"This is an excellent and very timely text, presenting the modern tools of high-dimensional geometry and probability in a very accessible and applications-oriented manner, with plenty of informative exercises. The book is infused with the author's insights and intuition in this field, and has extensive references to the latest developments in the area. It will be an extremely useful resource both for newcomers to this subject and for expert researchers."

– Terence Tao, University of California, Los Angeles

"Methods of high-dimensional probability have become indispensable in numerous problems of probability theory and its applications in mathematics, statistics, computer science, and electrical engineering. Roman Vershynin's wonderful text fills a major gap in the literature by providing a highly accessible introduction to this area. Starting with no prerequisites beyond a first course in probability and linear algebra, Vershynin takes the reader on a guided tour through the subject and consistently illustrates the utility of the material through modern data science applications. This book should be essential reading for students and researchers in probability theory, data science, and related fields."

– Ramon van Handel, Princeton University

"This very welcome contribution to the literature gives a concise introduction to several topics in 'high-dimensional probability' that are of key relevance in contemporary statistical science and machine learning. The author achieves a fine balance between presenting deep theory and maintaining readability for a non-specialist audience – this book is thus highly recommended for graduate students and researchers alike who wish to learn more about this by-now-indispensable field of modern mathematics."

– Richard Nickl, University of Cambridge

"Vershynin is one of the world's leading experts in the area of high-dimensional probability, and his textbook provides a gentle yet thorough treatment of many of the key tools in the area and their applications to the field of data science. The topics covered here are a must-know for anyone looking to do mathematical work in the field, covering subjects important in machine learning, algorithms and theoretical computer science, signal processing, and applied mathematics."

– Jelani Nelson, Harvard University

"High-Dimensional Probability is an excellent treatment of modern methods in probability and data analysis. Vershynin's perspective is unique and insightful, informed by his expertise as both a probabilist and a functional analyst. His treatment of the subject is gentle, thorough, and inviting, providing a great resource for both newcomers and those familiar with the subject. I believe, as the author does, that the topics covered in this book are indeed essential ingredients of the developing foundations of data science."

– Santosh Vempala, Georgia Institute of Technology

"Renowned for his deep contributions to high-dimensional probability, Roman Vershynin is to be commended for the clarity of his progressive exposition of the important concepts, tools, and techniques of the field. Advanced students and practitioners interested in the mathematical foundations of data science will enjoy the many relevant worked examples and lively use of exercises. This book is the reference I had been waiting for."

– Rémi Gribonval, IEEE & EURASIP Fellow, Directeur de Recherche, Inria, France

"High-dimensional probability is a fascinating mathematical theory that has grown rapidly in recent years. It is fundamental to high-dimensional statistics, machine learning, and data science. In this book, Roman Vershynin, who is a leading researcher in high-dimensional probability and a master of exposition, provides the basic tools and some of the main results and applications of high-dimensional probability. This book is an excellent textbook for a graduate course that will be appreciated by mathematics, statistics, computer science, and engineering students. It will also serve as an excellent reference book for researchers working in high-dimensional probability and statistics."

– Elchanan Mossel, Massachusetts Institute of Technology

"This book on the theory and application of high-dimensional probability is a work of exceptional clarity that will be valuable to students and researchers interested in the foundations of data science. A working knowledge of high-dimensional probability is essential for researchers at the intersection of applied mathematics, statistics, and computer science. The widely accessible presentation will make this book a classic that everyone in foundational data science will want to have on their bookshelf."

– Alfred Hero, University of Michigan

"Vershynin's book is a brilliant introduction to the mathematics which is at the core of modern signal processing and data science. The focus is on concentration of measure and its applications to random matrices, random graphs, dimensionality reduction, and suprema of random process. The treatment is remarkably clean, and the reader will learn beautiful and deep mathematics without unnecessary formalism."

– Andrea Montanari, Stanford University

High-Dimensional Probability

An Introduction with Applications in Data Science

High-Dimensional Probability offers insight into the behavior of random vectors, random matrices, random subspaces, and objects used to quantify uncertainty in high dimensions. Drawing on ideas from probability, analysis, and geometry, it lends itself to applications in mathematics, statistics, theoretical computer science, signal processing, optimization, and more. It is the first text to integrate theory, key tools, and modern applications of high-dimensional probability. Concentration inequalities form the core, and it covers both classical results such as Hoeffding's and Chernoff's inequalities and modern developments such as the matrix Bernstein inequality. It then introduces powerful methods based on stochastic processes, including such tools as Slepian's, Sudakov's, and Dudley's inequalities, as well as generic chaining and bounds based on VC dimension. A broad range of illustrations is embedded throughout, including classical and modern results for covariance estimation, clustering, networks, semidefinite programming, coding, dimension reduction, matrix completion, machine learning, compressed sensing, and sparse regression. Hints for many of the exercises are given at the back of the book.

ROMAN VERSHYNIN is Professor of Mathematics at the University of California, Irvine. He studies random geometric structures across mathematics and data sciences, in particular in random matrix theory, geometric functional analysis, convex and discrete geometry, geometric combinatorics, high-dimensional statistics, information theory, machine learning, signal processing, and numerical analysis. His honors include an Alfred Sloan Research Fellowship in 2005, an invited talk at the International Congress of Mathematicians in Hyderabad in 2010, and a Bessel Research Award from the Humboldt Foundation in 2013. His "Introduction to the non-asymptotic analysis of random matrices" has become a popular educational resource for many new researchers in probability and data science.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at www.cambridge.org/statistics.
Recent titles include the following:

21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler
23. *Applied Asymptotics*, by A. R. Brazzale, A. C. Davison and N. Reid
24. *Random Networks for Communication*, by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments*, by R. A. Bailey
26. *Symmetry Studies*, by Marlos A. G. Viana
27. *Model Selection and Model Averaging*, by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics*, edited by Nils Lid Hjort *et al.*
29. *From Finite Sample to Asymptotic Methods in Statistics*, by Pranab K. Sen, Julio M. Singer and Antonio C. Pedrosa de Lima
30. *Brownian Motion*, by Peter Mörters and Yuval Peres
31. *Probability (Fourth Edition)*, by Rick Durrett
33. *Stochastic Processes*, by Richard F. Bass
34. *Regression for Categorical Data*, by Gerhard Tutz
35. *Exercises in Probability (Second Edition)*, by Loïc Chaumont and Marc Yor
36. *Statistical Principles for the Design of Experiments*, by R. Mead, S. G. Gilmour and A. Mead
37. *Quantum Stochastics*, by Mou-Hsiung Chang
38. *Nonparametric Estimation under Shape Constraints*, by Piet Groeneboom and Geurt Jongbloed
39. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, by Jianfeng Yao, Shurong Zheng and Zhidong Bai
40. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, by Evarist Giné and Richard Nickl
41. *Confidence, Likelihood, Probability*, by Tore Schweder and Nils Lid Hjort
42. *Probability on Trees and Networks*, by Russell Lyons and Yuval Peres
43. *Random Graphs and Complex Networks (Volume 1)*, by Remco van der Hofstad
44. *Fundamentals of Nonparametric Bayesian Inference*, by Subhashis Ghosal and Aad van der Vaart
45. *Long-Range Dependence and Self-Similarity*, by Vladas Pipiras and Murad S. Taqqu
46. *Predictive Statistics*, by Bertrand S. Clarke and Jennifer L. Clarke
47. *High-Dimensional Probability*, by Roman Vershynin

High-Dimensional Probability

An Introduction with Applications in Data Science

Roman Vershynin
University of California, Irvine

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108415194
DOI: 10.1017/9781108231596

First published 2018

Printed in the United States of America by Sheridan Books, Inc.

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

Library of Congress Cataloging-in-Publication Data
Names: Vershynin, Roman, 1974– author.
Title: High-dimensional probability : an introduction with applications in
data science / Roman Vershynin, University of Michigan.
Description: Cambridge : Cambridge University Press, 2018. |
Series: Cambridge series in statistical and probabilistic mathematics ; 47 |
Includes bibliographical references and index.
Identifiers: LCCN 2018016910 | ISBN 9781108415194
Subjects: LCSH: Probabilities. | Stochastic processes. | Random variables.
Classification: LCC QA273 .V4485 2018 | DDC 519.2–dc23
LC record available at https://lccn.loc.gov/2018016910

ISBN 978-1-108-41519-4 Hardback

Contents

Foreword

This book begins with an appetizer: the empirical method of B. Maurey for approximating points in the convex hull of a set by averages. It is a beautiful and fascinating example where probability theory can elegantly solve problems which at first sight have nothing to do with probabilities. It is a very gratifying experience to learn from this book that probability theory opens up a whole world of other mathematical areas (which may have been found very difficult to access before).

After presenting the necessary background material, the book goes straight into the heart of the matter in Chapter 3. Concentration in high dimensions is treated in an enlightening way. For example, the formula

$$\sqrt{n \pm O(\sqrt{n})} = \sqrt{n} \pm O(1)$$

in Remark 3.1.2 says it all in all its simplicity. Likewise for Figure 3.6, where a Gaussian point cloud is shown in high dimensions: it concentrates on a sphere with radius $\sqrt{n}$. This shape has hardly anything in common with the bell shape in dimension 2 or 3 – our low-dimensional intuition is useless! As another example where probability theory can make life easier, the book provides an insightful proof of Grothendieck's inequality. To understand any of the other proofs of Grothendieck's inequality (with "good" constants), I would probably need several years.

Let me mention another theme that is extremely well explained in the book: the isoperimetric inequality and how it leads to blow up. If a subset of the sphere covers at least 50 percent, then its coverage is exponentially close to 100 percent. The book also presents several extensions to other metric spaces, for example concentration on the Grassmannian. In this way it provides a first entrance into this area, and one would like to learn more. The supplied pointers allow one to do so.

The book is a joy to read. The author conveys the material as an exciting story and one keeps on reading. Participation in the development of the storyline is encouraged by the many exercises that are scattered throughout the text.

Other topics treated in this book are random matrices, empirical process theory, and sparse recovery, to name a few. The results are important for research in data science but also simply of beauty on their own. Many students and researchers may have heard the key words, and this is the book to find out what they are really about.

Sara van de Geer, ETH Zürich

Preface

Who is This Book For?

This is a textbook in probability in high dimensions with a view toward applications in data sciences. It is intended for doctoral and advanced masters students and beginning researchers in mathematics, statistics, electrical engineering, computational biology, and related areas who are looking to expand their knowledge of theoretical methods used in modern research in the data sciences.

Why This Book?

The data sciences are moving fast, and probabilistic methods often provide a foundation and inspiration for such advances. Today, a typical graduate probability course is no longer sufficient to acquire the level of mathematical sophistication that is expected from a beginning researcher in data sciences. The book is intended to partially cover this gap. It presents some key probabilistic methods and results that form an essential toolbox for a mathematical data scientist. It can be used as a textbook for a basic second course in probability with a view toward data science applications. It is also suitable for self-study.

What is This Book About?

High-dimensional probability is an area of probability theory that studies random objects in $\mathbb{R}^n$, where the dimension n can be very large. The book places particular emphasis on random vectors, random matrices, and random projections. It teaches basic theoretical skills for the analysis of these objects, which include concentration inequalities, covering and packing arguments, decoupling and symmetrization tricks, chaining and comparison techniques for stochastic processes, combinatorial reasoning based on the VC dimension, and a lot more.

The study of high-dimensional probability provides vital theoretical tools for applications in data science. The book integrates theory with applications for covariance estimation, semidefinite programming, networks, elements of statistical learning, error correcting codes, clustering, matrix completion, dimension reduction, sparse signal recovery, and sparse regression.

Prerequisites

The essential prerequisites for reading this book are a rigorous course in probability theory (of the Masters or Ph.D. level), an excellent command of undergraduate linear algebra, and general familiarity with basic notions about metrics, normed and Hilbert spaces, and linear operators. A knowledge of measure theory is not essential but would be helpful.

A Word on the Exercises

The exercises are integrated into the text. The reader can do them immediately to check his or her understanding of the material just presented, and to prepare better for later developments. The difficulty of the exercises is indicated by the number of coffee cups; it ranges from easy (☕) to hard (☕☕☕☕). A pointing hand (☞) means that a hint is available at the end of the book.

Related Reading

The book covers only a fraction of the theoretical apparatus of high-dimensional probability and illustrates it with only a sample of data science applications. Each chapter in this book concludes with a Notes section, which has pointers to other texts on the subject matter of the chapter. A few particularly useful sources are noted here. The now classical book [8] showcases the probabilistic method in applications to discrete mathematics and computer science. The forthcoming book [19] will present a panorama of mathematical data science, focusing on applications in computer science. Both these books will be accessible to graduate and advanced undergraduate students. The lecture notes [206] are pitched at graduate students and present more theoretical material in high-dimensional probability.

Acknowledgements

The feedback from my many colleagues was instrumental in preparing this book. My special thanks go to Florent Benaych-Georges, Jennifer Bryson, Lukas Grätz, Rémi Gribonval, Ping Hsu, Mike Izbicki, George Linderman Cong Ma, Galyna Livshyts, Jelani Nelson, Ekkehard Schnoor, Martin Spindler, Dominik Stöger, Tim Sullivan, Terence Tao, Joel Tropp, Katarzyna Wyczesany, Yifei Shen, and Haoshu Xu, for many valuable suggestions and corrections, and in particular to Sjoerd Dirksen, Larry Goldstein, Wu Han, Han Wu, and Mahdi Soltanolkotabi for detailed proofreading of the book. I am grateful to Can Le, Jennifer Bryson, and my son Ivan Vershynin for their help with many of the pictures.