Mathematical Foundations of Computer Networking

Srinivasan Keshav

ADDISON-WESLEY PROFESSIONAL COMPUTING SERIES

Mathematical Foundations of Computer Networking

Mathematical Foundations of Computer Networking

Srinivasan Keshav

Addison-Wesley

Upper Saddle River, NJ • Boston • Indianapolis • San Francisco
New York • Toronto • Montreal • London • Munich • Paris • Madrid
Capetown • Sydney • Tokyo • Singapore • Mexico City

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact:

U.S. Corporate and Government Sales
(800) 382-3419
corpsales@pearsontechgroup.com

For sales outside the United States, please contact:

International Sales
international@pearson.com

Visit us on the Web: informit.com/aw

Library of Congress Cataloging-in-Publication Data

Keshav, Srinivasan.
Mathematical foundations of computer networking / Srinivasan Keshav.
p. cm.
Includes index.
ISBN 978-0-321-79210-5 (pbk. : alk. paper)—ISBN 0-321-79210-6 (pbk. : alk. paper) 1. Computer networks—Mathematics—Textbooks. I. Title.
TK5105.5.K484 2012
004.601'519—dc23

2011052203

ISBN-13: 978-0-321-79210-5
ISBN-10: 0-321-79210-6
Text printed in the United States on recycled paper at RR Donnelley in Crawfordsville, Indiana.
First printing, April 2012

Editor-in-Chief
Mark L. Taub

Senior Acquisitions Editor
Trina MacDonald

Managing Editor
John Fuller

Full-Service Production Manager
Julie B. Nahil

Copy Editor
Evelyn Pyle

Indexer
Ted Laux

Proofreader
Linda Begley

Technical Reviewers
Alan Kaplan
Abraham Matta
Johnny Wong

Publishing Coordinator
Olivia Basegio

Compositor
Rob Mauhar

Contents

Preface

Motivation

Graduate students, researchers, and professionals in the field of computer networking often require a firm conceptual understanding of its theoretical foundations. Knowledge of optimization, information theory, game theory, control theory, and queueing theory is assumed by research papers in the field. Yet these subjects are not taught in a typical computer science undergraduate curriculum. This leaves only two alternatives: to either study these topics on one's own from standard texts or take a remedial course. Neither alternative is attractive. Standard texts pay little attention to computer networking in their choice of problem areas, making it a challenge to map from the text to the problem at hand, and it is inefficient to require students to take an entire course when all that is needed is an introduction to the topic.

This book addresses these problems by providing a single source to learn about the mathematical foundations of computer networking. Assuming only a rudimentary grasp of calculus, the book provides an intuitive yet rigorous introduction to a wide range of mathematical topics. The topics are covered in sufficient detail so that the book will usually serve as both the first and ultimate reference. Note that the topics are selected to be *complementary* to those found in a typical undergraduate computer science curriculum. The book, therefore, does not cover network foundations, such as discrete mathematics, combinatorics, or graph theory.

Each concept in the book is described in four ways: intuitively, using precise mathematical notation, providing a carefully chosen numerical example, and offering a numerical exercise to be done by the reader. This progression is designed to gradually deepen understanding. Nevertheless, the depth of coverage provided here is not a substitute for that found in standard textbooks. Rather, I hope to provide enough intuition to allow a student to grasp the essence of a research paper that uses these theoretical foundations.

Organization

The chapters in this book fall into two broad categories: foundations and theories. The first five chapters are foundational, covering probability, statistics, linear algebra, optimization, and signals, systems, and transforms. These chapters provide the basis for the four theories covered in the latter half of the book: queueing theory, game theory, control theory, and information theory. Each chapter is written to be as self-contained as possible. Nevertheless, some dependencies do exist, as shown in Figure P.1, where dashed arrows show weak dependencies and solid arrows show strong dependencies.

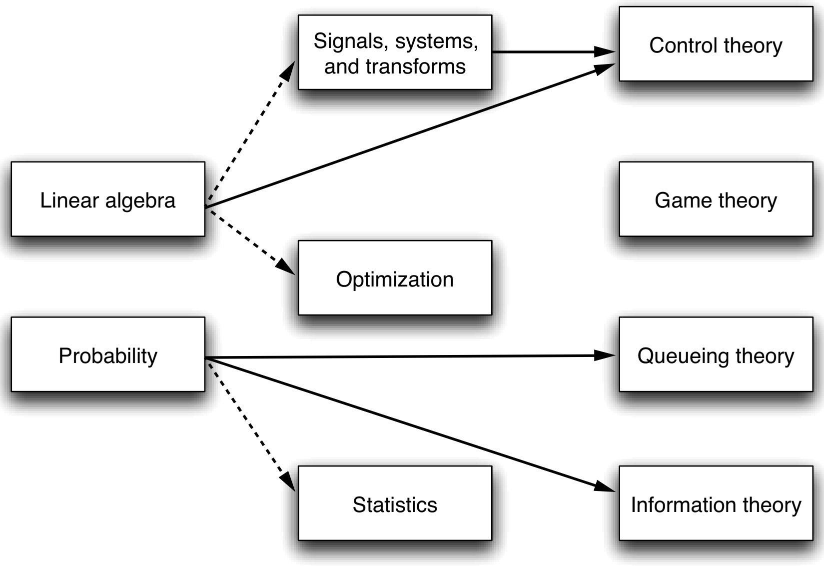

Figure P.1 Chapter organization

Using This Book

The material in this book can be completely covered in a sequence of two graduate courses, with the first course focusing on the first five chapters and the second course on the latter four. For a single-semester course, some possible alternatives are to cover

- Probability, statistics, queueing theory, and information theory
- Linear algebra; signals, systems, and transforms; control theory; and game theory
- Linear algebra; signals, systems, and transforms; control theory; selected portions of probability; and information theory
- Linear algebra; optimization, probability, queueing theory, and information theory

This book is designed for self-study. Each chapter has numerous solved examples and exercises to reinforce concepts. My aim is to ensure that every topic in the book is accessible to the perservering reader.

Acknowledgments

I have benefited immensely from the comments of dedicated reviewers on drafts of this book. Two reviewers in particular who stand out are Alan Kaplan, whose careful and copious comments improved every aspect of the book, and Johnny Wong, who not only reviewed multiple drafts of the chapters on probability and statistics but also used a draft to teach two graduate courses at the University of Waterloo.

I would also like to acknowledge the support I received from experts who reviewed individual chapters: Augustin Chaintreau, Columbia (probability and queueing theory); Tom Coleman, Waterloo (optimization); George Labahn, Waterloo (linear algebra); Kate Larson, Waterloo (game theory); Abraham Matta, Boston University (statistics; signals, systems, and transforms; and control theory); Sriram Narasimhan, Waterloo (control theory); and David Tse, UC Berkeley (information theory).

I received many corrections from my University of Waterloo students who took two courses based on book drafts in Fall 2008 and Fall 2011: Andrew Arnold, Nasser Barjesteh, Omar Beg, Abhirup Chakraborty, Betty Chang, Leila Chenaei, Francisco Claude, Andy Curtis, Hossein Falaki, Leong Fong, Bo Hu, Tian Jiang, Milad Khalki, Robin Kothari, Alexander Laplante, Constantine Murenin, Earl Oliver, Sukanta Pramanik, Ali Rajabi, Aaditeshwar Seth, Jakub Schmidtke, Kanwaljit Singh, Kellen Steffen, Chan Tang, Alan Tsang, Navid Vafaei, and Yuke Yang.

I would like to thank the staff of Addison-Wesley responsible for publishing this book, especially my editor, Trina MacDonald, and production editor, Julie Nahil.

Last but not the least, I would never have completed this book were it not for the unstinting support and encouragement from every member of my family—in particular, my wife, Nicole, and my daughters, Maya and Leela—for the last five years. Thank you.

—S. Keshav
Waterloo, February 2012

1

Probability

1.1 Introduction

The concept of probability pervades every aspect of our lives. Weather forecasts are couched in probabilistic terms, as are economic predictions and even outcomes of our own personal decisions. Designers and operators of computer networks need to often think probabilistically, for instance, when anticipating future traffic workloads or computing cache hit rates. From a mathematical standpoint, a good grasp of probability is a necessary foundation to understanding statistics, game theory, and information theory. For these reasons, the first step in our excursion into the mathematical foundations of computer networking is to study the concepts and theorems of probability.

This chapter is a self-contained introduction to the theory of probability. We begin by introducing the elementary concepts of outcomes, events, and sample spaces, which allows us to precisely define the conjunctions and disjunctions of events. We then discuss concepts of conditional probability and Bayes's rule. This is followed by a description of discrete and continuous random variables, expectations and other moments of a random variable, and the moment generating function. We discuss some standard discrete and continuous distributions and conclude with some useful theorems of probability and a description of Bayesian networks.

Note that in this chapter, as in the rest of the book, the solved examples are an essential part of the text. They provide a concrete grounding for otherwise abstract concepts and are necessary to understand the material that follows.

1.1.1 Outcomes

The mathematical theory of probability uses terms such as *outcome* and *event* with meanings that differ from those in common practice. Therefore, we first introduce a standard set of terms to precisely discuss probabilistic processes. These terms are shown in boldface. We will use the same convention to introduce other mathematical terms in the rest of the book.

Probability measures the degree of uncertainty about the potential **outcomes** of a **process**. Given a set of **distinct** and **mutually exclusive** outcomes of a process, denoted $\{o_1, o_2, \ldots\}$, called the **sample space** S, the probability of any outcome, denoted $P(o_i)$, is a real number between 0 and 1, where 1 means that the outcome will surely occur, 0 means that it surely will not occur, and intermediate values reflect the degree to which one is confident that the outcome will or will not occur.[1] We assume that it is certain that *some* element in S occurs. Hence, the elements of S describe all possible outcomes, and the sum of probability of all the elements of S is always 1.

EXAMPLE 1.1: SAMPLE SPACE AND OUTCOMES

Imagine rolling a six-faced die numbered 1 through 6. The process is that of rolling a die, and an outcome is the number shown on the upper horizontal face when the die comes to rest. Note that the outcomes are distinct and mutually exclusive because there can be only one upper horizontal face corresponding to each throw.

The sample space is $S = \{1, 2, 3, 4, 5, 6\}$, which has a size $|S| = 6$. If the die is fair, each outcome is equally likely, and the probability of each outcome is $\frac{1}{|S|} = \frac{1}{6}$.

EXAMPLE 1.2: INFINITE SAMPLE SPACE AND ZERO PROBABILITY

Imagine throwing a dart at random onto a dartboard of unit radius. The process is that of throwing a dart, and the outcome is the point where the dart penetrates the dartboard. We will assume that this point is vanishingly small, so that it can be thought of as a point on a two-dimensional real plane. Then, the outcomes are distinct and mutually exclusive.

The sample space S is the infinite set of points that lie within a unit circle in the real plane. If the dart is thrown truly randomly, every outcome is equally likely; because the outcomes are infinite, every outcome has a **probability of zero**. We need special care in dealing with such outcomes. It turns

1. Strictly speaking, S must be a measurable σ field.

out that, in some cases, it is necessary to interpret the probability of the occurrence of such an event as being vanishingly small rather than exactly zero. We consider this situation in greater detail in Section 1.1.5. Note that although the probability of any particular outcome is zero, the probability associated with any *subset* of the unit circle with area a is given by $\frac{a}{\pi}$, which tends to zero as a tends to zero.

1.1.2 Events

The definition of probability naturally extends to any subset of elements of S, which we call an **event**, denoted E. If the sample space is discrete, every event E is an element of the power set of S, which is the set of all possible subsets of S. The probability associated with an event, denoted $P(E)$, is a real number $0 \le P(E) \le 1$ and is the sum of the probabilities associated with the outcomes in the event.

EXAMPLE 1.3: EVENTS

Continuing with Example 1.1, we can define the event "the roll of a die results in an odd-numbered outcome." This corresponds to the set of outcomes {1,3,5}, which has a probability of $\frac{1}{6}+\frac{1}{6}+\frac{1}{6} = \frac{1}{2}$. We write $P(\{1,3,5\}) = 0.5$.

1.1.3 Disjunctions and Conjunctions of Events

Consider an event E that is considered to have occurred if either or both of two other events E_1 or E_2 occur, where both events are defined in the same sample space. Then, E is said to be the **disjunction**, or logical OR, of the two events denoted $E = E_1 \vee E_2$ and read "E_1 or E_2."

EXAMPLE 1.4: DISJUNCTION OF EVENTS

Continuing with Example 1.1, we define the events E_1 = "the roll of a die results in an odd-numbered outcome" and E_2 = "the roll of a die results in an outcome numbered less than 3." Then, $E_1 = \{1, 3, 5\}$ and $E_2 = \{1, 2\}$ and $E = E_1 \vee E_2 = \{1, 2, 3, 5\}$.

In contrast, consider event E that is considered to have occurred only if *both* of two other events E_1 or E_2 occur, where both are in the same sample space. Then, E

is said to be the **conjunction**, or logical AND, of the two events denoted $E = E_1 \wedge E_2$ and read "E_1 and E_2." When the context is clear, we abbreviate this to $E = E_1E_2$.

EXAMPLE 1.5: CONJUNCTION OF EVENTS

Continuing with Example 1.4, $E = E_1 \wedge E_2 = E_1E_2 = \{1\}$.

Two events E_i and E_j in S are **mutually exclusive** if only one of the two may occur simultaneously. Because the events have no outcomes in common, $P(E_i \wedge E_j) = P(\{\ \}) = 0$. Note that outcomes are *always* mutually exclusive, but events need not be so.

1.1.4 Axioms of Probability

One of the breakthroughs in modern mathematics was the realization that the theory of probability can be derived from just a handful of intuitively obvious axioms. Several variants of the axioms of probability are known. We present the three axioms as stated by Kolmogorov to emphasize the simplicity and elegance that lie at the heart of probability theory.

1. $0 \le P(E) \le 1$; that is, the probability of an event lies between 0 and 1.
2. $P(S) = 1$, that is, it is certain that at least some event in S will occur.
3. Given a potentially infinite set of *mutually exclusive* events $E_1, E_2, \ldots$

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \qquad \textbf{(EQ 1.1)}$$

That is, the probability that any *one* of the events in the set of mutually exclusive events occurs is the sum of their individual probabilities. For any finite set of n mutually exclusive events, we can state the axiom equivalently as

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) \qquad \textbf{(EQ 1.2)}$$

An alternative form of axiom 3 is:

$$P(E_1 \vee E_2) = P(E_1) + P(E_2) - P(E_1 \wedge E_2) \qquad \textbf{(EQ 1.3)}$$

This alternative form applies to non–mutually exclusive events.

EXAMPLE 1.6: PROBABILITY OF UNION OF MUTUALLY EXCLUSIVE EVENTS

Continuing with Example 1.1, we define the mutually exclusive events {1, 2} and {3, 4}, which both have a probability of 1/3. Then, $P(\{1, 2\} \cup \{3, 4\}) = P(\{1, 2\}) + P(\{3, 4\}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

EXAMPLE 1.7: PROBABILITY OF UNION OF NON–MUTUALLY EXCLUSIVE EVENTS

Continuing with Example 1.1, we define the non–mutually exclusive events {1, 2} and {2, 3}, which both have a probability of 1/3. Then, $P(\{1, 2\} \cup \{2, 3\}) = P(\{1, 2\}) + P(\{2, 3\}) - P(\{1, 2\} \wedge \{2, 3\}) = \frac{1}{3} + \frac{1}{3} - P(\{2\}) = \frac{2}{3} - \frac{1}{6} = \frac{1}{2}$.

1.1.5 Subjective and Objective Probability

The axiomatic approach is indifferent as to *how* the probability of an event is determined. It turns out that there are two distinct ways in which to determine the probability of an event. In some cases, the probability of an event can be derived from counting arguments. For instance, given the roll of a fair die, we know that only six outcomes are possible and that all outcomes are equally likely, so that the probability of rolling, say, a 1, is 1/6. This is called its **objective** probability. Another way of computing objective probabilities is to define the probability of an event as being the limit of a counting process, as the next example shows.

EXAMPLE 1.8: PROBABILITY AS A LIMIT

Consider a measurement device that measures the packet header types of every packet that crosses a link. Suppose that during the course of a day, the device samples 1,000,000 packets, of which 450,000 are UDP packets, 500,000 are TCP packets, and the rest are from other transport protocols. Given the large number of underlying observations, to a first approximation, we can consider the probability that a randomly selected packet uses the UDP protocol to be 450,000/1,000,000 = 0.45. More precisely, we state

$$P(UDP) = \underset{t \to \infty}{Lim} (UDPCount(t))/(TotalPacketCoun(t)),$$

where $UDPCount(t)$ is the number of UDP packets seen during a measurement interval of duration t, and $TotalPacketCount(t)$ is the total number of packets seen during the same measurement interval. Similarly, $P(TCP) = 0.5$.

Note that in reality, the mathematical limit cannot be achieved, because no packet trace is infinite. Worse, over the course of a week or a month, the underlying workload could change, so that the limit may not even exist. Therefore, in practice, we are forced to choose "sufficiently large" packet counts and hope that the ratio thus computed corresponds to a probability. This approach is also called the **frequentist** approach to probability.

In contrast to an objective assessment of probability, we can also use probabilities to characterize events **subjectively**.

EXAMPLE 1.9: SUBJECTIVE PROBABILITY AND ITS MEASUREMENT

Consider a horse race in which a favored horse is likely to win, but this is by no means assured. We can associate a subjective probability with the event, say, 0.8. Similarly, a doctor may look at a patient's symptoms and associate them with a 0.25 probability of a particular disease. Intuitively, this measures the degree of confidence that an event will occur, based on expert knowledge of the situation that is not (or cannot be) formally stated.

How is subjective probability to be determined? A common approach is to measure the odds that a knowledgeable person would bet on that event. Continuing with the example, a bettor who really thought that the favorite would win with a probability of 0.8, should be willing to bet $1 under the terms: If the horse wins, the bettor gets $1.25; if the horse loses, the bettor gets $0. With this bet, the bettor expects to not lose money; if the reward is greater than $1.25, the bettor will expect to make money. We can elicit the implicit subjective probability by offering a high reward and then lowering it until the bettor is just about to walk away, which would be at the $1.25 mark.

The subjective and frequentist approaches interpret zero-probability events differently. Consider an infinite sequence of successive events. Any event that occurs only a finite number of times in this infinite sequence will have a frequency that can be made arbitrarily small. In number theory, we do not and cannot differentiate between a number that can be made arbitrarily small and zero. So, from this perspective, such an event can be considered to have a probability of occurrence of zero *even though it may occur a finite number of times* in the sequence.

From a subjective perspective, a zero-probability event is defined as an event E such that a rational person would be willing to bet an arbitrarily large but finite amount that E will not occur. More concretely, suppose that this person were to receive a reward of $1 if E did not occur but would have to forfeit a sum of \$$F$ if E occurred. Then, the bet would be taken for any finite value of F.

1.2 Joint and Conditional Probability

Thus far, we have defined the terms used in studying probability and considered single events in isolation. Having set this foundation, we now turn our attention to the interesting issues that arise when studying **sequences of events**. In doing so, it is very important to keep track of the sample space in which the events are defined: A common mistake is to ignore the fact that two events in a sequence may be defined on different sample spaces.

1.2.1 Joint Probability

Consider two processes with sample spaces S_1 and S_2 that occur one after the other. The two processes can be viewed as a single **joint process** whose outcomes are the tuples chosen from the **product space** $S_1 \times S_2$. We refer to the subsets of the product space as **joint events**. Just as before, we can associate probabilities with outcomes and events in the product space. To keep things straight, in this section, we denote the sample space associated with a probability as a subscript, so that $P_{S_1}(E)$ denotes the probability of event E defined over sample space S_1, and $P_{S_1 \times S_2}(E)$ is an event defined over the product space $S_1 \times S_2$.

EXAMPLE 1.10: JOINT PROCESS AND JOINT EVENTS

Consider sample space $S_1 = \{1, 2, 3\}$ and sample space $S_2 = \{a, b, c\}$. Then, the product space is given by $\{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c), (3, a), (3, b), (3, c)\}$. If these events are equiprobable, the probability of each tuple is $\frac{1}{9}$. Let $E = \{1, 2\}$ be an event in S_1 and $F = \{b\}$ be an event in S_2. Then, the event EF is given by the tuples $\{(1, b), (2, b)\}$ and has probability $\frac{1}{9} + \frac{1}{9} = \frac{2}{9}$.

We will return to the topic of joint processes in Section 1.8. We now turn our attention to the concept of conditional probability.

1.2.2 Conditional Probability

Common experience tells us that if a sky is sunny, there is no chance of rain in the immediate future but that if the sky is cloudy, it may or may not rain soon. Knowing that the sky is cloudy, therefore, increases the chance that it may rain soon, compared to the situation when it is sunny. How can we formalize this intuition?

To keep things simple, first consider the case when two events E and F share a common sample space S and occur one after the other. Suppose that the probability of E is $P_S(E)$ and the probability of F is $P_S(F)$. Now, suppose that we are informed that event E actually occurred. By definition, the **conditional probability** of the event F conditioned on the occurrence of event E is denoted $P_{S \times S}(F|E)$ (read "the probability of F given E") and computed as

$$P_{S \times S}(F|E) = \frac{P_{S \times S}(E \wedge F)}{P_S(E)} = \frac{P_{S \times S}(EF)}{P_S(E)} \qquad \text{(EQ 1.4)}$$

If knowing that E occurred does not affect the probability of F, E and F are said to be **independent** and

$$P_{S \times S}(EF) = P_S(E)P_S(F)$$

EXAMPLE 1.11: CONDITIONAL PROBABILITY OF EVENTS DRAWN FROM THE SAME SAMPLE SPACE

Consider sample space $S = \{1, 2, 3\}$ and events $E = \{1\}$ and $F = \{3\}$. Let $P_S(E) = 0.5$ and $P_S(F) = 0.25$. Clearly, the space $S \times S = \{(1, 1), (1, 2), ..., (3, 2), (3, 3)\}$. The joint event $EF = \{(1, 3)\}$. Suppose that $P_{S \times S}(EF) = 0.3$. Then,

$$P_{S \times S}(F|E) = \frac{P_{S \times S}(EF)}{P_S(E)} = \frac{0.3}{0.5} = 0.6$$

We interpret this to mean that if event E occurred, the probability that event F occurs is 0.6. This is higher than the probability of F occurring on its own (which is 0.25). Hence, the fact the E occurred improves the chances of F occurring, so the two events are not independent. This is also clear from the fact that $P_{S \times S}(EF) = 0.3 \neq P_S(E)P_S(F) = 0.125$.

The notion of conditional probability generalizes to the case in which events are defined on more than one sample space. Consider a sequence of two processes with sample spaces S_1 and S_2 that occur one after the other. (This could be the condition of the sky now, for instance, and whether it rains after 2 hours.) Let event E be a subset of S_1 and event F a subset of S_2. Suppose that the probability of E is $P_{S_1}(E)$ and the probability of F is $P_{S_2}(F)$. Now, suppose that we are informed that event E occurred. We define the probability $P_{S_1 \times S_2}(F|E)$ as the **conditional probability** of the event F conditional on the occurrence of E as

$$P_{S_1 \times S_2}(F|E) = \frac{P_{S_1 \times S_2}(EF)}{P_{S_1}(E)} \qquad \text{(EQ 1.5)}$$

If knowing that E occurred does not affect the probability of F, E and F are said to be **independent** and

$$P_{S_1 \times S_2}(EF) = P_{S_1}(E) \times P_{S_2}(F) \quad \text{(EQ 1.6)}$$

EXAMPLE 1.12: CONDITIONAL PROBABILITY OF EVENTS DRAWN FROM DIFFERENT SAMPLE SPACES

Consider sample space $S_1 = \{1, 2, 3\}$ and sample space $S_2 = \{a, b, c\}$ with product space $\{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c), (3, a), (3, b), (3, c)\}$. Let $E = \{1, 2\}$ be an event in S_1 and $F = \{b\}$ be an event in S_2. Also, let $P_{S_1}(E) = 0.5$, and let $P_{S_1 \times S_2}(EF) = P_{S_1 \times S_2}(\{(1, b), (2, b)\}) = 0.05$.

If E and F are independent,

$$P_{S_1 \times S_2}(EF) = P_{S_1 \times S_2}(\{(1, b), (2, b)\}) = P_{S_1}(\{1, 2\}) \times P_{S_2}(\{b\})$$

$$0.05 = 0.5 \times P_{S_2}(\{b\})$$

$$P_{S_2}(\{b\}) = 0.1$$

Otherwise,

$$P_{S_1 \times S_2}(F|E) = \frac{P_{S_1 \times S_2}(EF)}{P_{S_1}(E)} = \frac{0.05}{0.5} = 0.1$$

It is important not to confuse $P(F | E)$ and $P(F)$. The conditional probability is defined in the product space $S_1 \times S_2$ and the unconditional probability in the space S_2. Explicitly keeping track of the underlying sample space can help avoid apparent contradictions such as the one discussed in Example 1.14.

EXAMPLE 1.13: USING CONDITIONAL PROBABILITY

Consider a device that samples packets on a link, as in Example 1.8. Suppose that measurements show that 20% of the UDP packets have a packet size of 52 bytes. Let $P(UDP)$ denote the probability that the packet is of type UDP, and let $P(52)$ denote the probability that the packet is of length 52 bytes. Then, $P(52 | UDP) = 0.2$. In Example 1.8, we computed that $P(UDP) = 0.45$. Therefore, $P(UDP \text{ AND } 52) = P(52 | UDP) * P(UDP) = 0.2 * 0.45 = 0.09$. That is, if we were to pick a packet at random from the sample, there is a 9% chance that it is a UDP packet of length 52 bytes, but it has a 20% chance of being of length 52 bytes if we know already that it is a UDP packet.

EXAMPLE 1.14: THE MONTY HALL PROBLEM

Consider a television show (loosely modeled on a similar show hosted by Monty Hall) in which three identical doors hide two goats and a luxury car. You, the contestant, can pick any door and obtain the prize behind it. Assume that you prefer the car to the goat. If you did not have any further information, your chance of picking the winning door is clearly 1/3. Now, suppose that after you pick one of the doors—say, Door 1—the host opens one of the other doors—say, Door 2—and reveals a goat behind it. Should you switch your choice to Door 3 or stay with Door 1?

Solution:

We can view the Monty Hall problem as a sequence of three processes: (1) the placement of a car behind one of the doors, (2) the selection of a door by the contestant, and (3) the revelation of what lies behind one of the other doors. The sample space for the first process is {Door 1, Door 2, Door 3}, abbreviated {1, 2, 3}, as are the sample spaces for the second and third processes. So, the product space is {(1, 1, 1), (1, 1, 2), (1, 1, 3), (1, 2, 1),..., (3, 3, 3)}.

Without loss of generality, assume that you pick Door 1. The game show host is now forced to pick either Door 2 or Door 3. Without loss of generality, suppose that the host picks Door 2, so that the set of possible outcomes that constitutes the reduced sample space is {(1, 1, 2), (2, 1, 2), (3, 1, 2)}. However, we know that the game show host will never open a door with a car behind it. Therefore, the outcome (2, 1, 2) is not possible. So, the reduced sample space is just the set {(1, 1, 2), (3, 1, 2)}. What are the associated probabilities?

To determine this, note that the initial probability space is {1, 2, 3} with equiprobable outcomes. Therefore, the outcomes {(1, 1, 2), (2, 1, 2), (3, 1, 2)} are also equiprobable. When moving to open Door 2, the game show host reveals private information that the outcome (2, 1, 2) is impossible, so the probability associated with this outcome is 0. The show host's forced move cannot affect the probability of the outcome (1, 1, 2), because the host never had the choice of opening Door 1 once you selected it. Therefore, its probability in the reduced sample space continues to be 1/3. This means that $P(\{(3, 1, 2)\}) = 2/3$, so it doubles your chances for you to switch doors.

One way to understand this somewhat counterintuitive result is to realize that the game show host's actions reveal private information, that is, the location of the car. Two-thirds of the time, the prize is behind the door you did not choose. The host always opens a door that does not have a prize behind it.

Therefore, the residual probability (2/3) must all be assigned to Door 3. Another way to think of it is that if you repeat a large number of experiments with two contestants—one who never switches doors and the other who always switches doors—the latter would win twice as often.

1.2.3 Bayes's Rule

One of the most widely used rules in the theory of probability is due to an English country minister: Thomas Bayes. Its significance is that it allows us to infer "backwards" from effects to causes rather than from causes to effects. The derivation of his rule is straightforward, though its implications are profound.

We begin with the definition of conditional probability (Equation 1.4):

$$P_{S \times S}(F|E) = \frac{P_{S \times S}(EF)}{P_S(E)}$$

If the underlying sample spaces can be assumed to be implicitly known, we can rewrite this as

$$P(EF) = P(F|E)P(E) \tag{EQ 1.7}$$

We interpret this to mean that the probability that both E and F occur is the product of the probabilities of two events: first, that E occurs; second, that conditional on E, F occurs.

Recall that $P(F|E)$ is defined in terms of the event F following event E. Now, consider the converse: F is known to have occurred. What is the probability that E occurred? This is similar to the problem: If there is fire, there is smoke, but if we see smoke, what is the probability that it was due to a fire? The probability we want is $P(E|F)$. Using the definition of conditional probability, it is given by

$$P(E|F) = \frac{P(EF)}{P(F)} \tag{EQ 1.8}$$

Substituting for $P(F)$ from Equation 1.7, we get

$$P(E|F) = \frac{P(F|E)}{P(F)}P(E) \tag{EQ 1.9}$$

which is **Bayes's rule**. One way of interpreting this is that it allows us to compute the degree to which some effect, or **posterior** F, can be attributed to some cause, or **prior** E.

EXAMPLE 1.15: BAYES'S RULE

Continuing with Example 1.13, we want to compute the following quantity: Given that a packet is 52 bytes long, what is the probability that it is a UDP packet?

Solution:

From Bayes's rule:

$$P(UDP|52) = \frac{P(52|UDP)P(UDP)}{P(52)} = \frac{0.2(0.45)}{0.54} = 0.167$$

We can generalize Bayes's rule when a posterior can be attributed to more than one prior. Consider a posterior F that is due to some set of n priors E_i such that the priors are mutually exclusive and exhaustive: That is, at least one of them occurs, and only one of them can occur. This implies that $\sum_{i=1}^{n} P(E_i) = 1$. Then,

$$P(F) = \sum_{i=1}^{n} P(FE_i) = \sum_{i=1}^{n} P(F|E_i)P(E_i) \qquad \textbf{(EQ 1.10)}$$

This is also called the **law of total probability**.

EXAMPLE 1.16: LAW OF TOTAL PROBABILITY

Continuing with Example 1.13, let us compute $P(52)$, that is, the probability that a packet sampled at random has a length of 52 bytes. To compute this, we need to know the packet sizes for all other traffic types. For instance, if $P(52|TCP) = 0.9$ and all other packets were known to be of length other than 52 bytes, then $P(52) = P(52|UDP) * P(UDP) + P(52|TCP) * P(TCP) + P(52|other) * P(other) = 0.2 * 0.45 + 0.9 * 0.5 + 0 = 0.54$.

The law of total probability allows one further generalization of Bayes's rule to obtain **Bayes's theorem**. From the definition of conditional probability, we have

$$P(E_i|F) = \frac{P(E_iF)}{P(F)}$$

From Equation 1.7, we have

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{P(F)}$$

Substituting Equation 1.10, we get

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{\left(\sum_{i=1}^{n} P(F|E_i)P(E_i)\right)} \tag{EQ 1.11}$$

This is called the **generalized Bayes's rule**, or Bayes's theorem. It allows us to compute the probability of any one of the priors E_i, conditional on the occurrence of the posterior F. This is often interpreted as follows: We have some set of mutually exclusive and exhaustive hypotheses E_i. We conduct an experiment, whose outcome is F. We can then use Bayes's formula to compute the revised estimate for each hypothesis.

EXAMPLE 1.17: BAYES'S THEOREM

Continuing with Example 1.15, consider the following situation: We pick a packet at random from the set of sampled packets and find that its length is *not* 52 bytes. What is the probability that it is a UDP packet?

Solution:

As in Example 1.6, let *UDP* refer to the event that a packet is of type UDP and *52* refer to the event that the packet is of length 52 bytes. Denote the complement of the latter event, that is, that the packet is not of length 52 bytes by 52^c.

From Bayes's rule:

$$\begin{aligned} P(\text{UDP}|52^c) &= \frac{P(52^c|\text{UDP})P(\text{UDP})}{P(52^c|\text{UDP})P(\text{UDP}) + P(52^c|\text{TCP})P(\text{TCP}) + P(52^c|other)P(other)} \\ &= \frac{0.8(0.45)}{0.8(0.45) + 0.1(0.5) + 1(0.05)} \\ &= 0.78 \end{aligned}$$

Thus, if we see a packet that is *not* 52 bytes long, it is quite likely a UDP packet. Intuitively, this must be true because most TCP packets are 52 bytes long, and there aren't very many non-UDP and non-TCP packets.

1.3 Random Variables

So far, we have restricted ourselves to studying events, which are collections of outcomes of experiments or observations. However, we are often interested in abstract quantities or outcomes of experiments that are derived from events and observations but are not themselves events or observations. For example, if we throw a fair die, we may want to compute the probability that the square of the face value is smaller than 10. This is random and can be associated with a probability and, moreover, depends on some underlying random events. Yet, it is neither an event nor an observation: It is a **random variable**. Intuitively, a random variable is a quantity that can assume any one of a set of values, called its **domain *D***, and whose value can be stated only probabilistically. In this section, we will study random variables and their distributions.

More formally, a **real random variable**—the one most commonly encountered in applications having to do with computer networking—is a mapping from events in a sample space S to the domain of real numbers. The probability associated with each value assumed by a real random variable[2] is the probability of the underlying event in the sample space, as illustrated in Figure 1.1.

A random variable is **discrete** if the set of values it can assume is finite and countable. The elements of D should be *mutually exclusive*—that is, the random variable cannot simultaneously take on more than one value—and *exhaustive*—the random variable cannot assume a value that is not an element of D.

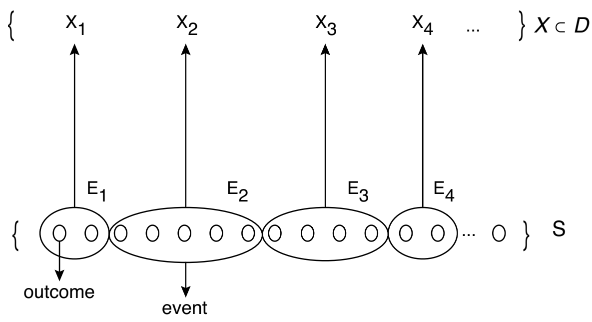

Figure 1.1 The random variable X takes on values from the domain D. Each value taken on by the random variable is associated with a probability corresponding to an event E, which is a subset of outcomes in the sample space S.

2. We deal with only real random variables in this text, so at this point will drop the qualifier "real."

EXAMPLE 1.18: A DISCRETE RANDOM VARIABLE

Consider a random variable I defined as the size of an IP packet rounded up to closest kilobyte. Then, I assumes values from the domain $D = \{1,2,3,..., 64\}$. This set is both mutually exclusive and exhaustive. The underlying sample space S is the set of potential packet sizes and is therefore identical to D. The probability associated with each value of I is the probability of seeing an IP packet of that size in some collection of IP packets, such as a measurement trace.

A random variable is **continuous** if the values it can take on are a subset of the real line.

EXAMPLE 1.19: A CONTINUOUS RANDOM VARIABLE

Consider a random variable T defined as the time between consecutive packet arrivals at a port of a switch, also called the packet interarrival time. Although each packet's arrival time is quantized by the receiver's clock, so that the set of interarrival times are finite and countable, given the high clock speeds of modern systems, modeling T as a continuous random variable is a good approximation of reality. The underlying sample space S is the subset of the real line that spans the smallest and largest possible packet interarrival times. As in the previous example, the sample space is identical to the domain of T.

1.3.1 Distribution

In many cases, we are not interested in the actual value taken by a random variable but in the probabilities associated with each such value that it can assume. To make this more precise, consider a discrete random variable X_d that assumes distinct values $D = \{x_1, x_2,..., x_n\}$. We define the value $p(x_i)$ to be the probability of the event that results in X_d assuming the value x_i. The function $p(X_d)$, which characterizes the probability that X_d will take on each value in its domain, is called the **probability mass function** of X_d.[3] It is also sometimes called the **distribution** of X_d.

3. Note the subtlety in this standard notation. Recall that $P(E)$ is the probability of an event E. In contrast, $p(X)$ refers to the distribution of a random variable X, and $p(X = x_i) = p(x_i)$ refers to the probability that random variable X takes on the value x_i.

EXAMPLE 1.20: PROBABILITY MASS FUNCTION

Consider a random variable H defined as 0 if fewer than 100 packets are received at a router's port in a particular time interval T and 1 otherwise. The sample space of outcomes consists of all possible numbers of packets that could arrive at the router's port during T, which is simply the set $S = \{1, 2, ..., M\}$, where M is the maximum number of packets that can be received in time T. Assuming that $M > 99$, we define two events $E_0 = \{0, 1, 2, ..., 99\}$ and $E_1 = \{100, 101, ..., M\}$. Given the probability of each outcome in S, we can compute the probability of each event, $P(E_0)$ and $P(E_1)$. By definition, $p(\mathrm{H} = 0) = p(0) = P(E_0)$ and $p(H = 1) = p(1) = P(E_1)$. The set $\{p(0), p(1)\}$ is the probability mass function of H. Notice how the probability mass function is closely tied to events in the underlying sample space.

Unlike a discrete random variable, which has nonzero probability of taking on any particular value in its domain, the probability that a continuous real random variable X_c will take on any specific value in its domain is 0. Nevertheless, in nearly all cases of interest in the field of computer networking, we will be able to assume that we can define the **density** function $f(x)$ of X_c as follows: The probability that X_c takes on a value between two reals, x_1 and x_2, $p(x_1 \le x \le x_2)$, is given by the integral $\int_{x_1}^{x_2} f(x)dx$. Of course, we need to ensure that $\int_{-\infty}^{\infty} f(x)dx = 1$. Alternatively, we can think of $f(x)$ being implicitly defined by the statement that a variable x chosen randomly in the domain of X_c has probability $f(a)\Delta$ of lying in the range $\left[a - \frac{\Delta}{2}, a + \frac{\Delta}{2}\right]$ when Δ is very small.

EXAMPLE 1.21: DENSITY FUNCTION

Suppose that we know that packet interarrival times are distributed *uniformly* in the range [0.5s, 2.5s]. The corresponding density function is a constant c over the domain. It is easy to see that $c = 0.5$ because we require $\int_{-\infty}^{\infty} f(x)dx = \int_{0.5}^{2.5} c\,dx = 2c = 1$. The probability that the interarrival time is in the interval $\left[1 - \frac{\Delta}{2}, 1 + \frac{\Delta}{2}\right]$ is therefore 0.5Δ.

1.3.2 Cumulative Density Function

The domain of a discrete real random variable X_d is totally ordered; that is, for any two values x_1 and x_2 in the domain, either $x_1 > x_2$ or $x_2 > x_1$. We define the **cumulative density function** $F(X_d)$ by

$$F(x) = \sum_{i | x_i \leq x} p(x_i) = p(X_d \leq x) \qquad \text{(EQ 1.12)}$$

Note the difference between $F(X_d)$, which denotes the cumulative distribution of random variable X_d, and $F(x)$, which is the value of the cumulative distribution for the value $X_d = x$.

Similarly, the cumulative density function of a continuous random variable X_c, denoted $F(X_c)$, is given by

$$F(x) = \int_{-\infty}^{x} f(y)dy = p(X_c \leq x) \qquad \text{(EQ 1.13)}$$

By definition of probability, in both cases, $0 \leq F(X_d) \leq 1$, $0 \leq F(X_c) \leq 1$.

EXAMPLE 1.22: CUMULATIVE DENSITY FUNCTIONS

Consider a discrete random variable D that can take on values {1, 2, 3, 4, 5} with probabilities {0.2, 0.1, 0.2, 0.2, 0.3}, respectively. The latter set is also the probability mass function of D. Because the domain of D is totally ordered, we compute the cumulative density function $F(D)$ as $F(1) = 0.2$, $F(2) = 0.3$, $F(3) = 0.5$, $F(4) = 0.7$, $F(5) = 1.0$.

Now, consider a continuous random variable C defined by the density function $f(x) = 1$ in the range [0,1]. The cumulative density function $F(C) = \int_{-\infty}^{x} f(y)dy = \int_{-\infty}^{x} dy = y|_0^x = x$. We see that, although, for example, $f(0.1) = 1$, this does not mean that the value 0.1 is certain!

Note that, by definition of cumulative density function, it is necessary that it achieve a value of 1 at right extreme value of the domain.

1.3.3 Generating Values from an Arbitrary Distribution

The cumulative density function $F(X)$, where X is either discrete or continuous, can be used to generate values drawn from the underlying discrete or continuous distribution $p(X_d)$ or $f(X_c)$, as illustrated in Figure 1.2. Consider a discrete random

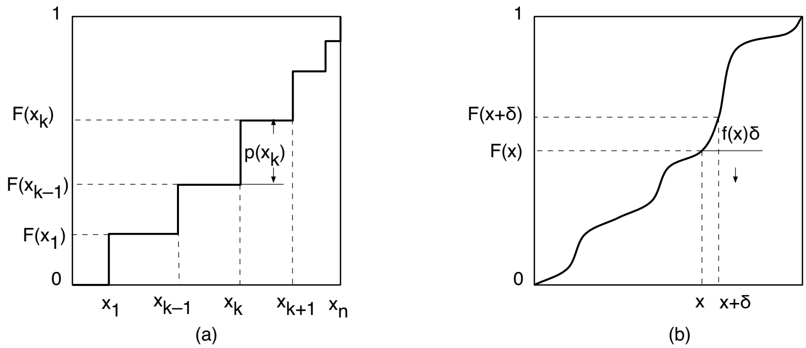

Figure 1.2 Generating values from an arbitrary (a) discrete or (b) continuous distribution

variable X_d that takes on values $x_1, x_2, \ldots, x_n$ with probabilities $p(x_i)$. By definition, $F(x_k) = F(x_{k-1}) + p(x_k)$. Moreover, $F(X_d)$ always lies in the range [0,1]. Therefore, if we were to generate a random number u with uniform probability in the range [0,1], the probability that u lies in the range $[F(x_{k-1}), F(x_k)]$ is $p(x_k)$. Moreover, $x_k = F^{-1}(u)$. Therefore, the procedure to generate values from the discrete distribution $p(X_d)$ is as follows: First, generate a random variable u uniformly in the range [0,1]; second, compute $x_k = F^{-1}(u)$.

We can use a similar approach to generate values from a continuous random variable X_c with associated density function $f(X_c)$. By definition, $F(x + \delta) = F(x) + f(x)\delta$ for very small values of δ. Moreover, $F(X_c)$ always lies in the range [0,1]. Therefore, if we were to generate a random number u with uniform probability in the range [0,1], the probability that u lies in the range $[F(x), F(x+\delta)]$ is $f(x)\delta$, which means that $x = F^{-1}(u)$ is distributed according to the desired density function $f(X_c)$. Therefore, the procedure to generate values from the continuous distribution $f(X_c)$ is as follows: First, generate a random variable u uniformly in the range [0,1]; second, compute $x = F^{-1}(u)$.

1.3.4 Expectation of a Random Variable

The **expectation**, **mean**, or **expected value** $E[X_d]$ of a discrete random variable X_d that can take on n values x_i with probability $p(x_i)$ is given by

$$E[X_d] = \sum_{i=1}^{n} x_i p(x_i) \qquad \textbf{(EQ 1.14)}$$

Similarly, the expectation $E[X_c]$ of a continuous random variable X_c with density function $f(x)$ is given by

$$E[X_c] = \int_{-\infty}^{\infty} xf(x)dx \quad \textbf{(EQ 1.15)}$$

Intuitively, the expected value of a random variable is the value we expect it to take, knowing nothing else about it. For instance, if you knew the distribution of the random variable corresponding to the time it takes for you to travel from your home to work, you expect your commute time on a typical day to be the expected value of this random variable.

EXAMPLE 1.23: EXPECTATION OF A DISCRETE AND A CONTINUOUS RANDOM VARIABLE

Continuing with the random variables C and D defined in Example 1.22, we find

$E[D] = 1*0.2 + 2*0.1 + 3*0.2 + 4*0.2 + 5*0.3 = 0.2 + 0.2 + 0.6 + 0.8 + 1.5 = 3.3.$

Note that the expected value of D is in fact a value it cannot assume! This is often true of discrete random variables. One way to interpret this is that D will take on values close to its expected value: in this case, 3 or 4.

Similarly,

$$E[C] = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{1} xdx = \left.\frac{x^2}{2}\right|_0^1 = \frac{1}{2}$$

C is the *uniform* distribution, and its expected value is the midpoint of the domain: 0.5.

The expectation of a random variable gives us a reasonable idea of how it behaves in the long run. It is important to remember, however, that two random variables with the same expectation can have rather different behaviors.

We now state, without proof, four useful properties of expectations.

1. For constants a and b:

$$E[aX + b] = aE[X] + b \quad \textbf{(EQ 1.16)}$$

2. $E[X+Y] = E[X] + E[Y]$, or, more generally, for any set of random variables X_i:

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] \quad \textbf{(EQ 1.17)}$$

3. For a discrete random variable X_d with probability mass function $p(x_i)$ and any function $g(.)$:

$$E[g(X_d)] = \sum_i g(x_i)p(x_i) \quad \text{(EQ 1.18)}$$

4. For a continuous random variable X_c with density function $f(x)$ and any function $g(.)$:

$$E[g(X_c)] = \int_{-\infty}^{\infty} g(x)f(x)dx \quad \text{(EQ 1.19)}$$

Note that, in general, $E[g(X)]$ is not the same as $g(E[X])$; that is, a function cannot be taken out of the expectation.

EXAMPLE 1.24: EXPECTED VALUE OF A FUNCTION OF A DISCRETE RANDOM VARIABLE

Consider a discrete random variable D that can take on values {1, 2, 3, 4, 5} with probabilities {0.2, 0.1, 0.2, 0.2, 0.3}, respectively. Then, $E[e^D] = 0.2e^1 + 0.1e^2 + 0.2e^3 + 0.2e^4 + 0.3e^5 = 60.74$.

EXAMPLE 1.25: EXPECTED VALUE OF A FUNCTION OF A CONTINUOUS RANDOM VARIABLE

Let X be a random variable that has equal probability of lying anywhere in the interval [0,1]. Then, $f(x) = 1; 0 \le x \le 1$. $E[X^2] = \int_0^1 x^2 f(x)dx = \frac{1}{3}x^3\Big|_0^1 = \frac{1}{3}$.

1.3.5 Variance of a Random Variable

The **variance** of a random variable is defined by $V(X) = E[(X - E[X])^2]$. Intuitively, it shows how far away the values taken on by a random variable would be from its expected value. We can express the variance of a random variable in terms of two expectations as $V(X) = E[X^2] - E[X]^2$. For

$$\begin{aligned}
V[X] &= E[(X - E[X])^2] \\
&= E[X^2 - 2XE[X] + E[X]^2] \\
&= E[X^2] - 2E[XE[X]] + E[X]^2 \\
&= E[X^2] - 2E[X]E[X] + E[X]^2 \\
&= E[X^2] - E[X]^2
\end{aligned}$$

In practical terms, the distribution of a random variable over its domain D—this domain is also called the **population**—is not usually known. Instead, the best we can do is to sample the values it takes on by observing its behavior over some period of time. We can estimate the variance of the random variable by keeping running counters for $\sum x_i$ and $\sum x_i^2$. Then,

$$V[X] \approx \left(\frac{\sum x_i^2 - (\sum x_i)^2}{n}\right),$$

where this approximation improves with n, the size of the sample, as a consequence of the law of large numbers, discussed in Section 1.7.4.

The following properties of the variance of a random variable can be easily shown for both discrete and continuous random variables.

1. For constant a:

$$V[X + a] = V[X] \qquad \textbf{(EQ 1.20)}$$

2. For constant a:

$$V[aX] = a^2 V[X] \qquad \textbf{(EQ 1.21)}$$

3. If X and Y are independent random variables:

$$V[X + Y] = V[X] + V[Y] \qquad \textbf{(EQ 1.22)}$$

1.4 Moments and Moment Generating Functions

Thus far, we have focused on elementary concepts of probability. To get to the next level of understanding, it is necessary to dive into the somewhat complex topic of moment generating functions. The *moments* of a distribution generalize its mean and variance. In this section, we will see how we can use a moment generating function (MGF) to compactly represent *all* the moments of a distribution. The moment generating function is interesting not only because it allows us to prove some useful results, such as the central limit theorem but also because it is similar in form to the Fourier and Laplace transforms, discussed in Chapter 5.

1.4.1 Moments

The **moments** of a distribution are a set of parameters that summarize it. Given a random variable X, its first **moment about the origin**, denoted M_0^1, is defined to be $E[X]$. Its **second moment about the origin**, denoted M_0^2, is defined as the expected value of the random variable X^2, or $E[X^2]$. In general, the rth moment of X about the *origin*, denoted M_0^r, is defined as $M_0^r = E[X^r]$.

We can similarly define the ***r*th moment about the *mean***, denoted M_μ^r, by $E[(X-\mu)^r]$. Note that the **variance** of the distribution, denoted by σ^2, or $V[X]$, is the same as M_μ^2. The third moment about the mean, M_μ^3, is used to construct a measure of **skewness**, which describes whether the probability mass is more to the left or the right of the mean, compared to a normal distribution. The fourth moment about the mean, M_μ^4, is used to construct a measure of peakedness, or **kurtosis**, which measures the "width" of a distribution.

The two definitions of a moment are related. For example, we have already seen that the variance of X, denoted $V[X]$, can be computed as $V[X] = E[X^2] - (E[X])^2$. Therefore, $M_\mu^2 = M_0^2 - (M_0^1)^2$. Similar relationships can be found between the higher moments by writing out the terms of the binomial expansion of $(X-\mu)^r$.

1.4.2 Moment Generating Functions

Except under some pathological conditions, a distribution can be thought to be uniquely represented by its moments. That is, if two distributions have the same moments, they will be identical except under some rather unusual circumstances. Therefore, it is convenient to have an expression, or "fingerprint," that compactly represents all the moments of a distribution. Such an expression should have terms corresponding to M_0^r for all values of r.

We can get a hint regarding a suitable representation from the expansion of e^x:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \qquad \textbf{(EQ 1.23)}$$

We see that there is one term for each power of x. This suggests the definition of the **moment generating function** of a random variable X as the expected value of e^{tX}, where t is an auxiliary variable:

$$M(t) = E[e^{tX}]. \qquad \textbf{(EQ 1.24)}$$

To see how this represents the moments of a distribution, we expand $M(t)$ as

$$\begin{aligned} M(t) &= E[e^{tX}] = E\left[1 + (tX) + \left(\frac{t^2X^2}{2!}\right) + \left(\frac{t^3X^3}{3!}\right) + \ldots\right] \\ &= 1 + E[tX] + E\left[\frac{t^2X^2}{2!}\right] + E\left[\frac{t^3X^3}{3!}\right] + \ldots \\ &= 1 + tE[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \ldots \\ &= 1 + tM_0^1 + \frac{t^2}{2!}M_0^2 + \frac{t^3}{3!}M_0^3 + \ldots \end{aligned} \qquad \textbf{(EQ 1.25)}$$

Thus, the MGF represents all the moments of the random variable X in a single compact expression. Note that the MGF of a distribution is undefined if one or more of its moments are infinite.

We can extract all the moments of the distribution from the MGF as follows: If we differentiate $M(t)$ once, the only term that is not multiplied by t or a power of t is M_0^1. So, $\left.\frac{dM(t)}{dt}\right|_{t=0} = M_0^1$. Similarly, $\left.\frac{d^2M(t)}{dt^2}\right|_{t=0} = M_0^2$. Generalizing, it is easy to show that to get the rth moment of a random variable X about the origin, we need to differentiate only its MGF r times with respect to t and then set t to 0.

It is important to remember that the "true" form of the MGF is the series expansion in Equation 1.25. The exponential is merely a convenient representation that has the property that operations on the series (as a whole) result in corresponding operations being carried out in the compact form. For example, it can be shown that the series resulting from the product of

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \quad \text{and} \quad e^y = 1 + y + \frac{y^2}{2!} + \frac{y^3}{3!} + \ldots \quad \text{is}$$

$$1 + (x+y) + \frac{(x+y)^2}{2!} + \frac{(x+y)^3}{3!} + \ldots = e^{x+y}.$$

This simplifies the computation of operations on the series. However, it is sometimes necessary to revert to the series representation for certain operations. In particular, if the compact notation of $M(t)$ is not differentiable at $t = 0$, we must revert to the series to evaluate $M(0)$, as shown next.

EXAMPLE 1.26: MGF OF A STANDARD UNIFORM DISTRIBUTION

Let X be a uniform random variable defined in the interval [0,1]. This is also called a **standard uniform distribution**. We would like to find all its moments. We find that $M(t) = E[e^{tX}] = \int_0^1 e^{tx}dx = \left.\frac{1}{t}e^{tx}\right|_0^1 = \frac{1}{t}[e^t - 1]$. However, this function is not defined—and therefore not differentiable—at $t = 0$. Instead, we revert to the series:

$$\frac{1}{t}[e^t - 1] = \frac{1}{t}\left[t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right] = 1 + \frac{t}{2!} + \frac{t^2}{3!} + \ldots$$

which *is* differentiable term by term. Differentiating r times and setting t to 0, we find that $M_0^r = 1/(r+1)$. So, $M_0^1 = \mu = 1/(1+1) = 1/2$ is the mean, and $M_0^2 = 1/(1+2) = 1/3 = E[X^2]$. Note that we found the expression for $M(t)$ by using the compact notation, but reverted to the series for differentiating it. The justification is that the integral of the compact form is identical to the summation of the integrals of the individual terms.

1.4.3 Properties of Moment Generating Functions

We now prove two useful properties of MGFs.

First, if X and Y are two independent random variables, the MGF of their sum is the product of their MGFs. If their individual MGFs are $M_1(t)$ and $M_2(t)$, respectively, the MGF of their sum is

$$M(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] \text{ (from independence)}$$
$$= M_1(t).M_2(t) \qquad \textbf{(EQ 1.26)}$$

EXAMPLE 1.27: MGF OF THE SUM

Find the MGF of the sum of two independent [0,1] uniform random variables.

Solution:

From Example 1.26, the MGF of a standard uniform random variable is $\frac{1}{t}[e^t - 1]$, so the MGF of random variable X defined as the sum of two independent uniform variables is $\frac{1}{t^2}[e^t - 1]^2$.

Second, if random variable X has MGF $M(t)$, the MGF of random variable $Y = a+bX$ is $e^{at}M(bt)$ because

$$E[e^{tY}] = E[e^{t(a+bX)}] = E[e^{at}e^{bXt}] = e^{at}E[e^{btX}] = e^{at}M(bt) \qquad \textbf{(EQ 1.27)}$$

As a corollary, if $M(t)$ is the MGF of a random variable X, the MGF of $(X - \mu)$ is given by $e^{-\mu t}M(t)$. The moments about the origin of $(X - \mu)$ are the moments about the mean of X. So, to compute the rth moment about the mean for a random variable X, we can differentiate $e^{-\mu t}M(t)$ r times with respect to t and set t to 0.

EXAMPLE 1.28: VARIANCE OF A STANDARD UNIFORM RANDOM VARIABLE

The MGF of a standard uniform random variable X is $\frac{1}{t}[e^t - 1]$. So, the MGF of $(X - \mu)$ is given by $\frac{e^{-\mu t}}{t}[e^t - 1]$. To find the variance of a standard uniform random variable, we need to differentiate twice with respect to t and then set t

to 0. Given the t in the denominator, it is convenient to rewrite the expression as $\left(1 - \mu t + \frac{\mu^2 t^2}{2!} - \ldots\right)\left(1 + \frac{t}{2!} + \frac{t^2}{3!} + \ldots\right)$, where the ellipses refer to terms with third and higher powers of t, which will reduce to 0 when t is set to 0. In this product, we need consider only the coefficient of t^2, which is $\frac{1}{3!} - \frac{\mu}{2!} + \frac{\mu^2}{2!}$. Differentiating the expression twice results in multiplying the coefficient by 2, and when we set t to zero, we obtain $E[(X - \mu)^2] = V[X] = 1/12$.

These two properties allow us to compute the MGF of a complex random variable that can be decomposed into the linear combination of simpler variables. In particular, it allows us to compute the MGF of independent, identically distributed (i.i.d.) random variables, a situation that arises frequently in practice.

1.5 Standard Discrete Distributions

We now present some discrete distributions that frequently arise when studying networking problems.

1.5.1 Bernoulli Distribution

A discrete random variable X is called a **Bernoulli** random variable if it can take only two values, 0 or 1, and its probability mass function is defined as $p(0) = 1 - a$ and $p(1) = a$. We can think of X as representing the result of some experiment, with X=1 being success, with probability a. The expected value of a Bernoulli random variable is a and variance is $p(1 - a)$.

1.5.2 Binomial Distribution

Consider a series of n Bernoulli experiments where the result of each experiment is *independent* of the others. We would naturally like to know the number of successes in these n trials. This can be represented by a discrete random variable X with parameters (n,a) and is called a **binomial** random variable. The probability mass function of a binomial random variable with parameters (n,a) is given by

$$p(i) = \binom{n}{i} a^i (1-a)^{n-i} \qquad \textbf{(EQ 1.28)}$$

If we set $b = 1 - a$, then these are just the terms of the expansion $(a+b)^n$. The expected value of a variable that is binomially distributed with parameters (n,a) is na.

EXAMPLE 1.29: BINOMIAL RANDOM VARIABLE

Consider a local area network with ten stations. Assume that, at a given moment, each node can be active with probability $p = 0.1$. What is the probability that (a) one station is active, (b) five stations are active, (c) all ten stations are active?

Solution:

Assuming that the stations are independent, the number of active stations can be modeled by a binomial distribution with parameters (10, 0.1). From the formula for $p(i)$, we get

a. $p(1) = \binom{10}{1} 0.1^1 0.9^9 = 0.38$

b. $p(5) = \binom{10}{5} 0.1^5 0.9^5 = 1.49 \times 10^{-3}$

c. $p(10) = \binom{10}{10} 0.1^{10} 0.9^0 = 1 \times 10^{-10}$

This is shown in Figure 1.3. Note how the probability of one station being active is 0.38, which is *greater* than the probability of any single station being active. Note also how rapidly the probability of multiple active stations drops. This is what drives **spatial statistical multiplexing**: the provisioning of a link with a capacity smaller than the sum of the demands of the stations.

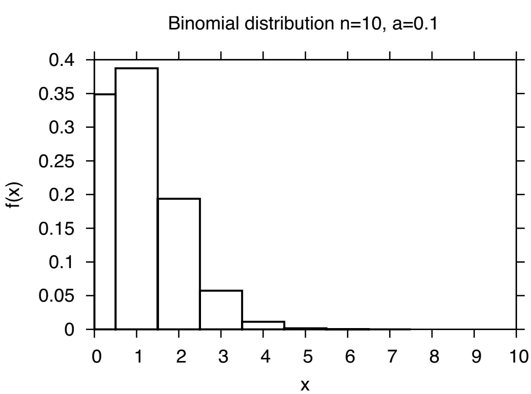

Figure 1.3 Example binomial distribution

1.5.3 Geometric Distribution

Consider a sequence of independent Bernoulli experiments, each of which succeeds with probability a. In section 1.5.2, we wanted to count the number of successes; now, we want to compute the probability mass function of a random variable X that represents the number of trials before the first success. Such a variable is called a **geometric** random variable and has a probability mass function

$$p(i) = (1-a)^{i-1}a \tag{EQ 1.29}$$

The expected value of a geometrically distributed variable with parameter a is $1/a$.

EXAMPLE 1.30: GEOMETRIC RANDOM VARIABLE

Assume that a link has a loss probability of 10% and that *packet losses are independent*, although this is rarely true in practice. Suppose that when a packet gets lost, this is detected and the packet is retransmitted until it is correctly received. What is the probability that it would be transmitted exactly one, two, and three times?

Solution:

Assuming that the packet transmissions are independent events, we note that the probability of success = p = 0.9. Therefore, $p(1) = 0.1^0 * 0.9 = 0.9$; $p(2) = 0.1^1 * 0.9 = 0.09$; $p(3) = 0.1^2 * 0.9 = 0.009$. Note the rapid decrease in the probability of more than two transmissions, even with a fairly high packet loss rate of 10%. Indeed, the expected number of transmissions is only $1/0.9 = 1.\bar{1}$.

1.5.4 Poisson Distribution

The **Poisson** distribution is widely encountered in networking situations, usually to model the arrival of packets or new end-to-end connections to a switch or a router. A discrete random variable X with the domain {0, 1, 2, 3,...} is said to be a Poisson random variable with parameter λ if, for some $\lambda > 0$:

$$P(X = i) = e^{-\lambda}\left(\frac{\lambda^i}{i!}\right) \tag{EQ 1.30}$$

Poisson variables are often used to model the number of events that happen in a fixed time interval. If the events are reasonably rare, the probability that multiple events occur in a fixed time interval drops off rapidly, due to the $i!$ term in the denominator. The first use of Poisson variables, indeed, was to investigate the number of soldier deaths due to being kicked by a horse in Napoleon's army!

The Poisson distribution, which has only a single parameter λ, can be used to model a binomial distribution with two parameters (n and a) when n is "large" and a is "small." In this case, the Poisson variable's parameter λ corresponds to the product of the two binomial parameters (i.e., $\lambda = n_{Binomial}{}^{*}\ a_{Binomial}$). Recall that a binomial distribution arises naturally when we conduct independent trials. The Poisson distribution, therefore, arises when the number of such independent trials is large, and the probability of success of each trial is small. The expected value of a Poisson distributed random variable with parameter λ is also λ.

Consider an endpoint sending a packet on a link. We can model the transmission of a packet by the endpoint in a given time interval as a trial as follows: If the source sends a packet in a particular interval, we will call the trial a success; if the source does not send a packet, we will call the trial a failure. When the load generated by each source is light, the probability of success of a trial defined in this manner, which is just the packet transmission probability, is small. Therefore, as the number of endpoints grows, and if we can assume the endpoints to be independent, the sum of their loads will be well modeled by a Poisson random variable. This is heartening because systems subjected to a Poisson load are mathematically tractable, as we will see in our discussion of queueing theory. Unfortunately, over the last two decades, numerous measurements have shown that actual traffic can be far from Poisson. Therefore, this modeling assumption should be used with care and only as a rough approximation to reality.

EXAMPLE 1.31: POISSON RANDOM VARIABLE

Consider a link that can receive traffic from one of 1,000 independent endpoints. Suppose that each node transmits at a uniform rate of 0.001 packets/second. What is the probability that we see at least one packet on the link during an arbitrary 1-second interval?

Solution:

Given that each node transmits packets at the rate of 0.001 packets/second, the probability that a node transmits a packet in any 1-second interval is $p_{Binomial} = 0.001$. Thus, the Poisson parameter $\lambda = 1000{*}0.001 = 1$. The probability that we see at least one packet on the link during any 1-second interval is therefore

$$\begin{aligned} 1 - p(0) & \\ &= 1 - e^{-1}1^{0}/0! \\ &= 1 - 1/e \\ &= 0.64 \end{aligned}$$

That is, there is a 64% chance that, during an arbitrary 1-second interval, we will see one or more packets on the link.

It turns out that a Poisson random variable is a good approximation to a binomial random variable even if the trials are weakly dependent. Indeed, we do not even require the trials to have equal probabilities, as long as the probability of success of each individual trial is "small." This is another reason why the Poisson random variable is frequently used to model the behavior of aggregates.

1.6 Standard Continuous Distributions

This section presents some standard continuous distributions. Recall from Section 1.3 that, unlike discrete random variables, the domain of a continuous random variable is a subset of the real line.

1.6.1 Uniform Distribution

A random variable X is said to be uniformly randomly distributed in the domain $[a,b]$ if its density function $f(x) = 1/(b - a)$ when x lies in $[a,b]$ and is 0 otherwise. The expected value of a uniform random variable with parameters a,b is $(a+b)/2$.

1.6.2 Gaussian, or Normal, Distribution

A random variable is **Gaussian**, or **normally** distributed, with parameters μ and σ^2 if its density is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \textbf{(EQ 1.31)}$$

We denote a Gaussian random variable X with parameters μ and σ^2 as $X \sim N(\mu,\sigma^2)$, where we read the "~" as "is distributed as."

The Gaussian distribution can be obtained as the limiting case of the binomial distribution as n tends to infinity and p is kept constant. That is, if we have a very large number of independent trials, such that the random variable measures the number of trials that succeed, the random variable is Gaussian. Thus, Gaussian random variables naturally occur when we want to study the statistical properties of aggregates.

The Gaussian distribution is called *normal* because many quantities, such as the heights of people, the slight variations in the size of a manufactured item, and the time taken to complete an activity approximately follow the well-known bell-shaped curve.[4]

4. With the caveat that many variables in real life are never negative, but the Gaussian distribution extends from $-\infty$ to ∞.

When performing experiments or simulations, it is often the case that the same quantity assumes different values during different trials. For instance, if five students were each measuring the pH of a reagent, it is likely that they would get five slightly different values. In such situations, it is common to assume that these quantities, which are supposed to be the same, are in fact normally distributed about some mean. Generally speaking, if you know that a quantity is supposed to have a certain standard value but you also know that there can be small variations in this value due to many small and independent random effects, it is reasonable to assume that the quantity is a Gaussian random variable with its mean centered on the expected value.

The expected value of a Gaussian random variable with parameters μ and σ^2 is μ and its variance is σ^2. In practice, it is often convenient to work with a **standard Gaussian distribution**, which has a zero mean and a variance of 1. It is possible to convert a Gaussian random variable X with parameters μ and σ^2 to a Gaussian random variable Y with parameters 0,1 by choosing $Y = (X - \mu)/\sigma$.

The Gaussian distribution is symmetric about the mean and asymptotes to 0 at $+\infty$ and $-\infty$. The σ^2 parameter controls the width of the central "bell": The larger this parameter, the wider the bell, and the lower the maximum value of the density function as shown in Figure 1.4. The probability that a Gaussian random variable X lies between $\mu - \sigma$ and $\mu + \sigma$ is approximately 68.26%; between $\mu - 2\sigma$ and $\mu + 2\sigma$ is approximately 95.44%; and between $\mu - 3\sigma$ and $\mu + 3\sigma$ is approximately 99.73%.

It is often convenient to use a Gaussian continuous random variable to approximately model a discrete random variable. For example, the number of packets arriving on a link to a router in a given fixed time interval will follow a discrete distribution. Nevertheless, by modeling it using a continuous Gaussian random variable, we can get quick estimates of its expected extremal values.

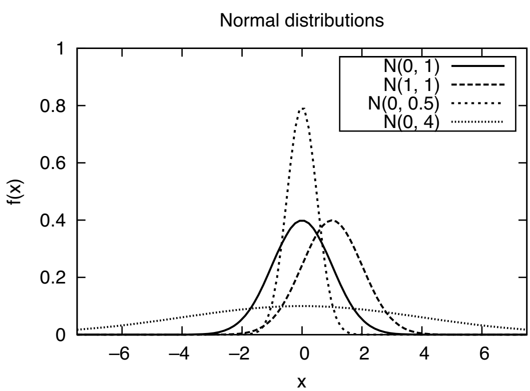

Figure 1.4 Gaussian distributions for different values of the mean and variance

EXAMPLE 1.32: GAUSSIAN APPROXIMATION OF A DISCRETE RANDOM VARIABLE

Suppose that the number of packets arriving on a link to a router in a 1-second interval can be modeled accurately by a normal distribution with parameters (20, 4). How many packets can we expect to see with at least 99% confidence?

Solution:

The number of packets are distributed (20, 4), so that $\mu = 20$ and $\sigma = 2$. We have more than 99% confidence that the number of packets seen will be $\mu \pm 3\sigma$, or between 14 and 26. That is, if we were to measure packets' arrivals over a long period of time, fewer than 1% of the 1-second intervals would have packet counts fewer than 14 or more than 26.

The MGF of the normal distribution is given by

$$\begin{aligned} M(t) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx \\ &= \frac{e^{\mu t + \frac{1}{2}\sigma^2 t^2}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{(x-\mu-\sigma^2 t)^2}{\sigma^2}} dx \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \end{aligned}$$

where in the last step, we recognize that the integral is the area under a normal curve, which evaluates to $\sigma\sqrt{2\pi}$. Note that the MGF of a normal variable with zero mean and a variance of 1 is therefore

$$M(t) = e^{\frac{1}{2}t^2} \quad \textbf{(EQ 1.32)}$$

We can use the MGF of a normal distribution to prove some elementary facts about it.

a. If $X \sim N(\mu,\sigma^2)$, then $a+bX \sim N(a+b\mu, b^2\sigma^2)$, because the MGF of $a+bX$ is

$$\begin{aligned} e^{at}M(bt) &= e^{at}e^{\mu bt + \frac{1}{2}\sigma^2 (bt)^2} \\ &= e^{(a+\mu b)t + \frac{1}{2}(\sigma^2 b^2)t^2} \end{aligned}$$

which can be seen to be a normally distributed random variable with mean $a+ b\mu$ and variance $b^2\sigma^2$.

b. If $X \sim N(\mu,\sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0,1)$. This is obtained trivially by substituting for a and b in expression (a). Z is called the **standard normal variable**.

c. If $X \sim N(\mu_1,\sigma_1^2)$ and $Y \sim N(\mu_2,\sigma_2^2)$ and X and Y are independent, $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$, because the MGF of their sum is the product of their individual MGFs $= e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2} = e^{(\mu_1+\mu_2)t + \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2}$. As a generalization, the sum of any number of independent normal variables is also normally distributed with the mean as the sum of the individual means and the variance as the sum of the individual variances.

1.6.3 Exponential Distribution

A random variable X is exponentially distributed with parameter λ, where $\lambda > 0$, if its density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{EQ 1.33}$$

Note than when $x = 0$, $f(x) = \lambda$ (see Figure 1.5). The expected value of such a random variable is $\frac{1}{\lambda}$ and its variance is $\frac{1}{\lambda^2}$. The exponential distribution is the continuous analog of the geometric distribution. Recall that the geometric distribution measures the number of trials until the first success. Correspondingly, the exponential distribu-

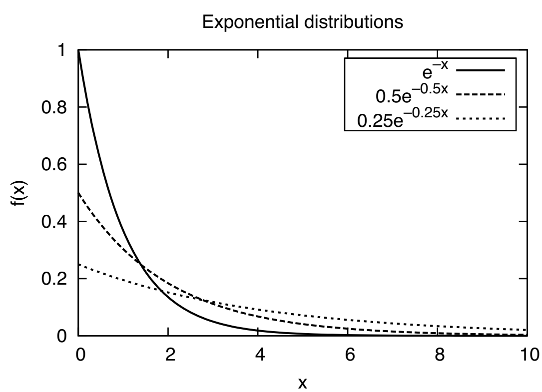

Figure 1.5 Exponentially distributed random variables with $\lambda = \{1, 0.5, 0.25\}$

tion arises when we are trying to measure the duration of time before some event happens (i.e., achieves success). For instance, it is used to model the time between two consecutive packet arrivals on a link.

The cumulative density function of the exponential distribution, $F(X)$, is given by

$$F(X) = p(X \le x) = 1 - e^{-\lambda x} \tag{EQ 1.34}$$

EXAMPLE 1.33: EXPONENTIAL RANDOM VARIABLE

Suppose that measurements show that the average length of a phone call is 3 minutes. Assuming that the length of a call is an exponential random variable, what is the probability that a call lasts more than 6 minutes?

Solution:

Clearly, the λ parameter for this distribution is 1/3. Therefore, the probability that a call lasts more than six minutes is $1 - F(6) = 1 - e^{-6/3} = 1 - e^{-2} = 13.5\%$.

An important property of the exponential distribution is that, like the geometric distribution, it is **memoryless** and, in fact, is the *only* memoryless continuous distribution. Intuitively, this means that the expected remaining time until the occurrence of an event with an exponentially distributed waiting time is *independent* of the time at which the observation is made. More precisely, $P(X > s+t \mid X>s) = P(X>t)$ for all s, t. From a geometric perspective, if we truncate the distribution to the left of any point on the positive X axis and then rescale the remaining distribution so that the area under the curve is 1, we will obtain the original distribution. The following examples illustrate this useful property.

EXAMPLE 1.34: MEMORYLESSNESS 1

Suppose that the time a bank teller takes is an exponentially distributed random variable with an expected value of 1 minute. When you arrive at the bank, the teller is already serving a customer. If you join the queue now, you can expect to wait 1 minute before being served. However, suppose that you decide to run an errand and return to the bank. If the same customer is still being served (i.e., the condition $X>s$), and if you join the queue now, the expected waiting time for you to be served would *still* be 1 minute!

EXAMPLE 1.35: MEMORYLESSNESS 2

Suppose that a switch has two parallel links to another switch and that packets can be routed on either link. Consider a packet A that arrives when both links are already in service. Therefore, the packet will be sent on the first link that becomes free. Suppose that this is link 1. Now, assuming that link service times are exponentially distributed, which packet is likely to finish transmission first: packet A on link 1 or the packet continuing service on link 2?

Solution:

Because of the memorylessness of the exponential distribution, the expected remaining service time on link 2 at the time that A starts transmission on link 1 is exactly the same as the expected service time for A, so we expect both to finish transmission at the same time. Of course, we are assuming that we don't know the service time for A. If a packet's service time is proportional to its length, and if we know A's length, we no longer have an expectation for its service time: We know it precisely, and this equality no longer holds.

1.6.4 Power-Law Distribution

A random variable described by its minimum value x_{min} and a scale parameter $\alpha > 1$ is said to obey the power-law distribution if its density function is given by

$$f(x) = \frac{(\alpha - 1)}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha} \qquad \textbf{(EQ 1.35)}$$

Typically, this function needs to be normalized for a given set of parameters to ensure that $\int_{-\infty}^{\infty} f(x)dx = 1$.

Note that $f(x)$ decreases rapidly with x. However, the decline is not as rapid as with an exponential distribution (see Figure 1.6). This is why a power-law distribution is also called a **heavy-tailed distribution**. When plotted on a log-log scale, the graph of $f(x)$ versus x shows a linear relationship with a slope of $-\alpha$, which is often used to quickly identify a potential power-law distribution in a data set.

Intuitively, if we have objects distributed according to an exponential or power law, a few "elephants" occur frequently and are common, and many "mice" are relatively uncommon. The elephants are responsible for most of the probability mass. From an engineering perspective, whenever we see such a distribution, it makes sense to build a system that deals well with the elephants, even at the expense of

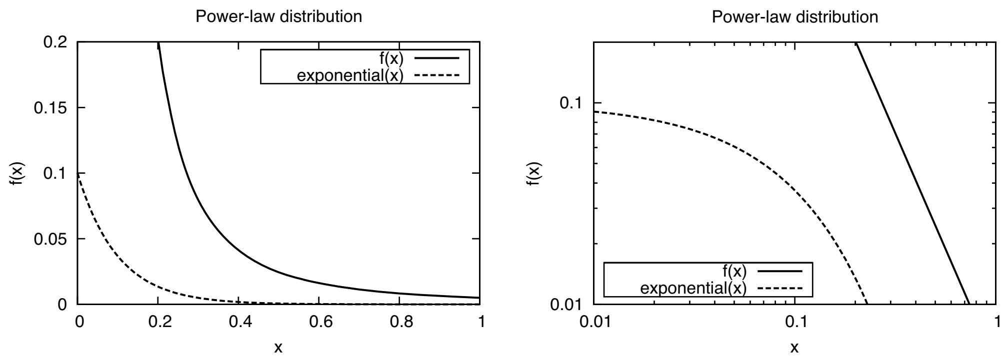

Figure 1.6 A typical power-law distribution with parameters $x_{min} = 0.1$ and $\alpha = 2.3$ compared to an exponential distribution using a linear-linear (left) and a log-log (right) scale

ignoring the mice. Two rules of thumb that reflect this are the *90/10 rule*—90% of the output is derived from 10% of the input—and the dictum *optimize for the common case*.

When $\alpha < 2$, the expected value of the random variable is infinite. A system described by such a random variable is unstable (i.e., its value is unbounded). On the other hand, when $\alpha > 2$, the tail probabilities fall rapidly enough that a power-law random variable can usually be well approximated by an exponential random variable.

A widely studied example of power-law distribution is the random variable that describes the number of users who visit one of a collection of Web sites on the Internet on any given day. Traces of Web site accesses almost always show that all but a microscopic fraction of Web sites get fewer than one visitor a day: Traffic is garnered mostly by a handful of well-known Web sites.

1.7 Useful Theorems

This section discusses some useful theorems: Markov's and Chebyshev's inequality theorems allow us to bound the amount of mass in the tail of a distribution, knowing nothing more than its expected value (Markov) and variance (Chebyshev). Chernoff's bound allows us to bound both the lower and upper tails of distributions arising from independent trials. The law of large numbers allows us to relate real-world measurements with the expectation of a random variable. Finally, the central limit theorem shows why so many real-world random variables are normally distributed.

1.7.1 Markov's Inequality

If X is a *non-negative* random variable with mean μ, then for any constant $a > 0$,

$$p(X \geq a) \leq \frac{\mu}{a} \quad \textbf{(EQ 1.36)}$$

Thus, we can bound the probability mass to the right of any constant a by a value proportional to the expected value of X and inversely proportional to a (Figure 1.7). Markov's inequality requires knowledge only of the mean of the distribution. Note that this inequality is trivial if $a < \mu$ (why?). Note also that the Markov inequality does not apply to some standard distributions, such as the normal distribution, because they are not always non-negative.

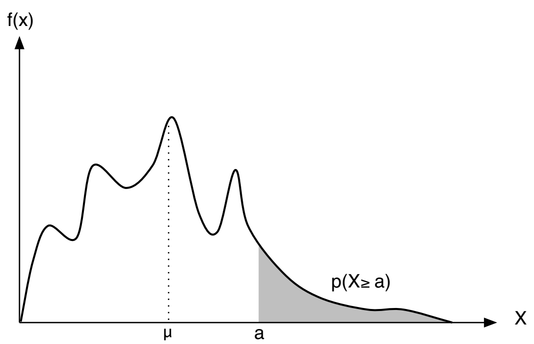

Figure 1.7 Markov's inequality

EXAMPLE 1.36: MARKOV INEQUALITY

Use the Markov inequality to bound the probability mass to the right of the value 0.75 of a uniform (0,1) distribution.

Solution:

The mean of this distribution is 0.5, so $p(X \geq 0.75) \leq \frac{0.5}{0.75} = 0.66$. The actual probability mass is only 0.25, so the Markov bound is quite loose. This is typical of a Markov bound.

1.7.2 Chebyshev's Inequality

If X is a random variable with a finite mean μ and variance σ^2, then for any constant $a > 0$,

$$p(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \qquad \textbf{(EQ 1.37)}$$

Chebyshev's inequality bounds the "tails" of a distribution on both sides of the mean, given the variance. Roughly, the farther away we get from the mean (the larger a is), the less mass there is in the tail (because the right-hand size decreases by a factor quadratic in a), as shown in Figure 1.8.

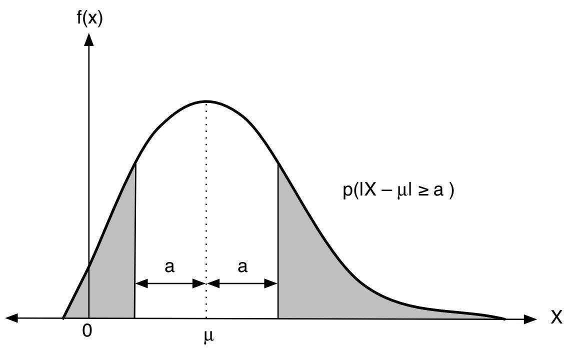

Figure 1.8 Chebyshev's inequality

EXAMPLE 1.37: CHEBYSHEV BOUND

Use the Chebyshev bound to compute the probability that a standard normal random variable has a value greater than 3.

Solution:

For a standard normal variable, $\mu = 0$ and $\sigma = 1$. We have $a = 3$. So, $p(|X| \geq 3) \leq \frac{1}{9}$, so that $p(X > 3) \leq \frac{1}{18}$, or about 5.5%. Compare this to the tight bound of 0.135% (Section 1.6.2).

1.7.3 Chernoff Bound

Let the random variable X_i denote the outcome of the ith iteration of a process, with $X_i = 1$ denoting success and $X_i = 0$ denoting failure. Assume that the probability of success of each iteration is independent of the others (this is critical!). Denote the probability of success of the ith trial by $p(X_i = 1) = p_i$. Let X be the number of successful trials in a run of n trials. Clearly,

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

Let $E[X] = \mu$ be the expected value of X (the expected number of successes). Then, we can state two Chernoff bounds that tell us the probability that there are too few or too many successes.

The **lower bound** is given by

$$p(X < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}, \qquad 0 < \delta \le 1 \tag{EQ 1.38}$$

This is somewhat hard to compute. A weaker but more tractable bound is

$$p(X < (1-\delta)\mu) < e^{\frac{-\mu\delta^2}{2}}, \qquad 0 < \delta \le 1 \tag{EQ 1.39}$$

Note that both equations bound the area under the density distribution of X between $-\infty$ and $(1-\delta)\mu$. The second form makes it clear that the probability of too few successes declines quadratically with δ.

The **upper bound** is given by

$$p(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}, \qquad \delta > 0 \tag{EQ 1.40}$$

A weaker but more tractable bound is

$$\begin{aligned} p(X > (1+\delta)\mu) &< e^{\frac{-\mu\delta^2}{4}} && \text{if } \delta < 2e-1 \\ p(X > (1+\delta)\mu) &< 2^{-\delta\mu} && \text{if } \delta > 2e-1 \end{aligned} \tag{EQ 1.41}$$

EXAMPLE 1.38: CHERNOFF BOUND

Use the Chernoff bound to compute the probability that a packet source that suffers from independent packet losses, where the probability of each loss is 0.1, suffers from more than four packet losses when transmitting ten packets.

Solution:

We define a successful event to be a packet loss, with the probability of success being $p_i = 0.1 \quad \forall i$. We have $E[X] = (10)(0.1) = 1 = \mu$. Also, we want to compute $p(X > 4) = p(X > (1+3)\mu)$ so that $\delta = 3$. So,

$$p(X > 4) < \left(\frac{e^3}{(1+3)^{1+3}}\right)^1 = \frac{e^3}{256} = 0.078$$

As with all bounds, this is looser than the exact value computed from the binomial theorem, given by

$$
\begin{aligned}
&(1-p(X=0)+p(X=1)+p(X=2)+p(X=3)+p(X=4))\\
&= 1-\binom{10}{0}(0.9)^{10}-\binom{10}{1}(0.1)(0.9)^{9}-\binom{10}{2}(0.1)^{2}(0.9)^{8}-\binom{10}{3}(0.1)^{3}(0.9)^{7}\\
&= 0.0033
\end{aligned}
$$

1.7.4 Strong Law of Large Numbers

The law of large numbers relates the **sample mean**—the average of a set of observations of a random variable—with the **population**, or **true mean**, which is its expected value. The **strong** law of large numbers, the better-known variant, states that if X_1, X_2,..., X_n are n independent, identically distributed random variables with the same expected value μ , then

$$P\left(\lim_{n \to \infty} (X_1 + X_2 + \ldots + X_n)/n = \mu\right) = 1 \qquad \textbf{(EQ 1.42)}$$

No matter how X is distributed, by computing an average over a sufficiently large number of observations, this average can be made to be as close to the true mean as we wish. This is the basis of a variety of statistical techniques for hypothesis testing, as described in Chapter 2.

We illustrate this law in Figure 1.9, which shows the average of 1,2,3,..., 500 successive values of a random variable drawn from a uniform distribution in the range

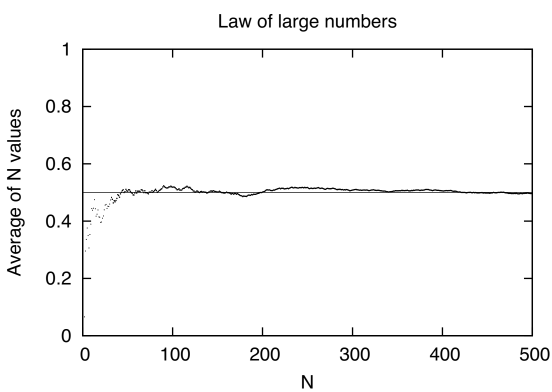

Figure 1.9 Strong law of large numbers: As N increases, the average value of sample of N random values converges to the expected value of the distribution.

[0, 1]. The expected value of this random variable is 0.5, and the average converges to this expected value as the sample size increases.

1.7.5 Central Limit Theorem

The central limit theorem deals with the sum of a *large* number of *independent* random variables that are arbitrarily distributed. The theorem states that no matter how each random variable is distributed, as long as its contribution to the total is small, the sum is well described by a Gaussian random variable.

More precisely, let $X_1, X_2,..., X_n$ be n independent, identically distributed random variables, each with a finite mean μ and variance σ^2. Then, the distribution of the normalized sum given by $\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$ tends to the standard (0,1) normal as $n \to \infty$. The central limit theorem is the reason why the Gaussian distribution is the limit of the binomial distribution.

In practice, the central limit theorem allows us to model aggregates by a Gaussian random variable if the size of the aggregate is large and the elements of the aggregate are independent.

The Gaussian distribution plays an important role in statistics because of the central limit theorem. Consider a set of measurements of a physical system. Each measurement can be modeled as an independent random variable whose mean and variance are those of the population. From the central limit theorem, their sum, and therefore their mean, which is just the normalized sum, is approximately normally distributed. As we will study in Chapter 2, this allows us to infer the population mean from the sample mean, which forms the foundation of statistical confidence. We now prove the central limit theorem by using MGFs.

The proof proceeds in three stages. First, we compute the MGF of the sum of n random variables in terms of the MGFs of each of the random variables. Second, we find a simple expression for the MGF of a random variable when the variance is large: a situation we expect when adding together many independent random variables. Finally, we plug this simple expression back into the MGF of the sum to obtain the desired result.

Consider a random variable $Y = X_1 + X_2 + \ldots + X_n$, the sum of n *independent* random variables X_i. Let μ_i and σ_i denote the mean and standard deviation of X_i, and let μ and σ denote the mean and standard deviation of Y. Because all the X_i s are independent,

$$\mu = \sum \mu_i \; ; \; \sigma^2 = \sum \sigma_i^2 \tag{EQ 1.43}$$

Define the random variable W_i to be $(X_i - \mu_i)$: It represents the distance of an instance of the random variable X_i from its mean. By definition, the rth moment of W_i about the origin is the rth moment of X_i about its mean. Also, because the X_i are independent, so are the W_i. Denote the MGF of X_i by $M_i(t)$ and the MGF of W_i by $N_i(t)$.

Note that $Y - \mu = X_1 + X_2 + \ldots + X_n - \sum \mu_i = \sum (X_i - \mu_i) = \sum W_i$. So, the MGF of $Y - \mu$ is the product of the MGFs of the $W_i = \prod_{i=1}^{n} N_i(t)$. Therefore, the MGF of $(Y - \mu)/\sigma$ denoted $N^*(t)$ is given by

$$N^*(t) = \prod_{i=1}^{n} N_i\left(\frac{t}{\sigma}\right) \quad \textbf{(EQ 1.44)}$$

Consider the MGF $N_i(t/\sigma)$, which is given by $E\left(e^{\frac{W_i t}{\sigma}}\right)$. Expanding the exponential, we find that

$$N_i\left(\frac{t}{\sigma}\right) = E\left(e^{\frac{W_i t}{\sigma}}\right) = 1 + E(W_i)\frac{t}{\sigma} + \frac{E(W_i^2)}{2!}\left(\frac{t}{\sigma}\right)^2 + \frac{E(W_i^3)}{3!}\left(\frac{t}{\sigma}\right)^3 + \ldots \quad \textbf{(EQ 1.45)}$$

Now, $E(W_i) = E(X_i - \mu_i) = E(X_i) - \mu_i = \mu_i - \mu_i = 0$, so we can ignore the second term in the expansion. Recall that σ is the standard deviation of the sum of n random variables. When n is large, so too is σ, which means that, to first order, we can ignore terms that have σ^3 and higher powers of σ in the denominator in Equation 1.45. Therefore, for large n, we can write

$$N^i\left(\frac{t}{\sigma}\right) \approx \left(1 + \frac{E(W_i^2)}{2!}\left(\frac{t}{\sigma}\right)^2\right) = 1 + \frac{\sigma_i^2}{2}\left(\frac{t}{\sigma}\right)^2 \quad \textbf{(EQ 1.46)}$$

where we have used the fact that $E(W_i^2) = E(X_i - \mu)^2$ which is the variance of $X_i = \sigma_i^2$.

Returning to the expression in Equation 1.44, we find that

$$\log N^*(t) = \log\left(\prod_{i=1}^{n} N_i\left(\frac{t}{\sigma}\right)\right) = \sum_{i=1}^{n} \log\left(N_i\left(\frac{t}{\sigma}\right)\right) \approx \sum_{i=1}^{n} \log\left(1 + \frac{\sigma_i^2}{2}\left(\frac{t}{\sigma}\right)^2\right) \quad \textbf{(EQ 1.47)}$$

It is easily shown by the Taylor series expansion that when h is small—so that h^2 and higher powers of h can be ignored—$\log(1+h)$ can be approximated by h. So, when n is large and σ is large, we can further approximate

$$\sum_{i=1}^{n} \log\left(1+\frac{\sigma_i^2}{2}\left(\frac{t}{\sigma}\right)^2\right) \approx \sum_{i=1}^{n} \frac{\sigma_i^2}{2}\left(\frac{t}{\sigma}\right)^2 = \frac{1}{2}\left(\frac{t}{\sigma}\right)^2 \sum_{i=1}^{n} \sigma_i^2 = \frac{1}{2}t^2 \qquad \textbf{(EQ 1.48)}$$

where, for the last simplification, we used Equation 1.43. Thus, log $N^*(t)$ is approximately 1/2 t^2, which means that

$$N^*(t) \approx e^{\frac{t^2}{2}} \qquad \textbf{(EQ 1.49)}$$

But this is just the MGF of a standard normal variable with 0 mean and a variance of 1 (Equation 1.32). Therefore, $(Y - \mu)/\sigma$ is a standard normal variable, which means that $Y \sim N(\mu, \sigma^2)$. We have therefore shown that the sum of a large number of independent random variables is distributed as a normal variable whose mean is the sum of the individual means and whose variance is the sum of the individual variances (Equation 1.43), as desired.

1.8 Jointly Distributed Random Variables

So far, we have considered distributions of one random variable. We now consider the distribution of two random variables simultaneously.

EXAMPLE 1.39: JOINT PROBABILITY DISTRIBUTION

Consider the two events: "rain today" and "rain tomorrow." Let the random variable X be 0 if it does not rain today and 1 if it does. Similarly, let the random variable Y be 0 if it does not rain tomorrow and 1 if it does. The four possible values for the random variables X and Y considered together are 00, 01, 10, and 11, corresponding to four joint events. We can associate probabilities with these events with the usual restrictions that these probabilities lie in [0,1] and that their sum be 1. For instance, consider the following distribution:

$$p(00) = 0.2,$$
$$p(01) = 0.4,$$
$$p(10) = 0.3,$$
$$p(11) = 0.1,$$

where the 00 is now interpreted as shorthand for $X = 0$ AND $Y = 0$, and so on. This defines the **joint probability** distribution of X and Y, which is denoted $p_{XY}(xy)$ or sometimes $p(X,Y)$. Given this joint distribution, we can extract the distribution of X alone, which is the probability of $X = 0$ and of $X = 1$, as follows: $p(X = 0) = p(00) + p(01) = 0.2 + 0.4 = 0.6$. Similarly, $p(X = 1) = 0.3 + 0.1 = 0.4$. As expected, $p(X = 0) + p(X = 1) = 1$. Similarly, note that $p(Y = 0) = 0.5$ and $p(Y = 1) = 0.5$.

We call the distribution of X alone as the **marginal** distribution of X and denote it p_X. Similarly, the marginal distribution of Y is denoted p_Y. Generalizing from the preceding example, we see that to obtain the marginal distribution of X, we should set X to each value in its domain and then sum over *all possible values of Y*. Similarly, to obtain the marginal distribution of Y, we set Y to each value in its domain and sum over all possible values of X.

An important special case of a joint distribution is when the two variables X and Y are **independent**. Then, $p_{XY}(xy) = p(X = x \text{ AND } Y = y) = p(X = x\ ) * p(Y = y) = p_X(x)p_Y(y)$. That is, each entry in the joint distribution is obtained simply as the product of the marginal distributions corresponding to that value. We sometimes denote this as $= p_X(x)p_Y(y)$.

EXAMPLE 1.40: INDEPENDENCE

In Example 1.39, $p_{XY}(00) = 0.2$, $p_X(0) = 0.6$, and $p_Y(0) = 0.5$, so X and Y are *not* independent: We *cannot* decompose the joint distribution into the product of the marginal distributions.

Given the joint distribution, we define the **conditional probability mass function of X**, denoted by $p_{X|Y}(x|y)$ by $p(X = x \mid Y = y) = p(X = x \text{ AND } Y = y)/p(Y = y) = \frac{p_{XY(xy)}}{p_Y(y)}$.

EXAMPLE 1.41: CONDITIONAL PROBABILITY MASS FUNCTION

Continuing with Example 1.39, suppose that we want to compute the probability that it will rain tomorrow, given that it rained today: $p_{Y|X}(1|1) = p_{XY}(11)/p_X(1) = 0.1/0.4 = 0.25$. Thus, knowing that it rained today makes it less probable that it will rain tomorrow because $p_{(Y=1)} = 0.5$ and $p_{(Y=1|X=1)} = 0.25$.

We can generalize the notion of joint probability in three ways. We outline these generalizations next. Note that the concepts we have developed for the simple preceding case continue to hold for these generalizations.

1. Instead of having only two values, 0 and 1, X and Y could assume any number of finite discrete values. In this case, if there are n values of X and m values of Y, we would need to specify, for the joint distribution, a total of nm values. If X and Y are independent, however, we need to specify only n+m values to completely specify the joint distribution.
2. We can generalize this further and allow X and Y to be continuous random variables. Then, the joint probability distribution $p_{XY}(xy)$ is implicitly defined by

$$p(a \le X \le a + \alpha, b \le Y \le b + \beta) = \int_{b}^{(b+\beta)} \int_{a}^{(a+\alpha)} p_{XY}(xy)dxdy \qquad \text{(EQ 1.50)}$$

   Intuitively, this is the probability that a randomly chosen two-dimensional vector will be in the vicinity of (a,b).
3. As a further generalization, consider the joint distribution of n random variables, $X_1, X_2,..., X_n$, where each variable is either discrete or continuous. If they are all discrete, we need to define the probability of each possible choice of each value of X_i. This grows exponentially with the number of random variables and with the size of each domain of each random variable. Thus, it is impractical to completely specify the joint probability distribution for a large number of variables. Instead, we exploit pairwise independence between the variables, using the construct of a Bayesian network, which is described next.

1.8.1 Bayesian Networks

Bayes's rule allows us to compute the degree to which one of a set of mutually exclusive prior events contributes to a posterior condition. Suppose that the posterior condition was itself a prior to yet another posterior, and so on. We could then imagine tracing this chain of conditional causation back from the final condition to the initial causes. This, in essence, is a Bayesian network. We will study one of the simplest forms of a Bayesian network next.

A Bayesian network with n nodes is a directed acyclic graph whose vertices represent random variables and whose edges represent conditional causation between these random variables: There is an edge from a random variable E_i, called the *parent*, or *cause*, to every random variable E_j whose outcome depends on it, called its *children*, or *effects*. If there is no edge between E_i and E_j, they are independent.

Each node in the Bayesian network stores the conditional probability distribution $p(E_j \mid \text{parents}(E_j))$, also called its **local distribution**. Note that if the node has no parents, its distribution is unconditionally known. The network allows us to compute the joint probability $p(E_1E_2...E_n)$ as

$$p(E_1E_2...E_n) = \prod_i p(E_i|parents(E_i)) \qquad \textbf{(EQ 1.51)}$$

That is, the joint distribution is simply the product of the local distributions. This greatly reduces the amount of information required to describe the joint probability distribution of the random variables. Choosing the Bayesian graph is a non-trivial problem and one that we will not discuss further. An overview can be found in the text by Russell and Norvig cited in Section 1.9.

Note that, because the Bayesian network encodes the full joint distribution, we can in principle extract any probability we want from it. Usually, we want to compute something much simpler. A Bayesian network allows us to compute probabilities of interest without having to compute the entire joint distribution, as the next example demonstrates.

EXAMPLE 1.42: BAYESIAN NETWORK

Consider the Bayesian network in Figure 1.10. Each circle shows a discrete random variable that can assume only two values: true or false. Each random variable is associated with an underlying event in the appropriate sample space, as shown in the figure. The network shows that if L, the random variable

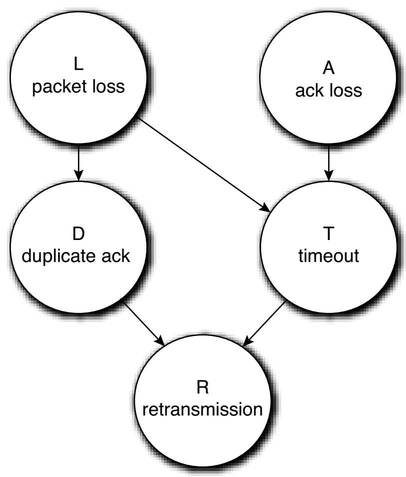

Figure 1.10 A Bayesian network to represent TCP retransmissions

representing packet loss event, has the value true (the cause), this may lead to a timeout event at the TCP transmitter (effect), so that the random variable representing this T, has a higher probability of having the value true. Similarly, the random variable denoting the loss of an acknowledgment packet may also increase the probability that T assumes the value true. The node marked T, therefore, stores the probabilty that it assumes the value true conditional on the parents, assuming the set of values {(true, true), (true, false), (false, true), (false, false)}.

The network also represents the fact that a packet loss event affects the likelihood of a duplicate acknowledgment event. However, packet and ack loss events are mutually exclusive, as are duplicate acks and timeouts. Finally, if there is either a duplicate ack or a timeout at the transmitter, it will surely retransmit a packet.

The joint distribution of the random variables (L, A, D, T, R) would assign a probability to every possible combination of the variables, such as *p*(*packet loss AND no ack loss AND no duplicate ack AND timeout AND no retransmission*). In practice, we rarely need the joint distribution. Instead, we may be interested only in computing the following probability: *p*(*packet loss* | *retransmission*) = $p(L|R)$. That is, we observe the event that the transmitter has retransmitted a packet. What is the probability that the event packet loss occurred: What is $p(L|R)$?

For notational simplicity, let $p(R = \text{true}) = p(R) = r$, $p(L = \text{true}) = p(L) = l$, $p(T = \text{true}) = p(T) = t$, $p(A = \text{true}) = p(A) = a$ and $p(D = \text{true}) = p(D) = d$. From the network, it is clear that we can write $p(R)$ as $p(R|T)t + p(R|D)d$. Similarly, $t = p(T|L)l + p(T|A)a$ and $d = p(D|L)l$. Therefore,

$$p(R) = r = p(R|T)(p(T|L)l + p(T|A)a) + p(R|D)p(D|L)l$$

If we know a and l and the conditional probabilities stored at each node, we can therefore compute r.

From the definition of conditional probabilities:

$$p(L|R) = \frac{p(LR)}{r} \qquad \textbf{(EQ 1.52)}$$

We have already seen how to compute the denominator. To compute the numerator, we sum across all possibilities for L and R as follows:

$$p(LR) = p(LRTD) + p(LRT\bar{D}) + p(LR\bar{T}D) + p(LR\bar{T}\bar{D})$$

where the overbar represents the probability that the random variable assumes the value false. However, note that T and D are mutually exclusive, so

$$p(TD) = 0$$
$$p(T\overline{D}) = p(T)$$
$$p(\overline{T}D) = p(D)$$

Thus,

$$p(LR) = p(LRT) + p(LRD) + p(LR\overline{T}\overline{D})$$

The last term is 0 because we do not have a retransmission unless there is either a timeout or a duplicate ack. Thus, $p(LR) = P(LRT) + P(LRD)$.

Replacing this in Equation 1.52, we get

$$p(PLR) = \frac{p(LRT) + p(LRD)}{p(R|T)(p(T|L)l + p(T|A)a) + p(R|D)p(D|L)l}$$

All these variables can be computed by observations over sufficiently long durations of time. For instance, to compute $p(LRT)$, we can compute the ratio of all retransmissions where there was both a packet loss and timeout event to the number of transmissions. Similarly, to compute $p(R\,|\,T)$, we can compute the ratio of the number of times a retransmission happens due to a timeout to the number of times a timeout happens. This allows us to compute $p(L\,|\,R)$ in practice.

1.9 Further Reading

A number of excellent introductory texts on probability treat this subject in more detail, such as S. Ross, *A First Course in Probability*, 7th ed., Prentice Hall, 2006. A more sophisticated treatment is the classic text by W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed., Wiley, 1968. Bayesian analysis is described in the standard textbook on artificial intelligence: S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010.

1.10 Exercises

1. **Sample space**

   In the IEEE 802.11 protocol, the congestion window (CW) parameter is used as follows: Initially, a terminal waits for a random time period, or *backoff*, chosen in the range $[1, 2^{\mathrm{CW}}]$ before sending a packet. If an acknowledgment for the packet is not received in time, CW is doubled, and the process is repeated until

CW reaches the value CWMAX. The initial value of CW is CWMIN. What are the sample spaces for the value of CW and the value of the backoff?

2. **Interpretations of probability**

   Consider the statement: Given the conditions right now, the probability of a snowstorm tomorrow morning is 25%. How would you interpret this statement from the perspective of an objective, frequentist, and subjective interpretation of probability, assuming that these are possible?

3. **Conditional probability**

   Consider a device that samples packets on a link.

   a. Suppose that measurements show that 20% of packets are UDP and that 10% of all packets are UDP packets with a packet size of 100 bytes.What is the conditional probability that a UDP packet has size 100 bytes?
   b. Suppose that 50% of packets were UDP, and 50% of UDP packets were 100 bytes long. What fraction of all packets are 100-byte UDP packets?

4. **Conditional probability again**

   Continuing with Exercise 3: How does the knowledge of the protocol type change the sample space of possible packet lengths? In other words, what is the sample space before and after you know the protocol type of a packet?

5. **Bayes's rule**

   For Exercise 3(a), what additional information do you need to compute P(UDP | 100)? Setting that value to x, express P(UDP | 100) in terms of x.

6. **Cumulative distribution function (CDF)**

   a. Suppose that *discrete* random variable D take values {1, 2, 3,...,i,...} with probability $1/2^i$. What is its CDF?
   b. Suppose continuous random variable C is uniform in the range $[x_1, x_2]$. What is its CDF?

7. **Expectations**

   Compute the expectations of the D and C in Exercise 6.

8. **Variance**

   Prove that $V[aX] = a^2V[X]$.

9. Moments

Prove that $M_\mu^3 = M_0^3 - 3M_0^2 M_0^1 + 2(M_0^1)^3$.

10. MGFs

Prove that the MGF of a uniform random variable, expressed in terms of its series expansion, is $E(e^{tx}) = \int_0^1 \left(1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \ldots\right) dx = \frac{1}{t}[e^t - 1]$.

11. MGFs

Prove that the rth moment of the uniform distribution about the origin is $1/(r+1)$.

12. MGF of a sum of two variables

Use MGFs to find the variance of the sum of two independent uniform standard random variables.

13. MGF of a normal distribution

Prove that if $X \sim N(\mu,\sigma^2)$, then $(X - \mu)/\sigma \sim N(0,1)$.

14. Bernoulli distribution

A hotel has 20 guest rooms. Assuming that outgoing calls are independent and that a guest room makes 10 minutes worth of outgoing calls during the busiest hour of the day, what is the probability that 5 calls are simultaneously active during the busiest hour? What is the probability of 15 simultaneous calls?

15. Geometric distribution

Consider a link that has a packet loss rate of 10%. Suppose that every packet transmission has to be acknowledged. Compute the expected number of data transmissions for a successful packet+ack transfer.

16. Poisson distribution

Consider a binomially distributed random variable X with parameters $n = 10$, $p = 0.1$.

a. Compute the value of $P(X = 8)$, using both the binomial distribution and the Poisson approximation.

b. Repeat for $n = 100$, $p = 0.1$.

17. Gaussian distribution

Prove that if X is Gaussian with parameters (μ, σ^2), the random variable $Y = aX + b$, where a and b are constants, is also Gaussian, with parameters $(a\mu + b, (a\sigma)^2)$.

18. Exponential distribution

Suppose that customers arrive at a bank with an exponentially distributed interarrival time with mean 5 minutes. A customer walks into the bank at 3 p.m. What is the probability that the next customer arrives no sooner than 3:15?

19. Exponential distribution

It is late August and you are watching the Perseid meteor shower. You are told that the time between meteors is exponentially distributed with a mean of 200 seconds. At 10:05 p.m., you see a meteor, after which you head to the kitchen for a bowl of ice cream, returning outside at 10:08 p.m. How long do you expect to wait to see the next meteor?

20. Power law

Consider a power-law distribution with $x_{min} = 1$ and $\alpha = 2$ and an exponential distribution with $\lambda = 2$. Fill in the following table:

x	$f_{power\_law}(x)$	$f_{exponential}(x)$
1		
5		
10		
50		
100		

It should now be obvious why a power-law distribution is called heavy-tailed!

21. Markov's inequality

Consider a random variable X that exponentially distributed with parameter $\lambda = 2$. What is the probability that $X > 10$ using (a) the exponential distribution and (b) Markov's inequality?

22. Joint probability distribution

Consider the following probability mass function defined jointly over the random variables X, Y, and Z:

$$p(000) = 0.05;\ p(001) = 0.05;\ p(010) = 0.1;\ p(011) = 0.3;$$
$$p(100) = 0.05;\ p(101) = 0.05;\ p(110) = 0.1;\ p(111) = 0.3.$$

a. Write down $p_X, p_Y, p_Z, p_{XY}, p_{XZ}, p_{YZ}$.

b. Are X and Y, X and Z, or Y and Z independent?

c. What is the probability that $X = 0$ given that $Z = 1$.

This page intentionally left blank

This page intentionally left blank

Index

E

J

K

Q

R