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 emphasize 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
Dubhashi: To the genes before me (my respected parents) and after me
(Vinus and Minoo)

Panconesi: To the memory of my beloved father
## Contents

Preface  xi

1 Chernoff–Hoeffding Bounds  1
   1.1 What Is “Concentration of Measure”?  1
   1.2 The Binomial Distribution  2
   1.3 The Chernoff Bound  3
   1.4 Heterogeneous Variables  5
   1.5 The Hoeffding Extension  6
   1.6 Useful Forms of the Bound  6
   1.7 A Variance Bound  8
   1.8 Pointers to the Literature  10
   1.9 Problems  10

2 Applications of the Chernoff–Hoeffding Bounds  16
   2.1 Probabilistic Amplification  16
   2.2 Load Balancing  17
   2.3 Skip Lists  18
   2.4 Quicksort  22
   2.5 Low-Distortion Embeddings  23
   2.6 Pointers to the Literature  29
   2.7 Problems  29

3 Chernoff–Hoeffding Bounds in Dependent Settings  34
   3.1 Negative Dependence  34
   3.2 Local Dependence  38
   3.3 Janson’s Inequality  39
   3.4 Limited Independence  43
   3.5 Markov Dependence  45
## Contents

3.6 Pointers to the Literature 49  
3.7 Problems 49  

4 **Interlude: Probabilistic Recurrences** 51  
4.1 Problems 56  

5 **Martingales and the Method of Bounded Differences** 58  
5.1 Review of Conditional Probabilities and Expectations 59  
5.2 Martingales and Azuma’s Inequality 61  
5.3 Generalising Martingales and Azuma’s Inequality 65  
5.4 The Method of Bounded Differences 67  
5.5 Pointers to the Literature 71  
5.6 Problems 72  

6 **The Simple Method of Bounded Differences in Action** 74  
6.1 Chernoff–Hoeffding Revisited 74  
6.2 Stochastic Optimisation: Bin Packing 74  
6.3 Balls and Bins 75  
6.4 Distributed Edge Colouring: Take 1 76  
6.5 Models for the Web Graph 78  
6.6 Game Theory and Blackwell’s Approachability Theorem 80  
6.7 Pointers to the Literature 82  
6.8 Problems 82  

7 **The Method of Averaged Bounded Differences** 85  
7.1 Hypergeometric Distribution 85  
7.2 Occupancy in Balls and Bins 86  
7.3 Stochastic Optimisation: Travelling Salesman Problem 88  
7.4 Coupling 90  
7.5 Handling Rare Bad Events 99  
7.6 Quicksort 101  
7.7 Pointers to the Literature 103  
7.8 Problems 103  

8 **The Method of Bounded Variances** 106  
8.1 A Variance Bound for Martingale Sequences 107  
8.2 Applications 110  
8.3 Pointers to the Literature 117  
8.4 Problems 118  

9 **Interlude: The Infamous Upper Tail** 121  
9.1 Motivation: Non-Lipschitz Functions 121  
9.2 Concentration of Multivariate Polynomials 121
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.3</td>
<td>The Deletion Method</td>
<td>123</td>
</tr>
<tr>
<td>9.4</td>
<td>Problems</td>
<td>124</td>
</tr>
<tr>
<td>10</td>
<td>Isoperimetric Inequalities and Concentration</td>
<td>126</td>
</tr>
<tr>
<td>10.1</td>
<td>Isoperimetric Inequalities</td>
<td>126</td>
</tr>
<tr>
<td>10.2</td>
<td>Isoperimetry and Concentration</td>
<td>127</td>
</tr>
<tr>
<td>10.3</td>
<td>The Hamming Cube</td>
<td>130</td>
</tr>
<tr>
<td>10.4</td>
<td>Martingales and Isoperimetric Inequalities</td>
<td>131</td>
</tr>
<tr>
<td>10.5</td>
<td>Pointers to the Literature</td>
<td>132</td>
</tr>
<tr>
<td>10.6</td>
<td>Problems</td>
<td>133</td>
</tr>
<tr>
<td>11</td>
<td>Talagrand’s Isoperimetric Inequality</td>
<td>136</td>
</tr>
<tr>
<td>11.1</td>
<td>Statement of the Inequality</td>
<td>136</td>
</tr>
<tr>
<td>11.2</td>
<td>The Method of Non-Uniformly Bounded Differences</td>
<td>139</td>
</tr>
<tr>
<td>11.3</td>
<td>Certifiable Functions</td>
<td>144</td>
</tr>
<tr>
<td>11.4</td>
<td>Pointers to the Literature</td>
<td>148</td>
</tr>
<tr>
<td>11.5</td>
<td>Problems</td>
<td>148</td>
</tr>
<tr>
<td>12</td>
<td>Isoperimetric Inequalities and Concentration via Transportation Cost Inequalities</td>
<td>151</td>
</tr>
<tr>
<td>12.1</td>
<td>Distance between Probability Distributions</td>
<td>151</td>
</tr>
<tr>
<td>12.2</td>
<td>Transportation Cost Inequalities Imply Isoperimetric Inequalities and Concentration</td>
<td>153</td>
</tr>
<tr>
<td>12.3</td>
<td>Transportation Cost Inequality in Product Spaces with the Hamming Distance</td>
<td>154</td>
</tr>
<tr>
<td>12.4</td>
<td>An Extension to Non-Product Measures</td>
<td>158</td>
</tr>
<tr>
<td>12.5</td>
<td>Pointers to the Literature</td>
<td>159</td>
</tr>
<tr>
<td>12.6</td>
<td>Problems</td>
<td>159</td>
</tr>
<tr>
<td>13</td>
<td>Quadratic Transportation Cost and Talagrand’s Inequality</td>
<td>161</td>
</tr>
<tr>
<td>13.1</td>
<td>Introduction</td>
<td>161</td>
</tr>
<tr>
<td>13.2</td>
<td>Review and Road Map</td>
<td>161</td>
</tr>
<tr>
<td>13.3</td>
<td>An $L_2$ (Pseudo)-Metric on Distributions</td>
<td>163</td>
</tr>
<tr>
<td>13.4</td>
<td>Quadratic Transportation Cost</td>
<td>165</td>
</tr>
<tr>
<td>13.5</td>
<td>Talagrand’s Inequality via Quadratic Transportation Cost</td>
<td>167</td>
</tr>
<tr>
<td>13.6</td>
<td>Extension to Dependent Processes</td>
<td>168</td>
</tr>
<tr>
<td>13.7</td>
<td>Pointers to the Literature</td>
<td>169</td>
</tr>
<tr>
<td>13.8</td>
<td>Problems</td>
<td>169</td>
</tr>
<tr>
<td>14</td>
<td>Log-Sobolev Inequalities and Concentration</td>
<td>171</td>
</tr>
<tr>
<td>14.1</td>
<td>Introduction</td>
<td>171</td>
</tr>
<tr>
<td>14.2</td>
<td>A Discrete Log-Sobolev Inequality on the Hamming Cube</td>
<td>172</td>
</tr>
</tbody>
</table>
Contents

14.3 Tensorisation 174
14.4 Modified Log-Sobolev Inequalities in Product Spaces 175
14.5 The Method of Bounded Differences Revisited 177
14.6 Self-Bounding Functions 179
14.7 Talagrand’s Inequality Revisited 179
14.8 Pointers to the Literature 181
14.9 Problems 181

Appendix A Summary of the Most Useful Bounds 185
A.1 Chernoff–Hoeffding Bounds 185
A.2 Bounds for Well-Behaved Functions 185

Bibliography 189
Index 195
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
Preface

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.
Preface

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

Probabilistic combinatorics  together with the classic

Graph colouring  together with

Random graphs  together with

Large-deviation theory  together with

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