Introduction to Probability for Computing

MOR HARCHOL-BALTER

Cambridge University Press
Illustrated by Elin Zhou

“Based on 20 years of teaching Computer Science and Operations Research at Carnegie Mellon University, Professor Harchol-Balter provides a unique presentation of probability and statistics that is both highly engaging and also strongly motivated by real-world computing applications that students will encounter in industry. This book is approachable and fun for undergraduate students, while also covering advanced concepts relevant to graduate students.”

Eytan Modiano, Massachusetts Institute of Technology

“This book provides a fantastic introduction to probability for computer scientists and computing professionals, addressing concepts and techniques crucial to the design and analysis of randomized algorithms, to performance well-designed simulations, to statistical inference and machine learning, and more. Also contains many great exercises and examples. Highly recommend!”

Avrim Blum, Toyota Technological Institute at Chicago

“Mor Harchol-Balter’s new book does a beautiful job of introducing students to probability! The book is full of great computer science-relevant examples, wonderful intuition, simple and clear explanations, and mathematical rigor. I love the question-answer style she uses, and could see using this book for students ranging from undergraduate students with zero prior exposure to probability all the way to graduate students (or researchers of any kind) who need to brush up and significantly deepen (and/or broaden) their knowledge of probability.”

Anna Karlin, University of Washington

“Probability is at the heart of modeling, design, and analysis of computer systems and networks. This book by a pioneer in the area is a beautiful introduction to the topic for undergraduate students. The material in the book introduces theoretical topics rigorously, but also motivates each topic with practical applications. This textbook is an excellent resource for budding computer scientists who are interested in probability.”

R. Srikant, University of Illinois at Urbana-Champaign

“I know probability, and have taught it to undergrads and grads at MIT, UC Berkeley, and Carnegie Mellon University. Yet this book has taught me some wonderfully interesting important material that I did not know. Mor is a great thinker, lecturer, and writer. I would love to have learned from this book as a student – and to have taught from it as an instructor!”

Manuel Blum, U.C. Berkeley and Carnegie Mellon University

To the students at CMU's
School of Computer Science
whose curiosity and drive
inspire me every day
to keep writing.

Contents

Part IV Computer Systems Modeling and Simulation

Part V Statistical Inference

Preface

Probability theory has become indispensable in computer science. It is at the core of machine learning and statistics, where one often needs to make decisions under stochastic uncertainty. It is also integral to computer science theory, where most algorithms today are randomized algorithms, involve random coin flips. It is a central part of performance modeling in computer networks and systems, where probability is used to predict delays, schedule jobs and resources, and provision capacity.

Why This Book?

This book gives an introduction to probability as it is used in computer science theory and practice, drawing on applications and current research developments as motivation and context. This is not a typical counting and combinatorics book, but rather it is a book centered on distributions and how to work with them.

Every topic is driven by what computer science students need to know. For example, the book covers distributions that come up in computer science, such as heavy-tailed distributions. There is a large emphasis on variability and higher moments, which are very important in empirical computing distributions. Computer systems modeling and simulation are also discussed, as well as statistical inference for estimating parameters of distributions. Much attention is devoted to tail bounds, such as Chernoff bounds. Chernoff bounds are used for confidence intervals and also play a big role in the analysis of randomized algorithms, which themselves comprise a large part of the book. Finally, the book covers Markov chains, as well as a bit of queueing theory, both with an emphasis on their use in computer systems analysis.

Intended Audience

The material is presented at the advanced undergraduate level. The book is based on an undergraduate class, Probability and Computing (PnC), which I have been teaching at Carnegie Mellon University (CMU) for almost 20 years. While PnC is primarily taken by undergraduates, several Masters and PhD students choose to take the class. Thus we imagine that instructors can use the book for different levels of classes, perhaps spanning multiple semesters.

Question/Answer Writing Style

The book uses a style of writing aimed at engaging the reader to be active, rather than passive. Instead of large blocks of text, we have short "Questions" and "Answers." In working through the book, you should cover up the answers, and write down your own answer to each question, before looking at the given answer. The goal is "thinking" rather than "reading," where each chapter is intended to feel like a conversation.

Exercises

The exercises in this book are an integral part of learning the material. They also introduce many of the computer science and statistics applications. Very few of the exercises are rote. Every problem has important insights, and the insights often build on each other. Exercises are (very roughly) organized from easier to harder. Several of the exercises in the book were contributed by students in the class!

To aid in teaching, solutions to a large subset of the exercises are available *for instructors only* at www.cambridge.org/harchol-balter. Instructors who need solutions to the remaining exercises can request these from the author. The solutions are for the personal use of the instructor only. They should not be distributed or posted online, so that future generations can continue to enjoy the exercises.

Organization of the Material

The book consists of eight parts. Parts I, II, and III provide an introduction to basic probability. Part IV provides an introduction to computer systems modeling and simulation. Part V provides an introduction to statistical inference. Parts VI and VII comprise a course in randomized algorithms, starting with tail bound inequalities and then applying these to analyze a long list of randomized algorithms. Part VIII provides an introduction to stochastic processes as they're used in computing.

Before we describe the parts in more detail, it is worth looking at the *dependency structure* for the book, given in Figure P1. Aside from Parts I, II, and III, most of the parts can be taught in any order.

In particular, it is possible to imagine at least *four different courses* being taught from this book, depending on the parts that an instructor might choose to teach. Figure P2 depicts different courses that one might teach. All the courses start with Parts I, II, and III, but then continue with Simulation, or Statistics, or Randomized Algorithms, or Stochastic Processes, depending on the particular course.

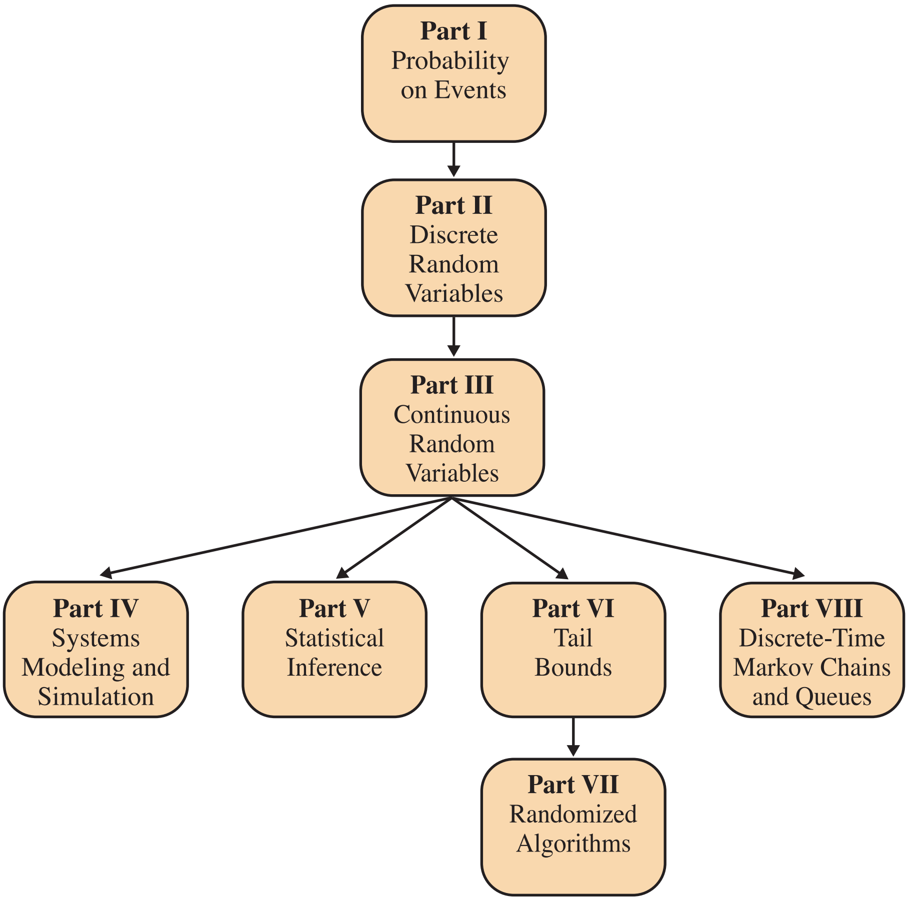

Figure P1 *The dependency structure between the parts of this book. Most parts are independent of other parts and can be taught in any order.*

Description of Each Part

Part I: Foundations and Probability on Events: Part I starts by reviewing the prerequisites for the book. These include series, calculus, elementary combinatorics, and asymptotic notation. Exercises and examples are provided to help in reviewing the prerequisites. The main focus of Part I is on defining probability on events, including conditioning on events, independence of events, the Law of Total Probability, and Bayes' Law. Some examples of applications covered in Part I are: faulty computer networks, Bayesian reasoning for healthcare testing, modeling vaccine efficacy, the birthday paradox, Monty Hall problems, and modeling packet corruption in the Internet.

Part II: Discrete Random Variables: Part II introduces the most common discrete random variables (Bernoulli, Binomial, Geometric, and Poisson), and then

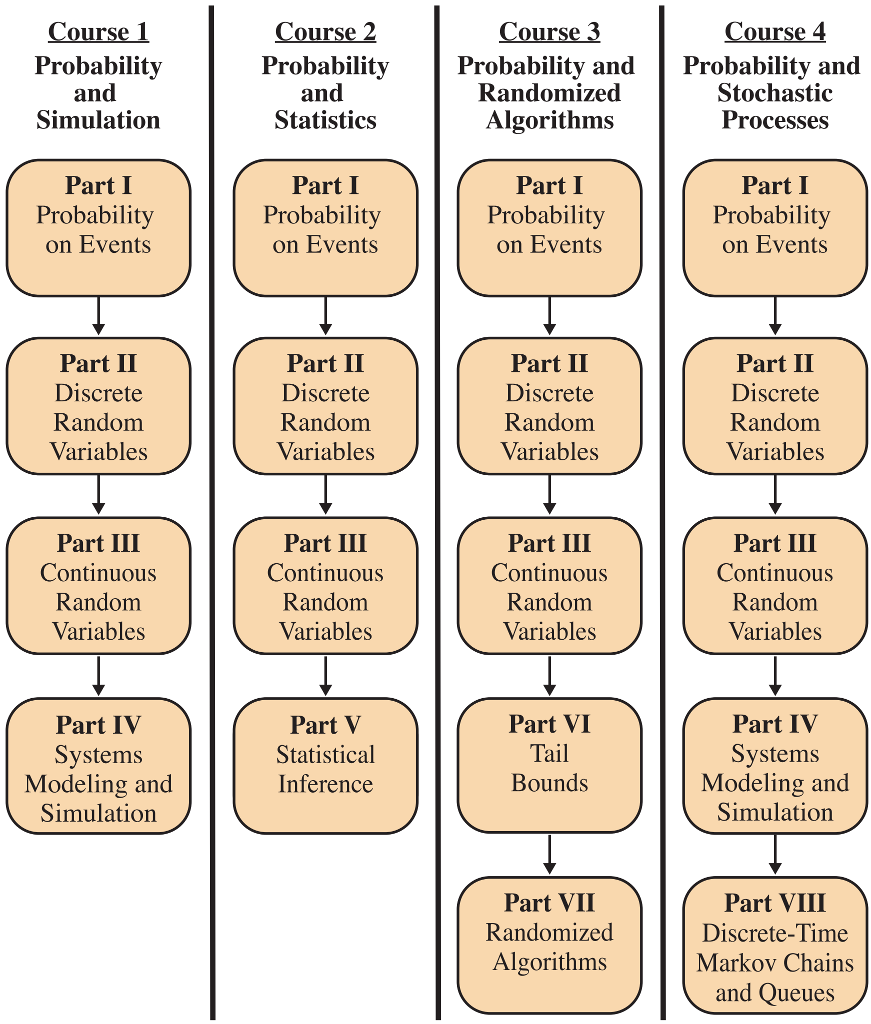

Figure P2 *Four different courses that one can teach out of this book.*

continues with the standard material on random variables, such as linearity of expectation, conditioning, conditional probability mass functions, joint distributions, and marginal distributions. Some more advanced material is also included, such as: variance and higher moments of random variables; moment-generating functions (specifically z-transforms) and their use in solving recurrence relations; Jensen's inequality; sums of a random number of random variables; tail orderings, and simple tail inequalities. Both Simpson's paradox and the inspection paradox are covered. Some examples of applications covered in Part II are: noisy reading from a flash storage, the binary symmetric channel, approximating a Binomial distribution by a Poisson, the classical marriage algorithm, modeling the time until a disk fails, the coupon collector problem, properties of random graphs, time until k consecutive failures, computer virus propagation, epidemic growth modeling, hypothesis testing in data analysis, stopping times, total variation distance, and polygon triangulation.

Part III: Continuous Random Variables: Part III repeats the material in Part II, but this time with continuous random variables. We introduce the Uniform, Exponential, and Normal distributions, as well as the Central Limit Theorem. In addition, we introduce the Pareto heavy-tailed distribution, which is most relevant for empirical computing workloads, and discuss its relevance to today's data center workloads. We cover failure rate functions and the heavy-tail property and their relevance to computing workloads. We again cover moment-generating functions, but this time via Laplace transforms, which are more commonly used with continuous random variables. Some applications covered in Part II are: classifying jobs in a supercomputing center, learning the bias of a coin, dart throwing, distributions whose parameters are random variables, relating laptop quality to lifetime, modeling disk delays, modeling web file sizes, modeling compute usage, modeling IP flow durations, and Internet node degrees.

Part IV: Computer Systems Modeling and Simulation: Part IV covers the basics of what is needed to run simulations of computer systems. We start by defining and analyzing the Poisson process, which is the most commonly used model for the arrival process of jobs into computer systems. We then study how to generate random variables for simulation, using the inverse transform method and the accept–reject method. Finally, we discuss how one would program a simple event-driven or trace-driven simulator. Some applications that we cover include: Malware detection of infected hosts, population modeling, reliability theory, generating a Normal random variable, generating Pareto and Bounded Pareto random variables, generating a Poisson random variable, simulation of heavy-tailed distributions, simulation of high-variance distributions, simulation of jointly distributed random variables, simulation of queues, and simulation of networks of queues.

Part V: Statistical Inference: Part V switches gears to statistics, particularly statistical inference, where one is trying to estimate some parameters of an experiment. We start with the most traditional estimators, the sample mean and sample variance. We also cover desirable properties of estimators, including zero bias, low mean squared error, and consistency. We next cover maximum likelihood estimation and linear regression. We complete this part with a discussion of maximum a posterior (MAP) estimators and minimum mean square error (MMSE) estimators. Some applications that we cover include: estimating voting probabilities, deducing the original signal in a noisy environment, estimating true job sizes from user estimates, estimation in interaction graphs, and estimation in networks with error correcting codes.

Part VI: Tail Bounds and Applications: Part VI starts with a discussion of tail bounds and concentration inequalities (Markov, Chebyshev, Chernoff), for which we provide full derivations. We provide several immediate applications for these tail bounds, including a variety of classic balls-and-bins applications. The balls and bins framework has immediate application to dispatching tasks to servers in a server farm, as well as immediate application to hashing algorithms, which we also study extensively. We cover applications of tail bounds to defining confidence intervals in statistical estimation, and well as bias estimation, polling schemes, crowd sourcing, and other common settings from computing and statistics.

Part VII: Randomized Algorithms: Part VII introduces a wide range of randomized algorithms. The randomized algorithms include Las Vegas algorithms, such as randomized algorithms for sorting and median finding, as well as Monte Carlo randomized algorithms such as MinCut, MaxCut, matrix multiplication checking, polynomial multiplication, and primality testing. The exercises in this part are particularly relevant because they introduce many additional randomized algorithms such as randomized dominating set, approximate median finding, independent set, AND/OR tree evaluation, knockout tournaments, addition of n-bit numbers, randomized string exchange, path-finding in graphs, and more. We use the tail bounds that we derived earlier in Part VI to analyze the runtimes and accuracy of our randomized algorithms.

Part VIII: Markov Chains with a Side of Queueing Theory: Part VIII provides an introduction to stochastic processes as they come up in computer science. Here we delve deeply into discrete-time Markov chains (both finite and infinite). We discuss not only how to solve for limiting distributions, but also when they exist and why. Ergodicity, positive-recurrence and null-recurrence, passage times, and renewal theory are all covered. We also cover time averages versus ensemble averages and the impact of these different types of averages on running simulations. Queueing theory is integral to Part VIII. We define the performance metrics that computer scientists care about: throughput, response time, and load. We cover Little's Law, stability, busy periods, and capacity provisioning. A huge number of applications are covered in Part VIII, including, for example, the classic PageRank algorithm for ranking web pages, modeling of epidemic spread, modeling of caches, modeling processors with failures, Brownian motion, estimating the spread of malware, reliability theory applications, population modeling, server farm and data center modeling, admission control, and capacity provisioning.

Acknowledgments

Most textbooks begin with a class, and this book is no exception. I created the Probability and Computing (called "PnC" for short) class 20 years ago, with the aim of teaching computer science undergraduates the probability that they need to know to be great computer scientists. Since then I have had a few opportunities to co-teach PnC with different colleagues, and each such opportunity has led to my own learning. I would like to thank my fantastic co-instructors: John Lafferty, Klaus Sutner, Rashmi Vinayak, Ryan O'Donnell, Victor Adamchik, and Weina Wang. I'm particularly grateful to Weina, who collaborated with me on three of the chapters of the book and who is a kindred spirit in Socratic teaching. The book has also benefited greatly from many spirited TAs and students in the class, who proposed fun exercises for the book, many referencing CMU or Pittsburgh.

I would also like to thank my illustrator, Elin Zhou, who painstakingly created every image and figure in the book, while simultaneously managing her undergraduate classes at CMU. I chose Elin as my illustrator because her artwork embodies the spirit of fun and inclusiveness that permeates the PnC class. One of the themes of PnC is chocolate, which is tossed out throughout the class to students who answer questions. This chocolate would not be possible if it weren't for our class sponsor, Citadel, who even paid to have chocolate mailed directly to student homes throughout the pandemic, while classes were online.

I have been fortunate to have several excellent editors at Cambridge University Press: Julie Lancashire, Ilaria Tassistro, and Rachel Norridge. Thanks to their recommendations, the statistics chapters were added, redundant material was removed, and the style and layout of the book improved immensely. My copy editor, Gary Smith, was also fantastic to work with and meticulous!

On a personal note, I want to thank my family. In particular, I'm grateful to my son, Danny Balter, for always telling me that I'm good at explaining things. I'm also grateful to my mom, Irit Harchol, who is one of my best friends, and who takes the time to talk with me every day as I walk to and from work. Thanks to my inlaws, Ann and Richard Young, who are my cheering squad. Finally, I have infinite love and gratitude for my husband, Ary Young, for always making me their top priority and for never leaving my side, even if it means sleeping on my sofa as I sit here typing away.