Codes and Automata

This major revision of Berstel and Perrin's classic *Theory of Codes* has been rewritten with a more modern focus and a much broader coverage of the subject. The concept of unambiguous automata, which is intimately linked with that of codes, now plays a significant role throughout the book, reflecting developments of the last 20 years. This is complemented by a discussion of the connection between codes and transducers, and new material from the field of symbolic dynamics. The authors have also explored links with more practical applications, including data compression and text processing. The treatment remains self-contained: there is background material on discrete mathematics, algebra and theoretical computer science. The wealth of exercises and examples make it ideal for self-study or courses. In sum this is a comprehensive reference on the theory of variable-length codes and their relation to unambiguous automata.

Jean Berstel is Emeritus Professor of Computer Science at the Université Paris-Est.

Dominique Perrin is Professor in Computer Science at the Université Paris-Est, and director of ESIEE Paris.

Christophe Reutenauer is Professor of Mathematics in the Combinatorics and Mathematical Computer Science Laboratory (LaCIM) at the University of Québec, Montréal.

ENCYCLOPEDIA OF MATHEMATICS AND ITS APPLICATIONS

All the titles listed below can be obtained from good booksellers or from Cambridge University Press. For a complete series listing visit

http://www.cambridge.org/uk/series/sSeries.asp?code=EOM

70 A. Pietsch and J. Wenzel *Orthonormal Systems and Banach Space Geometry*
71 G. E. Andrews, R. Askey and R. Roy *Special Functions*
72 R. Ticciati *Quantum Field Theory for Mathematicians*
73 M. Stern *Semimodular Lattices*
74 I. Lasiecka and R. Triggiani *Control Theory for Partial Differential Equations I*
75 I. Lasiecka and R. Triggiani *Control Theory for Partial Differential Equations II*
76 A. A. Ivanov *Geometry of Sporadic Groups I*
77 A. Schinzel *Polynomials with Special Regard to Reducibility*
78 T. Beth, D. Jungnickel and H. Lenz *Design Theory II, 2nd edn*
79 T. W. Palmer *Banach Algebras and the General Theory of *-Algebras II*
80 O. Stormark *Lie's Structural Approach to PDE Systems*
81 C. F. Dunkl and Y. Xu *Orthogonal Polynomials of Several Variables*
82 J. P. Mayberry *The Foundations of Mathematics in the Theory of Sets*
83 C. Foias, O. Manley, R. Rosa and R. Temam *Navier–Stokes Equations and Turbulence*
84 B. Polster and G. F. Steinke *Geometries on Surfaces*
85 R. B. Paris and D. Kaminski *Asymptotics and Mellin–Barnes Integrals*
86 R. McEliece *The Theory of Information and Coding, Student edition*
87 B. A. Magurn *An Algebraic Introduction to K-Theory*
88 T. Mora *Solving Polynomial Equation Systems I*
89 K. Bichteler *Stochastic Integration with Jumps*
90 M. Lothaire *Algebraic Combinatorics on Words*
91 A. A. Ivanov and S. V. Shpectorov *Geometry of Sporadic Groups II*
92 P. McMullen and E. Schulte *Abstract Regular Polytopes*
93 G. Gierz *et al. Continuous Lattices and Domains*
94 S. R. Finch *Mathematical Constants*
95 Y. Jabri *The Mountain Pass Theorem*
96 G. Gasper and M. Rahman *Basic Hypergeometric Series, 2nd edn*
97 M. C. Pedicchio and W. Tholen (eds.) *Categorical Foundations*
98 M. E. H. Ismail *Classical and Quantum Orthogonal Polynomials in One Variable*
99 T. Mora *Solving Polynomial Equation Systems II*
100 E. Olivieri and M. Eulália Vares *Large Deviations and Metastability*
101 A. Kushner, V. Lychagin and V. Rubtsov *Contact Geometry and Nonlinear Differential Equations*
102 L. W. Beineke and R. J. Wilson (eds.) with P. J. Cameron *Topics in Algebraic Graph Theory*
103 O. J. Staffans *Well-Posed Linear Systems*
104 J. M. Lewis, S. Lakshmivarahan and S. K. Dhall *Dynamic Data Assimilation*
105 M. Lothaire *Applied Combinatorics on Words*
106 A. Markoe *Analytic Tomography*
107 P. A. Martin *Multiple Scattering*
108 R. A. Brualdi *Combinatorial Matrix Classes*
110 M.-J. Lai and L. L. Schumaker *Spline Functions on Triangulations*
111 R. T. Curtis *Symmetric Generation of Groups*
112 H. Salzmann, T. Grundhöfer, H. Hähl and R. Löwen *The Classical Fields*
113 S. Peszat and J. Zabczyk *Stochastic Partial Differential Equations with Lévy Noise*
114 J. Beck *Combinatorial Games*
115 L. Barreira and Y. Pesin *Nonuniform Hyperbolicity*
116 D. Z. Arov and H. Dym *J-Contractive Matrix Valued Functions and Related Topics*
117 R. Glowinski, J.-L. Lions and J. He *Exact and Approximate Controllability for Distributed Parameter Systems*
118 A. A. Borovkov and K. A. Borovkov *Asymptotic Analysis of Random Walks*
119 M. Deza and M. Dutour Sikirić *Geometry of Chemical Graphs*
120 T. Nishiura *Absolute Measurable Spaces*
121 M. Prest *Purity, Spectra and Localisation*
122 S. Khrushchev *Orthogonal Polynomials and Continued Fractions*
123 H. Nagamochi and T. Ibaraki *Algorithmic Aspects of Graph Connectivity*
124 F. W. King *Hilbert Transforms I*
125 F. W. King *Hilbert Transforms II*
126 O. Calin and D.-C. Chang *Sub-Riemannian Geometry*
127 M. Grabisch *et al. Aggregation Functions*
128 L. W. Beineke and R. J. Wilson (eds.) with J. L. Gross and T. W. Tucker *Topics in Topological Graph Theory*
129 J. Berstel, D. Perrin and C. Reutenauer *Codes and Automata*
130 T. G. Faticoni *Modules over Endomorphism Rings*

Codes and Automata

JEAN BERSTEL

DOMINIQUE PERRIN
Université Paris-Est

CHRISTOPHE REUTENAUER
Université du Québec à Montréal

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521888318

First published 2010

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

ISBN 978-0-521-88831-8 Hardback

Contents

Preface

This book presents a comprehensive study of the theory of variable length codes. It is a complete reworking of the book *Theory of Codes* published by the first two authors more than twenty years ago. The present text includes many new results and also contains several additional chapters. Its focus is also broader, in the sense that more emphasis is given to algorithmic questions and to relations with other fields.

The theory of codes takes its origin in the theory of information devised by Shannon in the 1950s. As presented here, it makes use more of combinatorial and algebraic methods than of information theory. Due to the nature of the questions that are raised and solved, this theory has now become clearly a part of theoretical computer science and is strongly related to combinatorics on words, automata theory, formal languages, and the theory of semigroups.

The object of the theory of codes is, from an elementary point of view, the study of the properties concerning factorizations of words into sequences of words taken from a given set. One of the basic techniques used in this book is constructing special automata that perform this kind of parsing. We will show how properties of codes are reflected in combinatorial or algebraic properties of the associated devices.

It is quite remarkable that the problem of encoding as treated here admits a rather simple mathematical formulation: it is the study of embeddings of a free monoid into another. This may be considered to be a basic problem of algebra. There are related problems in other algebraic structures. For instance, if we replace free monoids by free groups, the study of codes reduces to that of subgroups of a free group. However, the situation is quite different at the very beginning since, according to the Nielsen–Schreier theorem, any subgroup of a free group is itself free, whereas the corresponding statement is false for free monoids. Nevertheless the relationship between codes and groups is more than an analogy, and we shall see in this book how the study of a group associated with a code can reveal some of its properties. It was M.-P. Schützenberger's discovery that coding theory is closely related to classical algebra. He has been the main architect of this theory. The main basic results are due to him and most further developments were stimulated by his conjectures.

The aim of the theory of codes is to give a structural description of codes in a way that allows their construction. This is easily accomplished for prefix codes, as shown in Chapter 3. The case of bifix codes is already much more difficult, and the complete structural description given in Chapter 6 is one of the highlights of

the theory. However, the structure of general codes (neither prefix nor suffix) still remains unknown to a large extent. For example, no systematic method is known for constructing all finite codes. The result given in Chapter 14 about the factorization of the polynomial of a code must be considered (despite the difficulty of its proof) as an intermediate step toward the understanding of codes.

Many of the results given in this book are concerned with extremal properties, the interest in which comes from the interconnection that appears between different concepts. But it also goes back to the initial investigations on codes considered as communication tools. Indeed, these extremal properties in general reflect some optimization in the encoding process. Thus a maximal code uses, in this sense, the whole capacity of the transmission channel.

Primarily, two types of methods are used in this book: direct methods on words on one hand, and automata and semigroups on the other hand. Direct methods consist of a more or less refined analysis of the sequencing of letters and factors within a word as it occurs in combinatorics on words. Automata and semigroups as used in Chapters 9–14, include the study of special automata associated with codes, called unambiguous automata and of the corresponding monoids of relations (unambiguous monoids of relations).

There are also many connections between the field of codes and automata and the field of symbolic dynamics. This aspect was not covered in *Theory of Codes*, and it is one of the new features of this volume. Symbolic dynamics focuses on the study of symbolic dynamical systems and, in particular of those defined by finite automata. The main point of intersection with codes is the notion of unambiguous automaton which coincides with the notion of *finite-to-one map* between symbolic systems. This relation is spread over several chapters. For example, the solution of the road coloring problem is presented in Chapter 10 and the notion of topological entropy is introduced in Chapter 13. The connections are explained in each chapter in the Notes section.

Codes and automata are related to algorithms on words and graphs. The computational complexity of algorithms related to codes is one of the topics of the book and is considered at various places in the text. We consider in particular algorithms related to tests for codes and to the construction of optimal prefix codes for several criteria.

The degree of generality of the exposition was influenced by the observation that many facts that hold for finite codes remain true for recognizable codes and even for the larger class of thin codes. In general, the transition from finite to recognizable codes does not imply major changes in the proof. However, changing to thin codes may imply some rather delicate computations. This is clearly demonstrated in Chapters 9 and 13, where the summations to be made become infinite when the codes are no longer recognizable. But this approach leads to a greater generality and, as we believe, to a better understanding by focusing attention on the main argument. Moreover, the characterization of the monoids associated with thin codes given in Chapter 9 may be considered to be a justification of our choice.

The organization of the book is as follows: A preliminary chapter (Chapter 1) is intended mainly to fix notation and should be consulted only when necessary. The

book is composed of two major parts: part one consisting of Chapters 2–8 and part two formed of Chapters 9–14.

Chapters 2–8 constitute an elementary introduction to the theory of codes in the sense that they primarily make use of direct methods. Chapter 2 contains the definition, the relationship with submonoids, the first results on Bernoulli distributions, and the introduction of the notions of complete, maximal, and thin codes.

Chapter 3 is devoted to a systematic study of prefix codes, developed at an elementary level. Indeed, this is the most intuitive and easy part of the theory of codes and certainly deserves considerable discussion. We believe that its interest largely goes beyond the theory of codes. We consider optimal prefix codes under various constraints. In particular, we give a full proof of the Garsia–Wachs algorithm.

Chapter 4 describes the automata used for representing codes, and for encoding and decoding words. The flower automaton is the basic tool for a syntactic study of codes. It is also helpful in an efficient algorithm for testing whether a rational set of words is a code. Encoders and decoders are transducers. We show how to construct deterministic transducers whenever it is possible.

Chapter 5 introduces the deciphering delay, the family of weakly prefix codes and their relation with weakly deterministic automata. The chapter contains the well-known theorem on maximal codes with finite deciphering delay.

Chapter 6 also is elementary, although it is more dense. Its aims are to describe the structure of maximal bifix codes and to give methods for constructing the finite ones. The use of formal power series is here of great help.

Chapter 7 is combinatorial in nature. It contains a description of length distributions of circular codes which is related to classical enumerative combinatorics. It contains also a systematic theory that leads to the study of the well-known comma-free codes.

Chapter 8 introduces the factorizations of a free monoid and more importantly of the characterization of the codes that may appear as factors. We present complete descriptions of finite factorizations for up to five factors.

The next five chapters contain what is known about codes but can be proved only by syntactic methods.

Chapter 9 is devoted to these techniques, using a more systematic treatment. Instead of the frequently encountered monoids of functions we study unambiguous monoids of relations which do not favor left or right. Chapter 9 contains an important result, already mentioned above: the characterization of thin maximal codes by a finiteness condition on the transition monoid of an unambiguous automaton.

Chapter 10 presents several results linked to the notion of synchronized codes. The notion of locally parsable code is related to that of local automaton. It contains also a proof of the road coloring problem, which has been recently solved. Chapter 11 deals with the groups of codes. It contains in particular the proof of the theorem of synchronization of semaphore codes announced in Chapter 3. Several results on the groups of finite maximal bifix codes are proved.

Chapter 12 presents elements of the theory of factorizations of cyclic groups. Several particular classes of these factorizations are described, such as those due to Hajós and Rédei. The relation with codes is developed.

Chapter 13 starts with a presentation of basics on probability spaces, and contains a proof of Kolmogorov's extension theorem. Next, it shows how to compute the density of the submonoid generated by a code by transferring the computation into the associated unambiguous monoid of relations. The formula of densities, linking together the density of the submonoids, the degree of the code, and the densities of the contexts, is the most striking result.

Chapter 14 contains the proof and discussion of the theorem of the factorization of the polynomial of a finite maximal code. Many of the results of the preceding chapters are used in the proof of this theorem, which contains the most current detailed information about the structure of general codes. The book ends with the connection between maximal bifix codes and semisimple algebras.

In an appendix, we gather, for the convenience of the reader, the conjectures mentioned in the book and present some additional open problems.

The book is written at an elementary level. In particular, the knowledge required is covered by a basic mathematical culture. Complete proofs are given and the necessary results of automata theory or theory of semigroups are presented in Chapter 1. Many examples are given which come from practical applications and illustrate the notions.

Each chapter is followed by a section of exercises. These frequently complement the material covered in the text. Solutions for this set of some 200 exercises are proposed at the end of the book. Each chapter ends with notes containing references, bibliographic discussions, complementary material, and references for the exercises.

It seems impossible to cover the whole text in a one-year course. However, the book contains enough material for several courses, at various levels, in undergraduate or graduate curricula.

A one-semester course at graduate level in discrete mathematics may be composed of Chapter 2, Chapter 3, Chapter 6, and Chapter 4. A one-semester course at undergraduate level may be composed of Chapter 2, Chapter 3 without the last section, and Chapter 4.

Several chapters are largely independent and can be lectured on separately. As an example, a course based solely on Chapter 7 has been taught by one of us. A course based on algorithms may contain the beginning of Chapter 2, the last section of Chapter 3, and Chapter 4.

Because of the extensive use of trees and of the algorithms described there, Chapter 3 by itself might constitute an interesting complement to a programming course.

Chapters 9 and 11, which rely on the structure of unambiguous monoids of relations, are an excellent illustration for a course in algebra. Similarly, Chapter 13 can be used as an adjunct to a course on probability theory.

The present volume is a new version of *Theory of Codes*, for which we have received help and collaboration from many people. It is a pleasure for us to renew our thanks to people who helped us during the preparation of the ancestor book: Aldo De Luca, Georges Hansel, Maurice Nivat, Jean-Eric Pin, Antonio Restivo, Stuart W. Margolis and Paul E. Schupp. The authors are greatly indebted to M.-P. Schützenberger (1920–1996). The project of writing the book stems from him and he has encouraged us constantly in many discussions.

The authors wish to thank, for help and comments on the present text, Marie-Pierre Béal, Jean-Marie Boë, Véronique Bruyère, Arturo Carpi, Christian Choffrut, Clelia De Felice, Sylvain Lavallée, Aaron Lauve, Yun Liu, Roberto Mantaci, Brian H. Marcus, Wojciek Plandowski, Jacques Sakarovitch, Alessandra Savelli, Paul H. Siegel, Sandor Szabó, Stephanie van Willigenburg and Ken Zeger. Special thanks are due to Jean Néraud who has carefully read all exercises and solutions.