SIMULATION AND THE MONTE CARLO METHOD

WILEY SERIES IN PROBABILITY AND STATISTICS
Established by *Walter A. Shewhart and Samuel S. Wilks*

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Geof H. Givens, Harvey Goldstein, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *J. Stuart Hunter, Iain M. Johnstone, Joseph B. Kadane, Jozef L. Teugels*

The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at http://www.wiley.com/go/wsps

SIMULATION AND THE MONTE CARLO METHOD

Third Edition

Reuven Y. Rubinstein
Technion

Dirk P. Kroese
University of Queensland

WILEY

Copyright © 2017 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Names: Rubinstein, Reuven Y. | Kroese, Dirk P.
Title: Simulation and the Monte Carlo method.
Description: Third edition / Reuven Rubinstein, Dirk P. Kroese. | Hoboken, New Jersey : John Wiley & Sons, Inc., [2017] | Series: Wiley series in probability and statistics | Includes bibliographical references and index.
Identifiers: LCCN 2016020742 (print) | LCCN 2016023293 (ebook) | ISBN 9781118632161 (cloth) | ISBN 9781118632208 (pdf) | ISBN 9781118632383 (epub)
Subjects: LCSH: Monte Carlo method. | Digital computer simulation. | Mathematical statistics. | Sampling (Statistics)
Classification: LCC QA298 .R8 2017 (print) | LCC QA298 (ebook) | DDC 518/.282--dc23
LC record available at https://lccn.loc.gov/2016020742

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To the memory of

Reuven Y. Rubinstein (1938–2012)

Reuven Rubinstein passed away during the writing of this third edition. Reuven was one of the pioneers of Monte Carlo simulation and remained at the forefront of research in this area right up to the end of his life. In 2011 he received the highest honor given by INFORMS Simulation Society: The Lifetime Professional Achievement Award, where his achievements were summarized as follows:

> Professor Rubinstein has been a pivotal figure in the theory and practice of simulation as we know it today. His career reflects a high level of creativity and contribution, with a willingness to explore new areas and an amazing ability to suggest surprising new avenues of research and to influence subsequent work.

May his contagious enthusiasm and curiosity live on through his books and the many people whom he inspired.

Dirk P. Kroese

CONTENTS

Preface xiii

Acknowledgments xvii

1 Preliminaries 1

1.1 Introduction 1
1.2 Random Experiments 1
1.3 Conditional Probability and Independence 2
1.4 Random Variables and Probability Distributions 4
1.5 Some Important Distributions 5
1.6 Expectation 6
1.7 Joint Distributions 7
1.8 Functions of Random Variables 11
  1.8.1 Linear Transformations 12
  1.8.2 General Transformations 13
1.9 Transforms 14
1.10 Jointly Normal Random Variables 15
1.11 Limit Theorems 16
1.12 Poisson Processes 17
1.13 Markov Processes 19
  1.13.1 Markov Chains 19
  1.13.2 Classification of States 21

1.13.3 Limiting Behavior 22
1.13.4 Reversibility 24
1.13.5 Markov Jump Processes 25
1.14 Gaussian Processes 27
1.15 Information 28
1.15.1 Shannon Entropy 29
1.15.2 Kullback–Leibler Cross-Entropy 31
1.15.3 Maximum Likelihood Estimator and Score Function 32
1.15.4 Fisher Information 33
1.16 Convex Optimization and Duality 34
1.16.1 Lagrangian Method 35
1.16.2 Duality 37
Problems 41
References 46

2 Random Number, Random Variable, and Stochastic Process Generation 49

2.1 Introduction 49
2.2 Random Number Generation 49
2.2.1 Multiple Recursive Generators 51
2.2.2 Modulo 2 Linear Generators 52
2.3 Random Variable Generation 55
2.3.1 Inverse-Transform Method 55
2.3.2 Alias Method 57
2.3.3 Composition Method 58
2.3.4 Acceptance–Rejection Method 59
2.4 Generating from Commonly Used Distributions 62
2.4.1 Generating Continuous Random Variables 62
2.4.2 Generating Discrete Random Variables 67
2.5 Random Vector Generation 70
2.5.1 Vector Acceptance–Rejection Method 71
2.5.2 Generating Variables from a Multinormal Distribution 72
2.5.3 Generating Uniform Random Vectors over a Simplex 73
2.5.4 Generating Random Vectors Uniformly Distributed over a Unit Hyperball and Hypersphere 74
2.5.5 Generating Random Vectors Uniformly Distributed inside a Hyperellipsoid 75
2.6 Generating Poisson Processes 75
2.7 Generating Markov Chains and Markov Jump Processes 77
2.7.1 Random Walk on a Graph 78
2.7.2 Generating Markov Jump Processes 79
2.8 Generating Gaussian Processes 80

2.9 Generating Diffusion Processes 81
2.10 Generating Random Permutations 83
Problems 85
References 89

3 Simulation of Discrete-Event Systems 91

3.1 Introduction 91
3.2 Simulation Models 92
 3.2.1 Classification of Simulation Models 94
3.3 Simulation Clock and Event List for DEDS 95
3.4 Discrete-Event Simulation 97
 3.4.1 Tandem Queue 97
 3.4.2 Repairman Problem 101
Problems 103
References 106

4 Statistical Analysis of Discrete-Event Systems 107

4.1 Introduction 107
4.2 Estimators and Confidence Intervals 108
4.3 Static Simulation Models 110
4.4 Dynamic Simulation Models 112
 4.4.1 Finite-Horizon Simulation 114
 4.4.2 Steady-State Simulation 114
4.5 Bootstrap Method 126
Problems 127
References 130

5 Controlling the Variance 133

5.1 Introduction 133
5.2 Common and Antithetic Random Variables 134
5.3 Control Variables 137
5.4 Conditional Monte Carlo 139
 5.4.1 Variance Reduction for Reliability Models 141
5.5 Stratified Sampling 144
5.6 Multilevel Monte Carlo 146
5.7 Importance Sampling 149
 5.7.1 Weighted Samples 149
 5.7.2 Variance Minimization Method 150
 5.7.3 Cross-Entropy Method 154
5.8 Sequential Importance Sampling 159
5.9 Sequential Importance Resampling 165

5.10 Nonlinear Filtering for Hidden Markov Models 167
5.11 Transform Likelihood Ratio Method 171
5.12 Preventing the Degeneracy of Importance Sampling 174
Problems 179
References 184

6 Markov Chain Monte Carlo 187

6.1 Introduction 187
6.2 Metropolis–Hastings Algorithm 188
6.3 Hit-and-Run Sampler 193
6.4 Gibbs Sampler 194
6.5 Ising and Potts Models 197
6.5.1 Ising Model 197
6.5.2 Potts Model 198
6.6 Bayesian Statistics 200
6.7 Other Markov Samplers 202
6.7.1 Slice Sampler 204
6.7.2 Reversible Jump Sampler 205
6.8 Simulated Annealing 208
6.9 Perfect Sampling 212
Problems 214
References 219

7 Sensitivity Analysis and Monte Carlo Optimization 221

7.1 Introduction 221
7.2 Score Function Method for Sensitivity Analysis of DESS 224
7.3 Simulation-Based Optimization of DESS 231
7.3.1 Stochastic Approximation 232
7.3.2 Stochastic Counterpart Method 237
7.4 Sensitivity Analysis of DEDS 246
Problems 252
References 255

8 Cross-Entropy Method 257

8.1 Introduction 257
8.2 Estimation of Rare-Event Probabilities 258
8.2.1 Root-Finding Problem 267
8.2.2 Screening Method for Rare Events 268
8.2.3 CE Method Combined with Sampling from the Zero-Variance Distribution 271
8.3 CE Method for Optimization 272

8.4 Max-Cut Problem 276
8.5 Partition Problem 282
8.5.1 Empirical Computational Complexity 283
8.6 Traveling Salesman Problem 283
8.6.1 Incomplete Graphs 288
8.6.2 Node Placement 289
8.6.3 Case Studies 290
8.7 Continuous Optimization 291
8.8 Noisy Optimization 292
8.9 MinxEnt Method 294
Problems 298
References 303

9 Splitting Method 307

9.1 Introduction 307
9.2 Counting Self-Avoiding Walks via Splitting 308
9.3 Splitting with a Fixed Splitting Factor 310
9.4 Splitting with a Fixed Effort 313
9.5 Generalized Splitting 314
9.6 Adaptive Splitting 318
9.7 Application of Splitting to Network Reliability 321
9.8 Applications to Counting 322
9.9 Case Studies for Counting with Splitting 325
9.9.1 Satisfiability (SAT) Problem 325
9.9.2 Independent Sets 330
9.9.3 Permanent and Counting Perfect Matchings 332
9.9.4 Binary Contingency Tables 334
9.9.5 Vertex Coloring 336
9.10 Splitting as a Sampling Method 337
9.11 Splitting for Optimization 340
9.11.1 Continuous Optimization 343
Problems 344
References 348

10 Stochastic Enumeration Method 351

10.1 Introduction 351
10.2 Tree Search and Tree Counting 352
10.3 Knuth's Algorithm for Estimating the Cost of a Tree 355
10.4 Stochastic Enumeration 357
10.4.1 Combining SE with Oracles 359
10.5 Application of SE to Counting 360

10.5.1 Counting the Number of Paths in a Network 360
10.5.2 Counting SATs 363
10.5.3 Counting the Number of Perfect Matchings in a Bipartite Graph 366
10.6 Application of SE to Network Reliability 368
10.6.1 Numerical Results 370
Problems 373
References 375

Appendix 377

A.1 Cholesky Square Root Method 377
A.2 Exact Sampling from a Conditional Bernoulli Distribution 378
A.3 Exponential Families 379
A.4 Sensitivity Analysis 382
A.4.1 Convexity Results 383
A.4.2 Monotonicity Results 384
A.5 A Simple CE Algorithm for Optimizing the Peaks Function 385
A.6 Discrete-Time Kalman Filter 385
A.7 Bernoulli Disruption Problem 387
A.8 Complexity 389
A.8.1 Complexity of Rare-Event Algorithms 389
A.8.2 Complexity of Randomized Algorithms: FPRAS and FPAUS 390
A.8.3 SATs in CNF 394
A.8.4 Complexity of Stochastic Programming Problems 395
Problems 402
References 403

Abbreviations and Acronyms 405

List of Symbols 407

Index 409

PREFACE

Since the publication in 2008 of the second edition of *Simulation and the Monte Carlo Method*, significant changes have taken place in the field of Monte Carlo simulation. This third edition gives a fully updated and comprehensive account of the major topics in Monte Carlo simulation.

The book is based on an undergraduate course on Monte Carlo methods given at the Israel Institute of Technology (Technion) and the University of Queensland for the last five years. It is aimed at a broad audience of students in engineering, physical and life sciences, statistics, computer science, mathematics, and simply anyone interested in using Monte Carlo simulation in their study or work. Our aim is to provide an accessible introduction to modern Monte Carlo methods, focusing on the main concepts, while providing a sound foundation for problem solving. For this reason most ideas are introduced and explained via concrete examples, algorithms, and experiments.

Although we assume that the reader has some basic mathematical background, such as an elementary course in probability and statistics, we nevertheless review the basic concepts of probability, stochastic processes, information theory, and convex optimization in Chapter 1.

In a typical stochastic simulation, randomness is introduced into simulation models via independent uniformly distributed random variables. These random variables are then used as building blocks to simulate more general stochastic systems. Chapter 2 deals with the generation of such random numbers, random variables, and stochastic processes.

Many real-world complex systems can be modeled as discrete-event systems. Examples of discrete-event systems include traffic systems, flexible manufacturing systems, computer-communications systems, inventory systems, production lines, coherent lifetime systems, PERT networks, and flow networks. The behavior of such systems is identified via a sequence of discrete "events" that causes the system to change from one "state" to another. We discuss how to model such systems on a computer in Chapter 3.

Chapter 4 treats the statistical analysis of the output data from static and dynamic simulation models. The main difference is that the former do not evolve in time whereas the latter do. For dynamic models, we distinguish between finite-horizon and steady-state simulations. Two popular methods for estimating steady-state performance measures — the batch means and regenerative methods — are discussed as well.

Chapter 5 deals with variance reduction techniques in Monte Carlo simulation, such as antithetic and common random numbers, control random variables, conditional Monte Carlo, stratified sampling, and importance sampling. Using importance sampling, one can often achieve substantial (sometimes dramatic) variance reduction, in particular when estimating rare-event probabilities. While dealing with importance sampling, we present two alternative approaches, called the *variance minimization* and the *cross-entropy* methods. Special attention is paid to importance sampling algorithms in which paths are generated in a sequential manner. Further improvements of such algorithms are obtained by resampling successful paths, giving rise to *sequential importance resampling* algorithms. We illustrate their use via a nonlinear filtering example. In addition, this chapter contains two new importance sampling based methods, called *the transform likelihood ratio method* and *the screening method* for variance reduction. The former presents a simple, convenient and unifying way of constructing efficient importance sampling estimators, whereas the latter ensures lowering of the dimensionality of the importance sampling density. This is accomplished by identifying (screening out) the most important (bottleneck) parameters to be used in the importance sampling distribution. As results, the accuracy of the importance sampling estimator increases substantially.

Chapter 6 gives a concise treatment of the generic *Markov chain Monte Carlo* (MCMC) method for *approximately* generating samples from an arbitrary distribution. We discuss the classic Metropolis–Hastings algorithm and the Gibbs sampler. In the former, one simulates a Markov chain such that its stationary distribution coincides with the target distribution, while in the latter the underlying Markov chain is constructed on the basis of a sequence of conditional distributions. We also deal with applications of MCMC in Bayesian statistics, and explain how MCMC is used to sample from the Boltzmann distribution for the Ising and Potts models, which are extensively used in statistical mechanics. Moreover, we show how MCMC is used in the simulated annealing method to find the global minimum of a multi-extremal function. We also show that both the Metropolis–Hastings and Gibbs samplers can be viewed as special cases of a general MCMC algorithm and then present two more modifications, namely the *slice* and the *reversible jump* samplers.

Chapter 7 is on sensitivity analysis and Monte Carlo optimization of simulated systems. Because of their complexity, the performance evaluation of discrete-event systems is usually studied by simulation, and the simulation is often associated with the estimation of the performance function with respect to some controllable parameters. Sensitivity analysis is concerned with evaluating sensitivities (gradients, Hessians, etc.) of the performance function with respect to system parameters.

This provides guidance to operational decisions and to selecting system parameters that optimize the performance measures. Monte Carlo optimization deals with solving stochastic programs, that is, optimization problems where the objective function and some of the constraints are unknown and need to be obtained via simulation. We deal with sensitivity analysis and optimization of both static and dynamic models. We introduce the celebrated *score function* method for sensitivity analysis, and two alternative methods for Monte Carlo optimization, the so-called *stochastic approximation* and *stochastic counterpart* methods. In particular, in the latter method, we show how using a single simulation experiment one can approximate quite accurately the true unknown optimal solution of the original deterministic program.

Chapter 8 deals with the *cross-entropy* (CE) method, which was introduced by the first author in 1997 as an adaptive algorithm for rare-event estimation using a cross-entropy minimization technique. It was soon realized that the underlying ideas had a much wider range of application than just in rare-event simulation: they could be readily adapted to tackle quite general combinatorial and multi-extremal optimization problems, including many problems associated with learning algorithms and neural computation. We provide a gradual introduction to the CE method, and show its elegance and versatility. In particular, we present a general CE algorithm for the estimation of rare-event probabilities and then slightly modify it for solving combinatorial optimization problems. We discuss applications of the CE method to several combinatorial optimization problems, such as the max-cut problem and the traveling salesman problem, and provide supportive numerical results on its effectiveness. Due to its versatility, tractability, and simplicity, the CE method has potentially a diverse range of applications, for example, in computational biology, DNA sequence alignment, graph theory, and scheduling. Over the last 10 years many hundreds of papers have been written on the theory and applications of CE. For more details see the site `www.cemethod.org`, our book *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning* (Springer, 2004), and search in the wikipedia under “cross-entropy method”. The chapter concludes with a discussion of the minimum cross-entropy (MinxEnt) optimization program.

Chapter 9 introduces the *splitting* method, which uses a sequential sampling plan to decompose a “difficult” problem into a sequence of “easy” problems. The method was originally designed for rare-event simulation, but it has developed into a highly versatile “particle MCMC” algorithm that can be used for rare-event estimation, optimization, and sampling. The chapter presents various splitting algorithms for dynamic and static simulation models, and demonstrates how they can be used to (1) estimate rare-event probabilities, (2) solve hard counting problems, (3) find solutions to challenging optimization problems, and (4) sample from complicated probability distributions. The chapter features a wide variety of case studies and numerical experiments, demonstrating the effectiveness of the method.

Many combinatorial problems can be formulated in terms of searching or counting the total cost of a tree. Chapter 10 presents a new Monte Carlo called *stochastic enumeration* (SE) that is well suited to solve such problems by generating random paths through the tree in a parallel fashion. The SE algorithm can be viewed as a sequential importance sampling method on a “hyper-tree” whose vertices are sets of vertices of the original tree. By combining SE with fast polynomial decision algorithms, we show how it can be used for counting #P-complete problems, such as the number of satisfiability assignments, number of paths in a general network, and the number of perfect matchings in a graph. The usefulness of the method is illustrated via a suite of numerical examples.

The appendix features a variety of supplementary topics, including a brief introduction to exponential families, the discrete-time Kalman filter, and the Cholesky square root method. The computational complexity of randomized algorithms is also discussed. An extensive range of exercises is provided at the end of each chapter.

In addition to two brand-new chapters (Chapters 9 and 10), this third edition offers substantial updates on a range of topics. The material on random number generation has been extensively revised by including state-of-the-art combined multiple recursive generators and the Mersenne Twister. The material on stochastic process generation has been extended by including the simulation of Gaussian processes, Brownian motion, and diffusion processes. The variance reduction chapter now includes a discussion of the novel multi-level Monte Carlo method. Our treatment of sequential importance has been significantly modified by emphasizing the significance of importance resampling. This addition also prepares for the particle MCMC approach in the new splitting chapter. The cross-entropy chapter is further enhanced by adding new insights into likelihood ratio degeneration, leading to the single-level improved CE algorithm. Twenty-five more questions have been added, along with their solutions in the *online solutions manual* that accompanies this book. Finally, to facilitate their implementation, most algorithms have been (re-)written in pseudo-code with flow control.

REUVEN RUBINSTEIN AND DIRK KROESE

Haifa and Brisbane
July 2016

ACKNOWLEDGMENTS

We thank all who contributed to this book. Robert Smith and Zelda Zabinski read and provided useful suggestions on Chapter 6. Alex Shapiro gave a detailed account of the complexity of stochastic programming problems (Section A.8.4). Pierre L'Ecuyer communicated the counterexample in Section 9.10 and Zdravko Botev suggested the multilevel Monte Carlo example in Section 5.6. Jim Spall kindly provided feedback on the second edition, which we have incorporated in the third edition. Josh Chan, Tom Taimre, and Zdravko Botev helped proofread the new material in this book, which is highly appreciated.

We are particularly grateful to the many undergraduate and graduate students at the Technion and the University of Queensland who helped make this book possible and whose valuable ideas and experiments were extremely encouraging and motivating. Qibin Duan, Morgan Grant, Robert Salomone, Rohan Shah, and Erli Wang read through the new material, solved the new exercises, and provided excellent feedback. Special thanks goes out to Slava Vaisman for his help on many computational problems that we encountered, and for the many fruitful discussions we had on the splitting and stochastic enumeration methods.

This book was supported by by the Australian Research Council *Centre of Excellence for Mathematical & Statistical Frontiers*, under grant number CE140100049.

RYR, DPK

CHAPTER 1

PRELIMINARIES

1.1 INTRODUCTION

The purpose of this chapter is to review some basic facts from probability, information theory, and optimization. In particular, Sections 1.2–1.11 summarize the main points from probability theory. Sections 1.12–1.14 describe various fundamental stochastic processes, such as Poisson, Markov, and Gaussian processes. Elements of information theory are given in Section 1.15, and Section 1.16 concludes with an outline of convex optimization theory.

1.2 RANDOM EXPERIMENTS

The basic notion in probability theory is that of a *random experiment*: an experiment whose outcome cannot be determined in advance. The most fundamental example is the experiment where a fair coin is tossed a number of times. For simplicity suppose that the coin is tossed three times. The *sample space*, denoted Ω, is the set of all possible outcomes of the experiment. In this case Ω has eight possible outcomes:

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} ,$$

where, for example, HTH means that the first toss is heads, the second tails, and the third heads.

Simulation and the Monte Carlo Method, Third Edition. By R. Y. Rubinstein and D. P. Kroese

Copyright © 2017 John Wiley & Sons, Inc. Published 2017 by John Wiley & Sons, Inc.

Subsets of the sample space are called *events*. For example, the event A that the third toss is heads is

$$A = \{HHH, HTH, THH, TTH\} \ .$$

We say that event A *occurs* if the outcome of the experiment is one of the elements in A. Since events are sets, we can apply the usual set operations to them. For example, the event $A \cup B$, called the *union* of A and B, is the event that A or B or both occur, and the event $A \cap B$, called the *intersection* of A and B, is the event that A and B both occur. Similar notation holds for unions and intersections of more than two events. The event A^c, called the *complement* of A, is the event that A does not occur. Two events A and B that have no outcomes in common, that is, their intersection is empty, are called *disjoint* events. The main step is to specify the probability of each event.

Definition 1.2.1 (Probability) A *probability* $\mathbb{P}$ is a rule that assigns a number $0 \leqslant \mathbb{P}(A) \leqslant 1$ to each event A, such that $\mathbb{P}(\Omega) = 1$, and such that for any sequence $A_1, A_2, \ldots$ of disjoint events

$$\mathbb{P}\Big(\bigcup_i A_i\Big) = \sum_i \mathbb{P}(A_i) \ . \tag{1.1}$$

Equation (1.1) is referred to as the *sum rule* of probability. It states that if an event can happen in a number of different ways, but not simultaneously, the probability of that event is simply the sum of the probabilities of the comprising events.

For the fair coin toss experiment the probability of any event is easily given. Namely, because the coin is fair, each of the eight possible outcomes is equally likely, so that $\mathbb{P}(\{HHH\}) = \cdots = \mathbb{P}(\{TTT\}) = 1/8$. Since any event A is the union of the "elementary" events $\{HHH\}, \ldots, \{TTT\}$, the sum rule implies that

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \ , \tag{1.2}$$

where $|A|$ denotes the number of outcomes in A and $|\Omega| = 8$. More generally, if a random experiment has finitely many and equally likely outcomes, the probability is always of the form (1.2). In that case the calculation of probabilities reduces to counting.

1.3 CONDITIONAL PROBABILITY AND INDEPENDENCE

How do probabilities change when we know that some event $B \subset \Omega$ has occurred? Given that the outcome lies in B, the event A will occur if and only if $A \cap B$ occurs, and the relative chance of A occurring is therefore $\mathbb{P}(A \cap B)/\mathbb{P}(B)$. This leads to the definition of the *conditional probability* of A given B:

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \ . \tag{1.3}$$

For example, suppose that we toss a fair coin three times. Let B be the event that the total number of heads is two. The conditional probability of the event A that the first toss is heads, given that B occurs, is $(2/8)/(3/8) = 2/3$.

Rewriting (1.3) and interchanging the role of A and B gives the relation $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \,|\, A)$. This can be generalized easily to the *product rule* of probability, which states that for any sequence of events $A_1, A_2, \ldots, A_n$,

$$\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}) \;, \tag{1.4}$$

using the abbreviation $A_1 A_2 \cdots A_k \;\equiv\; A_1 \cap A_2 \cap \cdots \cap A_k$.

Suppose that $B_1, B_2, \ldots, B_n$ is a *partition* of Ω. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is Ω. Then, by the sum rule, $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \cap B_i)$ and hence, by the definition of conditional probability, we have the *law of total probability*:

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i) \;. \tag{1.5}$$

Combining this with the definition of conditional probability gives *Bayes' rule*:

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\,\mathbb{P}(B_j)}{\sum_{i=1}^n \mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i)} \;. \tag{1.6}$$

Independence is of crucial importance in probability and statistics. Loosely speaking, it models the lack of information between events. Two events A and B are said to be *independent* if the knowledge that B has occurred does not change the probability that A occurs. That is, A, B independent $\Leftrightarrow \mathbb{P}(A \,|\, B) = \mathbb{P}(A)$. Since $\mathbb{P}(A \,|\, B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$, an alternative definition of independence is

$$A,\, B \text{ independent} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \;.$$

This definition covers the case where $B = \emptyset$ (empty set). We can extend this definition to arbitrarily many events.

Definition 1.3.1 (Independence) The events $A_1, A_2, \ldots,$ are said to be *independent* if for any k and any choice of distinct indexes $i_1, \ldots, i_k$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\,\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) \;.$$

Remark 1.3.1 In most cases independence of events is a model assumption. That is, we assume that there exists a $\mathbb{P}$ such that certain events are independent.

■ EXAMPLE 1.1

We toss a biased coin n times. Let p be the probability of heads (for a fair coin $p = 1/2$). Let A_i denote the event that the i-th toss yields heads, $i = 1, \ldots, n$. Then $\mathbb{P}$ should be such that the events $A_1, \ldots, A_n$ are independent, and $\mathbb{P}(A_i) = p$ for all i. These two rules completely specify $\mathbb{P}$. For example, the probability that the first k throws are heads and the last $n - k$ are tails is

$$\begin{aligned} \mathbb{P}(A_1 \cdots A_k A_{k+1}^c \cdots A_n^c) &= \mathbb{P}(A_1) \cdots \mathbb{P}(A_k)\,\mathbb{P}(A_{k+1}^c) \cdots \mathbb{P}(A_n^c) \\ &= p^k (1-p)^{n-k}. \end{aligned}$$

1.4 RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

Specifying a model for a random experiment via a complete description of Ω and $\mathbb{P}$ may not always be convenient or necessary. In practice, we are only interested in certain observations (i.e., numerical measurements) in the experiment. We incorporate these into our modeling process via the introduction of *random variables*, usually denoted by capital letters from the last part of the alphabet (e.g., X, $X_1, X_2, \ldots, Y, Z$).

■ EXAMPLE 1.2

We toss a biased coin n times, with p the probability of heads. Suppose that we are interested only in the number of heads, say X. Note that X can take any of the values in $\{0, 1, \ldots, n\}$. The *probability distribution* of X is given by the *binomial formula*

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n\,. \tag{1.7}$$

Namely, by Example 1.1, each elementary event $\{HTH \cdots T\}$ with exactly k heads and $n - k$ tails has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ such events.

The probability distribution of a general random variable X — identifying such probabilities as $\mathbb{P}(X = x), \mathbb{P}(a \leqslant X \leqslant b)$, and so on — is completely specified by the *cumulative distribution function* (cdf), defined by

$$F(x) = \mathbb{P}(X \leqslant x),\ \ x \in \mathbb{R}\,.$$

A random variable X is said to have a *discrete* distribution if, for some finite or countable set of values $x_1, x_2, \ldots$, $\mathbb{P}(X = x_i) > 0$, $i = 1, 2, \ldots$ and $\sum_i \mathbb{P}(X = x_i) = 1$. The function $f(x) = \mathbb{P}(X = x)$ is called the *probability mass function* (pmf) of X — but see Remark 1.4.1.

■ EXAMPLE 1.3

Toss two fair dice and let M be the largest face value showing. The pmf of M is given by

m	1	2	3	4	5	6	$\sum$
$f(m)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

For example, to get $M = 3$, either $(1, 3), (2, 3), (3, 3), (3, 2)$, or $(3, 1)$ has to be thrown, each of which happens with probability $1/36$.

A random variable X is said to have a *continuous* distribution if there exists a positive function f with total integral 1, such that for all a, b,

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f(u)\,\mathrm{d}u\,. \tag{1.8}$$

The function f is called the *probability density function* (pdf) of X. Note that in the continuous case the cdf is given by

$$F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u\ ,$$

and f is the derivative of F. We can interpret $f(x)$ as the probability "density" at $X = x$ in the sense that

$$\mathbb{P}(x \leqslant X \leqslant x+h) = \int_{x}^{x+h} f(u)\,\mathrm{d}u \approx h\, f(x)\ .$$

Remark 1.4.1 (Probability Density) Note that we have deliberately used the *same* symbol, f, for both pmf and pdf. This is because the pmf and pdf play very similar roles and can, in more advanced probability theory, both be viewed as particular instances of the general notion of *probability density.* To stress this viewpoint, we will call f in *both* the discrete and continuous case the pdf or (probability) density (function).

1.5 SOME IMPORTANT DISTRIBUTIONS

Tables 1.1 and 1.2 list a number of important continuous and discrete distributions. We will use the notation $X \sim f$, $X \sim F$, or $X \sim \mathsf{Dist}$ to signify that X has a pdf f, a cdf F or a distribution Dist. We sometimes write f_X instead of f to stress that the pdf refers to the random variable X. Note that in Table 1.1, Γ is the gamma function: $\Gamma(\alpha) = \int_0^\infty \mathrm{e}^{-x} x^{\alpha-1}\,\mathrm{d}x, \quad \alpha > 0$.

Table 1.1: Commonly used continuous distributions.

Name	Notation	$f(x)$	$x \in$	Parameters
Uniform	$\mathsf{U}[\alpha,\beta]$	$\dfrac{1}{\beta-\alpha}$	$[\alpha,\beta]$	$\alpha < \beta$
Normal	$\mathsf{N}(\mu,\sigma^2)$	$\dfrac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{R}$	$\sigma > 0,\ \mu \in \mathbb{R}$
Gamma	$\mathsf{Gamma}(\alpha,\lambda)$	$\dfrac{\lambda^\alpha x^{\alpha-1}\mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}$	$\mathbb{R}_+$	$\alpha,\lambda > 0$
Exponential	$\mathsf{Exp}(\lambda)$	$\lambda\,\mathrm{e}^{-\lambda x}$	$\mathbb{R}_+$	$\lambda > 0$
Beta	$\mathsf{Beta}(\alpha,\beta)$	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}$	$[0,1]$	$\alpha,\beta > 0$
Weibull	$\mathsf{Weib}(\alpha,\lambda)$	$\alpha\lambda\,(\lambda x)^{\alpha-1}\,\mathrm{e}^{-(\lambda x)^\alpha}$	$\mathbb{R}_+$	$\alpha,\lambda > 0$
Pareto	$\mathsf{Pareto}(\alpha,\lambda)$	$\alpha\lambda\,(1+\lambda x)^{-(\alpha+1)}$	$\mathbb{R}_+$	$\alpha,\lambda > 0$

Table 1.2: Commonly used discrete distributions.

Name	Notation	$f(x)$	$x \in$	Parameters
Bernoulli	$\mathsf{Ber}(p)$	$p^x(1-p)^{1-x}$	$\{0,1\}$	$0 \leqslant p \leqslant 1$
Binomial	$\mathsf{Bin}(n,p)$	$\binom{n}{x} p^x(1-p)^{n-x}$	$\{0,1,\ldots,n\}$	$0 \leqslant p \leqslant 1$, $n \in \mathbb{N}$
Discrete uniform	$\mathsf{DU}\{1,\ldots,n\}$	$\frac{1}{n}$	$\{1,\ldots,n\}$	$n \in \{1,2,\ldots\}$
Geometric	$\mathsf{G}(p)$	$p(1-p)^{x-1}$	$\{1,2,\ldots\}$	$0 \leqslant p \leqslant 1$
Poisson	$\mathsf{Poi}(\lambda)$	$\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}$	$\mathbb{N}$	$\lambda > 0$

1.6 EXPECTATION

It is often useful to consider different kinds of numerical characteristics of a random variable. One such quantity is the expectation, which measures the mean value of the distribution.

Definition 1.6.1 (Expectation) Let X be a random variable with pdf f. The *expectation* (or expected value or mean) of X, denoted by $\mathbb{E}[X]$ (or sometimes μ), is defined by

$$\mathbb{E}[X] = \begin{cases} \sum_x x\, f(x) & \text{discrete case,} \\ \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x & \text{continuous case.} \end{cases}$$

If X is a random variable, then a function of X, such as X^2 or $\sin(X)$, is again a random variable. Moreover, the expected value of a function of X is simply a weighted average of the possible values that this function can take. That is, for any real function h

$$\mathbb{E}[h(X)] = \begin{cases} \sum_x h(x)\, f(x) & \text{discrete case,} \\ \int_{-\infty}^{\infty} h(x)\, f(x)\, \mathrm{d}x & \text{continuous case.} \end{cases}$$

Another useful quantity is the variance, which measures the spread or dispersion of the distribution.

Definition 1.6.2 (Variance) The *variance* of a random variable X, denoted by $\mathrm{Var}(X)$ (or sometimes σ^2), is defined by

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \,.$$

The square root of the variance is called the *standard deviation*. Table 1.3 lists the expectations and variances for some well-known distributions.

Table 1.3: Expectations and variances for some well-known distributions.

Dist.	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	Dist.	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
$\mathsf{Bin}(n,p)$	np	$np(1-p)$	$\mathsf{Gamma}(\alpha,\lambda)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$\mathsf{G}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\mathsf{N}(\mu,\sigma^2)$	μ	σ^2
$\mathsf{Poi}(\lambda)$	λ	λ	$\mathsf{Beta}(\alpha,\beta)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$
$\mathsf{U}(\alpha,\beta)$	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$	$\mathsf{Weib}(\alpha,\lambda)$	$\frac{\Gamma(1/\alpha)}{\alpha\,\lambda}$	$\frac{2\Gamma(2/\alpha)}{\alpha} - \left(\frac{\Gamma(1/\alpha)}{\alpha\,\lambda}\right)^2$
$\mathsf{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$			

The mean and the variance do not give, in general, enough information to completely specify the distribution of a random variable. However, they may provide useful bounds. We discuss two such bounds. Suppose X can only take nonnegative values and has pdf f. For any $x > 0$, we can write

$$\begin{aligned}\mathbb{E}[X] &= \int_0^x t f(t)\,\mathrm{d}t + \int_x^\infty t f(t)\,\mathrm{d}t \geqslant \int_x^\infty t f(t)\,\mathrm{d}t \\ &\geqslant \int_x^\infty x f(t)\,\mathrm{d}t = x\,\mathbb{P}(X \geqslant x)\ ,\end{aligned}$$

from which follows the *Markov inequality:* if $X \geqslant 0$, then for all $x > 0$,

$$\mathbb{P}(X \geqslant x) \leqslant \frac{\mathbb{E}[X]}{x}\ . \tag{1.9}$$

If we also know the variance of a random variable, we can give a tighter bound. Namely, for any random variable X with mean μ and variance σ^2, we have

$$\mathbb{P}(|X-\mu| \geqslant x) \leqslant \frac{\sigma^2}{x^2}\ . \tag{1.10}$$

This is called the *Chebyshev inequality.* The proof is as follows: Let $D^2 = (X-\mu)^2$; then, by the Markov inequality (1.9) and the definition of the variance,

$$\mathbb{P}(D^2 \geqslant x^2) \leqslant \frac{\sigma^2}{x^2}\ .$$

Also, note that the event $\{D^2 \geqslant x^2\}$ is equivalent to the event $\{|X-\mu| \geqslant x\}$, so that (1.10) follows.

1.7 JOINT DISTRIBUTIONS

Often a random experiment is described by more than one random variable. The theory for multiple random variables is similar to that for a single random variable.

Let $X_1, \ldots, X_n$ be random variables describing some random experiment. We can accumulate these into a *random vector* $\mathbf{X} = (X_1, \ldots, X_n)$. More generally, a collection $\{X_t, t \in \mathscr{T}\}$ of random variables is called a *stochastic process*. The set $\mathscr{T}$ is called the *parameter set* or *index set* of the process. It may be discrete (e.g., $\mathbb{N}$ or $\{1, \ldots, 10\}$) or continuous (e.g., $\mathbb{R}_+ = [0, \infty)$ or $[1, 10]$). The set of possible values for the stochastic process is called the *state space*.

The joint distribution of $X_1, \ldots, X_n$ is specified by the *joint cdf*

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) .$$

The *joint pdf* f is given, in the discrete case, by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$, and in the continuous case f is such that

$$\mathbb{P}(\mathbf{X} \in \mathscr{B}) = \int_{\mathscr{B}} f(x_1, \ldots, x_n)\; \mathrm{d}x_1 \ldots \mathrm{d}x_n$$

for any (measurable) region $\mathscr{B}$ in $\mathbb{R}^n$. The marginal pdfs can be recovered from the joint pdf by integration or summation. For example, in the case of a continuous random vector (X, Y) with joint pdf f, the pdf f_X of X is found as

$$f_X(x) = \int f(x, y)\, \mathrm{d}y .$$

Suppose that X and Y are both discrete or both continuous, with joint pdf f, and suppose that $f_X(x) > 0$. Then the *conditional pdf* of Y given $X = x$ is given by

$$f_{Y|X}(y \,|\, x) = \frac{f(x, y)}{f_X(x)} \quad \text{for all } y .$$

The corresponding *conditional expectation* is (in the continuous case)

$$\mathbb{E}[Y \,|\, X = x] = \int y\, f_{Y|X}(y \,|\, x)\, \mathrm{d}y .$$

Note that $\mathbb{E}[Y \,|\, X = x]$ is a function of x, say $h(x)$. The corresponding random variable $h(X)$ is written as $\mathbb{E}[Y \,|\, X]$. It can be shown (see, for example, [3]) that its expectation is simply the expectation of Y, that is,

$$\mathbb{E}[\mathbb{E}[Y \,|\, X]] = \mathbb{E}[Y] . \tag{1.11}$$

When the conditional distribution of Y given X is identical to that of Y, X and Y are said to be independent. More precisely:

Definition 1.7.1 (Independent Random Variables) The random variables $X_1, \ldots, X_n$ are called *independent* if for all events $\{X_i \in A_i\}$ with $A_i \subset \mathbb{R}$, $i = 1, \ldots, n$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n) .$$

A direct consequence of the definition above for independence is that random variables $X_1, \ldots, X_n$ with joint pdf f (discrete or continuous) are independent if and only if

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \tag{1.12}$$

for all $x_1, \ldots, x_n$, where $\{f_{X_i}\}$ are the marginal pdfs.

■ EXAMPLE 1.4 Bernoulli Sequence

Consider the experiment where we flip a biased coin n times, with probability p of heads. We can model this experiment in the following way. For $i = 1, \ldots, n$, let X_i be the result of the i-th toss: $\{X_i = 1\}$ means heads (or success), $\{X_i = 0\}$ means tails (or failure). Also, let

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0), \quad i = 1, 2, \ldots, n \, .$$

Last, assume that $X_1, \ldots, X_n$ are independent. The sequence $\{X_i, i = 1, 2, \ldots\}$ is called a *Bernoulli sequence* or *Bernoulli process* with success probability p. Let $X = X_1 + \cdots + X_n$ be the total number of successes in n trials (tosses of the coin). Denote by $\mathscr{B}$ the set of all binary vectors $\mathbf{x} = (x_1, \ldots, x_n)$ such that $\sum_{i=1}^n x_i = k$. Note that $\mathscr{B}$ has $\binom{n}{k}$ elements. We now have

$$\begin{aligned}
\mathbb{P}(X = k) &= \sum_{\mathbf{x} \in \mathscr{B}} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \\
&= \sum_{\mathbf{x} \in \mathscr{B}} \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) = \sum_{\mathbf{x} \in \mathscr{B}} p^k (1-p)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k} \, .
\end{aligned}$$

In other words, $X \sim \mathsf{Bin}(n, p)$. Compare this with Example 1.2.

Remark 1.7.1 An *infinite* sequence $X_1, X_2, \ldots$ of random variables is called independent if for any finite choice of parameters $i_1, i_2, \ldots, i_n$ (none of them the same) the random variables $X_{i_1}, \ldots, X_{i_n}$ are independent. Many probabilistic models involve random variables $X_1, X_2, \ldots$ that are *independent and identically distributed*, abbreviated as *iid*. We will use this abbreviation throughout this book.

Similar to the one-dimensional case, the expected value of any real-valued function h of $X_1, \ldots, X_n$ is a weighted average of all values that this function can take. Specifically, in the continuous case,

$$\mathbb{E}[h(X_1, \ldots, X_n)] = \int \cdots \int h(x_1, \ldots, x_n) \, f(x_1, \ldots, x_n) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n \, .$$

As a direct consequence of the definitions of expectation and independence, we have

$$\mathbb{E}[a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n] = a + b_1 \mu_1 + \cdots + b_n \mu_n \tag{1.13}$$

for any sequence of random variables X_1, X_2, $\ldots$, X_n with expectations $\mu_1, \mu_2, \ldots, \mu_n$, where a, b_1, b_2, $\ldots$, b_n are constants. Similarly, for *independent* random variables, we have

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mu_1 \, \mu_2 \cdots \mu_n \, .$$

The *covariance* of two random variables X and Y with expectations $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$, respectively, is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \, .$$

This is a measure for the amount of linear dependency between the variables. A scaled version of the covariance is given by the *correlation coefficient*,

$$\varrho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\,\sigma_Y},$$

where $\sigma_X^2 = \operatorname{Var}(X)$ and $\sigma_Y^2 = \operatorname{Var}(Y)$. It can be shown that the correlation coefficient always lies between -1 and 1; see Problem 1.13.

For easy reference, Table 1.4 lists some important properties of the variance and covariance. The proofs follow directly from the definitions of covariance and variance and the properties of the expectation.

Table 1.4: Properties of variance and covariance.

1	$\operatorname{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
2	$\operatorname{Var}(aX+b) = a^2\operatorname{Var}(X)$
3	$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$
4	$\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$
5	$\operatorname{Cov}(aX+bY,Z) = a\operatorname{Cov}(X,Z) + b\operatorname{Cov}(Y,Z)$
6	$\operatorname{Cov}(X,X) = \operatorname{Var}(X)$
7	$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)$
8	X and Y indep. $\Longrightarrow \operatorname{Cov}(X,Y) = 0$

As a consequence of properties 2 and 7, for any sequence of *independent* random variables $X_1, \ldots, X_n$ with variances $\sigma_1^2, \ldots, \sigma_n^2$,

$$\operatorname{Var}(a + b_1X_1 + b_2X_2 + \cdots + b_nX_n) = b_1^2\,\sigma_1^2 + \cdots + b_n^2\,\sigma_n^2 \tag{1.14}$$

for any choice of constants a and $b_1, \ldots, b_n$.

For random vectors, such as $\mathbf{X} = (X_1, \ldots, X_n)^\top$, it is convenient to write the expectations and covariances in vector notation.

Definition 1.7.2 (Expectation Vector and Covariance Matrix) For any random vector $\mathbf{X}$, we define the *expectation vector* as the vector of expectations

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])^\top .$$

The *covariance matrix* Σ is defined as the matrix whose (i,j)-th element is

$$\operatorname{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] .$$

If we define the expectation of a vector (matrix) to be the vector (matrix) of expectations, then we can write

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$$

and

$$\Sigma = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] .$$

Note that $\boldsymbol{\mu}$ and Σ take on the same role as μ and σ^2 in the one-dimensional case.

Remark 1.7.2 Note that any covariance matrix Σ is *symmetric*. In fact (see Problem 1.16), it is *positive semidefinite*, that is, for any (column) vector $\mathbf{u}$,

$$\mathbf{u}^\top \Sigma \mathbf{u} \geqslant 0 .$$

1.8 FUNCTIONS OF RANDOM VARIABLES

Suppose that $X_1, \ldots, X_n$ are measurements of a random experiment. Often we are only interested in certain *functions* of the measurements rather than the individual measurements. Here are some examples.

■ EXAMPLE 1.5

Let X be a continuous random variable with pdf f_X and let $Z = aX + b$, where $a \neq 0$. We wish to determine the pdf f_Z of Z. Suppose that $a > 0$. We have for any z

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}\big(X \leqslant (z-b)/a\big) = F_X\big((z-b)/a\big) .$$

Differentiating this with respect to z gives $f_Z(z) = f_X\big((z-b)/a\big)/a$. For $a < 0$ we similarly obtain $f_Z(z) = f_X\big((z-b)/a\big)/(-a)$. Thus, in general,

$$f_Z(z) = \frac{1}{|a|} f_X\left(\frac{z-b}{a}\right) . \tag{1.15}$$

■ EXAMPLE 1.6

Generalizing the previous example, suppose that $Z = g(X)$ for some monotonically increasing function g. To find the pdf of Z from that of X we first write

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}\left(X \leqslant g^{-1}(z)\right) = F_X\left(g^{-1}(z)\right) ,$$

where g^{-1} is the inverse of g. Differentiating with respect to z now gives

$$f_Z(z) = f_X(g^{-1}(z)) \frac{\mathrm{d}}{\mathrm{d}z} g^{-1}(z) = \frac{f_X(g^{-1}(z))}{g'(g^{-1}(z))} . \tag{1.16}$$

For monotonically decreasing functions, $\frac{\mathrm{d}}{\mathrm{d}z} g^{-1}(z)$ in the first equation needs to be replaced with its negative value.

■ EXAMPLE 1.7 Order Statistics

Let $X_1, \ldots, X_n$ be an iid sequence of random variables with common pdf f and cdf F. In many applications one is interested in the distribution of the *order statistics* $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, where $X_{(1)}$ is the smallest of the $\{X_i, i = 1, \ldots, n\}$, $X_{(2)}$ is the second smallest, and so on. The cdf of $X_{(n)}$ follows from

$$\mathbb{P}(X_{(n)} \leqslant x) = \mathbb{P}(X_1 \leqslant x, \ldots, X_n \leqslant x) = \prod_{i=1}^{n} \mathbb{P}(X_i \leqslant x) = (F(x))^n \ .$$

Similarly,

$$\mathbb{P}(X_{(1)} > x) = \mathbb{P}(X_1 > x, \ldots, X_n > x) = \prod_{i=1}^{n} \mathbb{P}(X_i > x) = (1 - F(x))^n \ .$$

Moreover, because all orderings of $X_1, \ldots, X_n$ are equally likely, it follows that the joint pdf of the ordered sample is, on the wedge $\{(x_1, \ldots, x_n) : x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n\}$, simply $n!$ times the joint density of the unordered sample and zero elsewhere.

1.8.1 Linear Transformations

Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ be a column vector in $\mathbb{R}^n$ and A an $m \times n$ matrix. The mapping $\mathbf{x} \mapsto \mathbf{z}$, with $\mathbf{z} = A\mathbf{x}$, is called a *linear transformation*. Now consider a *random* vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$, and let

$$\mathbf{Z} = A\mathbf{X} \ .$$

Then $\mathbf{Z}$ is a random vector in $\mathbb{R}^m$. In principle, if we know the joint distribution of $\mathbf{X}$, then we can derive the joint distribution of $\mathbf{Z}$. Let us first see how the expectation vector and covariance matrix are transformed.

Theorem 1.8.1 *If* $\mathbf{X}$ *has an expectation vector* $\boldsymbol{\mu}_{\mathbf{X}}$ *and covariance matrix* $\Sigma_{\mathbf{X}}$*, then the expectation vector and covariance matrix of* $\mathbf{Z} = A\mathbf{X}$ *are given by*

$$\boldsymbol{\mu}_{\mathbf{Z}} = A\boldsymbol{\mu}_{\mathbf{X}} \tag{1.17}$$

and

$$\Sigma_{\mathbf{Z}} = A\, \Sigma_{\mathbf{X}}\, A^\top \ . \tag{1.18}$$

Proof: We have $\boldsymbol{\mu}_{\mathbf{Z}} = \mathbb{E}[\mathbf{Z}] = \mathbb{E}[A\mathbf{X}] = A\,\mathbb{E}[\mathbf{X}] = A\boldsymbol{\mu}_{\mathbf{X}}$ and

$$\begin{aligned} \Sigma_{\mathbf{Z}} &= \mathbb{E}[(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})^\top] = \mathbb{E}[A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}))^\top] \\ &= A\,\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top]A^\top \\ &= A\,\Sigma_{\mathbf{X}}\,A^\top \ . \end{aligned}$$

□

Suppose that A is an invertible $n \times n$ matrix. If $\mathbf{X}$ has a joint density $f_{\mathbf{X}}$, what is the joint density $f_{\mathbf{Z}}$ of $\mathbf{Z}$? Consider Figure 1.1. For any fixed $\mathbf{x}$, let $\mathbf{z} = A\mathbf{x}$. Hence, $\mathbf{x} = A^{-1}\mathbf{z}$. Consider the n-dimensional cube $C = [z_1, z_1 + h] \times \cdots \times [z_n, z_n + h]$. Let D be the image of C under A^{-1}, that is, the parallelepiped of all points $\mathbf{x}$ such that $A\mathbf{x} \in C$. Then,

$$\mathbb{P}(\mathbf{Z} \in C) \approx h^n\, f_{\mathbf{Z}}(\mathbf{z}) \ .$$

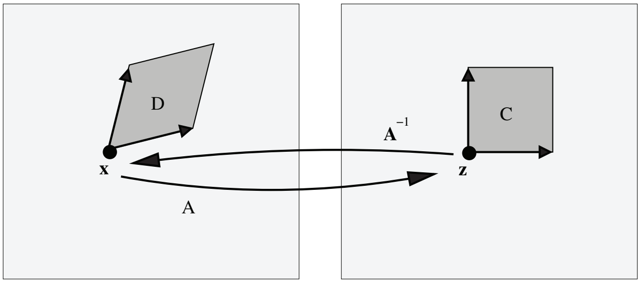


Figure 1.1: Linear transformation.

Now recall from linear algebra (e.g., [5]) that any matrix B linearly transforms an n-dimensional rectangle with volume V into an n-dimensional parallelepiped with volume $V\,|B|$, where $|B| = |\det(B)|$. Thus,

$$\mathbb{P}(\mathbf{Z} \in C) = \mathbb{P}(\mathbf{X} \in D) \approx h^n |A^{-1}|\, f_{\mathbf{X}}(\mathbf{x}) = h^n |A|^{-1}\, f_{\mathbf{X}}(\mathbf{x})\ .$$

Letting h go to 0, we obtain

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{f_{\mathbf{X}}(A^{-1}\mathbf{z})}{|A|},\ \ \mathbf{z} \in \mathbb{R}^n. \tag{1.19}$$

1.8.2 General Transformations

We can apply reasoning similar to that above to deal with general transformations $\mathbf{x} \mapsto \boldsymbol{g}(\mathbf{x})$, written out as

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{pmatrix}.$$

For a fixed $\mathbf{x}$, let $\mathbf{z} = \boldsymbol{g}(\mathbf{x})$. Suppose that $\boldsymbol{g}$ is invertible; hence $\mathbf{x} = \boldsymbol{g}^{-1}(\mathbf{z})$. Any infinitesimal n-dimensional rectangle at $\mathbf{x}$ with volume V is transformed into an n-dimensional parallelepiped at $\mathbf{z}$ with volume $V\,|J_{\mathbf{x}}(\boldsymbol{g})|$, where $J_{\mathbf{x}}(\boldsymbol{g})$ is the *matrix of Jacobi* at $\mathbf{x}$ of the transformation $\boldsymbol{g}$, that is,

$$J_{\mathbf{x}}(\boldsymbol{g}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix}.$$

Now consider a random column vector $\mathbf{Z} = \boldsymbol{g}(\mathbf{X})$. Let C be a small cube around $\mathbf{z}$ with volume h^n. Let D be the image of C under $\boldsymbol{g}^{-1}$. Then, as in the linear case,

$$\mathbb{P}(\mathbf{Z} \in C) \approx h^n\, f_{\mathbf{Z}}(\mathbf{z}) \approx h^n |J_{\mathbf{z}}(\boldsymbol{g}^{-1})|\, f_{\mathbf{X}}(\mathbf{x})\ .$$

Hence we have the transformation rule

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\boldsymbol{g}^{-1}(\mathbf{z}))\, |J_{\mathbf{z}}(\boldsymbol{g}^{-1})|,\ \ \mathbf{z} \in \mathbb{R}^n. \tag{1.20}$$

(Note: $|J_{\mathbf{z}}(\boldsymbol{g}^{-1})| = 1/|J_{\mathbf{x}}(\boldsymbol{g})|$.)

Remark 1.8.1 In most coordinate transformations, it is $\boldsymbol{g}^{-1}$ that is given — that is, an expression for $\mathbf{x}$ as a function of $\mathbf{z}$ rather than $\boldsymbol{g}$.

1.9 TRANSFORMS

Many calculations and manipulations involving probability distributions are facilitated by the use of transforms. Two typical examples are the *probability generating function* of a positive integer-valued random variable N, defined by

$$G(z) = \mathbb{E}[z^N] = \sum_{k=0}^{\infty} z^k \, \mathbb{P}(N = k) \; , \quad |z| \leqslant 1 \; ,$$

and the *Laplace transform* of a positive random variable X defined, for $s \geqslant 0$, by

$$L(s) = \mathbb{E}[\mathrm{e}^{-sX}] = \begin{cases} \sum_x \mathrm{e}^{-sx} \, f(x) & \text{discrete case,} \\ \int_0^\infty \mathrm{e}^{-sx} \, f(x) \, \mathrm{d}x & \text{continuous case.} \end{cases}$$

All transforms share an important *uniqueness property*: two distributions are the same if and only if their respective transforms are the same.

■ EXAMPLE 1.8

Let $M \sim \mathsf{Poi}(\mu)$; then its probability generating function is given by

$$G(z) = \sum_{k=0}^{\infty} z^k \, \mathrm{e}^{-\mu} \, \frac{\mu^k}{k!} = \mathrm{e}^{-\mu} \sum_{k=0}^{\infty} \frac{(z\mu)^k}{k!} = \mathrm{e}^{-\mu} \mathrm{e}^{z\mu} = \mathrm{e}^{-\mu(1-z)} \; . \tag{1.21}$$

Now let $N \sim \mathsf{Poi}(\nu)$ independently of M. Then the probability generating function of $M + N$ is given by

$$\mathbb{E}[z^{M+N}] = \mathbb{E}[z^M] \, \mathbb{E}[z^N] = \mathrm{e}^{-\mu(1-z)} \mathrm{e}^{-\nu(1-z)} = \mathrm{e}^{-(\mu+\nu)(1-z)} \; .$$

Thus, by the uniqueness property, $M + N \sim \mathsf{Poi}(\mu + \nu)$.

■ EXAMPLE 1.9

The Laplace transform of $X \sim \mathsf{Gamma}(\alpha, \lambda)$ is given by

$$\begin{aligned} \mathbb{E}[\mathrm{e}^{-sX}] &= \int_0^\infty \frac{\mathrm{e}^{-\lambda x} \, \lambda^\alpha \, x^{\alpha-1}}{\Gamma(\alpha)} \, \mathrm{e}^{-sx} \, \mathrm{d}x \\ &= \left(\frac{\lambda}{\lambda + s} \right)^\alpha \int_0^\infty \frac{\mathrm{e}^{-(\lambda+s)x} \, (\lambda + s)^\alpha \, x^{\alpha-1}}{\Gamma(\alpha)} \, \mathrm{d}x \\ &= \left(\frac{\lambda}{\lambda + s} \right)^\alpha \; . \end{aligned}$$

As a special case, the Laplace transform of the $\mathsf{Exp}(\lambda)$ distribution is given by $\lambda/(\lambda + s)$. Now let $X_1, \ldots, X_n$ be iid $\mathsf{Exp}(\lambda)$ random variables. The Laplace transform of $S_n = X_1 + \cdots + X_n$ is

$$\mathbb{E}[\mathrm{e}^{-sS_n}] = \mathbb{E}[\mathrm{e}^{-sX_1} \cdots \mathrm{e}^{-sX_n}] = \mathbb{E}[\mathrm{e}^{-sX_1}] \, \cdots \, \mathbb{E}[\mathrm{e}^{-sX_n}] = \left(\frac{\lambda}{\lambda + s} \right)^n \; ,$$

which shows that $S_n \sim \mathsf{Gamma}(n, \lambda)$.

1.10 JOINTLY NORMAL RANDOM VARIABLES

It is helpful to view normally distributed random variables as simple transformations of *standard normal* — that is, $\mathsf{N}(0,1)$-distributed — random variables. In particular, let $X \sim \mathsf{N}(0,1)$. Then X has density f_X given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{x^2}{2}}.$$

Now consider the transformation $Z = \mu + \sigma X$. Then, by (1.15), Z has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}}.$$

In other words, $Z \sim \mathsf{N}(\mu,\sigma^2)$. We can also state this as follows: if $Z \sim \mathsf{N}(\mu,\sigma^2)$, then $(Z-\mu)/\sigma \sim \mathsf{N}(0,1)$. This procedure is called *standardization*.

We now generalize this to n dimensions. Let $X_1,\dots,X_n$ be independent and standard normal random variables. The joint pdf of $\mathbf{X} = (X_1,\dots,X_n)^\top$ is given by

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}\mathrm{e}^{-\frac{1}{2}\mathbf{x}^\top\mathbf{x}}, \quad \mathbf{x}\in\mathbb{R}^n. \tag{1.22}$$

Consider the *affine* transformation (i.e., a linear transformation plus a constant vector)

$$\mathbf{Z} = \boldsymbol{\mu} + B\,\mathbf{X} \tag{1.23}$$

for some $m\times n$ matrix B. Note that, by Theorem 1.8.1, $\mathbf{Z}$ has expectation vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma = BB^\top$. Any random vector of the form (1.23) is said to have a *jointly normal* or *multivariate normal* distribution. We write $\mathbf{Z} \sim \mathsf{N}(\boldsymbol{\mu},\Sigma)$. Suppose that B is an invertible $n\times n$ matrix. Then, by (1.19), the density of $\mathbf{Y} = \mathbf{Z} - \boldsymbol{\mu}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|B|\sqrt{(2\pi)^n}}\,\mathrm{e}^{-\frac{1}{2}(B^{-1}\mathbf{y})^\top B^{-1}\mathbf{y}} = \frac{1}{|B|\sqrt{(2\pi)^n}}\,\mathrm{e}^{-\frac{1}{2}\mathbf{y}^\top(B^{-1})^\top B^{-1}\mathbf{y}}.$$

We have $|B| = \sqrt{|\Sigma|}$ and $(B^{-1})^\top B^{-1} = (B^\top)^{-1}B^{-1} = (BB^\top)^{-1} = \Sigma^{-1}$, so that

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\,\mathrm{e}^{-\frac{1}{2}\mathbf{y}^\top\Sigma^{-1}\mathbf{y}}.$$

Because $\mathbf{Z}$ is obtained from $\mathbf{Y}$ by simply adding a constant vector $\boldsymbol{\mu}$, we have $f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{Y}}(\mathbf{z}-\boldsymbol{\mu})$, and therefore

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\,\mathrm{e}^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})}, \quad \mathbf{z}\in\mathbb{R}^n. \tag{1.24}$$

Note that this formula is very similar to that of the one-dimensional case.

Conversely, given a covariance matrix $\Sigma = (\sigma_{ij})$, there exists a unique lower triangular matrix

$$B = \begin{pmatrix} b_{11} & 0 & \cdots & 0 \\ b_{21} & b_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix} \tag{1.25}$$

such that $\Sigma = BB^\top$. This matrix can be obtained efficiently via the *Cholesky square root method*; see Section A.1 of the Appendix.

1.11 LIMIT THEOREMS

We briefly discuss two of the main results in probability: the law of large numbers and the central limit theorem. Both are associated with sums of independent random variables.

Let $X_1, X_2, \ldots$ be iid random variables with expectation μ and variance σ^2. For each n, let $S_n = X_1 + \cdots + X_n$. Since $X_1, X_2, \ldots$ are iid, we have $\mathbb{E}[S_n] = n\,\mathbb{E}[X_1] = n\mu$ and $\mathrm{Var}(S_n) = n\,\mathrm{Var}(X_1) = n\sigma^2$.

The law of large numbers states that S_n/n is close to μ for large n. Here is the more precise statement.

Theorem 1.11.1 (Strong Law of Large Numbers) *If $X_1, \ldots, X_n$ are iid with expectation μ, then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1 \; .$$

The central limit theorem describes the limiting distribution of S_n (or S_n/n), and it applies to both continuous and discrete random variables. Loosely, it states that the random sum S_n has a distribution that is approximately normal, when n is large. The more precise statement is given next.

Theorem 1.11.2 (Central Limit Theorem) *If $X_1, \ldots, X_n$ are iid with expectation μ and variance $\sigma^2 < \infty$, then for all $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) = \Phi(x) \; ,$$

where Φ is the cdf of the standard normal distribution.

In other words, S_n has a distribution that is approximately normal, with expectation $n\mu$ and variance $n\sigma^2$. To see the central limit theorem in action, consider Figure 1.2. The left part shows the pdfs of $S_1, \ldots, S_4$ for the case where the $\{X_i\}$ have a $\mathsf{U}[0,1]$ distribution. The right part shows the same for the $\mathsf{Exp}(1)$ distribution. We clearly see convergence to a bell-shaped curve, characteristic of the normal distribution.

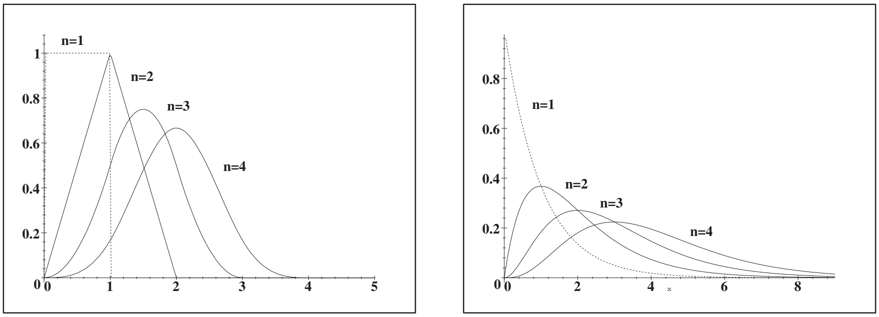


Figure 1.2: Illustration of the central limit theorem for (left) the uniform distribution and (right) the exponential distribution.

A direct consequence of the central limit theorem and the fact that a $\mathsf{Bin}(n,p)$ random variable X can be viewed as the sum of n iid $\mathsf{Ber}(p)$ random variables, $X = X_1 + \cdots + X_n$, is that for large n

$$\mathbb{P}(X \leqslant k) \approx \mathbb{P}(Y \leqslant k) \;, \tag{1.26}$$

with $Y \sim \mathsf{N}(np, np(1-p))$. As a rule of thumb, this *normal approximation to the binomial distribution* is accurate if both np and $n(1-p)$ are larger than 5.

There is also a central limit theorem for random vectors. The multidimensional version is as follows: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be iid random vectors with expectation vector $\boldsymbol{\mu}$ and covariance matrix Σ. Then for large n the random vector $\mathbf{X}_1 + \cdots + \mathbf{X}_n$ has approximately a multivariate normal distribution with expectation vector $n\boldsymbol{\mu}$ and covariance matrix $n\Sigma$.

1.12 POISSON PROCESSES

The Poisson process is used to model certain kinds of arrivals or patterns. Imagine, for example, a telescope that can detect individual photons from a faraway galaxy. The photons arrive at random times $T_1, T_2, \ldots$. Let N_t denote the number of arrivals in the time interval $[0,t]$, that is, $N_t = \sup\{k : T_k \leqslant t\}$. Note that the number of arrivals in an interval $I = (a,b]$ is given by $N_b - N_a$. We will also denote it by $N(a,b]$. A sample path of the arrival counting process $\{N_t, t \geqslant 0\}$ is given in Figure 1.3.

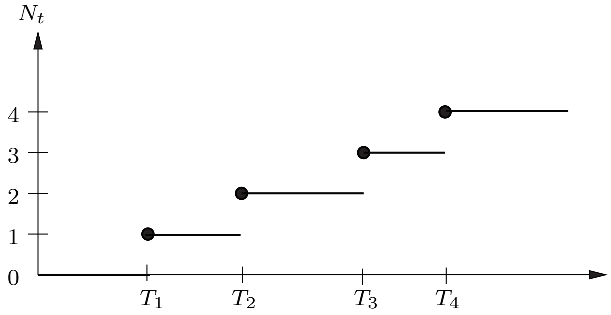


Figure 1.3: A sample path of the arrival counting process $\{N_t, t \geqslant 0\}$.

For this particular arrival process, one would assume that the number of arrivals in an interval (a,b) is independent of the number of arrivals in interval (c,d) when the two intervals do not intersect. Such considerations lead to the following definition:

Definition 1.12.1 (Poisson Process) An arrival counting process $N = \{N_t\}$ is called a *Poisson process* with *rate* $\lambda > 0$ if

(a) The numbers of points in nonoverlapping intervals are independent.

(b) The number of points in interval I has a Poisson distribution with mean $\lambda \times \text{length}(I)$.

Combining (a) and (b) we see that the number of arrivals in any small interval $(t, t+h]$ is independent of the arrival process up to time t and has a $\mathsf{Poi}(\lambda h)$ distribution. In particular, the conditional probability that exactly one arrival occurs during the time interval $(t, t+h]$ is $\mathbb{P}(N(t, t+h] = 1 \mid N_t) = \mathrm{e}^{-\lambda h}\, \lambda\, h \approx \lambda\, h$. Similarly, the probability of no arrivals is approximately $1 - \lambda h$ for small h. In other words, λ is the *rate* at which arrivals occur. Notice also that since $N_t \sim \mathsf{Poi}(\lambda t)$, the expected number of arrivals in $[0, t]$ is λt, that is, $\mathbb{E}[N_t] = \lambda t$. In Definition 1.12.1 N is seen as a random counting measure, where $N(I)$ counts the random number of arrivals in set I.

An important relationship between N_t and T_n is

$$\{N_t \geqslant n\} = \{T_n \leqslant t\} \; . \tag{1.27}$$

In other words, the number of arrivals in $[0, t]$ is at least n if and only if the n-th arrival occurs at or before time t. As a consequence, we have

$$\begin{aligned} \mathbb{P}(T_n \leqslant t) &= \mathbb{P}(N_t \geqslant n) = 1 - \sum_{k=0}^{n-1} \mathbb{P}(N_t = k) \\ &= 1 - \sum_{k=0}^{n-1} \frac{\mathrm{e}^{-\lambda\, t} (\lambda\, t)^k}{k!} \; , \end{aligned}$$

which corresponds exactly to the cdf of the $\mathsf{Gamma}(n, \lambda)$ distribution; see Problem 1.17. Thus

$$T_n \sim \mathsf{Gamma}(n, \lambda) \; . \tag{1.28}$$

Hence each T_n has the same distribution as the sum of n independent $\mathsf{Exp}(\lambda)$-distributed random variables. This corresponds with the second important characterization of a Poisson process:

> *An arrival counting process $\{N_t\}$ is a Poisson process with rate λ if and only if the interarrival times $A_1 = T_1, A_2 = T_2 - T_1, \ldots$ are independent and $\mathsf{Exp}(\lambda)$-distributed random variables.*

Poisson and Bernoulli processes are akin, and much can be learned about Poisson processes via the following *Bernoulli approximation*. Let $N = \{N_t\}$ be a Poisson process with parameter λ. We divide the time axis into small time intervals $[0, h), [h, 2h), \ldots$ and count how many arrivals occur in each interval. Note that the number of arrivals in any small time interval of length h is, with high probability, either 1 (with probability $\lambda\, h\, \mathrm{e}^{-\lambda h} \approx \lambda h$) or 0 (with probability $\mathrm{e}^{-\lambda\, h} \approx 1 - \lambda h$). Next, define $X = \{X_n\}$ to be a Bernoulli process with success parameter $p = \lambda\, h$. Put $Y_0 = 0$ and let $Y_n = X_1 + \cdots + X_n$ be the total number of successes in n trials. $Y = \{Y_n\}$ is called the *Bernoulli approximation* to N. We can view N as a limiting case of Y as we decrease h.

As an example of the usefulness of this interpretation, we now demonstrate that the Poisson property (b) in Definition 1.12.1 follows basically from the *independence* assumption (a). For small h, N_t should have approximately the same distribution as Y_n, where n is the integer part of t/h (we write $n = \lfloor t/h \rfloor$). Hence,

$$\begin{aligned}
\mathbb{P}(N_t = k) &\approx \mathbb{P}(Y_n = k) \\
&= \binom{n}{k} (\lambda\, h)^k (1 - (\lambda\, h))^{n-k} \\
&\approx \binom{n}{k} (\lambda t/n)^k (1 - (\lambda t/n))^{n-k} \\
&\approx \mathrm{e}^{-\lambda\, t}\, \frac{(\lambda\, t)^k}{k!} \; .
\end{aligned} \tag{1.29}$$

Equation (1.29) follows from the Poisson approximation to the binomial distribution; see Problem 1.22.

Another application of the Bernoulli approximation is the following. For the Bernoulli process, given that the total number of successes is k, the positions of the k successes are uniformly distributed over points $1, \ldots, n$. The corresponding property for the Poisson process N is that given $N_t = n$, the arrival times $T_1, \ldots, T_n$ are distributed according to the order statistics $X_{(1)}, \ldots, X_{(n)}$, where $X_1, \ldots, X_n$ are iid $\mathsf{U}[0, t]$.

1.13 MARKOV PROCESSES

Markov processes are stochastic processes whose futures are conditionally independent of their pasts given their present values. More formally, a stochastic process $\{X_t, t \in \mathscr{T}\}$, with $\mathscr{T} \subseteq \mathbb{R}$, is called a *Markov process* if, for every $s > 0$ and t,

$$(X_{t+s} \,|\, X_u, u \leqslant t) \quad \sim \quad (X_{t+s} \,|\, X_t) \; . \tag{1.30}$$

In other words, the conditional distribution of the future variable X_{t+s}, given the entire past of the process $\{X_u, u \leqslant t\}$, is the same as the conditional distribution of X_{t+s} given only the present X_t. That is, in order to predict future states, we only need to know the present one. Property (1.30) is called the *Markov property.*

Depending on the index set $\mathscr{T}$ and state space $\mathscr{E}$ (the set of all values the $\{X_t\}$ can take), Markov processes come in many different forms. A Markov process with a discrete index set is called a *Markov chain.* A Markov process with a discrete state space and a continuous index set (such as $\mathbb{R}$ or $\mathbb{R}_+$) is called a *Markov jump process.*

1.13.1 Markov Chains

Consider a Markov chain $X = \{X_t, t \in \mathbb{N}\}$ with a discrete (i.e., countable) state space $\mathscr{E}$. In this case the Markov property (1.30) is

$$\mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_0 = x_0, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_t = x_t) \tag{1.31}$$

for all $x_0, \ldots, x_{t+1}, \in \mathscr{E}$ and $t \in \mathbb{N}$. We restrict ourselves to Markov chains for which the conditional probabilities

$$\mathbb{P}(X_{t+1} = j \,|\, X_t = i), \;\; i, j \in \mathscr{E} \tag{1.32}$$

are independent of the time t. Such chains are called *time-homogeneous.* The probabilities in (1.32) are called the *(one-step) transition probabilities* of X. The distribution of X_0 is called the *initial distribution* of the Markov chain. The one-step transition probabilities and the initial distribution completely specify the distribution of X. Namely, we have by the product rule (1.4) and the Markov property (1.30),

$$\begin{aligned}
&\mathbb{P}(X_0 = x_0, \ldots, X_t = x_t) \\
&\quad = \mathbb{P}(X_0 = x_0)\,\mathbb{P}(X_1 = x_1 \,|\, X_0 = x_0) \cdots \mathbb{P}(X_t = x_t \,|\, X_0 = x_0, \ldots X_{t-1} = x_{t-1}) \\
&\quad = \mathbb{P}(X_0 = x_0)\,\mathbb{P}(X_1 = x_1 \,|\, X_0 = x_0) \cdots \mathbb{P}(X_t = x_t \,|\, X_{t-1} = x_{t-1})\ .
\end{aligned}$$

Since $\mathscr{E}$ is countable, we can arrange the one-step transition probabilities in an array. This array is called the (one-step) *transition matrix* of X. We usually denote it by P. For example, when $\mathscr{E} = \{0, 1, 2, \ldots\}$, the transition matrix P has the form

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} .$$

Note that the elements in every row are positive and sum up to unity.

Another convenient way to describe a Markov chain X is through its *transition graph.* States are indicated by the nodes of the graph, and a strictly positive (> 0) transition probability p_{ij} from state i to j is indicated by an arrow from i to j with weight p_{ij}.

■ EXAMPLE 1.10 Random Walk on the Integers

Let p be a number between 0 and 1. The Markov chain X with state space $\mathbb{Z}$ and transition matrix P defined by

$$P(i, i+1) = p, \quad P(i, i-1) = q = 1 - p, \qquad \text{for all } i \in \mathbb{Z}$$

is called a *random walk on the integers.* Let X start at 0; thus, $\mathbb{P}(X_0 = 0) = 1$. The corresponding transition graph is given in Figure 1.4. Starting at 0, the chain takes subsequent steps to the right with probability p and to the left with probability q.

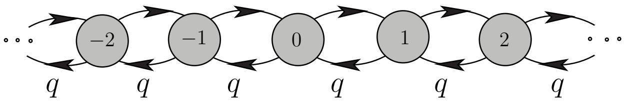


Figure 1.4: Transition graph for a random walk on $\mathbb{Z}$.

We show next how to calculate the probability that, starting from state i at some (discrete) time t, we are in j at (discrete) time $t + s$, that is, the probability $\mathbb{P}(X_{t+s} = j \mid X_t = i)$. For clarity, let us assume that $\mathscr{E} = \{1, 2, \ldots, m\}$ for some fixed m, so that P is an $m \times m$ matrix. For $t = 0, 1, 2, \ldots$, define the row vector

$$\boldsymbol{\pi}^{(t)} = (\mathbb{P}(X_t = 1), \ldots, \mathbb{P}(X_t = m)).$$

We call $\boldsymbol{\pi}^{(t)}$ the *distribution vector*, or simply the *distribution*, of X at time t and $\boldsymbol{\pi}^{(0)}$ the *initial distribution* of X. The following result shows that the t-step probabilities can be found simply by matrix multiplication.

Theorem 1.13.1 *The distribution of X at time t is given by*

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^{(0)} P^t \tag{1.33}$$

for all $t = 0, 1, \ldots$. (Here P^0 denotes the identity matrix.)

Proof: The proof is by induction. Equality (1.33) holds for $t = 0$ by definition. Suppose that this equality is true for some $t = 0, 1, \ldots$. We have

$$\mathbb{P}(X_{t+1} = k) = \sum_{i=1}^{m} \mathbb{P}(X_{t+1} = k \mid X_t = i)\, \mathbb{P}(X_t = i)\ .$$

But (1.33) is assumed to be true for t, so $\mathbb{P}(X_t = i)$ is the i-th element of $\boldsymbol{\pi}^{(0)} P^t$. Moreover, $\mathbb{P}(X_{t+1} = k \mid X_t = i)$ is the (i, k)-th element of P. Therefore, for every k,

$$\sum_{i=1}^{m} \mathbb{P}(X_{t+1} = k \mid X_t = i)\, \mathbb{P}(X_t = i) = \sum_{i=1}^{m} P(i, k)(\boldsymbol{\pi}^{(0)} P^t)(i)\ ,$$

which is just the k-th element of $\boldsymbol{\pi}^{(0)} P^{t+1}$. This completes the induction step, and thus the theorem is proved. □

By taking $\boldsymbol{\pi}^{(0)}$ as the i-th unit vector, $\mathbf{e}_i$, the t-step transition probabilities can be found as $\mathbb{P}(X_t = j \,|\, X_0 = i) = (\mathbf{e}_i\, P^t)(j) = P^t(i, j)$, which is the (i, j)-th element of matrix P^t. Thus, to find the t-step transition probabilities, we just have to compute the t-th power of P.

1.13.2 Classification of States

Let X be a Markov chain with discrete state space $\mathscr{E}$ and transition matrix P. We can characterize the relations between states in the following way: If states i and j are such that $P^t(i, j) > 0$ for some $t \geqslant 0$, we say that i *leads to* j and write $i \to j$. We say that i and j *communicate* if $i \to j$ and $j \to i$, and write $i \leftrightarrow j$. Using the relation "$\leftrightarrow$", we can divide $\mathscr{E}$ into *equivalence classes* such that all the states in an equivalence class communicate with each other but not with any state outside that class. If there is only one equivalent class ($= \mathscr{E}$), the Markov chain is said to be *irreducible*. If a set of states $\mathscr{A}$ is such that $\sum_{j \in \mathscr{A}} P(i, j) = 1$ for all $i \in \mathscr{A}$, then $\mathscr{A}$ is called a *closed* set. A state i is called an *absorbing* state if $\{i\}$ is closed. For example, in the transition graph depicted in Figure 1.5, the equivalence classes are $\{1, 2\}$, $\{3\}$, and $\{4, 5\}$. Class $\{1, 2\}$ is the only closed set: the Markov chain cannot escape from it. If state 1 were missing, state 2 would be absorbing. In Example 1.10 the Markov chain is irreducible since all states communicate.

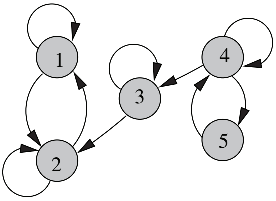


Figure 1.5: A transition graph with three equivalence classes.

Another classification of states is obtained by observing the system from a local point of view. In particular, let T denote the time the chain first visits state j, or first returns to j if it started there, and let N_j denote the total number of visits to j from time 0 on. We write $\mathbb{P}_j(A)$ for $\mathbb{P}(A \,|\, X_0 = j)$ for any event A. We denote the corresponding expectation operator by $\mathbb{E}_j$. State j is called a *recurrent* state if $\mathbb{P}_j(T < \infty) = 1$; otherwise, j is called *transient*. A recurrent state is called *positive recurrent* if $\mathbb{E}_j[T] < \infty$; otherwise, it is called *null recurrent*. Finally, a state is said to be *periodic, with period* δ, if $\delta \geqslant 2$ is the largest integer for which $\mathbb{P}_j(T = n\delta \text{ for some } n \geqslant 1) = 1$; otherwise, it is called *aperiodic*. For example, in Figure 1.5 states 1 and 2 are recurrent, and the other states are transient. All these states are aperiodic. The states of the random walk of Example 1.10 are periodic with period 2.

It can be shown that recurrence and transience are class properties. In particular, if $i \leftrightarrow j$, then i recurrent (transient) $\Leftrightarrow$ j recurrent (transient). Thus, in an irreducible Markov chain, one state being recurrent implies that all other states are also recurrent. And if one state is transient, then so are all the others.

1.13.3 Limiting Behavior

The limiting or "steady-state" behavior of Markov chains as $t \to \infty$ is of considerable interest and importance, and this type of behavior is often simpler to describe and analyze than the "transient" behavior of the chain for fixed t. It can be shown (see, for example, [3]) that in an irreducible, aperiodic Markov chain with transition matrix P the t-step probabilities converge to a constant that does not depend on the initial state. More specifically,

$$\lim_{t \to \infty} P^t(i,j) = \pi_j \tag{1.34}$$

for some number $0 \leqslant \pi_j \leqslant 1$. Moreover, $\pi_j > 0$ if j is positive recurrent and $\pi_j = 0$ otherwise. The intuitive reason behind this result is that the process "forgets" where it was initially if it goes on long enough. This is true for both finite and countably infinite Markov chains. The numbers $\{\pi_j, j \in \mathscr{E}\}$ form the *limiting distribution* of the Markov chain, provided that $\pi_j \geqslant 0$ and $\sum_j \pi_j = 1$. Note that these conditions are not always satisfied: they are clearly not satisfied if the Markov chain is transient, and they may not be satisfied if the Markov chain is recurrent (i.e., when the states are null-recurrent). The following theorem gives a method for obtaining limiting distributions. Here we assume for simplicity that $\mathscr{E} = \{0, 1, 2, \ldots\}$. The limiting distribution is identified with the row vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$.

Theorem 1.13.2 *For an irreducible, aperiodic Markov chain with transition matrix P, if the limiting distribution $\boldsymbol{\pi}$ exists, then it is uniquely determined by the solution of*

$$\boldsymbol{\pi} = \boldsymbol{\pi} P\,, \tag{1.35}$$

with $\pi_j \geqslant 0$ and $\sum_j \pi_j = 1$. Conversely, if there exists a positive row vector $\boldsymbol{\pi}$ satisfying (1.35) *and summing up to 1, then $\boldsymbol{\pi}$ is the limiting distribution of the Markov chain. Moreover, in that case, $\pi_j > 0$ for all j and all states are positive recurrent.*

Proof: (Sketch) For the case where $\mathscr{E}$ is finite, the result is simply a consequence of (1.33). Namely, with $\boldsymbol{\pi}^{(0)}$ being the i-th unit vector, we have

$$P^{t+1}(i,j) = \left(\boldsymbol{\pi}^{(0)}\, P^t\, P\right)(j) = \sum_{k \in \mathscr{E}} P^t(i,k) P(k,j)\,.$$

Letting $t \to \infty$, we obtain (1.35) from (1.34), provided that we can change the order of the limit and the summation. To show uniqueness, suppose that another vector $\mathbf{y}$, with $y_j \geqslant 0$ and $\sum_j y_j = 1$, satisfies $\mathbf{y} = \mathbf{y}P$. Then it is easy to show by induction that $\mathbf{y} = \mathbf{y}P^t$, for every t. Hence, letting $t \to \infty$, we obtain for every j

$$y_j = \sum_i y_i\, \pi_j = \pi_j\,,$$

since the $\{y_j\}$ sum up to unity. We omit the proof of the converse statement. □

■ EXAMPLE 1.11 Random Walk on the Positive Integers

This is a slightly different random walk than the one in Example 1.10. Let X be a random walk on $\mathscr{E} = \{0, 1, 2, \ldots\}$ with transition matrix

$$P = \begin{pmatrix} q & p & 0 & \ldots & & \\ q & 0 & p & 0 & \ldots & \\ 0 & q & 0 & p & 0 & \ldots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where $0 < p < 1$ and $q = 1 - p$. X_t could represent, for example, the number of customers who are waiting in a queue at time t.

All states can be reached from each other, so the chain is irreducible and every state is either recurrent or transient. The equation $\boldsymbol{\pi} = \boldsymbol{\pi} P$ becomes

$$\begin{aligned} \pi_0 &= q\,\pi_0 + q\,\pi_1\,, \\ \pi_1 &= p\,\pi_0 + q\,\pi_2\,, \\ \pi_2 &= p\,\pi_1 + q\,\pi_3\,, \\ \pi_3 &= p\,\pi_2 + q\,\pi_4\,, \end{aligned}$$

and so on. We can solve this set of equation sequentially. If we let $r = p/q$, then we can express the $\pi_1, \pi_2, \ldots$ in terms of π_0 and r as

$$\pi_j = r^j\, \pi_0, \;\; j = 0, 1, 2, \ldots\,.$$

> If $p < q$, then $r < 1$ and $\sum_{j=0}^{\infty} \pi_j = \pi_0/(1-r)$, and by choosing $\pi_0 = 1 - r$, we can make the sum $\sum \pi_j = 1$. Hence, for $r < 1$, we have found the limiting distribution $\boldsymbol{\pi} = (1-r)(1, r, r^2, r^3, \ldots)$ for this Markov chain, and all the states are therefore positive recurrent. However, when $p \geqslant q$, $\sum \pi_j$ is either 0 or infinite, and hence all states are either null-recurrent or transient. (It can be shown that only the case $p = q$ leads to null-recurrent states.)

Let X be a Markov chain with limiting distribution $\boldsymbol{\pi}$. Suppose $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$. Then, combining (1.33) and (1.35), we have $\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}$. Thus, if the initial distribution of the Markov chain is equal to the limiting distribution, then the distribution of X_t is the same for all t (and is given by this limiting distribution). In fact, it is not difficult to show that for any k the distribution of $X_k, X_{k+1}, X_{k+2} \ldots$ is the same as that of $X_0, X_1, \ldots$. In other words, when $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$, the Markov chain is a stationary stochastic process. More formally, a stochastic process $\{X_t, t \in \mathbb{N}\}$ is called *stationary* if, for any positive $\tau, t_1, \ldots, t_n$, the vector $(X_{t_1}, \ldots, X_{t_n})$ has the same distribution as $(X_{t_1+\tau}, \ldots, X_{t_n+\tau})$. Similar definitions hold when the index set is $\mathbb{Z}$, $\mathbb{R}_+$, or $\mathbb{R}$. For this reason any distribution $\boldsymbol{\pi}$ for which (1.35) holds is called a *stationary distribution.*

Noting that $\sum_j p_{ij} = 1$, we can rewrite (1.35) as the system of equations

$$\sum_j \pi_i \, p_{ij} = \sum_j \pi_j \, p_{ji} \quad \text{for all } i \in \mathscr{E} \; . \tag{1.36}$$

These are called the *global balance equations.* We can interpret (1.35) as the statement that the "probability flux" out of i is balanced by the probability flux into i. An important generalization, which follows directly from (1.36), states that the same balancing of probability fluxes holds for an arbitrary set $\mathscr{A}$. That is, for every set $\mathscr{A}$ of states we have

$$\sum_{i \in \mathscr{A}} \sum_{j \notin \mathscr{A}} \pi_i \, p_{ij} = \sum_{i \in \mathscr{A}} \sum_{j \notin \mathscr{A}} \pi_j \, p_{ji} \; . \tag{1.37}$$

1.13.4 Reversibility

Reversibility is an important notion in the theory of Markov and more general processes. A stationary stochastic process $\{X_t\}$ with index set $\mathbb{Z}$ or $\mathbb{R}$ is said to be *reversible* if, for any positive integer n and for all $t_1, \ldots, t_n$, the vector $(X_{t_1}, \ldots, X_{t_n})$ has the same distribution as $(X_{-t_1}, \ldots, X_{-t_n})$. One way to visualize this is to imagine that we have taken a video of the stochastic process, which we may run in forward and reverse time. If we cannot determine whether the video is running forward or backward, the process is reversible. The main result for reversible Markov chains is that a stationary Markov process is reversible if and only if there exists a collection of positive numbers $\{\pi_i, i \in \mathscr{E}\}$ summing to unity that satisfy the *detailed (or local) balance equations*

$$\pi_i \, p_{ij} = \pi_j \, p_{ji} \; , \quad i, j \in \mathscr{E}. \tag{1.38}$$

Whenever such a collection $\{\pi_j\}$ exists, it is the stationary distribution of the process.

A good way to think of the detailed balance equations is that they balance the probability flux from state i to state j with that from state j to state i. Contrast this with the equilibrium equations (1.36), which balance the probability flux out of state i with that into state i.

Kolmogorov's criterion is a simple criterion for reversibility based on the transition probabilities. It states that a stationary Markov process is reversible if and only if its transition rates satisfy

$$p(i_1, i_2)\, p(i_2, i_3) \dots p(i_{n-1}, i_n)\, p(i_n, i_1) = p(i_1, i_n)\, p(i_n, i_{n-1}) \dots p(i_2, i_1) \tag{1.39}$$

for all finite loops of states $i_1, \dots, i_n, i_1$. (For clarity, we have used the notation $p(i, j)$ rather than p_{ij} for the transition probabilities.) The idea is quite intuitive: if the process in forward time is more likely to traverse a certain closed loop in one direction than in the opposite direction, then in backward time it will exhibit the opposite behavior, and hence we have a criterion for detecting the direction of time. If such "looping" behavior does not occur, the process must be reversible.

1.13.5 Markov Jump Processes

A *Markov jump process* $X = \{X_t, t \geqslant 0\}$ can be viewed as a continuous-time generalization of a Markov chain and also of a Poisson process. The Markov property (1.30) now reads

$$\mathbb{P}(X_{t+s} = x_{t+s} \mid X_u = x_u, u \leqslant t) = \mathbb{P}(X_{t+s} = x_{t+s} \mid X_t = x_t)\ . \tag{1.40}$$

As in the Markov chain case, one usually assumes that the process is *time-homogeneous*, that is, $\mathbb{P}(X_{t+s} = j \,|\, X_t = i)$ does not depend on t. Denote this probability by $P_s(i, j)$. An important quantity is the *transition rate* q_{ij} from state i to j, defined for $i \neq j$ as

$$q_{ij} = \lim_{t \downarrow 0} \frac{P_t(i, j)}{t}\ .$$

The sum of the rates out of state i is denoted by q_i. A typical sample path of X is shown in Figure 1.6. The process jumps at times $T_1, T_2, \dots$ to states $Y_1, Y_2, \dots$, staying some length of time in each state.

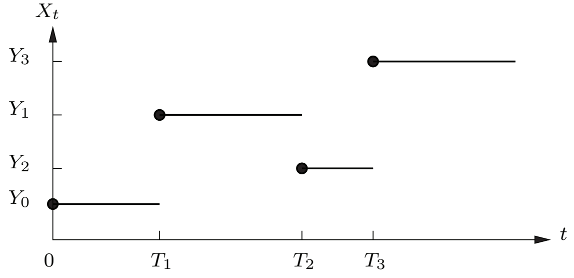


Figure 1.6: A sample path of a Markov jump process $\{X_t, t \geqslant 0\}$.

More precisely, a Markov jump process X behaves (under suitable regularity conditions; see [3]) as follows:

1. Given its past, the probability that X jumps from its current state i to state j is $K_{ij} = q_{ij}/q_i$.

2. The amount of time that X spends in state j has an exponential distribution with mean $1/q_j$, independent of its past history.

The first statement implies that the process $\{Y_n\}$ is in fact a Markov chain, with transition matrix $K = (K_{ij})$.

A convenient way to describe a Markov jump process is through its *transition rate graph*. This is similar to a transition graph for Markov chains. The states are represented by the nodes of the graph, and a transition rate from state i to j is indicated by an arrow from i to j with weight q_{ij}.

■ EXAMPLE 1.12 Birth-and-Death Process

A *birth-and-death process* is a Markov jump process with a transition rate graph of the form given in Figure 1.7. Imagine that X_t represents the total number of individuals in a population at time t. Jumps to the right correspond to births, and jumps to the left to deaths. The *birth rates* $\{b_i\}$ and the *death rates* $\{d_i\}$ may differ from state to state. Many applications of Markov chains involve processes of this kind. Note that the process jumps from one state to the next according to a Markov chain with transition probabilities $K_{0,1} = 1$, $K_{i,i+1} = b_i/(b_i + d_i)$, and $K_{i,i-1} = d_i/(b_i + d_i)$, $i = 1, 2, \ldots$. Moreover, it spends an $\mathsf{Exp}(b_0)$ amount of time in state 0 and $\mathsf{Exp}(b_i + d_i)$ in the other states.

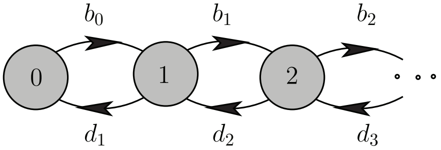


Figure 1.7: The transition rate graph of a birth-and-death process.

Limiting Behavior We now formulate the continuous-time analogues of (1.34) and Theorem 1.13.2. Irreducibility and recurrence for Markov jump processes are defined in the same way as for Markov chains. For simplicity, we assume that $\mathscr{E} = \{1, 2, \ldots\}$. If X is a recurrent and irreducible Markov jump process, then regardless of i,

$$\lim_{t\to\infty} \mathbb{P}(X_t = j \mid X_0 = i) = \pi_j \tag{1.41}$$

for some number $\pi_j \geqslant 0$. Moreover, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ is the solution to

$$\sum_{j\neq i} \pi_i\, q_{ij} = \sum_{j\neq i} \pi_j\, q_{ji}, \quad \text{for all } i = 1, \ldots, m \tag{1.42}$$

with $\sum_j \pi_j = 1$, if such a solution exists, in which case all states are positive recurrent. If such a solution does not exist, all π_j are 0.

As in the Markov chain case, $\{\pi_j\}$ is called the *limiting distribution* of X and is usually identified with the row vector $\boldsymbol{\pi}$. Any solution $\boldsymbol{\pi}$ of (1.42) with $\sum_j \pi_j = 1$ is called a *stationary distribution*, since taking it as the initial distribution of the Markov jump process renders the process stationary.

Equations (1.42) are again called the *global balance equations* and are readily generalized to (1.37), replacing the transition probabilities with transition rates. More important, if the process is reversible, then, as with Markov chains, the stationary distribution can be found from the *local balance equations*:

$$\pi_i \, q_{ij} = \pi_j \, q_{ji} \; , \;\; i, j \in \mathscr{E} \; . \tag{1.43}$$

Reversibility can be easily verified by checking that looping does not occur, that is, via Kolmogorov's criterion (1.39), replacing the probabilities p with rates q.

■ **EXAMPLE 1.13** ***M/M/1* Queue**

Consider a service facility where customers arrive at certain random times and are served by a single server. Arriving customers who find the server busy wait in the queue. Customers are served in the order in which they arrive. The interarrival times are exponential random variables with rates λ, and the service times of customers are iid exponential random variables with rates μ. Last, the service times are independent of the interarrival times. Let X_t be the number of customers in the system at time t. By the memoryless property of the exponential distribution (see Problem 1.7), it is not difficult to see that $X = \{X_t, t \geqslant 0\}$ is a Markov jump process, and in fact a birth-and-death process with birth rates $b_i = \lambda$, $i = 0, 1, 2, \ldots$ and death rates $d_i = \mu$, $i = 1, 2, \ldots$.

Solving the global balance equations (or, more easily, the local balance equations, since X is reversible), we see that X has a limiting distribution given by

$$\lim_{t \to \infty} \mathbb{P}(X_t = n) = (1 - \varrho) \, \varrho^n, \quad n = 0, 1, 2, \ldots, \tag{1.44}$$

provided that $\varrho = \lambda/\mu < 1$. This means that the expected service time needs to be less than the expected interarrival time for a limiting distribution to exist. In that case, the limiting distribution is also the stationary distribution. In particular, if X_0 is distributed according to (1.44), then X_t has the same distribution for all $t > 0$.

1.14 GAUSSIAN PROCESSES

The normal distribution is also called the *Gaussian* distribution. Gaussian processes are generalizations of multivariate normal random vectors (discussed in Section 1.10). Specifically, a stochastic process $\{X_t, t \in \mathscr{T}\}$ is said to be *Gaussian* if all its finite-dimensional distributions are Gaussian. That is, if for any choice of n and $t_1, \ldots, t_n \in \mathscr{T}$, it holds that

$$(X_{t_1}, \ldots, X_{t_n})^\top \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma) \tag{1.45}$$

for some expectation vector $\boldsymbol{\mu}$ and covariance matrix Σ (both of which depend on the choice of $t_1, \ldots, t_n$). Equivalently, $\{X_t, t \in \mathscr{T}\}$ is Gaussian if any linear combination $\sum_{i=1}^n b_i X_{t_i}$ has a normal distribution. Note that a Gaussian process is determined completely by its *expectation function* $\mu_t = \mathbb{E}[X_t], \;\; t \in \mathscr{T}$, and *covariance function* $\Sigma_{s,t} = \mathrm{Cov}(X_s, X_t), \;\; s, t \in \mathscr{T}$.

■ EXAMPLE 1.14 Wiener Process (Brownian Motion)

The quintessential Gaussian process is the *Wiener process* or (standard) *Brownian motion*. It can be viewed as a continuous version of a random walk process. Figure 1.8 gives a typical sample path. The Wiener process plays a central role in probability and forms the basis of many other stochastic processes.

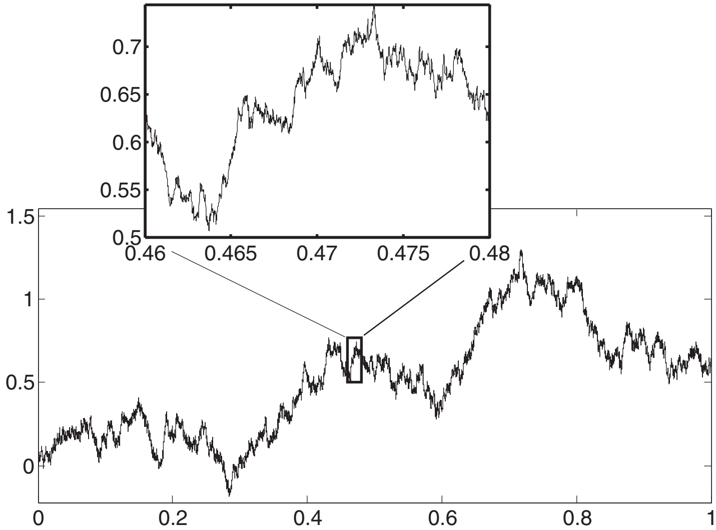


Figure 1.8: A sample path of the Wiener process. The inset shows a magnification of the path over a small time interval.

The Wiener process can be defined as a Gaussian process $\{X_t, t \geqslant 0\}$ with expectation function $\mu_t = 0$ for all t and covariance function $\Sigma_{s,t} = s$ for $0 \leqslant s \leqslant t$. The Wiener process has many fascinating properties (e.g., [11]). For example, it is a Markov process (i.e., it satisfies the Markov property (1.30)) with continuous sample paths that are *nowhere differentiable*. Moreover, the increments $X_t - X_s$ over intervals $[s, t]$ are independent and normally distributed. Specifically, for any $t_1 < t_2 \leqslant t_3 < t_4$,

$$X_{t_4} - X_{t_3} \quad \text{and} \quad X_{t_2} - X_{t_1}$$

are independent random variables, and for all $t \geqslant s \geqslant 0$,

$$X_t - X_s \sim \mathsf{N}(0, t - s) \ .$$

This leads to a simple simulation procedure for Wiener processes, which is discussed in Section 2.8.

1.15 INFORMATION

In this section we discuss briefly various measures of information in a random experiment. Suppose that we describe the measurements in a random experiment via a random vector $\mathbf{X} = (X_1, \dots, X_n)$ with pdf f. Then all the information about the experiment (all of our probabilistic knowledge) is obviously contained in the pdf f. However, in most cases we would want to characterize our information about the experiments with just a few key numbers, such as the *expectation* and the *covariance matrix* of $\mathbf{X}$, which provide information about the mean measurements and the variability of the measurements, respectively. Another informational measure comes from coding and communications theory, where the *Shannon entropy* characterizes the average number of bits needed to transmit a message $\mathbf{X}$ over a (binary) communication channel. Yet another approach to information can be found in statistics. Specifically, in the theory of point estimation, the pdf f depends on a parameter vector $\boldsymbol{\theta}$. The question is how well $\boldsymbol{\theta}$ can be estimated via an outcome of $\mathbf{X}$ — in other words, how much information about $\boldsymbol{\theta}$ is contained in the "data" $\mathbf{X}$. Various measures for this type of information are associated with the *maximum likelihood*, the *score*, and the *(Fisher) information matrix*. Finally, the amount of information in a random experiment can often be quantified via a *distance* concept, such as the *Kullback–Leibler* "distance" (divergence), also called the *cross-entropy*.

1.15.1 Shannon Entropy

One of the most celebrated measures of uncertainty in information theory is the *Shannon entropy*, or simply *entropy*. A good reference is [4], where the entropy of a discrete random variable X with density f is defined as

$$\mathbb{E}\left[\log_2 \frac{1}{f(X)}\right] = -\mathbb{E}\left[\log_2 f(X)\right] = -\sum_{\mathscr{X}} f(x) \log_2 f(x) \ .$$

Here X is interpreted as a random character from an alphabet $\mathscr{X}$, such that $X = x$ with probability $f(x)$. We will use the convention $0 \ln 0 = 0$.

It can be shown that the most efficient way to transmit characters sampled from f over a binary channel is to encode them such that the number of bits required to transmit x is equal to $\log_2(1/f(x))$. It follows that $-\sum_{\mathscr{X}} f(x) \log_2 f(x)$ is the expected bit length required to send a random character $X \sim f$; see [4].

A more general approach, which includes continuous random variables, is to define the entropy of a random variable X with density f by

$$\mathcal{H}(X) = -\mathbb{E}[\ln f(X)] = \begin{cases} -\sum f(x) \ln f(x) & \text{discrete case,} \\ -\int f(x) \ln f(x) \, \mathrm{d}x & \text{continuous case.} \end{cases} \tag{1.46}$$

Definition (1.46) can easily be extended to random vectors $\mathbf{X}$ as (in the continuous case)

$$\mathcal{H}(\mathbf{X}) = -\mathbb{E}[\ln f(\mathbf{X})] = -\int f(\mathbf{x}) \ln f(\mathbf{x}) \, \mathrm{d}\mathbf{x} \ . \tag{1.47}$$

$\mathcal{H}(\mathbf{X})$ is often called the *joint* entropy of the random variables $X_1, \dots, X_n$, and it is also written as $\mathcal{H}(X_1, \dots, X_n)$. In the continuous case, $\mathcal{H}(\mathbf{X})$ is frequently referred to as the *differential entropy* to distinguish it from the discrete case.

■ EXAMPLE 1.15

Let X have a $\mathsf{Ber}(p)$ distribution for some $0 \leqslant p \leqslant 1$. The density f of X is given by $f(1) = \mathbb{P}(X = 1) = p$ and $f(0) = \mathbb{P}(X = 0) = 1 - p$ so that the entropy of X is

$$\mathcal{H}(X) = -p \ln p - (1-p) \ln(1-p) \; .$$

The graph of the entropy as a function of p is depicted in Figure 1.9. Note that the entropy is maximal for $p = 1/2$, which gives the "uniform" density on $\{0, 1\}$.

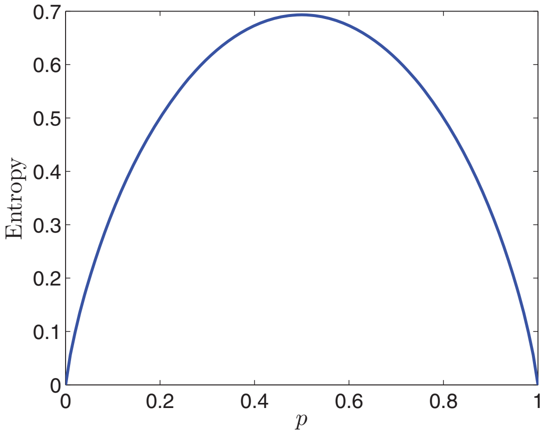


Figure 1.9: The entropy for the $\mathsf{Ber}(p)$ distribution as a function of p.

Next, consider a sequence $X_1, \ldots, X_n$ of iid $\mathsf{Ber}(p)$ random variables. Let $\mathbf{X} = (X_1, \ldots, X_n)$. The density of $\mathbf{X}$, say g, is simply the product of the densities of the X_i, so that

$$\mathcal{H}(\mathbf{X}) = -\mathbb{E}\left[\ln g(\mathbf{X})\right] = -\mathbb{E}\left[\ln \prod_{i=1}^{n} f(X_i)\right] = \sum_{i=1}^{n} -\mathbb{E}\left[\ln f(X_i)\right] = n\,\mathcal{H}(X) \; .$$

The properties of $\mathcal{H}(\mathbf{X})$ in the continuous case are somewhat different from those in the discrete one. In particular:

1. The differential entropy can be negative, whereas the discrete entropy is always positive.
2. The discrete entropy is insensitive to invertible transformations, whereas the differential entropy is not. Specifically, if $\mathbf{X}$ is discrete, $\mathbf{Y} = g(\mathbf{X})$, and g is an invertible mapping, then $\mathcal{H}(\mathbf{X}) = \mathcal{H}(\mathbf{Y})$ because $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))$. However, in the continuous case, we have an additional factor due to the Jacobian of the transformation.

It is not difficult to see that of any density f, the one that gives the maximum entropy is the uniform density on $\mathscr{X}$. That is,

$$\mathcal{H}(\mathbf{X}) \text{ is maximal } \Leftrightarrow \; f(\mathbf{x}) = \frac{1}{|\mathscr{X}|} \text{ (constant)} \; . \tag{1.48}$$

For two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with joint pdf f, we define the *conditional entropy* of $\mathbf{Y}$ given $\mathbf{X}$ as

$$\mathcal{H}(\mathbf{Y}\,|\,\mathbf{X}) = -\mathbb{E}\left[\ln \frac{f(\mathbf{X},\mathbf{Y})}{f_{\mathbf{X}}(\mathbf{X})}\right] = \mathcal{H}(\mathbf{X},\mathbf{Y}) - \mathcal{H}(\mathbf{X})\;, \tag{1.49}$$

where $f_{\mathbf{X}}$ is the pdf of $\mathbf{X}$ and $\frac{f(\mathbf{x},\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}$ is the conditional density of $\mathbf{Y}$ (at $\mathbf{y}$), given $\mathbf{X} = \mathbf{x}$. It follows that

$$\mathcal{H}(\mathbf{X},\mathbf{Y}) = \mathcal{H}(\mathbf{X}) + \mathcal{H}(\mathbf{Y}\,|\,\mathbf{X}) = \mathcal{H}(\mathbf{Y}) + \mathcal{H}(\mathbf{X}\,|\,\mathbf{Y})\;. \tag{1.50}$$

It is reasonable to require that any sensible additive measure describing the average amount of uncertainty should satisfy at least (1.50) and (1.48). It follows that the uniform density carries the least amount of information, and the entropy (average amount of uncertainty) of $(\mathbf{X},\mathbf{Y})$ is equal to the sum of the entropy of $\mathbf{X}$ and the amount of entropy in $\mathbf{Y}$ after the information in $\mathbf{X}$ has been accounted for. It is argued in [10] that any concept of entropy that includes the general properties (1.48) and (1.50) must lead to the definition (1.47).

The *mutual information* of $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathcal{M}(\mathbf{X},\mathbf{Y}) = \mathcal{H}(\mathbf{X}) + \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{X},\mathbf{Y})\;, \tag{1.51}$$

which, as the name suggests, can be interpreted as the amount of information shared by $\mathbf{X}$ and $\mathbf{Y}$. An alternative expression, which follows from (1.50) and (1.51), is

$$\mathcal{M}(\mathbf{X},\mathbf{Y}) = \mathcal{H}(\mathbf{X}) - \mathcal{H}(\mathbf{X}\,|\,\mathbf{Y}) = \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{Y}\,|\,\mathbf{X})\;, \tag{1.52}$$

which can be interpreted as the reduction of the uncertainty of one random variable due to the knowledge of the other. It is not difficult to show that the mutual information is always positive. It is also related to the cross-entropy concept, which follows.

1.15.2 Kullback–Leibler Cross-Entropy

Let g and h be two densities on $\mathscr{X}$. The Kullback–Leibler cross-entropy between g and h (compare with (1.47)) is defined (in the continuous case) as

$$\begin{aligned}\mathcal{D}(g,h) &= \mathbb{E}_g\left[\ln \frac{g(\mathbf{X})}{h(\mathbf{X})}\right] \\ &= \int g(\mathbf{x}) \ln g(\mathbf{x})\, \mathrm{d}\mathbf{x} - \int g(\mathbf{x}) \ln h(\mathbf{x})\, \mathrm{d}\mathbf{x}\;.\end{aligned} \tag{1.53}$$

$\mathcal{D}(g,h)$ is also called the *Kullback–Leibler divergence*, the *cross-entropy*, and the *relative entropy.* If not stated otherwise, we will call $\mathcal{D}(g,h)$ the *cross-entropy* (CE) between g and h. Notice that $\mathcal{D}(g,h)$ is not a distance between g and h in the formal sense, since in general $\mathcal{D}(g,h) \neq \mathcal{D}(h,g)$. Nonetheless, it is often useful to think of $\mathcal{D}(g,h)$ as a distance because

$$\mathcal{D}(g,h) \geqslant 0$$

and $\mathcal{D}(g,h) = 0$ if and only if $g(x) = h(x)$. This follows from Jensen's inequality (if ϕ is a convex function, such as $-\ln$, then $\mathbb{E}[\phi(X)] \geqslant \phi(\mathbb{E}[X])$). Namely

$$\mathcal{D}(g,h) = \mathbb{E}_g\left[-\ln\frac{h(\mathbf{X})}{g(\mathbf{X})}\right] \geqslant -\ln\left\{\mathbb{E}_g\left[\frac{h(\mathbf{X})}{g(\mathbf{X})}\right]\right\} = -\ln 1 = 0\,.$$

It can be readily seen that the mutual information $\mathcal{M}(\mathbf{X},\mathbf{Y})$ of vectors $\mathbf{X}$ and $\mathbf{Y}$ defined in (1.51) is related to the CE in the following way:

$$\mathcal{M}(\mathbf{X},\mathbf{Y}) = \mathcal{D}(f, f_{\mathbf{X}} f_{\mathbf{Y}}) = \mathbb{E}_f\left[\ln\frac{f(\mathbf{X},\mathbf{Y})}{f_{\mathbf{X}}(\mathbf{X})\, f_{\mathbf{Y}}(\mathbf{Y})}\right]\,,$$

where f is the (joint) pdf of $(\mathbf{X},\mathbf{Y})$ and $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$ are the (marginal) pdfs of $\mathbf{X}$ and $\mathbf{Y}$, respectively. In other words, the mutual information can be viewed as the CE that measures the distance between the joint pdf f of $\mathbf{X}$ and $\mathbf{Y}$ and the product of their marginal pdfs $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$, that is, under the assumption that the vectors $\mathbf{X}$ and $\mathbf{Y}$ are *independent*.

1.15.3 Maximum Likelihood Estimator and Score Function

We introduce here the notion of the *score function* (SF) via the classical *maximum likelihood estimator*. Consider a random vector $\mathbf{X} = (X_1,\ldots,X_n)$ that is is distributed according to a fixed pdf $f(\cdot;\boldsymbol{\theta})$ with unknown parameter (vector) $\boldsymbol{\theta} \in \Theta$. Say that we want to estimate $\boldsymbol{\theta}$ on the basis of a given outcome $\mathbf{x}$ (the data) of $\mathbf{X}$. For a given $\mathbf{x}$, the function $\mathcal{L}(\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta})$ is called the *likelihood function*. Note that $\mathcal{L}$ is a function of $\boldsymbol{\theta}$ for a fixed parameter $\mathbf{x}$, whereas for the pdf f it is the other way around. The maximum likelihood *estimate* $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$ of $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta};\mathbf{x})\,. \tag{1.54}$$

Because the function ln is monotone increasing, we also have

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \ln\mathcal{L}(\boldsymbol{\theta};\mathbf{x})\,. \tag{1.55}$$

The random variable $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ with $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$ is the corresponding maximum likelihood *estimator*, which is again written as $\widehat{\boldsymbol{\theta}}$. Note that often the data $X_1,\ldots,X_n$ form a random sample from some pdf $f_1(\cdot;\boldsymbol{\theta})$, in which case $f(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^N f_1(x_i;\boldsymbol{\theta})$ and

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^N \ln f_1(X_i;\boldsymbol{\theta})\,. \tag{1.56}$$

If $\mathcal{L}(\boldsymbol{\theta};\mathbf{x})$ is a continuously differentiable concave function with respect to $\boldsymbol{\theta}$ and the maximum is attained in the interior of Θ, then we can find the maximum likelihood estimator of $\boldsymbol{\theta}$ by solving

$$\nabla_{\boldsymbol{\theta}} \ln\mathcal{L}(\boldsymbol{\theta};\mathbf{x}) = \mathbf{0}\,.$$

The function $\mathcal{S}(\cdot;\mathbf{x})$ defined by

$$\mathcal{S}(\boldsymbol{\theta};\mathbf{x}) = \nabla_{\boldsymbol{\theta}} \ln\mathcal{L}(\boldsymbol{\theta};\mathbf{x}) = \frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta})} \tag{1.57}$$

is called the *score function.* For the exponential family (A.9) it is easy to see that

$$\mathcal{S}(\boldsymbol{\theta};\mathbf{x}) = \frac{\nabla c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})} + \mathbf{t}(\mathbf{x}) \,. \tag{1.58}$$

The *random vector* $\mathcal{S}(\boldsymbol{\theta}) = \mathcal{S}(\boldsymbol{\theta};\mathbf{X})$ with $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$ is called the *(efficient) score.* The expected score is always equal to the zero vector, that is,

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{S}(\boldsymbol{\theta})] = \int \nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})\,\mu(\mathrm{d}\mathbf{x}) = \nabla_{\boldsymbol{\theta}} \int f(\mathbf{x};\theta)\,\mu(\mathrm{d}\mathbf{x}) = \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0}\,,$$

where the interchange of differentiation and integration is justified via the bounded convergence theorem.

1.15.4 Fisher Information

The covariance matrix $\mathcal{I}(\boldsymbol{\theta})$ of the score $\mathcal{S}(\boldsymbol{\theta})$ is called the *Fisher information matrix.* Since the expected score is always $\mathbf{0}$, we have

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{S}(\boldsymbol{\theta})\mathcal{S}(\boldsymbol{\theta})^{\top}\right] \,. \tag{1.59}$$

In the one-dimensional case, we thus have

$$\mathcal{I}(\theta) = \mathbb{E}_{\theta}\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right] \,.$$

Because

$$\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta) = \frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)}{f(x;\theta)} - \left(\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}\right)^2 ,$$

we see that (under straightforward regularity conditions) the Fisher information is also given by

$$\mathcal{I}(\theta) = -\mathbb{E}_{\theta}\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right] \,.$$

In the multidimensional case we have similarly

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\nabla \mathcal{S}(\boldsymbol{\theta})\right] = -\mathbb{E}_{\boldsymbol{\theta}}\left[\nabla^2 \ln f(\mathbf{X};\boldsymbol{\theta})\right] \,, \tag{1.60}$$

where $\nabla^2 \ln f(\mathbf{X};\boldsymbol{\theta})$ denotes the *Hessian* of $\ln f(\mathbf{X};\boldsymbol{\theta})$, that is, the (random) matrix

$$\left(\frac{\partial^2 \ln f(\mathbf{X};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right) \,.$$

The importance of the Fisher information in statistics is corroborated by the famous *Cramér–Rao inequality*, which (in a simplified form) states that the variance of any unbiased estimator Z of $g(\boldsymbol{\theta})$ is bounded from below via

$$\mathrm{Var}(Z) \geqslant (\nabla g(\boldsymbol{\theta}))^{\top}\, \mathcal{I}^{-1}(\boldsymbol{\theta})\, \nabla g(\boldsymbol{\theta}) \,. \tag{1.61}$$

For more details, see [12].

1.16 CONVEX OPTIMIZATION AND DUALITY

Let $f(x)$, $x \in \mathbb{R}$, be a real-valued function with continuous derivatives — also called a C^1 function. The standard approach to minimizing $f(x)$ is to solve the equation

$$f'(x) = 0\,. \tag{1.62}$$

The solutions to (1.62) are called *stationary points*. If, in addition, the function has continuous second derivatives (a so-called C^2 function), the condition

$$f''(x^*) > 0 \tag{1.63}$$

ensures that a stationary point x^* is a *local minimizer*, that is, $f(x^*) < f(x)$ for all x in a small enough neighborhood of x^*.

For a C^1 function on $\mathbb{R}^n$, (1.62) generalizes to

$$\nabla f(\mathbf{x}) \equiv \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix} = \mathbf{0}\,, \tag{1.64}$$

where $\nabla f(\mathbf{x})$ is the *gradient* of f at $\mathbf{x}$. Similarly, a stationary point $\mathbf{x}^*$ is a local minimizer of f if the *Hessian matrix* (or simply *Hessian*) at $\mathbf{x}^*$,

$$\nabla^2 f(\mathbf{x}^*) \equiv \begin{pmatrix} \frac{\partial^2 f(\mathbf{x}^*)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\mathbf{x}^*)}{\partial x_1 \partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 f(\mathbf{x}^*)}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f(\mathbf{x}^*)}{\partial x_n^2} \end{pmatrix}, \tag{1.65}$$

is *positive definite*, that is, $\mathbf{x}^\top \left[\nabla^2 f(\mathbf{x}^*)\right] \mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

The situation can be further generalized by introducing *constraints*. A general constrained optimization problems can be written as

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & f(\mathbf{x}) & (1.66) \\ \text{subject to:} \quad & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, m\,, & (1.67) \\ & g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \ldots, k\,. & (1.68) \end{aligned}$$

Here f, g_i, and h_i are given functions, $f(\mathbf{x})$ is called the *objective function*, and $h_i(\mathbf{x}) = 0$ and $g_i(\mathbf{x}) \leqslant 0$ represent the *equality* and *inequality* constraints, respectively.

The region of the domain where the objective function is defined and where all the constraints are satisfied is called the *feasible region*. A *global solution* to the optimization problem is a point $\mathbf{x}^* \in \mathbb{R}^n$ such that there exists no other point $\mathbf{x} \in \mathbb{R}^n$ for which $f(\mathbf{x}) < f(\mathbf{x}^*)$. Alternative names are *global minimizer* and *global minimum*, although the latter could be confused with the minimum value of the function. Similarly, for a *local* solution/minimizer, the condition $f(\mathbf{x}) < f(\mathbf{x}^*)$ only needs to hold in some neighborhood of $\mathbf{x}^*$.

Within this formulation fall many of the traditional optimization problems. An optimization problem in which the objective function and the equality and inequality constraints are linear functions, is called a *linear program*. An optimization problem in which the objective function is quadratic, while the constraints are linear functions is called a *quadratic program.* Convexity plays an important role in many practical optimization problems.

Definition 1.16.1 (Convex Set) A set $\mathscr{X} \in \mathbb{R}^n$ is called *convex* if, for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$ and $\theta \in (0,1)$, the point $(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \in \mathscr{X}$.

Definition 1.16.2 (Convex Function) A function $f(\mathbf{x})$ on a convex set $\mathscr{X}$ is called *convex* if, for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$ and $\theta \in (0,1)$,

$$f\big(\theta\mathbf{x} + (1-\theta)\mathbf{y}\big) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \,. \tag{1.69}$$

If a strict inequality in (1.69) holds, the function is said to be *strictly convex.* If a function f is (strictly) convex, then $-f$ is said to be (strictly) *concave.* Assuming $\mathscr{X}$ is an open set, convexity for $f \in C^1$ is equivalent to

$$f(\mathbf{y}) \geqslant f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^\top \nabla f(\mathbf{x}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathscr{X} \,.$$

Moreover, for $f \in C^2$, convexity is equivalent to the Hessian matrix being positive semidefinite for all $\mathbf{x} \in \mathscr{X}$:

$$\mathbf{y}^\top \left[\nabla^2 f(\mathbf{x})\right] \mathbf{y} \geqslant 0, \quad \text{for all } \mathbf{y} \in \mathbb{R}^n \,.$$

The problem (1.66) is said to be a *convex programming problem* if

1. the objective function f is convex,
2. the inequality constraint functions $\{g_i(\mathbf{x})\}$ are convex, and
3. the equality constraint functions $\{h_i(\mathbf{x})\}$ are *affine*, i.e., of the form $\boldsymbol{a}_i^\top \mathbf{x} - b_i$.

Note that the last requirement follows from the fact that an equality constraint $h_i(\mathbf{x}) = 0$ can be viewed as a combination of the inequality constraints $h_i(\mathbf{x}) \leqslant 0$ and $-h_i(\mathbf{x}) \leqslant 0$, so that both h_i and $-h_i$ need to be convex. Both the linear and quadratic programs (with positive definite matrix C) are convex.

1.16.1 Lagrangian Method

The main components of the Lagrangian method are the Lagrange multipliers and the Lagrange function. The method was developed by Lagrange in 1797 for the optimization problem (1.66) with equality constraints (1.67). In 1951 Kuhn and Tucker extended Lagrange's method to inequality constraints.

Definition 1.16.3 (Lagrange Function) Given an optimization problem (1.66) containing only equality constraints $h_i(\mathbf{x}) = 0,\ i = 1, \ldots, m$, the *Lagrange function*, or *Lagrangian*, is defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_i \beta_i \, h_i(\mathbf{x}) \,,$$

where the coefficients $\{\beta_i\}$ are called the *Lagrange multipliers.*

A necessary condition for a point $\mathbf{x}^*$ to be a local minimizer of $f(\mathbf{x})$ subject to the equality constraints $h_i(\mathbf{x}) = 0,\ i = 1, \ldots, m$, is

$$\begin{aligned} \nabla_{\mathbf{x}}\, \mathcal{L}(\mathbf{x}^*, \boldsymbol{\beta}^*) &= \mathbf{0}\ , \\ \nabla_{\boldsymbol{\beta}}\, \mathcal{L}(\mathbf{x}^*, \boldsymbol{\beta}^*) &= \mathbf{0}\ , \end{aligned}$$

for some value $\boldsymbol{\beta}^*$. The conditions above are also sufficient if $\mathcal{L}(\mathbf{x}, \boldsymbol{\beta}^*)$ is a convex function of $\mathbf{x}$.

■ EXAMPLE 1.16 Maximum Entropy Distribution

Let $p = \{p_i, i = 1, \ldots, n\}$ be a probability distribution. Consider the following program, which maximizes the (Shannon) entropy:

$$\begin{aligned} \max_{\mathbf{p}} \quad & -\sum_{i=1}^{n} p_i \ln p_i \\ \text{subject to:} \quad & \sum_{i=1}^{n} p_i = 1\ . \end{aligned}$$

The Lagrangian is

$$\mathcal{L}(\mathbf{p}, \beta) = \sum_{i=1}^{n} p_i \ln p_i + \beta \left(\sum_{i=1}^{n} p_i - 1 \right)$$

over the domain $\{(\mathbf{p}, \beta) : p_i \geq 0, i = 1, \ldots, n,\ \beta \in \mathbb{R}\}$. The optimal solution $\mathbf{p}^*$ of the problem is the uniform distribution, that is, $\mathbf{p}^* = (1/n, \ldots, 1/n)$; see Problem 1.35.

Definition 1.16.4 (Generalized Lagrange Function) Given the original optimization problem (1.66), containing both the equality and inequality constraints, the *generalized Lagrange function*, or simply *Lagrangian*, is defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{i=1}^{k} \alpha_i\, g_i(\mathbf{x}) + \sum_{i=1}^{m} \beta_i\, h_i(\mathbf{x})\ .$$

A necessary condition for a point $\mathbf{x}^*$ to be a local minimizer of $f(\mathbf{x})$ in the optimization problem (1.66) is the existence of an $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ such that

$$\begin{aligned} \nabla_{\mathbf{x}}\, \mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) &= \mathbf{0}\ , \\ \nabla_{\boldsymbol{\beta}}\, \mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) &= \mathbf{0}\ , \\ g_i(\mathbf{x}^*) &\leqslant 0, \quad i = 1, \ldots, k\ , \\ \alpha_i^* &\geqslant 0, \quad i = 1, \ldots, k\ , \\ \alpha_i^*\, g_i(\mathbf{x}^*) &= 0, \quad i = 1, \ldots, k\ . \end{aligned}$$

These equations are usually referred as the *Karush–Kuhn–Tucker (KKT) conditions.* For *convex* programs we have the following important results:

1. Every local solution $\mathbf{x}^*$ to a convex programming problem is a global solution and the set of global solutions is convex. If, in addition, the objective function is strictly convex, then any global solution is unique.

2. For a strictly convex programming problem with C^1 objective and constraint functions, the KKT conditions are necessary and sufficient for a unique global solution.

1.16.2 Duality

The aim of duality is to provide an alternative formulation of an optimization problem that is often more computationally efficient or has some theoretical significance (see [7], page 219). The original problem (1.66) is referred to as the *primal* problem, whereas the reformulated problem, based on Lagrange multipliers, is referred to as the *dual* problem. Duality theory is most relevant to convex optimization problems. It is well known that if the primal optimization problem is (strictly) convex, then the dual problem is (strictly) concave and has a (unique) solution from which the optimal (unique) primal solution can be deduced.

Definition 1.16.5 (Lagrange Dual Program) The *Lagrange dual program* of the primal program (1.66), is

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \mathcal{L}^*(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ \text{subject to:} \quad & \boldsymbol{\alpha} \geqslant 0 \ , \end{aligned}$$

where $\mathcal{L}^*$ is the *Lagrange dual function*:

$$\mathcal{L}^*(\boldsymbol{\alpha},\boldsymbol{\beta}) = \inf_{\mathbf{x}\in\mathscr{X}} \mathcal{L}(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \ . \tag{1.70}$$

It is not difficult to see that if f^* is the minimal value of the primal problem, then $\mathcal{L}^*(\boldsymbol{\alpha},\boldsymbol{\beta}) \leqslant f^*$ for any $\boldsymbol{\alpha} \geqslant 0$ and any $\boldsymbol{\beta}$. This property is called *weak duality.* The Lagrangian dual program thus determines the best lower bound on f^*. If d^* is the optimal value for the dual problem, then $d^* < f^*$. The difference $f^* - d^*$ is called the *duality gap.*

The duality gap is extremely useful for providing lower bounds for the solutions of primal problems that may be impossible to solve directly. It is important to note that for linearly constrained problems, if the primal is infeasible (does not have a solution satisfying the constraints), then the dual is either infeasible or unbounded. Conversely, if the dual is infeasible, then the primal has no solution. Of crucial importance is the *strong duality* theorem, which states that for convex programs (1.66) with linear constrained functions h_i and g_i the duality gap is zero, and any $\mathbf{x}^*$ and $(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$ satisfying the KKT conditions are (global) solutions to the primal and dual programs, respectively. In particular, this holds for linear and convex quadratic programs (note that not all quadratic programs are convex).

For a convex primal program with C^1 objective and constraint functions, the Lagrangian dual function (1.70) can be obtained by simply setting the gradient (with respect to $\mathbf{x}$) of the Lagrangian $\mathcal{L}(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta})$ to zero. One can further simplify the dual program by substituting into the Lagrangian the relations between the variables thus obtained.

■ EXAMPLE 1.17 Linear Programming Problem

Consider the following linear programming problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{c}^\top \mathbf{x} \\ \text{subject to:} \quad & A\mathbf{x} \geqslant \mathbf{b} \,. \end{aligned}$$

The Lagrangian is $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{c}^\top \mathbf{x} - \boldsymbol{\alpha}^\top (A\mathbf{x} - \mathbf{b})$. The Lagrange dual function is the infimum of $\mathcal{L}$ over all $\mathbf{x}$; thus

$$\mathcal{L}^*(\boldsymbol{\alpha}) = \begin{cases} \mathbf{b}^\top \boldsymbol{\alpha} & \text{if} \quad A^\top \boldsymbol{\alpha} = \mathbf{c} \,, \\ -\infty & \text{otherwise,} \end{cases}$$

so that the Lagrange dual program becomes

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \mathbf{b}^\top \boldsymbol{\alpha} \\ \text{subject to:} \quad & A^\top \boldsymbol{\alpha} = \mathbf{c} \,, \\ & \boldsymbol{\alpha} \geqslant \mathbf{0} \,. \end{aligned}$$

An interesting fact to note here is that for the linear programming problem the dual of the dual problem always gives back the primal problem.

■ EXAMPLE 1.18 Quadratic Programming Problem

Consider the following quadratic programming problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2} \mathbf{x}^\top C \mathbf{x} \\ \text{subject to:} \quad & C\mathbf{x} \geqslant \mathbf{b} \,, \end{aligned}$$

where the $n \times n$ matrix C is assumed to be positive definite (for a general quadratic programming problem the matrix C can always be assumed to be symmetric, but it is not necessarily positive definite). The Lagrangian is $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{x}^\top C\mathbf{x} - \boldsymbol{\alpha}^\top (C\mathbf{x} - \mathbf{b})$. We can minimize this by taking its gradient with respect to $\mathbf{x}$ and setting it to zero. This gives $C\mathbf{x} - C\boldsymbol{\alpha} = C(\mathbf{x} - \boldsymbol{\alpha}) = \mathbf{0}$. The positive definiteness of C implies that $\mathbf{x} = \boldsymbol{\alpha}$. The maximization of the Lagrangian is now reduced to maximizing $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\alpha}) = \frac{1}{2}\,\boldsymbol{\alpha}^\top C\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top (C\boldsymbol{\alpha} - \mathbf{b}) = -\frac{1}{2}\,\boldsymbol{\alpha}^\top C\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{b}$ subject to $\boldsymbol{\alpha} \geqslant \mathbf{0}$. Hence we can write the dual problem as

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2}\,\boldsymbol{\alpha}^\top C\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{b} \\ \text{subject to:} \quad & \boldsymbol{\alpha} \geqslant \mathbf{0} \,. \end{aligned}$$

Notice that the dual problem involves only simple nonnegativity constraints.

Now suppose that we are given the Cholesky factorization $C = BB^\top$. It turns out (see Problem 1.36) that the Lagrange dual of the dual problem above can be written as

$$\begin{aligned} \min_{\boldsymbol{\mu}} \quad & \frac{1}{2}\,\boldsymbol{\mu}^\top \boldsymbol{\mu} \\ \text{subject to:} \quad & B\boldsymbol{\mu} \geqslant \mathbf{b} \,, \end{aligned} \tag{1.71}$$

with $\boldsymbol{\mu} = B^\top \boldsymbol{\alpha}$. This is a so-called *least distance* problem, which, provided that we know the Cholesky factorization of C, is easier to solve than the original quadratic programming problem.

A final example of duality is provided by the widely used *minimum cross-entropy method* [9].

■ EXAMPLE 1.19 Minimum Cross-Entropy (MinxEnt) Method

Let $\mathbf{X}$ be a discrete random variable (or vector) taking values $\mathbf{x}_1, \ldots, \mathbf{x}_r$, and let $\mathbf{q} = (q_1, \ldots, q_r)^\top$ and $\mathbf{p} = (p_1, \ldots, p_r)^\top$ be two strictly positive distribution (column) vectors for $\mathbf{X}$. Consider the following primal program of minimizing the cross-entropy of $\mathbf{p}$ and $\mathbf{q}$, that is, $\sum_{i=1}^n p_i \ln(p_i/q_i)$, for a fixed $\mathbf{q}$, subject to linear equality constraints:

$$\min_{\mathbf{p}} \quad \sum_{k=1}^r p_k \ln \frac{p_k}{q_k} \tag{1.72}$$

$$\text{subject to:} \quad \mathbb{E}_{\mathbf{p}}[S_i(\mathbf{X})] = \sum_{k=1}^r S_i(\mathbf{x}_k)\, p_k = \gamma_i, \quad i = 1, \ldots, m \tag{1.73}$$

$$\sum_{k=1}^r p_k = 1 \, , \tag{1.74}$$

where $S_1, \ldots, S_m$ are arbitrary functions.

Here the objective function is convex, since it is a linear combination of functions of the form $p \ln(p/c)$, which are convex on $\mathbb{R}_+$, for any $c > 0$. In addition, the equality constraint functions are affine (of the form $\mathbf{a}^\top \mathbf{p} - \gamma$). Therefore, this problem is convex. To derive the optimal solution $\mathbf{p}^*$ of the primal program above, it is typically easier to solve the associated *dual* program [9]. Below we present the corresponding procedure.

1. The Lagrangian of the primal problem is given by

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta) = \sum_{k=1}^r p_k \ln \frac{p_k}{q_k} - \sum_{i=1}^m \lambda_i \Big(\sum_{k=1}^r S_i(\mathbf{x}_k)\, p_k - \gamma_i \Big) + \beta \Big(\sum_{k=1}^r p_k - 1 \Big) \, , \tag{1.75}$$

   where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ is the Lagrange multiplier vector corresponding to (1.73) and β is the Lagrange multiplier corresponding to (1.74). Note that we can use either a plus or a minus sign in the second sum of (1.75). We choose the latter because later we generalize the very same problem to inequality ($\geqslant$) constraints in (1.73), giving rise to a minus sign in the Lagrangian.

2. Solve (for fixed $\boldsymbol{\lambda}$ and β)

$$\min_{\mathbf{p}} \mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta) \tag{1.76}$$

   by solving

$$\nabla_{\mathbf{p}} \mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta) = \mathbf{0} \, ,$$

   which gives the set of equations

$$\nabla_{p_k} \mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta) = \ln \frac{p_k}{q_k} + 1 - \sum_{i=1}^m \lambda_i \, S_i(\mathbf{x}_k) + \beta = 0, \quad k = 1, \ldots, r \, .$$

Denote the optimal solution and the optimal function value obtained from the program (1.76) as $\mathbf{p}(\boldsymbol{\lambda}, \beta)$ and $\mathcal{L}^*(\boldsymbol{\lambda}, \beta)$, respectively. The latter is the Lagrange dual function. So we write

$$p_k(\boldsymbol{\lambda}, \beta) = q_k \exp\left(-\beta - 1 + \sum_{i=1}^{m} \lambda_i\, S_i(\mathbf{x}_k)\right), \quad k = 1, \ldots, r\,. \tag{1.77}$$

Since the sum of the $\{p_k\}$ must be 1, we obtain

$$\mathrm{e}^{\beta} = \sum_{k=1}^{r} q_k \exp\left(-1 + \sum_{i=1}^{m} \lambda_i\, S_i(\mathbf{x}_k)\right)\,. \tag{1.78}$$

Substituting $\mathbf{p}(\boldsymbol{\lambda}, \beta)$ back into the Lagrangian gives

$$\mathcal{L}^*(\boldsymbol{\lambda}, \beta) = -1 + \sum_{i=1}^{m} \lambda_i\, \gamma_i - \beta\,. \tag{1.79}$$

3. Solve the *dual* program

$$\max_{\boldsymbol{\lambda}, \beta} \mathcal{L}^*(\boldsymbol{\lambda}, \beta)\,. \tag{1.80}$$

Since β and $\boldsymbol{\lambda}$ are related via (1.78), we can solve (1.80) by substituting the corresponding $\beta(\boldsymbol{\lambda})$ into (1.79) and optimizing the resulting function:

$$D(\boldsymbol{\lambda}) = -1 + \sum_{i=1}^{m} \lambda_i\, \gamma_i - \ln\left\{\sum_{k=1}^{r} q_k\, \exp\{-1 + \textstyle\sum_{i=1}^{m} \lambda_i\, S_i(\mathbf{x}_k)\}\right\}. \tag{1.81}$$

Since $D(\boldsymbol{\lambda})$ is continuously differentiable and concave with respect to $\boldsymbol{\lambda}$, we can derive the optimal solution, $\boldsymbol{\lambda}^*$, by solving

$$\nabla_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda}) = \mathbf{0}\,, \tag{1.82}$$

which can be written componentwise in the following explicit form:

$$\begin{aligned} \nabla_{\lambda_j} D(\boldsymbol{\lambda}) &= \gamma_i - \frac{\sum_{k=1}^{r} S_i(\mathbf{x}_k)\, q_k \exp\left\{-1 + \sum_{j=1}^{m} \lambda_j\, S_j(\mathbf{x}_k)\right\}}{\sum_{k=1}^{r} q_k \exp\left\{-1 + \sum_{j=1}^{m} \lambda_j\, S_j(\mathbf{x}_k)\right\}} \\ &= \gamma_i - \frac{\mathbb{E}_{\mathbf{q}}\left[S_i(\mathbf{X}) \exp\left\{-1 + \sum_{j=1}^{m} \lambda_j\, S_j(\mathbf{X})\right\}\right]}{\mathbb{E}_{\mathbf{q}}\left[\exp\left\{-1 + \sum_{j=1}^{m} \lambda_j\, S_j(\mathbf{X})\right\}\right]} = 0 \end{aligned} \tag{1.83}$$

for $j = 1, \ldots, m$. The optimal vector $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ can be found by solving (1.83) numerically. Note that if the primal program has a nonempty interior optimal solution, then the dual program has an optimal solution $\boldsymbol{\lambda}^*$.

4. Finally, substitute $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ and $\beta = \beta(\boldsymbol{\lambda}^*)$ back into (1.77) to obtain the solution to the original MinxEnt program.

It is important to note that we do not need to explicitly impose the conditions $p_i \geqslant 0$, $i = 1, \ldots, n$, because the quantities $\{p_i\}$ in (1.77) are automatically strictly positive. This is a crucial property of the CE distance; see also

[1]. It is instructive (see Problem 1.37) to verify how adding the nonnegativity constraints affects the procedure above.

When inequality constraints $\mathbb{E}_{\mathbf{p}}[S_i(\mathbf{X})] \geqslant \gamma_i$ are used in (1.73) instead of equality constraints, the solution procedure remains almost the same. The only difference is that the Lagrange multiplier vector $\boldsymbol{\lambda}$ must now be nonnegative. It follows that the dual program becomes

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & D(\boldsymbol{\lambda}) \\ \text{subject to:} \quad & \boldsymbol{\lambda} \geqslant \mathbf{0} \,, \end{aligned}$$

with $D(\boldsymbol{\lambda})$ given in (1.81).

A further generalization is to replace the above discrete optimization problem with a *functional* optimization problem. This topic will be discussed in Chapter 8. In particular, Section 8.9 deals with the MinxEnt method, which involves a functional MinxEnt problem.

PROBLEMS

Probability Theory

1.1 Prove the following results, using the properties of the probability measure in Definition 1.2.1 (here A and B are events):

- **a)** $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- **b)** $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

1.2 Prove the product rule (1.4) for the case of three events.

1.3 We draw three balls consecutively from a bowl containing exactly five white and five black balls, without putting them back. What is the probability that all drawn balls will be black?

1.4 Consider the random experiment where we toss a biased coin until heads comes up. Suppose that the probability of heads on any one toss is p. Let X be the number of tosses required. Show that $X \sim \mathsf{G}(p)$.

1.5 In a room with many people, we ask each person his/her birthday (day and month). Let N be the number of people queried until we get a “duplicate” birthday.

- **a)** Calculate $\mathbb{P}(N > n)$, $n = 0, 1, 2, \ldots$.
- **b)** For which n do we have $\mathbb{P}(N \leqslant n) \geqslant 1/2$?
- **c)** Use a computer to calculate $\mathbb{E}[N]$.

1.6 Let X and Y be independent standard normal random variables, and let U and V be random variables that are derived from X and Y via the linear transformation

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} \sin\alpha & -\cos\alpha \\ \cos\alpha & \sin\alpha \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} .$$

- **a)** Derive the joint pdf of U and V.
- **b)** Show that U and V are independent and standard normally distributed.

1.7 Let $X \sim \mathsf{Exp}(\lambda)$. Show that the *memoryless property* holds: for all $s, t \geqslant 0$,

$$\mathbb{P}(X > t + s \,|\, X > t) = \mathbb{P}(X > s) .$$

1.8 Let X_1, X_2, X_3 be independent Bernoulli random variables with success probabilities 1/2, 1/3, and 1/4, respectively. Give their conditional joint pdf, given that $X_1 + X_2 + X_3 = 2$.

1.9 Verify the expectations and variances in Table 1.3.

1.10 Let X and Y have joint density f given by

$$f(x,y) = c\,x\,y, \quad 0 \leqslant y \leqslant x, \quad 0 \leqslant x \leqslant 1\,.$$

a) Determine the normalization constant c.
b) Determine $\mathbb{P}(X + 2\,Y \leqslant 1)$.

1.11 Let $X \sim \mathsf{Exp}(\lambda)$ and $Y \sim \mathsf{Exp}(\mu)$ be independent. Show that

a) $\min(X,Y) \sim \mathsf{Exp}(\lambda + \mu)$,
b) $\mathbb{P}(X < Y \mid \min(X,Y)) = \dfrac{\lambda}{\lambda + \mu}$.

1.12 Verify the properties of variance and covariance in Table 1.4.

1.13 Show that the correlation coefficient always lies between −1 and 1. [Hint: Use the fact that the variance of $aX + Y$ is always nonnegative, for any a.]

1.14 Consider Examples 1.1 and 1.2. Define X as the function that assigns the number $x_1 + \cdots + x_n$ to each outcome $\omega = (x_1, \ldots, x_n)$. The event that there are exactly k heads in n throws can be written as

$$\{\omega \in \Omega : X(\omega) = k\}\,.$$

If we abbreviate this to $\{X = k\}$, and further abbreviate $\mathbb{P}(\{X = k\})$ to $\mathbb{P}(X = k)$, then we obtain exactly (1.7). Verify that one can always view random variables in this way, that is, as real-valued functions on Ω, and that probabilities such as $\mathbb{P}(X \leqslant x)$ should be interpreted as $\mathbb{P}(\{\omega \in \Omega : X(\omega) \leqslant x\})$.

1.15 Show that

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j)\,.$$

1.16 Let Σ be the covariance matrix of a random column vector $\mathbf{X}$. Write $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the expectation vector of $\mathbf{X}$. Hence $\Sigma = \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top]$. Show that Σ is positive semidefinite. That is, for any vector $\mathbf{u}$, we have $\mathbf{u}^\top \Sigma \mathbf{u} \geqslant 0$.

1.17 Suppose $Y \sim \mathsf{Gamma}(n, \lambda)$. Show that for all $x \geqslant 0$

$$\mathbb{P}(Y \leqslant x) = 1 - \sum_{k=0}^{n-1} \frac{\mathrm{e}^{-\lambda\,x} (\lambda\,x)^k}{k!}\,. \tag{1.84}$$

1.18 Consider the random experiment where we draw uniformly and independently n numbers, $X_1, \ldots, X_n$, from the interval [0,1].

a) Let M be the smallest of the n numbers. Express M in terms of $X_1, \ldots, X_n$.

b) Determine the pdf of M.

1.19 Let $Y = \mathrm{e}^X$, where $X \sim \mathsf{N}(0,1)$.

a) Determine the pdf of Y.

b) Determine the expected value of Y.

1.20 We select a point (X,Y) from the triangle $(0,0) - (1,0) - (1,1)$ in such a way that X has a uniform distribution on $(0,1)$ and the conditional distribution of Y given $X = x$ is uniform on $(0,x)$.

a) Determine the joint pdf of X and Y.

b) Determine the pdf of Y.

c) Determine the conditional pdf of X given $Y = y$ for all $y \in (0,1)$.

d) Calculate $\mathbb{E}[X \mid Y = y]$ for all $y \in (0,1)$.

e) Determine the expectations of X and Y.

Poisson Processes

1.21 Let $\{N_t, t \geqslant 0\}$ be a Poisson process with rate $\lambda = 2$. Find

a) $\mathbb{P}(N_2 = 1, N_3 = 4, N_5 = 5)$,

b) $\mathbb{P}(N_4 = 3 \mid N_2 = 1, N_3 = 2)$,

c) $\mathbb{E}[N_4 \mid N_2 = 2]$,

d) $\mathbb{P}(N[2,7] = 4, N[3,8] = 6)$,

e) $\mathbb{E}[N[4,6] \mid N[1,5] = 3]$.

1.22 Show that for any fixed $k \in \mathbb{N}$, $t > 0$ and $\lambda > 0$,

$$\lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} = \frac{(\lambda t)^k}{k!}\, \mathrm{e}^{-\lambda t} \,.$$

[Hint: Write out the binomial coefficient and use the fact that $\lim_{n\to\infty}\left(1 - \frac{\lambda t}{n}\right)^n = \mathrm{e}^{-\lambda t}$.]

1.23 Consider the Bernoulli approximation in Section 1.12. Let $U_1, U_2, \ldots$ denote the times of success for the Bernoulli process X.

a) Verify that the "intersuccess" times $U_1, U_2 - U_1, \ldots$ are independent and have a geometric distribution with parameter $p = \lambda h$.

b) For small h and $n = \lfloor t/h \rfloor$, show that the relationship $\mathbb{P}(A_1 > t) \approx \mathbb{P}(U_1 > n)$ leads in the limit, as $n \to \infty$, to

$$\mathbb{P}(A_1 > t) = \mathrm{e}^{-\lambda t}.$$

1.24 If $\{N_t, t \geqslant 0\}$ is a Poisson process with rate λ, show that for $0 \leqslant u \leqslant t$ and $j = 0, 1, 2, \ldots, n$,

$$\mathbb{P}(N_u = j \mid N_t = n) = \binom{n}{j} \left(\frac{u}{t}\right)^j \left(1 - \frac{u}{t}\right)^{n-j},$$

that is, the conditional distribution of N_u given $N_t = n$ is binomial with parameters n and u/t.

Markov Processes

1.25 Determine the (discrete) pdf of each X_n, $n = 0, 1, 2, \ldots$, for the random walk in Example 1.10. Also, calculate $\mathbb{E}[X_n]$ and the variance of X_n for each n.

1.26 Let $\{X_n, n \in \mathbb{N}\}$ be a Markov chain with state space $\{0, 1, 2\}$, transition matrix

$$P = \begin{pmatrix} 0.3 & 0.1 & 0.6 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.7 & 0.2 \end{pmatrix},$$

and initial distribution $\pi = (0.2, 0.5, 0.3)$. Determine

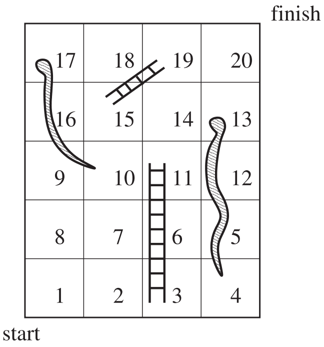
- **a)** $\mathbb{P}(X_1 = 2)$,
- **b)** $\mathbb{P}(X_2 = 2)$,
- **c)** $\mathbb{P}(X_3 = 2 \,|\, X_0 = 0)$,
- **d)** $\mathbb{P}(X_0 = 1 \,|\, X_1 = 2)$,
- **e)** $\mathbb{P}(X_1 = 1, X_3 = 1)$.

1.27 Two dogs harbor a total number of m fleas. Spot initially has b fleas and Lassie has the remaining $m - b$. The fleas have agreed on the following immigration policy: at every time $n = 1, 2 \ldots$, a flea is selected at random from the total population and that flea will jump from one dog to the other. Describe the flea population on Spot as a Markov chain and find its stationary distribution.

1.28 Classify the states of the Markov chain with the following transition matrix:

$$P = \begin{pmatrix} 0.0 & 0.3 & 0.6 & 0.0 & 0.1 \\ 0.0 & 0.3 & 0.0 & 0.7 & 0.0 \\ 0.3 & 0.1 & 0.6 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 & 0.9 & 0.0 \\ 0.1 & 0.1 & 0.2 & 0.0 & 0.6 \end{pmatrix}.$$

1.29 Consider the following snakes-and-ladders game. Let N be the number of tosses required to reach the finish using a fair die. Calculate the expectation of N using a computer.



1.30 Ms. Ella Brum walks back and forth between her home and her office every day. She owns three umbrellas, which are distributed over two umbrella stands (one at home and one at work). When it is not raining, Ms. Brum walks without an umbrella. When it is raining, she takes one umbrella from the stand at the place of her departure, provided there is one available. Suppose that the probability that it is raining at the time of any departure is p. Let X_n denote the number of umbrellas available at the place where Ella arrives after walk number n; $n = 1, 2, \ldots$, including the one that she possibly brings with her. Calculate the limiting probability that it rains and no umbrella is available.

1.31 A mouse is let loose in the maze of Figure 1.10. From each compartment the mouse chooses one of the adjacent compartments with equal probability, independent of the past. The mouse spends an exponentially distributed amount of time in each compartment. The mean time spent in each of the compartments 1, 3, and 4 is two seconds; the mean time spent in compartments 2, 5, and 6 is four seconds. Let $\{X_t, t \geqslant 0\}$ be the Markov jump process that describes the position of the mouse for times $t \geqslant 0$. Assume that the mouse starts in compartment 1 at time $t = 0$.

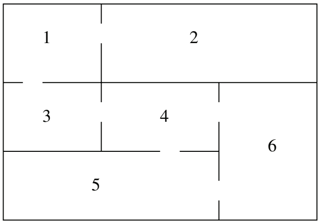


Figure 1.10: A maze.

What are the probabilities that the mouse will be found in each of the compartments $1, 2, \ldots, 6$ at some time t far away in the future?

1.32 In an $M/M/\infty$-queueing system, customers arrive according to a Poisson process with rate a. Every customer who enters is immediately served by one of an infinite number of servers; hence there is no queue. The service times are exponentially distributed, with mean $1/b$. All service and interarrival times are independent. Let X_t be the number of customers in the system at time t. Show that the limiting distribution of X_t, as $t \to \infty$, is Poisson with parameter a/b.

Optimization

1.33 Let $\mathbf{a}$ and let $\mathbf{x}$ be n-dimensional column vectors. Show that $\nabla_{\mathbf{x}}\, \mathbf{a}^\top \mathbf{x} = \mathbf{a}$.

1.34 Let A be a symmetric $n \times n$ matrix and $\mathbf{x}$ be an n-dimensional column vector. Show that $\nabla_{\mathbf{x}}\, \frac{1}{2}\, \mathbf{x}^\top A \mathbf{x} = A\mathbf{x}$. What is the gradient if A is not symmetric?

1.35 Show that the optimal distribution $\mathbf{p}^*$ in Example 1.16 is given by the uniform distribution.

1.36 Derive the program (1.71).

1.37 Consider the MinxEnt program

$$\min_{\mathbf{p}} \quad \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

$$\text{subject to:} \quad \mathbf{p} \geqslant \mathbf{0}, \quad A\mathbf{p} = \mathbf{b}, \quad \sum_{i=1}^{n} p_i = 1 \ ,$$

where $\mathbf{p}$ and $\mathbf{q}$ are probability distribution vectors and A is an $m \times n$ matrix.

a) Show that the Lagrangian for this problem is of the form

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta, \boldsymbol{\mu}) = \mathbf{p}^\top \boldsymbol{\xi}(\mathbf{p}) - \boldsymbol{\lambda}^\top (A\mathbf{p} - \mathbf{b}) - \boldsymbol{\mu}^\top \mathbf{p} + \beta(\mathbf{1}^\top \mathbf{p} - 1) \ .$$

b) Show that $p_i = q_i \exp(-\beta - 1 + \mu_i + \sum_{j=1}^{m} \lambda_j \, a_{ji})$, for $i = 1, \ldots, n$.

c) Explain why, as a result of the KKT conditions, the optimal $\boldsymbol{\mu}^*$ must be equal to the zero vector.

d) Show that the solution to this MinxEnt program is exactly the same as for the program where the nonnegativity constraints are omitted.

Further Reading

An easy introduction to probability theory with many examples is [13], and a more detailed textbook is [8]. A classical reference is [6]. An accurate and accessible treatment of various stochastic processes is given in [3]. For convex optimization we refer to [2] and [7].

REFERENCES

1. Z. I. Botev, D. P. Kroese, and T. Taimre. Generalized cross-entropy methods for rare-event simulation and optimization. *Simulation: Transactions of the Society for Modeling and Simulation International*, 83(11):785–806, 2007.
2. S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, Cambridge, UK, 2004.
3. E. Çinlar. *Introduction to Stochastic Processes.* Prentice Hall, Englewood Cliffs, NJ, 1975.
4. T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.
5. C. W. Curtis. *Linear Algebra: An Introductory Approach.* Springer-Verlag, New York, 1984.
6. W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley & Sons, New York, 2nd edition, 1970.
7. R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, New York, 1987.
8. G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes.* Oxford University Press, Oxford, 3rd edition, 2001.
9. J. N. Kapur and H. K. Kesavan. *Entropy Optimization Principles with Applications.* Academic Press, New York, 1992.