The Fundamentals of Heavy Tails

Heavy tails – extreme events or values more common than expected – emerge everywhere: the economy, natural events, and social and information networks are just a few examples. Yet after decades of progress, they are still treated as mysterious, surprising, and even controversial, primarily because the necessary mathematical models and statistical methods are not widely known.

This book, for the first time, provides a rigorous introduction to heavy-tailed distributions accessible to anyone who knows elementary probability. It tackles and tames the zoo of terminology for models and properties, demystifying topics such as the generalized central limit theorem and regular variation. It tracks the natural emergence of heavy-tailed distributions from a wide variety of general processes, building intuition. And it reveals the controversy surrounding heavy tails to be the result of flawed statistics, then equips readers to identify and estimate with confidence. Over 100 exercises complete this engaging package.

JAYAKRISHNAN NAIR is Associate Professor in Electrical Engineering at IIT Bombay. His research focuses on modeling, performance evaluation, and design issues in online learning environments, communication networks, queueing systems, and smart power grids. He is the recipient of best paper awards at IFIP Performance (2010 and 2020) and ACM e-Energy (2020).

ADAM WIERMAN is Professor of Computing and Mathematical Sciences at the California Institute of Technology (Caltech). His research develops tools in machine learning, optimization, control, and economics with the goal of making the networked systems that govern our world sustainable and resilient. He is best known for his work spearheading the design of algorithms for sustainable data centers and he is the recipient of numerous awards including the ACM Sigmetrics Rising Star award, the ACM Sigmetrics Test of Time award, the IEEE Communication Society William Bennet Prize, and multiple teaching and best paper awards.

BERT ZWART is group leader at CWI Amsterdam and Professor of Mathematics at Eindhoven University of Technology. He has expertise in stochastic operations research, queueing theory, and large deviations, and in the context of heavy tails, he has focused on sample path properties, designing Monte Carlo methods and applications to computer-communication and energy networks. He was area editor of *Operations Research*, the flagship journal of his profession, from 2009 to 2017, and was the recipient of the INFORMS Applied Probability Society Erlang prize, awarded every two years to an outstanding young applied probabilist.

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:

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: Theory and Examples* (*Fourth Edition*), by Rick Durrett
32. *Analysis of Multivariate and High-Dimensional Data*, by Inge Koch
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
48. *High-Dimensional Statistics*, by Martin J. Wainwright
49. *Probability: Theory and Examples* (*Fifth Edition*), by Rick Durrett
50. *Statistical Model-based Clustering and Classification*, by Charles Bouveyron *et al*.
51. *Spectral Analysis for Univariate Time Series*, by Donald B. Percival and Andrew T. Walden
52. *Statistical Hypothesis Testing in Context*, by Michael P. Fay and Erica Brittain

The Fundamentals of Heavy Tails

Properties, Emergence, and Estimation

Jayakrishnan Nair
Indian Institute of Technology, Bombay

Adam Wierman
California Institute of Technology

Bert Zwart
CWI Amsterdam and Eindhoven University of Technology

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

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

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/9781316511732
DOI: 10.1017/9781009053730

First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

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

ISBN 978-1-316-51173-2 Hardback

Contents

Preface

Heavy tails are a continual source of excitement and confusion as they are repeatedly "discovered" in new contexts. This is true across the social sciences, the physical sciences, and especially the information sciences, where heavy-tails pop up seemingly everywhere – from degree distributions in the web and social networks to file sizes and interarrival times of compute jobs. However, despite nearly a century of work on heavy tails, they are still treated as *mysterious*, due to their counterintuitive properties; *surprising*, since the central limit theorem trains us to expect the "Normal" distribution; and even *controversial*, since intuitive statistical tools can lead to false discoveries.

The goal of this book is to show that heavy-tailed distributions need not be mysterious and should not be surprising or controversial either. The book strives to demystify heavy-tailed distributions by showing how to reason formally about their counterintuitive properties by demonstrating that the emergence of heavy-tailed phenomena should be expected (not surprising) and by revealing that most of the controversy surrounding heavy-tails is the result of bad statistics and can be avoided by using the proper tools. It proceeds by offering a coherent collection of bite-sized introductions to the big questions:

How can we understand the seemingly counterintuitive properties of heavy-tailed distributions?
Why are heavy-tailed distributions so common?
How can heavy-tailed distributions be estimated from data?

These questions are important not just to mathematicians but also to computer scientists, economists, social scientists, astronomers, and beyond. Heavy-tailed distributions truly come up everywhere in the world around us, and understanding their consequences is crucial to an understanding of the world. Despite this, many students learn probability and statistics without being taught how to think carefully about the heavy tails. Our mission is to make this practice a thing of the past.

This book aims to provide an accessible introduction to the fundamentals of heavy tails. It covers mathematically deep concepts that are typically taught in graduate-level courses, such as the generalized central limit theorem, extreme value theory, and regular variation; but it does so using only elementary mathematical tools in order to make these topics accessible to anyone who has had an introductory probability course. To make this possible we have explicitly chosen, at times, not to prove results in full generality. Instead, we have worked to rigorously present the core idea in as simple a manner as possible, pushing aside technicalities

that are necessary to extend the analysis to the most general cases. However, we always provide pointers to other books and papers where proofs of the most general results can be found.

How to Use This Book

The book is designed for teaching as well as for independent study by anyone who wants to develop the intuition and mathematical tools for reasoning about heavy-tailed distributions. In teaching from the book, we have found that each chapter corresponds to approximately two 80-minute lectures (though we do not cover all the material in each chapter during lectures) and that it is exciting to include a project component in the course where students explore heavy tails in their areas of interest. Due to the interdisciplinary nature of heavy tails, we have found that students tend to come from widely varying backgrounds and so the projects can be extremely diverse.

The first chapter of the book introduces the class of heavy-tailed distributions formally, and discusses a few examples of common heavy-tailed distributions. These examples illustrate how the behavior of heavy-tailed distributions contrasts with that of light-tailed distributions, but do not yet build intuition for these differences or explain why heavy-tailed distributions are so common in the world around us. The chapter also begins to introduce the controversy that often surrounds heavy-tailed distributions because intuitive statistical approaches for identifying heavy tails in data are flawed.

The remainder of the book is organized to first provide intuition, both qualitative and mathematical, about the defining properties of heavy-tailed distributions (Part I: Properties), then explain why heavy-tailed distributions are so common in the world around us (Part II: Emergence), and finally develop the statistical tools for the estimation of heavy-tailed distributions (Part III: Estimation).

Given the mystique and excitement that surrounds the discovery of heavy-tailed phenomena, detection and estimation of heavy tails in data is an activity that is often (over)zealously pursued. You may be tempted to skip directly to Part III on estimation. However, the book is written so that the tools used in Part II are developed in Part I, and the tools used in Part III are developed in Parts I and II. Thus, we encourage readers to work through the book in order. That said, we have organized the material in each chapter so that there is a main body that presents the core ideas important for later chapters, followed by sections that present examples and/or variations of the main topic. These later sections can be viewed as enrichment opportunities that can be skipped as desired if the goal is to move quickly to Part III. The quickest path for acquiring the background needed before digging into Part III consists of Chapter 2 from Part I, then Chapters 5 and 7 from Part II.

Finally, we would like to emphasize that the aim of this book is to present the fundamentals of heavy-tailed distributions in a way that is both rigorous and accessible to readers who have taken an introductory undergraduate course in probability. Given that the theory of heavy tails uses advanced mathematical tools and is typically presented in a way that is accessible only to graduate students in mathematics, our target has required us to rethink and reprove many classical results in the area with the goal of providing a simple, intuitive presentation. This often means stating theorems that have more restrictive assumptions than the most general results known and presenting proofs that convey the key ideas but address some of the difficult technical details via either an assumption or a reference to a technical

lemma in another source. To provide interested readers with references to the full generality of the results we discuss, each chapter ends with an "Additional Notes" section that includes references to more detail on the topics presented in the chapter. We encourage readers to follow up on the references in these sections. Additionally, many of the exercises at the ends of the chapters ask the reader to derive extensions or fill in technical details that we have left out of the main body of the chapters, so we encourage readers to work through the exercises.

Our goal for this book is that, through reading it, heavy-tailed distributions will be demystified for you. That their properties will be intuitive, not mysterious. That their emergence will be expected, not surprising. And, that you will have the proper statistical tools for studying heavy-tailed phenomena and so will help resolve (or avoid) controversies rather than feed them. Happy reading!

Acknowledgments

When first teaching an iteration of our course on heavy tails at Caltech in 2011 we had hopes of finishing a book on the topic within a couple years. Instead, the writing process continued off and on over a decade, during which time the authors added eight children to their families! The process of writing a book like this is joyous. You get to return to the core results of a field you love and think deeply about how to simplify them to their core; attempting to isolate a form of the results that may be more restrictive than the most general formulation, but that expose the essence of the intuition for the results using as simple an argument as possible. In many cases, this process led us to develop new intuition for classical results, which has fed back into our own research over the years.

Given this lengthy on-again/off-again process of writing this book, it is not surprising that this book owes a great deal of thanks to a great number of people. Our approach to thinking about heavy tails has been shaped in many ways by our research collaborators and mentors throughout our careers. The book draws on perspectives that we have developed from our conversations with colleagues such as Onno Boxma, Sem Borst, Mor Harchol-Balter, Alan Scheller-Wolf, Sergey Foss, Sid Resnick, Gennady Samorodnitsky, Thomas Mikosch, John Doyle, K. Mani Chandy, Steven Low, Mike Pinedo, Alessandro Zocca, and Guannan Qu, Peter Glynn, Ramesh Johari, and others.

The book grew out of courses taught at Caltech and IIT Bombay, and we owe a huge debt of thanks to the teaching assistants and students of those courses for their feedback and suggestions. We also thank the faculty who have integrated parts of this book into their courses at other schools all around the world. The comments and suggestions from their experience have been invaluable. More broadly, thanks goes to all the students in the Rigorous Systems Research Group (RSRG) at Caltech, many of whom read early versions of the chapters and shared thoughts and insights at formative stages of the book.

We are extremely appreciative of Longbo Huang and Chenye Wu, who hosted Adam Wierman for a summer of intensive writing at Tsinghua University, which was crucial to finally finishing the book. The conversations with Longbo and Chenye about the book and research more broadly were inspirational, and the input on the book from students at Tsinghua was irreplaceable.

As we have progressed in writing the book, we have kept a version of our draft available on our webpages and we want to thank the hundreds of readers who carefully studied those versions and provided us comments, found typos, and shared applications where they applied the ideas of the book. It was extremely motivating and stimulating to hear about the wide

variety of places where the intuitions and techniques introduced in the book proved to be impactful.

Finally, we are grateful to the editorial team at Cambridge University Press, especially Diana Gillooly, for their patience with us as we prepared the book and for the advice and help they provided along the way.