Introduction to Probability

SECOND EDITION

Dimitri P. Bertsekas and John N. Tsitsiklis

Massachusetts Institute of Technology

WWW site for book information and orders
http://www.athenasc.com

Athena Scientific, Belmont, Massachusetts
To the memory of

Pantelis Bertsekas and Nikos Tsitsiklis
Preface

*Probability is common sense reduced to calculation*

*Laplace*

This book is an outgrowth of our involvement in teaching an introductory probability course ("Probabilistic Systems Analysis") at the Massachusetts Institute of Technology.

The course is attended by a large number of students with diverse backgrounds, and a broad range of interests. They span the entire spectrum from freshmen to beginning graduate students, and from the engineering school to the school of management. Accordingly, we have tried to strike a balance between simplicity in exposition and sophistication in analytical reasoning. Our key aim has been to develop the ability to construct and analyze probabilistic models in a manner that combines intuitive understanding and mathematical precision.

In this spirit, some of the more mathematically rigorous analysis has been just sketched or intuitively explained in the text, so that complex proofs do not stand in the way of an otherwise simple exposition. At the same time, some of this analysis is developed (at the level of advanced calculus) in theoretical problems, that are included at the end of the corresponding chapter. Furthermore, some of the subtler mathematical issues are hinted at in footnotes addressed to the more attentive reader.

The book covers the fundamentals of probability theory (probabilistic models, discrete and continuous random variables, multiple random variables, and limit theorems), which are typically part of a first course on the subject. It also contains, in Chapters 4-6 a number of more advanced topics, from which an instructor can choose to match the goals of a particular course. In particular, in Chapter 4, we develop transforms, a more advanced view of conditioning, sums of random variables, least squares estimation, and the bivariate normal distribu-
tion. Furthermore, in Chapters 5 and 6, we provide a fairly detailed introduction to Bernoulli, Poisson, and Markov processes.

Our M.I.T. course covers all seven chapters in a single semester, with the exception of the material on the bivariate normal (Section 4.7), and on continuous-time Markov chains (Section 6.5). However, in an alternative course, the material on stochastic processes could be omitted, thereby allowing additional emphasis on foundational material, or coverage of other topics of the instructor’s choice.

Our most notable omission in coverage is an introduction to statistics. While we develop all the basic elements of Bayesian statistics, in the form of Bayes’ rule for discrete and continuous models, and least squares estimation, we do not enter the subjects of parameter estimation, or non-Bayesian hypothesis testing.

The problems that supplement the main text are divided in three categories:

(a) *Theoretical problems:* The theoretical problems (marked by *) constitute an important component of the text, and ensure that the mathematically oriented reader will find here a smooth development without major gaps. Their solutions are given in the text, but an ambitious reader may be able to solve many of them, especially in earlier chapters, before looking at the solutions.

(b) *Problems in the text:* Besides theoretical problems, the text contains several problems, of various levels of difficulty. These are representative of the problems that are usually covered in recitation and tutorial sessions at M.I.T., and are a primary mechanism through which many of our students learn the material. Our hope is that students elsewhere will attempt to solve these problems, and then refer to their solutions to calibrate and enhance their understanding of the material. The solutions are posted on the book’s www site

   http://www.athenasc.com/probbook.html

(c) *Supplementary problems:* There is a large (and growing) collection of additional problems, which is not included in the book, but is made available at the book’s www site. Many of these problems have been assigned as homework or exam problems at M.I.T., and we expect that instructors elsewhere will use them for a similar purpose. While the statements of these additional problems are publicly accessible, the solutions are made available from the authors only to course instructors.

We would like to acknowledge our debt to several people who contributed in various ways to the book. Our writing project began when we assumed responsibility for a popular probability class at M.I.T. that our colleague Al Drake had taught for several decades. We were thus fortunate to start with an organization of the subject that had stood the test of time, a lively presentation of the various topics in Al’s classic textbook, and a rich set of material that had been used in recitation sessions and for homework. We are thus indebted to Al Drake
for providing a very favorable set of initial conditions.

We are thankful to the several colleagues who have either taught from the draft of the book at various universities or have read it, and have provided us with valuable feedback. In particular, we thank Ibrahim Abou Faycal, Gustavo de Veciana, Eugene Feinberg, Bob Gray, Muriel Médard, Jason Papastavrou, Ilya Pollak, David Tse, and Terry Wagner.

The teaching assistants for the M.I.T. class have been very helpful. They pointed out corrections to various drafts, they developed problems and solutions suitable for the class, and through their direct interaction with the student body, they provided a robust mechanism for calibrating the level of the material.

Reaching thousands of bright students at M.I.T. at an early stage in their studies was a great source of satisfaction for us. We thank them for their valuable feedback and for being patient while they were taught from a textbook-in-progress.

Last but not least, we are grateful to our families for their support throughout the course of this long project.

Dimitri P. Bertsekas, dimitrib@mit.edu
John N. Tsitsiklis, jnt@mit.edu

Cambridge, Mass., May 2002
Preface to the Second Edition

This is a substantial revision of the 1st edition, involving a reorganization of old material and the addition of new material. The length of the book has increased by about 25 percent. The main changes are the following:

(a) Two new chapters on statistical inference have been added, one on Bayesian and one on classical methods. Our philosophy has been to focus on the main concepts and to facilitate understanding of the main methodologies through some key examples.

(b) Chapters 3 and 4 have been revised, in part to accommodate the new material of the inference chapters and in part to streamline the presentation. Section 4.7 of the 1st edition (bivariate normal distribution) has been omitted from the new edition, but is available at the book’s website.

(c) A number of new examples and end-of-chapter problems have been added.

The main objective of the new edition is to provide flexibility to instructors in their choice of material, and in particular to give them the option of including an introduction to statistical inference. Note that Chapters 6-7, and Chapters 8-9 are mutually independent, thus allowing for different paths through the book. Furthermore, Chapter 4 is not needed for Chapters 5-7, and only Sections 4.2-4.3 from Chapter 4 are needed for Chapters 8 and 9. Thus, some possible course offerings based on this book are:

(a) Probability and introduction to statistical inference: Chapters 1-3, Sections 4.2-4.3, Chapter 5, Chapters 8-9.

(b) Probability and introduction to stochastic processes: Chapters 1-3 and 5-7, with possibly a few sections from Chapter 4.

We would like to express our thanks to various colleagues who have contributed valuable comments on the material in the 1st edition and/or the organization of the material in the new chapters. Ed Coffman, Munther Dahleh, Vivek Goyal, Anant Sahai, David Tse, George Verghese, Alan Willsky, and John Wyatt have been very helpful in this regard. Finally, we thank Mengdi Wang for her help with figures and problems for the new chapters.

Dimitri P. Bertsekas, dimitrib@mit.edu
John N. Tsitsiklis, jnt@mit.edu
Cambridge, Mass., June 2008
# Contents

1. Sample Space and Probability ......................................................... p. 1
   1.1. Sets ...................................................................................... p. 3
   1.2. Probabilistic Models .............................................................. p. 6
   1.3. Conditional Probability .......................................................... p. 18
   1.4. Total Probability Theorem and Bayes’ Rule ................................. p. 28
   1.5. Independence ........................................................................ p. 34
   1.6. Counting ............................................................................... p. 44
   1.7. Summary and Discussion ......................................................... p. 51
       Problems ................................................................................. p. 53

2. Discrete Random Variables .............................................................. p. 71
   2.1. Basic Concepts ...................................................................... p. 72
   2.2. Probability Mass Functions ................................................... p. 74
   2.3. Functions of Random Variables .............................................. p. 80
   2.4. Expectation, Mean, and Variance .......................................... p. 81
   2.5. Joint PMFs of Multiple Random Variables ............................... p. 92
   2.6. Conditioning ......................................................................... p. 97
   2.7. Independence ........................................................................ p. 109
   2.8. Summary and Discussion ....................................................... p. 115
       Problems ................................................................................. p. 119

3. General Random Variables .............................................................. p. 139
   3.1. Continuous Random Variables and PDFs ................................. p. 140
   3.2. Cumulative Distribution Functions ........................................ p. 148
   3.3. Normal Random Variables ..................................................... p. 153
   3.4. Joint PDFs of Multiple Random Variables ............................... p. 158
   3.5. Conditioning ......................................................................... p. 164
   3.6. The Continuous Bayes’ Rule .................................................... p. 178
3.7. Summary and Discussion ........................................ p. 182
Problems ......................................................... p. 184

4. Further Topics on Random Variables .......................... p. 201
4.1. Derived Distributions ........................................ p. 202
4.2. Covariance and Correlation ................................ p. 217
4.3. Conditional Expectation and Variance Revisited ....... p. 222
4.4. Transforms .................................................. p. 229
4.5. Sum of a Random Number of Independent Random Variables ................................. p. 240
4.6. Summary and Discussion .................................... p. 244
Problems ......................................................... p. 246

5. Limit Theorems .................................................. p. 263
5.1. Markov and Chebyshev Inequalities ....................... p. 265
5.2. The Weak Law of Large Numbers ........................ p. 269
5.3. Convergence in Probability ................................ p. 271
5.4. The Central Limit Theorem ................................ p. 273
5.5. The Strong Law of Large Numbers ....................... p. 280
5.6. Summary and Discussion .................................... p. 282
Problems ......................................................... p. 284

6. The Bernoulli and Poisson Processes ......................... p. 295
6.1. The Bernoulli Process ....................................... p. 297
6.2. The Poisson Process ........................................ p. 309
6.3. Summary and Discussion .................................... p. 324
Problems ......................................................... p. 326

7.1. Discrete-Time Markov Chains .............................. p. 340
7.2. Classification of States .................................... p. 346
7.3. Steady-State Behavior ...................................... p. 352
7.4. Absorption Probabilities and Expected Time to Absorption .................. p. 362
7.5. Continuous-Time Markov Chains ......................... p. 369
7.6. Summary and Discussion .................................... p. 378
Problems ......................................................... p. 380

8. Bayesian Statistical Inference ................................ p. 407
8.1. Bayesian Inference and the Posterior Distribution .... p. 412
8.2. Point Estimation, Hypothesis Testing, and the MAP Rule ................................. p. 420
8.3. Bayesian Least Mean Squares Estimation ................ p. 430
8.4. Bayesian Linear Least Mean Squares Estimation ................................. p. 437
8.5. Summary and Discussion ........................................ p. 444
      Problems ....................................................... p. 446

   9.1. Classical Parameter Estimation ............................ p. 462
   9.2. Linear Regression .......................................... p. 477
   9.3. Binary Hypothesis Testing ................................. p. 486
   9.4. Significance Testing ....................................... p. 496
   9.5. Summary and Discussion .................................. p. 505
      Problems ....................................................... p. 507

Index ................................................................. p. 521
1

Sample Space and Probability

Contents

1.1. Sets ........................................ p. 3
1.2. Probabilistic Models ......................... p. 6
1.3. Conditional Probability ....................... p. 18
1.4. Total Probability Theorem and Bayes’ Rule ........ p. 28
1.5. Independence ................................... p. 34
1.6. Counting ..................................... p. 44
1.7. Summary and Discussion ....................... p. 51
    Problems ..................................... p. 53
“Probability” is a very useful concept, but can be interpreted in a number of ways. As an illustration, consider the following.

A patient is admitted to the hospital and a potentially life-saving drug is administered. The following dialog takes place between the nurse and a concerned relative.

RELATIVE: Nurse, what is the probability that the drug will work?
NURSE: I hope it works, we’ll know tomorrow.
RELATIVE: Yes, but what is the probability that it will?
NURSE: Each case is different, we have to wait.
RELATIVE: But let’s see, out of a hundred patients that are treated under similar conditions, how many times would you expect it to work?
NURSE (somewhat annoyed): I told you, every person is different, for some it works, for some it doesn’t.
RELATIVE (insisting): Then tell me, if you had to bet whether it will work or not, which side of the bet would you take?
NURSE (cheering up for a moment): I’d bet it will work.
RELATIVE (somewhat relieved): OK, now, would you be willing to lose two dollars if it doesn’t work, and gain one dollar if it does?
NURSE (exasperated): What a sick thought! You are wasting my time!

In this conversation, the relative attempts to use the concept of probability to discuss an uncertain situation. The nurse’s initial response indicates that the meaning of “probability” is not uniformly shared or understood, and the relative tries to make it more concrete. The first approach is to define probability in terms of frequency of occurrence, as a percentage of successes in a moderately large number of similar situations. Such an interpretation is often natural. For example, when we say that a perfectly manufactured coin lands on heads “with probability 50%,” we typically mean “roughly half of the time.” But the nurse may not be entirely wrong in refusing to discuss in such terms. What if this was an experimental drug that was administered for the very first time in this hospital or in the nurse’s experience?

While there are many situations involving uncertainty in which the frequency interpretation is appropriate, there are other situations in which it is not. Consider, for example, a scholar who asserts that the Iliad and the Odyssey were composed by the same person, with probability 90%. Such an assertion conveys some information, but not in terms of frequencies, since the subject is a one-time event. Rather, it is an expression of the scholar’s subjective belief. One might think that subjective beliefs are not interesting, at least from a mathematical or scientific point of view. On the other hand, people often have to make choices in the presence of uncertainty, and a systematic way of making use of their beliefs is a prerequisite for successful, or at least consistent, decision making.
In fact, the choices and actions of a rational person can reveal a lot about the inner-held subjective probabilities, even if the person does not make conscious use of probabilistic reasoning. Indeed, the last part of the earlier dialog was an attempt to infer the nurse’s beliefs in an indirect manner. Since the nurse was willing to accept a one-for-one bet that the drug would work, we may infer that the probability of success was judged to be at least 50%. Had the nurse accepted the last proposed bet (two-for-one), this would have indicated a success probability of at least 2/3.

Rather than dwelling further on philosophical issues about the appropriateness of probabilistic reasoning, we will simply take it as a given that the theory of probability is useful in a broad variety of contexts, including some where the assumed probabilities only reflect subjective beliefs. There is a large body of successful applications in science, engineering, medicine, management, etc., and on the basis of this empirical evidence, probability theory is an extremely useful tool.

Our main objective in this book is to develop the art of describing uncertainty in terms of probabilistic models, as well as the skill of probabilistic reasoning. The first step, which is the subject of this chapter, is to describe the generic structure of such models and their basic properties. The models we consider assign probabilities to collections (sets) of possible outcomes. For this reason, we must begin with a short review of set theory.

1.1 SETS

Probability makes extensive use of set operations, so let us introduce at the outset the relevant notation and terminology.

A set is a collection of objects, which are the elements of the set. If \( S \) is a set and \( x \) is an element of \( S \), we write \( x \in S \). If \( x \) is not an element of \( S \), we write \( x \notin S \). A set can have no elements, in which case it is called the empty set, denoted by \( \emptyset \).

Sets can be specified in a variety of ways. If \( S \) contains a finite number of elements, say \( x_1, x_2, \ldots, x_n \), we write it as a list of the elements, in braces:

\[
S = \{x_1, x_2, \ldots, x_n\}.
\]

For example, the set of possible outcomes of a die roll is \( \{1, 2, 3, 4, 5, 6\} \), and the set of possible outcomes of a coin toss is \( \{H, T\} \), where \( H \) stands for "heads" and \( T \) stands for "tails."

If \( S \) contains infinitely many elements \( x_1, x_2, \ldots \), which can be enumerated in a list (so that there are as many elements as there are positive integers) we write

\[
S = \{x_1, x_2, \ldots\},
\]

and we say that \( S \) is countably infinite. For example, the set of even integers can be written as \( \{0, 2, -2, 4, -4, \ldots\} \), and is countably infinite.
Alternatively, we can consider the set of all \( x \) that have a certain property \( P \), and denote it by
\[
\{ x \mid x \text{ satisfies } P \}.
\]
(The symbol "|" is to be read as "such that"). For example, the set of even integers can be written as \( \{ k \mid k / 2 \text{ is integer} \} \). Similarly, the set of all scalars \( x \) in the interval \([0, 1]\) can be written as \( \{ x \mid 0 \leq x \leq 1 \} \). Note that the elements \( x \) of the latter set take a continuous range of values, and cannot be written down in a list (a proof is sketched in the end-of-chapter problems); such a set is said to be uncountable.

If every element of a set \( S \) is also an element of a set \( T \), we say that \( S \) is a subset of \( T \), and we write \( S \subseteq T \) or \( T \supseteq S \). If \( S \subseteq T \) and \( T \subseteq S \), the two sets are equal, and we write \( S = T \). It is also expedient to introduce a universal set, denoted by \( \Omega \), which contains all objects that could conceivably be of interest in a particular context. Having specified the context in terms of a universal set \( \Omega \), we only consider sets \( S \) that are subsets of \( \Omega \).

**Set Operations**

The complement of a set \( S \), with respect to the universe \( \Omega \), is the set \( \{ x \in \Omega \mid x \notin S \} \) of all elements of \( \Omega \) that do not belong to \( S \), and is denoted by \( S^c \). Note that \( \Omega^c = \emptyset \).

The union of two sets \( S \) and \( T \) is the set of all elements that belong to \( S \) or \( T \) (or both), and is denoted by \( S \cup T \). The intersection of two sets \( S \) and \( T \) is the set of all elements that belong to both \( S \) and \( T \), and is denoted by \( S \cap T \). Thus,
\[
S \cup T = \{ x \mid x \in S \text{ or } x \in T \},
\]
and
\[
S \cap T = \{ x \mid x \in S \text{ and } x \in T \}.
\]

In some cases, we will have to consider the union or the intersection of several, even infinitely many sets, defined in the obvious way. For example, if for every positive integer \( n \), we are given a set \( S_n \), then
\[
\bigcup_{n=1}^{\infty} S_n = S_1 \cup S_2 \cup \cdots = \{ x \mid x \in S_n \text{ for some } n \},
\]
and
\[
\bigcap_{n=1}^{\infty} S_n = S_1 \cap S_2 \cap \cdots = \{ x \mid x \in S_n \text{ for all } n \}.
\]

Two sets are said to be disjoint if their intersection is empty. More generally, several sets are said to be disjoint if no two of them have a common element. A collection of sets is said to be a partition of a set \( S \) if the sets in the collection are disjoint and their union is \( S \).
If $x$ and $y$ are two objects, we use $(x, y)$ to denote the ordered pair of $x$ and $y$. The set of scalars (real numbers) is denoted by $\mathbb{R}$; the set of pairs (or triplets) of scalars, i.e., the two-dimensional plane (or three-dimensional space, respectively) is denoted by $\mathbb{R}^2$ (or $\mathbb{R}^3$, respectively).

Sets and the associated operations are easy to visualize in terms of Venn diagrams, as illustrated in Fig. 1.1.

![Venn Diagrams](image)

**Figure 1.1:** Examples of Venn diagrams. (a) The shaded region is $S \cap T$. (b) The shaded region is $S \cup T$. (c) The shaded region is $S \cap T^c$. (d) Here, $T \subseteq S$. The shaded region is the complement of $S$. (e) The sets $S$, $T$, and $U$ are disjoint. (f) The sets $S$, $T$, and $U$ form a partition of the set $\Omega$.

### The Algebra of Sets

Set operations have several properties, which are elementary consequences of the definitions. Some examples are:

$$
S \cup T = T \cup S,
S \cap (T \cup U) = (S \cap T) \cup (S \cap U),
(S^c)^c = S,
S \cup \Omega = \Omega.
$$

Two particularly useful properties are given by De Morgan’s laws which state that

$$(\bigcup_n S_n)^c = \bigcap_n S_n^c, \quad \left(\bigcap_n S_n\right)^c = \bigcup_n S_n^c.$$ 

To establish the first law, suppose that $x \in (\bigcup_n S_n)^c$. Then, $x \notin \bigcup_n S_n$, which implies that for every $n$, we have $x \notin S_n$. Thus, $x$ belongs to the complement
of every $S_n$. and $x \in \cap_n S_n$. This shows that $(\cup_n S_n)^c \subset \cap_n S_n^c$. The converse inclusion is established by reversing the above argument, and the first law follows. The argument for the second law is similar.

1.2 PROBABILISTIC MODELS

A probabilistic model is a mathematical description of an uncertain situation. It must be in accordance with a fundamental framework that we discuss in this section. Its two main ingredients are listed below and are visualized in Fig. 1.2.

Elements of a Probabilistic Model

- The sample space $\Omega$, which is the set of all possible outcomes of an experiment.

- The probability law, which assigns to a set $A$ of possible outcomes (also called an event) a nonnegative number $P(A)$ (called the probability of $A$) that encodes our knowledge or belief about the collective "likelihood" of the elements of $A$. The probability law must satisfy certain properties to be introduced shortly.

![Figure 1.2: The main ingredients of a probabilistic model.](image)

Sample Spaces and Events

Every probabilistic model involves an underlying process, called the experiment, that will produce exactly one out of several possible outcomes. The set of all possible outcomes is called the sample space of the experiment, and is denoted by $\Omega$. A subset of the sample space, that is, a collection of possible
outcomes, is called an event.† There is no restriction on what constitutes an experiment. For example, it could be a single toss of a coin, or three tosses, or an infinite sequence of tosses. However, it is important to note that in our formulation of a probabilistic model, there is only one experiment. So, three tosses of a coin constitute a single experiment, rather than three experiments.

The sample space of an experiment may consist of a finite or an infinite number of possible outcomes. Finite sample spaces are conceptually and mathematically simpler. Still, sample spaces with an infinite number of elements are quite common. As an example, consider throwing a dart on a square target and viewing the point of impact as the outcome.

Choosing an Appropriate Sample Space

Regardless of their number, different elements of the sample space should be distinct and mutually exclusive, so that when the experiment is carried out there is a unique outcome. For example, the sample space associated with the roll of a die cannot contain “1 or 3” as a possible outcome and also “1 or 4” as another possible outcome. If it did, we would not be able to assign a unique outcome when the roll is a 1.

A given physical situation may be modeled in several different ways, depending on the kind of questions that we are interested in. Generally, the sample space chosen for a probabilistic model must be collectively exhaustive, in the sense that no matter what happens in the experiment, we always obtain an outcome that has been included in the sample space. In addition, the sample space should have enough detail to distinguish between all outcomes of interest to the modeler, while avoiding irrelevant details.

Example 1.1. Consider two alternative games, both involving ten successive coin tosses:

Game 1: We receive $1 each time a head comes up.

Game 2: We receive $1 for every coin toss, up to and including the first time a head comes up. Then, we receive $2 for every coin toss, up to the second time a head comes up. More generally, the dollar amount per toss is doubled each time a head comes up.

† Any collection of possible outcomes, including the entire sample space Ω and its complement, the empty set ∅, may qualify as an event. Strictly speaking, however, some sets have to be excluded. In particular, when dealing with probabilistic models involving an uncountably infinite sample space, there are certain unusual subsets for which one cannot associate meaningful probabilities. This is an intricate technical issue, involving the mathematics of measure theory. Fortunately, such pathological subsets do not arise in the problems considered in this text or in practice, and the issue can be safely ignored.
In game 1, it is only the total number of heads in the ten-toss sequence that matters, while in game 2, the order of heads and tails is also important. Thus, in a probabilistic model for game 1, we can work with a sample space consisting of eleven possible outcomes, namely, 0, 1, ..., 10. In game 2, a finer grain description of the experiment is called for, and it is more appropriate to let the sample space consist of every possible ten-long sequence of heads and tails.

### Sequential Models

Many experiments have an inherently sequential character; for example, tossing a coin three times, observing the value of a stock on five successive days, or receiving eight successive digits at a communication receiver. It is then often useful to describe the experiment and the associated sample space by means of a tree-based sequential description, as in Fig. 1.3.

![Figure 1.3](image)

**Figure 1.3:** Two equivalent descriptions of the sample space of an experiment involving two rolls of a 4-sided die. The possible outcomes are all the ordered pairs of the form \((i,j)\), where \(i\) is the result of the first roll, and \(j\) is the result of the second. These outcomes can be arranged in a 2-dimensional grid as in the figure on the left, or they can be described by the tree on the right, which reflects the sequential character of the experiment. Here, each possible outcome corresponds to a leaf of the tree and is associated with the unique path from the root to that leaf. The shaded area on the left is the event \(\{(1,4), (2,4), (3,4), (4,4)\}\) that the result of the second roll is 4. That same event can be described by the set of leaves highlighted on the right. Note also that every node of the tree can be identified with an event, namely, the set of all leaves downstream from that node. For example, the node labeled by a 1 can be identified with the event \(\{(1,1), (1,2), (1,3), (1,4)\}\) that the result of the first roll is 1.

### Probability Laws

Suppose we have settled on the sample space \(\Omega\) associated with an experiment. To complete the probabilistic model, we must now introduce a probability law.
Intuitively, this specifies the “likelihood” of any outcome, or of any set of possible outcomes (an event, as we have called it earlier). More precisely, the probability law assigns to every event \( A \), a number \( P(A) \), called the probability of \( A \), satisfying the following axioms.

**Probability Axioms**

1. **(Nonnegativity)** \( P(A) \geq 0 \), for every event \( A \).
2. **(Additivity)** If \( A \) and \( B \) are two disjoint events, then the probability of their union satisfies

\[
P(A \cup B) = P(A) + P(B).
\]

More generally, if the sample space has an infinite number of elements and \( A_1, A_2, \ldots \) is a sequence of disjoint events, then the probability of their union satisfies

\[
P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots.
\]

3. **(Normalization)** The probability of the entire sample space \( \Omega \) is equal to 1, that is, \( P(\Omega) = 1 \).

In order to visualize a probability law, consider a unit of mass which is “spread” over the sample space. Then, \( P(A) \) is simply the total mass that was assigned collectively to the elements of \( A \). In terms of this analogy, the additivity axiom becomes quite intuitive: the total mass in a sequence of disjoint events is the sum of their individual masses.

A more concrete interpretation of probabilities is in terms of relative frequencies: a statement such as \( P(A) = 2/3 \) often represents a belief that event \( A \) will occur in about two thirds out of a large number of repetitions of the experiment. Such an interpretation, though not always appropriate, can sometimes facilitate our intuitive understanding. It will be revisited in Chapter 5, in our study of limit theorems.

There are many natural properties of a probability law, which have not been included in the above axioms for the simple reason that they can be derived from them. For example, note that the normalization and additivity axioms imply that

\[
1 = P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset).
\]

and this shows that the probability of the empty event is 0:

\[
P(\emptyset) = 0.
\]
As another example, consider three disjoint events \( A_1, A_2, \) and \( A_3 \). We can use the additivity axiom for two disjoint events repeatedly, to obtain

\[
P(A_1 \cup A_2 \cup A_3) = P(A_1 \cup (A_2 \cup A_3))
\]

\[
= P(A_1) + P(A_2 \cup A_3)
\]

\[
= P(A_1) + P(A_2) + P(A_3).
\]

Proceeding similarly, we obtain that the probability of the union of finitely many disjoint events is always equal to the sum of the probabilities of these events. More such properties will be considered shortly.

**Discrete Models**

Here is an illustration of how to construct a probability law starting from some common sense assumptions about a model.

**Example 1.2.** Consider an experiment involving a single coin toss. There are two possible outcomes, heads \((H)\) and tails \((T)\). The sample space is \( \Omega = \{H,T\} \), and the events are

\[
\{H,T\}, \{H\}, \{T\}, \emptyset.
\]

If the coin is fair, i.e., if we believe that heads and tails are “equally likely,” we should assign equal probabilities to the two possible outcomes and specify that \( P(\{H\}) = P(\{T\}) = 0.5 \). The additivity axiom implies that

\[
P(\{H,T\}) = P(\{H\}) + P(\{T\}) = 1,
\]

which is consistent with the normalization axiom. Thus, the probability law is given by

\[
P(H,T) = 1, \quad P(H) = 0.5, \quad P(T) = 0.5, \quad P(\emptyset) = 0,
\]

and satisfies all three axioms.

Consider another experiment involving three coin tosses. The outcome will now be a 3-long string of heads or tails. The sample space is

\[
\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.
\]

We assume that each possible outcome has the same probability of \( 1/8 \). Let us construct a probability law that satisfies the three axioms. Consider, as an example, the event

\[
A = \{\text{exactly 2 heads occur}\} = \{HHT, HTH, THH\}.
\]

Using additivity, the probability of \( A \) is the sum of the probabilities of its elements:

\[
P(HHT, HTH, THH) = P(HHT) + P(HTH) + P(THH)
\]

\[
= \frac{1}{8} + \frac{1}{8} + \frac{1}{8}
\]

\[
= \frac{3}{8}.
\]
Similarly, the probability of any event is equal to 1/8 times the number of possible outcomes contained in the event. This defines a probability law that satisfies the three axioms.

By using the additivity axiom and by generalizing the reasoning in the preceding example, we reach the following conclusion.

**Discrete Probability Law**

If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. In particular, the probability of any event \( \{s_1, s_2, \ldots, s_n\} \) is the sum of the probabilities of its elements:

\[
P(\{s_1, s_2, \ldots, s_n\}) = P(s_1) + P(s_2) + \ldots + P(s_n).
\]

Note that we are using here the simpler notation \( P(s_i) \) to denote the probability of the event \( \{s_i\} \), instead of the more precise \( P(\{s_i\}) \). This convention will be used throughout the remainder of the book.

In the special case where the probabilities \( P(s_1), \ldots, P(s_n) \) are all the same (by necessity equal to \( 1/n \), in view of the normalization axiom), we obtain the following.

**Discrete Uniform Probability Law**

If the sample space consists of \( n \) possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event \( A \) is given by

\[
P(A) = \frac{\text{number of elements of } A}{n}.
\]

Let us provide a few more examples of sample spaces and probability laws.

**Example 1.3.** Consider the experiment of rolling a pair of 4-sided dice (cf. Fig. 1.4). We assume the dice are fair, and we interpret this assumption to mean that each of the sixteen possible outcomes [pairs \((i,j)\), with \(i, j = 1, 2, 3, 4\)] has the same probability of 1/16. To calculate the probability of an event, we must count the number of elements of the event and divide by 16 (the total number of possible
outcomes. Here are some event probabilities calculated in this way:

\[
P(\{\text{the sum of the rolls is even}\}) = \frac{8}{16} = \frac{1}{2},
\]
\[
P(\{\text{the sum of the rolls is odd}\}) = \frac{8}{16} = \frac{1}{2},
\]
\[
P(\{\text{the first roll is equal to the second}\}) = \frac{4}{16} = \frac{1}{4},
\]
\[
P(\{\text{the first roll is larger than the second}\}) = \frac{6}{16} = \frac{3}{8},
\]
\[
P(\{\text{at least one roll is equal to 4}\}) = \frac{7}{16}.
\]

Figure 1.4: Various events in the experiment of rolling a pair of 4-sided dice, and their probabilities, calculated according to the discrete uniform law.

Continuous Models

Probabilistic models with continuous sample spaces differ from their discrete counterparts in that the probabilities of the single-element events may not be sufficient to characterize the probability law. This is illustrated in the following examples, which also indicate how to generalize the uniform probability law to the case of a continuous sample space.

Example 1.4. A wheel of fortune is continuously calibrated from 0 to 1, so the possible outcomes of an experiment consisting of a single spin are the numbers in the interval \( \Omega = [0, 1] \). Assuming a fair wheel, it is appropriate to consider all outcomes equally likely, but what is the probability of the event consisting of a single element? It cannot be positive, because then, using the additivity axiom, it would follow that events with a sufficiently large number of elements would have
probability larger than 1. Therefore, the probability of any event that consists of a single element must be 0.

In this example, it makes sense to assign probability $b - a$ to any subinterval $[a, b]$ of $[0, 1]$, and to calculate the probability of a more complicated set by evaluating its “length.” This assignment satisfies the three probability axioms and qualifies as a legitimate probability law.

**Example 1.5.** Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour, with all pairs of delays being equally likely. The first to arrive will wait for 15 minutes and will leave if the other has not yet arrived. What is the probability that they will meet?

Let us use as sample space the unit square, whose elements are the possible pairs of delays for the two of them. Our interpretation of “equally likely” pairs of delays is to let the probability of a subset of $\Omega$ be equal to its area. This probability law satisfies the three probability axioms. The event that Romeo and Juliet will meet is the shaded region in Fig. 1.5, and its probability is calculated to be $7/16$.

![Diagram showing the event M that Romeo and Juliet will arrive within 15 minutes of each other.](image)

**Figure 1.5:** The event $M$ that Romeo and Juliet will arrive within 15 minutes of each other (cf. Example 1.5) is

$$M = \left\{ (x, y) \mid |x - y| \leq 1/4, 0 \leq x \leq 1, 0 \leq y \leq 1 \right\},$$

and is shaded in the figure. The area of $M$ is 1 minus the area of the two unshaded triangles, or $1 - (3/4) \cdot (3/4) = 7/16$. Thus, the probability of meeting is $7/16$.

† The “length” of a subset $S$ of $[0, 1]$ is the integral $\int_S dt$, which is defined, for “nice” sets $S$, in the usual calculus sense. For unusual sets, this integral may not be well defined mathematically, but such issues belong to a more advanced treatment of the subject. Incidentally, the legitimacy of using length as a probability law hinges on the fact that the unit interval has an uncountably infinite number of elements. Indeed, if the unit interval had a countable number of elements, with each element having zero probability, the additivity axiom would imply that the whole interval has zero probability, which would contradict the normalization axiom.
Properties of Probability Laws

Probability laws have a number of properties, which can be deduced from the axioms. Some of them are summarized below.

<table>
<thead>
<tr>
<th>Some Properties of Probability Laws</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consider a probability law, and let $A$, $B$, and $C$ be events.</td>
</tr>
<tr>
<td>(a) If $A \subset B$, then $P(A) \leq P(B)$.</td>
</tr>
<tr>
<td>(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.</td>
</tr>
<tr>
<td>(c) $P(A \cup B) \leq P(A) + P(B)$.</td>
</tr>
<tr>
<td>(d) $P(A \cup B \cup C) = P(A) + P(A^c \cap B) + P(A^c \cap B^c \cap C)$.</td>
</tr>
</tbody>
</table>

These properties, and other similar ones, can be visualized and verified graphically using Venn diagrams, as in Fig. 1.6. Note that property (c) can be generalized as follows:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} P(A_i).$$

To see this, we apply property (c) to the sets $A_1$ and $A_2 \cup \cdots \cup A_n$, to obtain

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2 \cup \cdots \cup A_n).$$

We also apply property (c) to the sets $A_2$ and $A_3 \cup \cdots \cup A_n$, to obtain

$$P(A_2 \cup \cdots \cup A_n) \leq P(A_2) + P(A_3 \cup \cdots \cup A_n).$$

We continue similarly, and finally add.

Models and Reality

The framework of probability theory can be used to analyze uncertainty in a wide variety of physical contexts. Typically, this involves two distinct stages.

(a) In the first stage, we construct a probabilistic model by specifying a probability law on a suitably defined sample space. There are no hard rules to guide this step, other than the requirement that the probability law conform to the three axioms. Reasonable people may disagree on which model best represents reality. In many cases, one may even want to use a somewhat “incorrect” model, if it is simpler than the “correct” one or allows for tractable calculations. This is consistent with common practice in science.
and engineering, where the choice of a model often involves a tradeoff between accuracy, simplicity, and tractability. Sometimes, a model is chosen on the basis of historical data or past outcomes of similar experiments, using statistical inference methods, which will be discussed in Chapters 8 and 9.

![Venn Diagrams](image)

**Figure 1.6:** Visualization and verification of various properties of probability laws using Venn diagrams. If $A \subset B$, then $B$ is the union of the two disjoint events $A$ and $A^c \cap B$; see diagram (a). Therefore, by the additivity axiom, we have

$$P(B) = P(A) + P(A^c \cap B) \geq P(A),$$

where the inequality follows from the nonnegativity axiom and verifies property (a).

From diagram (b), we can express the events $A \cup B$ and $B$ as unions of disjoint events:

$$A \cup B = A \cup (A^c \cap B), \quad B = (A \cap B) \cup (A^c \cap B).$$

Using the additivity axiom, we have

$$P(A \cup B) = P(A) + P(A^c \cap B), \quad P(B) = P(A \cap B) + P(A^c \cap B).$$

Subtracting the second equality from the first and rearranging terms, we obtain

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

verifying property (b). Using also the fact $P(A \cap B) \geq 0$ (the nonnegativity axiom), we obtain

$$P(A \cup B) \leq P(A) + P(B),$$

verifying property (c).

From diagram (c), we see that the event $A \cup B \cup C$ can be expressed as a union of three disjoint events:

$$A \cup B \cup C = A \cup (A^c \cap B) \cup (A^c \cap B^c \cap C),$$

so property (d) follows as a consequence of the additivity axiom.
(b) In the second stage, we work within a fully specified probabilistic model and derive the probabilities of certain events, or deduce some interesting properties. While the first stage entails the often open-ended task of connecting the real world with mathematics, the second one is tightly regulated by the rules of ordinary logic and the axioms of probability. Difficulties may arise in the latter if some required calculations are complex, or if a probability law is specified in an indirect fashion. Even so, there is no room for ambiguity: all conceivable questions have precise answers and it is only a matter of developing the skill to arrive at them.

Probability theory is full of "paradoxes" in which different calculation methods seem to give different answers to the same question. Invariably though, these apparent inconsistencies turn out to reflect poorly specified or ambiguous probabilistic models. An example, Bertrand's paradox, is shown in Fig. 1.7.

Figure 1.7: This example, presented by L. F. Bertrand in 1889, illustrates the need to specify unambiguously a probabilistic model. Consider a circle and an equilateral triangle inscribed in the circle. What is the probability that the length of a randomly chosen chord of the circle is greater than the side of the triangle? The answer here depends on the precise meaning of "randomly chosen." The two methods illustrated in parts (a) and (b) of the figure lead to contradictory results.

In (a), we take a radius of the circle, such as $AB$, and we choose a point $C$ on that radius, with all points being equally likely. We then draw the chord through $C$ that is orthogonal to $AB$. From elementary geometry, $AB$ intersects the triangle at the midpoint of $AB$, so the probability that the length of the chord is greater than the side is $1/2$.

In (b), we take a point on the circle, such as the vertex $V$, we draw the tangent to the circle through $V$, and we draw a line through $V$ that forms a random angle $\Phi$ with the tangent, with all angles being equally likely. We consider the chord obtained by the intersection of this line with the circle. From elementary geometry, the length of the chord is greater than the side of the triangle if $\Phi$ is between $\pi/3$ and $2\pi/3$. Since $\Phi$ takes values between $0$ and $\pi$, the probability that the length of the chord is greater than the side is $1/3$. 
A Brief History of Probability

- **B.C.E.** Games of chance were popular in ancient Greece and Rome, but no scientific development of the subject took place, possibly because the number system used by the Greeks did not facilitate algebraic calculations. The development of probability based on sound scientific analysis had to await the development of the modern arithmetic system by the Hindus and the Arabs in the second half of the first millennium, as well as the flood of scientific ideas generated by the Renaissance.

- **16th century.** Girolamo Cardano, a colorful and controversial Italian mathematician, publishes the first book describing correct methods for calculating probabilities in games of chance involving dice and cards.

- **17th century.** A correspondence between Fermat and Pascal touches upon several interesting probability questions and motivates further study in the field.

- **18th century.** Jacob Bernoulli studies repeated coin tossing and introduces the first law of large numbers, which lays a foundation for linking theoretical probability concepts and empirical fact. Several mathematicians, such as Daniel Bernoulli, Leibnitz, Bayes, and Lagrange, make important contributions to probability theory and its use in analyzing real-world phenomena. De Moivre introduces the normal distribution and proves the first form of the central limit theorem.

- **19th century.** Laplace publishes an influential book that establishes the importance of probability as a quantitative field and contains many original contributions, including a more general version of the central limit theorem. Legendre and Gauss apply probability to astronomical predictions, using the method of least squares, thus pointing the way to a vast range of applications. Poisson publishes an influential book with many original contributions, including the Poisson distribution. Chebyshev, and his students Markov and Lyapunov, study limit theorems and raise the standards of mathematical rigor in the field. Throughout this period, probability theory is largely viewed as a natural science, its primary goal being the explanation of physical phenomena. Consistently with this goal, probabilities are mainly interpreted as limits of relative frequencies in the context of repeatable experiments.

- **20th century.** Relative frequency is abandoned as the conceptual foundation of probability theory in favor of a now universally used axiomatic system, introduced by Kolmogorov. Similar to other branches of mathematics, the development of probability theory from the axioms relies only on logical correctness, regardless of its relevance to physical phenomena. Nonetheless, probability theory is used pervasively in science and engineering because of its ability to describe and interpret most types of uncertain phenomena in the real world.
1.3 CONDITIONAL PROBABILITY

Conditional probability provides us with a way to reason about the outcome of an experiment, based on partial information. Here are some examples of situations we have in mind:

(a) In an experiment involving two successive rolls of a die, you are told that the sum of the two rolls is 9. How likely is it that the first roll was a 6?

(b) In a word guessing game, the first letter of the word is a “t”. What is the likelihood that the second letter is an “h”?

(c) How likely is it that a person has a certain disease given that a medical test was negative?

(d) A spot shows up on a radar screen. How likely is it to correspond to an aircraft?

In more precise terms, given an experiment, a corresponding sample space, and a probability law, suppose that we know that the outcome is within some given event \( B \). We wish to quantify the likelihood that the outcome also belongs to some other given event \( A \). We thus seek to construct a new probability law that takes into account the available knowledge: a probability law that for any event \( A \) specifies the conditional probability of \( A \) given \( B \), denoted by \( P(A \mid B) \).

We would like the conditional probabilities \( P(A \mid B) \) of different events \( A \) to constitute a legitimate probability law, which satisfies the probability axioms. The conditional probabilities should also be consistent with our intuition in important special cases, e.g., when all possible outcomes of the experiment are equally likely. For example, suppose that all six possible outcomes of a fair die roll are equally likely. If we are told that the outcome is even, we are left with only three possible outcomes, namely, 2, 4, and 6. These three outcomes were equally likely to start with, and so they should remain equally likely given the additional knowledge that the outcome was even. Thus, it is reasonable to let

\[
P(\text{the outcome is 6} \mid \text{the outcome is even}) = \frac{1}{3}.
\]

This argument suggests that an appropriate definition of conditional probability when all outcomes are equally likely, is given by

\[
P(A \mid B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.
\]

Generalizing the argument, we introduce the following definition of conditional probability:

\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}.
\]
where we assume that \( P(B) > 0 \); the conditional probability is undefined if the conditioning event has zero probability. In words, out of the total probability of the elements of \( B \), \( P(A | B) \) is the fraction that is assigned to possible outcomes that also belong to \( A \).

### Conditional Probabilities Specify a Probability Law

For a fixed event \( B \), it can be verified that the conditional probabilities \( P(A | B) \) form a legitimate probability law that satisfies the three axioms. Indeed, non-negativity is clear. Furthermore,

\[
P(\Omega | B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1,
\]

and the normalization axiom is also satisfied. To verify the additivity axiom, we write for any two disjoint events \( A_1 \) and \( A_2 \),

\[
P(A_1 \cup A_2 | B) = \frac{P((A_1 \cup A_2) \cap B)}{P(B)} = \frac{P((A_1 \cap B) \cup (A_2 \cap B))}{P(B)} = \frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)} = \frac{P(A_1)}{P(B)} + \frac{P(A_2)}{P(B)} = P(A_1 | B) + P(A_2 | B),
\]

where for the third equality, we used the fact that \( A_1 \cap B \) and \( A_2 \cap B \) are disjoint sets, and the additivity axiom for the (unconditional) probability law. The argument for a countable collection of disjoint sets is similar.

Since conditional probabilities constitute a legitimate probability law, all general properties of probability laws remain valid. For example, a fact such as \( P(A \cup C) \leq P(A) + P(C) \) translates to the new fact

\[
P(A \cup C | B) \leq P(A | B) + P(C | B).
\]

Let us also note that since we have \( P(B | B) = P(B)/P(B) = 1 \), all of the conditional probability is concentrated on \( B \). Thus, we might as well discard all possible outcomes outside \( B \) and treat the conditional probabilities as a probability law defined on the new universe \( B \).

Let us summarize the conclusions reached so far.
Properties of Conditional Probability

- The conditional probability of an event $A$, given an event $B$ with $P(B) > 0$, is defined by
  \[ P(A | B) = \frac{P(A \cap B)}{P(B)} , \]
  and specifies a new (conditional) probability law on the same sample space $\Omega$. In particular, all properties of probability laws remain valid for conditional probability laws.

- Conditional probabilities can also be viewed as a probability law on a new universe $B$, because all of the conditional probability is concentrated on $B$.

- If the possible outcomes are finitely many and equally likely, then
  \[ P(A | B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B} . \]

Example 1.6. We toss a fair coin three successive times. We wish to find the conditional probability $P(A | B)$ when $A$ and $B$ are the events

$A = \{\text{more heads than tails come up}\}, \quad B = \{\text{1st toss is a head}\}.$

The sample space consists of eight sequences.

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

which we assume to be equally likely. The event $B$ consists of the four elements $HHH, HHT, HTH, HTT$, so its probability is

$P(B) = \frac{4}{8}.$

The event $A \cap B$ consists of the three elements $HHH, HHT, HTH$, so its probability is

$P(A \cap B) = \frac{3}{8}.$

Thus, the conditional probability $P(A | B)$ is

$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{3/8}{4/8} = \frac{3}{4}.$

Because all possible outcomes are equally likely here, we can also compute $P(A | B)$ using a shortcut. We can bypass the calculation of $P(B)$ and $P(A \cap B)$, and simply
divide the number of elements shared by \( A \) and \( B \) (which is 3) with the number of elements of \( B \) (which is 4) to obtain the same result \( 3/4 \).

**Example 1.7.** A fair 4-sided die is rolled twice and we assume that all sixteen possible outcomes are equally likely. Let \( X \) and \( Y \) be the result of the 1st and the 2nd roll, respectively. We wish to determine the conditional probability \( P(A \mid B) \), where

\[
A = \{ \max(X,Y) = m \}, \quad B = \{ \min(X,Y) = 2 \}.
\]

and \( m \) takes each of the values 1, 2, 3, 4.

As in the preceding example, we can first determine the probabilities \( P(A \cap B) \) and \( P(B) \) by counting the number of elements of \( A \cap B \) with the number of elements of \( B \); see Fig. 1.8.

\[
P(\{\max(X,Y) = m\} \mid B) = \begin{cases} 
2/5, & \text{if } m = 3 \text{ or } m = 4, \\
1/5, & \text{if } m = 2, \\
0, & \text{if } m = 1. 
\end{cases}
\]

**Example 1.8.** A conservative design team, call it \( C \), and an innovative design team, call it \( N \), are asked to separately design a new product within a month. From past experience we know that:

(a) The probability that team \( C \) is successful is \( 2/3 \).

(b) The probability that team \( N \) is successful is \( 1/2 \).

(c) The probability that at least one team is successful is \( 3/4 \).
Assuming that exactly one successful design is produced, what is the probability that it was designed by team $N$?

There are four possible outcomes here, corresponding to the four combinations of success and failure of the two teams:

- $SS$: both succeed, $FF$: both fail,
- $SF$: $C$ succeeds, $N$ fails, $FS$: $C$ fails, $N$ succeeds.

We were given that the probabilities of these outcomes satisfy

$$P(SS) + P(SF) = \frac{2}{3}, \quad P(SS) + P(FS) = \frac{1}{2}, \quad P(SS) + P(SF) + P(FS) = \frac{3}{4}.$$

From these relations, together with the normalization equation

$$P(SS) + P(SF) + P(FS) + P(FF) = 1,$$

we can obtain the probabilities of individual outcomes:

$$P(SS) = \frac{5}{12}, \quad P(SF) = \frac{1}{4}, \quad P(FS) = \frac{1}{12}, \quad P(FF) = \frac{1}{4}.$$

The desired conditional probability is

$$P(FS \mid \{SF, FS\}) = \frac{\frac{12}{12}}{\frac{1}{4} + \frac{12}{12}} = \frac{1}{4}.$$

Using Conditional Probability for Modeling

When constructing probabilistic models for experiments that have a sequential character, it is often natural and convenient to first specify conditional probabilities and then use them to determine unconditional probabilities. The rule $P(A \cap B) = P(B)P(A \mid B)$, which is a restatement of the definition of conditional probability, is often helpful in this process.

**Example 1.9. Radar Detection.** If an aircraft is present in a certain area, a radar detects it and generates an alarm signal with probability 0.99. If an aircraft is not present, the radar generates a (false) alarm, with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of no aircraft presence and a false alarm? What is the probability of aircraft presence and no detection?

A sequential representation of the experiment is appropriate here, as shown in Fig. 1.9. Let $A$ and $B$ be the events

$$A = \{\text{an aircraft is present}\},$$

$$B = \{\text{the radar generates an alarm}\},$$
and consider also their complements
\[ A^c = \{ \text{an aircraft is not present} \}, \]
\[ B^c = \{ \text{the radar does not generate an alarm} \}. \]

The given probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.9. Each possible outcome corresponds to a leaf of the tree, and its probability is equal to the product of the probabilities associated with the branches in a path from the root to the corresponding leaf. The desired probabilities are

\[
P(\text{not present, false alarm}) = P(A^c \cap B) = P(A^c)P(B \mid A^c) = 0.95 \cdot 0.10 = 0.095, \]
\[
P(\text{present, no detection}) = P(A \cap B^c) = P(A)P(B^c \mid A) = 0.05 \cdot 0.01 = 0.0005. \]

\[ \text{Figure 1.9: Sequential description of the experiment for the radar detection problem in Example 1.9.} \]

Extending the preceding example, we have a general rule for calculating various probabilities in conjunction with a tree-based sequential description of an experiment. In particular:

(a) We set up the tree so that an event of interest is associated with a leaf. We view the occurrence of the event as a sequence of steps, namely, the traversals of the branches along the path from the root to the leaf.

(b) We record the conditional probabilities associated with the branches of the tree.

(c) We obtain the probability of a leaf by multiplying the probabilities recorded along the corresponding path of the tree.
In mathematical terms, we are dealing with an event \( A \) which occurs if and only if each one of several events \( A_1, \ldots, A_n \) has occurred, i.e., \( A = A_1 \cap A_2 \cap \cdots \cap A_n \). The occurrence of \( A \) is viewed as an occurrence of \( A_1 \), followed by the occurrence of \( A_2 \), then of \( A_3 \), etc., and it is visualized as a path with \( n \) branches, corresponding to the events \( A_1, \ldots, A_n \). The probability of \( A \) is given by the following rule (see also Fig. 1.10).

### Multiplication Rule

Assuming that all of the conditioning events have positive probability, we have

\[
P(\bigcap_{i=1}^{n} A_i) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid \bigcap_{i=1}^{n-1} A_i).
\]

The multiplication rule can be verified by writing

\[
P(\bigcap_{i=1}^{n} A_i) = P(A_1) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdots \frac{P(\bigcap_{i=1}^{n} A_i)}{P(\bigcap_{i=1}^{n-1} A_i)}.
\]

![Figure 1.10: Visualization of the multiplication rule.](image)

The final node of the path corresponds to the intersection event \( A \), and its probability is obtained by multiplying the conditional probabilities recorded along the branches of the path

\[
P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1) \cdots P(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}).
\]

Note that any intermediate node along the path also corresponds to some intersection event and its probability is obtained by multiplying the corresponding conditional probabilities up to that node. For example, the event \( A_1 \cap A_2 \cap A_3 \) corresponds to the node shown in the figure, and its probability is

\[
P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2).
\]
and by using the definition of conditional probability to rewrite the right-hand side above as

\[ P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid \cap_{i=1}^{n-1} A_i). \]

For the case of just two events, \(A_1\) and \(A_2\), the multiplication rule is simply the definition of conditional probability.

**Example 1.10.** Three cards are drawn from an ordinary 52-card deck without replacement (drawn cards are not placed back in the deck). We wish to find the probability that none of the three cards is a heart. We assume that at each step, each one of the remaining cards is equally likely to be picked. By symmetry, this implies that every triplet of cards is equally likely to be drawn. A cumbersome approach, which we will not use, is to count the number of all card triplets that do not include a heart, and divide it with the number of all possible card triplets. Instead, we use a sequential description of the experiment in conjunction with the multiplication rule (cf. Fig. 1.11).

Define the events

\[ A_i = \{ \text{the } i\text{th card is not a heart} \}. \quad i = 1, 2, 3. \]

We will calculate \(P(A_1 \cap A_2 \cap A_3)\), the probability that none of the three cards is a heart, using the multiplication rule

\[ P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2). \]

We have

\[ P(A_1) = \frac{39}{52}, \]

since there are 39 cards that are not hearts in the 52-card deck. Given that the first card is not a heart, we are left with 51 cards. 38 of which are not hearts, and

\[ P(A_2 \mid A_1) = \frac{38}{51}. \]

Finally, given that the first two cards drawn are not hearts, there are 37 cards which are not hearts in the remaining 50-card deck, and

\[ P(A_3 \mid A_1 \cap A_2) = \frac{37}{50}. \]

These probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.11. The desired probability is now obtained by multiplying the probabilities recorded along the corresponding path of the tree:

\[ P(A_1 \cap A_2 \cap A_3) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50}. \]
Note that once the probabilities are recorded along the tree, the probability of several other events can be similarly calculated. For example,

\[
P(1\text{st is not a heart and 2nd is a heart}) = \frac{39}{52} \cdot \frac{13}{51},
\]

\[
P(1\text{st and 2nd are not hearts, and 3rd is a heart}) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{13}{50}.
\]

**Example 1.11.** A class consisting of 4 graduate and 12 undergraduate students is randomly divided into 4 groups of 4. What is the probability that each group includes a graduate student? We interpret "randomly" to mean that given the assignment of some students to certain slots, any of the remaining students is equally likely to be assigned to any of the remaining slots. We then calculate the desired probability using the multiplication rule, based on the sequential description shown in Fig. 1.12. Let us denote the four graduate students by 1, 2, 3, 4, and consider the events

\[A_1 = \{\text{students 1 and 2 are in different groups}\},\]
\[A_2 = \{\text{students 1, 2, and 3 are in different groups}\},\]
\[A_3 = \{\text{students 1, 2, 3, and 4 are in different groups}\}.
\]

We will calculate \(P(A_3)\) using the multiplication rule:

\[
P(A_3) = P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 | A_1)P(A_3 | A_1 \cap A_2).
\]

We have

\[
P(A_1) = \frac{12}{15},
\]

since there are 12 student slots in groups other than the one of student 1, and there are 15 student slots overall, excluding student 1. Similarly,

\[
P(A_2 | A_1) = \frac{8}{14},
\]
since there are 8 student slots in groups other than those of students 1 and 2, and there are 14 student slots, excluding students 1 and 2. Also,

$$P(A_3 | A_1 \cap A_2) = \frac{4}{13},$$

since there are 4 student slots in groups other than those of students 1, 2, and 3, and there are 13 student slots, excluding students 1, 2, and 3. Thus, the desired probability is

$$\frac{12}{15} \cdot \frac{8}{14} \cdot \frac{4}{13},$$

and is obtained by multiplying the conditional probabilities along the corresponding path of the tree in Fig. 1.12.

**Example 1.12. The Monty Hall Problem.** This is a much discussed puzzle, based on an old American game show. You are told that a prize is equally likely to be found behind any one of three closed doors in front of you. You point to one of the doors. A friend opens for you one of the remaining two doors, after making sure that the prize is not behind it. At this point, you can stick to your initial choice, or switch to the other unopened door. You win the prize if it lies behind your final choice of a door. Consider the following strategies:

(a) Stick to your initial choice.

(b) Switch to the other unopened door.

(c) You first point to door 1. If door 2 is opened, you do not switch. If door 3 is opened, you switch.

Which is the best strategy? To answer the question, let us calculate the probability of winning under each of the three strategies.

Under the strategy of no switching, your initial choice will determine whether you win or not, and the probability of winning is $1/3$. This is because the prize is equally likely to be behind each door.

Under the strategy of switching, if the prize is behind the initially chosen door (probability $1/3$), you do not win. If it is not (probability $2/3$), and given that
another door without a prize has been opened for you, you will get to the winning door once you switch. Thus, the probability of winning is now 2/3, so (b) is a better strategy than (a).

Consider now strategy (c). Under this strategy, there is insufficient information for determining the probability of winning. The answer depends on the way that your friend chooses which door to open. Let us consider two possibilities.

Suppose that if the prize is behind door 1, your friend always chooses to open door 2. (If the prize is behind door 2 or 3, your friend has no choice.) If the prize is behind door 1, your friend opens door 2, you do not switch, and you win. If the prize is behind door 2, your friend opens door 3, you switch, and you win. If the prize is behind door 3, your friend opens door 2, you do not switch, and you lose. Thus, the probability of winning is 2/3. so strategy (c) in this case is as good as strategy (b).

Suppose now that if the prize is behind door 1, your friend is equally likely to open either door 2 or 3. If the prize is behind door 1 (probability 1/3), and if your friend opens door 2 (probability 1/2), you do not switch and you win (probability 1/6). But if your friend opens door 3, you switch and you lose. If the prize is behind door 2, your friend opens door 3, you switch, and you win (probability 1/3). If the prize is behind door 3, your friend opens door 2, you do not switch and you lose. Thus, the probability of winning is 1/6 + 1/3 = 1/2, so strategy (c) in this case is inferior to strategy (b).

1.4 TOTAL PROBABILITY THEOREM AND BAYES’ RULE

In this section, we explore some applications of conditional probability. We start with the following theorem, which is often useful for computing the probabilities of various events, using a “divide-and-conquer” approach.

**Total Probability Theorem**

Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events $A_1, \ldots, A_n$) and assume that $P(A_i) > 0$, for all $i$. Then, for any event $B$, we have

\[
P(B) = P(A_1 \cap B) + \cdots + P(A_n \cap B)
= P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n).
\]

The theorem is visualized and proved in Fig. 1.13. Intuitively, we are partitioning the sample space into a number of scenarios (events) $A_i$. Then, the probability that $B$ occurs is a weighted average of its conditional probability under each scenario, where each scenario is weighted according to its (unconditional) probability. One of the uses of the theorem is to compute the probability of various events $B$ for which the conditional probabilities $P(B \mid A_i)$ are known or