
function U = inv(T)
n = dimension of T
if n = 1, u11 = t11^{-1}, return
Partition T according to (1), where T11 has
   dimension ⌈n/2⌉.
U11 = inv(T11)    % Recursive call
U22 = inv(T22)    % Recursive call
U12 = −U11 T12 U22.
```

Let us now work out the computational cost of this algorithm, in *flops*, where a flop is a multiplication, addition, subtraction, or division. Denote the cost of calling inv for an $n\times n$ matrix by c_n and assume for simplicity that $n = 2^k$. We then have $c_n = 2c_{n/2} + 2(n/2)^3$, where the second term is the cost of forming the triangular–full–triangular product $U_{11}T_{12}U_{22}$ of matrices of dimension $n/2$. Solving this recurrence gives $c_n = n^3/3 + O(n^2)$, which is the same as the cost of inverting a triangular matrix by standard techniques such as solving $TX = I$ by substitution.

As these examples show, recursion is a powerful way to express algorithms. But it is not always the right tool. To illustrate, consider the Fibonacci numbers, 1, 1, 2, 3, 5, ..., which satisfy the recurrence $f_n = f_{n-1} + f_{n-2}$ for $n \geqslant 2$, with $f_0 = f_1 = 1$. The obvious way to express the computation of the f_i is as a loop:

```
f0 = 1, f1 = 1
for i = 2:n
    fi = f_{i−1} + f_{i−2}
end
```

If just f_n is required then an alternative is the recursive function

```
function f = fib(n)
if n ⩽ 1
   f = 1
else
   f = fib(n − 1) + fib(n − 2)
end
```

The problem with this recursion is that it computes $\mathrm{fib}(n-1)$ and $\mathrm{fib}(n-2)$ independently instead of obtaining $\mathrm{fib}(n-1)$ from $\mathrm{fib}(n-2)$ with one addition as in the previous algorithm. In fact, the evaluation of $\mathrm{fib}(n)$ requires $f_n \approx 1.6^n$ operations, so the recursive algorithm is exponential in cost versus the linear cost of the first algorithm. It is possible to compute f_n with only logarithmic cost. The idea is to write

$$\begin{bmatrix} f_n \\ f_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} f_{n-1} \\ f_{n-2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^2 \begin{bmatrix} f_{n-2} \\ f_{n-3} \end{bmatrix}$$
$$= \cdots = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^{n-1} \begin{bmatrix} f_1 \\ f_0 \end{bmatrix}.$$

The matrix $\left[\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right]^{n-1}$ can be computed in $O(\log_2 n)$ operations using the analogue for matrices of algorithm 3.

A divide and conquer algorithm can break the problem into more than two subproblems. An example is the Karatsuba algorithm for multiplying two n-digit integers x and y. Suppose n is a power of 2 and write $x = x_1 10^{n/2} + x_2$, $y = y_1 10^{n/2} + y_2$, where x_1, x_2, y_1, and y_2 are $n/2$-digit integers. Then

$$xy = x_1 y_1 10^n + (x_1 y_2 + x_2 y_1) 10^{n/2} + x_2 y_2.$$

Copyrighted material

Computing xy has been reduced to computing three half-sized products because $x_1y_2 + x_2y_1 = (x_1 + x_2)(y_1 + y_2) - x_1y_1 - x_2y_2$. This procedure can be applied recursively. Denoting by C_n the number of arithmetic operations (on single-digit numbers) to form the product of two n-digit integers by this algorithm, we have $C_n = 3C_{n/2} + kn$ and $C_1 = 1$, where kn is the cost of the additions. Then

$$\begin{aligned} C_n &= 3(3C_{n/4} + kn/2) + kn \\ &= 3(3(3C_{n/8} + kn/4) + kn/2) + kn \\ &= kn(1 + 3/2 + (3/2)^2 + \cdots + (3/2)^{\log_2 n}) \\ &\approx 3kn^{\log_2 3} \approx 3kn^{1.58}, \end{aligned}$$

where the approximation is obtained by assuming that n is a power of 2. The cost is asymptotically less than the $O(n^2)$ cost of forming xy by the usual long multiplication method taught in school.

4 Computational Complexity

The computational cost of an algorithm is usually defined as the total number of arithmetic operations it requires, though it can also be defined as the execution time, under some assumption on the time required for each arithmetic operation. The cost is usually a function of the problem size, n say, and since the growth with n is of particular interest, the cost is usually approximated by the highest-order term, with lower-order terms ignored.

The algorithms considered so far all have the property that their computational cost is straightforward to evaluate and essentially independent of the data. For many algorithms the cost can vary greatly with the data. For example, an algorithm to sort a list of numbers might run more quickly when the list is nearly sorted. In this case it is desirable to find a bound that applies in all cases (a worst-case bound)—preferably one that is attainable for some set of data. It is also useful to have estimates of cost under certain assumptions on the distribution of the data. In *average-case analysis*, a probability distribution is assumed for the data and the expected cost is determined. *Smoothed analysis*, developed since 2000, interpolates between worst-case analysis and average-case analysis by measuring the expected performance of algorithms under small random perturbations of worst-case inputs. A number of algorithms are known for which the worst-case cost is exponential in the problem dimension n whereas the smoothed cost is polynomial in n, a prominent example being the SIMPLEX METHOD [IV.11 §3.1] for linear programming.

A good example of a problem for which different algorithms can have widely varying cost is the solution of a linear system $Ax = b$, where A is an $n \times n$ matrix. Cramer's rule states that $x_i = \det(A_i(b))/\det(A)$, where $A_i(b)$ denotes A with its ith column replaced by b. If the determinant is evaluated from the usual textbook formula involving EXPANSION BY MINORS [I.2 §18], the cost of computing x is about $(n + 1)!$ operations, making this method impractical unless n is very small. By contrast, Gaussian elimination solves the system in $2n^3/3 + O(n^2)$ operations, with mere polynomial growth of the operation count with n. However, Gaussian elimination is by no means of optimal complexity, as we now explain.

The complexity of matrix inversion can be shown to be the same as that of matrix multiplication, so it suffices to consider the matrix multiplication problem $C = AB$ for $n \times n$ matrices A and B. The usual formula for matrix multiplication yields C in $2n^3$ operations. In a 1969 paper Volker Strassen showed that when $n = 2$ the product can be computed from the formulas

$$\begin{aligned} p_1 &= (a_{11} + a_{22})(b_{11} + b_{22}), \\ p_2 &= (a_{21} + a_{22})b_{11}, \qquad p_3 = a_{11}(b_{12} - b_{22}), \\ p_4 &= a_{22}(b_{21} - b_{11}), \qquad p_5 = (a_{11} + a_{12})b_{22}, \\ p_6 &= (a_{21} - a_{11})(b_{11} + b_{12}), \\ p_7 &= (a_{12} - a_{22})(b_{21} + b_{22}), \end{aligned}$$

$$C = \begin{bmatrix} p_1 + p_4 - p_5 + p_7 & p_3 + p_5 \\ p_2 + p_4 & p_1 + p_3 - p_2 + p_6 \end{bmatrix}.$$

The evaluation requires seven multiplications and eighteen additions instead of eight multiplications and eight additions for the usual formulas. At first sight, this does not appear to be an improvement. However, these formulas do not rely on commutativity so are valid when the a_{ij} and b_{ij} are matrices, in which case for large dimensions the saving of one multiplication greatly outweighs the extra ten additions. Assuming n is a power of 2, we can partition A and B into four blocks of size $n/2$, apply Strassen's formulas for the multiplication, and then apply the same formulas recursively on the half-sized matrix products. The resulting algorithm requires $O(n^{\log_2 7}) = O(n^{2.81})$ operations. Strassen's work sparked interest in finding matrix multiplication algorithms of even lower complexity. Since there are $O(n^2)$ elements of data, which must each participate in at least one operation, the

Copyrighted material

Table 1 The cost of solving an $n \times n$ linear system obtained by discretizing the two-dimensional Poisson equation.

Year	Method	Cost	Type
1948	Banded Cholesky	n^2	Direct
1948	Jacobi, Gauss–Seidel	n^2	Iterative
1950	SOR (optimal parameter)	$n^{3/2}$	Iterative
1952	Conjugate gradients	$n^{3/2}$	Iterative
1965	Fast Fourier transform	$n \log n$	Direct
1965	Block cyclic reduction	$n \log n$	Direct
1977	Multigrid	n	Iterative

exponent of n must be at least 2. The current world record upper bound on the exponent is 2.3728639, proved by François Le Gall in 2014. However, all existing algorithms with exponent less than that of Strassen's algorithm are extremely complicated and not of practical interest.

An area that has undergone many important algorithmic developments over the years is the solution of linear systems arising from the discretization of partial differential equations (PDEs). Consider the POISSON EQUATION [III.18] on a square with the unknown function specified on the boundary. When discretized on an $N \times N$ grid by centered differences, a system of $n = N^2$ equations in n unknowns is obtained with a banded, symmetric positive-definite coefficient matrix containing $O(n)$ nonzeros. Table 1 gives the dominant term in the operation count (ignoring the multiplicative constant) for different methods, some of which are described in NUMERICAL LINEAR ALGEBRA AND MATRIX ANALYSIS [IV.10]. For the iterative algorithms it is assumed that the iteration is terminated when the error is of order 10^{-6}. The year is the year of first publication, or, for the first two methods, the year that the first stored-program computer was operational. Since there are n elements in the solution vector and at least one operation is required to compute each element, a lower bound on the cost is $O(n)$, and this is achieved by the multigrid method. The algorithmic speedups shown in the table are of a similar magnitude to the speedups in computer hardware over the same period.

4.1 Complexity Classes

The algorithms we have described so far all have a cost that is bounded by a polynomial in the problem dimension, n. For some problems the existence of algorithms with polynomial complexity is unclear. In analyzing this question mathematicians and computer scientists use a classification of problems that makes a distinction finer than whether there is or is not an algorithm of polynomial run time. This classification is phrased in terms of decision problems: ones that have a yes or no answer. The problem class *P* comprises those problems that can be solved in polynomial time in the problem dimension. The class *NP* comprises those problems for which a yes answer can be verified in polynomial time. An example of a problem in NP is a jigsaw puzzle: it is easy to check that a claimed solution is a correctly assembled puzzle, but solving the puzzle in the first place appears to be much harder.

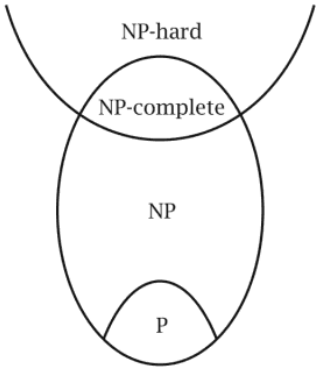


Figure 2 Complexity classes. It is not known whether the classes P and NP are equal.

A problem is *NP-complete* if it is in NP and it is possible to reduce any other NP problem to it in polynomial time. Hence if a polynomial-time algorithm exists for an NP-complete problem then all NP problems can be solved in polynomial time. Many NP-complete problems are known, including Boolean satisfiability, graph coloring, choosing optimal page breaks in a document, and the Battleship game or puzzle.

A problem (not necessarily a decision problem) is *NP-hard* if it is at least as hard as any NP problem, in the sense that there is an NP-complete problem that is reducible to it in polynomial time. Thus the NP-hard problems are even harder than the NP-complete problems. Examples of NP-hard problems are the TRAVELING SALESMAN PROBLEM [VI.18], SPARSE APPROXIMATION [VII.10], and nonconvex QUADRATIC PROGRAMMING [IV.11 §1.3]. Figure 2 shows the relation among the classes.

An excellent example of the subtleties of computational complexity is provided by the determinant and the permanent of a matrix. The permanent of an $n \times n$

Copyrighted material

matrix A is

$$\operatorname{perm}(A) = \sum_{\sigma} \prod_{i=1}^{n} a_{i,\sigma_i},$$

where the vector σ ranges over all permutations of the set of integers $\{1, 2, \ldots, n\}$. The determinant has a similar expression differing only in that the product term is multiplied by the sign (± 1) of the permutation. Yet while the determinant can be computed in $O(n^3)$ operations, by Gaussian elimination, no polynomial-time algorithm has ever been discovered for computing the permanent. Leslie Valiant gave insight into this disparity when he showed in 1979 that the problem of computing the permanent is complete for a complexity class of counting problems called #P that extends NP.

The most famous open problem in computer science is "is P equal to NP?" It was posed by Stephen Cook in 1971 and is one of the seven Clay Institute Millennium Problems, for each of which a $1 million prize is available for a solution. Informally, the question is whether the "easy to solve" problems are equal to the "easy to check" problems. It is known that P $\subseteq$ NP, so the question is whether or not the inclusion is strict.

5 Trade-off between Speed and Accuracy

In designing algorithms that run in floating-point arithmetic it frequently happens that an increase in speed is accompanied by a decrease in accuracy. A classic example is the computation of the *sample variance* of n numbers $x_1, \ldots, x_n$, which is defined as

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{2}$$

where the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Computing s_n^2 from this formula requires two passes through the data, one to compute $\bar{x}$ and the other to accumulate the sum of squares. A two-pass computation is undesirable for large data sets or when the sample variance is to be computed as the data is generated. An alternative formula, found in statistics textbooks (and implemented on many pocket calculators and spreadsheets over the years), uses about the same number of operations but requires only one pass through the data:

$$s_n^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right). \tag{3}$$

However, this formula behaves badly in floating-point arithmetic. For example, if $n = 3$ and $x_1 = 10\,000$, $x_2 = 10\,001$, and $x_3 = 10\,002$, then, in IEEE single-precision arithmetic (with unit roundoff $u \approx 6 \times 10^{-8}$), the sample variance is computed as 1.0 by the two-pass formula (relative error 0) but 0.0 by the one-pass formula (relative error 1). The reason for the poor accuracy of the one-pass formula is that there is massive SUBTRACTIVE CANCELATION [II.13] in (3). The original formula (2) always yields a computed result with error $O(nu)$. Is there a way of combining the speed of the one-pass formula with the accuracy of the two-pass one? Yes: the recurrence

$$\left.\begin{aligned} M_1 &= x_1, \qquad Q_1 = 0, \\ M_k &= M_{k-1} + \frac{x_k - M_{k-1}}{k} \\ Q_k &= Q_{k-1} + \frac{(k-1)(x_k - M_{k-1})^2}{k} \end{aligned}\right\} \quad k = 2\colon n$$

calculates Q_n, which yields $s_n^2 = Q_n/(n-1)$ and produces an accurate result in floating-point arithmetic.

6 Choice of Algorithm

Much research in numerical analysis and scientific computing is about finding the best algorithm for solving a given problem, and for classic problems such as solving a PDE or finding the eigenvalues of a matrix there are many possibilities, with improvements continually being developed. However, even for some quite elementary problems there are several possible algorithms, some of which are far from obvious.

A first example is the evaluation of a polynomial $p(x) = a_0 + a_1 x + \cdots + a_n x^n$. The most obvious way to evaluate the polynomial is by directly forming the powers of x.

```
1  p = a_0 + a_1 x, w = x
2  for i = 2:n
3      w = wx
4      p = p + a_i w
5  end
```

This algorithm requires $2n$ multiplications and n additions (ignoring the constant term in the operation count).

An alternative method is *Horner's method* (nested multiplication). It is derived by writing the polynomial in nested form:

$$p(x) = (\cdots((a_n x + a_{n-1})x + a_{n-2})x + \cdots + a_1)x + a_0.$$

Copyrighted material

Table 2 Some algorithms mentioned in this book.

Algorithm	Reference	Key early figures
Gaussian elimination	IV.10 §2	Ancient Chinese (ca. 1 C.E.), Gauss (1809); formulated as LU factorization by various authors from 1940s
Newton's method	II.28	Newton (1669), Raphson (1690)
Fast Fourier transform	II.10	Gauss (1805), Cooley and Tukey (1965)
Cholesky factorization	IV.10 §2	Cholesky (1910)
Remez algorithm	IV.9 §3.5, VI.11 §2	Remez (1934)
Simplex method (linear programming)	IV.11 §3.1	Dantzig (1947)
Conjugate gradient and Lanczos methods	IV.10 §9	Hestenes and Stiefel (1952), Lanczos (1952)
Ford–Fulkerson algorithm	IV.37 §7	Ford and Fulkerson (1956)
k-means algorithm	IV.17 §5.3	Lloyd (1957), Steinhaus (1957)
QR factorization	IV.10 §2	Givens (1958), Householder (1958)
Dijkstra's algorithm	VI.10	Dijkstra (1959)
Quasi-Newton methods	IV.11 §4.2	Davidon (1959), Broyden, Fletcher, Goldfarb, Powell, Shanno (early 1960s)
QR algorithm	IV.10 §5.5	Francis (1961), Kublanovskaya (1962)
QZ algorithm	IV.10 §5.8	Moler and Stewart (1973)
Singular value decomposition	II.32	Golub and Kahan (1965), Golub and Reinsch (1970)
Strassen's method	I.4 §4	Strassen (1968)
Multigrid	IV.10 §9, IV.13 §3, IV.16	Fedorenko (1964), Brandt (1973), Hackbusch (1977)
Interior point methods	IV.11 §3.2	Karmarkar (1984)
Generalized minimal residual method	IV.10 §9	Saad and Schulz (1986)
Fast multipole method	VI.17	Greengard and Rokhlin (1987)
JPEG	VII.7 §5, VII.8	Members of the Joint Photographic Experts Group (1992)
PageRank	VI.9	Brin and Page (1998)
HITS	I.1	Kleinberg (1999)

existing mathematical ideas to practical problems: new results are continually being developed, usually building on old ones. Applied mathematicians are always innovating, and the constant arrival of new or modified problems provides direction and motivation for their research.

In this article we describe some goals of research in applied mathematics from the perspectives of the ancient problem of solving equations, the more contemporary theme of exploiting structure, and the practically important tasks of modeling and prediction. We also discuss the strategy behind research.

1 Solving Equations

A large proportion of applied mathematics research papers are about analyzing or solving equations. The equations may be algebraic, such as linear or nonlinear equations in one or more variables. They may be ordinary differential equations (ODEs), partial differential equations (PDEs), integral equations, or differential-algebraic equations.

The wide variety of equations reflects the many different ways in which one can attempt to capture the behavior of the system being modeled. Whatever the equation, an applied mathematician is interested in answering a number of questions.

1.1 Does the Equation Have a Solution?

We are interested in whether there is a unique solution and, if there is more than one solution, how many there are and how they are characterized. Existence of solutions may not be obvious, and one occasionally

Copyrighted material

hears tales of mathematicians who have solved equations for which a proof is later given that no solution exists. Such a circumstance may sound puzzling: is it not easy to check that a putative solution actually is a solution? Unfortunately, checking satisfaction of the equation may not be easy, especially if one is working in a function space. Moreover, the problem specification may require the solution to have certain properties, such as existence of a certain number of derivatives, and the claimed solution might satisfy the equation but fail to have some of the required properties. Instead of analyzing the problem in the precise form in which it is given, it may be better to investigate what additional properties must be imposed for an equation to have a unique solution.

1.2 Is the Equation Well-Posed?

A problem is *well-posed* if it has a unique solution and the solution changes continuously with the data that define the problem. A problem that is not well-posed is *ill-posed*. For an ill-posed problem an arbitrarily small perturbation of the data can produce an arbitrarily large change in the solution, which is clearly an unsatisfactory situation.

An example of a well-posed problem is to determine the weight supported by each leg of a three-legged table. Assuming that the table and its legs are perfectly symmetric and the ground is flat, the answer is that each leg carries one-third of the total weight. For a table with four legs each leg supports one-quarter of the total weight, but if one leg is shortened by a tiny amount then it leaves the ground and the other three legs support the weight of the table (a phenomenon many of us have experienced in restaurants). For four-legged tables the problem is therefore ill-posed.

For finite-dimensional problems, uniqueness of the solution implies well-posedness. For example, a linear system $Ax = b$ of n equations in n unknowns with a nonsingular coefficient matrix A is well-posed. Even so, if A is nearly singular then a small perturbation of A can produce a large change in the solution, albeit not arbitrarily large: the CONDITION NUMBER [I.2 §22] $\kappa(A) = \|A\| \, \|A^{-1}\|$ bounds the relative change. But for infinite-dimensional problems the existence of a unique solution does not imply that the problem is well-posed; examples are given in the article on INTEGRAL EQUATIONS [IV.4 §6].

The notion of well-posedness was introduced by Jacques Hadamard at the beginning of the twentieth century. He believed that physically meaningful problems should be well-posed. Today it is recognized that many problems are ill-posed, and they are routinely solved by reformulating them so that they are well-posed, typically by a process called REGULARIZATION [IV.15 §2.6] (see also INTEGRAL EQUATIONS [IV.4 §7]).

An important source of ill-posed problems is INVERSE PROBLEMS [IV.15]. Consider a mathematical model in which the inputs are physical variables that can be adjusted and the output variables are the result of an experiment. The *forward problem* is to predict the outputs from a given set of inputs. The *inverse problem* is to make deductions about the inputs that could have produced a given set of outputs. In practice, the measurements of the outputs may be subject to noise and the model may be imperfect, so UNCERTAINTY QUANTIFICATION [II.34] needs to be carried out in order to estimate the uncertainty in the predictions and deductions.

1.3 What Qualitative Properties Does a Solution Have?

It may be of more interest to know the behavior of a solution than to know the solution itself. One may be interested in whether the solution, $f(t)$ say, decays as $t \to \infty$, whether it is monotonic in t, or whether it oscillates and, if so, with what fixed or time-varying frequency. If the problem depends on parameters, it may be possible to answer these questions for a range of values of the parameters.

1.4 Does an Iteration Converge?

As we saw in METHODS OF SOLUTION [I.3], solutions are often computed from iterative processes, and we therefore need to understand these processes. Various facets of convergence may be of interest.

- Is the iteration always defined, or can it break down (e.g., because of division by zero)?
- For what starting values, and for what class of problems, does the iteration converge?
- To what does the iteration converge, and how does this depend on the starting value (if it does at all)?
- How fast does the iteration converge?
- How are errors (in the initial data, or rounding errors introduced during the iteration) propagated? In particular, are they bounded?

Copyrighted material

To illustrate some of these points we consider the iteration

$$x_{k+1} = \frac{1}{p}[(p-1)x_k + x_k^{1-p}a], \tag{1}$$

with p a positive integer and $a \in \mathbb{C}$, which is Newton's method for computing a pth root of a. We ask for which a and which starting values x_0 the iteration converges and to what root it converges. The analysis is simplified by defining $y_k = \theta^{-1}x_k$, where θ is a pth root of a, as the iteration can then be rewritten

$$y_{k+1} = \frac{1}{p}[(p-1)y_k + y_k^{1-p}], \qquad y_0 = \theta^{-1}x_0, \tag{2}$$

which is Newton's method for computing a pth root of unity. The original parameters a and x_0 have been combined into the starting value y_0.

Figure 1 illustrates the convergence of the iteration for $p = 2, 3, 5$. For y_0 ranging over a 400×400 grid with $\operatorname{Re} y_0, \operatorname{Im} y_0 \in [-2.5, 2.5]$, it plots the root to which y_k from (2) converges, with each root denoted by a different grayscale from white (the principal root, 1) to black. Convergence is declared if after fifty iterations the iterate is within relative distance 10^{-13} of a root; the relatively small number of points for which convergence was not observed are plotted white. For $p = 2$ the figure suggests that the iteration converges to 1 if started in the open right half-plane and -1 if started in the open left half-plane, and this can be proved to be true. But for $p = 3, 5$ the regions of convergence have a much more complicated structure, involving sectors with petal-like boundaries.

The complexity of the convergence for $p \geqslant 3$ was first noticed by Arthur Cayley in 1879, and an analysis of convergence for all starting values requires the theory of Julia sets of rational maps. However, for practical purposes it is usually principal roots that need to be computed, so from a practical viewpoint the main implication to be drawn from the figure is that for $p = 3, 5$ Newton's method converges to 1 for y_0 sufficiently close to the positive real axis—and it can be proved that this is true.

We see from this example that the convergence analysis depends very much on the precise question that is being asked. The iteration (1) generalizes in a natural way to matrices and operators, for which the convergence results for the scalar case can be exploited.

2 Preserving Structure

Many mathematical problems have some kind of structure. An example with explicit structure is a linear system $Ax = b$ in which the $n \times n$ matrix A is a TOEPLITZ MATRIX [I.2 §18]. This system has $n^2 + n$ numbers in A and b but only $3n - 1$ independent parameters. On the other hand, if for the vector ODE $y' = f(t, y)$ there is a vector v such that $v^{\mathrm{T}} f(t, y) = 0$ for all t and y, then $(\mathrm{d}/\mathrm{d}t)v^{\mathrm{T}}y(t) = v^{\mathrm{T}}f(t, y) = 0$, so $v^{\mathrm{T}}y(t)$ is constant for all t. This conservation or invariance property is a form of structure, though one more implicit than for the Toeplitz system.

An example of a nonlinear conservation property is provided by the system of ODEs

$$\begin{aligned} u'(t) &= v(t), \\ v'(t) &= -u(t). \end{aligned}$$

For this system,

$$\frac{\mathrm{d}}{\mathrm{d}t}(u^2 + v^2) = 2(u'u + v'v) = 2(vu - uv) = 0,$$

so there is a quadratic invariant. In particular, for the initial values $u(0) = 1$ and $v(0) = 0$ the solution is $u(t) = \cos t$ and $v(t) = -\sin t$, which lies on the unit circle centered at the origin in the uv-plane. If we solve the system using a numerical method, we would like the numerical solution also to lie on the circle. In fact, one potential use of this differential equation is to provide a method for plotting circles that avoids the relatively expensive evaluation of sines and cosines. Consider the following four standard numerical methods applied to our ODE system. Here, $u_k \approx u(kh)$ and $v_k \approx v(kh)$, where h is a given step size, and $u_0 = 1$ and $v_0 = 0$:

$$\begin{array}{ll} \text{Forward Euler} & \begin{cases} u_{k+1} = u_k + hv_k, \\ v_{k+1} = v_k - hu_k, \end{cases} \\ \text{Backward Euler} & \begin{cases} u_{k+1} = u_k + hv_{k+1}, \\ v_{k+1} = v_k - hu_{k+1}, \end{cases} \\ \text{Trapezium method} & \begin{cases} u_{k+1} = u_k + h(v_k + v_{k+1})/2, \\ v_{k+1} = v_k - h(u_k + u_{k+1})/2, \end{cases} \\ \text{Leapfrog method} & \begin{cases} u_{k+1} = u_k + hv_k, \\ v_{k+1} = v_k - hu_{k+1}. \end{cases} \end{array}$$

Figure 2 plots the numerical solutions computed with 32 steps of length $h = 2\pi/32$. We see that the forward Euler solution spirals outward while the backward Euler solution spirals inward. The trapezium method solution stays nicely on the unit circle. The leapfrog method solution traces an ellipse. This behavior is easy to explain if we write each method in the form

$$z_{k+1} = Gz_k, \qquad z_k = \begin{bmatrix} u_k \\ v_k \end{bmatrix},$$

where $G = \left[\begin{smallmatrix} 1 & h \\ -h & 1 \end{smallmatrix}\right]$ for the Euler method, for example. Then the behavior of the sequence z_k depends on the

Copyrighted material

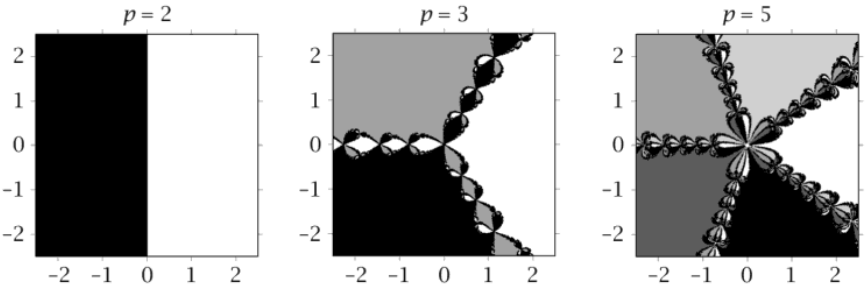


Figure 1 Newton iteration for a pth root of unity. Each point y_0 in the region is shaded according to the root to which the iteration converges, with white denoting the principal root, 1.

eigenvalues of the matrix G. It turns out that the spectral radius of G is greater than 1 for forward Euler and less than 1 for backward Euler, which explains the spiraling. For the trapezium rule G is orthogonal, so $\|z_{k+1}\|_2 = \|z_k\|_2$ and the trapezium solutions stay exactly on the unit circle. For the leapfrog method the determinant of G is 1, which means that areas are preserved, but G is not orthogonal so the leapfrog solution drifts slightly off the circle.

The subject of GEOMETRIC INTEGRATION [IV.12 §5] is concerned more generally with methods for integrating nonlinear initial-value ODEs and PDEs in a way that preserves the invariants of the system, while also providing good accuracy in the usual sense. This includes, in particular, SYMPLECTIC INTEGRATORS [IV.12 §1.3] for Hamiltonian systems.

3 Modeling and Prediction

AS WHAT IS APPLIED MATHEMATICS? [I.1 §1] explains, modeling is the first step in solving a physical problem. Models are necessarily simplifications because it is impractical to incorporate every detail. But simple models can still be useful as tools to explore the broad consequences of physical laws. Moreover, the more complex a model is the more parameters it has (all of which need estimating) and the harder it is to analyze.

In their 1987 book *Empirical Model-Building and Response Surfaces*, Box and Draper ask us to

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

Road maps illustrate this statement. They are always a simplified representation of reality due to representing a three-dimensional world in two dimensions and displaying wiggly roads as straight lines. But road maps are very useful. Moreover, there is no single "correct" map but rather many possibilities depending on resolution and purpose. Another example is the approximation of π. The approximation $\pi \approx 3.14$ is a model for π that is wrong in that it is not exact, but it is good enough for many purposes.

It is difficult to give examples of the modeling process because knowledge of the problem domain is usually required and derivations can be lengthy. We will use for illustration a very simple model of population growth, based on the *logistic equation*

$$\frac{\mathrm{d}N}{\mathrm{d}t} = rN\left(1 - \frac{N}{K}\right).$$

Here, $N(t)$ is a representation in a continuous variable of the number of individuals in a population at time t, $r > 0$ is the growth rate of the population, and $K > 0$ is the carrying capacity. For $K = \infty$, the model says that the rate of change of the population, $\mathrm{d}N/\mathrm{d}t$, is rN; that is, it is proportional to the size of the population through the constant r, so the population grows exponentially. For finite K, the model attenuates this rate of growth by a subtractive term rN^2/K, which can be interpreted as representing the increasing effects of competition for food as the population grows. The logistic equation can be solved exactly for $N(t)$ (see ORDINARY DIFFERENTIAL EQUATIONS [IV.2 §2]). Laboratory experiments have shown that the model can predict reasonably well the growth of protozoa feeding on bacteria. However, for some organisms the basic logistic equation is not a good model because it assumes instant responses to changes in population size and so does not account for gestation periods, the time taken for young to reach maturity, and other delays. A more realistic model may therefore be

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = rN(t)\left(1 - \frac{N(t-\tau)}{K}\right),$$

Copyrighted material

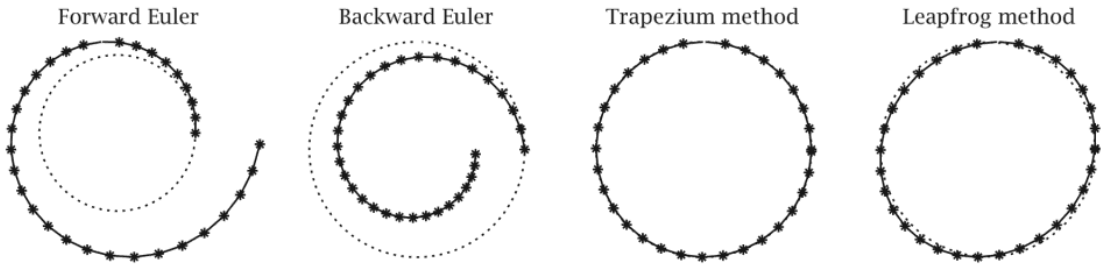


Figure 2 Approximations to the unit circle computed by four different numerical integrators with step size $h = 2\pi/32$. The dotted line is the unit circle; asterisks denote numerical approximations.

where $\tau > 0$ is a delay parameter. At time t, part of the quadratic term is now evaluated at an earlier time, $t - \tau$. This delay differential equation has oscillatory solutions and has been found to model well the population of lemmings in the Arctic. Note that in contrast to the PREDATOR–PREY MODEL [I.2 §10], the delayed logistic model can produce oscillations in a population without the need for a second species acting as predator. There is no suggestion that either of these logistic models is perfect, but with appropriate fitting of parameters they can provide useful approximations to actual populations and can be used to predict future behavior.

3.1 Errors

A lot of research is devoted to understanding the errors that arise at the different stages of the modeling process. These can broadly be categorized as follows.

Errors in the mathematical model. Setting up the model introduces errors, since the model is never exact. These are the hardest errors to estimate.

Approximation errors. These are the errors incurred when infinite-dimensional equations are replaced by a finite-dimensional system (that is, a continuous problem is replaced by a discrete one: the process of discretization), or when simpler approximations to the equations are developed (e.g., by MODEL REDUCTION [II.26]). These errors include errors in replacing one approximating space by another (e.g., replacing continuous functions by polynomials), errors in FINITE-DIFFERENCE [II.11] approximations, and errors in truncating power series and other expansions.

Rounding errors. Once the problem has been put in a form that can be solved by an algorithm implemented in a computer program, the effects of the rounding errors introduced by working in finite-precision arithmetic need to be determined.

Analysis of errors may include looking at the effects of uncertainties in the model data, including in any parameters in the model that must be estimated. This might be tackled in a statistical sense using techniques from UNCERTAINTY QUANTIFICATION [II.34]—indeed, if the model has incompletely known data then probabilistic techniques may already be in use to estimate the missing data. Sensitivity of the solution of the model may also be analyzed by obtaining worst-case error bounds with the aid of CONDITION NUMBERS [I.2 §22].

3.2 Multiphysics and Multiscale Modeling

Scientists are increasingly tackling problems with one or both of the following characteristics: (a) the system has multiple components, each governed by its own physical principles; and (b) the relevant processes develop over widely different time and space scales. These are called *multiphysics* and *multiscale* problems, respectively. An example of both is the problem of modeling how *space weather* affects the Earth, and in particular modeling the interaction of the solar wind (the flow of charged particles emitted by the sun) with the Earth's magnetic field. Different physical models describe the statistical distribution of the plasma, which consists of charged particles, and the evolution of the electric and magnetic fields, and these form a coupled nonlinear system of PDEs. The length scales range from millions of kilometers (the Earth–sun distance) to hundreds of meters, and the timescales range from hours down to 10^{-5} seconds. Problems such as this pose challenges both for modeling and for computational solution of the models. The computations require HIGH-PERFORMANCE COMPUTERS [VII.12], and a particular task is to present the vast quantities of data generated in such a way that users, such as forecasters of space

Copyrighted material

Not everyone will agree with Bonnor. Some take a more methodological approach and almost equate applied mathematics with "mathematical modeling." Others are of a more concrete, mathematical mind and insist that there are parts of mathematics that are per se more or less applicable than others. We find Bonnor's definition appealing because it stresses the social dimension of the mathematical working process and allows a historical understanding of the notion of applied mathematics.

The importance of "attitudes" notwithstanding, by any definition applied mathematics has to be "genuine" mathematics in the sense that it aims at and/or uses general statements (theorems) even if the piece of mathematics in question has not yet been fully logically established. In fact, the applicability of mathematics is mainly based on its "generality," which in relation to fields of application often appears as "abstractness." This applies even to relatively elementary applications such as the use of positional number systems.

Applications of mathematics, even on a nonelementary level, have been possible because certain practices and properties, such as algorithms for approximations or geometrical constructions, have always existed within mathematics itself and have led to spontaneous or deliberate applications. While, as the universal mathematician John von Neumann observed in 1947, in pure mathematics many problems and methods are selected for aesthetic reasons, in applied mathematics, problems considered at the time as urgent have priority, and the choice of methods often has to be subordinated to the goals in question. However, attitudes and values, which often had and continue to have strong political and economic overtones, have always been instrumental in deciding exactly which parts of mathematics should be emphasized and developed. Since attitudes have to be promoted through education, this puts a great responsibility on teaching and training and makes developments in that area an important topic for a history of applied mathematics.

Of course, many modern and recent applications rely on older mathematical ideas in differential equations, topology, and discrete mathematics and on established notation and symbolism (matrices, quaternions, Laplace transforms, etc.), while important new developments in INTEGRAL EQUATIONS [IV.4], measure theory, vector and tensor analysis, etc., at the turn of the twentieth century have added to these ideas.

However, in the twentieth century, three major scientific and technical innovations both changed and enlarged the notion of applied mathematics. In rough chronological order, these are mathematical *modeling* in a broad, modern sense, *stochastics* (modern probability and statistics), and the *digital computer*. These three innovations have, through their interactions, restructured applied mathematics. They were principally established after World War II, and it was also only then that the term "mathematical modeling" came to be more frequently used for activities that had hitherto usually been expressed by less concise words such as "problem formulation and evaluation." In addition to these innovations, which are essentially concerned with methodology, several totally new fields of application, such as electrical engineering, economics, biology, meteorology, etc., emerged in the twentieth century.

While in 1914 one of the pioneers of modern applied mathematics, Carl Runge, still doubted whether "the name of 'applied mathematics' was chosen appropriately, because when applied to empirical sciences it still remains pure mathematics," the three major innovations listed above would radically alter and extend the notion of mathematics and, in particular, that of applied mathematics. Due to these innovations, the modern disciplines at the interface of mathematics and engineering, such as cybernetics, control theory, computer science, and optimization, were all able to emerge in the 1940s and 1950s in the United States (Wiener, Shannon, Dantzig) and the Soviet Union (Andronov, Kolmogorov, Pontryagin, Kantorovich) independently, and to a somewhat lesser degree in England (Turing, Southwell, Wilkinson), France (Couffignal), and elsewhere. These innovations also gradually changed "hybrid disciplines," such as electrical engineering and aerodynamics, that had originated at the turn of the century. In the case of aerodynamics, not only were statistical explanations of TURBULENCE [V.21] increasingly proposed after World War II, but also CONFORMAL MAPPINGS [II.5] gradually lost importance in favor of computational FLUID DYNAMICS [IV.28]. Within operations research, with its various approaches and techniques (linear programming, optimization methods, statistical quality control, inventory control, queuing analysis, network flow analysis), mathematical concepts, especially mathematical models, acquired an even stronger foothold than in the more traditional industrial engineering.

One typical modern mathematical discipline that intimately combines pure and applied aspects of the subject and that is intertwined with various other scientific (physical and biological) and engineering disciplines

Copyrighted material

is the theory of DYNAMICAL SYSTEMS [IV.20]. After initial work in the field by Poincaré, Lyapunov, and Birkhoff, the theory fell into oblivion until the 1960s. This falling away can be explained by fashion (such as the trend toward the mathematics of Bourbaki), by new demands in applications connected to dissipative systems, and by the partial invisibility of the Russian school in the West. With the advent of modern computing devices, the shape of the discipline changed dramatically. Mathematicians were empowered computationally and graphically, the visualization of new objects such as fractals was made possible, and applications in fields such as CONTROL THEORY [IV.34] and meteorology—quantitative applications as well as qualitative ones—began to proliferate. The philosophical discussion about mathematics and applications has also been enriched by this discipline, with the public being confronted by catchwords such as CHAOS [II.3], catastrophe, and self-organization. However, the process whereby the various streams of problems converged and led to the subject's modern incarnation is complex:

> In the 1930s, for example, what could the socio-professional worlds of the mathematician Birkhoff (professor at Harvard), the "grand old man of radio" van der Pol (at the Philips Research Lab), and the Soviet "physico [engineer] mathematician" Andronov at Gorki have had in common? What, in the 1950s, had Kolmogorov's school in common with Lefschetz's? It is precisely this manifold character of social and epistemic landscapes that poses problem[s] in this history.
>
> Aubin and Dahan (2002)

The role played by the three major innovations continues largely unabated today, as is evident, for instance, from a 2012 report from the Society for Industrial and Applied Mathematics (SIAM) on industrial mathematics:

> Roughly half of all mathematical scientists hired into business and industry are statisticians. The second-largest group by academic specialty is applied mathematics. Compared to the 1996 survey, fewer graduates reported "modeling and simulation" as an important academic specialty for their jobs, and more reported "statistics." Programming and computer skills continue to be the most important technical skill that new hires bring to their jobs.

By separating statistics from applied mathematics, the SIAM report follows a certain tradition, caused in part by institutional boundaries, such as the existence of separate statistics departments in universities. This distinction is also partly followed in the present volume and in this article, although there is no doubt about the crucial role of probability and statistics in applications. For example, one need only consider the Monte Carlo method—notably developed at Los Alamos in the 1940s by Stanislaw Ulam and von Neumann, and continued by Nicholas Metropolis—which is now used in a wide variety of different contexts including numerical integration, optimization, and inverse problems. In addition, the combination of stochastics and modeling in biological and physical applications has had a philosophical dimension, contributing to the abandonment of rigid causality in science, e.g., through Karl Pearson's correlation coefficient and Werner Heisenberg's uncertainty principle. However, by the end of the 1960s the limitations of stochastics in helping us to understand the nature of disorder had become apparent, particularly in connection with the study of complex ("chaotic") dynamical systems. Nevertheless, stochastics continues to play an important role in the development of big theories with relevance for applications, including statistical models for weather forecasting.

1.1 Further Themes and Some Limitations

Putting stochastics on the sidelines is but one of several limitations of this article—limitations that are the result of a lack of space, a lack of distance, and more general methodological considerations. A further thematic restriction concerns industrial mathematics, which figures separately from applied mathematics in the very name of SIAM, although there are obvious connections between the two, in particular with respect to training and in developing attitudes toward applications. Knowing that industrial mathematics has changed, and above all expanded, from its origins in the early twentieth century to move beyond its purely industrial context, these connections become even clearer. Industrial mathematics is, today, an established subdiscipline, loosely described as the modeling of problems of direct and immediate interest to industry, performed partly in industrial surroundings and partly in academic ones.

The history of mathematical instruments, including both numerical and geometrical devices, and their underlying mathematical principles is another topic we have had to leave out almost completely. Some discussion of the history of mathematical table projects

Copyrighted material

is the farthest we reach in this respect. This limitation also applies to the technological basis of modern computing and the development of software technology (covered by Cortada's excellent bibliography), which has provided, and continues to provide, an important stimulus for the development (and funding) of applied mathematics. In 2000, in the *Journal of Computational and Applied Mathematics*, it was estimated that of the increase in computational power, half should be attributed to improved algorithms and half should be attributed to the increase in computational hardware speeds. Computing technology has continued to advance rapidly, and companies are making more and more aggressive use of HIGH-PERFORMANCE COMPUTING [VII.12]

A detailed discussion of the fields of application of mathematics themselves—be it in (pure) mathematics, the sciences, engineering, economics and finance, industry, or the military—is absent from this article for a number of reasons, both practical and methodological, above all the huge variation of specific conditions in these fields.

This particularly affects the role of mathematics in the military, to which we will devote only scattered remarks and no systematic discussion. While there are still considerable lacunae in the literature on mathematics during World War I (although some of these have been filled by publications prepared for the centenary), there is more to be found in print about mathematics in World War II, not least because of the increased role of that discipline in it. (We recommend Booß-Bavnbek and Høyrup (2003) as a good place to start to find out more than is covered in our article.)

Another topic that deserves broader coverage than is possible here is the history of philosophical reflection about mathematical applications. This is particularly true for the notion of "mathematical modeling" taken in the sense of problem formulation. According to the *Oxford Encyclopedic Dictionary* (1996), the new notion of a mathematical model was used first in a statistical context in 1901. At about the same time, the French physicist and philosopher Pierre Duhem accused British physicists of still using the term "model" only in the older and narrower sense of material, mechanical, or visualizable models. Duhem therefore preferred the word "analogy" for expressing the relationship between a theory and some other set of statements. Particularly with the upswing of "mathematical modeling" since the 1980s, a broad literature, often with a philosophical bent, has discussed the specificity of mathematics as a language, as an abstract unifier and a source of concepts and principles for various scientific and societal domains of application. Another (though not unrelated) development in the philosophy of applied mathematics concerns the growing importance of algorithmic aspects within mathematics as a whole. It was no coincidence that in the 1980s, with the rise of scientific computing, several "maverick" philosophers of mathematics, such as Philip Kitcher and Thomas Tymosczko, entered the scene. They introduced the notion of "mathematical practice," by which they meant more than simply applications. One of the features of the maverick tradition was the polemic against the ambitions of mathematical logic to be a canon for the philosophy of mathematics, ambitions that have dominated much of the philosophy of mathematics in the twentieth century. The change was inspired by the work of both those mathematicians (such as Philip Davis and Reuben Hersh) and those philosophers (including Imre Lakatos and David Corfield) who were primarily interested in the actual working process of mathematicians, or what they sometimes called "real mathematics." Meanwhile, the philosophical discussion of mathematical practice has been professionalized and reconnected to the foundationalist tradition. It usually avoids premature discussion of "big questions" such as "Why is mathematics applicable?" or "Is the growth of mathematics rational?" restricting its efforts to themes of mathematical practice in a broader sense, like visualization, explanation, purity of methods, philosophical aspects of the uses of computer science in mathematics, and so on. An overview of the more recent developments in the philosophy of mathematical practice is given in the introduction to Mancosu (2008).

Unfortunately, there is also little space for biographical detail in this article, and thus no bow can be given to the great historical heroes of applied mathematics, such as Archimedes, Ptolemy, Newton, Euler, Laplace, and Gauss. Nor is there room to report on the conversions of pure mathematicians into applied mathematicians, such as those undergone by Alexander Ostrowski, John von Neumann, Solomon Lefschetz, Ralph Fowler, Garrett Birkhoff, and David Mumford, all personal trajectories that paralleled the global development of mathematics. In any case, any systematic inclusion of biographies could not be restricted to mathematicians, considering the term in its narrowest sense. In an influential report on industrial mathematics in the *American Mathematical Monthly* of

Copyrighted material

1941, Thornton Fry spoke about a "contrast between the ubiquity of mathematics and the fewness of the mathematicians." Indeed, historically, engineers such as Theodore von Kármán, Richard von Mises, Ludwig Prandtl, and Oliver Heaviside; physicists such as Walter Ritz, Aleksandr Andronov, Cornelius Lanczos, and Werner Romberg; and industrial mathematicians such as Balthasar van der Pol have, by any measure, made significant contributions to applied mathematics. In addition, several pioneers of applied mathematics, such as Gaspard Monge, Felix Klein, Mauro Picone, Vladimir Steklov, Vannevar Bush, and John von Neumann, actively used political connections. The actions of nonscientists, and particularly politicians, have also therefore played a part in the development of the subject. For a full history of applied mathematics the concrete interplay of the interests of mathematicians, physicists, engineers, the military, industrialists, politicians, and other appliers of mathematics would have to be analyzed, but this is a task that goes well beyond the scope of this article.

In general, the historical origin of individual notions or methods of applied mathematics, which often have a history spanning several centuries, will not be traced here; pertinent historical information is often included in the specialized articles elsewhere in this volume. By and large, then, this article will focus on the broader methodological trends and the institutional advances that have occurred in applied mathematics since the early nineteenth century.

1.2 Periodization

From the point of view of applications, the history of mathematics can be roughly divided into five main periods that reveal five qualitatively different levels of applied mathematics, the first two of which can be considered as belonging to the prehistory of the subject.

(1) ca. 4000 B.C.E.–1400 C.E. Emergence of mathematical thinking, and establishment of theoretical mathematics with spontaneous applications.

(2) ca. 1400–1800. Period of "mixed mathematics" centered on the Scientific Revolution of the seventeenth century and including "rational mechanics" of the eighteenth century (dominated by Euler).

(3) 1800–1890. Applied mathematics between the Industrial Revolution and the start of what is often called the second industrial (or scientific–technical) revolution. Gradual establishment of both the term and the notion of "applied mathematics." France and Britain dominate applied mathematics, while Germany focuses more on pure.

(4) 1890–1945. The so-called resurgence of applications and increasing internationalization of mathematics. The rise of new fields of application (electrical communication, aviation, economics, biology, psychology), and the development of new methods, particularly those related to mathematical modeling and statistics.

(5) 1945–2000. Modern internationalized applied mathematics after World War II, inextricably linked with industrial mathematics and high-speed digital computing, led largely by the United States and the Soviet Union, the new mathematical superpowers.

Arguably, one could single out at least two additional subperiods of applied mathematics: the eighteenth century, with Euler's "rational mechanics," and the technological revolution of the present age accompanied by the rise of computer science since the 1980s. However, in the first of these subperiods, which will be described in some detail below, mathematics as a discipline was still not yet fully established, either institutionally or with respect to its goals and values, so distinguishing between her pure and applied aspects is not straightforward. As to the second of the two subperiods, we believe that these events are so recent that they escape an adequate historical description. Moreover, World War II had such strong repercussions on mathematics as a whole—particularly on institutionalization (journals, institutes, professionalization), on material underpinning (state funding, computers, industry), and not least on the massive migration of mathematicians to the United States—that it can be considered a watershed in the worldwide development of both pure and applied mathematics. However, the dramatic prediction by James C. Frauenthal—in an editorial of *SIAM News* in 1980 on what he considered the "revolutionary" change in applied mathematics brought about by the invention of the computer—that by 2025 "in only a few places will there remain centers for research in pure mathematics as we know it today" seems premature.

2 Mathematics before the Industrial Revolution

Since the emergence of mathematical thinking around 4000 B.C.E., through antiquity and up to the start of the Renaissance (ca. 1400 C.E.), and embracing the cultures

Copyrighted material

of Mesopotamia, Egypt, and ancient Greece, as well as those of China, India, the Americas, and the Islamic–Arabic world, applications arose as a result of various societal, technical, philosophical, and religious needs. Well before the emergence of the Greek notion of a mathematical proof around 500 B.C.E., areas of application of mathematics in various cultures included accountancy, agricultural surveying, teaching at scribal schools, religious ceremonies, and (somewhat later) astronomy. Among the methods used were practical arithmetic, basic geometry, elementary combinatorics, approximations (e.g., π), and solving quadratic equations. Instruments included simple measuring and calculation devices: measuring rods, compasses, scales, knotted ropes, counting rods, abaci, etc.

The six classical sciences—geometry, arithmetic, astronomy, music, statics, and optics—existed from the time of Greek antiquity and were based on mathematical theory, with the Greek word "mathemata" broadly referring to anything teachable and learnable. The first four of the classical sciences constituted the quadrivium within the Pythagorean–Platonic tradition. The theories of musical harmony (as applied arithmetic) and astronomy (as applied geometry) can thus be considered the two historically earliest branches of applied mathematics. The two outstanding applied mathematicians of Greek antiquity were Archimedes (statics, hydrostatics, mechanics) and Ptolemy (astronomy, optics, geography). Since the early Middle Ages, the Computus (Latin for computation)—the calculation of the date of Easter in terms of first the Julian calendar and later the Gregorian calendar—was considered to be the most important computation in Europe. In medieval times, particularly from the seventh century, the development of algebraic and calculative techniques and of trigonometry in the hands of Islamic and Indian mathematicians constituted considerable theoretical progress and a basis for further applications, with significant consequences for European mathematics. Particularly notable was the *Liber Abaci* (1202) of Leonardo of Pisa (Fibonacci), which heralded the gradual introduction of the decimal positional system into Europe, one of the broadest and most important applications of mathematics during the period. Chinese mathematics remained more isolated from other cultures at the time and is in need of further historical investigation, as are some developments within Christian scholastics. In spite of their relative fewness and their thematic restrictions, we consider the early applications to be a deep and historically important root for the emergence of theoretical mathematics and not as a mere follow-up of the latter.

From the beginning of the fifteenth century to the end of the eighteenth century, applications of mathematics were successively based on the dissemination of the decimal system, the rise of symbolic algebra, the theory of perspective, functional thinking (Descartes's coordinates), the calculus, and natural philosophy (physics). The teaching of practical arithmetic, including the decimal system, by professional "reckoning masters," such as the German Adam Ries in the sixteenth century, remained on the agenda for several centuries. Meanwhile, the first systematic discussion of decimal fractions appeared in a book by the Dutch engineer Simon Stevin in 1585.

During this period, and connected to the new demands of society, there emerged various hybrid disciplines combining elements of mathematics and engineering: architecture, ballistics, navigation, dynamics, hydraulics, and so on. Their origins can be traced back, at least in part, to medieval times. For example, partly as a result of fourteenth-century scholastic analysis, the subject of local motion was separated from the traditional philosophical problem of general qualitative change, thus becoming a subject of study in its own right.

The term "mixed mathematics" as a catch-all for the various hybrid disciplines seems to have been introduced by the Italian Marsilio Ficino during the fifteenth century in his commentary on Plato's *Republic*. It was first used in English by Francis Bacon in 1605. In his *Mathematicall Præface* to the first English translation of Euclid's *Elements* (1570), John Dee set out a "groundplat" or plan of the "sciences and artes mathematicall," which included astronomy and astrology. Due to the broad meaning of the original Greek word, the Latin name "mathematicus" was used for almost every European practitioner or artisan within one of these hybrid disciplines. As late as 1716, the loose use of "mathematicus" was deplored by the philosopher Christian Wolff (a follower of Leibniz) in his *Mathematisches Lexicon*, an influential dictionary of mathematics, because in his opinion it diminished the role of mathematics.

The emergence of the new Baconian sciences (magnetism, electricity, chemistry, etc.)—which went beyond mixed mathematics and were even partially opposed to the mathematical spirit of the classical sciences (Bacon's acknowledgment of the future of mixed mathematics, as expressed in the epigraph, was coupled with a certain distrust of pure mathematics)—signaled

Copyrighted material

In our opinion (deviating slightly from Truesdell), rational mechanics, in hindsight, bears almost all the characteristics of applied mathematics in the modern sense. At the time, however, in a predominantly utilitarian environment, it was the pinnacle of mathematics. It was then rarely counted as mixed mathematics, notwithstanding some occasional remarks by d'Alembert. The term mixed mathematics was more frequently used for the mathematically less advanced engineering mechanics (Bernard Forest de Bélidor and Charles-Augustin de Coulomb, etc.) of the time and for other fields of application.

Toward the end of the eighteenth century rational mechanics was somewhat narrowed down, both thematically and with respect to possible applications (although still including continuum mechanics), by further mathematical formalization, particularly at the hands of Lagrange, Euler's successor in Berlin, whose *Méchanique Analitique* first appeared in 1788. The towering figure of Pierre Simon Laplace in Paris—with his pioneering work since the late 1770s in celestial and terrestrial mechanics and in probability theory—foreshadowed much of the important French work in applied mathematics, such as that done by Poisson, Fourier, Cauchy, and others in the century that followed. To Laplace (generating functions, difference equations) and to his great younger contemporary Carl Friedrich Gauss in Göttingen (numerical integration, elimination, least-squares method) we owe much of the foundations of future numerical analysis. Parts of their work overlapped (interpolation), while parts were supplemented by Adrien-Marie Legendre (least-squares method), details of which can be traced from Goldstine's *A History of Numerical Analysis.*

3 Applied Mathematics in the Nineteenth Century

Around 1800, in the age of the Industrial Revolution and of continued nation building, state funding and political and ideological support (revolution in France, Neo-Humanism in Germany) led (mainly through teaching and journals) to a new level of recognition for mathematics as a discipline. The older bifurcation of pure/mixed mathematics was replaced in France and Germany (although not yet in England) by that of pure/applied. The difference was mainly that before 1800 only mixed mathematics together with rational mechanics had the support of patrons, while now, around 1800, the whole of mathematics was beginning to be supported and recognized. Somewhat paradoxically, then, in spite of the general importance of the Industrial Revolution as a historical background, it is pure mathematics that increasingly gets systematic public support for the first time. Indeed, for most of the nineteenth century, mathematics would not be strongly represented in either engineering or industrial environments.

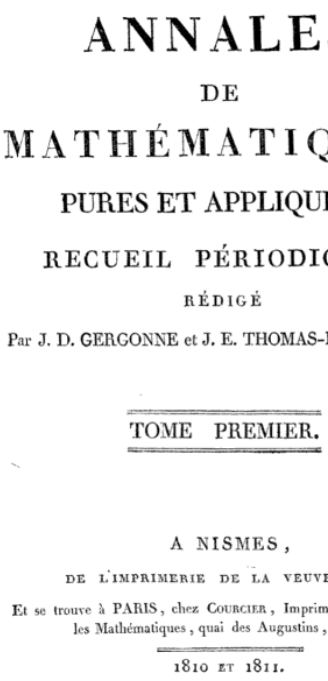
The foundation of the École Polytechnique (EP) in Paris in 1794 is a good point of reference for the beginning of our third period. The EP, where military and civil engineers were trained, became the leading and "most mathematical" institution within a system of technical education. This included several "schools of applications," such as the École des Mines and the École Nationale des Ponts et Chaussées, to which the students of the EP proceeded. The EP became an example to be emulated by many technical colleges, particularly in German-speaking regions, throughout the nineteenth century. The most influential mathematician in the early history of the EP was Gaspard Monge, and it was in accordance with his ideas that mathematics became one of the bases of the EP curriculum. In 1795, in the introduction to his lectures on descriptive geometry, the theory that became the "language of the engineer" for more than a century, Monge wrote:

> In order to reduce the dependence of the French nation on foreign industry one has to direct public education to those subjects which require precision.

Monge's aspirations for a use of higher mathematics in industrial production remained largely unfulfilled at the time, except for the use of descriptive geometry. However, developments in industry and in educational systems led to a stronger focus on the criteria for precision and exactitude in the sciences (most notably in academic physics) and in engineering, preparing the ground for an increased use of mathematics in these fields of application at the beginning of the twentieth century. In fact, it could be argued that it required a logical consolidation and a more theoretical phase of the development of mathematical analysis before a new phase of more sophisticated applied mathematics could set in.

The first concrete institutional confirmation of the notion of "applied mathematics" was the appearance of the term in the names of journals. Again, the Germans were quicker than the French here. Two short-lived journals cofounded by the influential combinatorialist Carl Friedrich Hindenburg were the *Leipziger*

Copyrighted material



ANNALES

DE

MATHÉMATIQUES

PURES ET APPLIQUÉES.

RECUEIL PÉRIODIQUE,

RÉDIGÉ

Par J. D. GERGONNE et J. E. THOMAS-LAVERNÈDE.

TOME PREMIER.

A NISMES,

DE L'IMPRIMERIE DE LA VEUVE BELLE.

Et se trouve à PARIS, chez COURCIER, Imprimeur-Libraire pour les Mathématiques, quai des Augustins, n.° 57.

1810 ET 1811.

Figure 4 Gergonne's *Annales de Mathématiques Pures et Appliquées* (1810–11).

Magazin für reine und angewandte Mathematik (1786–89) and the *Archiv für reine und angewandte Mathematik* (1795–99). A somewhat longer career was had by *Annales de Mathématiques Pures et Appliquées*, founded by Joseph Diaz Gergonne in 1810 (figure 4). While this journal survived only until 1832, the German *Journal für die reine und angewandte Mathematik* (which, according to the preface by its founder August Leopold Crelle in 1826, was largely modeled after Gergonne's journal) is still extant today. This is true too of the French *Journal de Mathématiques Pures et Appliquées*, founded by Joseph Liouville in 1836, and of the Italian *Annali di Matematica Pura ed Applicata*, launched by Francesco Brioschi and Barnaba Tortolini in Italy in 1858 as an immediate successor to the *Annali di Scienze Matematiche e Fisiche*. On the other hand, James Joseph Sylvester's *Quarterly Journal of Pure and Applied Mathematics*, which was founded in 1855, survived only until 1927.

The inclusion of "applied mathematics" in the names of these nineteenth-century journals did not necessarily guarantee a strong representation of applied topics, however, either in the journals themselves or in the mathematical culture at large. But neither were these journals the only outlets for articles on applied topics. Journals associated with national academies, such as the *Philosophical Transactions of the Royal Society*, carried articles on applied topics, while the *Philosophical Magazine* (launched in 1798) was the journal of choice for several leading nineteenth-century British applied mathematicians.

This was also the period in which positions explicitly devoted to applications were created at universities. In Norway, which had just introduced a constitution and was emancipating itself from Danish rule, Christopher Hansteen's position as "lecturer for applied mathematics" ("Lector i den anvendte Mathematik") at the newly founded university in Christiania was expressly justified in May 1814 by "the broad scope of applied mathematics and its importance for Norway." In 1815 Hansteen was promoted to "Professor Matheseos applicatae."

Throughout the nineteenth century, the mathematization of mechanics continued largely in the tradition of Lagrange's analytical mechanics, with a division of labor between physicists and mathematicians such as William Rowan Hamilton and Carl Jacobi, arguably neglecting some of the topics and insights of Euler's rational mechanics, particularly in continuum mechanics. However, from the 1820s, although the EP still gave preference to analytical mechanics in its courses, there were efforts among the professors there, and at the more practically oriented French engineering schools ("écoles d'application"), to develop a mechanics for the special needs of engineers, a discipline that would today be called technical mechanics. The latter drew strongly on traditions in mixed mathematics, such as the work of de Bélidor in hydraulics from the 1730s to the 1750s and that of de Coulomb in mechanics and electromagnetism from the 1780s onward. It found its first energetic proponents in Claude Navier, Jean Victor Poncelet, and Gaspard Gustave de Coriolis.

Around 1820, Poncelet separately developed his projective geometry, which became part of the mathematically rather sophisticated engineering education at several continental technical colleges. It led to methods such as graphical statics, founded by the German–Swiss Carl Culmann in the middle of the century, with applications in crystallography and civil engineering, the latter exemplified by the construction of the Eiffel Tower in 1889. Also in the 1820s, influenced by Euler's hydrodynamics and possibly by Navier's work in engineering,

Copyrighted material

Augustin-Louis Cauchy was the first to base the theory of elasticity on a general definition of internal stresses. The work of Cauchy, who was at the same time known for his efforts to introduce rigor into analysis, underscores the dominance of the French in both pure and applied mathematics during the early nineteenth century, with the singular work of Gauss in Göttingen being the only notable exception.

In England the development of both pure and applied mathematics during the nineteenth century showed marked differences from that in Continental Europe. One of the goals of the short-lived Analytical Society (1812–19), founded in Cambridge by Charles Babbage and others, was to promote Leibnizian calculus over Newtonian calculus, or, in Babbage's words, to promote "The Principles of pure D-ism in opposition to the Dotage of the University." The members of the Analytical Society were impressed by the new rigor in analysis achieved in France, especially in the work of Lagrange, and lobbied for a change in teaching and research in Cambridge mathematics, and in particular in the examinations of the Mathematical Tripos, which were very much based on traditional mixed and physical mathematics, as well as on Euclid's *Elements*. If anything, though, this aspect of the French influence led away from applications and toward a gradual purification of British mathematics.

Babbage was impressed by the French mathematical tables project directed by Gaspard de Prony at the end of the eighteenth century. In a similar vein to Monge before him, Babbage pointed to increased competition between nations in the age of industrialization, and he stressed the need for the development of calculating techniques. In *On the Economy of Machinery and Manufactures* (1832) he wrote:

> It is the science of *calculation*,—which becomes continually more necessary at each step of our progress, and which must ultimately govern the whole of the applications of science to the arts of life.

Another (at least indirect) impact of the Industrial Revolution on mathematics was the Russian Pafnuty Lvovich Chebyshev's study of James Watt's steam engine, in particular of the "governor," the theory of which proved to be a stimulus for the notion of feedback in control theory, and the modern theory of servomechanisms. Chebyshev's interest in the technical mechanics of links was also one of the stimuli for his studies concerning mathematical approximation theory in the 1850s. In addition, he was impressed with Poncelet's technical mechanics. As a result, and due to Chebyshev's great influence within Russian mathematics, applied mathematics remained much more part of mainstream mathematics in Russia during the latter half of the nineteenth century than it was in other parts of Europe, especially in Germany.

In the middle of the nineteenth century, the French engineering schools, in particular the EP, lost their predominant position in mathematics, due to slow industrial development in France and problems with the overcentralized and elitist educational system. The lead was taken by the German-speaking technical colleges ("Technische Hochschulen") in Prague, Vienna, Karlsruhe, and Zurich, in particular with respect to the mathematization of the engineering sciences. This was true for their emulation of the general axiomatic spirit of mathematics even more than for their concern for the actual mathematical details. Ferdinand Redtenbacher (in his analytical machine theory (1852)) and Franz Reuleaux (in his kinematics (1875)) aimed at "designing invention and construction deductively." So convinced of the important future role of mathematics were leading engineers at the Technische Hochschulen that they supported the appointment of academically trained mathematicians from the classical universities. In this way pure mathematicians, such as Richard Dedekind, Alfred Clebsch, and later Felix Klein, assumed positions at Technische Hochschulen in which they were responsible for the education of engineers.

In parallel, and also from the middle of the nineteenth century, mathematics at the leading German universities that did not have engineering departments increasingly developed into a pure science, detached from practical applications. Supported by the ideology of "Neo-Humanism" within a politically unmodernized environment, the discipline's educational goal (and its legitimation in society) was the training of high school teachers, who during their studies were often introduced to the frontiers of recent (pure) mathematical research. The result of this was that professors at the Technische Hochschulen who were hired from the traditional universities were not really prepared for training engineers. In the long run, the strategy of appointing university mathematicians backfired and this, together with general controversies about the social status of technical schools, led to the so-called anti-mathematical movement of engineers in Germany in the 1890s.

British and Irish applied mathematics, in the sense of mathematical physics, remained strong through the

Copyrighted material

nineteenth century, with work by George Green, George Stokes, William Rowan Hamilton, James Clerk Maxwell, William Thomson (Lord Kelvin), Lord Rayleigh, William Rankine, Oliver Heaviside, Karl Pearson, and others, while in the works of James Joseph Sylvester and Arthur Cayley in the 1850s, the foundations of modern matrix theory were laid. That being said, there was no systematic, state-supported technical or engineering education in the British system until late in the nineteenth century. What was taught in this respect at schools and traditional universities was increasingly questioned by engineers such as John Perry (see below), particularly with respect to the mathematics involved. In England the term "mixed mathematics" was occasionally used interchangeably with "applied mathematics" up until the end of the century, a prominent example of this being the tribute by Richard Walker (then president of the London Mathematical Society) to Lord Rayleigh on winning the society's De Morgan Medal in 1890.

Applications of mathematics also featured among the activities of the British Association for the Advancement of Science, which, in 1871, formed a Mathematical Tables Committee for both cataloguing and producing numerical tables; the committee lasted, with varying levels of intensity, until 1948, when the Royal Society took over. A prime example of joint enterprise between pure and applied mathematicians—the original committee consisted of Cayley, Henry Smith, Stokes, and Thomson—the project catered for all tastes, its products including both factor tables and Bessel function tables, among others. As J. W. L. Glaisher, the project secretary, wrote in 1873: "one of the most valuable uses of numerical tables is that they connect mathematics and physics, and enable the extension of the former to bear fruit practically in aiding the advance of the latter." The project was finally dissolved in 1965, with some of the greatest British mathematicians, both pure and applied, having been active in its work.

With the upswing of electrical engineering, more sophisticated mathematics (operational calculus, complex numbers, vectors) finally entered industry in around 1890, e.g., through the work of Heaviside in England and of the German immigrant Charles P. Steinmetz in the United States. Mechanical engineering, on the other hand, e.g., in the construction of turbines, remained free of advanced mathematics until well into the twentieth century.

It was also not until the end of the nineteenth century that applied mathematics finally began to lose its almost exclusive bond to mathematical physics and mechanics; new fields of application, new methods such as statistics, and new professions such as actuarial and industrial mathematicians were largely matters for the twentieth century.

This change is nicely captured through the example of the English applied mathematician Karl Pearson. In his philosophical book *Grammar of Science* (1892), Pearson, who at the time was mainly known for his work on elasticity, defined as the "topic of Applied Mathematics ... the process of analyzing inorganic phenomena by aid of ideal elementary motions." At the time Pearson was already working on biometrics, the subject that would lead him to found, together with Francis Galton, the journal *Biometrika* in 1901. Therefore, although Pearson was effectively extending the realm of applications of mathematics to the statistical analysis of biological (i.e., organic) phenomena, he apparently did not consider what he was doing to be applied mathematics.

4 The "Resurgence of Applications" and New Developments up to World War II

From the 1890s, the University of Göttingen (pure) mathematician Felix Klein saw the importance of taking the diverging interests of the engineering professors at technical colleges and those of German university mathematicians into account. Not only did different professions (teaching and engineering) require different education, but the gradual emergence of industrial mathematics had to be considered as well. Klein recognized the need for reform, including in teaching at high school level, and he developed Göttingen into a center of mathematics and the exact sciences (figure 5). Chairs for applied mathematics and applied mechanics were created there in 1904, with Carl Runge and Ludwig Prandtl being the first appointees. Meanwhile, from 1901, and under the editorship of Runge, the transformation of *Zeitschrift für Mathematik und Physik* into a journal exclusively for applied mathematics had begun. These events in Germany, contrasting with those of the period before, led to talk about a "resurgence of applications" ("Wiederhervorkommen der Anwendungen").

From 1898 and for several decades afterward, the famous German multivolume *Encyclopedia of the Mathematical Sciences including Their Applications* was edited by Klein together with Walther von Dyck and Arnold Sommerfeld, both from Munich, and others.

Copyrighted material


FELIX KLEIN
ELEMENTARMATHEMATIK
VOM HÖHEREN STANDPUNKTE AUS
DRITTE AUFLAGE
DRITTER BAND
PRÄZISIONS- UND
APPROXIMATIONSMATHEMATIK
AUSGEARBEITET VON
C. H. MÜLLER
FÜR DEN DRUCK FERTIG GEMACHT
UND MIT ZUSÄTZEN VERSEHEN VON
FR. SEYFARTH
MIT 156 ABBILDUNGEN
BERLIN
VERLAG VON JULIUS SPRINGER
1928

Figure 5 F. Klein, *Präzisions- und Approximationsmathematik* (1928). Posthumous publication of Klein's 1901 lectures in Göttingen where he tried to differentiate between a mathematics of precision and one of approximation.

The articles in it, all written in German but including authors from France and Britain, such as Paul Painlevé and Edmund Taylor Whittaker, contain valuable historical references that are still worth consulting today. "Applications" as emphasized in the title and in the program of the *Encyclopedia* meant areas of application, such as mechanics, electricity, optics, and geodesy. The articles were assigned to volumes IV–VI, which were in themselves divided into several voluminous books each. There were also articles on mechanical engineering, such as those by von Mises and von Kármán. However, topics that would today be classified as core subjects of applied mathematics—such as numerical calculation (Rudolph Mehmke), difference equations (Dmitri Seliwanoff), and interpolation and error compensation (both by Julius Bauschinger)—appeared as appendices within volume I, which was devoted to pure mathematics (arithmetic, algebra, and probability). Runge's contribution on "separation and approximation of roots" (1899) was subsumed under "algebra."

Klein also succeeded in introducing a state examination in applied mathematics for mathematics teachers, which focused on numerical methods, geodesy, statistics, and astronomy. In addition, he inspired educational reform of mathematics in high schools that he designed around the notions of "functional thinking" and "intuition," thereby trying to counteract the overly logical and arithmetical tendencies that had until then permeated mathematics education. Klein and his allies insisted on taking into account international developments in teaching and research, for instance by initiating a series of comparative international reports on mathematical education; these reports in turn led to the creation of what has now become the International Commission on Mathematical Instruction (ICMI). The *Encyclopedia* also provided evidence of the increasing significance of the international dimension. In his "introductory report" in 1904, von Dyck stressed the importance for the project of securing foreign authors in applied mathematics. Later, a French translation of the *Encyclopedia* began to appear in a considerably enlarged version, although the project was never completed due to the outbreak of World War I.

Around 1900, reform movements reacting to problems in mathematics education similar to those in Germany existed in almost all industrialized nations. In England, the engineer John Perry had initiated a reform of engineering education in the 1890s, and this reform played into the ongoing critical discussions of the antiquated Cambridge Mathematical Tripos examinations and their traditional reliance on Euclid. The "Perry Movement" was noticed in Germany and in the United States. On the pages of *Science* in 1903, the founding father of modern American mathematics, Eliakim Hastings Moore, himself very much a pure mathematician, declared himself to be in "agreement with Perry" and proposed a "laboratory method of instruction in mathematics and physics." At about the same time (1905), similar ideas "de créer de vrais laboratoires de Mathématiques" were proposed by Émile Borel in France. In Edinburgh, Whittaker instituted a "Mathematical Laboratory" in 1913 and later, together with George Robinson, published the influential *The Calculus of Observations: A Treatise on Numerical Mathematics* (1924), which derived from Whittaker's lectures given in the Mathematical Laboratory. In Germany, the

Copyrighted material

Research Council's National Institute for the Application of the Calculus under Mauro Picone in Naples and later in Rome.

In contrast, Britain created no new institutes for mathematics. Even G. H. Hardy, who had left Cambridge for Oxford shortly after the war, was unable to persuade his new university to build one. Nevertheless, the war left a tangible legacy for applied mathematics. Imperial College received a substantial grant to finance its Department of Aeronautics, while Cambridge established a new chair in aeronautical engineering. As a result of increased funding after the war, establishments such as the Royal Aircraft Establishment and the National Physical Laboratory were able to retain a number of their wartime staff, several of whom were mathematicians. Notable inclusions were Hermann Glauert, who made a career in aerodynamics at the Royal Aircraft Establishment, and Robert Frazer, who worked on wing flutter at the National Physical Laboratory. In the 1930s Frazer and his colleagues W. J. Duncan and A. R. Collar were "the first to use matrices in applied mathematics." In addition, theoreticians and practitioners who were brought together because of the war worked together afterward. Sometimes, as in the case of the Cambridge mathematician Arthur Berry and the aeronautical engineer Leonard Bairstow, the end of hostilities meant only the end of working in the same location, it did not mean the end of collaboration.

In the interwar period the degree of industrialization in a particular country was without doubt one of the defining factors in that country's support of applied mathematics. This is well exemplified by the solid development of applied mathematics in industrialized Czechoslovakia compared with the strong tradition in pure mathematics in less industrialized Poland.

Indeed, it became increasingly obvious after the war that engineering mathematics and insurance mathematics, both of which corresponded to the developing needs of the new professions and industries, had become legitimate parts of applied mathematics. Not only were they the most promising areas of the subject, but they were economically the most rewarding. Students trained at von Mises's institute in Berlin and at Prandtl's institute in Göttingen found jobs in various aerodynamic laboratories and proving grounds, as well as in industry. Von Mises himself both undertook governmental assignments and acted as an advisor for industry. At Siemens, AEG, and Zeiss (all in Germany), the General Electric Company (in Britain), Philips (in the Netherlands), and General Electric and Bell Laboratories (in the United States) (Millman 1984), industrial laboratories (mainly in electrical engineering but also, for instance, in the optical and aviation industries) developed a demand for trained mathematicians. It was the study of the propagation of radio waves and of the electrical devices required to generate them that led in 1920 to the Dutchman van der Pol working out the equation that is to this day considered as the prototype of the nonlinear feedback oscillator. VAN DER POL'S EQUATION [IV.2 §10]) and his modeling approach have repeatedly been cited as exemplars for modern applied mathematics. Van der Pol's contribution, together with theoretical work by Henri Poincaré on limit cycles, strongly influenced Russian work on nonlinear mechanics. Its mathematical depth gained the approval (albeit somewhat reluctant approval) even of André Weil, a foremost member of the Bourbaki group of French mathematicians, who in 1950 called it "one of the few interesting problems which contemporary physics has suggested to mathematics." Van der Pol, who worked at the Philips Laboratories in Eindhoven from 1922, also contributed to the justification of the Heaviside operational calculus in electrical engineering. Around 1929 he used integral transformation methods similar to those developed before him by the English mathematician Thomas Bromwich and the American engineer John Carson at Bell Laboratories, who in 1926 wrote the influential book *Electric Circuit Theory and the Operational Calculus*. Somewhat later, the German Gustav Doetsch provided a more systematic justification of Heaviside's calculus based on the theory of the Laplace transform in his well-received book *Theorie und Anwendung der Laplace-Transformation* (1937). In another influential book, *Economic Control of Quality of Manufactured Product* (1931), the physicist Walter Shewart, a colleague of Carson's at Bell Laboratories, was one of the first to promote statistics for industrial quality control using so-called control charts.

However, many of these developments in applied and industrial mathematics, both in Europe and America, occurred outside their national academic institutions, notwithstanding the beginnings of systematic academic training in applied mathematics in new institutes such as the one led by von Mises. A number of academically trained mathematicians and physicists were impressed by the spectacular and revolutionary ideas of relativity theory and quantum theory, but they were slow to recognize the importance of those new applications, often in engineering, that relied on classical mechanics.

Copyrighted material

This aloofness of academic scientists prevailed in the United States. The undisputed leader of American mathematicians, George Birkhoff, was aware of that when he addressed the American Mathematical Society at its semicentennial in 1938 with the following words:

> The field of applied mathematics always will remain of the first order of importance inasmuch as it indicates those directions of mathematical effort to which nature herself has given approval.
>
> Unfortunately, American mathematicians have shown in the last fifty years a disregard for this most authentically justified field of all.

There were exceptions, such as Norbert Wiener at the Massachusetts Institute of Technology, who was based in the mathematics department but interacted with the electrical engineering department run by Vannevar Bush (the inventor of the "differential analyzer," an analogue computer) and through it with Bell Laboratories, and there were also the individual efforts of a number of mathematicians with a European background. One of the most successful of the latter was Harry Bateman, Professor of Aeronautical Research and Mathematical Physics at Caltech in Pasadena, who became a champion of special functions during the 1930s and 1940s, and who had earlier (immediately prior to his emigration from England in 1910) discussed the Laplace transform and applications to differential equations. However, American academia was late in recognizing applied mathematics, as exemplified by the abovementioned report on *Numerical Integration of Differential Equations* (1933), in which the authors write that the report was produced "without special grant for relief from teaching from any of the institutions represented." Mathematical physicist Warren Weaver (who later, in World War II, would lead the Applied Mathematics Panel within the American war effort) was surprised, as late as 1930, "at the emphasis given, in the discussion [on a planned journal for applied mathematics], to the field between mathematics and engineering." During the 1920s and 1930s, Rockefeller money had primarily been geared toward supporting pure academic mathematical and physical research, leaving applied research in the hands of industry. It was left to the clever negotiations of Richard Courant (Göttingen's adherent to applied mathematics) to win Rockefeller fellowships for the applied candidates under his tutelage, such as Wilhelm Cauer and Alwin Walther.

The 1920s and 1930s were also a time in which mathematical modeling came to the fore, although the term "mathematical modeling" was rarely used before World War II. In a 1993 article on the emergence of biomathematics in *Science in Context*, the author Giorgio Israel emphasizes the increasing role of mathematical modeling in the nonphysical sciences:

> Another important characteristic of the new trends of mathematical modeling and applied mathematics is interest in the mathematization of the nonphysical sciences. The 1920s offer in fact an extraordinary concentration of new research in these fields, which is developed from points of view more or less reflecting the modeling approach. So the systematic use of mathematics in economics (both in the context of microeconomics and game theory) is found in the work of K. Menger, J. Von Neumann, O. Morgenstern, and A. Wald, starting from 1928. The basic mathematical model of the spread of an epidemic (following the research of R. Ross on malaria) was published in 1927 [by W. O. Kermack and A. G. McKendrick]; the first papers by S. Wright, R. A. Fisher and J. B. S. Haldane on mathematical theory of population genetics appeared in the early twenties; the first contributions of Volterra and Lotka to population dynamics and the mathematical theory of the struggle for existence were published in 1925 and 1926; and many isolated contributions (such as van der Pol's model) also appeared in these years.

Moreover, during the twentieth century there was a certain tendency for mathematicians to be less inspired by physics and to resort instead to less rigorous or less complete models from other sciences, including engineering. In 1977 Garrett Birkhoff, George Birkhoff's son, wrote:

> Engineers and physicists create and adopt mathematical models for very different purposes. Physicists are looking for universal laws (of "natural philosophy"), and want their models to be exact, universally valid, and philosophically consistent. Engineers, whose complex artifacts are usually designed for a limited range of operating conditions, are satisfied if their models are reasonably realistic under those conditions. On the other hand, since their artifacts do not operate in sterilized laboratories, they must be "robust" with respect to changes in many variables. This tends to make engineering models somewhat fuzzy yet kaleidoscopic. In fluid mechanics, Prandtl's "mixing length" theory and von Kármán's theory of "vortex streets" are good examples; the "jet streams" and "fronts" of meteorologists are others.

The same author, himself a convert from abstract algebra to hydrodynamics, explains resistance to mathematical models in economics, pointing to the fact that they did not fit well into Bourbaki's "conventional

Copyrighted material

framework of pure mathematics." The latter has often been described as in some respects being inimical to applications and (since the "New Math" of the 1960s) as being pedagogically disastrous. However, as detailed by Israel in the paper quoted above, the relationship between Bourbaki and the new practices in modeling has not necessarily been negative. Some mathematicians considered Bourbaki's notion of mathematics as an "abstract scheme of possible realities" to be the right way to liberate mathematics from the classical reductionist mechanistic approach that had often relied on linearization methods. There have even been efforts, for instance by logicians, to introduce "planned artificial language" into the sciences, as exemplified in J. H. Woodger's *The Axiomatic Method in Biology* (1937). However, these efforts seem to have had limited success. It took another step in the development of computers in the 1980s before necessarily simplified models of biological processes could be abandoned, and investigations of cellular automata, membrane computing, simulation of ecological systems, and similar tasks from modern mathematical biology could be undertaken.

During the 1920s and 1930s, many further results in different fields of application were obtained. Well-known examples include Alan Turing's work during the 1930s on the theory of algorithms and computability, and the Russian Leonid Kantorovich's work on linear programming within an economic context (1939), which escaped the attention of Western scholars for several decades.

This was also a period in which some of the foundations were laid for what would, from the late 1940s on, be called numerical analysis. In 1928 Courant and his students Kurt Friedrichs and Hans Lewy, all three of whom eventually emigrated to the United States, published "On the partial difference equations of mathematical physics" in *Mathematische Annalen*. The paper was translated in the *IBM Journal of Research and Development* as late as 1967 on the grounds that it was "one of the most prophetically stimulating developments in numerical analysis ... before the appearance of electronic digital computers.... The ideas exposed still prevail." In the history of numerical analysis, the paper gained special importance because it contains the germ of the notion of numerical stability and involves the problem of well-posedness of partial differential equations (as proposed by Hadamard in 1902).

5 Applied Mathematics during and after World War II

World War II, like World War I, was not a mathematicians' war. Indeed, in early 1942 the chemist, and Harvard president, James Conant said: "The last was a war of chemistry but this one is a war of physics." This of course partially reflected the increasing role of mathematics in World War II, revealed by the use of ballistics, operations research, statistics, and cryptography throughout the conflict. In fact, the president of the American National Academy of Sciences, the physicist Frank B. Jewett, responded to Conant with the words: "It may be a war of physics but the physicists say it is a war of mathematics." However, at the time, due to lingering tradition, mathematics was not given the same high priority as the other sciences either in the preparation for warfare nor in war-related research. In the early 1940s within the leading research organizations in the United States, in Germany, and in other countries, mathematics was still subordinate to other fields, such as engineering and physics. In addition, the mathematicians themselves were not prepared for a new and broader social role, e.g., as professionals in industry, such as might be demanded by the war. When considering the future of their field during and after the war, many pure mathematicians were worried that mathematics would suffer from a too utilitarian point of view. This is exemplified by the well-known essay *A Mathematician's Apology* written by the leading English mathematician G. H. Hardy in 1940.

But not long after Hardy's essay was written, another Cambridge mathematician, Alan Turing, demonstrated the potential of sophisticated mathematics—a mix of logic, number theory, and Bayesian statistics—for warfare, when he and his collaborators at Bletchley Park broke the code of the German Enigma machine.

In Germany, the Diplommathematiker (mathematics degree with diploma), which was designed for careers in industry and the civil service, was officially introduced in 1942, and teaching as a career for mathematicians began to lose its monopoly.

The entry of the United States into World War II in December 1941 brought with it deep changes in the way mathematicians worked together with industry, the military, and government. In the *American Mathematical Monthly*, rich memoirs on the state of industrial mathematics and (academic) applied mathematics in the United States by Thornton Fry (1941) and Roland

Copyrighted material

Richardson (1943), respectively, were published. Probably the most spectacular development in communications mathematics took place in the 1940s at Bell Laboratories with the formulation of information theory by Claude Shannon.

Based on their European experiences, immigrants to the United States such as von Kármán, Jerzy Neyman, von Neumann, and Courant contributed substantially to a new kind of collaboration between mathematicians and users of mathematics. In the mathematical war work organized by the Applied Mathematics Panel, where the leading positions were occupied by Americans, with Warren Weaver at the head, the applied mathematicians cooperated with mathematicians of an originally purer persuasion, natives of the United States (Oswald Veblen, Marston Morse) and immigrants (von Neumann) alike.

As well as their political and administrative experience, the immigrants brought to their new environment European research traditions from engineering mathematics, classical analysis, and discrete mathematics. Ideas, such as those of von Neumann in theoretical computing, could gradually mature and materialize within the industrial infrastructure of the United States (Bell Laboratories, etc.), aided during the war by seemingly unlimited public money (Los Alamos, etc.). In March 1945, while the war was still on, von Neumann sent a famous memo on the "Use of variational methods in hydrodynamics" to Veblen. Von Neumann recommended the "great virtue of Ritz's method" and deplored that before, and even during, the war mathematical work had not been sufficiently centralized for a systematic attack on the nonlinear equations occurring in fluid mechanics and related fields. In the same memo von Neumann pointed to the "increasing availability of high-power computing devices," a development to which he had of course contributed substantially. As mentioned in the introduction to the SIAM "History of numerical analysis and scientific computing" Web pages:

> Modern numerical analysis can be credibly said to begin with the 1947 paper by John von Neumann and Herman Goldstine, "Numerical inverting of matrices of high order" (*Bulletin of the AMS*, Nov. 1947). It is one of the first papers to study rounding error and include discussion of what today is called scientific computing.

Von Neumann and Goldstine's results were soon followed up and critically discussed by English mathematicians (Leslie Fox and James Wilkinson, as well as Alan Turing) at the National Physical Laboratory at Teddington.

After the war, the increased level of U.S. federal funding for mathematics was maintained. Although partly fueled by the beginning of the Cold War, it was nevertheless no longer restricted to applications. Much of it was channeled through the department of defense (e.g., by the Office of Naval Research) and the new National Science Foundation (NSF), which was founded in 1950. The NSF was initiated by the electrical engineer Vannevar Bush, who had led the Office for Scientific Research and Development, the American war-research organization. The concerns about exaggerated utilitarianism that were harbored by pure mathematicians before the war therefore turned out to be groundless.

As George Dantzig, the creator of the SIMPLEX METHOD [IV.11 §3.1], observed, the outpouring of papers in linear programming between 1947 and 1950 coincided with the building of the first digital computers, which made applications in the field possible. Mathematical approaches to logistics, warehousing, and facility location were practiced from at least the 1950s, with early results in optimization by Dantzig, William Karush, Harold Kuhn, and Albert Tucker being enthusiastically received by (and utilized in the logistics programs of) the United States Air Force and the Office of Naval Research. These optimization techniques are still highly relevant to industry today.

The papers of the 1950 Symposium on Electromagnetic Waves, sponsored by the United States Air Force and published in the new journal *Communications on Pure and Applied Mathematics*, summarized the war effort in the field. Richard Courant, by then at New York University, pointed to the importance of a new approach to classical electromagnetism, where "a great number of new problems were suggested by engineers." This strengthened the feeling, already evident before the war, that the predominance of academic mathematical physics as the main source of inspiration for mathematical applications had begun to wane. Ironically, Courant's paper of 1943 on variational methods for the solution of problems of equilibrium and vibrations, which would later be widely considered to be one of the starting points for the FINITE-ELEMENT METHOD [II.12] (the name being coined by R. W. Clough in 1960), lay in obscurity for many years because Courant, not being an engineer, did not link the idea to networks of discrete elements. Another reason for the later breakthrough was a development of variational methods of

Copyrighted material

approximation in the theory of partial differential equations, which would ultimately prove to be vital to the development of finite-element methods in the 1960s (see J. T. Oden's chapter in Nash (1990)). A thoughtful historical look at the different ways mathematicians and engineers use finite-element methods is given in Babuška (1994).

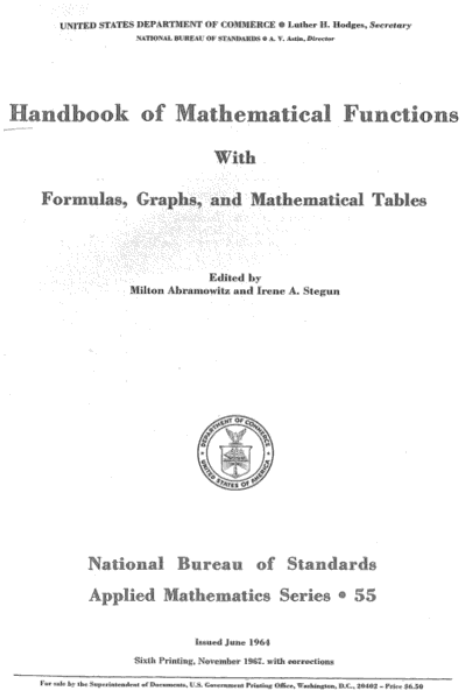
New institutions for mathematical research, both pure and applied, were created after the war. Among them the institute under Richard Courant at New York University developed the most strongly. In a 1954 government report it was stated that Courant's institute had an

> enrollment of over 400 graduate students in mathematics of which about half have a physics or engineering background.... The next largest figure, reported from Brown University, is a whole order of magnitude smaller! In the way of a rough estimate this means that New York University alone provides about one third of this country's annual output of applied mathematicians with graduate training.

The figures are based on a questionnaire prepared in connection with a conference organized in 1953 at Columbia University in New York by F. Joachim Weyl, the son of Hermann Weyl, as part of a Survey of Training and Research in Applied Mathematics sponsored by the American Mathematical Society and by the National Research Council under contract with the NSF. The conference proceedings and the report both included discussions on not only the training of applied mathematicians (particularly for industry, and including international comparisons) but also the increasing use of electronic computing; a summary was published in the *Bulletin of the American Mathematical Society* in 1954.

Between 1947 and 1954 the Institute for Numerical Analysis at the University of California, Los Angeles, sponsored by the National Bureau of Standards, played a special role in training university staff in numerical analysis and computer operations. The institute was closed in 1954, a victim of McCarthyism.

Brown University's summer school of applied mechanics, which were organized by Richardson from 1941 onward, had relied heavily on the contributions of immigrants. This is also partly true of the first American journal of applied mathematics, the *Quarterly of Applied Mathematics*, which began in 1943, and of *Mathematical Tables and Other Aids to Computation*, another Brown journal, which started the same year under Raymond Archibald.


UNITED STATES DEPARTMENT OF COMMERCE • Luther H. Hodges, *Secretary*
NATIONAL BUREAU OF STANDARDS • A. V. Astin, *Director*

Handbook of Mathematical Functions

With

Formulas, Graphs, and Mathematical Tables

Edited by
Milton Abramowitz and Irene A. Stegun

National Bureau of Standards
Applied Mathematics Series • 55

Issued June 1964
Sixth Printing, November 1967, with corrections

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 20402 - Price $6.50

Figure 6 M. Abramowitz and I. A. Stegun, eds, *Handbook of Mathematical Functions* (1964).

Various projects on mathematical tables and special functions that had their origins early in the twentieth century in various countries (the United Kingdom, Germany, and the United States) received a boost from the war. At Caltech, Arthur Erdélyi, with financial support from the Office of Naval Research, oversaw the Bateman Manuscript Project—the collation and publication of material collected by Harry Bateman, who had died in 1946—which led to the three-volume *Higher Transcendental Functions* (1953–55).

The Mathematical Tables Project, which had been initiated by the Works Progress Administration in New York in 1938, with Gertrude Blanch as its technical director, was disbanded after the war but many of its members moved to Washington in 1947 to become part of the new National Applied Mathematics Laboratories of the National Bureau of Standards. The latter's conference of 1952 resulted in one of the best-selling applied mathematics books of all time, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables* (1964) by Milton Abramowitz and Irene

Copyrighted material

the Abel Prize in Mathematics "for his groundbreaking contributions to the theory and application of partial differential equations and to the computation of their solutions"—signaled a new level of acceptance for applied mathematics in his widely distributed paper in *SIAM Review* on "The flowering of applied mathematics in America":

> Whereas in the not so distant past a mathematician asserting "applied mathematics is bad mathematics" or "the best applied mathematics is pure mathematics" could count on a measure of assent and applause, today a person making such statements would be regarded as ignorant.

The publication of articles by working applied mathematicians in Metropolis et al. (1980) and Nash (1990), and the extensive seven-volume historical project undertaken by the *Journal of Computational and Applied Mathematics* (2000), which was republished in Brezinski and Wuytack (2001), seem to testify to a growing self-confidence of applied mathematics within mathematics more widely. Some efforts, such as the publication of the "Top ten algorithms" in *Computing in Science and Engineering* in 2000 (six of which are contained in [I.5, table 2]), provoked controversial discussions. Many practitioners of applied mathematics in these and other publications reveal awareness of problems regarding the rigor and reliability of their methods, showing that the links between pure and applied mathematics exist and continue to stimulate the field. However, the standard philosophical approaches to mathematics—circling repetitively around formalism, logicism, and intuitionism, with no consideration of applications and doing no justice to the ever-increasing range of mathematical practice—are no longer satisfying either to mathematicians or to the public.

The unabated loyalty to pure mathematics as the mother discipline sometimes leads to overcautious reflection on the part of the applied mathematician. A nice example is provided by Trefethen (again in *The Princeton Companion to Mathematics*) in the context of rounding errors and the problem of numerical stability:

> These men, including von Neumann, Wilkinson, Forsythe, and Henrici, took great pains to publicize the risks of careless reliance on machine arithmetic. These risks are very real, but the message was communicated all too successfully, leading to the current widespread impression that the main business of numerical analysis is coping with rounding errors. In fact, the main business of numerical analysis is designing algorithms that converge quickly; rounding-error analysis, while often a part of the discussion, is rarely the central issue. If rounding errors vanished, 90% of numerical analysis would remain.

But it is not only these methodological concerns that hold back applied mathematics. For example, Lax (1989) mentions persisting problems in education and the need to maintain training in classical analysis:

> The applied point of view is essential for the much-needed reform of the undergraduate curriculum, especially its sorest spot, calculus. The teaching of calculus has been in the doldrums ever since research mathematicians gave up responsibility for undergraduate courses.

The education and training of applied mathematicians remains a central concern, and it is not even clear whether the situation has changed significantly since the Columbia University Conference of 1953. At that time the applied mathematician and statistician John Wilder Tukey, best known for the development of the FFT algorithm and the box plot, declared with reference to what is now called modeling:

> Formulation is the most important part of applied mathematics, yet no one has started to work on the theory of formulation—if we had one, perhaps we could teach applied mathematics.

A 1998 report by the NSF states that:

> Careers in mathematics have become less attractive to U.S. students. [Several]... factors contribute to this change: (i) students mistakenly believe that the only jobs available are collegiate teaching jobs, a job market which is saturated (more than 1,100 new Ph.D.s compete for approximately 600 academic tenure-track openings each year); (ii) academic training in the mathematical sciences tends to be narrow and to leave students poorly prepared for careers outside academia; (iii) neither students nor faculty understand the kinds of positions available outside academia to those trained in the mathematical sciences.

The same report underscores the undiminished dependence of American pure and applied mathematics on immigration from Europe and (now) from Asia, South America, and elsewhere:

> Although the United States is the strongest national community in the mathematical sciences, this strength is somewhat fragile. If one took into account only

Copyrighted material

> home-grown experts, the United States would be weaker than Western Europe. Interest by native-born Americans in the mathematical sciences has been steadily declining. Many of the strongest U.S. mathematicians were trained outside the United States and even more are not native born. A very large number of them emigrated from the former Soviet Union following its collapse. (Russia's strength in mathematics has been greatly weakened with the disappearance of research funding and the exodus of most of its leading mathematicians.) Western Europe is nearly as strong in mathematics as the United States, and leads in important areas. It has also benefited by the presence of émigré Soviet mathematical scientists.

The Fields Medals for Pierre-Louis Lions (son of Jacques-Louis Lions) (1994), Jean-Christophe Yoccoz (1994), Stanislav Smirnov (2010), and Cédric Villani (2010) testify to the growing strength of European applied mathematics and to the changed status of the field within mathematics. Likewise, the awarding of the Abel Prize of the Norwegian Academy of Science and Letters to Peter Lax (2005), Srinivasa Varadhan (2007), and Endre Szemerédi (2012) for predominantly applied topics is a further indication of this shift. In addition, prestigious prizes devoted specifically to applications, with particular emphasis on connections to technological developments, have been founded in recent decades. The fact that several of these prizes have been named for mathematicians of outstanding theoretical ability—the ACM A. M. Turing Award (starting in 1966), the IMU Rolf Nevanlinna Prize (1981), the DMV and IMU Carl Friedrich Gauss Prize (2006)—underscores the unity of mathematics in its pure and applied aspects.

Meanwhile, problems remain in the academic–industrial relationship and, connected to it, in the professional image of the applied mathematician, as described in the two most recent reports on "Mathematics in Industry" (1996 and 2012) published by SIAM. The report for 2012 summarizes the situation:

> Industrial mathematics is a specialty with a curious case of double invisibility. In the academic world, it is invisible because so few academic mathematicians actively engage in work on industrial problems. Research in industrial mathematics may not find its way into standard research journals, often because the companies where it is conducted do not want it to. (Some companies encourage publication and others do not; policies vary widely.) And advisors of graduates who go into industry may not keep track of them as closely as they keep track of their students who stay in academia.

However, most of the problems mentioned in this article with respect to academic applied mathematics (research funding, the lack of applications in mathematics education, the need for migration between national cultures) concern pure and applied mathematics alike. On the purely cognitive and theoretical level, the difference between the two aspects of mathematics—for all its interesting and important historical and sociological dimensions—hardly exists, as the above-quoted NSF report of 1998 underscores:

> Nowadays all mathematics is being applied, so the term applied mathematics should be viewed as a different cross cut of the discipline.

Further Reading

Aspray, W. 1990. *John von Neumann and the Origins of Modern Computing*. Cambridge, MA: MIT Press.

Aubin, D., and A. Dahan Dalmedico. 2002. Writing the history of dynamical systems and chaos: longue durée and revolution, disciplines and cultures. *Historia Mathematica* 29:273–339.

Babuška, I. 1994. Courant element: before and after. In *Finite Element Methods: Fifty Years of the Courant Element*, edited by M. Křížek, P. Neittaanmäki, and R. Stenberg, pp. 37–51. New York: Marcel Dekker.

Bennett, S. 1979/1993. *A History of Control Engineering: Volume 1, 1800–1930; Volume 2, 1930–1950*. London: Peter Peregrinus.

Birkhoff, G. 1977. Applied mathematics and its future. In *Science and Technology in America: An Assessment*, edited by R. M. Thomson, pp. 82–103. Gaithersburg, MD: National Bureau of Standards.

Booß-Bavnbek, B., and J. Høyrup, eds. 2003. *Mathematics and War*. Basel: Birkhäuser.

Brezinski, C., and L. Wuytack, eds. 2001. *Numerical Analysis: Historical Developments in the 20th Century*. Amsterdam: Elsevier.

Campbell-Kelly, M., M. Croarken, R. Flood, and E. Robson, eds. 2003. *The History of Mathematical Tables: From Sumer to Spreadsheets*. Oxford: Oxford University Press.

Chabert, J.-L. 1999. *A History of Algorithms: From the Pebble to the Microchip*. Berlin: Springer.

Cortada, J. W. 1990/1996. *A Bibliographic Guide to the History of Computing, Computers, and the Information Processing Industry*, two volumes. New York: Greenwood Press. (Available online at www.cbi.umn.edu/.)

Darrigol, O. 2005. *Worlds of Flow: A History of Hydrodynamics from the Bernoullis to Prandtl*. Oxford: Oxford University Press.

Goldstine, H. H. 1977. *A History of Numerical Analysis from the 16th through the 19th Century*. New York: Springer.

Copyrighted material

Grattan-Guinness, I., ed. 1994. *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences*, two volumes. London: Routledge.

Grötschel, M., ed. 2012. Optimization Stories (21st International Symposium on Mathematical Programming, Berlin, August 19–24, 2012). *Documenta Mathematica* (extra volume).

Lax, P. 1989. The flowering of applied mathematics in America. *SIAM Review* 31:533–41.

Lenstra, J. K., A. Rinnooy Kan, and A. Schrijver, eds. 1991. *History of Mathematical Programming: A Collection of Personal Reminiscences*. Amsterdam: North-Holland.

Lucertini, M., A. Millán Gasca, and F. Nicolò, eds. 2004. *Technological Concepts and Mathematical Methods in the Evolution of Modern Engineering Systems: Controlling, Managing, Organizing*. Basel: Birkhäuser.

Mancosu, P., ed. 2008. *The Philosophy of Mathematical Practice*. Oxford: Oxford University Press.

Mehrtens, H., H. Bos, and I. Schneider, eds. 1981. *Social History of Nineteenth Century Mathematics*. Boston, MA: Birkhäuser.

Metropolis, N., J. Howlett, and G.-C. Rota. 1980. *A History of Computing in the Twentieth Century*. New York: Academic Press.

Millman, S., ed. 1984. *A History of Engineering and Science in the Bell System: Communication Sciences (1925–1980)*. Murray Hill, NJ: AT&T Bell Laboratories.

Nash, S. G., ed. 1990. *A History of Scientific Computing*. New York: ACM Press.

Tournès, D. 1998. L'origine des méthodes multipas pour l'intégration numérique des équations différentielles ordinaires. *Revue d'Histoire des Mathématiques* 4:5–72.

Truesdell, C. 1960. A program toward rediscovering the rational mechanics of the age of reason. *Archive for History of Exact Sciences* 1:3–36.

Copyrighted material

Copyrighted material

Part II
Concepts

II.1 Asymptotics

P. A. Martin

When sketching the graph of a function, $y = f(x)$, we may notice (or look for) lines that the graph approaches, often as $x \to \pm\infty$. For example, the graph of $y = x^2/(x^2+1)$ approaches the straight line $y = 1$ as $x \to \infty$ (and as $x \to -\infty$). This line is called an *asymptote*. Asymptotes need not be horizontal or straight, and they may be approached as $x \to x_0$ for some finite x_0. For example, $y = x^4/(x^2+1)$ approaches the parabola $y = x^2$ as $x \to \pm\infty$, and $y = \log x$ approaches the vertical line $x = 0$ as $x \to 0$ through positive values. Another example is that $\sinh x = \frac{1}{2}(e^x + e^{-x})$ approaches $\frac{1}{2}e^x$ as $x \to \infty$: we say that $\sinh x$ grows exponentially with x.

The qualitative notions exemplified above can be made much more quantitative. One feature that we want to retain when we say something like "$y = f(x)$ approaches $y = g(x)$ as $x \to \infty$" is that, to be useful, $g(x)$ should be simpler than $f(x)$, where "simpler" will depend on the context. This is a familiar idea; for example, we can approximate a smooth curve near a chosen point on the curve by the tangent line through that point.

When $\lim_{x\to x_0}[f(x)/g(x)] = 1$, we write $f(x) \sim g(x)$ as $x \to x_0$, and we say that $g(x)$ is an *asymptotic approximation* to $f(x)$ as $x \to x_0$. For example, $\sinh x \sim x$ as $x \to 0$ and $\tanh x \sim 1$ as $x \to \infty$. A famous asymptotic approximation of this kind is Stirling's formula from 1730: $n! \sim (2\pi n)^{1/2}(n/e)^n$ as $n \to \infty$.

According to our definition, we have $e^x \sim 1$, $e^x \sim 1 + x$, and $e^x \sim 1 + 2x$, all as $x \to 0$. On the other hand, we have the Maclaurin expansion, $e^x = 1 + x + \frac{1}{2}x^2 + \cdots$, which converges for all x; truncating this infinite series gives good approximations to e^x near $x = 0$, and these approximations improve if we take more terms in the series. This suggests that we should select $1+x$ and not $1 + 2x$, so our definition of "$\sim$" is too crude. We want asymptotic approximations to be approximations, and we want to be able to improve them by taking more terms, if possible. With this in mind, suppose we have a sequence of functions, $\phi_n(x)$, $n = 0, 1, 2, \ldots$, with the property that $\phi_{n+1}(x)/\phi_n(x) \to 0$ as $x \to x_0$. Standard examples are $\phi_n(x) = x^n$ as $x \to 0$ and $\phi_n(x) = x^{-n}$ as $x \to \infty$. Let $R_N(x) = \sum_{n=0}^{N} a_n\phi_n(x)$ for some coefficients a_n. We write

$$f(x) \sim \sum_{n=0}^{\infty} a_n\phi_n(x) \quad \text{as } x \to x_0,$$

and say that the series is an *asymptotic expansion* of $f(x)$ as $x \to x_0$ when, for each $N = 0, 1, 2, \ldots$,

$$[f(x) - R_N(x)]/\phi_N(x) \to 0 \quad \text{as } x \to x_0. \qquad (1)$$

In words, the "error" $f - R_N$ is comparable to the first term omitted, the one with $n = N + 1$. Note that the definition does not require the infinite series to be convergent (so that $R_N(x)$ may not have a limit as $N \to \infty$ for fixed x). Instead, for each fixed N, we impose a requirement on the error as $x \to x_0$, namely (1).

Asymptotic approximations may be convergent. For example, we have $e^x \sim 1 + x + \frac{1}{2}x^2 + \cdots$ as $x \to 0$. However, many interesting and useful asymptotic expansions are divergent. As an example, the *complementary error function*

$$\begin{aligned}\operatorname{erfc}(x) &= \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,dt \\ &\sim \frac{e^{-x^2}}{x\sqrt{\pi}}\left[1 + \sum_{n=1}^{\infty}(-1)^n\frac{1\cdot 3\cdots(2n-1)}{(2x^2)^n}\right]\end{aligned}$$

as $x \to \infty$, where the series is obtained by repeated integration by parts of the defining integral. The series is divergent, but taking a few terms gives a good approximation to $\operatorname{erfc} x$, an approximation that improves as x becomes larger.

Many techniques have been devised for obtaining asymptotic expansions. Some are designed for functions defined by integrals (such as $\operatorname{erfc} x$), others for functions that solve differential equations. Asymptotic methods can also be used to estimate the complexity

Copyrighted material

It is often hard to do much more than numerical simulations to determine the behavior of systems, leading to a somewhat limited phenomenological description of their behavior. There are two central methods in the study of complex systems that go further than this, though again with limited concrete predictive power. These are graph theory to characterize how components influence each other, and dimension reduction methods to capture (where applicable) any lower-dimensional approximations that determine the evolution of the system.

If the system has real variables x_i, $i = 1, \ldots, N$, then each variable can be identified with the node of a graph labeled by i, with an edge from i to j if the dynamics of x_j is directly influenced by x_i (see GRAPH THEORY [II.16]). For example, if the evolution is determined by a differential equation then $\dot{x}_i = f_i(x_1, x_2, \ldots, x_N)$, but not every variable need appear explicitly in the argument of f_i, so there is an edge from x_j to x_i only if $\partial f_i / \partial x_j$ is not identically zero. This graph can be represented by an adjacency matrix (a_{ij}) with $a_{ij} = 1$ if there is an edge from i to j and $a_{ij} = 0$ otherwise. The degree of a node is the number of edges at the node (this can be split into the in-degree (respectively, out-degree) if only edges ending (respectively, starting) at the node are counted). The proportion of nodes with degree k is the degree distribution of the network. Properties of the degree distribution are often used to characterize the network. For example, if the degree distribution obeys a power law, the network is said to be scale free (the Internet is supposedly of this type; see NETWORK ANALYSIS [IV.18]).

By analyzing subgraphs of biological models it was found that some subgraphs appear in examples much more often than would be expected on the basis of a statistical analysis. This has led to the conjecture that these *motifs* may have associated functional properties.

In many complex systems the individual components of the system behave according to very simple, though often nonlinear, rules. For example, a bird in a flock may change its direction of flight as a function of the average direction of flight of nearby birds. Although this is a local rule, the effect across the entire flock of birds is to produce coherent movement of the flock as a whole. This effect, whereby simple local rules lead to interesting global results, is called *emergent behavior*. The emergent behavior resulting from given local rules is often unclear until the system is simulated numerically.

In some cases the dimension of the problem can be reduced, so fewer variables need to be considered, making the system easier to simulate and more amenable to analysis. The methods of dimension reduction often rely on SINGULAR VALUE DECOMPOSITION [II.32] techniques to identify the more dynamically active directions in phase space, and then an attempt is made to project the system onto these directions and analyze the resulting system.

In some systems the mean-field theory of theoretical physics can be used to understand collective behavior.

Since complexity theory encompasses so many different models, the range of possible dynamic phenomena is vast, even before further complications such as stochastic effects or network evolution are included. Complex systems describing neuron interactions in the brain can model pattern recognition and memory (see MATHEMATICAL NEUROSCIENCE [VII.21]). Numerical models of partial differential equations are complex systems, and the dynamical behavior can include synchronization, in which all components lock on to a similar pattern of behavior, and PATTERN FORMATION [IV.27]. Different parts of the system may behave in dynamically different ways, with regions of frustration (or fronts) separating them. Interactions may have different strengths, leading to different timescales in the problem. This is particularly true of many biological models and adds to the difficulty of modeling phenomena accurately.

Further Reading

Ball, R., V. Kolokoltsov, and R. S. MacKay, eds. 2013. *Complexity Science*. Cambridge: Cambridge University Press.

Estrada, E. 2011. *The Structure of Complex Networks: Theory and Applications*. Oxford: Oxford University Press.

Watts, D. J. 1999. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton, NJ: Princeton University Press.

II.5 Conformal Mapping

Darren Crowdy

1 What Is a Conformal Mapping?

Conformal mapping is the name given to the idea of interpreting an analytic function of a complex variable in a geometric fashion. Let $z = x + \mathrm{i}y$ and suppose that another complex variable w is defined by

$$w = f(z) = \phi(x, y) + \mathrm{i}\psi(x, y),$$

where ϕ and ψ are, respectively, the real and imaginary parts of some function $f(z)$, an analytic function

Copyrighted material

of z. One can think of this relation as assigning a correspondence between points in the complex z-plane and points in the complex w-plane. Under this function a designated region of the z-plane is transplanted, or "mapped," to some region in the w-plane, as illustrated in figure 1. The shape of the image will depend on f. The fact that f is an *analytic* function implies certain special properties of this mapping of regions. If the mapping is to be one-to-one, then a necessary, but not sufficient, condition is that the derivative $f'(z) = \mathrm{d}f/\mathrm{d}z$ does not vanish in the z-region of interest.

A simple example is the *Cayley mapping*

$$w = f(z) = \frac{1+z}{1-z}.$$

This maps the interior of the unit disk $|z| < 1$ in the z-plane to the right half w-plane $\operatorname{Re} w > 0$. The point $z = 1$ maps to $w = \infty$, and $z = -1$ maps to $w = 0$. The unit circle $|z| = 1$ maps to the imaginary w-axis. Conformal mappings clearly preserve neither area nor perimeters; their principal geometrical feature is that they locally preserve angles. To see this, note that since $f(z)$ is analytic at a point z_0, it has a local Taylor expansion there:

$$w = f(z_0) + f'(z_0)(z - z_0) + \cdots.$$

If $\delta z = z - z_0$ is an infinitesimal line element through z_0 in the z-plane, its image δw under the mapping defined as $\delta w = w - w_0$, where $w_0 = f(z_0)$, is, to leading order,

$$\delta w \approx f'(z_0)\delta z.$$

But $f'(z_0)$ is just a nonzero complex number so, under a conformal mapping, all infinitesimal line elements through z_0 are transplanted to line elements through w_0 in the w-plane that are simply rescaled by the modulus of $f'(z_0)$ and rotated by its argument. In particular, the angle between two given line elements through z_0 is preserved by the mapping.

2 The Riemann Mapping Theorem

The Riemann mapping theorem is considered by many to be the pinnacle of achievement of nineteenth-century mathematics. It is an existence theorem: it states that there exists a conformal mapping from the unit z-disk to *any* given simply connected region (no holes) in the w-plane, so long as it is not the entire plane.

3 Conformal Invariance

One reason why conformal mappings are an important tool in applied mathematics is the property of *conformal invariance* of certain boundary-value problems

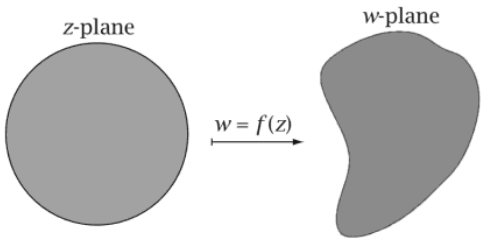


Figure 1 A conformal mapping from a region in a complex z-plane to a region in a complex w-plane.

arising in applications. An example is the boundary-value problem determining Green's function $G(z; z_0)$ for the Laplace equation in a region D in $\mathbb{R}^2$ with boundary ∂D, which can be written as

$$\nabla^2 G = \delta^{(2)}(z - z_0) \quad \text{in } D \text{ with } G = 0 \text{ on } \partial D,$$

where z_0 is some point inside D and $\delta^{(2)}$ is the two-dimensional DIRAC DELTA FUNCTION [III.7]. The Green function for the unit disk $|z| < 1$ is known to be

$$G(z; z_0) = \operatorname{Im}\left[\frac{\mathrm{i}}{2\pi}\log\left(\frac{z - z_0}{|z_0|(z - 1/\overline{z_0})}\right)\right],$$

where $\overline{z_0}$ is the complex conjugate of z_0. Now if D is any other simply connected region of a complex w-plane, the corresponding Green function in D is nothing other than $G(f^{-1}(w); f^{-1}(w_0))$, where $f^{-1}(w)$ is the inverse function of the conformal mapping taking the unit z-disk to D. Geometrically, $f^{-1}(w)$ is just the inverse conformal mapping transplanting D to the unit disk $|z| < 1$. The Green function in any simply connected region D is therefore known immediately provided the conformal mapping between D and the unit disk can be found.

4 Schwarz–Christoffel Mappings

The Riemann mapping theorem is nonconstructive and, while the existence of a conformal mapping between given simply connected regions is guaranteed, the practical matter of actually constructing it is another story. One of the few general constructions often used in applications is the *Schwarz–Christoffel mapping*. This is a conformal mapping from a standard region such as the unit z-disk $|z| < 1$ to the region interior or exterior to an N-sided polygon. At the preimage of any vertex of the polygon (a *prevertex*), the local argument outlined earlier demonstrating the preservation of angles between infinitesimal line elements must fail. Indeed, at any such prevertex it can be argued that the derivative $f'(z)$ of the conformal mapping must have a simple

Copyrighted material

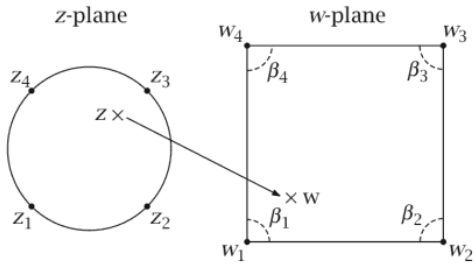


Figure 2 A Schwarz–Christoffel mapping from the unit z-disk to the interior of a square in a w-plane. A function of the form (1) with $N = 4$ identifies a point w with a point z. Here, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \pi/2$.

zero, a simple pole, or a branch point singularity. The general formula for a mapping from $|z| < 1$ to the interior of a bounded polygon in a w-plane is

$$w = f(z) = A + B \int^z \prod_{k=1}^{N} \left(1 - \frac{z'}{z_k}\right)^{(\beta_k/\pi - 1)} \mathrm{d}z', \quad (1)$$

while the formula for a mapping from $|z| < 1$ to the exterior of a bounded polygon in a w-plane, with $z = 0$ mapping to $w = \infty$, is

$$w = f(z) = A + B \int^z \prod_{k=1}^{N} \left(1 - \frac{z'}{z_k}\right)^{(\beta_k/\pi - 1)} \frac{\mathrm{d}z'}{z'^2}. \quad (2)$$

The parameters $\{\beta_k \mid k = 1, 2, \ldots, N\}$ are the *turning angles* shown in figure 2; the points $\{z_k \mid k = 1, 2, \ldots, N\}$ are the prevertices. A and B are complex constants. These so-called *accessory parameters* are usually computed numerically by fixing geometrical features such as ensuring that the sides of the polygon have the required length. A famous mapping of Schwarz–Christoffel type known for its use in aerodynamics is the *Joukowski mapping*,

$$w = f(z) = \frac{1}{2}\left(z + \frac{1}{z}\right),$$

which maps the unit disk $|z| < 1$ to the infinite region exterior to a flat plate, or airfoil, lying on the real w-axis between $w = -1$ and $w = 1$. It is a simple matter to derive it from (2) with the prevertices $z_1 = 1$, $z_2 = -1$ and turning angles $\beta_1 = \beta_2 = 2\pi$. Since it is natural, given any two-dimensional shape, to approximate it by taking a set of points on the boundary and joining them with straight line segments to form a polygon, the Schwarz–Christoffel formula has found many uses in applied mathematics. Versatile numerical software to compute the accessory parameters has also been developed.

Further Reading

Courant, R. 1950. *Dirichlet's Principle, Conformal Mapping and Minimal Surfaces.* New York: Interscience.
Driscoll, T. A., and L. N. Trefethen. 2002. *Schwarz–Christoffel Mapping.* Cambridge: Cambridge University Press.
Nehari, Z. 1975. *Conformal Mapping.* New York: Dover.

II.6 Conservation Laws

Barbara Lee Keyfitz

1 Quasilinear Hyperbolic Partial Differential Equations

A system of first-order partial differential equations (PDEs) in the form

$$\boldsymbol{u}_t + \sum_{i=1}^{d} A_i(x, t, \boldsymbol{u})\boldsymbol{u}_{x_i} + \boldsymbol{b}(x, t, \boldsymbol{u}) = 0, \quad (1)$$

where $\boldsymbol{u} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^n$, the A_i are $n \times n$ matrices, and $\boldsymbol{u}_t \equiv \partial \boldsymbol{u}/\partial t$ and $\boldsymbol{u}_{x_i} \equiv \partial \boldsymbol{u}/\partial x_i$, is said to be *quasilinear*; the system is nonlinear as defined in the article PARTIAL DIFFERENTIAL EQUATIONS [IV.3], but the terms containing derivatives of $\boldsymbol{u}$ appear only in linear combination. Identifying t as a time variable and $x = (x_1, \ldots, x_d)$ as a space variable, the *Cauchy problem* asks for a solution to (1) for $t > 0$ with the initial condition

$$\boldsymbol{u}(x, 0) = \boldsymbol{u}_0(x). \quad (2)$$

By analogy with the theory of linear PDEs, one expects this problem to be well-posed only if the system is *hyperbolic*, which means that all the roots $\tau(\xi)$ (known as *characteristics*) of the polynomial equation

$$\det\left(\tau I + \sum_{i=1}^{d} A_i \xi_i\right) = 0 \quad (3)$$

are real for all $\xi \in \mathbb{R}^d$ and, as eigenvalues of the matrix $\sum_{i=1}^{d} A_i \xi_i$, each has equal ALGEBRAIC AND GEOMETRIC MULTIPLICITIES [II.22].

In 1974 Fritz John showed that if $d = 1$, and the system is *genuinely nonlinear* (meaning that $\nabla_\xi \tau_i \cdot \boldsymbol{r}_i \neq 0$ for each root τ_i of (3) and corresponding eigenvector $\boldsymbol{r}_i$), then for smooth Cauchy data at least one component of $\nabla \boldsymbol{u}$ tends to infinity in finite time, exactly as in the BURGERS EQUATION [III.4] (see also PARTIAL DIFFERENTIAL EQUATIONS [IV.3 §3.6]).

Characteristics in hyperbolic systems define the speed of propagation of signals in specific directions (normal to ξ), so genuine nonlinearity says that this speed is a nontrivial function of the state $\boldsymbol{u}$. This has physical significance as a description of the phenomena

Copyrighted material

modeled by conservation laws, and it has mathematical implications for the existence of smooth solutions. Specifically, the behavior seen in solutions of the Burgers equation typifies solutions of genuinely nonlinear hyperbolic systems.

Furthermore, despite the fact that DISTRIBUTION SOLUTIONS [IV.3 §5.2] are well defined for linear hyperbolic equations, the concept fails for quasilinear systems since, in the first place, A_i and $A_i \boldsymbol{u}_{x_i}$ are not defined if $\boldsymbol{u}$ lacks sufficient smoothness, and, in the second, the standard procedure of creating the weak form of an equation (multiply by a smooth test function and integrate by parts) does not usually succeed in eliminating $\nabla \boldsymbol{u}$ from the system when $A = A(\boldsymbol{u})$ depends in a nontrivial way on $\boldsymbol{u}$. The exception is when each $A_i \boldsymbol{u}_{x_i}$ is itself a derivative: $A_i \boldsymbol{u}_{x_i} = \partial_{x_i} f_i(\boldsymbol{u})$. This happens if each row of each A_i is a gradient, and that happens only if the requisite mixed partial derivatives are equal. In this case, we have a *system of balance laws*:

$$\boldsymbol{u}_t + \sum_{i=1}^{d} (f_i(x,t,\boldsymbol{u}))_{x_i} + \boldsymbol{b}(x,t,\boldsymbol{u}) = 0. \tag{4}$$

In the important case in which $\boldsymbol{b} \equiv 0$, we have a system of *conservation laws*. The weak form of (4) is

$$\iint \left[\boldsymbol{u}\varphi_t + \sum_{i=1}^{d} (f_i(x,t,\boldsymbol{u}))\varphi_{x_i} - \boldsymbol{b}(x,t,\boldsymbol{u})\varphi \right] \mathrm{d}x\, \mathrm{d}t = 0. \tag{5}$$

Since this is the only case in which solutions to (1) can be unambiguously defined, the subject of quasilinear hyperbolic systems is often referred to as "conservation laws."

A mathematical challenge in conservation laws is to find spaces of functions that are inclusive enough to admit weak solutions for general classes of conservation laws but regular enough that solutions and their approximations can be analyzed. At this time, there is a satisfactory well-posedness theory only in a single space dimension.

2 How Conservation Laws Arise

Problems of importance in physics, engineering, and technology lead to systems of conservation laws; a sample selection of these problems follows.

2.1 Compressible Flow

The basic equations of compressible fluid flow, derived from the principles of conservation of mass, momentum, and energy, along with constitutive equations relating thermodynamic quantities, take the form

$$\left.\begin{aligned} \rho_t + \operatorname{div}(\rho \boldsymbol{u}) &= 0, \\ (\rho \boldsymbol{u})_t + \operatorname{div}(\rho \boldsymbol{u} \otimes \boldsymbol{u}) + \nabla p &= 0, \\ (\rho E)_t + \operatorname{div}(\rho \boldsymbol{u} H) &= 0, \end{aligned}\right\} \tag{6}$$

where ρ represents density, $\boldsymbol{u}$ velocity, p pressure, E energy, and H enthalpy, with

$$E = \tfrac{1}{2}|\boldsymbol{u}|^2 + \frac{1}{\gamma - 1}\frac{p}{\rho}, \qquad H = \gamma E,$$

and γ a constant that depends on the fluid ($\gamma = 1.4$ for air). To obtain the first equation in (6), one notes that the total amount of mass in an arbitrary control volume D is the integral over D of the density, and this changes in time if there is flux through the boundary Γ of D. Furthermore, the flux is precisely the product of the density and the velocity normal to that boundary, from which we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \iint_D \rho \, \mathrm{d}V = -\int_\Gamma \rho \boldsymbol{u} \cdot \nu \, \mathrm{d}A. \tag{7}$$

(The negative sign will remind the reader of the convention that ν is the outward normal, and flow out of D will decrease the mass contained in D.) Interchanging differentiation and integration on the left in (7), along with an application of the DIVERGENCE THEOREM [I.2 §24] on the right, immediately yields

$$\iint_D (\rho_t + \operatorname{div}(\rho \boldsymbol{u}))\, \mathrm{d}V = 0. \tag{8}$$

Finally, the observation that D is an arbitrary domain in the region allows one to pass to the infinitesimal version in (6). The integral version (8) also justifies the weak form (5), since if (8) holds on arbitrary domains then it is possible to form weighted averages with arbitrary differentiable functions φ and to integrate by parts, which produces (5).

In compressible flow, the speed of sound is finite; in (6) it is one of the characteristics. Steady flow at speeds that exceed the speed of sound also gives a hyperbolic system of conservation laws (6) with the time derivatives absent. In this case, the hyperbolic direction (the time-like variable) is given by the flow direction.

Conservation principles also lead to equations for ELASTICITY [IV.26 §3.3] and MAGNETOHYDRODYNAMICS [IV.29]. Industrial applications include continuum models for multiphase flow (e.g., water mixed with steam in nuclear reactor cooling systems, or multicomponent flows in oil reservoirs).

The necessity of solving, or at least approximating, conservation laws for many of these applications has

Copyrighted material

resulted in extensive techniques for numerical simulation of solutions, even when existence of solutions remains an open question.

2.2 Chromatography

Chromatography is a widely used industrial process for separating chemical components of a mixture by differential adsorption on a substrate. Modeling a chromatographic column leads to a system of conservation laws in a single space variable that takes the form

$$\boldsymbol{c}_x + (\boldsymbol{f}(\boldsymbol{c}))_t = 0,$$

where $\boldsymbol{c} = (c_1, \ldots, c_n)$ is a vector of component concentrations and $\boldsymbol{f}$ is the equilibrium column isotherm. A common model for $\boldsymbol{f}$ uses the Langmuir isotherm and gives, with positive parameters α_i measuring the relative adsorption rates,

$$f_i = c_i + \frac{\alpha_i c_i}{1 + \sum c_j}, \quad 1 \leqslant i \leqslant n.$$

2.3 Other Models

Many other physical phenomena lead naturally to conservation laws. For example, a continuum model for vehicular traffic on a one-way road is the scalar equation

$$u_t + q(u)_x = 0,$$

where u represents the linear density of traffic and $q(u) = uv(u)$ the flux, where v is velocity. As in (7), this equation is a conservation law, the "law of conservation of cars." This model assumes that the velocity at which traffic moves depends only on the traffic density. Although this model is too simple to be of much practical use, it is appealing as a pedagogical tool. Adaptations of it are of interest in current research.

Further Reading

Bellomo, N., and C. Dogbe. 2011. On the modeling of traffic and crowds: a survey of models, speculations, and perspectives. *SIAM Review* 53:409–63.

Courant, R., and K. O. Friedrichs. 1948. *Supersonic Flow and Shock Waves*. New York: Wiley-Interscience.

Dafermos, C. M. 2000. *Hyperbolic Conservation Laws in Continuum Physics*. Berlin: Springer.

Godlewski, E., and P.-A. Raviart. 1996. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. New York: Springer.

John, F. 1974. Formation of singularities in one-dimensional nonlinear wave propagation. *Communications on Pure and Applied Mathematics* 27:377–405.

Rhee, H.-K., R. Aris, and N. R. Amundson. 1989. *First-Order Partial Differential Equations: Volume II. Theory and Application of Hyperbolic Systems and Quasilinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Whitham, G. B. 1974. *Linear and Nonlinear Waves*. New York: Wiley-Interscience.

II.7 Control

A *system* is a collection of objects that interact and produce various outputs in response to different inputs. Systems arise in a wide variety of situations and include chemical plants, cars, the human body, and a country's economy. Control problems associated with these systems include the production of a chemical, control of self-driving cars, the regulation of bodily functions such as temperature and heartbeat, and the control of debt. In each case one wants to have a way of controlling these processes automatically without direct human intervention.

A general control system is depicted in figure 1. The state of the system is described by n *state variables* x_i, and these span the *state space*. In general, the x_i cannot be observed or measured directly, but p *output variables* y_i, which depend on the x_i, are known. The system is controlled by manipulating m *control variables* u_i.

The system might be expressed as a system of difference equations (discrete time) or differential equations (continuous time). In the latter case a linear, time-invariant control problem takes the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = Ax(t) + Bu(t),$$
$$y(t) = Cx(t) + Du(t),$$

where A, B, C, and D are $n \times n$, $n \times m$, $p \times n$, and $p \times m$ matrices, respectively. This is known as a *state-space system*. In some cases an additional $n \times n$ matrix E, which is usually singular, premultiplies the $\mathrm{d}x/\mathrm{d}t$ term; these so-called *descriptor systems* or *generalized state-space systems* lead to DIFFERENTIAL-ALGEBRAIC EQUATIONS [I.2 §12].

A natural question is whether, given a starting value $x(0)$, the input u can be chosen so that x takes a given value at time t. Questions of this form are fundamental in classical control theory.

If feedback occurs from the outputs or state variables to the controller, then the system is called a *closed-loop system*. In *output feedback*, illustrated in figure 1, u depends on y, while in *state feedback* u depends on x.

Copyrighted material

physical quantities defining a model. In order to explain how this works, we need to introduce some definitions and establish the notation that will be used below. Consider a system that contains n physical quantities, $q_1, q_2, \ldots, q_n$, that we believe to be relevant for describing the system's behavior, the quantities being expressed using r fundamental units, or dimensions, denoted by $d_1, d_2, \ldots, d_r$. The generally accepted SI unit system consists of $r = 7$ basic dimensions and numerous derived dimensions. More precisely, d_1 is the meter (m), d_2 the second (s), d_3 the kilogram (kg), d_4 the ampere (A), d_5 the mole (mol), d_6 the kelvin (K), and d_7 the candela (cd). The number r can be smaller, since not all units are always needed. The physical dimension of a quantity q is denoted by $[q]$.

A meaningful mathematical relation between the quantities q_j should obey the *principle of dimensional homogeneity*, which can be summarized as follows: *summing up quantities is meaningful only if all the terms have the same dimension.* Furthermore, any functional relation of the type

$$f(q_1, q_2, \ldots, q_n) = 0 \tag{1}$$

should remain valid if expressed in different units. In other words, since *dimensional scaling must not change the equation*, it is natural to seek to express the relations in terms of dimensionless quantities. It is therefore not a surprise that dimensionless quantities, known as *Π-numbers*, have a central role in dimensional analysis. A canonical example of a Π-number is π, the invariant ratio of the circumference and the diameter of circles of all sizes.

Given a system described by the physical quantities $q_1, q_2, \ldots, q_n$, we will define a Π-number, or a *dimensionless group*, to be any combination of those quantities of the form

$$R = q_1^{\mu_1} q_2^{\mu_2} \cdots q_n^{\mu_n}, \tag{2}$$

where the μ_j are *rational* numbers, not all equal to zero, and R is dimensionless. If, in such a system, we are able to identify k Π-numbers, $R_1, \ldots, R_k$, that characterize it, we can describe it with a dimensionless version of (1) of the form

$$\varphi(R_1, R_2, \ldots, R_k) = 0.$$

The advantage of the latter formulation is that it automatically satisfies the dimensional homogeneity; moreover, it does not change with any scaling of the model that leaves the values of the Π-numbers invariant. These points are best clarified by a classical example of dimensional analysis.

Consider steady fluid flow in a pipe of constant diameter D. The fluid is assumed to be incompressible, having density ρ and viscosity μ. By denoting the pressure drop across a distance L by Δp and the (average) velocity by v, we may assume that there is an algebraic relation between the quantities:

$$f(L, D, \rho, \mu, v, \Delta p) = 0. \tag{3}$$

In SI units, the dimensions of the variables involved are

$$\left.\begin{aligned} [L] = [D] = \mathrm{m}, \qquad [\rho] = \frac{\mathrm{kg}}{\mathrm{m}^3}, \qquad [\mu] = \frac{\mathrm{kg}}{\mathrm{m}\cdot\mathrm{s}}, \\ [v] = \frac{\mathrm{m}}{\mathrm{s}}, \qquad [\Delta p] = \frac{\mathrm{kg}}{\mathrm{m}\cdot\mathrm{s}^2}. \end{aligned}\right\} \tag{4}$$

In his classic paper of 1883, Osborne Reynolds suggested a scaling law of the form

$$\Delta p = \rho v^2 \frac{L}{D} F\left(\frac{\rho v D}{\mu}\right), \tag{5}$$

where F is some function; Reynolds himself considered the power law $F(R) = cR^{-n}$ with different values of n and experimentally validated it. Equation (5) can be seen as a dimensionless version of (3),

$$\varphi(R_1, R_2, R_3) = 0,$$

where

$$R_1 = \frac{\rho v D}{\mu}, \qquad R_2 = \frac{\Delta p}{\rho v^2}, \qquad R_3 = \frac{L}{D}.$$

The quantities R_1 and R_2 are known as the *Reynolds number* and the *Euler number*, respectively, and it is a straightforward matter to check that R_1, R_2, and R_3 are dimensionless. The scaling law (5) has been experimentally validated in a range of geometric settings. An example of its use is the design of miniature models. If the dimensions are scaled by a factor α, $L \to \alpha L$, $D \to \alpha D$, we may assume that the flow in the miniature model gives a good prediction for the actual system if we scale the velocity and pressure as $v \to v/\alpha$ and $\Delta p \to \Delta p/\alpha^2$, leaving the dimensionless quantities intact.

In view of the above example it is natural to ask how many Π-numbers characterize a given system and if there is a systematic way of finding them. To address these questions it is important to identify possible redundancy among the physical quantities, on the one hand, and the dimensions, on the other. With this in mind we introduce the concepts of *independency* and *relevance* of the dimensions.

The dimensions $d_1, \ldots, d_r$ are *independent* if none can be expressed as a rational product of the others, that is,

$$d_1^{\alpha_1} d_2^{\alpha_2} \cdots d_r^{\alpha_r} = 1 \tag{6}$$

Copyrighted material

if and only if $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$. The dimensions d_j may be the fundamental dimensions of the SI unit system or derived dimensions, such as the newton ($\mathrm{N} = \mathrm{kg/m \cdot s^2}$).

It is not a coincidence that this definition strongly resembles that of linear independency in linear algebra, as will become evident later.

Let a system be described by n quantities, $q_1, \ldots, q_n$, and r dimensions, $d_1, \ldots, d_r$, with the dimensional dependency

$$[q_j] = d_1^{\mu_{j1}} d_2^{\mu_{j2}} \cdots d_r^{\mu_{jr}}, \quad 1 \leqslant j \leqslant n. \tag{7}$$

We say that the dimensions d_k, $1 \leqslant k \leqslant r$, are *relevant* if for each d_k there are rational coefficients α_{kj} such that

$$d_k = [q_1]^{\alpha_{k1}} [q_2]^{\alpha_{k2}} \cdots [q_n]^{\alpha_{kn}}. \tag{8}$$

In other words, the dimensions d_k are relevant if they can be expressed in terms of the dimensions of the variables q_k. It follows immediately that, if the quantities q_j can be measured, then there must exist an operational description of all units in terms of the measurements. Identifying relevant quantities may be more subtle than it seems.

For the sake of definiteness, assume that we adhere to the SI system, and denote the seven basic SI units by $e_1, e_2, \ldots, e_7$, the ordering being unimportant. We now proceed to define an associated dimension space: to each e_i we associate a vector $\boldsymbol{e}_i \in \mathbb{R}^7$, where $\boldsymbol{e}_i$ is the ith unit coordinate vector. Further, we define a group homomorphism between the $\mathbb{Q}$-moduli of dimensions and vectors; since any dimension d can be represented in the SI system in terms of the seven basic units e_i as

$$d = e_1^{\nu_1} \cdots e_7^{\nu_7},$$

we associate d with a vector $\boldsymbol{d}$, where

$$\boldsymbol{d} = \nu_1 \boldsymbol{e}_1 + \cdots + \nu_7 \boldsymbol{e}_7.$$

Along these lines, we associate with a quantity q with dimensions

$$[q] = d_1^{\mu_1} \cdots d_r^{\mu_r}$$

the vector

$$\boldsymbol{q} = \mu_1 \boldsymbol{d}_1 + \cdots + \mu_r \boldsymbol{d}_r.$$

It is straightforward to verify that the representation of $\boldsymbol{q}$ in terms of the basis vectors $\boldsymbol{e}_j$ is unambiguous.

We are now ready to revisit independency of units in the light of the associated vectors. In linear algebraic terms, condition (6) is equivalent to saying that

$$\alpha_1 \boldsymbol{d}_1 + \cdots + \alpha_r \boldsymbol{d}_r = \boldsymbol{0},$$

and therefore *the independency of dimensions is equivalent to the linear independency of the corresponding dimension vectors*.

Next we look for a connection with linear algebra to help us reinterpret the concept of relevance. In the dimension space, condition (7) can be expressed as

$$\boldsymbol{q}_j = \mu_{j1} \boldsymbol{d}_1 + \cdots + \mu_{jr} \boldsymbol{d}_r = \sum_{k=1}^{r} \mu_{jk} \boldsymbol{d}_k,$$

which implies that every $\boldsymbol{q}_j$ is in the subspace spanned by the vectors $\boldsymbol{d}_k$, while the linear algebraic formulation of condition (8),

$$\boldsymbol{d}_k = \alpha_{k1} \boldsymbol{q}_1 + \cdots + \alpha_{kn} \boldsymbol{q}_n = \sum_{\ell=1}^{n} \alpha_{k\ell} \boldsymbol{q}_\ell,$$

states that the vectors $\boldsymbol{d}_k$ are in the subspace spanned by the vectors $\boldsymbol{q}_\ell$. We therefore conclude that the relevance of dimensions is equivalent to the condition that

$$\operatorname{span}\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\} = \operatorname{span}\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_r\}.$$

It is obvious that when $n > r$, there must be redundancy among the quantities because the subspace can be spanned by fewer than n vectors. This redundancy is indeed the key to the theory of Π-numbers.

Let us take a second look at the definition of Π-number, (2). In order for a quantity to be dimensionless, the coefficients of the dimension vectors must all vanish, which, in the new formalism, is equivalent to the corresponding dimension vector being the zero vector. In other words, equation (2) is equivalent to

$$\mu_1 \boldsymbol{q}_1 + \mu_2 \boldsymbol{q}_2 + \cdots + \mu_n \boldsymbol{q}_n = \boldsymbol{R} = \boldsymbol{0}.$$

If we now define the *dimension matrix* of the quantities $q_1, \ldots, q_n$ to be

$$Q = \begin{pmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_n \end{pmatrix} \in \mathbb{R}^{r \times n},$$

we can immediately verify that the vector $\mu \in \mathbb{R}^n$ with entries μ_j must satisfy $Q\mu = \boldsymbol{0}$, so μ must be in the null space of Q, $\mathcal{N}(Q)$.

We can now restate the definition of Π-number in the language of linear algebra: $R = q_1^{\mu_1} \cdots q_n^{\mu_n}$ is a Π-number if and only if $\mu \in \mathcal{N}(Q)$.

It is a central question in dimensional analysis how many essentially different Π-numbers can be found that correspond to a given system. If R_1 and R_2 are two Π-numbers, their product and ratio are also Π-numbers, yet they are not independent. To find out how to determine which Π-numbers are independent, assume that R_1 and R_2 correspond to vectors μ and ν in the null space of Q. From the observation that

$$R_1 \times R_2 = q_1^{\mu_1} \cdots q_n^{\mu_n} \times q_1^{\nu_1} \cdots q_n^{\nu_n} = q_1^{\mu_1 + \nu_1} \cdots q_n^{\mu_n + \nu_n},$$

Copyrighted material

it follows that multiplication of two Π-numbers corresponds to addition of the corresponding vectors in the null space of the dimension vector. This naturally leads to the definition that the Π-numbers $\{R_1, \dots, R_k\}$ are *essentially different* if the corresponding coefficient vectors in $\mathcal{N}(Q)$ are linearly independent. In particular, the number of essentially different Π-numbers is equal to the dimension of $\mathcal{N}(Q)$, and a maximal set of essentially different Π-numbers corresponds to a basis for $\mathcal{N}(Q)$.

It is now easy to state the following central theorem of dimensional analysis, which is a corollary of the theorem about the dimensions of the FOUR FUNDAMENTAL SUBSPACES [I.2 §21] of a dimension matrix.

Buckingham's Π theorem. *If a physical problem is described by n variables, with every variable expressed in terms of r independent and relevant dimensions, the number of essentially different Π-numbers (dimensionless groups whose numerical values depend on the properties of the system) is at most $n - r$.*

It is important to stress that the number of essentially different Π-numbers is "at most" $n - r$ because the system may actually admit fewer. It is a nice corollary that the Π-numbers of a system can be found by computing a basis for the null space of the dimension matrix by Gaussian elimination, which results in rational coefficients.

Returning to our example from fluid dynamics, let a system be described by the five quantities length (L), a characteristic scalar velocity (v_0), density (ρ), viscosity (μ), and pressure (p), the dimensions of which were given in (4). We characterize the system with three SI units, m, s, and kg. The dimension matrix in this case is

$$Q = \begin{pmatrix} 1 & 1 & -3 & -1 & -1 \\ 0 & -1 & 0 & -1 & -2 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

To find a basis of the null space we reduce the matrix to its row echelon form by Gauss–Jordan elimination, which shows that its rank is three. This implies that the null space is two dimensional, with a basis consisting of the two vectors

$$\boldsymbol{u}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{u}_2 = \begin{pmatrix} 0 \\ -2 \\ -1 \\ 0 \\ 1 \end{pmatrix},$$

corresponding to the Reynolds number and the Euler number, respectively:

$$R_1 = L^1 v_0^1 \rho^1 \mu^{-1} p^0, \quad R_2 = L^0 v_0^2 \rho^{-1} \mu^0 p^1.$$

To appreciate the usefulness of finding these Π-numbers, consider the nondimensionalization of the NAVIER–STOKES EQUATION [III.23],

$$\rho\left(\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v}\right) = -\nabla p + \mu \Delta \boldsymbol{v},$$

where $\Delta = \nabla \cdot \nabla$. Assuming that a characteristic speed v_0 (e.g., an asymptotic value) and a characteristic length scale L are given, first we nondimensionalize the velocity and the spatial variable, writing

$$\boldsymbol{v} = v_0 \boldsymbol{\vartheta}, \qquad \boldsymbol{x} = L\boldsymbol{\xi},$$

and then we define a dimensionless pressure field based on the nondimensionality of the Euler number R_2,

$$\pi(\boldsymbol{\xi}) = \frac{1}{\rho v_0^2} p(L\boldsymbol{\xi}),$$

arriving at the scaled version of the equation:

$$\frac{\rho v_0^2}{L}\left(\frac{L}{v_0}\frac{\partial \boldsymbol{\vartheta}}{\partial t} + \boldsymbol{\vartheta} \cdot \nabla' \boldsymbol{\vartheta}\right) = -\frac{\rho v_0^2}{L}\nabla' \pi + \frac{\mu v_0}{L^2}\Delta' \boldsymbol{\vartheta},$$

where $\nabla' = \nabla_{\boldsymbol{\xi}}$ and $\Delta' = \nabla' \cdot \nabla'$. By going further and defining the time in terms of the characteristic timescale L/v_0,

$$t = \frac{L}{v_0}\tau,$$

the nondimensional version of the Navier–Stokes equation ensues:

$$\frac{\partial \boldsymbol{\vartheta}}{\partial \tau} + \boldsymbol{\vartheta} \cdot \nabla' \boldsymbol{\vartheta} = -\nabla' \pi + \frac{1}{R_1}\Delta' \boldsymbol{\vartheta}.$$

This form provides a natural justification for the different approximations corresponding to, for example, nonviscous fluid flow (R_1 large) or nonturbulent flow (R_1 small).

Further Reading

Barenblatt, G. I. 1987. *Dimensional Analysis.* New York: Gordon & Breach.

———. 2003. *Scaling.* Cambridge: Cambridge University Press.

Calvetti, D., and E. Somersalo. 2012. *Computational Mathematical Modeling: An Integrated Approach across Scales.* Philadelphia, PA: SIAM.

Mattheij, R. M. M., S. W. Rienstra, and J. H. M. ten Thije Boonkkamp. 2005. *Partial Differential Equations: Modeling, Analysis and Computation.* Philadelphia, PA: SIAM.

Copyrighted material

II.10 The Fast Fourier Transform

Daniel N. Rockmore

In 1965 James Cooley and John Tukey wrote a brief article (a note, really) that laid out an efficient method for computing the various trigonometric sums necessary for computing or approximating the Fourier transform of a function on the real line. While theirs was not the first such article (it was later discovered that the algorithm's fundamental step was first sketched in papers of Gauss), what was very different was the context. Newly invented analog-to-digital converters had now enabled the accumulation of (for the time) extraordinarily large data sets of sampled time series, whose analysis required the computation of the underlying signal's Fourier transform. In this new world of 1960s "big data," a clever reduction in computational complexity (a term not yet widely in use) could make a tremendous difference.[1]

While the Cooley–Tukey approach is what is usually associated with the phrase "fast Fourier transform" (or "FFT"), this term more correctly refers to a family of algorithms designed to accomplish the efficient calculation of the FOURIER TRANSFORM [II.19] (or an approximation thereof) of a real-valued function f sampled at points x_j (on either the real line, the unit interval, or the unit circle): samples go in and Fourier coefficients are returned. The discrete sums of interest

$$\hat{f}(k) = \sum_{j=0}^{n-1} f(j)\omega_n^{jk} \tag{1}$$

computed for each $k = 0, \dots, n-1$, where $\omega_n = \mathrm{e}^{2\pi \mathrm{i}/n}$ is a primitive nth root of unity and $f(j) = f(x_j)$, make up what is usually called the "discrete Fourier transform" (DFT). This can be written succinctly as the outcome of the matrix–vector multiplication

$$\hat{f} = \Omega f, \tag{2}$$

where the (j,k) element of Ω is ω_n^{jk}.

1 The Cooley–Tukey FFT

If computed directly, the DFT requires n^2 multiplications and $n(n-1)$ additions, or $2n^2 - n$ arithmetic operations (assuming the $f(j)$ values and the powers of the root of unity have been precomputed and stored). Note that this is approximately $2n^2$ (and, asymptotically, $O(n^2)$) operations. The "classical" FFT (i.e., the Cooley–Tukey FFT) can be employed in the case in which n can be factored, $n = pq$, whereupon we can take advantage of a concomitant factorization of the calculation (which, in turn, is a factorization of the matrix Ω) that can be cast as a DIVIDE AND CONQUER ALGORITHM [I.4 §3], writing the DFT of order n as p DFTs of order q (or q DFTs of order p). More explicitly, in this case we can write

$$\begin{aligned} j = j(a,b) &= aq + b, \quad 0 \leqslant a < p,\ 0 \leqslant b < q, \\ k = k(c,d) &= cp + d, \quad 0 \leqslant c < q,\ 0 \leqslant d < p, \end{aligned}$$

so that (1) can be rewritten as

$$\hat{f}(c,d) = \sum_{b=0}^{q-1} \omega_n^{b(cp+d)} \sum_{a=0}^{p-1} f(a,b)\omega_p^{ad} \tag{3}$$

using the fact that $\omega_n^{adq} = \omega_p^{ad}$.

Computation of $\hat{f}$ is now performed in two steps.

First, compute for each b the inner sums (for all d)

$$\tilde{f}(b,d) = \sum_{a=0}^{p-1} f(a,b)\omega_p^{ad}, \tag{4}$$

which have the form of DFTs of length p equispaced among multiples of q. In engineering language, (4) would be called "a subsampled DFT of length p."

Direct calculation of all the $\tilde{f}(b,d)$ requires $pq[p + (p-1)]$ arithmetic operations. Step two is to then compute an additional pq transforms of length q,

$$\hat{f}(c,d) = \sum_{b=0}^{q-1} \omega_n^{b(cp+d)} \tilde{f}(b,d),$$

requiring at most an additional $pq[q + (q-1)]$ operations to complete the calculation. Thus, instead of the approximately $2n^2 = 2(pq)^2$ operations required by direct computation, the above algorithm uses approximately $2(pq)(p+q)$ operations. If n can be factored further, this approach works even better. When n is a power of two, the successive splittings of the calculation give the well-known $O(n \log_2 n)$ complexity result (in comparison to $O(n^2)$).

Since $\Omega^*\Omega = nI$, from (2) we have $f = n^{-1}\Omega^*\hat{f}$, so the discretized function $f = (f(0), \dots, f(n-1))$ (sample values) can be recovered from its Fourier coefficients via

$$f(m) = \frac{1}{n}\sum_k \hat{f}(k)\omega_n^{-mk},$$

1. Many years later Cooley told me that he believed that the fast Fourier transform could be thought of as one of the inspirations for asymptotic algorithmic analysis and the study of computational complexity, as previous to the publication of his paper with Tukey very few people had considered data sets large enough to suggest the utility of an asymptotic analysis.

Copyrighted material

a so-called inverse transform. The inverse transform expresses f as a superposition of (sampled) exponentials or, equivalently, sines and cosines of frequencies that are multiples of $2\pi/n$, so that if we think of f as a function of time, the DFT is a change of basis from the "time domain" to the "frequency domain."

In the case in which $n = 2N - 1$ and the $f(x_j)$ represent equispaced samples of a bandlimited function on the circle (or, equivalently, on the unit interval), so that $x_j = j/n$, and of bandlimit N (i.e., $\hat{f}(k) = 0$ for all $k > N$), then (up to a normalization) the sums exactly compute the Fourier coefficients of the function f (suitably indexed). The form of the inverse transform can itself be restated as a DFT, so that an FFT enables the efficient change of basis between the time and frequency domains.

The utility of an efficient algorithm for computing these sums cannot be overstated—occupying as it does a central position in the world of SIGNAL PROCESSING [IV.35], IMAGE PROCESSING [VII.8], and INFORMATION PROCESSING [IV.36]—not only for the intrinsic interest in the Fourier coefficients (say, in various forms of spectral analysis, especially for time series) but also for their use in effecting an efficient convolution of data sequences via the relation (for two functions on n points)

$$\widehat{(f \star g)}(k) = \hat{f}(k)\hat{g}(k),$$

where

$$(f \star g)(k) = \sum_{m=0}^{n-1} f(k-m)g(m). \tag{5}$$

If computed directly for all k, (5) requires $n[n + (n-1)] = O(n^2)$ operations. An efficient FFT-based convolution is effected by first computing $\hat{f}$ and $\hat{g}$, then using n operations for pointwise multiplication of the transformed sequences, and then using another FFT for the efficient inverse transform back to the time domain.

This relationship is the key to FFTs that work for data streams of prime length p. The best-known ideas make use of rewriting the DFT at nonzero frequencies in terms of a convolution of length $p-1$ and then computing the DFT at the zero frequency directly. One well-known example is Rader's prime FFT, which uses the fact that we can find a generator g of $\mathbb{Z}/p\mathbb{Z}^\times$, a cyclic group (under multiplication) of order $p-1$, to write $\hat{f}(g^{-b})$ as

$$\hat{f}(g^{-b}) = f(0) + \sum_{a=0}^{p-2} f(g^a)\mathrm{e}^{2\pi \mathrm{i} g^{a-b}/p}. \tag{6}$$

The summation in (6) has the form of a convolution of length $p-1$ of the sequence $f'(a) = f(g^a)$ with the function $z(a) = \mathrm{e}^{2\pi \mathrm{i} g^a/p}$.

Through the use of these kinds of reductions—contributions by various members of the "FFT family"—we achieve a general $O(n \log_2 n)$ algorithm.

Further Reading

Brigham, E. O. 1988. *The Fast Fourier Transform and Its Applications.* Englewood Cliffs, NJ: Prentice-Hall.

Cooley, J. W. 1987. The re-discovery of the fast Fourier transform algorithm. *Mikrochimica Acta* 3:33–45.

Heideman, M. T., D. H. Johnson, and C. S. Burrus. 1985. Gauss and the history of the fast Fourier transform. *Archive for History of Exact Sciences* 34(3):265–77.

Maslen, D. K., and D. N. Rockmore. 2001. The Cooley–Tukey FFT and group theory. *Notices of the American Mathematical Society* 48(10):1151–61.

van Loan, C. F. 1992. *Computational Frameworks for the Fast Fourier Transform.* Philadelphia, PA: SIAM.

II.11 Finite Differences

In the definition of the derivative of a real function f of a real variable, $f'(x) = \lim_{\varepsilon \to 0}(f(x+\varepsilon) - f(x))/\varepsilon$, we can take a small positive $\varepsilon = h > 0$ and form the approximation

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}.$$

This process is called *discretization* and the approximation is called a *forward difference* because we evaluate f at a point to the right of x. We could instead take a small negative ε, so that with $h = -\varepsilon$ we have

$$f'(x) = \frac{f(x-h) - f(x)}{-h} = \frac{f(x) - f(x-h)}{h}.$$

The latter approximation is a *backward difference*. Higher derivatives can be approximated in a similar fashion. An example is the *centered second difference* approximation

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

The term *finite differences* is used to describe such approximations to derivatives by linear combinations of function values. One way to derive finite-difference approximations, and also to analyze their accuracy, is by manipulating TAYLOR SERIES [I.2 §9] expansions. A more systematic approach is through the *calculus of finite differences*, which is based on operators such as the *forward difference operator* $\Delta f(x) = f(x+h) - f(x)$ and its powers: $\Delta^2 f(x) = \Delta(\Delta f(x)) = f(x +$

Copyrighted material

$e^A e^B$ holds if A and B commute, but it does not hold in general.

More generally, for any function with a Taylor series expansion the scalar argument can be replaced by a square matrix as long as the eigenvalues of the matrix are within the radius of convergence of the Taylor series. Thus we have

$$\cos(A) = I - \frac{A^2}{2!} + \frac{A^4}{4!} - \frac{A^6}{6!} + \cdots,$$
$$\sin(A) = A - \frac{A^3}{3!} + \frac{A^5}{5!} - \frac{A^7}{7!} + \cdots,$$
$$\log(I + A) = A - \frac{A^2}{2} + \frac{A^3}{3} - \frac{A^4}{4} + \cdots, \quad \rho(A) < 1,$$

where ρ denotes the SPECTRAL RADIUS [I.2 §20]. The series for log raises two questions: does $X = \log(I + A)$ satisfy $e^X = I + A$ and, if so, which of the many matrix logarithms is produced (note that if $e^X = I + A$ then $e^{X+2k\pi iI} = e^X e^{2k\pi iI} = I + A$ for any integer k)? The answer to the first question is yes. The answer to the second question is that the logarithm produced is the *principal logarithm*, which for a matrix with no eigenvalues lying on the nonpositive real axis is the unique logarithm all of whose eigenvalues have imaginary parts lying in the interval $(-\pi, \pi)$.

Defining $f(A)$ via a power series may specify the function only for a certain range of A, as for the logarithm, and moreover, some functions do not have a (convenient) power series. For more general functions a different approach is needed.

If f is analytic on and inside a closed contour Γ that encloses the spectrum of A, then we can define

$$f(A) := \frac{1}{2\pi i} \int_\Gamma f(z)(zI - A)^{-1}\, dz,$$

which is a generalization to matrices of the CAUCHY INTEGRAL FORMULA [IV.1 §7]. Another definition can be given in terms of the JORDAN CANONICAL FORM [II.22] $Z^{-1}AZ = J = \operatorname{diag}(J_1, J_2, \ldots, J_p)$, where J_k is an $m_k \times m_k$ Jordan block with eigenvalue λ_k. The definition is

$$f(A) := Zf(J)Z^{-1} = Z \operatorname{diag}(f(J_k))Z^{-1},$$

where

$$f(J_k) := \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \cdots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{bmatrix}.$$

This definition does not require f to be analytic but merely requires the existence of the derivatives $f^{(j)}(\lambda_k)$ for j up to one less than the size of the largest block in which λ_k appears. Note that when A is diagonalizable, that is, $A = ZDZ^{-1}$ for $D = \operatorname{diag}(\lambda_i)$, the definition is simply $f(A) = Zf(D)Z^{-1}$, where $f(D) = \operatorname{diag}(f(\lambda_i))$.

The Cauchy integral and Jordan canonical form definitions are equivalent when f is analytic.

Some key properties that follow from the definitions are that $f(A)$ commutes with A, $f(X^{-1}AX) = X^{-1}f(A)X$ for any nonsingular X, and $f(A)$ is upper (lower) triangular if A is. It can also be shown that certain forms of identity carry over from the scalar case to the matrix case, under assumptions that ensure that all the relevant matrices are defined. Examples are $\exp(iA) = \cos(A) + i\sin(A)$ and $\cos^2(A) + \sin^2(A) = I$. However, care is needed when dealing with multivalued functions; for example, for the principal logarithm, $\log(e^A)$ cannot be guaranteed to equal A without restrictions on the spectrum of A.

Another important class of functions is the pth roots: the solutions of $X^p = A$, where p is a positive integer. For nonsingular A there are many pth roots. The one usually required in practice is the *principal pth root*, defined for A with no eigenvalues lying on the nonpositive real axis as the unique pth root whose eigenvalues lie strictly within the wedge making an angle π/p with the positive real axis, and denoted by $A^{1/p}$. Thus $A^{1/2}$ is the square root whose eigenvalues all lie in the open right half-plane.

The function $\operatorname{sign}(A) = A(A^2)^{-1/2}$, defined for any A having no pure imaginary eigenvalues, is the *matrix sign function*. It has applications in control theory, in particular for solving ALGEBRAIC RICCATI EQUATIONS [III.25], and corresponds to the scalar function mapping complex numbers in the open left and right half-planes to -1 and 1, respectively.

Matrix functions provide one-line solutions to many problems. For example, the second-order ordinary differential equation initial-value problem

$$\frac{d^2y}{dt^2} + Ay = 0, \quad y(0) = y_0,\ y'(0) = y_0',$$

with y an n-vector and A an $n \times n$ matrix, has solution

$$y(t) = \cos(\sqrt{A}t)y_0 + (\sqrt{A})^{-1}\sin(\sqrt{A}t)y_0',$$

where $\sqrt{A}$ denotes *any* square root of A. Alternatively, by writing $z = \left[\begin{smallmatrix} y' \\ y \end{smallmatrix}\right]$ we can convert the problem into two first-order differential equations:

$$z' = \begin{bmatrix} y'' \\ y' \end{bmatrix} = \begin{bmatrix} 0 & -A \\ I & 0 \end{bmatrix} \begin{bmatrix} y' \\ y \end{bmatrix} = \begin{bmatrix} 0 & -A \\ I & 0 \end{bmatrix} z,$$

Copyrighted material

from which follows the formula

$$\begin{bmatrix} y'(t) \\ y(t) \end{bmatrix} = \exp\left(\begin{bmatrix} 0 & -tA \\ tI & 0 \end{bmatrix} \right) \begin{bmatrix} y_0' \\ y_0 \end{bmatrix}.$$

There is an explicit formula for a function of a 2×2 triangular matrix:

$$f\left(\begin{bmatrix} \lambda_1 & \alpha \\ 0 & \lambda_2 \end{bmatrix} \right) = \begin{bmatrix} f(\lambda_1) & \alpha f[\lambda_1, \lambda_2] \\ 0 & f(\lambda_2) \end{bmatrix},$$

where

$$f[\lambda_1, \lambda_2] = \begin{cases} \dfrac{f(\lambda_2) - f(\lambda_1)}{\lambda_2 - \lambda_1}, & \lambda_1 \neq \lambda_2, \\ f'(\lambda_2), & \lambda_1 = \lambda_2, \end{cases}$$

is a first-order divided difference. This formula extends to $n \times n$ triangular matrices T, although the formula for the (i, j) element contains up to 2^n terms and so is not computationally useful unless n is very small. It is nevertheless possible to compute $F = f(T)$ for an $n \times n$ triangular matrix in $n^3/3$ operations using the *Parlett recurrence*, which is obtained by equating elements in the equation $TF = FT$.

Further Reading

Higham, N. J. 2008. *Functions of Matrices: Theory and Computation*. Philadelphia, PA: SIAM.

Higham, N. J., and A. H. Al-Mohy. 2010. Computing matrix functions. *Acta Numerica* 19:159–208.

II.15 Function Spaces

Hans G. Feichtinger

While in the early days of mathematics each function was treated individually, it became appreciated that it was more appropriate to make collective statements for all continuous functions, all integrable functions, or all continuously differentiable ones. Fortunately, most of these collections of functions (f_k) are closed under addition and allow the formation of linear combinations $\sum_{k=1}^{K} c_k f_k$ for real or complex coefficients c_k, $1 \leqslant k \leqslant K$. They are, therefore, vector spaces. In addition, most of these spaces are endowed with a suitable NORM [I.2 §19.3] $f \mapsto \|f\|$, allowing one to measure the size of their members and hence to introduce concepts of closeness by looking at the distance $d(f_1, f_2) := \|f_1 - f_2\|$. One can therefore say that a *function space* is a normed space consisting of (generalized) functions on some domain.

For the vector space $C_b(D)$ of bounded and continuous functions on some domain $D \subset \mathbb{R}^d$, the sup-norm $\|f\|_\infty := \sup_{z \in D} |f(z)|$ is the appropriate norm. With this norm, $C_b(D)$ is a Banach space (that is, a complete normed space), i.e., every CAUCHY SEQUENCE [I.2 §19.4] with respect to this norm is convergent to a unique limit element in the space. Hence, such Banach spaces of functions share many properties with the Euclidean spaces of vectors in $\mathbb{R}^d$, with the important distinction that they are not finite dimensional.

1 Lebesgue Spaces $L^p(\mathbb{R}^d)$

The completeness of the *Lebesgue space* $L^1(\mathbb{R}^d)$, consisting of all (measurable) functions with $\|f\|_1 := \int_{\mathbb{R}^d} |f(z)|\, \mathrm{d}z < \infty$, is the reason why the Lebesgue integral is preferred over the Riemann integral. Note that in order to ensure the property that $\|f\|_1 = 0$ implies $f = 0$ (the null function), one has to regard two functions f_1 and f_2 as equal if they are equal up to a set of measure zero, i.e., if the set $\{z \mid f_1(z) \neq f_2(z)\}$ has Lebesgue measure zero.

Another norm that is important for many applications is the L^2-norm, $\|f\|_2 := (\int_{\mathbb{R}^d} |f(z)|^2)^{1/2}$. It is related to an inner product defined as $\langle f, g \rangle := \int_{\mathbb{R}^d} f(z)\overline{g(z)}\, \mathrm{d}z$ via the formula $\|f\|_2 := \langle f, f \rangle^{1/2}$. $(L^2(\mathbb{R}^d), \|\cdot\|_2)$ is a Hilbert space, and one can talk about orthogonality and unitary linear mappings, comparable with the situation of the Euclidean space $\mathbb{R}^d$ with its standard inner product.

Having these three norms, namely $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$, it is natural to look for norms "in between." This leads to the L^p-spaces, defined by the finiteness of $\|f\|_p^p := \int_{\mathbb{R}^d} |f(z)|^p\, \mathrm{d}z$ for $1 \leqslant p < \infty$. The limiting case for $p \to \infty$ is the space $L^\infty(\mathbb{R}^d)$ of essentially bounded functions.

Since these spaces are not finite dimensional, it is necessary to work with the set of all bounded linear functionals, the so-called dual space, which is often, but not always, a function space. For $1 \leqslant p < \infty$ the dual space to L^p is L^q, with $1/p + 1/q = 1$, meaning that any continuous linear functional on $L^p(\mathbb{R}^d)$ has the form $f \mapsto \int_{\mathbb{R}^d} f(z) g(z)\, \mathrm{d}z$ for a unique function $g \in L^q(\mathbb{R}^d)$.

$L^1(\mathbb{R}^d)$ also appears as the natural domain for the Fourier transform, given for $s \in \mathbb{R}^d$ by

$$\mathcal{F}\colon f \mapsto \hat{f}(s) = \int_{\mathbb{R}^d} f(t) \exp\left\{ -2\pi \mathrm{i} \sum_{j=1}^{d} s_j t_j \right\} \mathrm{d}t,$$

while $(L^2(\mathbb{R}^d), \|\cdot\|_2)$ allows us to describe $\mathcal{F}$ as a *unitary* (and hence isometric) automorphism.

Copyrighted material

2 Related Function Spaces

The Lebesgue spaces are prototypical for a much larger class of *Banach function spaces* or *Banach lattices*, a class that also includes *Lorentz spaces* $L(p,q)$ or *Orlicz spaces* L^Φ, which are all *rearrangement invariant*. This means that for any transformation $\alpha\colon \mathbb{R}^d \to \mathbb{R}^d$ that has the property that it preserves the (Lebesgue) measure $|M|$ of a set, i.e., with $|\alpha(M)| = |M|$, one has $\|f\| = \|\alpha^*(f)\|$, where $\alpha^*(f)(z) := f(\alpha(z))$.

In contrast, *weighted spaces* such as $L^p_w(\mathbb{R}^d)$, characterized by $\|f\|_{p,w} = \|fw\|_p < \infty$, allow us to capture the decay of f at infinity using some strictly positive *weight function* w. For applications in the theory of partial differential equations (PDEs), polynomial weights such as $w_s(x) = (1+|x|^2)^{s/2}$ are important. For $s \geqslant 0$, *Sobolev spaces* $\mathcal{H}_s(\mathbb{R}^d)$ can be defined as inverse images of $L^2_{w_s}(\mathbb{R}^d)$ under the Fourier transform. For $s \in \mathbb{N}$ they consist of those functions that have (in a distributional sense) s derivatives in $L^2(\mathbb{R}^d)$. *Mixed norm* L^p spaces (using different p-norms in different directions) are also not invariant in this sense, but they are still very useful.

A large variety of function spaces arose out of the attempt to characterize smoothness, including fractional differentiability. Examples are *Besov spaces* $B^s_{p,q}(\mathbb{R}^d)$ and *Triebel–Lizorkin spaces* $F^s_{p,q}(\mathbb{R}^d)$; the classical Sobolev spaces are the only function spaces that belong to both families. The origin of this theory is in the theory of Lipschitz spaces $\mathrm{Lip}(\alpha)$, where the range $\alpha \in (0,1)$ allows us to express the degree of smoothness (differentiability corresponds intuitively to the case $\alpha = 1$).

3 Wavelets and Modulation Spaces

Many of the spaces mentioned above are highly relevant for PDEs, e.g., the description of elliptic PDEs. Their characterization using Paley–Littlewood (dyadic Fourier) decompositions has ignited wavelet theory. For $1 < p < \infty$ they can be characterized via (weighted) summability conditions of their wavelet coefficients with respect to (sufficiently "good") mother wavelets. In the limiting case, one obtains the *real Hardy space* $H^1(\mathbb{R}^d)$ and its dual, the BMO-space, which consists of functions of *bounded mean oscillation*. Both spaces are important for the study of Calderon–Zygmund operators or the Hardy–Littlewood maximal operator. Wavelets provide unconditional bases for these spaces, including *Besov* and *potential* spaces.

For the affine "$ax + b$"-group acting on the space $L^2(\mathbb{R}^d)$, function spaces are defined using the continuous wavelet transform, and atomic characterizations (involving *Banach frames*) of the above smoothness spaces are obtained. Alternatively, the Schrödinger representation of the Heisenberg group, again on $L^2(\mathbb{R}^d)$ via time–frequency shifts, gives rise to the family of *modulation spaces* $M^s_{p,q}$. They were introduced as *Wiener amalgam spaces* on the Fourier transform side, using uniform partitions of unity (instead of dyadic ones).

Using engineering terminology, the now-classical spaces $M^s_{p,q}(\mathbb{R}^d)$ are characterized by the behavior of the *short-time Fourier transform* of their members (replacing the continuous wavelet transform). They play an important role in time–frequency analysis, and their atomic characterizations use *Gabor expansions*.

A variety of Banach spaces of analytic or polyanalytic functions play an important role in complex analysis. Again, integrability conditions over their domain are typically used to define these spaces. The corresponding L^2-spaces are typically reproducing kernel Hilbert spaces, with good localization of these kernels allowing one to view them as continuous mappings on (weighted, mixed-norm) L^p-spaces as well. We mention some of the spaces that are important in the context of complex analysis or Toeplitz operators: Fock spaces, Bergman spaces, and Segal–Bargmann spaces.

4 Variations of the Theme

One of the first important examples of a Banach space of functions was the space BV of functions of *bounded variation*. One simple characterization of functions of this type (the so-called Jordan decomposition) is that they are the difference of two bounded and nondecreasing functions (the ascending part of the function minus the descending part of it). Via Fourier–Stieltjes integrals, F. Riesz showed that there is a one-to-one correspondence between the dual space of $(C[0,1], \|\cdot\|_\infty)$ and $BV[0,1]$ endowed with the variation norm. More recently, total variation in a two-dimensional setting has been fundamental to IMAGE RESTORATION ALGORITHMS [VII.8]. Another family of function spaces that captures variation at different scales are the Morrey–Campanato spaces.

In addition to Banach spaces of functions there are also topological vector spaces and Fréchet spaces of functions, among them the *spaces of test functions* that are used in distribution theory. *Generalized functions*

Copyrighted material

consist of the continuous linear functionals on such spaces. The "theory of function spaces" as developed by Hans Triebel includes a large variety of Banach spaces of such generalized functions (or distributions).

Further Reading

Ambrosio, L., N. Fusco, and D. Pallara. 2000. *Functions of Bounded Variation and Free Discontinuity Problems.* Oxford: Clarendon.

Gröchenig, K. 2001. *Foundations of Time–Frequency Analysis.* Boston, MA: Birkhäuser.

Leoni, G. 2009. *A First Course in Sobolev Spaces.* Providence, RI: American Mathematical Society.

Meyer, Y. 1992. *Wavelets and Operators*, translated by D. H. Salinger. Cambridge: Cambridge University Press.

Stein, E. M. 1970. *Singular Integrals and Differentiability Properties of Functions.* Princeton, NJ: Princeton University Press.

Triebel, H. 1983. *Theory of Function Spaces.* Basel: Birkhäuser.

II.16 Graph Theory

Timothy A. Davis and Yifan Hu

At the back of most airline magazines you will find a map of airports and the airline routes that connect them. This is just one example of a *graph*, a widely used mathematical entity that represents relationships between discrete objects. More precisely, a graph $G = (V, E)$ consists of a set of *nodes* V and a set of *edges* $E \subseteq \{(i, j) \mid i, j \in V\}$ that connect them. A graph is not a diagram but it can be drawn, as illustrated in figure 1.

Graphs arise in a vast array of applications, including social networks (a node is a person and an edge is a relationship between two people), computational fluid dynamics (a node is an unknown such as the pressure at a certain point and an edge is the physical connection between two unknowns), finding things on the web (a node is a web page and an edge is a link), circuit simulation (the wires are the edges), economics (a node is a financial entity and the edges represent trade between two entities), and many others.

In some problems, an edge connects in both directions, and in this case the graph is *undirected*. For example, friendship is mutual, so if Alice and Bob are friends, the edges (Alice, Bob) and (Bob, Alice) are the same. In other cases, the direction of the edge is important. If Alice follows Bob on Twitter, this does not mean that Bob follows Alice. In this *directed* graph, the edge (Alice, Bob) is not the same as the edge (Bob, Alice).

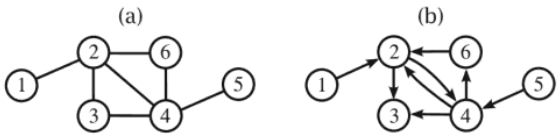


Figure 1 Two example graphs: (a) undirected and (b) directed.

In a *simple* graph, an edge (i, j) can appear just once, but in a *multigraph* it can appear multiple times (E becomes a multiset). Simple graphs do not have *self-edges* (i, i), but a *pseudograph* can have multiple edges and self-edges. The airline route map in the back of the magazine is an example of a simple undirected graph. Representing each flight for a whole airline would require a directed multigraph: the flight from Philadelphia to New York is not the same as the flight in the opposite direction, and there are many flights each day between the two airports. If sightseeing tours are added (self-edges), then a pseudograph would be needed.

The *adjacency set* of a node i, also called its *neighbors*, is the set of nodes j where edge (i, j) is in the graph. For a directed graph, this is the out-adjacency; the in-adjacency of node i is the set $\{j \mid (j, i) \in E\}$. A graph can be represented as a binary *adjacency matrix*, with entries $a_{ij} = 1$ if $(i, j) \in E$, and $a_{ij} = 0$ otherwise. The *degree* of a node is the size of its adjacency set.

Graphs can contain infinite sets of nodes and edges. Consider the directed graph on the natural numbers $\mathbb{N}$ with the edges (i, j), where j is an integer multiple of i. A prime number $j > 1$ in this graph has in-adjacency $\{1, j\}$ and an in-degree of 2 (including the self-edge (j, j)); a composite number $j > 1$ has a larger in-degree.

Nodes i and j are *incident* on the edge (i, j) and, likewise, the edge (i, j) is incident on its two nodes. A *subgraph* of G consists of a subset of its nodes and edges, $\bar{G} = (\bar{V}, \bar{E})$, where $\bar{V} \subseteq V$ and $\bar{E} \subseteq E$. If an edge (i, j) appears in $\bar{E}$, then its two incident nodes must also appear in $\bar{V}$, but the opposite need not hold. Two special kinds of subgraphs are *node-induced* and *edge-induced* subgraphs. A node-induced subgraph starts with a subset of nodes $\bar{V}$; the edges $\bar{E}$ are all those edges whose two incident nodes are both in $\bar{V}$. An *edge-induced* subgraph starts with a subset of edges $\bar{E}$ and then $\bar{V}$ consists of all nodes incident on those edges. A graph is *completely connected* if it has an edge between every pair of nodes. A *clique* is a completely connected subgraph.

Copyrighted material

A *path* from i to j is a list of nodes $(i, \ldots, j)$ with edges between adjacent pairs of nodes. The path cannot traverse a directed edge backward. The *length* of the path is the number of nodes in the list minus one. In a *simple* path, a node can appear only once. If there is a path from i to j, then node j is *reachable* from node i. The set of all nodes reachable from i is the *reach* of i. Among all paths from i to j, one with the shortest length is a *shortest path*, its length the *(geodesic) distance* from i to j. The *diameter* of a graph is the length of the longest possible shortest path. In a *small-world graph*, each node is a small distance (logarithmic in the number of nodes) away from any other node.

An undirected graph is *connected* if there is a path between each pair of nodes, but there are two kinds of connectivity in a directed graph. If a path exists between every pair of nodes, then a directed graph is *strongly connected*. A directed graph is *weakly connected* if its *underlying undirected graph* is connected; to obtain such a graph, all edge directions are dropped.

A *cycle* is a path that starts and ends at the same node i; the cycle is *simple* if no node is repeated (except for node i itself). There are no cycles in an *acyclic* graph. The acronym DAG is often used for a directed acyclic graph.

The undirected graph in figure 1(a) is connected. Nodes $\{2,3,4\}$ form a clique, as do $\{2,4,6\}$. The path $(1,2,4,3,2,6)$ has length 5 and is not simple. A simple path from 1 to 6 is $(1,2,4,6)$ of length 3, but the shortest path is $(1,2,6)$ of length 2, which traverses the edges $(1,2)$ and $(2,6)$. The path $(2,3,4,2)$ is a cycle of length 3. Node 2 has degree 4, with neighbors $\{1,3,4,6\}$. The diameter of the graph is 3. Since the graph is connected, the reach of node 2 is the whole graph. This graph is the underlying undirected graph of the directed graph in part (b) of the figure.

The largest clique in this directed graph has only two nodes: $\{2,4\}$. The out-adjacency of node 2 is the set $\{3,4\}$ and its in-adjacency is $\{1,4,6\}$. The reach of node 2 is $\{2,3,4,6\}$. The graph is not strongly connected since there is no path from 1 to 5, but it is weakly connected since its underlying undirected graph is connected.

Figure 2 illustrates a *bipartite* graph. The nodes of a bipartite graph are partitioned into two sets, and no edge in the graph is incident on a pair of nodes in the same partition. Bipartite graphs arise naturally when modeling a relationship between two very different sets. For example, in term/document analysis, a bipartite graph of m terms and n documents has an edge (i,j) if term i appears in document j. No edge connects two terms, nor two documents. An undirected bipartite graph is often represented as a rectangular $m \times n$ adjacency matrix, where $a_{ij} = 1$ if the edge (i,j) appears in the graph and $a_{ij} = 0$ otherwise.

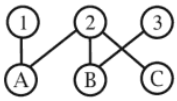


Figure 2 An undirected bipartite graph.

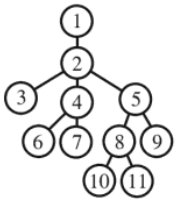


Figure 3 A tree of height 4, with node 1 as the root.

An undirected acyclic graph is a *forest*. An important special case is a *tree*, which is a connected forest. In a tree, there is a unique simple path between each pair of nodes. In a *rooted* tree, one node is designated as the *root*. The *ancestors* of node i are all the nodes on the path from i to the root (excluding i itself). The first node after i in this path is the *parent* of i, and node i is its *child*. The length of this path is the *level* of the node (the root has level zero). The *height* of a tree is the maximum level of its nodes.

In a tree, all nodes except the root have a single parent. Nodes can have any number of children, and a node with no children is a *leaf*. *Internal* nodes have at least one child. In a *binary* tree, nodes have at most two children, and in a *full binary* tree, all internal nodes have exactly two children. Node i is a *descendant* of all nodes in the path from i to the root (excluding i itself). The *subtree* rooted at node i is the subgraph induced by node i and its descendants.

In the example in figure 3, the parent of node 5 is 2, its descendants are $\{8,9,10,11\}$, its ancestors are $\{1,2\}$, and its children are $\{8,9\}$. Since node 2 has three children, the tree is not binary.

Sometimes a graph with its nodes and edges is not enough to fully represent a problem. Edges in a graph do not have a length, but this is useful for the airline route map, and thus nodes and edges are often augmented with additional data. Attaching a single numerical value to each node and/or edge is common; this

Copyrighted material

and the *Laplace transform*

$$(\mathcal{L}f)(s) = \int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t.$$

This definition of $\mathcal{L}$ is standard, but the definition of $\mathcal{F}$ is one of many; for example, some authors insert an extra factor $(2\pi)^{-1/2}$, and some use $\mathrm{e}^{-\mathrm{i}st}$. Always check the author's definition when reading a book or article in which Fourier transforms are used!

Many integral transforms have an associated *convolution*; given two functions of a real variable, f and g, their convolution is another function of a real variable, denoted by $f * g$. It is defined so that $J(f * g) = (Jf)(Jg)$; the transform of the convolution is the product of the transforms. For the Fourier transform

$$(f * g)(t) = \int_{-\infty}^\infty f(t-s)g(s)\,\mathrm{d}s,$$

while for the Laplace transform

$$(f * g)(t) = \int_0^t f(t-s)g(s)\,\mathrm{d}s.$$

It is easy to see that, in both cases, $f * g = g * f$. Convolution is an important operation in SIGNAL PROCESSING [IV.35] and in many applications involving Fourier analysis and INTEGRAL EQUATIONS [IV.4].

There are also discrete versions of integral transforms in which the integral is replaced by a finite sum of terms. The *discrete Fourier transform* is especially important because it can be computed rapidly using the FAST FOURIER TRANSFORM [II.10] (FFT).

II.20 Interval Analysis

Warwick Tucker

Interval analysis is a calculus based on set-valued mathematics. In its simplest (and by far most popular) form, it builds upon interval arithmetic, which is a natural extension of real-valued arithmetic. Despite its simplicity, this kind of set-valued mathematics has a very wide range of applications in computer-aided proofs for continuous problems. In a nutshell, interval arithmetic enables us to bound the range of a continuous function, i.e., it produces a set *enclosing* the range of a given function over a given domain. This, in turn, enables us to prove mathematical statements that use open conditions, such as strict inequalities, fixed-point theorems, etc.

1 Interval Arithmetic

In this section we will briefly describe the fundamentals of interval arithmetic. Let $\mathbb{IR}$ denote the set of closed intervals of the real line. For any element $\boldsymbol{a} \in \mathbb{IR}$, we use the notation $\boldsymbol{a} = [\underline{a}, \overline{a}]$. If $\star$ is one of the operators $+, -, \times, /$, we define arithmetic on elements of $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{IR}$ by

$$\boldsymbol{a} \star \boldsymbol{b} = \{a \star b \colon a \in \boldsymbol{a},\ b \in \boldsymbol{b}\}, \tag{1}$$

except that $\boldsymbol{a}/\boldsymbol{b}$ is undefined if $0 \in \boldsymbol{b}$. Working exclusively with closed intervals, the resulting interval can be expressed in terms of the endpoints of the arguments. This makes the arithmetic very easy to implement in software.

Note that a generic element in $\mathbb{IR}$ has no additive or multiplicative inverse. For example, we have $[1,2] - [1,2] = [-1,1] \neq [0,0]$, and $[1,2]/[1,2] = [\frac{1}{2}, 2] \neq [1,1]$. This is known as the *dependency problem*, and it can cause large overestimations. In practice, however, the use of high-order (e.g., Taylor series) representations greatly mitigates this problem.

A key feature of interval arithmetic is that it is *inclusion monotonic*; i.e., if $\boldsymbol{a} \subseteq \boldsymbol{a}'$ and $\boldsymbol{b} \subseteq \boldsymbol{b}'$, then by (1) we have

$$\boldsymbol{a} \star \boldsymbol{b} \subseteq \boldsymbol{a}' \star \boldsymbol{b}'.$$

This is of fundamental importance: it says that, if we can enclose the arguments, we can enclose the result.

More generally, when we extend a real-valued function f to an interval-valued one F, we demand that it satisfies the *inclusion principle*

$$\mathrm{range}(f; \boldsymbol{x}) = \{f(x) \colon x \in \boldsymbol{x}\} \subseteq F(\boldsymbol{x}). \tag{2}$$

If this can be arranged for a finite set of *standard* functions, then the inclusion principle will also hold for any *elementary* function constructed by arithmetic and composition applied to the set of standard functions.

Multivariate functions can be handled by working componentwise on interval vectors (boxes) $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

When implementing interval arithmetic on a computer, the endpoints must be FLOATING-POINT NUMBERS [II.13]. This introduces rounding errors, which must be properly dealt with. As an example, interval addition becomes

$$\boldsymbol{a} + \boldsymbol{b} = [\nabla(\underline{a} + \underline{b}), \triangle(\overline{a} + \overline{b})].$$

Here, $\nabla(x)$ is the largest floating-point number no greater than x, and $\triangle(x) = -\nabla(-x)$. The IEEE standard for floating-point computations guarantees that this type of *outward rounding* preserves the inclusion principle for $+$, $-$, $\times$, and $/$. For other operations (such as trigonometric functions) there are no such assurances; interval extensions of these functions must be built from scratch.

Copyrighted material

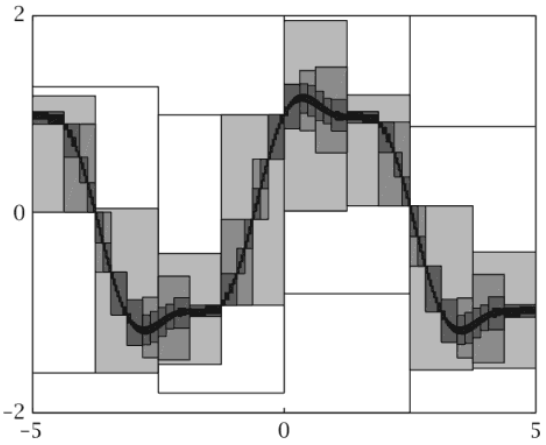


Figure 1 Successively tighter enclosures of a graph.

2 Interval Analysis

The inclusion principle (2) enables us to capture continuous properties of a function, using only a finite number of operations. Its most important use is to explicitly bound discretization errors that naturally arise in numerical algorithms.

As an example, consider the function $f(x) = \cos^3 x + \sin x$ on the domain $\boldsymbol{x} = [-5, 5]$. For any decomposition of the domain $\boldsymbol{x}$ into a finite set of subintervals $\boldsymbol{x} = \bigcup_{i=1}^{n} \boldsymbol{x}_i$, we can form the set-valued graph consisting of the pairs $(\boldsymbol{x}_1, F(\boldsymbol{x}_1)), \ldots, (\boldsymbol{x}_n, F(\boldsymbol{x}_n))$. As the partition is made finer (that is, as $\max_i \operatorname{diam}(\boldsymbol{x}_i)$ is made smaller), the set-valued graph tends to the graph of f (see figure 1). And, most importantly, every such set-valued graph contains the graph of f.

This way of incorporating the discretization errors is extremely useful for quadrature, optimization, and equation solving. As one example, suppose we wish to compute the definite integral $I = \int_0^8 \sin(x + e^x)\, dx$.

A MATLAB function `simpson` that implements a simple textbook adaptive Simpson quadrature algorithm produces the following result.

```
% Compute integral I with tolerance 1e-6.
>> I = simpson(@(x) sin(x + exp(x)), 0, 8)
I =
   0.251102722027180
```

A (*very naive*) set-valued approach to quadrature is to enclose the integral I via

$$I \in \sum_{i=1}^{n} F(\boldsymbol{x}_i) \operatorname{diam}(\boldsymbol{x}_i),$$

which, for a sufficiently fine partition, produces the integral enclosure

$$I \in 0.347400172\,6^{66}_{49}.$$

Thus, it turns out that the result from `simpson` was completely wrong! This is one example of the importance of rigorous computations.

3 Recent Developments

There is currently an ongoing effort within the IEEE community to standardize the implementation of interval arithmetic. The hope is that we will enable computer manufacturers to incorporate these types of computations at the hardware level. This would remove the large computational penalty incurred by repeatedly having to switch rounding modes—a task that central processing units were not designed to perform efficiently.

Further Reading

Alefeld, G., and J. Herzberger. 1983. *Introduction to Interval Computations.* New York: Academic Press.

Kulisch, U. W., and W. L. Miranker. 1981. *Computer Arithmetic in Theory and Practice.* New York: Academic Press.

Lerch, M., G. Tischler, J. Wolff von Gudenberg, W. Hofschuster, and W. Krämer. 2006. filib++, a fast interval library supporting containment computations. *ACM Transactions on Mathematical Software* 32(2):299–324.

Moore, R. E., R. B. Kearfott, and M. J. Cloud. 2009. *Introduction to Interval Analysis.* Philadelphia, PA: SIAM.

Rump, S. M. 1999. INTLAB—INTerval LABoratory. In *Developments in Reliable Computing*, edited by T. Csendes, pp. 77–104. Dordrecht: Kluwer Academic.

Tucker, W. 2011. *Validated Numerics: A Short Introduction to Rigorous Computations.* Princeton, NJ: Princeton University Press.

II.21 Invariants and Conservation Laws

Mark R. Dennis

As important as the study of *change* in the mathematical representation of physical phenomena is the study of *invariants*. Physical laws often depend only on the *relative* positions and times between phenomena, so certain physical quantities do not change; i.e., they are *invariant*, under continuous translation or rotation of the spatial axes. Furthermore, as the spatial configuration of a system evolves with time, quantities such as total energy may remain unchanged; that is, they are

Copyrighted material

conserved. The study of invariants has been a remarkably successful approach to the mathematical formulation of physical laws, and the study of continuous symmetries and conservation laws—which are related by the result known as *Noether's theorem*—has become a systematic part of our description of physics over the last century, from the atomic scale to the cosmic scale.

As an example of so-called *Galilean invariance*, Newton's force law keeps the same form when the velocity of the frame of reference (i.e., the coordinate system specified by x-, y-, and z-axes) is changed by adding a constant; this is equivalent to adding the same constant velocity to all the particles in a mechanical system. Other quantities *do* change under such a velocity transformation, such as the kinetic energy $\frac{1}{2}m|\boldsymbol{v}|^2$ (for a particle of mass m and velocity $\boldsymbol{v}$); however, for an evolving, nondissipative system such as a bouncing, perfectly elastic rubber ball, the total energy is constant in time—that is, energy is conserved.

The development of our understanding of fundamental laws of dynamics can be interpreted by progressively more sophisticated and general representations of space and time themselves: ancient Greek physical science assumed absolute space with a privileged spatial point (the center of the Earth), through static Euclidean space where all spatial points are equivalent, through CLASSICAL MECHANICS [IV.19] where all inertial frames, moving at uniform velocity with respect to each other, are equivalent according to Newton's first law, to the modern theories of special and general relativity. The theory of relativity (both general and special) is motivated by Einstein's principle of covariance, which is described below. In this theory, space and time in different frames of reference are treated as coordinate systems on a four-dimensional pseudo-Riemannian manifold (whose mathematical background is described in TENSORS AND MANIFOLDS [II.33]), which manifestly combines conservation laws and continuous geometric symmetries of space and time. In special relativity (described in some detail in this article), this manifold is flat *Minkowski space-time*, generalizing Euclidean space to include time in a physically natural way. In general relativity, described in detail in GENERAL RELATIVITY AND COSMOLOGY [IV.40], this manifold may be curved, depending in part on the distribution of matter and energy according to EINSTEIN'S FIELD EQUATIONS [III.10].

In quantum physics, the description of a system in terms of a complex vector in Hilbert space gives rise to new symmetries. An important example is the fact that physical phenomena do not depend on the overall phase (argument) of this vector. Extension of Noether's theorem here leads to the conservation of electric charge, and extension to Yang–Mills theories provides other conserved quantities associated with the nuclear forces studied in contemporary fundamental particle physics. Other phenomena, such as the Higgs mechanism (leading to the Higgs boson recently discovered in high-energy experiments), are a consequence of the breaking of certain quantum symmetries in certain low-energy regimes. Symmetry and symmetry breaking in quantum theory are discussed briefly at the end of this article.

Spatial vectors, such as $\boldsymbol{r} = (x, y, z)$, represent the spatial distance between a chosen point and the origin, and of course the vector between two such points $\boldsymbol{r}_2 - \boldsymbol{r}_1$ is independent of translations of this origin. Similarly, the scalar product $\boldsymbol{r}_2 \cdot \boldsymbol{r}_1$ is unchanged under rotation of the coordinate system by an orthogonal matrix $\boldsymbol{R}$, under which $\boldsymbol{r} \to \boldsymbol{R}\boldsymbol{r}$.

Continuous groups of transformations such as translation and rotation, and their matrix representations, are an important tool used in calculations of invariants. For example, the set of two-dimensional matrices $\left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$, representing rotations through angles θ, may be considered as a continuous one-parameter Abelian group of matrices generated by the MATRIX EXPONENTIAL [II.14] $\mathrm{e}^{\theta \boldsymbol{A}}$, where $\boldsymbol{A}$ is the generator $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. The generator itself is found as the derivative of the original matrix with respect to θ, evaluated at $\theta = 0$. Translations are less obviously represented by matrices; one approach is to append an extra dimension to the position vector with unit entry, such as $(1, x)$ specifying one-dimensional position x; a translation by X is thus represented by

$$\begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} = \exp\left(X \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \right). \tag{1}$$

When a physical system is invariant under a one-parameter group of transformations, the corresponding generator plays a role in determining the associated conservation law.

1 Mechanics in Euclidean Space

It is conventional in classical mechanics to define the positions of a set of interacting particles in a vector space. However, we do not observe any unique origin to the three-dimensional space we inhabit, which we therefore take to be the *Euclidean space* $\mathbb{E}^3$; only relative positions between different interacting subsystems

Copyrighted material

(i.e., positions relative to the common center of mass) enter the equations of motion. The entire system may be translated in space without any effect on the phenomena.

Of course, *external forces* acting on the system may prevent this, such as a rubber ball in a linear gravity field. (In such situations the source of the force, such as the Earth as the source of gravity, is not considered part of the system.) The gravitational force may be represented by a potential $V = gz$ for height z and gravitational acceleration g; the ball's mass m times the negative gradient, $-m\nabla V$, gives the downward force acting on the ball. The contours of V, given by $z = \text{const.}$, nevertheless have a symmetry: they are invariant to translations of the horizontal coordinates x and y. Since the gradient of the potential is proportional to the gravitational force—which, by virtue of Newton's law equals the rate of change of the particle's linear momentum—the horizontal component of momentum does not change and is therefore conserved even when the particle bounces due to an impulsive, upward force from the floor. The continuous, horizontal translational symmetry of the system therefore leads to conservation of linear momentum in the horizontal plane. In a similar argument employing Newton's laws in cylindrical polar coordinates, the invariance of the potential to rotations about the z-axis leads to the conservation of the vertical component a body's angular momentum, as observed for tops spinning frictionlessly.

2 Noether's Theorem

The Lagrangian framework for mechanics (CLASSICAL MECHANICS [IV.19 §2]), which describes systems acting under forces defined by gradients of potentials (as in the previous section), is a natural mathematical setting in which to explore the connection between a system's symmetries and its conservation laws. Here, a mechanical system evolving in time t is described by n generalized coordinates $q_j(t)$ and their time derivatives $\dot{q}_j$, for $j = 1, \dots, n$, where the initial values $q_j(t_0)$ at time t_0 and final values $q_j(t_1)$ at t_1 are fixed. The *action* of the system is the functional

$$S[\{q_j\}] = \int_{t_0}^{t_1} L(\{q_j\}, \{\dot{q}_j\}, t)\, \mathrm{d}t,$$

where $L(\{q_j\}, \{\dot{q}_j\}, t)$ is the *Lagrangian*; this is a function of the coordinates, their corresponding velocities, and maybe time, specified here by the total kinetic energy minus the total potential energy of the system (thereby capturing the forces as gradients of the potential energy). Using the CALCULUS OF VARIATIONS [IV.6], the functions $q_j(t)$ that satisfy the laws of mechanics are those that make the action stationary, and these satisfy Lagrange's equations of motion

$$\frac{\partial L}{\partial q_j} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_j} = 0, \quad j = 1, \dots, n. \tag{2}$$

The argument of the time derivative in this expression, $\partial L/\partial \dot{q}_j$, is called the *canonical momentum* p_j for each j. The set of equations (2) involves the combination of partial derivatives of the Lagrangian with respect to the coordinates and velocities, together with the total derivative with respect to time. By the chain rule, this total derivative affects explicit time dependence in L and the implicit time dependence in each q_j and $\dot{q}_j$. Many of the conservation laws involving Lagrangians involve such an interplay of explicit and implicit time dependence.

Any transformation of the coordinates q_j that does not change the Lagrangian is a symmetry of the system. If L does not have explicit dependence on a coordinate q_j, then the first term in (2) vanishes: $\mathrm{d}p_j/\mathrm{d}t = 0$, i.e., the corresponding canonical momentum is conserved in time. In the example from the last section of a particle in a linear gravitational field, the coordinates can be chosen to be Cartesian x, y, z, or cylindrical polars r, ϕ, z; L is independent of x and y, leading to conservation of horizontal momentum, and also ϕ, leading to conservation of angular momentum about the z-axis. The theorem is proved for symmetries of this type in CLASSICAL MECHANICS [IV.19 §2.3]: if a system is homogeneous in space (translation invariant), then linear momentum is conserved, and if it is isotropic (independent of rotations, such as the Newtonian gravitational potential around a massive point particle exerting a central force), then angular momentum is conserved (equivalent to Kepler's second law of planetary motion for gravity).

Since it is the equations (2) that represent the physical laws rather than the form of L or S, the system may admit a more general kind of symmetry whose transformation adds a time-dependent function to the Lagrangian L. If, under the transformation, the Lagrangian transforms $L \to L + \mathrm{d}\Lambda/\mathrm{d}t$ involving the total time derivative of some function Λ, the action transforms $S \to S + \Lambda(t_1) - \Lambda(t_0)$. Thus the transformed action is still made stationary by functions satisfying (2), so transformations of this kind are symmetries of the system, which are also continuous if Λ also depends on a continuous parameter s so that its time derivative is

Copyrighted material

zero when $s = 0$. It is not then difficult to see that the quantity

$$\sum_{j=1}^{n} p_j \frac{\partial q_j}{\partial s}\Big|_{s=0} - \frac{\partial \Lambda}{\partial s}\Big|_{s=0}, \tag{3}$$

defined in terms of the generators of the transformation on each coordinate and the Lagrangian, is constant in time. This is Noether's theorem for classical mechanics.

An important example is when the Lagrangian has no explicit dependence on time, $\partial L/\partial t = 0$. In this case, under an infinitesimal time translation $t \to t + \delta t$, $L \to L + \delta t \mathrm{d}L/\mathrm{d}t$, so here Λ is $L\delta t$, with L evaluated at t, and δt plays the role of s. Under the same infinitesimal transformation, $q_j \to q_j + \delta t \dot{q}_j$, so the relevant conserved quantity (3) is $\sum_j p_j \dot{q}_j - L$, which is the Hamiltonian of the system, which is equal to the total energy in many systems of interest. It is apparently a fundamental law of physics that the total energy in physical processes is conserved in time; energy can be in other forms such as electromagnetic, gravitational, or heat, as well as mechanical. Noether's theorem states that the law of conservation of energy is equivalent to the fact that the physical laws of the system, characterized by their Lagrangian, do not change with time.

The vanishing of the action functional's integrand (i.e., the Lagrangian L) is equivalent to the existence of a first integral for the system of Lagrange equations, which is interpreted in the mechanical setting as a constant of the motion of the system. In this sense, Noether's theorem may be applied more generally in other physical situations described by functionals whose physical laws are given by the corresponding Euler–Lagrange equations. In the case of the Lagrangian approach applied to fields (i.e., functions of space and time), Noether's theorem generalizes to give a continuous density ρ (such as mass or charge density) and a flux vector $\boldsymbol{J}$ satisfying the continuity equation $\dot{\rho} + \nabla \cdot \boldsymbol{J} = 0$ at every point in space and time.

3 Galilean Relativity

Newton's first law of motion can be paraphrased as "all inertial frames, traveling at uniform linear velocity with respect to each other, are equivalent for the formulation of mechanics"—that is, without action of external forces, a system will behave in the same way regardless of the motion of its center of mass. The behavior of a mechanical system is therefore independent of its overall velocity; this is a consequence of Newton's second law, that force is proportional to acceleration. According to pre-Newtonian physics, forces were thought to be proportional to *velocity* (as the effect of friction was not fully appreciated), and it was not until Galileo's thought experiments in friction-free environments that the proportionality of force to *acceleration* was appreciated. In spite of Galilean invariance, problems involving circular motion do in fact seem to require a privileged frame of reference, called *absolute space*. One example due to Newton himself is the problem of explaining, without absolute space, the meniscus formed by the surface of water in a spinning bucket; such problems are properly overcome only in general relativity.

With Galilean relativity, absolute position is no longer defined: events occurring at the same position but at different times in one frame (such as a moving train carriage) occur at different positions in other frames (such as the frame of the train track). However, changes to the state of motion, i.e., accelerations, have physical consequences and are related to forces. This is an example of a *covariance principle*, whose importance for physical theories was emphasized by Einstein. According to this principle, from the statement of physical laws in one frame of reference (such as the laws of motion), one can derive their statement in a different frame of reference from the application of the appropriate transformation rule between reference frames. The statement in the new frame should have the same *mathematical form* as in the previous frame, although quantities may not take the same values in different frames.

Transformations between different inertial frames are represented mathematically in a similar way to the translations of (1); events are labeled by their positions in space and time, such as (t, x) in one frame and (t, x') in another moving at velocity v with respect to the first. Since $x' = x - vt$, the transformation from (t, x) to (t, x') is represented by the matrix $\left(\begin{smallmatrix} 1 & 0 \\ -v & 1 \end{smallmatrix}\right)$. This Galilean transformation (or *Galilean boost*) differs from (1) in that time t is here appended to the position vector, since the translation from the boost is time-dependent. Galilean boosts in three spatial dimensions, together with regular translations and rotations, define the *Galilean group*. It can be shown that the Lagrangian of a free particle follows directly from the covariance of the corresponding action under the Galilean group.

Infinitesimal velocity boosts generate a Noetherian symmetry on systems of particles interacting via forces that depend only on the positions of the others. Consider N point particles of mass m_k and position $\boldsymbol{r}_k$ such that V depends only on $\boldsymbol{r}_k - \boldsymbol{r}_\ell$ for $k, \ell = 1, \ldots, N$. Under

Copyrighted material

Neuenschwander, D. E. 2010. *Emmy Noether's Wonderful Theorem*. Baltimore, MD: Johns Hopkins University Press.
Penrose, R. 2004. *The Road to Reality* (chapters 17–20 are particularly relevant). London: Random House.

II.22 The Jordan Canonical Form

Nicholas J. Higham

A canonical form for a class of matrices is a form of matrix—usually chosen to be as simple as possible—to which all members of the class can be reduced by transformations of a specified kind. The Jordan canonical form (JCF) is associated with similarity transformations on $n \times n$ matrices. A similarity transformation of a matrix A is a transformation from A to $X^{-1}AX$, where X is nonsingular. The JCF is the simplest form that can be achieved by similarity transformations, in the sense that it is the closest to a diagonal matrix.

The JCF of a complex $n \times n$ matrix A can be written $A = ZJZ^{-1}$, where Z is nonsingular and the Jordan matrix J is a block-diagonal matrix

$$\begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_p \end{bmatrix}$$

with diagonal blocks of the form

$$J_k = J_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix}.$$

Here, blanks denote zero blocks or zero entries. The matrix J is unique up to permutation of the diagonal blocks, but Z is not. Each λ_k is an eigenvalue of A and may appear in more than one Jordan block. All the EIGENVALUES [I.2 §20] of the Jordan block J_k are equal to λ_k. By definition, an eigenvector of J_k is a nonzero vector x satisfying $J_k x = \lambda_k x$, and all such x are nonzero multiples of the vector $x = [1\ 0\ \cdots\ 0]^{\mathrm{T}}$. Therefore J_k has only one linearly independent eigenvector. Expand x to a vector $\tilde{x}$ with n components by padding it with zeros in positions corresponding to each of the other Jordan blocks J_i, $i \neq k$. The vector $\tilde{x}$ has a single 1, in the rth component, say. A corresponding eigenvector of A is $Z\tilde{x}$, since $A(Z\tilde{x}) = ZJZ^{-1}(Z\tilde{x}) = ZJ\tilde{x} = \lambda_k Z\tilde{x}$; this eigenvector is the rth column of Z.

If every block J_k is 1×1 then J is diagonal and A is similar to a diagonal matrix; such matrices A are called *diagonalizable*. For example, real symmetric matrices are diagonalizable—and moreover the eigenvalues are real and the matrix Z in the JCF can be taken to be orthogonal. A matrix that is not diagonalizable is *defective*; such matrices do not have a complete set of linearly independent eigenvectors or, equivalently, their Jordan form has at least one block of dimension 2 or greater.

To give a specific example, the matrix

$$A = \frac{1}{2}\begin{bmatrix} 3 & 1 & 1 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

has a JCF with

$$Z = \begin{bmatrix} 0 & \frac{1}{2} & 1 \\ -1 & -\frac{1}{2} & 0 \\ 1 & 0 & 0 \end{bmatrix}, \qquad J = \left[\begin{array}{c|cc} \frac{1}{2} & 0 & 0 \\ \hline 0 & 1 & 1 \\ 0 & 0 & 1 \end{array}\right].$$

As the partitioning of J indicates, there are two Jordan blocks: a 1×1 block with eigenvalue $\frac{1}{2}$ and a 2×2 block with eigenvalue 1. The eigenvalue $\frac{1}{2}$ of A has an associated eigenvector equal to the first column of Z. For the double eigenvalue 1 there is only one linearly independent eigenvector, namely the second column, z_2, of Z. The third column, z_3, of Z is a generalized eigenvector: it satisfies $Az_3 = z_2 + z_3$.

The JCF provides complete information about the eigensystem. The *geometric multiplicity* of an eigenvalue, defined as the number of associated linearly independent eigenvectors, is the number of Jordan blocks in which that eigenvalue appears. The *algebraic multiplicity* of an eigenvalue, defined as its multiplicity as a zero of the characteristic polynomial $q(t) = \det(tI - A)$, is the number of copies of the eigenvalue among all the Jordan blocks. For the matrix (1) above, the geometric multiplicity of the eigenvalue 1 is 1 and the algebraic multiplicity is 2, while the eigenvalue $\frac{1}{2}$ has geometric and algebraic multiplicities both equal to 1.

The *minimal polynomial* of a matrix is the unique monic polynomial ψ of lowest degree such that $\psi(A) = 0$. The degree of ψ is certainly no larger than n because the CAYLEY–HAMILTON THEOREM [IV.10 §5.3] states that $q(A) = 0$. The minimal polynomial of an $m \times m$ Jordan block $J_k(\lambda_k)$ is $(t - \lambda_k)^m$. The minimal polynomial of A is therefore given by

$$\psi(t) = \prod_{i=1}^{s} (t - \lambda_i)^{m_i},$$

Copyrighted material

where $\lambda_1, \ldots, \lambda_s$ are the distinct eigenvalues of A and m_i is the dimension of the largest Jordan block in which λ_i appears. An $n \times n$ matrix is *derogatory* if the minimal polynomial has degree less than n. This is equivalent to some eigenvalue appearing in more than one Jordan block. The matrix A in (1) is defective but not derogatory. The $n \times n$ identity matrix is derogatory for $n > 1$: it has characteristic polynomial $(t-1)^n$ and minimal polynomial $t - 1$.

Two questions that arise in many situations are, "Do the powers of the matrix A converge to zero?" and "Are the powers of A bounded?" The answers to both questions are easily obtained using the JCF. If $A = ZJZ^{-1}$ then $A^2 = ZJZ^{-1} \cdot ZJZ^{-1} = ZJ^2Z^{-1}$ and, in general, $A^k = ZJ^kZ^{-1}$. Therefore the powers of A converge to zero precisely when the powers of J converge to zero, and this in turn holds when the powers of each individual Jordan block converge to zero. The powers of a 1×1 Jordan block $J_i = (\lambda_i)$ obviously converge to zero when $|\lambda_i| < 1$. In general, since $J_i(\lambda_i)^k$ has diagonal elements λ_i^k, for the powers of $J_k(\lambda_k)$ to converge to zero it is necessary that $|\lambda_k| < 1$, and this condition turns out to be sufficient. Therefore $A^k \to 0$ as $k \to \infty$ precisely when $\rho(A) < 1$, where ρ is the *spectral radius*, defined as the largest absolute value of any eigenvalue of A.

Turning to the question of whether the powers of A are bounded, by the argument in the previous paragraph it suffices to consider an individual Jordan block. The powers of $J_k(\lambda_k)$ are clearly bounded when $|\lambda_k| < 1$, as we have just seen, and unbounded when $|\lambda_k| > 1$. When $|\lambda_k| = 1$ the powers are bounded if the block is 1×1, but they are unbounded for larger blocks. For example, $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]^k = \left[\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right]$, which is unbounded as $k \to \infty$. The conclusion is that the powers of A are bounded as long as $\rho(A) \leqslant 1$ and any eigenvalues of modulus 1 are in Jordan blocks of size 1. Thus the powers of A in (1) are not bounded.

In one sense, defective matrices—those with nontrivial Jordan structure—are very rare because the diagonalizable matrices are dense in the set of all matrices. Therefore if you generate matrices randomly you will be very unlikely to generate one that is not diagonalizable (this is true even if you generate matrices with random integer entries). But in another sense, defective matrices are quite common. Certain types of BIFURCATIONS [IV.21] in dynamical systems are characterized by the presence of nontrivial Jordan blocks in the Jacobian matrix, while in problems where some function of the eigenvalues of a matrix is optimized the optimum often occurs at a defective matrix.

While the JCF provides understanding of a variety of matrix problems, it is not suitable as a computational tool. The JCF is not a continuous function of the entries of the matrix and can be very sensitive to perturbations. For example, for $\varepsilon \neq 0$, $\left[\begin{smallmatrix} 1 & \varepsilon \\ 0 & 1 \end{smallmatrix}\right]$ (one Jordan block) and $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ (two Jordan blocks) have different Jordan structures, even though the matrices can be made arbitrarily close by taking ε sufficiently small. In practice, it is very difficult to compute the JCF in floating-point arithmetic due to the unavoidable perturbations caused by rounding errors. As a general principle, the SCHUR DECOMPOSITION [IV.10 §5.5] is preferred for practical computations.

II.23 Krylov Subspaces

Valeria Simoncini

1 Definition and Properties

The mth Krylov subspace of the matrix $A \in \mathbb{C}^{n \times n}$ and the vector $v \in \mathbb{C}^n$ is

$$\mathcal{K}_m(A, v) = \operatorname{span}\{v, Av, \ldots, A^{m-1}v\}.$$

The dimension of $\mathcal{K}_m(A, v)$ is at most m, and it is less if an invariant subspace of A with respect to v is obtained for some $m_* < m$. In general, $\mathcal{K}_m(A, v) \subseteq \mathcal{K}_{m+1}(A, v)$ (the spaces are nested); if $m_* = n$, then $\mathcal{K}_n(A, v)$ spans the whole of $\mathbb{C}^n$.

Let $v_1 = v/\|v\|_2$, with $\|v\|_2 = (v^*v)^{1/2}$ the 2-norm, and let $\{v_1, v_2, \ldots, v_m\}$ be an orthonormal basis of $\mathcal{K}_m(A, v)$. Setting $\mathcal{V}_m = [v_1, v_2, \ldots, v_m]$, from the nesting property it follows that the next basis vector v_{m+1} can be computed by the following Arnoldi relation:

$$A\mathcal{V}_m = [\mathcal{V}_m, v_{m+1}]H_{m+1,m},$$

where $H_{m+1,m} \in \mathbb{C}^{(m+1)\times m}$ is an *upper Hessenberg* matrix (upper triangular plus nonzero entries immediately below the diagonal) whose columns contain the coefficients that make v_{m+1} orthogonal to the already available basis vectors $v_1, \ldots, v_m$.

Suppose we wish to approximate a vector y by a vector $x \in \mathcal{K}_m(A, v)$, measuring error in the 2-norm. Any such x can be written as a polynomial in A of degree at most $m - 1$ times v: $x = \sum_{i=0}^{m-1} \alpha_i A^i v$. If A is Hermitian, then by using a spectral decomposition we can reduce A to diagonal form by unitary transformations, which do not change the 2-norm, and it is then clear that the eigenvalues of A and the decomposition of v

Copyrighted material

in terms of the eigenvectors of A drive the approximation properties of the space. For non-Hermitian A the approximation error is harder to analyze, especially for highly nonnormal or nondiagonalizable matrices.

By replacing v by an $n \times s$ matrix V, with $s \geqslant 1$, spaces of dimension at most ms can be obtained. An immediate matrix counterpart is

$$\mathbb{K}_m(A, V) = \left\{ \sum_{i=0}^{m-1} \gamma_i A^i V : \gamma_i \in \mathbb{C} \text{ for all } i \right\}.$$

A richer version is obtained by working with linear combinations of all the available vectors:

$$\mathbb{K}_m^{\square}(A, V) = \text{range}([V, AV, \ldots, A^{m-1}V]).$$

Methods based on this latter space are called "block" methods, since all matrix structure properties are generalized to blocks (e.g., $H_{m+1,m}$ will be block upper Hessenberg, with $s \times s$ blocks). Block spaces are appropriate, for instance, in the presence of multiple eigenvalues or if the original application requires using the same A and different vectors v.

2 Applications and Generalizations

Krylov subspaces are used in projection methods for solving large algebraic linear systems, eigenvalue problems, and matrix equations; for approximating a wide range of matrix functions (analytic functions, trace, determinant, transfer functions, etc.); and in model order reduction.

The general idea is to project the original problem of size n onto the Krylov subspace of dimension $m \ll n$ and then solve the smaller $m \times m$ reduced problem with a more direct method (one that would be too computationally expensive if applied to the original $n \times n$ problem). If the Krylov subspace is good enough, then the projected problem retains sufficient information from the original problem that the sought after quantities are well approximated.

When equations are involved, Krylov subspaces usually play a role as approximation spaces, as well as test spaces. The actual test space used determines the resulting method and influences the convergence properties.

Generalized spaces have emerged as second-generation Krylov subspaces. In the eigenvalue context, the "shift-and-invert" Krylov subspace $\mathcal{K}_m((A - \sigma I)^{-1}, v)$ is able to efficiently approximate eigenvalues in a neighborhood of a fixed scalar $\sigma \in \mathbb{C}$; here I is the identity matrix of size n. In matrix function evaluations and matrix equations, the extended space $\mathcal{K}_m(A, v) + \mathcal{K}_m(A^{-1}, A^{-1}v)$ has shown some advantages over the classical space, while for $\sigma_1, \ldots, \sigma_m \in \mathbb{C}$, the use of the more general *rational* space

$$\text{span}\{(A - \sigma_1 I)^{-1}v, \ldots, (A - \sigma_m I)^{-1}v\}$$

has recently received a lot of attention for its potential in a variety of advanced applications beyond eigenvalue problems, where it was first introduced in the 1980s. All these generalized spaces require solving systems with some shifted forms of A, so that they are in general more expensive to build than the classical one, depending on the computational cost involved in solving these systems. However, the computed space is usually richer in spectral information, so that a much smaller space dimension is required to satisfactorily approximate the requested quantities. The choice among these variants thus depends on the spectral and sparsity properties of the matrix A.

Further Reading

Freund, R. W. 2003. Model reduction methods based on Krylov subspaces. *Acta Numerica* 12:126–32.

Liesen, J., and Z. Strakoš. 2013. *Krylov Subspace Methods: Principles and Analysis*. Oxford: Oxford University Press.

Saad, Y. 2003. *Iterative Methods for Sparse Linear Systems*, 2nd edn. Philadelphia, PA: SIAM.

Watkins, D. S. 2007. *The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods*. Philadelphia, PA: SIAM.

II.24 The Level Set Method

Fadil Santosa

1 The Basic Idea

The level set method is a numerical method for representing a closed curve or surface. It uses an implicit representation of the geometrical object in question. It has found widespread use in problems where the closed curve evolves or needs to be recomputed often. A main advantage of the method is that such a representation is very flexible and calculation can be done on a regular grid. In computations where surfaces evolve, changes in the topology of the surface are easily handled.

Consider an example in two dimensions in the (x, y)-plane. Suppose one is interested in the motion of a curve under external forcing terms. Let $C(t)$ denote the curve as a function of time t. One method for solving this problem is to track the curve, which can be done by choosing marker points, $(x_i(t), y_i(t)) \in C(t)$,

Copyrighted material

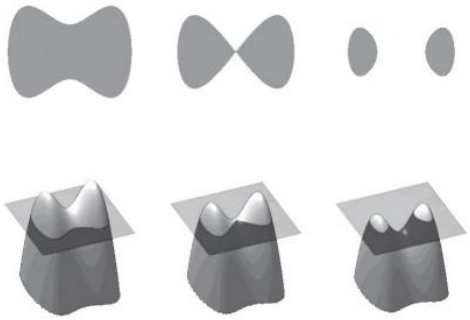


Figure 1 The graphs of a level set function $z = \varphi(x, y, t)$ for three values of t are shown at the bottom of the figure. The plane $z = 0$ intersects these functions. The domains $D(t) = \{(x, y)\colon \varphi(x, y, t) > 0\}$ are shown above. Note that, in this example, $D(t)$ has gone through a topological change as t varies.

$i = 1, \ldots, n$, whose motions are determined by the forcing. The curve itself may be recovered at any time t by a prescribed spline interpolation.

The level set method takes a different approach; it represents the curve as the zero level set of a function $\varphi(x, y, t)$. That is, the curve is given by

$$C(t) = \{(x, y)\colon \varphi(x, y, t) = 0\}.$$

One can set up the function so that the interior of the curve $C(t)$ is the set

$$D(t) = \{(x, y)\colon \varphi(x, y, t) > 0\}.$$

In figure 1 the level set function $z = \varphi(x, y, t)$ can be seen to intersect the plane $z = 0$ at various times t. The sets of $D(t)$ (not up to the same scale) are shown above each three-dimensional figure.

An advantage of the level set method is demonstrated in the figure. One can see that a topological change in $D(t)$ has occurred as t is varied. The level set method allows for such a change without the need to redefine the representation, as would be the case for the front-tracking method described previously.

2 Discretization

One of the attractive features of the level set method is that calculations are done on a regular Cartesian grid. Suppose we have discretized the computational domain and the nodes are at coordinates (x_i, y_j) for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The values of the level set function $\varphi(x, y, t)$ are then stored at coordinate points $x = x_i$ and $y = y_j$. At any time, if one is interested in the curve $C(t)$, the zero level set, the set

$$C(t) = \{(x, y)\colon \varphi(x, y, t) = 0\}$$

needs to be approximated from the data $\varphi(x_i, y_j, t)$. One is typically interested in such quantities as the normal to the curve and the curvature at a point on the curve. These quantities are easily calculated by evaluating finite-difference approximations of

$$\nu = \frac{\nabla\varphi}{|\nabla\varphi|}$$

and

$$\kappa = \nabla \cdot \nu,$$

where the gradient operator $\nabla = [\partial/\partial x \ \partial/\partial y]^{\mathrm{T}}$.

In practice, it is not necessary to keep all values of the level set function on the nodes. Since one is often interested only in the motion of the curve $C(t)$, the zero level set, one needs only the values of the level set function in the neighborhood of the curve. Such approaches have been dubbed "narrow-band methods" and can potentially reduce the amount of computation in a problem involving complex evolution of surfaces.

It must be noted that, in the two-dimensional example here, $C(t)$ is a one-dimensional object, whereas the level set function $\varphi(x, y, t)$ is a two-dimensional function. Thus, one might say that the ability to track topological changes is made at the cost of increased computational complexity.

3 Applications

A simple problem one may pose is that of tracking the motion of a curve for which every point on the curve is moving in the direction normal to the curve with a given velocity. If the velocity is v, then the equation for the level sets is given by

$$\frac{\partial\varphi}{\partial t} = v|\nabla\varphi|.$$

If one is interested in tracking the motion of the zero level set $C(t)$, then one must specify an initial condition

$$\varphi(x, y, 0) = \varphi_0(x, y),$$

where the initial zero level set is given by

$$C(0) = \{(x, y)\colon \varphi_0(x, y) = 0\}.$$

This evolution of such a curve may be very complicated and go through topological changes. The power of the level set method is demonstrated here because all one needs to do is solve the initial-value problem for $\varphi(x, y, t)$.

Another simple problem is the classical motion by mean curvature. In this "flow," one is interested in tracking the motion of a curve for which every point on the curve is moving normal to the curve at a velocity

Copyrighted material

proportional to the curvature. The evolution equation is given by

$$\frac{\partial \varphi}{\partial t} = \nabla \cdot \frac{\nabla \varphi}{|\nabla \varphi|}.$$

A numerical solution of this evolution equation can be used to demonstrate the classical Grayson theorem, which asserts that, if the closed curve starts out without self-intersections, then it will never form self-intersections and it will become convex in finite time.

Significant problems arising from applications from diverse fields have benefited from the level set treatment. The following is an incomplete list meant to give a sense of the range of applications.

Image processing. The level set method can be used for segmentation of objects in a two-dimensional scene. It has also been demonstrated to be effective in modeling surfaces from point clouds.

Fluid dynamics. Two-phase flows, which involve interfaces separating the two phases, can be approached by the level set method. It is particularly effective for problems in which one of the phases is dispersed in bubbles.

Inverse problems. Inverse problems exist in which the unknown that one wishes to reconstruct from data is the boundary of an object. Examples include inverse scattering.

Optimal shape design. When the object is to design a shape that maximizes certain attributes (design objectives), it is often very convenient to represent the shape by a level set function.

Computer animation. The need for physically based simulations in the animation industry has been partially met by solving equations of physics using the level set method to represent surfaces that are involved in the simulation.

Current research areas include improved accuracy in the numerical schemes employed and in applying the method to ever more complex physics.

Further Reading

Osher, S., and R. Fedkiw. 2003. *Level Set Methods and Dynamic Implicit Surfaces.* New York: Springer.

Osher, S., and J. A. Sethian. 1988. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics* 79:12–49.

Sethian, J. A. 1999. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science.* Cambridge: Cambridge University Press.

II.25 Markov Chains

Beatrice Meini

A Markov chain is a type of random process whose behavior in the future is influenced only by its current state and not by what happened in the past. A simple example is a random walk on $\mathbb{Z}$, where a particle moves among the integers of the real line and is allowed to move one step forward with probability $0 < p < 1$ and one step backward with probability $q = 1 - p$ (see figure 1). The position of the particle at time $n + 1$ (in the future) depends on the position of the particle at time n (at the present time), and what happened before time n (in the past) is irrelevant.

To give a precise definition of a Markov chain we will need some notation. Let E be a countable set representing the states, and let Ω be a set that represents the sample space. Let $X, Y \colon \Omega \to E$ be two random variables. We denote by $\mathbb{P}[X = j]$ the probability that X takes the value j, and we denote by $\mathbb{P}[X = j \mid Y = i]$ the probability that X takes the value j given that the random variable Y takes the value i. A discrete stochastic process is a family $(X_n)_{n\in\mathbb{N}}$ of random variables $X_n \colon \Omega \to E$.

A stochastic process $(X_n)_{n\in\mathbb{N}}$ is called a *Markov chain* if

$$\mathbb{P}[X_{n+1} = i_{n+1} \mid X_0 = i_0, \dots, X_n = i_n] = \mathbb{P}[X_{n+1} = i_{n+1} \mid X_n = i_n]$$

at any time $n \geqslant 0$ and for any states $i_0, \dots, i_{n+1} \in E$. This means that the state X_n at time n is sufficient to determine which state X_{n+1} might be occupied at time $n + 1$, and we may forget the past history $X_0, \dots, X_{n-1}$.

It is often required that the laws that govern the evolution of the system be time invariant. The Markov chain is said to be *homogeneous* if the transitions from one state to another state are independent of the time n, i.e., if

$$\mathbb{P}[X_{n+1} = j \mid X_n = i] = p_{ij}$$

at any time $n \geqslant 0$ and for any states $i, j \in E$. The number p_{ij} represents the probability of passing from state i to state j in one time step. The matrix $P = (p_{ij})_{i,j\in E}$ is called the *transition matrix* of the Markov chain. The matrix P is a *stochastic matrix*: that is, it has nonnegative entries and unit row sums ($\sum_{j\in E} p_{ij} = 1$ for all $i \in E$). The dynamic behavior of the Markov chain is governed by the transition matrix P. In particular, the problem of computing the probability that, after n

Copyrighted material

of (1), which in the linear time-invariant case is obtained by taking Laplace transforms and inserting the transformed equation (1a) into the transformed (1c). The transfer function represents the mapping of inputs u to outputs y. As a rational matrix-valued function of a complex variable, it can be approximated in different ways. In rational interpolation methods, V, W are computed so that

$$\frac{\mathrm{d}^j}{\mathrm{d}s^j} G(s_k) = \frac{\mathrm{d}^j}{\mathrm{d}s^j} \hat{G}(s_k), \quad k = 1, \ldots, K,\ j = 0, \ldots, J_k,$$

for K interpolation points s_k and derivatives up to order J_k at each point. Here, $\hat{G}$ denotes the transfer function of (2) and (3), defined by $\hat{A} = W^{\mathrm{T}}AV$, $\hat{B} = W^{\mathrm{T}}B$, and $\hat{C} = CV$.

In the nonlinear case, a further question is how to obtain functions $\hat{f}$ and $\hat{g}$ allowing for fast evaluation. Simply setting $\hat{f}(t, x, u, p) = W^{\mathrm{T}} f(t, V\hat{x}, u, p)$ obviously does not lead to faster simulation in general. Therefore, dedicated methods, such as (discrete) empirical interpolation, are needed to obtain a "reduced" $\hat{f}$ and $\hat{g}$.

2 An Example

As an example consider the mathematical model of a microgyroscope: a device used in stability control of vehicles. Finite-element discretization of this particular model leads to a linear time-invariant system of $n = 34\,722$ linear ordinary differential equations with $d = 4$ parameters, $m = 1$ input, and $q = 12$ outputs. Using a reduced-order model of size $r = 289$, a parameter study involving two parameters (defining x- and y-axes in figure 1) and the excitation frequency ω (i.e., the parametric transfer function $G(s, p)$ is evaluated for $s = \mathrm{i}\omega$ with varying ω) could be accelerated by a factor of approximately 90 without significant loss of accuracy. The output y was computed with an error of less than 0.01% in the whole frequency and parameter domain. Figure 1 shows the response surfaces of the full and reduced-order models at one frequency for variations of two parameters.

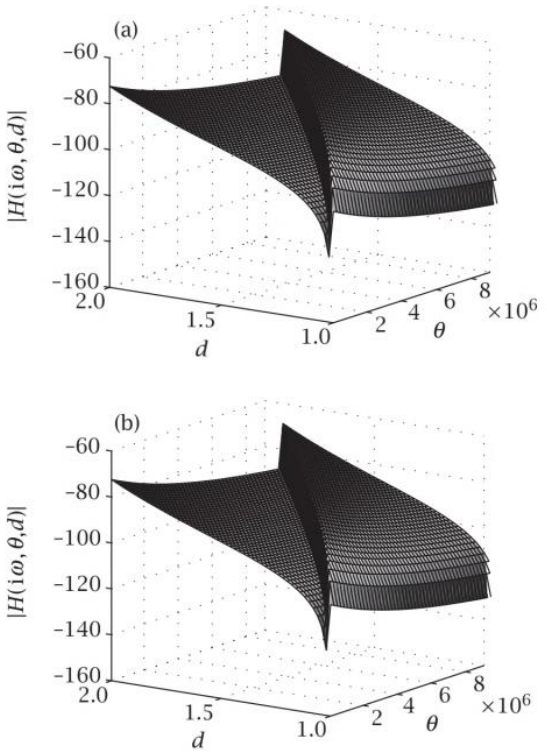


Figure 1 The parametric transfer function of a microdevice (at $\omega = 0.025$): results from (a) the full model with dimension 34 722 and (b) the reduced-order model with dimension 289. (Computations and graphics by L. Feng and T. Breiten.)

Further Reading

Antoulas, A. 2005. *Approximation of Large-Scale Dynamical Systems.* Philadelphia, PA: SIAM.

Benner, P., M. Hinze, and E. J. W. ter Maten, eds. 2011. *Model Reduction for Circuit Simulation.* Lecture Notes in Electrical Engineering, volume 74. Dordrecht: Springer.

Benner, P., V. Mehrmann, and D. C. Sorensen, eds. 2005. *Dimension Reduction of Large-Scale Systems.* Lecture Notes in Computational Science and Engineering, volume 45. Berlin: Springer.

Schilders, W. H. A., H. A. van der Vorst, and J. Rommes, eds. 2008. *Model Order Reduction: Theory, Research Aspects and Applications.* Mathematics in Industry, volume 13. Berlin: Springer.

II.27 Multiscale Modeling

Fadil Santosa

To accurately model physical, biological, and other phenomena, one is often confronted with the need to capture complex interactions occurring at distinct temporal and spatial scales. In the language of multiscale modeling, temporal scales are usually differentiated by slow, medium, and fast timescales. Spatially, the phenomena are separated into micro-, meso-, and

Copyrighted material

macroscales. In modeling the deformation of solids, for instance, the microscale phenomena could be atomistic interactions occurring on a femtosecond (10^{-15} second) timescale. At the mesoscopic scale, one could be interested in the behavior of the constituent macromolecules, e.g., a tangled bundle of polymers. Finally, at the macroscopic scale, one might be interested in how a body, whose size could be in meters, deforms under an applied force. The challenge in multiscale modeling is that the interactions at one scale communicate with interactions at other scales. Thus, in the example given, the question we wish to answer is how the applied forces affect the atomistic interactions, and how those interactions impact the behavior of the macromolecules, which in turn affects how the overall shape of the body deforms.

Multiscale modeling is a rapidly developing field because of its enormous importance in applications. The range of applications is staggering. It has been applied in geophysics, biology, chemistry, meteorology, materials science, and physics.

We give another concrete example that arises in solid mechanics. Suppose we have a block of pure aluminum whose crystalline structure is known. How can we calculate its elastic properties, i.e., its Lamé modulus and Poisson ratio, *ab initio* from knowledge of its atomistic structure? Such a calculation would start by considering Schrödinger's equation for the multiparticle system. By solving for the ground states of the system, one can then extract the desired macroscopic properties of the bulk aluminum.

A classical example of multiscale modeling in applied mathematics is the HOMOGENIZATION METHOD [II.17], which allows for extraction of effective properties of composite materials. Consider the steady-state distribution of temperature in a rod of length ℓ made up of a material with rapidly oscillating conductivity. The conductivity is described by a periodic function $a(y) > 0$, such that $a(y+1) = a(y)$. A small-scale ε is introduced to denote the actual period in the medium. The governing equation for temperature $u(x)$ is

$$\left(a\left(\frac{x}{\varepsilon}\right)u'\right)' = f, \quad 0 < x < \ell,$$

where a prime denotes differentiation with respect to x. Here, f is the heat source distribution, with x measuring distance along the rod. To solve the problem, the solution u is developed in powers of ε. The macroscopic behavior of u is identified with the zeroth order. This solution will be smooth as the small rapid oscillations are ignored.

Current research in multiscale modeling focuses on bridging the phenomena at the different scales and developing efficient numerical methods. There are efforts to develop rigorous multiscale models that agree with their continuum counterparts. CONTINUUM MODELS [IV.26] are macroscale models derived from first principles and where the material properties are usually measured. Other efforts concentrate more on developing accurate simulations, such as modeling the properties of Kevlar starting from the polymers in the resin and the carbon fibers used. All research in this area involves some numerical analysis and scientific computing.

II.28 Nonlinear Equations and Newton's Method

Marcos Raydan

Nonlinear equations appear frequently in the mathematical modeling of real-world processes. They are usually written as a zero-finding problem: find $x_j \in \mathbb{R}$, for $j = 1, 2, \ldots, n$, such that

$$f_i(x_1, \ldots, x_n) = 0 \quad \text{for } i = 1, 2, \ldots, n,$$

where the f_i are given functions of n variables. This system of equations is nonlinear if at least one of the functions f_i depends nonlinearly on at least one of the variables. Using vector notation, the problem can also be written as find $x = [x_1, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^n$ such that

$$F(x) = [f_1(x), \ldots, f_n(x)]^{\mathrm{T}} = 0.$$

If every function f_i depends linearly on all the variables, then it is usually written as a linear system of equations $Ax = b$, where $b \in \mathbb{R}^n$ and A is an $n \times n$ matrix.

The existence and uniqueness of solutions for nonlinear systems of equations is more complicated than for linear systems of equations. For solving $Ax = b$, the number of solutions must be either zero, infinity, or one (when A is nonsingular), whereas $F(x) = 0$ can have zero, infinitely many, or any finite number of solutions. Fortunately, in practice, it is usually sufficient to find a solution of the nonlinear system for which a reasonable initial approximation is known.

Even in the simple one-dimensional case ($n = 1$), most nonlinear equations cannot be solved by a closed formula, i.e., using a finite number of operations. A well-known exception is the problem of finding the roots of polynomials of degree less than or equal to four, for which closed formulas have been known for

Copyrighted material

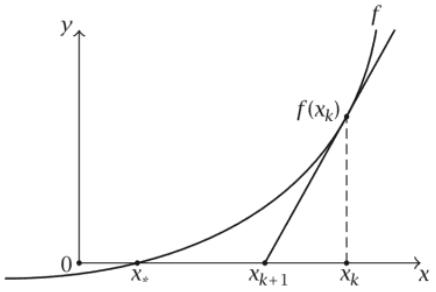


Figure 1 One iteration of Newton's method in one dimension for $f(x) = 0$.

centuries. As a consequence, in general iterative methods must be used to produce increasingly accurate approximations to the solution. One of the oldest iterative schemes, which has played an important role in the numerical methods literature for solving $F(x) = 0$, is Newton's method.

1 Newton's Method

Newton's method for solving nonlinear equations was born in one dimension. In that case, the problem is find $x \in \mathbb{R}$ such that $f(x) = 0$, where $f\colon \mathbb{R} \to \mathbb{R}$ is differentiable in the neighborhood of a solution x_*. Starting from a given x_0, on the kth iteration Newton's method constructs the tangent line passing through the point $(x_k, f(x_k))$,

$$M_k(x) = f(x_k) + f'(x_k)(x - x_k),$$

and defines the next iterate, x_{k+1}, as the root of the equation $M_k(x) = 0$ (see figure 1). Hence, from a given $x_0 \in \mathbb{R}$, Newton's method generates the sequence $\{x_k\}$ of approximations to x_* given by

$$x_{k+1} = x_k - f(x_k)/f'(x_k).$$

Notice that the tangent line or linear model $M_k(x)$ is equal to the first two terms of the Taylor series of f around x_k.

Newton's idea in one dimension can be extended to n-dimensional problems. In $\mathbb{R}^n$ the method approximates the solution of a square nonlinear system of equations by solving a sequence of square linear systems. As in the one-dimensional case, on the kth iteration the idea is to define x_{k+1} as a zero of the linear model given by

$$M_k(x) = F(x_k) + J(x_k)(x - x_k),$$

where the map $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is assumed to be differentiable in a neighborhood of a solution x_* and where $J(x_k)$ is the $n \times n$ Jacobian matrix with entries $J_{ij}(x_k) = \partial f_i/\partial x_j(x_k)$ for $1 \leqslant i, j \leqslant n$. Therefore, starting at a given $x_0 \in \mathbb{R}^n$, Newton's method carries out for $k = 1, 2, \ldots$ the following two steps.

- Solve $J(x_k)s_k = -F(x_k)$ for s_k.
- Set $x_{k+1} = x_k + s_k$.

Notice that Newton's method is scale invariant: if the method is applied to the nonlinear system $AF(x) = 0$, for any nonsingular $n \times n$ matrix A, the sequence of iterates is identical to the ones obtained when it is applied to $F(x) = 0$. Another interesting theoretical feature is its impressively fast local convergence. Under some standard assumptions—namely that $J(x_*)$ is nonsingular, $J(x)$ is Lipschitz continuous in a neighborhood of x_*, and the initial guess x_0 is sufficiently close to x_*—the sequence $\{x_k\}$ generated by Newton's method converges *q-quadratically* to x_*; i.e., there exist $c > 0$ and $\hat{k} \geqslant 0$ such that for all $k \geqslant \hat{k}$,

$$\|x_{k+1} - x_*\| \leqslant c\|x_k - x_*\|^2.$$

Hence Newton's method is theoretically attractive, but it may be difficult to use in practice for various reasons, including the need to calculate the derivatives, the need to have a good initial guess to guarantee convergence, and the cost of solving an $n \times n$ linear system per iteration.

2 Practical Variants

If the derivatives are not available, or are too expensive to compute, they can be approximated by finite differences. A standard option is to approximate the jth column of $J(x_k)$ by a forward difference quotient: $(F(x_k + h_k e_j) - F(x_k))/h_k$, where e_j denotes the jth unit vector and $h_k > 0$ is a suitable small number. Notice that, when using this finite-difference variant, the map F needs to be evaluated $n + 1$ times per iteration, once for each column of the Jacobian and one for the vector x_k. Therefore, this variant is attractive when the evaluation of F is not expensive.

Another option is to extend the well-known one-dimensional secant method to the n-dimensional problem $F(x) = 0$. The main idea, in these so-called *secant* or QUASI-NEWTON METHODS [IV.11 §4.2], is to generate not only a sequence of iterates $\{x_k\}$ but also a sequence of matrices $\{B_k\}$ that approximate $J(x_k)$ and satisfy the secant equation $B_k s_{k-1} = y_{k-1}$, where $s_{k-1} = x_k - x_{k-1}$

Copyrighted material

and $y_{k-1} = F(x_k) - F(x_{k-1})$. In this case, an initial matrix $B_0 \approx J(x_0)$ must be supplied. Clearly, infinitely many $n \times n$ matrices satisfy the secant equation. As a consequence, a wide variety of quasi-Newton methods (e.g., Broyden's method) with different properties have been developed.

When using Newton's method, or any of its derivative-free variants, a linear system needs to be solved at each iteration. This linear system can be solved by direct methods (e.g., LU or QR factorization), but if n is large and the Jacobian matrix has a sparse structure, it may be preferable to use an iterative method (e.g., a KRYLOV SUBSPACE METHOD [IV.10 §9]). For that, note that x_k can be used as the initial guess for the solution at iteration $k+1$. One of the important features of these so-called inexact variants of Newton's method is that modern iterative linear solvers do not require explicit knowledge of the Jacobian; instead, they require only the matrix–vector product $J(x_k)z$ for any given vector z. This product can be approximated using a forward finite difference:

$$J(x_k)z \approx (F(x_k + h_k z) - F(x_k))/h_k.$$

Hence, inexact variants of Newton's method are also suitable when derivatives are not available. In all the discussed variants, the local *q-quadratic convergence* is in general lost, but *q-superlinear convergence* can nevertheless be obtained, i.e., $\|x_{k+1} - x_*\|/\|x_k - x_*\| \to 0$.

Finally, in general Newton's method converges only locally, so it requires globalization strategies to be practically effective. The two most popular and best-studied options are line searches and trust regions. In any case, a merit function $\hat{f}\colon \mathbb{R}^n \to \mathbb{R}^+$ must be used to evaluate the quality of all possible iterates. When solving $F(x) = 0$, the natural choice is $\hat{f}(x) = F(x)^{\mathrm{T}}F(x)$.

Further Reading

Dennis, J. E., and R. Schnabel. 1983. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Englewood Cliffs, NJ: Prentice Hall. (Republished by SIAM (Philadelphia, PA) in 1996.)

Kelley, C. T. 2003. *Solving Nonlinear Equations with Newton's Method.* Philadelphia, PA: SIAM.

Ypma, T. J. 1995. Historical development of the Newton–Raphson method. *SIAM Review* 37(4):531–51.

II.29 Orthogonal Polynomials

Polynomials $p_0(x)$, $p_1(x)$, ..., where p_i has degree i, are *orthogonal polynomials* on an interval $[a, b]$ with respect to a nonnegative weight function $w(x)$ if

$$\int_a^b w(x)p_i(x)p_j(x)\,\mathrm{d}x = 0, \quad i \neq j,$$

that is, if all distinct pairs of polynomials are orthogonal on $[a, b]$ with respect to w. For a given weight function and interval, the orthogonality conditions determine the polynomials p_i uniquely up to a constant factor.

Table 1 Parameters in the three-term recurrence (1) for some classical orthogonal polynomials.

Polynomial	$[a, b]$	$w(x)$	a_j	b_j	c_j
Chebyshev	$[-1, 1]$	$(1-x^2)^{-1/2}$	2	0	1
Legendre	$[-1, 1]$	1	$\frac{2j+1}{j+1}$	0	$\frac{j}{j+1}$
Hermite	$(-\infty, \infty)$	e^{-x^2}	2	0	$2j$
Laguerre	$[0, \infty)$	e^{-x}	$-\frac{1}{j+1}$	$\frac{2j+1}{j+1}$	$\frac{j}{j+1}$

An important property of orthogonal polynomials is that they satisfy a three-term recurrence relation

$$p_{j+1}(x) = (a_j x + b_j)p_j(x) - c_j p_{j-1}(x), \quad j \geqslant 1. \quad (1)$$

The weight functions, interval, and recurrence coefficients for some classical orthogonal polynomials are summarized in table 1, in which is assumed the normalization $p_0(x) = 1$, with $p_1(x) = x$ for the Chebyshev and Legendre polynomials, $p_1(x) = 2x$ for the Hermite polynomials, and $p_1(x) = 1 - x$ for the Laguerre polynomials.

Orthogonal polynomials have many interesting properties and find use in many different settings, e.g., in numerical integration, Krylov subspace methods, and the theory of continued fractions. In this volume they arise in LEAST-SQUARES APPROXIMATION [IV.9 §3.3], NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS [IV.13 §6], RANDOM-MATRIX THEORY [IV.24], and as SPECIAL FUNCTIONS [IV.7 §7]. See SPECIAL FUNCTIONS [IV.7 §7] for more information.

II.30 Shocks

Barbara Lee Keyfitz

1 What Are Shocks?

"Shocks" (or "shock waves") is another name for the field of quasilinear hyperbolic PDEs, or CONSERVATION LAWS [II.6]. When the mathematical theory of supersonic flow was in its infancy, the first text on the subject named it this way; and the first modern monograph to

Copyrighted material

focus on the mathematical theory of quasilinear hyperbolic PDEs also used this terminology. Shocks are a dominant feature of the subject, for, as noted with reference to the BURGERS EQUATION [III.4] (see also PARTIAL DIFFERENTIAL EQUATIONS [IV.3 §3.6]), solutions to initial-value problems, even with smooth data, are not likely to remain smooth for all time. We return to the derivation to see what happens when solutions are not differentiable.

In the derivation of a system in one space dimension,

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \tag{1}$$

one typically invokes the conservation of each component u_i of $\boldsymbol{u}$. The rate of change of u_i over a control length $[x, x+h]$ is the net flux across the endpoints:

$$\partial_t \int_x^{x+h} u_i(y,t)\,\mathrm{d}y = f_i(\boldsymbol{u}(x,t)) - f_i(\boldsymbol{u}(x+h,t)). \tag{2}$$

Under the assumption that $\boldsymbol{u}$ is differentiable, the mean value theorem of calculus yields (1) in the limit $h \to 0$. However, (2) is also useful in a different case. If $\boldsymbol{u}$ approaches two different limits, $\boldsymbol{u}_{\mathrm{L}}(X(t),t)$ and $\boldsymbol{u}_{\mathrm{R}}(X(t),t)$, on the left and right sides of a curve of discontinuity, $x = X(t)$, then taking the limit $h \to 0$ in (2) with x and $x+h$ straddling the curve $X(t)$ yields a relationship among $\boldsymbol{u}_{\mathrm{L}}$, $\boldsymbol{u}_{\mathrm{R}}$, and the derivative of the curve:

$$\begin{aligned} X'(t)(\boldsymbol{u}_{\mathrm{R}}(X(t),t) - \boldsymbol{u}_{\mathrm{L}}(X(t),t)) \\ = \boldsymbol{f}(\boldsymbol{u}_{\mathrm{R}}(X(t),t)) - \boldsymbol{f}(\boldsymbol{u}_{\mathrm{L}}(X(t),t)). \end{aligned} \tag{3}$$

This is known as the (generalized) Rankine–Hugoniot relation. The quantity $X'(t)$ measures the speed of propagation of the discontinuity at $X(t)$.

Because solutions of conservation laws are not expected to be continuous for all time, even when the initial data are smooth, it is necessary to allow shocks in any formulation of what is meant by a "solution" of (1). Conservation law theory states that a solution of (1) may contain countably many shocks, the functions $X(t)$ may be no smoother than Lipschitz continuous, and there may be countably many points in physical (x,t)-space at which shock curves intersect. In the case of conservation laws in more than one space dimension, the notion of a "shock curve" can be generalized to that of a "shock surface" by supposing that the solution is piecewise differentiable on each side of such a surface. One obtains an equation similar to (3) that relates the states on either side of the surface to the normal to the surface at each point. However, as distinct from the case in a single space dimension, it is not known whether all solutions have this structure, or whether more singular behavior is possible.

2 Entropy, Admissibility, and Uniqueness

Although allowing for weak solutions, in the form of solutions containing shocks, is forced upon us by both mathematical considerations (they arise from almost all data) and physical considerations (they are seen in all the fluid systems modeled by conservation laws), a new difficulty arises: if shocks are admitted as solutions to a conservation law system, there may be *too many* solutions (this is also known, somewhat illogically, as "lack of uniqueness"). Here is a simple example, involving the Burgers equation. If at $t = 0$ we are given

$$u(x,0) = \begin{cases} 0, & x \leqslant 0, \\ 1, & x > 0, \end{cases}$$

then

$$u(x,t) = \begin{cases} 0, & x \leqslant \frac{1}{2}t, \\ 1, & x > \frac{1}{2}t, \end{cases}$$

is a shock solution in the sense of (3). But

$$u(x,t) = \begin{cases} 0, & x \leqslant 0, \\ x/t, & 0 < x \leqslant t, \\ 1, & x > t, \end{cases}$$

is also a solution, and in fact it is the latter, which is described as a "rarefaction wave," that is correct, while the former, known as a "RAREFACTION SHOCK [V.20 §2.2]," can be ruled out on both mathematical and physical grounds. A fluid that is rarefying (that is, one in which the force of pressure is decreasing), be it a gas or traffic, spreads out gradually and erases the initial discontinuity, while a fluid that is being compressed forms a shock.

Another mode of reasoning, which has both a mathematical and a physical basis, goes as follows. Suppose η is a convex function of $\boldsymbol{u}$ for which another function $q(\boldsymbol{u})$ exists such that

$$\eta(\boldsymbol{u})_t + q(\boldsymbol{u})_x = 0 \tag{4}$$

whenever $\boldsymbol{u}$ is a smooth solution of (1). When this is the case, we say that (1) "admits a convex entropy." A calculation (easy for the Burgers equation and true in general) shows that we should not expect (4) to be satisfied (in the weak sense, as an additional Rankine–Hugoniot relation like (3)) in regions containing shocks. But since η is convex, imposing the requirement that η decrease in time when shocks are present forces a bound on

Copyrighted material

II.32 The Singular Value Decomposition

Nicholas J. Higham

One of the most useful matrix factorizations is the *singular value decomposition* (SVD), which is defined for an arbitrary rectangular matrix $A \in \mathbb{C}^{m\times n}$. It takes the form

$$A = U\Sigma V^*, \quad \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_p) \in \mathbb{R}^{m\times n}, \quad (1)$$

where $p = \min(m, n)$, Σ is a diagonal matrix with diagonal elements $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_p \geqslant 0$, and $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary. The σ_i are the *singular values* of A, and they are the nonnegative square roots of the p largest eigenvalues of A^*A. The columns of U and V are the left and right *singular vectors* of A, respectively.

Postmultiplying (1) by V gives $AV = U\Sigma$ since $V^*V = I$, which shows that the ith columns of U and V are related by $Av_i = \sigma_i u_i$ for $i = 1\colon p$. Similarly, $A^*u_i = \sigma_i v_i$ for $i = 1\colon p$. A geometrical interpretation of the former equation is that the singular values of A are the lengths of the semiaxes of the hyperellipsoid $\{Ax\colon \|x\|_2 = 1\}$.

Assuming that $m \geqslant n$ for notational simplicity, from (1) we have

$$A^*A = V(\Sigma^*\Sigma)V^*, \quad (2)$$

with $\Sigma^*\Sigma = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$, which shows that the columns of V are eigenvectors of the matrix A^*A with corresponding eigenvalues the squares of the singular values of A. Likewise, the columns of U are eigenvectors of the matrix AA^*.

The SVD reveals a great deal about the matrix A and the key subspaces associated with it. The rank, r, of A is equal to the number of nonzero singular values, and the range and the null space of A are spanned by the first r columns of U and the last $n - r$ columns of V, respectively.

The SVD reveals not only the rank but also how close A is to a matrix of a given rank, as shown by a classic 1936 theorem of Eckart and Young.

Theorem 1 (Eckart–Young). *Let $A \in \mathbb{C}^{m\times n}$ have the SVD (1). If $k < r = \operatorname{rank}(A)$, then for the 2-norm and the Frobenius norm,*

$$\min_{\operatorname{rank}(B)=k} \|A - B\| = \|A - A_k\| = \begin{cases} \sigma_{k+1}, & 2\text{-norm}, \\ \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}, & F\text{-norm}, \end{cases}$$

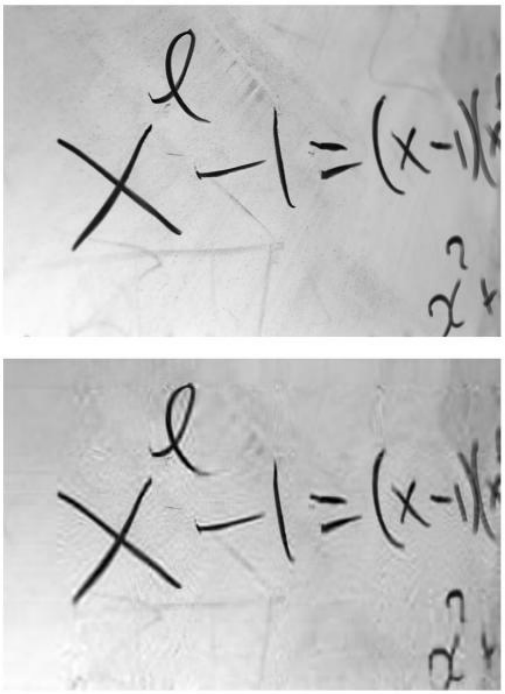


Figure 1 Photo of a blackboard, inverted so that white and black are interchanged in order to show more clearly the texture of the board: (a) original 1067×1600 image; (b) image compressed using rank-40 approximation A_{40} computed from SVD.

where

$$A_k = UD_kV^*, \quad D_k = \operatorname{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0).$$

In many situations the matrices that arise are necessarily of low rank but errors in the underlying data make the matrices actually obtained of full rank. The Eckart–Young result tells us that in order to obtain a lower-rank matrix we are justified in discarding (i.e., setting to zero) singular values that are of the same order of magnitude as the errors in the data.

The SVD (1) can be written as an outer product expansion

$$A = \sum_{i=1}^{p} \sigma_i u_i v_i^*,$$

and A_k in the Eckart–Young theorem is given by the same expression with p replaced by k. If $k \ll p$ then A_k requires much less storage than A and so the SVD can provide *data compression* (or *data reduction*). As an example, consider the monochrome image in figure 1(a) represented by a 1067×1600 array of RGB

Copyrighted material

values ($R = G = B$ since the image is monochrome). Let $A \in \mathbb{R}^{1067\times1600}$ contain the values from any one of the three channels. The singular values of A range from 8.4×10^4 down to 1.3×10^1. If we retain only the singular values down to the 40th, $\sigma_{40} = 2.1 \times 10^3$ (a somewhat arbitrary cutoff since there is no pronounced gap in the singular values), we obtain the image in figure 1(b). The reduced SVD requires only 6% of the storage of the original matrix. Some degradation is visible in the compressed image (and more can be seen when it is viewed at 100% size on screen), but it retains all the key features of the original image. While this example illustrates the power of the SVD, image compression is in general done much more effectively by the JPEG SCHEME [VII.7 §5].

A pleasing feature of the SVD is that the singular values are not unduly affected by perturbations. Indeed, if A is perturbed to $A + E$ then no singular value of A changes by more than $\|E\|_2$.

The SVD is a valuable tool in applications where two-sided orthogonal transformations can be carried out without "changing the problem," as it allows the matrix of interest to be diagonalized. Foremost among such problems is the LINEAR LEAST-SQUARES PROBLEM [IV.10 §7.1] $\min_{x\in\mathbb{C}^n} \|b - Ax\|_2$.

The SVD was first derived by Beltrami in 1873. The first reliable method for computing it was published by Golub and Kahan in 1965; this method applies two-sided unitary transformations to A and does not form and solve the equation (2), or its analogue for AA^*. Once software for computing the SVD became readily available, in the 1970s, the use of the SVD proliferated. Among the wide variety of uses of the SVD are for TEXT MINING [VII.24], deciphering encrypted messages, and image deblurring.

Further Reading

Eldén, L. 2007. *Matrix Methods in Data Mining and Pattern Recognition.* Philadelphia, PA: SIAM.

Golub, G. H., and C. F. Van Loan. 2013. *Matrix Computations*, 4th edn. Baltimore, MD: Johns Hopkins University Press.

Moler, C. B., and D. Morrison. 1983. Singular value analysis of cryptograms. *American Mathematical Monthly* 90:78–87.

II.33 Tensors and Manifolds

Mark R. Dennis

We know that the surface of the Earth is curved, despite the fact that it appears flat. This is easily understood from the fact that the Earth's radius of curvature is over 6000 km, vast on a human scale. This picture motivates the mathematical definition of a *manifold* (properly a *Riemannian manifold*): a space that appears to be Euclidean locally in a neighborhood of each point (or pseudo-Euclidean, as defined below) but globally may have curvature, such as the surface of a sphere.

Manifolds are most simply defined in terms of the coordinate systems on them, and of course there are uncountably many such systems. *Tensors* are mathematical objects defined on manifolds, such as vector fields, which are in a natural sense independent of the coordinate system used to define them and their components. The importance of tensors in physics stems from the fact that the description of physical phenomena ought to be independent of any coordinate system we choose to impose on space and hence should be tensorial.

Our description of manifolds and tensors will be rather informal. For instance, we will picture vector or tensor fields as defining a vector or tensor at each point of the manifold itself rather than more abstractly as a section of the appropriate tangent bundle. In applications, tensors are frequently used in the study of GENERAL RELATIVITY AND COSMOLOGY [IV.40], which involves describing the dynamics of matter and fields using any reference frame (coordinate system), assuming space-time is a four-dimensional pseudo-Riemannian curved manifold, as described below.

An n-dimensional manifold is a topological space such that a neighborhood around each point is equivalent (i.e., homeomorphic) to a neighborhood of a point in n-DIMENSIONAL EUCLIDEAN SPACE [I.2 §19.1]. More formally, it can be defined as the set of smooth coordinate systems that can be defined on the space, together with transformation rules between them. In a neighborhood around each point, a coordinate system can always be found that looks locally Cartesian, regardless of any global curvature (which can cause the system to fail to be Cartesian at other points).

In practice, each coordinate system on a Riemannian manifold has a *metric*, defined below, which is possibly position dependent. This enables inner products between pairs of vectors at each point in the space to be defined. The situation is complicated by the fact that, at each point, most coordinate systems are *oblique*, as in figure 1. The following description uses "index notation," which suggests the explicit choice of

Copyrighted material

force on the body. The conventional way to overcome the paradox is to bring back viscosity but only inside a thin BOUNDARY LAYER [II.2] attached to the body (see also FLUID MECHANICS [IV.28 §7.2]).

III.12 The Euler–Lagrange Equations

Paul Glendinning

The function $y(x)$ with derivative $y' = \mathrm{d}y/\mathrm{d}x$ that maximizes or minimizes the integral

$$\int F(y, y', x)\,\mathrm{d}x$$

with given endpoints satisfies the Euler–Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial y'}\right) - \frac{\partial F}{\partial y} = 0. \tag{1}$$

There are many variants of this equation to deal with further complications, e.g., if y or x or both are vectors, and more details are given in CALCULUS OF VARIATIONS [IV.6], but this simple version is sufficient to demonstrate the power and ubiquity of variational problems of this form.

If $F = F(y, y')$ has no explicit x-dependence (x is said to be *absent*), then the Euler–Lagrange equations can be simplified by finding a first integral. Using (1) it is straightforward to show that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(y'\frac{\partial F}{\partial y'} - F\right) = 0,$$

and hence that

$$y'\frac{\partial F}{\partial y'} - F = A \tag{2}$$

for some constant A.

Application 1: Potential Forces

Classical mechanics can be formulated as a problem of minimizing the integral of a function called the *Lagrangian*, $\mathcal{L}$, which is the kinetic energy minus the potential energy. For a particle moving in one dimension with position q (so the dependent variable q plays the role of y above and time t plays the role of the independent variable x) in a potential $V(q)$, the Lagrangian is $\mathcal{L} = \frac{1}{2}m\dot{q}^2 - V(q)$ and the Euler–Lagrange equation (1) is simply Newton's law for the acceleration, $m\ddot{q} = -V'(q)$ (where the prime denotes differentiation with respect to q), while the autonomous version (2) shows that $\frac{1}{2}m\dot{q}^2 + V(q)$ is constant, which is the conservation of energy (see CLASSICAL MECHANICS [IV.19]).

The power of this approach (and a related version due to Hamilton) is such that much of modern theoretical physics revolves round a generalization called an *action*.

Application 2: The Catenary

The problem of determining the curve describing the rest state of a heavy chain or cable with fixed endpoints can also be solved using the Euler–Lagrange formulation, although the original seventeenth-century solution uses simple mechanics. In the rest state the chain will assume a shape $y = y(x)$ that minimizes the potential energy $g\int y\,\mathrm{d}s$, where g is the acceleration due to gravity and s is the arc length along the chain. The length of the chain is $\int \mathrm{d}s$, and since this length is assumed to be constant, L say, $\int \mathrm{d}s = L$. This acts as a constraint on the solutions of the energy-minimization problem and so the full problem can be approached by introducing a LAGRANGE MULTIPLIER [I.3 §10], λ. Scaling out the constant g and noting that $\mathrm{d}s = \sqrt{1 + y'^2}\,\mathrm{d}x$, the shape of the curve minimizes

$$\int y\sqrt{1 + y'^2}\,\mathrm{d}x - \lambda\left(\int \sqrt{1 + y'^2}\,\mathrm{d}x - L\right). \tag{3}$$

(The second term represents the constraint and is zero when the constraint is satisfied.) The Euler–Lagrange equation with

$$F(y, y', \lambda) = \sqrt{1 + y'^2} - \lambda\sqrt{1 + y'^2}$$

can now be used since the λL term of (3) is constant with respect to variations in y. The Euler–Lagrange equation is supplemented by an additional equation obtained by extremizing with respect to the Lagrange multiplier, i.e., setting the derivative of (3) with respect to λ to zero, but this is just the length constraint again. Since x is absent, (2) implies that

$$(y - \lambda)\left(\frac{y'^2}{\sqrt{1 + y'^2}} - \sqrt{1 + y'^2}\right) = A.$$

Tidying up the left-hand side and rearranging gives $A^2(1 + y'^2) = (y - \lambda)^2$. Rewriting this as an expression for y' gives a differential equation that can be solved by separation of variables to give

$$y - \lambda = A\cosh\left(\frac{x - B}{A}\right),$$

where B is a further constant of integration. This is the catenary curve, and the constants are determined by the endpoints of the chain and the constraint that the total length is L.

Copyrighted material