Concentration of Measure for the Analysis of Randomized Algorithms

Randomized algorithms have become a central part of the algorithms curriculum based on their increasingly widespread use in modern applications.

This book presents a coherent and unified treatment of probabilistic techniques for obtaining high probability estimates on the performance of randomized algorithms. It covers the basic toolkit from the Chernoff–Hoeffding bounds to more sophisticated techniques like martingales and isoperimetric inequalities, as well as some recent developments like Talagrand's inequality, transportation cost inequalities and log-Sobolev inequalities. Along the way, variations on the basic theme are examined, such as Chernoff–Hoeffding bounds in dependent settings. The authors emphasise comparative study of the different methods, highlighting respective strengths and weaknesses in concrete example applications.

The exposition is tailored to discrete settings sufficient for the analysis of algorithms, avoiding unnecessary measure-theoretic details, thus making the book accessible to computer scientists as well as probabilists and discrete mathematicians.

DEVDATT P. DUBHASHI is Professor in the Department of Computer Science and Engineering at Chalmers University, Sweden. He earned a Ph.D. in computer science from Cornell University and held positions at the Max-Planck-Institute for Computer Science in Saarbruecken, BRICS, the University of Aarhus, and IIT Delhi. Dubhashi has published widely at international conferences and in journals, including many special issues dedicated to best contributions. His research interests span the range from combinatorics, to probabilistic analysis of algorithms and, more recently, to computational systems biology and distributed information systems such as the Web.

ALESSANDRO PANCONESI is Professor of Computer Science at Sapienza University of Rome. He earned a Ph.D. in computer science from Cornell University and is the recipient of the 1992 ACM Danny Lewin Award. Panconesi has published more than 50 papers in international journals and selective conference proceedings, and he is the associate editor of the *Journal of Discrete Algorithms* and the director of BiCi, the Bertinoro International Center of Informatics. His research spans areas of algorithmic research as diverse as randomized algorithms, distributed computing, complexity theory, experimental algorithmics, wireless networking and web information retrieval.

Concentration of Measure for the Analysis of Randomized Algorithms

DEVDATT P. DUBHASHI
Chalmers University

ALESSANDRO PANCONESI
Sapienza University of Rome

CAMBRIDGE
UNIVERSITY PRESS

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

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521884273

First published 2009

Printed in the United States of America

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

Library of Congress Cataloging in Publication data

Dubhashi, Devdatt.
Concentration of measure for the analysis of randomized algorithms / Devdatt Dubhashi.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-88427-3 (hardback : alk. paper)
1. Random variables. 2. Distribution (Probability theory). 3. Limit theorems
(Probability theory). 4. Algorithms. I. Title.
QA273.D765 2009
518′.1–dc22 2009009014

ISBN 978-0-521-88427-3 hardback

Dubhashi: To the genes before me (my respected parents) and after me
(Vinus and Minoo)
Panconesi: To the memory of my beloved father

Cambridge University Press
978-0-521-88427-3 - Concentration of Measure for the Analysis of Randomized Algorithms
Devdatt P. Dubhashi and Alessandro Panconesi
Frontmatter
More information

Contents

Preface

The aim of this book is to provide a body of tools for establishing concentration of measure that is accessible to researchers working in the design and analysis of randomized algorithms.

Concentration of measure refers to the phenomenon that a function of a large number of random variables tends to concentrate its values in a relatively narrow range (under certain conditions of smoothness of the function and under certain conditions of the dependence amongst the set of random variables). Such a result is of obvious importance to the analysis of randomized algorithms: for instance, the running time of such an algorithm can then be guaranteed to be concentrated around a pre-computed value. More generally, various other parameters measuring the performance of randomized algorithms can be provided tight guarantees via such an analysis.

In a sense, the subject of concentration of measure lies at the core of modern probability theory as embodied in the laws of large numbers, the central limit theorem and, in particular, the theory of large deviations [26]. However, these results are asymptotic: they refer to the limit as the number of variables n goes to infinity, for example. In the analysis of algorithms, we typically require quantitative estimates that are valid for finite (though large) values of n. The earliest such results can be traced back to the work of Azuma, Chernoff and Hoeffding in the 1950s. Subsequently, there have been steady advances, particularly in the classical setting of martingales. In the last couple of decades, these methods have taken on renewed interest, driven by applications in algorithms and optimisation. Also several new techniques have been developed.

Unfortunately, much of this material is scattered in the literature, and also rather forbidding for someone entering the field from a computer science or algorithms background. Often this is because the methods are couched in the technical language of analysis and/or measure theory. Although this may be strictly necessary to develop results in their full generality, it is not needed when the method is used in computer science applications (where the probability spaces are often finite and discrete), and indeed may serve only as a distraction or barrier.

Our main goal here is to give an exposition of the basic and more advanced methods for measure concentration in a manner that is accessible to the researcher in randomized algorithms and enables him or her to quickly start putting them to work in his or her application.

Book Outline

The book falls naturally into two parts. The first part contains the core bread-and-butter methods that we believe belong as an absolutely essential ingredient in the toolkit of a researcher in randomized algorithms today. Chapters 1 and 2 start with the basic Chernoff–Hoeffding bound on the sum of bounded independent random variables and give several applications. This topic is now covered in other recent books, and we therefore give several examples not covered there and refer the reader to these books, which can be read profitably together with this one (see suggestions given later). In Chapter 3, we give four versions of the Chernoff–Hoeffding bound in situations in which the random variables are not independent – this often is the case in the analysis of algorithms. Chapter 4 is a small interlude on probabilistic recurrences which can often give very quick estimates of tail probabilities based only on expectations.

The next series of chapters, Chapters 5–8, is devoted to a powerful extension of the Chernoff–Hoeffding bound to arbitrary functions of random variables (rather than just the sum) and where the assumption of independence can be relaxed somewhat. This is achieved via the concept of a *martingale*. These methods are by now rightly perceived as being fundamental in algorithmic applications and have begun to appear, albeit very scantily, in introductory books such as [74] and, more thoroughly, in the more recent [72]. Our treatment here is far more comprehensive and nuanced, and at the same time also very accessible to the beginner. We offer a host of relevant examples in which the various methods are seen in action.

Chapter 5 gives an introduction to the basic definition and theory of martingales leading to Azuma's inequality. The concept of martingales, as found in probability textbooks, poses quite a barrier to the computer scientist who is unfamiliar with the language of filters, partitions and measurable sets from measure theory. We are able to dispense with the measure-theoretic baggage entirely and keep to very elementary discrete probability. Chapters 6–8 are devoted to a set of nicely packaged inequalities based on martingales that are deployed with a host of applications. One of the special features of our exposition is our introduction of a very useful concept in probability called *coupling* and our demonstration of how it can be used to great advantage in working with these inequalities.

Chapter 9 is another short interlude containing an introduction to some recent specialised methods that were very successful in analysing certain key problems in random graphs.

We end Part I with Chapter 10, which is an introduction to isoperimetric inequalities that are a common setting for results on the concentration of measure. This lays the groundwork for the methods in Part II.

Part II of the book, Chapters 11–14, contains some more advanced techniques and recent developments. Here we systematise and make accessible some very useful tools that appear scattered in the literature and are couched in terms quite unfamiliar to computer scientists. From this (for a computer scientist) arcane body of work we distill out what is relevant and useful for algorithmic applications, using many non-trivial examples showing how these methods can be put to good use.

Chapter 11 is an introduction to Talagrand's isoperimetric theory, a theory developed in his 1995 epic, which proved a major landmark in the subject and led to the resolution of some outstanding open problems. We give a statement of the inequality that is simpler, at least conceptually, than the ones usually found in the literature. Yet, the simpler statement is sufficient for all the known applications, several of which are given in the book.

In Chapter 12, we give an introduction to an approach from information theory, via the so-called *transportation cost* inequalities, which yields very elegant proofs of the isoperimetric inequalities in Chapter 10. This approach, as shown by Kati Marton, extends in an elegant way to prove Talagrand's isoperimetric inequality, and we give an account of this in Chapter 13. In Chapter 14, we give an introduction to another approach from information theory that leads to concentration inequalities – the so-called *entropy* method or *log-Sobolev* inequalities. This approach too yields short proofs of Talagrand's inequality, and we also revisit the method of bounded differences in a different light.

How to Use the Book

This book is, we hope, a self-contained, comprehensive and quite accessible resource for any person with a typical computer science or mathematics background who is interested in applying concentration of measure methods in the design and analysis of randomized algorithms.

This book can also be used in an advanced course in randomized algorithms (or related courses) to supplement and complement some well-established textbooks. For instance, we recommend using it for a course in the following fields:

Randomized algorithms together with

- R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, 1995.
- M. Mitzenmacher and E. Upfal. *Probability and Computing*. Cambridge University Press, Cambridge, 2005.

Probabilistic combinatorics together with the classic

- N. Alon and J. Spencer. *The Probabilistic Method*, second edition. John Wiley, Hoboken, NJ, 2000.

Graph colouring together with

- M. Molloy and B. Reed. *Graph Coloring and the Probabilistic Method*. Springer, New York, 2002.

Random graphs together with

- S. Janson, T. Luczak, and A. Rucinski. *Random Graphs*. John Wiley, Hoboken, NJ, 2000.

Large-deviation theory together with

- F. den Hollander. *Large Deviations*. Fields Institute Monograph. American Mathematical Society, Providence, RI, 2000.

Acknowledgements

Several people have been helpful by providing much-needed encouragement, suggestions, corrections, comments and even drawings. We thank them all: Luigi Ambrosio, Flavio Chierichetti, Stefan Dziembowski, Alan Frieze, Rafael Frongillo, Bernd Gärtner, Michelangelo Grigni, Johan Hastad, Michal Karonski, Fred Kochman, Silvio Lattanzi, Alberto Marchetti-Spaccamela, Aravind Srinivasan and Sebastiano Vigna. Very special thanks go to Eli Upfal. The responsibility for any mistakes or omissions, alas, rests only upon us.

We thank our parents for their sacrifices to give us the best opportunities in life. Devdatt also thanks his family, Anna, Minoo and Vinus, for all the time that was rightfully theirs. Finally we thank Himanshu Abrol of Aptara and Lauren Cowles of Cambridge University Press for their kindness and effectiveness.

Concentration of Measure for the Analysis of Randomized Algorithms