

**The IMA Volumes  
in Mathematics  
and its Applications**

**Volume 107**

*Series Editors*

Avner Friedman   Robert Gulliver

# Institute for Mathematics and its Applications IMA

The **Institute for Mathematics and its Applications** was established by a grant from the National Science Foundation to the University of Minnesota in 1982. The IMA seeks to encourage the development and study of fresh mathematical concepts and questions of concern to the other sciences by bringing together mathematicians and scientists from diverse fields in an atmosphere that will stimulate discussion and collaboration.

The IMA Volumes are intended to involve the broader scientific community in this process.

Willard Miller, Jr., Professor and Director

\* \* \* \* \*

## IMA ANNUAL PROGRAMS

|           |  |
|-----------|--|
| 1982–1983 | Statistical and Continuum Approaches to Phase Transition                   |
| 1983–1984 | Mathematical Models for the Economics of Decentralized Resource Allocation |
| 1984–1985 | Continuum Physics and Partial Differential Equations                       |
| 1985–1986 | Stochastic Differential Equations and Their Applications                   |
| 1986–1987 | Scientific Computation   |
| 1987–1988 | Applied Combinatorics  |
| 1988–1989 | Nonlinear Waves  |
| 1989–1990 | Dynamical Systems and Their Applications                                   |
| 1990–1991 | Phase Transitions and Free Boundaries                                      |
| 1991–1992 | Applied Linear Algebra   |
| 1992–1993 | Control Theory and its Applications  |
| 1993–1994 | Emerging Applications of Probability                                       |
| 1994–1995 | Waves and Scattering   |
| 1995–1996 | Mathematical Methods in Material Science                                   |
| 1996–1997 | Mathematics of High Performance Computing                                  |
| 1997–1998 | Emerging Applications of Dynamical Systems                                 |
| 1998–1999 | Mathematics in Biology   |
| 1999–2000 | Reactive Flows and Transport Phenomena                                     |
| 2000–2001 | Mathematics in Multi-Media   |

Continued at the back

George Cybenko     Dianne P. O’Leary  
Jorma Rissanen  
Editors

# The Mathematics of Information Coding, Extraction, and Distribution

With 30 Illustrations



Springer

---

Mathematics Subject Classifications (1991): 68P20, 94A24, 68P10, 68P25, 94A60

---

Library of Congress Cataloging-in-Publication Data  
Cybenko, George

The mathematics of information coding, extraction, and distribution / George Cybenko, Dianne P. O'Leary, Jorma Rissanen.  
p. cm. — (The IMA volumes in mathematics and its applications ; 107)

Based on the proceedings of a workshop held in Nov. 1996 at the IMA.

Includes bibliographical references.

ISBN 978-1-4612-7178-9 ISBN 978-1-4612-1524-0 (eBook)

DOI 10.1007/978-1-4612-1524-0

1. High performance computing. 2. Coding theory. 3. Information theory. I. O'Leary, Dianne P. II. Rissanen, Jorma. III. Title.  
IV. Series: IMA volumes in mathematics and its applications ; v. 107.

QA76.88.C93 1999  
004.3—dc21

98-31464

Printed on acid-free paper.

© 1999 Springer Science+Business Media New York  
Originally published by Springer-Verlag New York, Inc. in 1999  
Softcover reprint of the hardcover 1st edition 1999

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher, Springer Science+Business Media, LLC except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone. Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Springer Science+Business Media, LLC provided that the appropriate fee is paid directly to Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, USA (Telephone: (508) 750-8400), stating the ISBN number, the title of the book, and the first and last page numbers of each article copied. The copyright owner's consent does not include copying for general distribution, promotion, new works, or resale. In these cases, specific written permission must first be obtained from the publisher.

Production managed by A. Orrantia; manufacturing supervised by Jacqui Ashri.  
Camera-ready copy prepared by the IMA.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4612-7178-9

## FOREWORD

This IMA Volume in Mathematics and its Applications

### THE MATHEMATICS OF INFORMATION CODING, EXTRACTION, AND DISTRIBUTION

is based on the proceedings of a workshop that was an integral part of the 1996–97 IMA program on “MATHEMATICS IN HIGH-PERFORMANCE COMPUTING.” The workshop brought together experts in various areas of mathematical and practical information theory and modeling to formulate the problems, explore new analytic methods and exchange ideas. It also addressed applications areas such as data mining, compression, database theory and machine learning, with special attention to the interactions between these areas from the analytical and mathematical points of view. Ideally, the workshop devoted half time to dissemination of new technical results and half time to the formulation of new paradigms and problems for future research.

We thank George Cybenko of Dartmouth College, Dianne O’Leary of University of Maryland, and Jorma Rissanen of IBM Almaden Research Center for their excellent work in organizing the workshop and editing the proceedings. We also take this opportunity to thank the National Science Foundation (NSF), the Office of Naval Research (ONR), and the Department of Energy (DOE), whose financial support made the workshop possible.

Willard Miller, Jr.

Robert Gulliver

## PREFACE

On November 11–15, 1996, a workshop on Information Coding, Extraction, and Distribution was held at the IMA as part of the Year of the Mathematics of High Performance Computing. There were approximately 30 attendees. The speakers included Michael Orchard, Robert Gray, Ahmed Tewfik, Jorma Rissanen, Julia Abrahams, Paul Siegel, Gil Strang, Cynthia Dwork, George Cybenko, Chris Atkeson, Dianne O’Leary, Geoff Davis, Eric Metois, Duncan Buell, and Manfred Opper, and this volume is a summary of some of their presentations.

High performance computing consumes and generates vast amounts of data, and the storage, retrieval, and transmission of this data are major obstacles to effective use of computing power. Challenges inherent in all of these operations are security, speed, reliability, authentication and reproducibility. This workshop focused on a wide variety of technical results aimed at meeting these challenges. Topics ranging from the mathematics of coding theory to the practicalities of copyright preservation for Internet resources drew spirited discussion and interaction among experts in diverse but related fields. We hope this volume contributes to continuing this dialogue.

George Cybenko  
Dianne P. O’Leary  
Jorma Rissanen

## CONTENTS

|  |     |
|--|-----|
| Foreword .....   | v   |
| Preface .....  | vii |
| Correspondences between variable length parsing<br>and coding problems .....                                       | 1   |
| <i>Julia Abrahams</i>  |     |
| The foundations of information push and pull .....   | 9   |
| <i>George Cybenko and Brian Brewington</i>   |     |
| Copyright? Protection? .....   | 31  |
| <i>Cynthia Dwork</i>   |     |
| Lossy compression, classification, and regression .....  | 49  |
| <i>Robert M. Gray</i>  |     |
| Latent semantic indexing via a semi-discrete<br>matrix decomposition .....   | 73  |
| <i>Tamara G. Kolda and Dianne P. O'Leary</i>   |     |
| Worst case prediction over sequences under log loss .....  | 81  |
| <i>Manfred Opper and David Haussler</i>  |     |
| Issues in multimedia databases: Coding for content-based<br>image retrieval and digital copyright protection ..... | 91  |
| <i>Mitchell D. Swanson and Ahmed H. Tewfik</i>   |     |