EK 381
Probability for Engineers
Class Notes

Prof. D. Castañón & Prof. B. Nazer
Dept. of Electrical and Computer Engineering

Boston University
College of Engineering
8 St. Mary's Street

Boston, MA 02215

Fall, 2020

Contents

List of Figures

List of Tables

Chapter 1

Foundations of Probability

1.1 Introduction

What is probability theory? It is an *axiomatic* theory which *describes and predicts* the outcomes of inexact, *repeated* experiments. Note the emphases in the above definition. The basis of probabilistic analysis is to determine or estimate the probabilities that certain known events occur, and then to use the axioms of probability theory to derive probabilities of other events of interest, and to predict the outcomes of certain experiments.

For example, consider any card game. The inexact experiment is the shuffling of a deck of cards, with the outcome being the order in which the cards appear. An estimate of the underlying probabilities would be that all orderings are equally likely; an event might be a collection of outcomes, such as all the outcomes where the ace of spades is the first card. The underlying events would then be assigned a given probability.

Based on the underlying probability of the events, you may wish to compute the probability that, if you are playing alone against a dealer, you would win a hand of blackjack. Certain orderings of the cards lead to winning hands, and the probability of winning can be computed from the combined information on the orderings.

While card games and other games of chance make fun illustrations for applications of probability, we are interested in using probability for engineering problems. Why do we use probability in such problems? First, we use probability to model phenomena whose outcomes are too hard to model because they involve too many microscopic factors. For instance, the temperature in a room is the result of kinetic energy events released from particle collisions, but modeling those events by representing trajectories of molecules in a room requires very large scale computations. Instead, we can use a probabilistic description of those collisions that forms the basis for thermodynamics.

A different reason we use probability in engineering is to model lack of precision in measurement. No measurement instrument is exact, and all measurements incur some degree of error. We use probability for representing the errors in what we measure versus the actual measured value. For instance, data received over communications channels are subject to unknown distortions whose effect is captured using probability models. A third reason to use probability in engineering problems arises in representing physical phenomena at atomic levels, in modern physics and quantum mechanics. Heisenberg's famous uncertainty principle uses probability to describe fundamental limits in knowing both the position and momentum of atomic particles.

Below is a brief list of examples of how probability models are used in different fields of engineering and science:

- Game Theory: We model outcomes of games of chance, such as cards, rolls of dice, landing of roulette balls, etc. We use those models to derive superior playing strategies that maximize our odds of winning.
- Weather: the evolution of weather fronts over time is subject to many unknown variations, so weather prediction uses probability to estimate likely weather patterns, including predicting hurricane trajectories and strenghts.
- Finance: Probability models are the foundation for mathematical finance, to represent the uncertain evolution of stock prices over time.

- Physics: probability is used to represent possible locations for electrons in orbits, and in statistical mechanics to represent macroscopic effects of numerous molecular motions.
- Molecular Biology: potential DNA mutations of a virus are represented using probability models.
- Science and Engineering Measurements: errors are represented using probability models. Max Born made the observation measured values are within a factor from true values. To quote Max Born, one of the pioneers of quantum mechanics, "The conception of chance enters in the very first steps of scientific activity in virtue of the fact that no observation is absolutely correct."
- Circuits: The true resistance, capacitance and inductance of circuit elements is variable, and these variations are often modeled using probability.
- Optics: The actual number of photons per unit time emitted by a source of given intensity is random, and modeled using probability.
- Transportation: The travel time on roads, the routes selected by traffic, and the wait times at intersections and toll booths are represented using probability models to predict traffic flow.
- Manufacturing: Production times of parts, demand for products, variations in supply chain deliveries are effects that are often modeled by probability.
- Robotics: Problems in determining robot position from sensor data are estimation problems solved using the theory of probability
- Medicine: Problems in diagnosis based on observed patient data are fundamental hypothesis testing problems best addressed using probability.
  item Nuclear Engineering: Failure analysis and diagnosis is based on probabilistic reasoning.
- Astronomy: Detecting and tracking the location of celestial objects using different instruments is based on the theory of probability.
- Data Science: The foundations of data science are the probabilistic theories of estimation and classification.

As the above list indicates, Probability Theory is useful across a wide range of engineering applications.

What do we mean by the probability of an event? This foundational question has been the focus of debate for several centuries, and has several possible answers: One interpretation is the *frequentist* interpretation that the probability of an event means that, if an experiment is repeated an infinite number of times, the probability of the event is the fraction of times that the event occurs in the repeated experiments. This is often used when dealing with simple physical processes, such as rolling dice, shuffling cards, and measurement systems, where experiments can be repeated a large number of times at low cost.

There is a different school of thought: the *subjectivist* interpretation of the probability of an event represents an individual belief that the event will occur, and reflects how much one would be willing to bet that the event will occur. This interpretation is most appropriate when experiments cannot be repeated, such as in economics and social situations. For instance, what is the probability that the New England Patriots will win the Super Bowl this year? That is not an experiment that can be repeated; furthermore, asking that question from different individuals can result in very different estimates of that probability. Similarly, the probability that a nuclear reactor will fail corresponds to events that are hard to repeat, and hence are often nothing but subjective estimates.

How are event probabilities estimated? In the *frequentist* approach, we use statistical observations: We perform an experiment a large number of times N, and count the number of times that the event A is observed, as N_A. The ratio of the two, $\frac{N_A}{N}$ is then estimated as the probability that event A occurs when the experiment is conducted. This estimate varies with the number of times you run the experiment! Ideally, you would like to conduct an infinite number of experiments, but that is impractical, and may not even give you a consistent answer. In this course, we will describe a theoretical foundation for this approach, which shows that as $N \to \infty$, the ratio approaches the underlying correct probability of the event A.

For experiments that are hard to repeat, a different approach at determining probabilties of outcomes is to apply some subjective beliefs based on principles of "equality" or "nonprejudice": If there is no reason to believe that some events are more probable than others, assume they are equally probable. This approach is typical for games of chance, where we assumed ideal balanced coins, dice, roulette wheels, etc. In these experiments, the number of outcomes is typically finite, and the probability of an event is proportionate to the number of outcomes of the experiment that are in the event. Thus, the probability that a roll of two six-sided dice totals 7 is proportional to the number of possible dice outcomes that total 7.

This course is based on the modern axiomatic theory of probability, espoused by mathematicians such as Andrey Kolmogorov: Treat any experiment as generating outcomes in a set (finite or not): the sample space. Events are subsets of the sample space, and probability is any function that assigns a number in [0,1] to an event in a *consistent* way that must satisfy some intuitively appealing properties that will be discussed later. Thus, in a subjectivist interpretation, we cannot assign arbitrary probabilities to different events (the probability that the Patriots will win the Super Bowl versus the probability that a different team will win the Super Bowl.) However, as long as the probabilities are assigned consistent with the axioms we will present, we can use the foundations of probability theory to analyze and predict outcomes in engineering applications in a rigorous manner.

In essence, probability theory provides us with a "calculus" for representing and reasoning about uncertainty that is consistent with basic axiomatic foundations. Probability Theory is an axiomatic theory that models uncertainty in a consistent manner for predictions and decisions. It allows for the computation of probabilities for compound events, chaining of events, derived events as well as conditional inferencing and information processing.

A common question is how probability is related to statistics. The two sciences are close: Probability often deals with predicting the likelihood of future events, while statistics often involves the analysis of the frequency of past events. They use similar axiomatic foundations, and the above distinction is not exclusive. Statistics focuses on the analysis of past data, collected from experiments that involve uncertainty, and is used to understand the results of experiments: validity of outcomes, typicality, cause-effect relationships, and correlations. It builds models based on observations. Probability provides the calculus that allows models built from statistics to be used for predictions and inferencing about future events.

The distinction is best highlighted by an anecdote: A probabilist and a statistician walk to a craps table. The probabilist sees the pair of dice and thinks: "Six-sided dice? Assume each face of the dice is equally likely to land face up. Now compute the chances that each possible number is rolled and bet accordingly." The statistician thinks: "Those dice may look OK, but how do I know that they are not loaded? I'll watch a while, and keep track of how often each number comes up. Then I can decide if my observations are consistent with the assumption of equal-probability faces. Once I'm confident enough that the dice are fair, I'll ask my friend the probabilist to tell me how to bet."

To paraphrase a quote attributed to Persi Diaconis, a Stanford professor, "the problems considered by probability and statistics are inverse to each other. In probability theory we consider some underlying process which has some randomness or uncertainty modeled by random variables, and we figure out what happens. In statistics we observe something that has happened, and try to figure out what underlying process would explain those observations."

1.1.1 A Brief History of Probability

Probability and games of chance arise in anecdotes in every ancient civilization, from Asia, Europe, Central and South Africa. Problems such as weather prediction were critical in estimating agricultural output and governed the pricing of commerce. Observers of astronomical events used astrology for subjective predictions of important events. In metallurgy, early makers of tools used formal rules to reason about mixtures of metals in alloys as well as heating and quenching times to strengthen their tools and reduce impurities. Arab mathematicians used permutations and combinations to list all possible Arabic words with and without vowels, and used early statistics concepts such as frequency analysis for statistical inference. However, none

of these early civilizations developed a consistent calculus for manipulating uncertainty across compound events.

It is fitting that the foundations of modern probability arose from a gambling dispute. These foundations were articulated in a series of articles between mathematicians Blaise Pascal and Pierre de Fermat in 1654. Their discussion concerned a game of chance involving multiple rounds with two players who have equal chances of winning each round. The players contribute equally to a prize pot, and agree in advance that the first player to win a certain number of rounds will collect the entire prize, say the first to win five games. Now suppose that the game is interrupted by external circumstances before either player has achieved victory, so that player 1 has won 3 games and player 2 has won 2. How does one then divide the pot fairly?

Pascal and Fermat articulated some desired properties of the solution: a player who is closer to winning should get a larger part of the pot. But the problem is not merely one of calculation; it also involves deciding what a "fair" division actually is. In their discussions, Pascal and Fermat provided a convincing, self-consistent solution to this problem, and also developed concepts that are still fundamental to probability theory.

To them, it was clear that a player with a 7–5 lead in a game to 10 has the same chance of eventually winning as a player with a 17–15 lead in a game to 20, so Pascal and Fermat therefore thought that interruption in either situation should lead to the same division of the pot.

Fermat now reasoned thus:[1] if one player needs r more rounds to win and the other needs s, the game will surely have been won by someone after $r + s - 1$ additional rounds. Fermat was thus able to compute the odds for each player to win, simply by writing down a table of all possible continuations and counting how many of them would lead to each player winning. Fermat now considered it obviously fair to divide the stakes in proportion to those odds.

Fermat's solution was improved by Pascal in two ways. First, Pascal produced a more elaborate argument why the resulting division should be considered fair. Second, he showed how to calculate the correct division more efficiently than Fermat's tabular method, using a recursive technique.

Shortly after, encouraged by Pascal, Christiann Huygens published the first book of Probability that used their axiomatic framework[2]. Because of the appeal of games of chance, probability theory soon became popular, and the subject developed rapidly during the 18th century. One of the major contributors during this period was Jacob Bernoulli, who studied games with uneven odds, and whose work[3] led to the law of large numbers and to the definition of stochastic convergence, which was the foundation for the frequentist approach to probability. His analysis of games led to the modern concept of Bernoulli random variables and the binomial distribution.

Later in the 18th century, mathematician Abraham de Moivre developed a technique for approximating binomial coefficients[4]. de Moivre's work led to the development of the Central Limit Theorem, and the use of the Gaussian distribution as a fundamental tool in probability and statistics.

Most of the early work in probability theory focused on games of chance. In 1812 Pierre-Simon, Marquis de Laplace, introduced many new ideas in his book, *Théorie Analytique des Probabilités*. Laplace applied probabilistic ideas to many scientific and practical problems, such as the theory of errors, statistical mechanics and actuarial mathematics. In a subsequent article[5], Laplace set out the principles for Bayesian reasoning and inference, and developed the use of characteristic functions and moment generating functions for estimation of moments of random variables. He also connected the principles of least squares estimation to probabilistic inferencing. In his book, Laplace wrote "We see that the theory of probability is at the bottom only common sense reduced to calculation; ... The most important questions in life are, for the most part, really only problems of probability."

[1] Keith Devlin: The Unfinished Game: Pascal, Fermat, and the Seventeenth-Century Letter that Made the World Modern.
[2] Rekeningh in Spelen van Gluck, translated as "On Reasoning in Games of Chance".
[3] Ars Conjectandi, literally translated as 'art of conjecturing', published after his death in 1713.
[4] "Approximatio ad Summam Terminorum Binomii $(a + b)^n$ in Seriem expansi," in "The Doctrine of Chance's" (1718).
[5] Essai philosophique sur les probabilités (1814).

Another important 19th century contributor was Siméon Denis Poisson, who was a mathematical physicist working on various electromagnetic and optics problems. Poisson published a memoir in 1830, where he discusses the ratio of female births and male births in France using the theory of Laplace and binomial distributions based on Bernoulli's work. Poisson proves the weak law of large numbers first. Then, he considers a different limit where the number of births n grows, but the probability of a female birth p diminishes so that pn is constant. He introduced the Poisson distribution as the limit distribution in this problem.

In Russia, Pafnuty Chebyshev is one of the founding fathers of Russian mathematics. His contributions to Probability Theory in the 19th century are extensive. He is best known for the the Chebyshev inequality[6] that bounds the probability that a random variable with known mean and standard deviation has an outcome that is more than a given number of standard deviations away from the mean, and is used to prove the weak law of large numbers in a general setting. He was also an academic mentor of Andrey Markov, another major contributor to the development of Probability Theory.

Andrey Markov is the developer of the theory of Markov Chains. He rigorously proved extensions of the central limit theorem and the law of large numbers to sequences of dependent random variables, a problem that he started working on with Chebychev. His extensive contributions are reflected by the many modern concepts that bear his name, including the Markov inequality, Markov chains, Markov processes, the Gauss-Markov theorem, and Markov random fields.

In the early 20th century, Ronald Fisher developed the foundations of modern statistical analysis, including maximum likelihood detection, analysis of variance, design of experiments, and Fisher information. He applied his principles to botany and genetics and became well-known as a biostatistician. Many modern techniques in data science and statistics carry his name, including the Fisher's linear discriminant and the Behrens-Fisher distribution.

In 1933, Andrey Kolmogorov published his book, *Foundations of the Theory of Probability*, laying the modern axiomatic foundations of probability theory that we teach today. Subsequently, Kolmogorov extended his work to develop the foundation for estimation, smoothing and prediction for stochastic processes, key techniques that are at the heart of modern navigation systems. In statistics, he is best known for the Kolmogorov-Smirnov test for testing whether a collection of independent samples corresponds to a given distribution for a random variable.

1.1.2 Probability at Boston University's College of Engineering

We briefly mention some of the research areas at Boston University that use probability as its foundations.

In the areas of communications and network systems, probability is used in the modeling of traffic, and in performance analysis of networks. It is also used in the physical layer processes of designing signaling strategies, along with coding and decoding. It provides the foundation for information theory and the design of efficient coding strategies for wired and wireless systems.

Probability is extensively used in the analysis of manufacturing systems and networks. Dynamic modeling of demand and production involves probabilistic principles. Key techniques for quality control and product development involve important concepts from design of experiments and statistics.

In aerospace and robotics systems, probability theory provides the foundations for estimation of the system conditions such as location and orientation using noisy sensors, so that effective control can be applied. For intelligent systems and autonomy, probability provides the foundations for machine learning algorithms for robot vision, classification and situation assessment.

In acoustics, we model propagation of waves through random media using the principles of probability, as well as in sonar signal detection and imaging. Probability also provides the foundation for acoustic imaging.

[6] P. L. Chebyshev, Des valeurs moyennes,J.Math.Pures Appl.(2), 1867.

In space physics, we use probability principles as the foundation for imaging the ionosphere with incoherent scatter radar, for tracking space objects using telescopes and radars, and for representing uncertainty in propagation of light through the atmosphere in adaptive optics.

In biomedical systems, we use probability to model the effectiveness of treatments and procedures, and to assess risks. We also use probability to reduce noise in signals and extract meaningful information from signals and images.

In signal and image processing, probability provides the foundation for signal enhancement, denoising, detection and inference, including some recent work on imaging with single-photon cameras. In photonics and nanotechnology, probability provides the foundation for photon propagation and quantum mechanics.

In computer engineering, probability is used for the analysis of algorithms, and for reliability, fault detection and isolation. Probability is also used for the design of novel algorithms for solving hard combinatorial problems that exploit randomness.

Figure 1.1 shows examples of some of the applications listed above.

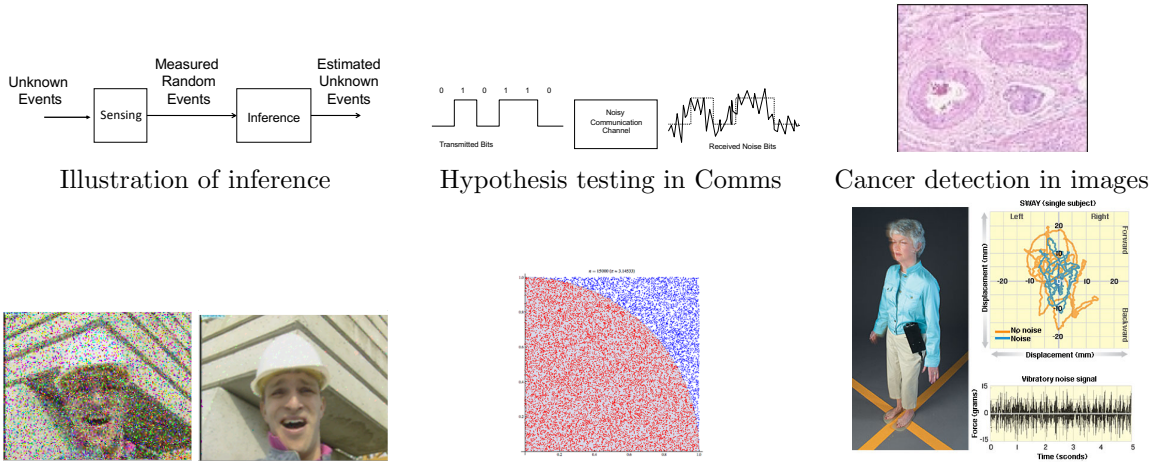


Illustration of inference | Hypothesis testing in Comms | Cancer detection in images

Noise mitigation | Monte Carlo Computation | Noise for stabilization

Figure 1.1: Applications of Probability

1.2 Axioms of Probability

A formal axiomatic theory of probability is necessary to deal with more complex issues such as chaining of events and derived events. At its foundations are fundamental definitions that allow a formulation, along with specific axioms that are accepted without proof as needed by the theory. What follows in the theory of probability are theorems, propositions and lemmas that are consequences of the axioms and definitions and allow the application of probability.

We begin with a review of set theory, which forms the mathematical basis for much of the axioms of probability theory.

1.2.1 Set Theory

Definition 1.1
A set is a collection of elements.

Elements can be anything you like: numbers, letters, people, movies, combinations of items, etc. We usually use capital letters (e.g. A) to denote sets, and lower case letters (e.g. e, x) to denote elements of the set.

A set can be empty, which is called the null set, also denoted by the symbol $\emptyset$. The collection can have a finite number of elements, a countably infinite number of elements, or an uncountable number of elements (e.g. an interval of the real line.) There are several ways to define a set, including

- List its elements: $A = \{1, 3, 5, 7\}$.
- Give a rule in words to generate the set: $A = \{\text{odd integers greater than } 2\}$.
- Give a rule using mathematical symbols: $A = \{x \text{ integer} : x > 2\}$.

In the last version, we use the variable x, and the colon ":" is used as a shortcut for the expression "such that". Hence the last rule reads: A is the set of all numbers such that the number is greater than 2 and the number is an integer. We refer to this version as "set-builder notation."

We use the following notation throughout this text:

- $x \in B$ means that "x is an element of the set B".
- $x \notin B$ means that "x is not an element of the set B".
- The **empty set** or **null set** is the set with no elements. Notation: $\emptyset$ or $\{\ \}$.
- We denote by Ω the **universal set**, i.e. the set of all possible elements.
- A **subset** A of the set B, denoted as $A \subset B$, is a collection of some (or none) of the elements that are in B.
- Two sets are **equal** if $A \subset B$ and $B \subset A$. Thus, the two sets contain the same elements.

A *Venn Diagram* can be used to illustrate relationships between sets. For instance, the figures in 1.2 illustrate Venn diagrams for different set operations. Understanding set operations is much easier if you can visualize the operations using a Venn diagram.

On sets, we define elementary set operations:

- Set complement $A^c = \{x : x \in \Omega \text{ and } x \notin A\}$. Note that $(A^c)^c = A$.
- Set union $A \cup B = \{x \in \Omega : x \in A \text{ or } x \in B\}$. This is sometimes written as $A + B$.
- Set intersection $A \cap B = \{x \in \Omega : x \in A \text{ and } x \in B\}$. This is sometimes written as $A \cdot B$.
- Set Difference $A - B = \{x \in \Omega : x \in A \text{ and } x \notin B\}$. Note $A - B = A \cap B^c$.

These operations are illustrated in 1.2.

Below are other important set concepts that we use in the course:

- A and B are disjoint, or mutually exclusive, sets if and only if $A \cap B = \emptyset$.
- A finite collection of sets $A_1, \ldots, A_n$ are *mutually exclusive* if and only if $A_i \cap A_j = \emptyset$ for any $i \neq j \in \{1, \ldots, n\}$.
- A finite collection of sets $A_1, \ldots, A_n$ is *collectively exhaustive* in Ω if and only if $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$.

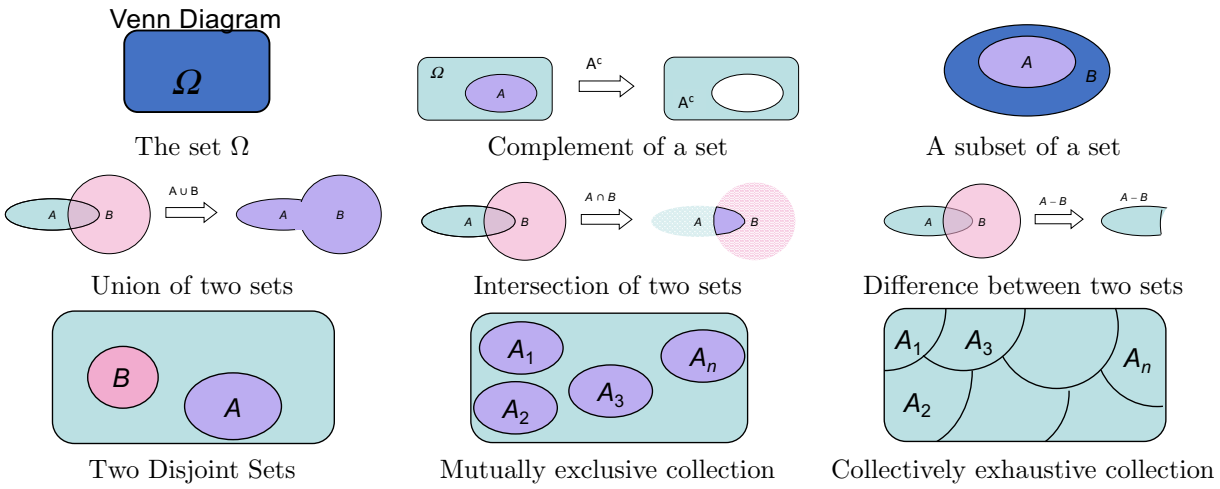


Figure 1.2: Illustration of Set Operations and Concepts

- A countable collection of sets $A_1, A_2, \ldots$, is *mutually exclusive* in Ω if and only if $A_i \cap A_j = \emptyset$ for all $i, j \in \{1, 2, \ldots\}$.

- A countable collection of sets $A_1, A_2, \ldots$, is *collectively exhaustive* in Ω if and only if $A_1 \cup A_2 \cup \ldots = \Omega$.

- A finite or countable collection of sets is a **partition** if it is both mutually exclusive and collectively exhaustive.

From the above definitions, there are several results that can be derived, known as De Morgan's Theorems. The proof of these is obvious; Figure 1.2.1 illustrates the proof of the first result.

- $(A \cup B)^c = A^c \cap B^c$. That is, an element that is not in (A or B) must be (not in A) and (not in b).

- $(A_1 \cup A_2 \cup A_3 \cup \ldots)^c = A_1^c \cap A_2^c \cap A_3^c \cap \ldots$

- $(A_1 \cap A_2 \cap A_3 \cap \ldots)^c = A_1^c \cup A_2^c \cup A_3^c \cup \ldots$

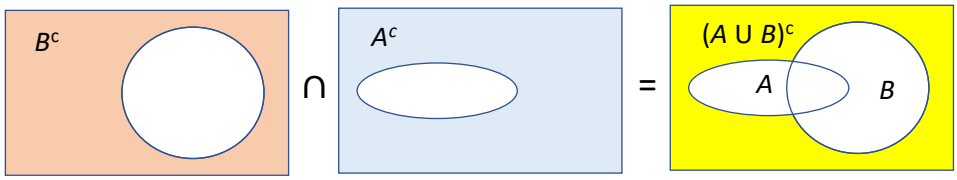


Figure 1.3: Illustration of De Morgan's First Theorem.

To complete this section, we review some mathematical notation that we use throughout these notes. We use the symbol $\forall$ to denote *for all*. Hence $\forall x \in A$ means for all elements x of the set A. The existential qualifier $\exists$ is used to denote that there exists an element. Hence $\exists x \in A$ means that there exists at least one element x that belongs to A. The negative of there exists is denoted $\nexists$.

1.2.2 Probability Axioms

The basic model for probability begins with the concept of a random experiment: An **experiment** is a procedure that generates an observable outcome. An **outcome** is a possible observation of an experiment. The **sample space** Ω of an experiment is the set of all possible outcomes ω of the experiment. Each possible **outcome** ω is an element of the sample space Ω.

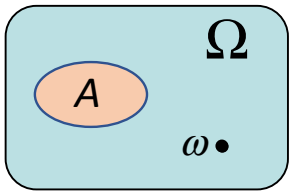


Figure 1.4: Illustration of outcomes ω and events A.

An **event** is subset of Ω: that is, a collection of outcomes. Note that an event may contain a single element, or be the empty set, or be all of Ω. An event is something we will assign probability to; in our axiomatic theory, we assign probabilities to events, not outcomes. Note that it is possible that not every subset of Ω is an event, as we will explain later.

Example 1.1
Experiment: roll a normal six-sided die once. An outcome is the number that shows up on top of the die, which is in $\{1,2,3,4,5,6\}$. The sample space $\Omega = \{1,2,3,4,5,6\}$. Examples of events are $E_1 = \{1,3,5\}$, the set of all odd outcomes; E_2 = set of all outcomes greater than 2 ($\{3,4,5,6\}$); and E_3 = set of all outcomes that are the square of an integer ($\{1,4\}$).

Example 1.2
Experiment: Perform 2 rolls of a quadrilateral (four-sided) die, record both numbers. An outcome is the ordered pair of numbers: $\{1,2,3,4\} \times \{1,2,3,4\}$, so we have Ω consisting of 16 ordered pairs. Examples of events are $E_1\{(1,3),(2,2),(3,1)\}$, the set of all outcomes where the two numbers sum to 4; E_2 = the set of all outcomes that sum to an odd number ($\{(1,2),(1,4),(2,1),(2,3),(3,2),(3,4),(4,1),(4,3)\}$.)

Example 1.3
Experiment: Go to the Green Line station on St. Mary's Street and Commonwealth Avenue, going West and wait for the train to arrive. The outcome is the number of minutes (as a real number) before the train arrives. Hence, an outcome is a number x in the sample space $\Omega = [0,\infty)$. Examples of events are $E_1 = \{\text{train arrives under five minutes}\} = \{\omega : \omega < 5\}$; and $E_2 = \{\text{train arrives in more than 20 minutes}\} = \{\omega : \omega > 20\}$.

Example 1.4
Experiment: Measure the arrival time of a pulse, arriving at a random time in the interval $[0,T]$. An outcome is the time of arrival, namely a number $t \in [0,T]$. The sample space $\Omega = [0,T]$ contains an uncountable number of outcomes. Examples of events are $E_1 = \{\omega = T/2\}$ which contains a single outcome, or $E_2 = \{\omega : 0 < a < \omega < b < T\}$ that contains an interval of outcomes.

Example 1.5
As an experiment, pick a point in the unit square $[0,1] \times [0,1]$. An outcome ω is the ordered pair consisting of the coordinates of the point, namely a pair $(x,y) \in [0,1] \times [0,1]$. The sample space Ω is an uncountable set of ordered pairs $\Omega = \{(x,y) \in [0,1] \times [0,1]\}$. Examples of events are $E_1 = \{(1/2, 2/3)\}$ which contains a single outcome, or $E_2 = \{(x,y) \in [0,1]^2 : x + y \leq 0.2\}$ that contains a region of outcomes.

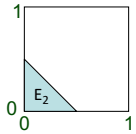


Figure 1.5: Event E_2 in ex. 1.5.

Note that there is a correspondence between the terminology of set theory and that of the probability axioms. We highlight the correspondence in the table below:

Set theory		Probability theory
Universal set	$\Longleftrightarrow$	Sample Space
Element	$\Longleftrightarrow$	Outcome
Subset	$\Longleftrightarrow$	Event

Let's define the collection of all events in an experiment as an event space $\mathcal{E}$. As we highlighted before, we may not want to define every subset of Ω as an event. Since events are sets A for which we want to compute the probability that an outcome is A, there are certain properties that the space of all events must have. We list them below.

Definition 1.2
The event space $\mathcal{E}$ is a collection of subsets of Ω which satisfies the following axioms:

1. $\Omega \in \mathcal{E}$. Thus, the sample space Ω is an event, and the probability that an outcome is in Ω should be 1.
2. If $A \in \mathcal{E}$, then $A^c \in \mathcal{E}$. The complement of an event is also an event, because we want to assign probability to the event that the outcome is not in A.
3. If $A_i \in \mathcal{E}, i = 1, 2, \ldots$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{E}$. The union of a countable number of events is an event. This implies that the space $\mathcal{E}$ is closed under the operation of countable unions.

Note that, by the properties of set theory, this implies that $\emptyset \in \mathcal{E}$, because $\Omega^c = \emptyset$. It also implies that the intersection of two events A, B is also an event $A \cap B$, because $A \cap B = (A^c \cup B^c)^c$ using De Morgan's theorems. Also, the set $A - B \equiv \{\omega \in A | \omega \notin B\} \in \mathcal{E}$, because $A - B = A \cap B^c$, which is the intersection of two elements of $\mathcal{E}$.

In terms of assigning probabilities, we will only consider events that are in the event space $\mathcal{E}$. This avoids some mathematical pitfalls that can arise if we consider $\mathcal{E}$ to be all of the subsets of Ω. We call a set *countable* if its elements can be indexed by the natural numbers $0, 1, 2, \ldots$ When the set Ω is countable, we can simply let $\mathcal{E}$ be the collection of all subsets of Ω, as these mathematical difficulties only arise for sets with uncountable numbers of outcomes, such as an interval of real numbers.

The event space $\mathcal{E}$ is often called a **σ-field** (or σ-algebra) in mathematics because it contains Ω, it is closed under countable unions and complementation. In many cases, we construct the set of events $\mathcal{E}$ by specifying some of the basic events that we want to compute probabilities for, and then finding the smallest collection of events that contains the basic events, and is closed under countable unions and complementation.

Example 1.6
Flip 2 coins, a penny and a dime. $\Omega = \{hh, ht, th, tt\}$, with 4 outcomes.

Events of interest $E_i = \{$outcomes with i heads$\}$. Thus, $E_0 = \{tt\}$ contains 1 outcome; $E_1 = \{ht, th\}$ contains 2 outcomes.

What is the smallest event space $\mathcal{E}$ that contains these events? It is $\mathcal{E} = \{\emptyset, E_0, E_1, E_2, E_0 \cup E_1, E_0 \cup E_2, E_1 \cup E_2, \Omega\}$.

Note that $\mathcal{E}$ contains the union of any collections of events, and the complement of each event! However, there are only 8 elements in $\mathcal{E}$, whereas the total number of subsets of Ω is 16. Thus, subsets such as $\{ht\}$ are not events in this event space.

Example 1.7
Consider an experiment consisting of selecting a real number in the interval $[0, 1]$. Consider as events of interest sets of the form $(a, b), a, b \in [0, 1]$. We can define the event space $\mathcal{E}$ as the smallest σ-field that contains these open intervals as events. Note that $\mathcal{E}$ contains the set with two points $\{0, 1\}$ because it is the complement of $(0, 1)$. With further thought, we realize that $\mathcal{E}$ will contain every closed interval $[a, b]$, $a, b \in [0, 1]$, as well as many other events of interest.

An event $A \in \mathcal{E}$ is called an **atom** if it contains only a single outcome; atoms are events of the form $A = \{\omega\}$ for some $\omega \in \Omega$. Events A_i indexed by a set I are called *mutually exclusive* if $A_i \cap A_j = \emptyset$ for all $i, j \in I, i \neq j$. Note that this index set can be infinite in the definition.

We have thus far defined two key components of the axioms of probability: the sample space Ω, which is a collection of outcomes, and the event space $\mathcal{E}$, which is a σ-field collection of subsets of Ω. The third component we need is a *probability measure* $\mathbb{P}$ that assigns a probability value in $[0, 1]$ to each event contained in $\mathcal{E}$; that is, it maps the set of events into the closed unit interval $[0, 1]$. This probability measure $\mathbb{P}[A]$ is interpreted as the probability that the outcome of the experiment is contained in the event $A \in \mathcal{E}$.

The axioms which a probability measure must satisfy are:

1. (**Non-negativity:**) For any event $A \in \mathcal{E}$, $\mathbb{P}[A] \geq 0$ Probabilities are non-negative.
2. (**Normalization:**) $\mathbb{P}[\Omega] = 1$. The probability that we generate an outcome in Ω is one.

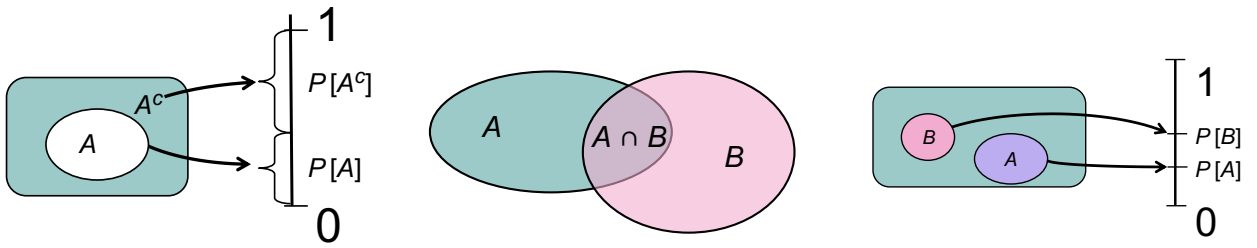
3. (**Countable Additivity**) For any countable collection of mutually exclusive events $A_i, i = 1, 2, \ldots$, we have $\mathbb{P}[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$.

Example 1.8
Consider a sample space $\Omega = [0, 1]$, the unit interval. How do we define events in this space? If we let $\mathcal{E} = \{A : A \subset \Omega\}$, this may become too large a set, with too many elements, such that it is difficult to define a probability measure on these events that satisfy the above axioms.

For instance, assume we want to define a probability measure so that all points have the same probability. For $\omega \in \Omega$, then $\mathbb{P}[\{\omega\}]$ must be a constant. However, that constant must be zero, because otherwise we could find an infinite number of disjoint sets that we could add and get a subset of Ω with infinite probability! Thus, knowing the probability of individual outcomes does not help us in defining a probability measure.

However, consider a different set of events, $E_{a,b} = \{0 \geq a < \omega < b \leq 1\}$. For this interval, we can easily assign a probability measure corresponding to the length of the interval, so that $\mathbb{P}[E_{a,b}] = b - a$. Define the event space $\mathcal{E}$ as the smallest σ-field that contains all the intervals $E_{a,b}$ and is closed under countable unions and complementations: this is known as the *Borel σ-field. Note that every event $A \in \mathcal{E}$ can be written in terms of countable unions and complements of intervals, for which we know how to compute the probability measure. We can extend the measure $\mathbb{P}[A]$ to all elements in $\mathcal{E}$ using the axioms of probability, including the countable additivity axiom. We will show that the countable additivity axiom implies that the probability measure is continuous, and hence we can extend the definition on open intervals to apply to all intervals, and to countable unions and intersections of intervals.*

We are now ready to define a probability space. A *probability space* is a triple $(\Omega, \mathcal{E}, \mathbb{P})$ which is used to describe the outcomes of a random experiment. The set Ω is the set of all possible elementary experiment outcomes ω. The set $\mathcal{E}$ is a $\sigma-$field of events that are subsets of Ω and satisfy the properties of event spaces. The probability measure $\mathbb{P} : \mathcal{E} \to [0, 1]$ satisfies the axioms of probability measures.



$\mathbb{P}[A] + \mathbb{P}[A^c] = 1$ $\quad$ $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$ $\quad$ If Disjoint, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$

Figure 1.6: Illustration of Probability Concepts

Based on the above definition, probability measures can be shown to satisfy additional properties, discussed below. We show brief proofs of selected properties to illustrate how to use the basic properties of probability measures to compute probabilities.

1. $\mathbb{P}[A] = 1 - \mathbb{P}[A^c]$. This follows because A and A^c are mutually exclusive, and $A \cup A^c = \Omega$, so $\mathbb{P}[A] + \mathbb{P}[A^c] = \mathbb{P}[\Omega] = 1$.

2. $\mathbb{P}[\emptyset] = 0$.

3. For any finite collection $A_1, A_2, \ldots, A_n$ of mutually exclusive events,

$$\mathbb{P}[\cup_{i=1}^{n} A_i] = \sum_{i=1}^{n} \mathbb{P}[A_i].$$

4. $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$. This follows because $A \cup B = A \cup (B - A)$, and $A, B - A$ are mutually exclusive. Hence, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B - A]$. Furthermore, $B = (B - A) \cup (A \cap B)$, and these two sets are mutually exclusive. Thus, $\mathbb{P}[B] = \mathbb{P}[B - A] + \mathbb{P}[A \cap B]$. Hence, $m[B - A] = \mathbb{P}[B] - \mathbb{P}[A \cap B]$. Substituting into the first equation yields the result.

5. If $B \subset A$, then $\mathbb{P}[B] \leq \mathbb{P}[A]$ and $\mathbb{P}[B] + \mathbb{P}[A - B] = \mathbb{P}[A]$.
6. $\mathbb{P}[A \cup B \cup C) = \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[C] - \mathbb{P}[A \cap B] - \mathbb{P}[A \cap C] - \mathbb{P}[B \cap C] + \mathbb{P}[A \cap B \cap C]$;
7. If $A_i, \ldots, A_n$ are mutually exclusive events, $\mathbb{P}[A_i] = \sum_{i=1}^{n} \mathbb{P}[A_i]$.
8. If $A_i, \ldots, A_n$ are mutually exclusive events and $\cup_{i=1}^{n} A_i = \Omega$, then $\sum_{i=1}^{n} \mathbb{P}[A_i] = 1$.
9. If $A_1, A_2, A_3, \ldots$ are mutually exclusive events and $\cup_{i=1}^{\infty} A_i = \Omega$, then for any event A, $\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \cap A_i]$.

This last result is important because it allows us to compute the probability of a complex event as the sum of probabilities of simpler, mutually exclusive events. That is the cornerstone of how most probabilities of events are computed: we break down the events into mutually exclusive subsets for which the probabilities are known, and we use the countable additivity property.

Example 1.9
Consider the example of a shuffle of a deck of cards. The sample space Ω consists of the possible orderings (permutations of 52 cards). While there are many outcomes, there is still a finite number of them. Thus we can make the event space $\mathcal{E}$ the set of all subsets of Ω. Assuming that all permutations are equally likely, the probability measure $\mathbb{P}[A]$ can be defined to be proportional to the number of outcomes in A. For instance, consider the event A consisting of all outcomes where the first card in the deck is the ace of spades. The number of outcomes in A is 51! (the first card is the ace of spades, the other 51 can be in any order), and the total number of elements in Ω is 52!. Hence, $\mathbb{P}[A] = \frac{1}{52}$.

Example 1.10
Consider the toss of a fair coin, with outcomes H, T. The set of outcomes $\Omega = \{H, T\}$. The σ-field $\mathcal{E}$ is

$$\mathcal{F} = \{\{H\}, \{T\}, \emptyset, \{H, T\}\}$$

If the coin is fair, the measure $\mathbb{P}$ will have the following properties:

$$\mathbb{P}[\{H\})] = \frac{1}{2}; \mathbb{P}[\{T\}] = \frac{1}{2}; \mathbb{P}[\{H, T\}] = 1; \mathbb{P}[\emptyset] = 0;$$

One of the important properties of probability measures is the continuity of probability, in the sense specified below. If we have a sequence $A_1 \subset A_2 \subset \cdots$ of increasing events in $\mathcal{E}$, the sequence A_j is monotone increasing and converging to the union $\cup_{i=1}^{\infty} A_i$. Will the probabilities converge also? They will; they are an increasing sequence $\mathbb{P}[A_i]$ of real numbers that are bounded above by one. This allows us to define probability measures on events that can be expressed as limits of events, as shown in the following lemma.

Lemma 1.1
Suppose $A_1, A_2, \ldots$ is a sequence of events in $\mathcal{E}$. Then,

1. If $A_1 \subset A_2 \subset \cdots$, then $\cup_{k=1}^{\infty} A_k \in \mathcal{E}$, $\lim_{k\to\infty} \mathbb{P}[A_k]$ exists, and one defines $\mathbb{P}[\cup_{k=1}^{\infty} A_k] = \lim_{k\to\infty} \mathbb{P}[A_k]$.
2. If $A_1 \supset A_2 \supset \cdots$, then $\cap_{k=1}^{\infty} A_k \in \mathcal{E}, \lim_{k\to\infty} \mathbb{P}[A_k]$ exists, and one defines $\lim_{k\to\infty} \mathbb{P}[A_k] = \mathbb{P}[\cap_{k=1}^{\infty} A_k]$.

proof For the first part, note that $\cup_{k=1}^{\infty} A_k$ is a countable union of events, and hence it is also an event, because event spaces are closed under countable unions and complementation. Let $D_1 = A_1, D_k = A_k - A_{k-1}, k \geq 2$. Note that $D_k \in \mathcal{E}$, because $D_k = A_k \cap A_{k-1}^c$ and intersections of events are also events. Furthermore, the the collection $D_1, D_2, D_3, \cdots$ is mutually exclusive. Then, by the countable additivity axiom of probability,

$$\mathbb{P}[A_k] = \mathbb{P}[\cup_{j=1}^{k} A_j] = \mathbb{P}[\cup_{j=1}^{k} D_j] = \sum_{j=1}^{k} \mathbb{P}[D_j],$$

and thus is an increasing sequence of numbers. Since $\mathbb{P}[A_k]$ is bounded by 1, the monotone convergence theorem guarantees it has a limit that is a number less than or equal to 1, so $\lim_{k\to\infty} \mathbb{P}[A_k]$ exists and is a probability. Thus, the probability of the event $\cup_{k=1}^{\infty} A_k$ is well-defined, as

$$\mathbb{P}[\cup_{i=1}^{\infty} A_k] = \lim_{k\to\infty} \mathbb{P}[A_k] = \sum_{j=1}^{\infty} \mathbb{P}[D_j].$$

For the second part, consider the sets $B_k = A_k^c$. Then, $\mathbb{P}[A_k] = 1 - \mathbb{P}[B_k]$. By the first part, we know $\cup_{i=1}^{\infty} B_k$ is an event, and that $\lim_{k\to\infty} \mathbb{P}[B_k]$ exists and is a probability, and that we define $\mathbb{P}[\cup_{i=1}^{\infty} B_k] = \lim_{k\to\infty} \mathbb{P}[B_k]$.

Now, note $\cap_{k=1}^{\infty} A_k = (\cup_{i=1}^{\infty} B_k)^c$, so it is also an event. Since it is the complement of an event,

$$\mathbb{P}[\cap_{k=1}^{\infty} A_k] = 1 - \mathbb{P}[\cup_{i=1}^{\infty} B_k] = 1 - \lim_{k\to\infty} \mathbb{P}[B_k] = \lim_{k\to\infty} (1 - \mathbb{P}[B_k]) = \lim_{k\to\infty} \mathbb{P}[A_k].$$

Example 1.11
Consider $\Omega = [0, 1]$, the unit interval. Let the set $\mathcal{E}$ be the Borel σ-field in Ω. Note that not all subsets of Ω will be Borel sets, although every interesting subsets we care about is likely to be Borel sets. For an open interval (a, b), we define the measure as its length:

$$\mathbb{P}[(a, b)] = b - a.$$

We can now use the axioms of probability to extend this definition to all Borel sets. It should be easy to see that every Borel set can be written as a countable union of intervals (closed or open, so that a set with only one element x can be written as the interval $[x, x]$). By lemma 1.1, we can now extend the measure $\mathbb{P}[A]$ to compute this uniquely using a limit process.

Why is the concept of event needed over and above the concept of outcome? There are many situations where we want to model the set of possible outcomes as continuous, rather than discrete. In those situations, we know that there are at most a finite number of mutually exclusive events that have probability at least ϵ. By defining probability measures on events, we are able to focus on a finite number of significant events instead of an uncountable number of outcomes. A Furthermore, not every subset of Ω can be considered an event, because it may be impossible to construct a probability measure satisfying the probability axioms. If you are interested in this topic, we show an example in Appendix B of a space with some subsets for which we cannot define a consistent probability measure that satisfy the probability axioms.

In many applications, we define the event space $\mathcal{E}$ by defining a collection of basic events for which we want to compute probabilities, and then finding the smallest σ-field that contains those events. By smallest, we mean the following: A σ-field $\mathcal{E}'$ is said to be a refinement of $\mathcal{E}$ (written as $\mathcal{E} \subset \mathcal{E}'$), if and only if, for any event $A \in \mathcal{E}$, said event is also $A \in \mathcal{E}'$. The smallest or coarsest σ-field that contains a collection of events $\{A_i\}$ is denoted as $\sigma(\{A_i\})$, and is not a refinement of any other σ-field that contains the collection of events $\{A_i\}$. We used this approach to define Borel sets over the unit interval, where the A_i were open intervals in $[0, 1]$. The definition of Borel sets can be generalized to the real line, or n-dimensional Euclidean spaces, or to many other spaces.

As a final note, in any probability space, there can be events which have no probability of occurring. Thus, the difference between two events is often negligible; in such cases, we would like to define a notion of equivalence of events. Two events $A, B \in \mathcal{E}$ are said to be equal with probability one if and only if $\mathbb{P}[A \cup B - A \cap B] = 0$.

The axiomatic theory of probability highlights the approach we need to compute the probability of any event of interest in applications. We outline the steps below, and then proceed to apply to solve probability questions in several examples:

- Identify the sample space from experiment description (the set of all outcomes).
- Describe probability law on events (atoms if finite).
- Identify event of interest
- Calculate the probability of this event as follows:
  - Partition the event of interest into disjoint events for which the probability measures are known.
  - Use axioms of probability to combine the disjoint event probabilities.

Example 1.12
Consider the experiment as one roll of a six-sided die, with balanced outcomes. In this experiment, $\Omega = \{1, \ldots, 6\}$. The problem is to compute the probability of getting an odd outcome (E_1), and the probability of getting an outcome greater than 2 (E_2).

Since we assume the die is balanced, we know $\mathbb{P}[\{\omega\}] = 1/6$, for $\omega \in \Omega$.

We identify $E_1 = \{1, 3, 5\} = \{1\} \cup \{3\} \cup \{5\}$. Given this disjoint decomposition,

$$\mathbb{P}[E_1] = \mathbb{P}[\{1\}] + \mathbb{P}[\{3\}] + \mathbb{P}[\{5\} = \frac{3}{6} = \frac{1}{2}$$

Similarly, $E_2 = \{3, 4, 5, 6\} = \{3\} \cup \{4\} \cup \{5\} \cup \{6\}$ which is another disjoint decomposition, so

$$\mathbb{P}[E_2] = \mathbb{P}[\{3\}] + \mathbb{P}[\{4\}] + \mathbb{P}[\{5\}] + \mathbb{P}[\{6\}] = \frac{2}{3}.$$

Example 1.13
Consider the experiment of 2 rolls of a quadrilateral (four-sided) die, record both numbers and their order. The sample space $\Omega = \{1, 2, 3, 4\}x\{1, 2, 3, 4\}$. We are asked to compute the probability of the following events: E_1 is the set of outcomes that sum to 5. E_2 is the set of outcomes that sum to a prime number not divisible by 3 or 5. E_3 is the set of outcomes such that the smallest of the two numbers is 2.

Proceeding as before, every ordered pair has an equal probability of occurring, which is $1/16$. Now,

$$E_1 = \{(1, 4), (2, 3), (3, 2), (4, 1)\} = \{(1, 4)\} \cup \{(2, 3)\} \cup \{(3, 2)\} \cup \{(4, 1)\}.$$

$$E_2 = \{(1, 1), (3, 4), (4, 3)\} = \{(1, 1)\} \cup \{(3, 4)\} \cup \{(4, 3)\}.$$

$$E_3 = \{(2, 2), (2, 3), (2, 4), (3, 2), (4, 2)\} = \{(2, 2)\} \cup \{(2, 3)\} \cup \{(2, 4)\} \cup \{(3, 2)\} \cup \{(4, 2)\}.$$

Thus,

$$\mathbb{P}[E_1] = \mathbb{P}[\{(1, 4)\}] + \mathbb{P}(\{(2, 3)\}] + \mathbb{P}(\{(3, 2)\}] + \mathbb{P}[\{(4, 1)\}] = \frac{4}{16} = \frac{1}{4}$$

$$\mathbb{P}[E_2) = \mathbb{P}[\{(1, 1)\}] + \mathbb{P}[\{(3, 4)\}] + \mathbb{P}[\{(4, 3))\}] = \frac{3}{16}$$

$$\mathbb{P}[E_3) = \mathbb{P}[\{(2, 2)\}] + \mathbb{P}[\{(2, 3)\}] + \mathbb{P}[\{(2, 4)\}] + \mathbb{P}[\{(3, 2)\}] + m[\{(4, 2)\}] = \frac{5}{16}$$

What about computing $\mathbb{P}[E_1 \cap E_3]$? Note $E_1 \cap E_3 = \{(2, 3), (3, 2)\}$ so $\mathbb{P}[E_1 \cap E_3] = \frac{2}{16}$.

Note that in the above examples, we have used symmetry of the measure $\mathbb{P}$ to simplify computations. The next two examples describe a more complex experiment.

Example 1.14
Our experiment consists of generating telephone calls. Calls can be long or brief, and can voice or data. Thus, an outcome BV denotes a brief voice call, and LD denotes a long data call. The set of outcomes $\Omega = \{LV, LD, BV, BD\}$. The event set $\mathcal{E}$ is the set of all subsets of Ω. Assume we know the following: the probability of a long voice call is 0.35, the probability of a voice call is 0.7, and the probability of a long call is 0.6. What is the probability of a brief data call? What is the probability of a brief voice call? What is the probability of a long data call?

We use the axioms of probability:

$$\mathbb{P}[\{LV, LD\}] = 0.6 = \mathbb{P}[\{LV\}] + \mathbb{P}[\{LD\}] = 0.35 + \mathbb{P}[\{LD\}] \Rightarrow \mathbb{P}[\{LD\}) = 0.25$$

$$\mathbb{P}[\{LV, BV\}] = 0.7 = \mathbb{P}[\{LV\}] + \mathbb{P}[\{BV\}] = 0.35 + \mathbb{P}[\{BV\}] \Rightarrow \mathbb{P}[\{BV\}] = 0.35$$

$$\mathbb{P}[\{BD\}) = 1 - (\mathbb{P}[\{LV\}]] + \mathbb{P}[\{BV\}] + \mathbb{P}[\{LD\}]) = 0.05$$

Example 1.15
Assume we have 2 factories, making the same part. However, they have different quality control. The probability that a part from factory 1 is OK is 0.9. The probability that a part from factory 2 is OK is 0.8.

Assume that factory 1 makes 70% of the parts that are sold, and factory 2 makes 30% of the parts. The outcome from the experiment is a part selected at the store, which will be either good (G) or bad (B), and will come from factory 1 or factory 2.

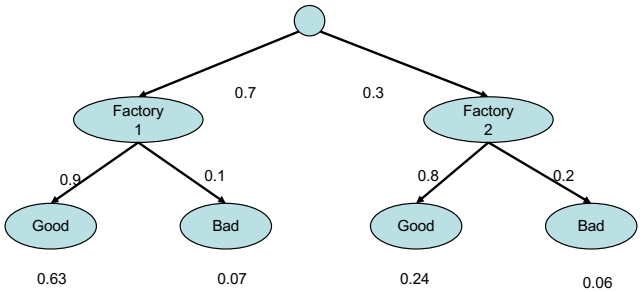


Figure 1.7: Example 1.15.

Given the above description, we can compute the probability of each of the atoms as follows: $\mathbb{P}[\{1G\}] = 0.7{*}0.9 = 0.63$, namely, the probability that the part comes from factory 1 times the probability that the part is good given it comes from factory 1. Similarly, $\mathbb{P}[\{1B\}] = 0.07, \mathbb{P}[\{2G\}] = 0.24, \mathbb{P}[\{2B\}] = 0.06$.

The sample space $\Omega = \{1G, 1B, 2G, 2B\}$. We can view the experiment graphically as shown in Figure 1.15.

What is the probability that the part you buy is good?

$$\mathbb{P}[\{1G, 2G\}] = \mathbb{P}[\{1G\}] + \mathbb{P}[\{2G\}] = 0.87$$

The above example illustrates how our probability calculus allows us to define complex compound experiments.

Note the following: if the sample Ω has a finite number of outcomes $\omega_1, \omega_2, \ldots, \omega_n$, we usually take the event space $\mathcal{E}$ to be the set of all subsets of Ω. In this case, we have the finest disjoint partition of Ω as $\Omega = \{\omega_1\} \cup \{\omega_2\} \cup \cdot \cup \omega_n\}$, and so we can define the probability measure $\mathbb{P}[\cdot]$ on any event by defining it on the atoms $\mathbb{P}[\{\omega_k\}], k = 1, \ldots, n$. In this case, to compute the probability of an event such as $A = \{\omega_1, \omega_3, \omega_5\}$, then we can recognize that A is the union of disjoint sets $\{\omega_1\}, \{\omega_3\}, \{\omega_5\}$, so

$$\mathbb{P}[A] = \mathbb{P}[\{\omega_1\}] + \mathbb{P}[\{\omega_3\}] + \mathbb{P}[\{\omega_5\}].$$

Example 1.16
Consider the following experiment: Ask a person which of the following cities they prefer to live in: Boston, Chicago, Los Angeles, New York, San Francisco. The answer is the outcome, which we denote in shorthand as $S = \{bo, ch, la, ny, sf\}$.

Since it is a finite space, we can assign probabilities to each atom containing a single outcome, as

$$\mathbb{P}[\{bo\}] = \frac{1}{3}; \mathbb{P}[\{sf\}] = \frac{1}{4}; \mathbb{P}[\{ch\}] = \frac{1}{6}; \mathbb{P}[\{la\}] = \frac{1}{8}; \mathbb{P}[\{ny\}] = \frac{1}{8}.$$

Consider the event $A = \{$ east coast city$\}$. Then, $A = \{bo, ny\}$. Using a disjoint decomposition,

$$\mathbb{P}[A] = \mathbb{P}[\{bo\}] + \mathbb{P}[\{ny\}] = \frac{1}{3} + \frac{1}{8} = \frac{11}{24}.$$

Consider the event $B = \{westcoastcity\} = \{sf, la\}$. Then, $\mathbb{P}[B] = \mathbb{P}[\{la\}] + \mathbb{P}[\{sf\}] = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$.

1.3 Conditional Probability and Independence of Events

Conditional probability is the foundation of many engineering applications, especially those involving inference and decision making. Examples of these involve deciding whether radar measurements correspond to the reflections from an aircraft, whether the observation of symptoms indicate the probability that a patient has a disease, and similar questions. In addition, conditional probability is useful for describing complex probability models, such as experiments where the outcome depends on conditions of earlier outcomes. For instance, if you are taking the SATs online, the next question that appears depends on whether you answer the current question correctly or not.

Consider a probability space, and a pair of events $A, B \in \mathcal{E}$ such that $\mathbb{P}[B) > 0$. We define the conditional probability of event A given that B has occurred as:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B)]}{\mathbb{P}[B]}. \tag{1.1}$$

Note that this is not defined when $\mathbb{P}[B)] = 0$. Such events have no probability of being observed in practice, which leads to the lack of a definition. Intuitively, we think of conditioning on event B as **restricting** the universe of possible outcomes to those in B. Hence, only outcomes in $A \cap B$ are now possible out of those in A. Furthermore, we need to **rescale** or normalize so that the conditional probability satisfies the normalization axiom: $\mathbb{P}[B|B] = 1$, which requires that we divide by $\mathbb{P}[B]$.

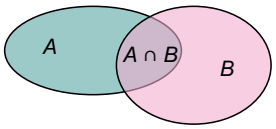


Figure 1.8: Conditional probability.

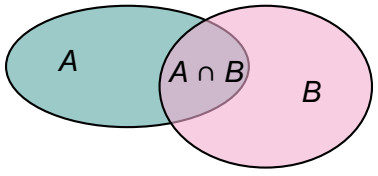
Note the following important relationships:

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A - B] + \mathbb{P}[A \cap B]}$$
$$\mathbb{P}[A \cap B] = \mathbb{P}[B|A]\mathbb{P}[A] = \mathbb{P}[A|B]\mathbb{P}[B]$$

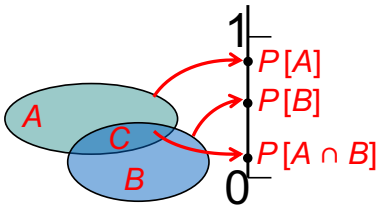
We can extend this to n events $A_1, A_2, \ldots, A_n$ recursively, as:

$$\mathbb{P}[\cap_{k=1}^{n} A_k] = \mathbb{P}[A_1]\mathbb{P}[A_2|A_1]\mathbb{P}[A_3|A_1 \cap A_2] \cdots \mathbb{P}[A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}].$$

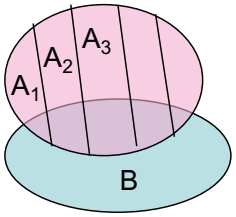
Note that this assumes $\mathbb{P}[A_1 \cap A_2 \cap \cdot \cap A_{n-1}] > 0$ so that conditional probabilities are defined.



$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$ Independence:



$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$



$\mathbb{P}[A|B] = \sum \mathbb{P}[A_i|B]$ for A_i partition A

Figure 1.9: Illustration of Conditional Probability Concepts

Conditional probability functions have an interesting property: they are also probability measures, and a conditional probability space can be defined! Hence, one way of understanding conditional probability is in terms of two operations: Restrict the set of outcomes to B, and compute the relative probability of $A \cap B$ in the restricted sample space B. Restrict means the conditional probability space focuses only on outcomes in B, so the new sample space $\Omega' = B$, and the events are $\mathcal{E}' = \{A \cap B, A \in \mathcal{E}\}$. The conditional probability defines a new measure $\mathbb{P}[\cdot|B]$ on these events that forms a probability space. Rescaling means the original measure $\mathbb{P}[\cdot)]$ must be rescaled (divided by $\mathbb{P}[B]$) so that $\mathbb{P}[\Omega'|B] = \mathbb{P}[B|B] = 1$.

Example 1.17
Consider the city example 1.16. What is the probability that someone's preferred city is Los Angeles, given that their preferred city is on the west coast?

$$\mathbb{P}[\{la\}|\{la,sf\}] = \frac{\mathbb{P}[\{la\} \cap \{la,sf\}]}{\mathbb{P}[\{la,sf\}]} = \frac{\mathbb{P}[\{la\}]}{\mathbb{P}[\{la,sf\}]} = \frac{\frac{1}{8}}{\frac{3}{8}} = \frac{1}{3}.$$

Example 1.18
Consider the previous example 1.17. What is the probability that, if you find the part is good, it was made in factory 1?

The observed event is $B = \{(1G, 2G)\}$, meaning the part is good. The event of interest is $A = \{1G, 1N\}$. We want $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{0.63}{0.87}$.

Here is a brief summary of the properties of conditional probability, which reflect the concept that it is a full probability measure on a restricted sample space corresponding to outcomes in B:

1. $\mathbb{P}[A|B] \in [0,1]$. It is a probability measure.
2. $\mathbb{P}[\Omega|B] = \mathbb{P}[B|B] = 1$.
3. If $A = A_1 \cup A_2 \cup \ldots$ where A_i are mutually exclusive, then
$$\mathbb{P}[A|B] = \mathbb{P}[A_1|B] + \mathbb{P}[A_2|B] + \ldots$$

To show this last item, note the following:

$$\begin{aligned}
\mathbb{P}[A|B] &= \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[(\cup_{i=1}^{\infty} A_i) \cap B]}{\mathbb{P}[B]} \\
&= \frac{\mathbb{P}[\cup_{i=1}^{\infty} [A_i \cap B)]}{\mathbb{P}[B]} \quad \text{(note } A_1 \cap B, A_2 \cap B, \ldots \text{ are mutually exclusive)} \\
&= \frac{\sum_{i=1}^{\infty} \mathbb{P}[A_i \cap B]}{\mathbb{P}[B)} = \sum_{i=1}^{\infty} \frac{\mathbb{P}[A_i \cap B]}{\mathbb{P}[B]} = \sum_{i=1}^{\infty} \mathbb{P}[A_i|B]
\end{aligned}$$

Example 1.19
Consider an experiment where we roll two 6-sided, balanced dice, so all 36 outcomes are equally likely. We consider the following events: B is the set of all outcomes where the smallest of the two numbers rolled is 3. Note that B has 7 elements. A is the set of all outcomes where the first die rolls a 3. Note that A has 6 elements, four of which are in B. In this case,

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{4}{7}.$$

Conditional probability is often used to describe the probability measure on complex experiments. In these experiments, you can define the overall probability as a sequence of conditional experiments. We have already seen this illustrated in example 1.15, where the probability that a part was good was dependent on which factory produced it. We illustrate this in the following example:

Example 1.20
We are going to draw three cards from a perfectly shuffled deck of cards, where each order is equally likely. What is the probability that we draw three hearts?

Let A be the event that the first card we draw is a heart. Since there are 13 of those in the 52 cards, $\mathbb{P}[A] = \frac{1}{4}$. Let B be the event that the second card we draw is a heart. Note that we can compute the conditional probability of B given A, because if A was observed, then there are only 12 out of 51 cards left that are hearts, so $\mathbb{P}[B|A] = \frac{12}{51}$. Let C be the event that the third card drawn is a heart. Then, $\mathbb{P}[C|B \cap A] = \frac{11}{50}$.

Then, using the multiplication rule,

$$\mathbb{P}[A \cap B \cap C] = \mathbb{P}[C|B \cap A]\mathbb{P}[B|A]\mathbb{P}[A] = \frac{1}{4} \cdot \frac{12}{51} \cdot \frac{11}{50} = \frac{11}{850}.$$

One of the foundations of inference is Bayes' theorem, which is a consequence of the definition of conditional probability.

Theorem 1.1
Bayes' Rule: Let A, B be events in a probability spae with $\mathbb{P}[A) > 0, \mathbb{P}[B) > 0$. Then,

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

Bayes' Rule provides a technique for evaluating the probability of cause, based on the observation of effect. For example, for medical diagnosis, we often to compute $\mathbb{P}[B|A]$, what is the probability of cause B, given the observed effects A. We often have a model for $\mathbb{P}[A|B]$, the probability that certain effects A are associated with cause B. Using Bayes' Rule, we reverse the implication.

An example of the application of Bayes' Rule was seen in example 1.18. In that example, we computed the probability that a good part was manufactured in factory 1, after seeing the effect that the manufactured part was good. However, one of the hardest part for applying Bayes' Rule is computing the denominator $\mathbb{P}[B]$.

One way of doing this is to use the Law of Total Probability, which can be stated as follows.

Theorem 1.2
Let $A_1, A_2, \ldots$ denote a countable set of pairwise mutually exclusive events with $\mathbb{P}[A_i)] > 0$ for $i = 1, 2, \ldots$, and assume that $A_1 \cup A_2 \cup \ldots = \Omega$, so the events are collectively exhaustive. Thus, $A_1, A_2, \ldots$ is a partition. Then, for any event $B \in \mathcal{E}$,

$$\mathbb{P}[B] = \mathbb{P}[B|A_1]\mathbb{P}[A_1] + \mathbb{P}[B|A_2]\mathbb{P}[A_2] + \ldots = \mathbb{P}[B \cap A_1] + \mathbb{P}[B \cap A_2] + \ldots$$

. The Law of Total Probability is useful for computing the denominator $\mathbb{P}[B]$ in Bayes' Rule, by decomposing B as the union of disjoint events.

Example 1.21
You go in for a diagnostic test for a specific diseases, and you test positive! You know that the test can have false positives, and the probability of a false positive (a positive diagnosis when you are not ill) is 0.05. However, you know that the test never misses a disease: the probability that the test returns a positive diagnosis if you are ill is 1.0. However, you know that this is a rare disease, that affects only 0.1% of the population.

Given all that information, what is the probability that you are actually ill?

Let's proceed as before: outcomes of the experiment are $\Omega = \{H+, H-, I+, I-\}$ where H indicates that you are healthy (I for ill), and $+, -$ are the possible outcomes of the diagnostic test. Define the events $H = \{H+, H-\}, I = \{I+, I-\}$ correspond to the person being healthy or ill, and the events $P = \{H+, I+\}, N = \{H-, I-\}$ to the event that the test outcome is positive or negative. Can we construct the probabilities of these outcomes given the information? Note what we are given the following in the problem description:

- Only 0.1% of the population has the disease: $\mathbb{P}[I] = 0.001, \mathbb{P}[H] = 0.999$.
- The probability of a false positive is 0.05: $\mathbb{P}[H \cap P|H] = 0.05, \mathbb{P}[H \cap N|H] = 0.95$.
- The test never has a missed detection: $\mathbb{P}[I \cap P|I] = 1; \mathbb{P}[| \cap N|I] = 0$.

The tree on the right illustrates a conditional diagram for this information, and helps us organize our computation.

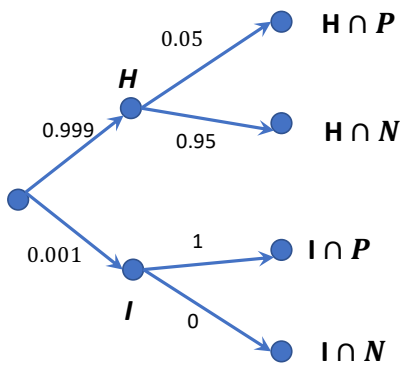


Figure 1.10: Figure for example 1.21.

We want to compute $\mathbb{P}[I|P]$. We use Bayes' rule as:

$$\mathbb{P}[I|P] = \frac{\mathbb{P}[I \cap P]}{\mathbb{P}[P]} = \frac{\mathbb{P}[P|I]\mathbb{P}[I]}{\mathbb{P}[P]}$$

From the diagram, we can see the following: $\mathbb{P}[I] = 0.001, \mathbb{P}[P|I] = 1$, so the terms in the numerator are readily evaluated. What about the denominator? For this, we use the Law of Total Probability, as

$$\mathbb{P}[P] = \mathbb{P}[P|I]\mathbb{P}[I] + \mathbb{P}[P|H]\mathbb{P}[H] = 1 * 0.001 + 0.05(0.999) = 0.05095.$$

Combining the numerator and denominator, we get $\mathbb{P}[I|P] = \frac{0.001}{0.05095} \approx 0.0196$.

Note how we used the Law of Total Probability to compute the denominator, since H, I form a partition of Ω. The message is that you should not be fast to assume you are sick...probability can help understand how to combine the different pieces of information!

Example 1.22
Consider a noisy communication channel, where binary bits are transmitted (values 0 or 1) but received occasionally with errors. Assume that a bit is received correctly with probability 0.95, and is received in error with probability 0.05. Assume that the probability of transmitting a 1 is 0.1, and a zero is 0.9. Given that the bit you received is 0, what is the probability that the transmitted bit was 0?

Define the following events:

Event A_1: Bit 0 transmitted. We are given $\mathbb{P}[A_1] = 0.9$.

Event A_2: Bit 1 transmitted, $\mathbb{P}[A_2] = 0.1$.

Event B_1: Bit 0 received. $\mathbb{P}[B_1|A_1] = 0.95, \mathbb{P}[B_1|A_2] = 0.05$.

Event B_2: Bit 1 received. $\mathbb{P}[B_2|A_1] = 0.05, \mathbb{P}[B_2|A_2] = 0.95$.

We wish to compute $\mathbb{P}[A_1|B_1]$. Using Bayes' Rule, and the Law of Total probability, this is

$$\mathbb{P}[A_1|B_1] = \frac{\mathbb{P}[B_1|A_1]\mathbb{P}[A_1]]}{\mathbb{P}[B_1]} = \frac{\mathbb{P}[B_1|A_1]\mathbb{P}[A_1]}{\mathbb{P}[B_1|A_1]\mathbb{P}[A_1] + \mathbb{P}[B_1|A_2]\mathbb{P}[A_2]} = \frac{0.9 * 0.95}{0.9 * 0.95 + 0.1 * 0.05} \approx 0.994$$

Example 1.23
Monty Hall Game Show: Here is a paradoxical example from the game show "Let's Make a Deal." You know there is a prize behind one of three doors. You are asked to pick one, which you do: door 1. The game show host, Monty Hall, opens door number 2 and shows that there is no prize behind that door. He then gives you a choice to keep your original door, or switch to door 3. Should you switch?

At first glance, the choice seems harmless: There are two doors left, and the prize is behind one of them. It seems like switching and not switching should give you equal chance of winning. However, during the few seasons of the show, those that switched wound up winning 2/3 of the time? Why?

Here is a quick explanation: The original door choice had only 1/3 chance of winning. Hence, switching to both of the other two door choices has 2/3 chance of winning. The fact that Monty opened one of those doors and showed it had no prize means that choosing the other door has the same chance of winning as choosing both doors, namely 2/3.

Let's analyze this using Bayes' Rule: Sample space $\Omega = \{1, 2, 3\}$ corresponding to which door has a prize. The measure on atoms is $\mathbb{P}[\{1\}] = \mathbb{P}[\{2\}] = \mathbb{P}[\{3\}] = 1/3$. Let event $E_1 =$ prize is behind door 1. Let event O be the event that Monty opens a door that does not have a prize behind it. What is $\mathbb{P}[E_1|O]$?

Bayes' Rule: $\mathbb{P}[E_1|O] = \frac{\mathbb{P}[O|E_1]\mathbb{P}[E_1]}{\mathbb{P}[O]}$. Here is the source of the paradox: Monty will always open a door that does not have a prize (or else the game ends!) Hence, $\mathbb{P}[O] = 1, \mathbb{P}[O|E_1] = 1$. Thus, Bayes' Rule yields

$$\mathbb{P}[E_1|O] = \frac{\mathbb{P}[O|E_1]\mathbb{P}[E_1]}{\mathbb{P}[O]} = \mathbb{P}[E_1] = 1/3.$$

which means that not switching wins the prize only 1/3 of the time. Thus, one should always switch!

Example 1.24
3 factories manufacture batteries for an electric car. However, the batteries from factory one meet the needed specification only 70% of the time, while the batteries from factories 2 and 3 meet specifications 80% and 85% of the time respectively. The car manufacturer buys 40% of its batteries from factory 1, 30% of its batteries from factory 2, and the remaining 30% of its batteries from factory 3.

An outcome of this experiment is the battery that is in the car you purchased. The battery was made by one of the three manufacturers, and it either met the specification (denote by G), or did not (denote by B). Thus, the sample space can be described with 6 outcomes, as $\Omega = \{1G, 1B, 2G, 2B, 3G, 3B\}$. Let A denote the event that the battery in your car meets specification. Let B_i denote the event that the battery in your car came from factory i, for $i = 1, 2, 3$. Note that B_1, B_2, B_3 are mutually disjoint and collectively exhaustive. Then, using the Law of Total Probability,

$$\mathbb{P}[A] = \sum_{i=1}^{3} \mathbb{P}[A|B_i]\mathbb{P}[B_i] = 0.7 * 0.4 + 0.8 * 0.3 + 0.85 * 0.3 = 0.775$$

Example 1.25
This is a longer example to show the use of Bayes' Rule and conditional probability. There is a new virus infecting smartphones and randomly compromising some of them. We know that 80% of the smartphones run Android OS, and 20% run a different operating system. Let A denote the event of phones that run Android, and let A^c denote the event of phones that run a different operating system. Let B be the event that the virus infects the phone. We are given that $\mathbb{P}[B|A] = 0.5$, so that $\mathbb{P}[B^c|A] = 0.5$ also. We are also given that $\mathbb{P}[B|A^c] = 1/3$, because non-Android phones are less common.

Let C be the event that the phone is compromised if it is infected. We are given that $\mathbb{P}[C|A \cap B] = 1/4$. The phone can still be compromised even if it is not infected! The probability of this is $\mathbb{P}[C|A\ B^c] = 1/8$. For non-android phones, the probability that the phone is compromised if it is infected is $\mathbb{P}[C|A^c \cap B] = 1/6$, and the probability that the phone is compromised if the phone is not infected is $\mathbb{P}[C|A^c \cap B^c] = 1/8$.

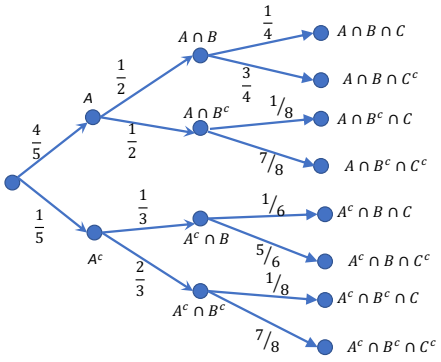


Figure 1.11: Figure for example 1.25.

We can combine all of these probabilities (and their complements) to obtain a complete event diagram that chains these events appropriately. This diagram is shown in Figure 1.11. For instanced, what is the probability that a phone that runs Android has no bug but is still compromised? This is $\mathbb{P}[A \cap B^c \cap C] = \mathbb{P}[A]\mathbb{P}[B^c|A]\mathbb{P}[C|A \cap B^C] = \frac{4}{5} \cdot \frac{1}{2} \cdot \frac{1}{8} = \frac{1}{20}$. Note that this is the product of the probabilities on the branches leading to the end node $A \cap B^c \cap C$.

What is the probability that a phone has a virus but is not compromised? That is $\mathbb{P}[B \cap C^c]$. We can compute this using the Law of Total Probability, as A and A^c form a partition of the sample space. Thus,

$$\mathbb{P}[B \cap C^c] = \mathbb{P}[B \cap C^c|A]\mathbb{P}[A] + \mathbb{P}[B \cap C^c|A^c]\mathbb{P}[A^c] = \mathbb{P}[A \cap B \cap C^c] + \mathbb{P}[A^c \cap B \cap C^c] = \frac{3}{10} + \frac{1}{18} = \frac{16}{45}.$$

Given that a phone is compromised, what is the probability that it is an Android phone? We answer this using Bayes' Rule:

$$\mathbb{P}[A|C] = \frac{\mathbb{P}[C|A]\mathbb{P}[A]}{\mathbb{P}[C]}$$

Note that B and B^c are also a partition of the sample space. Using the Law of Total Probability with the conditional probability $\mathbb{P}[\cdot|A]$, we get

$$\mathbb{P}[C|A] = \mathbb{P}[C|A \cap B]\mathbb{P}[B|A] + \mathbb{P}[C|A \cap B^c]\mathbb{P}[B^c|A] = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{8} \cdot \frac{1}{2} = \frac{3}{16}.$$

Similarly, note that $A \cap B, A \cap B^c, A^c \cap B, A^c \cap B^c$ form a partition of the sample space Y. Then,

$$\begin{aligned}\mathbb{P}[C] &= \mathbb{P}[C \cap A \cap B] + \mathbb{P}[C \cap A \cap B^c] + \mathbb{P}[C \cap A^c \cap B] + \mathbb{P}[C \cap A^c \cap B^c] \\ &= \frac{4}{5} \cdot \frac{1}{2} \cdot \frac{1}{4} + \frac{4}{5} \cdot \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{5} \cdot \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{5} \cdot \frac{2}{3} \cdot \frac{1}{8} = \frac{8}{45}\end{aligned}$$

1.3.1 Independence

Independence is an important concept in probability. It is one of the most common assumptions used in modeling experiments with multiple sources of randomness, and allows for efficient characterization of the resulting probability measures.

Mathematically, two events A, B in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ are **independent** if and only if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. Independence implies $\mathbb{P}[B|A] = \mathbb{P}[B], \mathbb{P}[A|B] = \mathbb{P}[A]$ as long as $\mathbb{P}[A] > 0, \mathbb{P}[B] > 0$, because for any two events, $\mathbb{P}[A \cap B] = \mathbb{P}[A|B]\mathbb{P}[B]$.

Independence has nothing to do with A, B being mutually exclusive, i. e. having no outcomes in common. If A, B have no outcomes in common, then $\mathbb{P}[A \cap B] = 0$, and $\mathbb{P}[A|B] = 0$ also, so it is impossible for $\mathbb{P}[A|B] = \mathbb{P}[A]$ as long as $\mathbb{P}[A] > 0$. If A, B are disjoint, knowing that the experiment outcome is in B implies it is cannot be in A and so this provides information about A, and thus the events are not independent. Independence is a property of the probability measure $\mathbb{P}[\cdot]$ and the specific sets A, B.

If A, B are independent events, then A^c, B are also independent, because

$$\mathbb{P}[A^c \cap B] + \mathbb{P}[A \cap B] = \mathbb{P}[B] = \mathbb{P}[A^c \cap B] + \mathbb{P}[A]\mathbb{P}[B]$$

This implies

$$\mathbb{P}[A^c \cap B] = (1 - \mathbb{P}[A])\mathbb{P}[B] = \mathbb{P}[A^c]\mathbb{P}[B]$$

Example 1.26
Consider a simple sample space $\Omega = \{0, 1, 2, 3\}$ with four elements, and define the probability measure on the atoms to be:

$$\mathbb{P}[\{0\}] = \frac{2}{9}; \mathbb{P}[\{1\}] = \frac{1}{9}; \mathbb{P}[\{2\}] = \frac{4}{9}; \mathbb{P}[\{3\}] = \frac{2}{9}.$$

Define events $A = \{0, 1\}, B = \{2, 3\}, C = \{1, 3\}$.

Are A, B independent? They are mutually exclusive, so they are not, because $\mathbb{P}[A \cap B] = 0$.

Are A, C independent? We have to check: $\mathbb{P}[A] = \mathbb{P}[\{0\}] + \mathbb{P}[\{1\}] = \frac{1}{3}; \mathbb{P}[C] = \mathbb{P}[\{1\}] + \mathbb{P}[\{3\}] = \frac{1}{3}$. $\mathbb{P}[A \cap C] = \mathbb{P}[\{1\}] = 1/9 = \mathbb{P}[A]\mathbb{P}[C] = \frac{1}{3}$. Thus, they are independent!

The concept of independence can be extended to a finite sequence of sets $A_1, \ldots, A_m$, which are mutually independent if

- Any collection of k of the sets $(k < m)$ $A_{j_1}, A_{j_2}, \ldots, A_{j_k}$ are mutually independent. item $\mathbb{P}[A_1 \cap A_2 \cap \ldots \cap A_m) = \mathbb{P}[A_1]\mathbb{P}[A_2] \cdots \mathbb{P}[A_m]$.

.

Note that the above concept of mutual independence implies much more than pairwise independence. For pairwise independence, any two sets A_i, A_j, $i \neq j, i, j \in \{1, \ldots, m\}$ are independent. It is easy to construct examples of events which are pairwise independent, but not mutually independent. Mutually independent means that no subset of the events can be used to predict the probability of occurrence of any of the other events.

Independence can be tedious to check. Often, as in the above example, it is easier to recognize lack of independence, as when two events are mutually exclusive. In many engineering applications, we will assume independence.

Example 1.27
The experiment in this example is to flip a coin twice, and record both faces. The sample space is thus $\Omega = \{HH, HT, TH, TT\}$. Assume the coins are fair, so each atom in the sample space has probability $\frac{1}{4}$. Define the following events: $A =$ {First flip is H}; $B =$ {Second flip is H}; $C =$ {Flips have different outcomes}. Note that $\mathbb{P}[A] = \mathbb{P}[B] = \mathbb{P}[C] = \frac{1}{4}$, as each of the events has two outcomes.

Observe the following: $\mathbb{P}[A \cap B] = \mathbb{P}[\{HH\}] = \frac{1}{4}; \mathbb{P}[A \cap C] = \mathbb{P}[\{HT\}] = \frac{1}{4}; \mathbb{P}[B \cap C] = \mathbb{P}[\{TH\}] = \frac{1}{4}$. Therefore, A and B are independent, and B and C are independent, and A and C are independent! Thus, A, B, C are pairwise independent. However, are they mutually independent?

Note that $A \cap B \cap C = \emptyset$, so $\mathbb{P}[A \cap B \cap C] = 0 \neq \mathbb{P}[A]\mathbb{P}[B]\mathbb{P}[C]$. Therefore, the events A, B, C are not mutually independent.

We now define the concept of conditional independence. Two events A, B are *conditionally independent* given event C if $\mathbb{P}[A \cap B|C] = \mathbb{P}[A|C)]\mathbb{P}[B|C]$. Basically, this defines independence in terms of the conditional probability measure $\mathbb{P}[\cdot|C]$. Note the following: two events that were independent in the original probability measure $\mathbb{P}[\cdot]$ can become conditionally dependent when a third event C is observed as true. Similarly, two events that were not independent originally can become independent given C is observed.

Example 1.28
We illustrate this concept with a simple encryption example. Assume we want to send a single bit M equally likely to be 0 or 1. We generate independently another bit K, equally likely to be 0 or 1, and we send the message

$$C = M \oplus K$$

where $\oplus$ is addition modulo 2. Thus, if $K = 1$, the original bit M is flipped. The sample space of these experiments is $\Omega = \{00, 01, 10, 11\}$ corresponding to possible pairs MK, and $\mathbb{P}[\{ij\}] = 1/4, i, j = 0, 1$.

Define following events are independent: $K_i = \{K = i\}, M_j = \{M = j\}$. Note that K_0, M_1 are independent. Consider the event $C_i = \{C = i\}$.

By construction, K_0 and M_1 are independent, which can be verified by

$$\mathbb{P}[K_0 \cap M_1] = \mathbb{P}[\{00, 10\} \cap \{10, 11\}] = \mathbb{P}[\{10\}] = 0.25 = \mathbb{P}[K_0]\mathbb{P}[M_1]$$

Note that $\mathbb{P}[C_0] = 0.5$. Then,

$$\begin{aligned}
\mathbb{P}[K_0|C_0] &= \frac{\mathbb{P}[K_0 \cap C_0]}{\mathbb{P}[C_0]} = \frac{\mathbb{P}[\{00, 10\} \cap \{00, 11\}]}{\mathbb{P}[C_0]} = 0.5, \\
\mathbb{P}[M_1|C_0] &= \frac{\mathbb{P}[M_1 \cap C_0]}{\mathbb{P}[C_0]} = \frac{\mathbb{P}[\{10, 11\} \cap \{00, 11\}]}{\mathbb{P}[C_0]} = 0.5, \\
\mathbb{P}[K_0 \cap M_1|C_0] &= \frac{\mathbb{P}[K_0 \cap M_1 \cap C_0]}{\mathbb{P}[C_0]} = \frac{m(\emptyset)}{\mathbb{P}[C_0]} = 0.
\end{aligned}$$

which shows that K_0 and M_1 are not conditionally independent given C_0.

Here is another surprising fact: C_0 and M_1 are independent! We show this by computation:

$$\mathbb{P}[C_0 \cap M_1] = \mathbb{P}[\{00, 11\} \cap \{10, 11\}] = 0.25 = \mathbb{P}[C_0]\mathbb{P}[M_1]$$

You can verify that C_i is independent of K_j, and M_k for any set of choices $i, j, k \in \{0, 1\}$! What this implies is that knowing C alone does not reveal anything about M, hence ensuring the privacy of M. In our simple scale, the possible weights are $\{0, 1, 2, 3\}$. In addition to its bias, the scale has an error measuring a weight which is independent for each time you weigh the object, and the error has equal probability in $\{0, 1\}$.

Example 1.29
Here is an example where two dependent events become conditionally independent. An acoustic microphone is listening to detect whether a particular sound waveform is present or not. However, the background noise in the room can either be "loud" or "soft". The probability that the background is "loud" is 0.5. If the background nose is "loud", the microphone will detect the presence of a sound with probability of error 0.4. That is, if the sound is present, the microphone will detect it with probability 0.6, and fail to detect it with probability 0.4.

If the background noise is "soft", the microphone will detect the presence of a sound with probability of error 0.2. The experiment consists of a room with the noise present with background chosen randomly from "loud" or "soft". The microphone will try twice to detect the presence of the sound twice, where the errors the microphone makes are independent for each try., but the background noise is the same in both measurements.

The sample space in this experiment can be stated in terms of 3 variables: l or s for whether the background is loud or soft, $d_1 = 0, 1$ as to whether the first mesurement is a detection, and $d_2 = 0, 1$ as to whether the second measurement is a detection. Thus, the sample space consists of 16 outcomes,

$$\Omega = \{l00, l01, l10, l11, s00, s01, s10, s11\}$$

Given the description of the experiment, we compute the probability of each atom as follows:

$$\mathbb{P}[\{l00\}] = 0.08; \mathbb{P}[\{l01\}] = \mathbb{P}[\{pl10\}] = 0.12; \mathbb{P}[\{l11\}] = 0.18$$

$$\mathbb{P}[\{s00\}] = 0.02; \mathbb{P}[\{s01\}] = \mathbb{P}[\{s10\}] = 0.08; \mathbb{P}[\{s11\}] = 0.32$$

Note that these add up to 1. Define the event A be the event that the first measurement is a detection, and B the event that the second measurement is a detection. By combining the atoms that form these events, we compute $\mathbb{P}[A] = 0.7, \mathbb{P}[B] = 0.7$. However, $\mathbb{P}[A \cap B] = 0.5$, so the events are not independent.

Define the event C to be the event that the background is "loud". By construction, $\mathbb{P}[B) = 0.5$. Now,

$$\mathbb{P}[A|C] = \frac{\mathbb{P}[A \cap C]}{\mathbb{P}[C]} = \frac{0.3}{0.5} = 0.6$$

$$\mathbb{P}[B|C] = \frac{\mathbb{P}[B \cap C]}{\mathbb{P}[C)} = \frac{0.3}{0.5} = 0.6$$

$$\mathbb{P}[A \cap B|C] = \frac{\mathbb{P}[A \cap B \cap C]}{\mathbb{P}[C]} = \frac{0.18}{0.5} = 0.36$$

which shows that A, B are conditionally independent given C, knowledge of the background state. Basically, the dependence in events A, B arises because they contain the common uncertainty from the same background state. This is removed if the background is observed so it is no longer uncertain.

Example 1.30
Consider the experiment of selecting an integer from 1 to 4, where each number is equally likely. Consider the events $\{1, 2\}, \{1, 3\}, \{1, 4\}$; Note that the above events are pairwise independent. However,

$$\mathbb{P}[\{1, 2\} \cap \{1, 3\} \cap \{1, 4\}] = 1/4 \neq \mathbb{P}[\{1, 2\}].\mathbb{P}[\{1, 3\}]\mathbb{P}[\{1, 4\}] = 1/8.$$

Example 1.31
Mutual independence is a much stronger condition that pairwise independence. For example suppose we were looking for 5 genetic markers in blood samples, denoted by A, B, C, D, E. We are given that the presence of marker A is in 1 out of every 100 persons, marker B in 1 of 50 persons, marker C in one of 40 persons, marker D in one of 5 persons and marker E in one of 170 persons. If the presence of each of the markers was mutually independent, the probability that all the markers were present in a blood sample is

$$\mathbb{P}[A \cap B \cap C \cap D \cap E] = \frac{1}{100 * 50 * 40 * 5 * 170} = \frac{1}{170,000,000}$$

However, if all we knew was that the presence of the markers was pairwise independent but not mutually independent, then all we can say is

$$\mathbb{P}[A \cap B \cap C \cap D \cap E) \leq \mathbb{P}[A \cap E) \leq \frac{1}{17,000}$$

Those three orders of magnitude matter!

The concepts of independence and conditional independence are used extensively in this course to construct complex compound experiments. We have already seen this in several examples previously. These experiments are sequences of sub-experiments, where the later subexperiments depend on the outcomes of the earlier subexperiments. That is, we are given an initial subexperiment with events of outcomes A_i, defined by probability measure $\mathbb{P}[A_i]$. Then, we define the next sub-experiment with events of outcomes B_j, defined by conditional probability measures $\mathbb{P}[B_j|A_i]$ depending on the events observed in the earlier subexperiments. Note that we can now define a probability measure on the compound experiment with

$$\mathbb{P}[B_j \cap A_i] = \mathbb{P}[B_j|A_i]\mathbb{P}[A_i]$$

This is particularly simple when the experiments have discrete outcomes and we define the probability measures on atoms. We illustrate this below with two examples.

Example 1.32
We have three factories B_i making a product. The experiment will generate a sample product, which may or may not be acceptable. As a first step, we select which factory will make the product. We describe this by a probability measure over the atoms of this first step, $\mathbb{P}[\{B_i\})$, with probability 0.3, 0.4, 0.3 for each of the outcomes B_1, B_2, B_3. In the next part of the experiment, the selected factory makes a sample product, which may turn out to be acceptable A or not N. We describe this with conditional probability based on the factory selected:

$$\mathbb{P}[A|B_1] = 0.8, \mathbb{P}[N|B_1] = 0.2; \mathbb{P}[A|B_2] = 0.9, \mathbb{P}[N|B_2] = 0.1; \mathbb{P}[A|B_3] = 0.6, \mathbb{P}[N|B_3] = 0.4.$$

We can illustrate this example with a tree diagram, as illustrated in Figure 1.12. The compound experiment has defined all the probabilities.

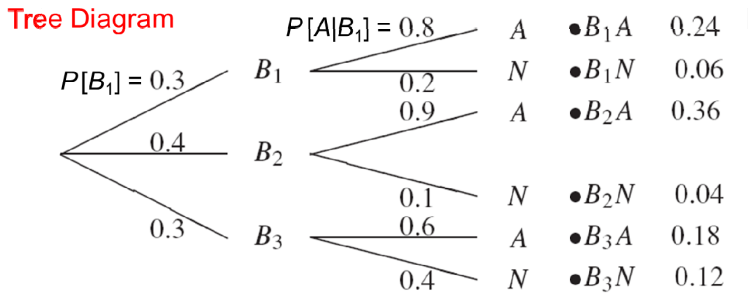


Figure 1.12: Tree diagram for example 1.32.

Example 1.33
Consider a communication channel where we are going to send a four-bit sequence of bits $a_1a_2a_3a_4$. Each bit is generated independently from $\{0, 1\}$ with $\mathbb{P}[\{a_i = 0\}] = 0.4$. The bits are input into the communication channel one at a time, where each bit can be flipped independently by the channel with probability 0.2, or left as is with probability 0.8.

This experiment is illustrated in figure 1.13.

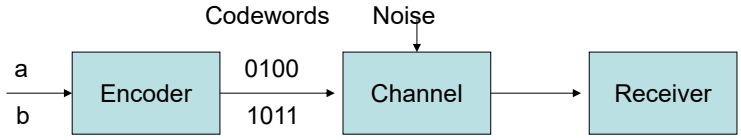


Figure 1.13: Illustration of communications channel in example 1.33.

We construct the probability model with a compound experiment. First, we generate the code word as one of 16 possible binary code words in the sample space

$$\Omega = \{0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111\}$$

The probability measure in this space is given by the signal generation: each atom will have probability $(0.6)^n(0.4)^{4-n}$ where n is the number of ones in the code.

Next, we generate the errors in the code word, based on the error description. Given the code word, the probability that we generate another code word that differs from the current one is $(0.8)^n(0.2)^{4-n}$, where n here refers to the number of bits that were not flipped in error. Hence, $\mathbb{P}[\{0100\}|\{0000\}] = (0.8)^3(0.2)$ We now have a complete probability model, and can answer the following question: If one receives 0010, what is the probability that 0010 was the transmitted message?

$$\mathbb{P}[\{0010\text{transmit}\}|\{0010\}\text{receive}] = \frac{\mathbb{P}[\{0010\text{receive}\}|\{0010\}\text{transmit}]\mathbb{P}[\{0010\text{transmit}\}]}{\mathbb{P}[\{0010\text{receive}\}]}$$

1.4 Computing probability measures for finite sample spaces with equally likely outcomes

There are many cases where one assumes that every outcome is equally likely in an experiment. In this case, the probability of an event $A \in \Omega$ can be computed as the following ratio:

$$\mathbb{P}[A] = \frac{\#\text{ outcomes in } A}{\#\text{ outcomes in } \Omega}.$$

However, we want to avoid enumeration of all the outcomes to figure out the counts! For instance, if we shuffle a deck of 52 cards, what is the number of possible outcomes, where an outcome is a particular order of the 52 cards? Fortunately, we know this is just a problem of permutations, and that number is 52!, which saves us the trouble of enumerating all the card orders! But, consider the event where you are the third player among 4 players in a game of Bridge, and your hand will be dealt all four aces. How many of the shuffle outcomes are in this event? That requires far more clever counting.

Unlike some versions of Probability courses, counting is not a major part of this course. However, the early history of probability up to the 19th century focused on games of chance where counting was the predominant tool for computation of probabilities. In this section, we review some basic formulas that can be used for effective counting.

1.4.1 Counting

The first set of formulas for counting involve permutations and combinations. Given a set of n unique elements, a possible order of these elements is a **permutation**. The number of possible permutations of a set of n elements is $n!$. If we have to select k of these n elements, and order is not important, the number of unique k element sets that can be chosen out of n elements is $\binom{n}{k}$, where

$$\binom{n}{k} = \frac{n!}{(n-k)!(k)!} = \binom{n}{n-k}$$

The number $\binom{n}{k}$ is also called the binomial coefficient. This is because the coefficients in the binomial theorem are given by

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

We will encounter this formula later in this course. For now, it can be used to derive some simple identities such as

$$2^n = \sum_{k=0}^{n} \binom{n}{k},$$

obtained by substituting $a = b = 1$ into the binomial theorem.

Example 1.34
Using the above formulas, we can answer the Bridge question asked earlier: what is the probability that you, sitting as third chair, will be dealt all four aces? We know the number of possible outcomes in Ω, which is the number of permutations of 52 cards: (52)! To compute the number of outcomes where all four aces lie in the cards dealt to the third chair, we proceed as follows:

Assume the four aces are in the cards dealt to the third chair. The number of possible orders for the aces among the 13 cards received by the third chair is $\binom{13}{4}(4!)$, where the first term represents the times you receive an ace, and the second term represents the order in which the aces are received. This number is thus $\frac{13!}{4!9!}4! = (13)(12)(11)(10)$. For each of these possible arrangements of the aces for the third chair, the other 48 cards can be arranged arbitrarily, so there are 48! arrangements of the remaining cards.

We have just computed the number of card shuffles where the third chair will be dealt all four aces as (13)(12)(11)(10) 48!. The probability that this event A happens when all shuffle outcomes are equally likely is:

$$\mathbb{P}[A] = \frac{(13)(12)(11)(10)48!}{52!} = \frac{(13)(12)(11)(10)}{(52)(51)(50)(49)} = \frac{(11)}{(17)(5)(49)} \approx 0.26\%$$

which means that you get all four aces approximately once in every 400 hands (if the dealer is honest.)

Another set of useful counting formulas focus on experiments that are composed of ordered subexperiments. Assume there are r subexperiments, and k^{th} subexperiment consists of n_k outcomes (that can be freely chosen). For instance, you are going to the pet store to buy one of 10 fishes in the small fish tank, one of the 6 dogs in the kennels and one of the 7 cats in the kennel. What are the total number of fish/dog/cat outcomes? The general formula is given by:

$$\#\text{ of outcomes } = n_1 \cdot n_2 \cdots n_r$$

which, for the case of fish/dog/cat outcomes, becomes: # of outcomes = 420.

Counting experiments of this type arise, for instance, in dice games with many dice, or coin games with many coins. If you roll 10 six-sided dice, what are the numbers of possible outcomes? In this case, $n_k = 6$ for each k and $r = 10$, thus it is 6^{10}.

Note that in the above count, we keep track of the outcome for each subexperiment. Thus, order matters in these counts. An implicit assumption is that each subexperiment has outcomes that are selected independently of each other, which is why the total number of outcomes is the product of the number of possible outcomes in each subexperiment.

1.4.2 Sampling

Sampling problems are popular problems in early courses in probability. The typical problem considers a bag with n unique balls (e.g. a lottery urn with 100 numbers). Given that you will take k balls out of the bag, how many possible ways are there to take k balls out? We make a distinction as to whether order matters or not. If order does not matter, this is the simple combination formula we discussed earlier; that is, the number of possible combinations of k balls is $\binom{n}{k}$. However, if order matters, then there are more ways: the right number is $k!\binom{n}{k} = \frac{n!}{(n-k)!}$.

What if the balls are replaced and put back in, so that the same ball can be taken out more than once? We refer to that as sampling with replacement. If order matters, then this is the same as running k subexperiments, each with n possible outcomes, so the total number of outcomes is n^k.

If order does not matter, the number of different outcomes is different. It is hardly obvious as to how to count in this case, as it is not a standard combination. Here is a different way to pose the problem. Assume there are n distinct items. Let $x_i, i = 1, \ldots, n$ denote the number of times an item appears in an outcome; note that this is order-independent. If we are to draw a total of k items, we must have $x_1 + x_2 + \ldots + x_n = k$ in any outcome. Thus, the total number of outcomes is the number of possible solutions of this equation where $x_i \in \{0, 1, \ldots, n\}, i = 1, \ldots, n$.

Let's furthermore represent multiplicities as numbers of ones: Hence, if $x_k = 3$, then $x_k = 111$. Similarly, $x_k = 0$ would be replaced by $x_k =$, that is, no entry. With this notation, the term $x_1 + x_2 + \ldots + x_n$ must be a sequence of length $k + (n-1)$ composed of exactly k digits 1, and $n - 1$ + signs! This is the hard part to visualize: we have reduced the problem to finding $n - 1$ positions for the + signs out of the $k + n - 1$ total positions. For instance, if $n = 3$ and $k = 2$, the sequence ++11 means $x_1 = x_2 = 0, x_3 = 2$. The sequence 1+1+ means $x_1 = x_2 = 1, x_3 = 0$. Once you understand this mapping, the final answer is just another combination formula:

$$\#\text{ order-independent } k \text{ out of } n \text{ samples with replacement} = \binom{n-1+k}{n-1} = \binom{n-1+k}{k}$$

Example 1.35
Suppose we have three balls, one red, one blue and one green. We put them in a bag, and sample them three times with replacement. How many order-independent sets of colors can be obtained from this sampling?

In this example, the number of draws $k = 3$, and the number of possible colors (values) is $n = 3$. Hence, the total number of order-independent outcomes is

$$\binom{n-1+k}{k} = \binom{3-1+3}{3} = \binom{5}{3} = 10.$$

For this small example, we can actually list all of the possible outcomes. We list outcome R for a red ball, G for a green ball, and B for a blue ball. The outcomes are

`RRR RRB RRG RBB RBG RGG BBB BBG BGG GGG`

We summarize these formulas in Table 1.1:

	Order Dependent	Order Independent
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Table 1.1: Sampling formulas

1.4.3 Partitions

Another popular set of examples in elementary probability courses consist of partition problems where we have n items and we want to divide them into r groups, so that the k^{th} group contains n_k elements, such that $n_1 + n_2 + \ldots + n_r = n$. For instance, we have 18 basketball players, and we want to divide them into six teams of three players. How many possible ways of forming teams are there?

This is an extension of the binomial coefficient formula. In that problem we wanted to divide n elements into two groups, one of size k and another of size $n-k$. In this extension each element appears in exactly one group because $\sum_{k=1}^{r} n_k = n$. The number of ways to form such a **partition** is given by the multinomial coefficient

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

The way to derive this is to perform sequential selection of combinations: First, select n_1 out of n. Then select n_2 out of the remaining $n - n_1$. Continue this until you select n_r out of the remaining n_r. This yields:

$$\begin{aligned}
\binom{n}{n_1, n_2, \ldots, n_r} &= \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_r}{n_r} \\
&= \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!}\cdots\frac{n_r!}{n_r!} \\
&= \frac{n!}{n_1! n_2! \cdots n_r!} \text{ because of all the cancelations with successive terms.}
\end{aligned}$$

This formula can be used in a generalization of the binomial theorem, as follows:

$$(x_1 + x_2 + x_3 + \ldots + x_r)^n = \sum_{n_1+n_2+\ldots+n_r=n} \binom{n}{n_1, n_2, \ldots, n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$$

which lead to some nice identities such as

$$r^n = \sum_{n_1+n_2+\ldots+n_r=n} \frac{n!}{n_1!n_2!\cdots n_r!}.$$

Example 1.36
Twelve people have a potluck party. Six people will be selected to bring a main dish, four people will bring drinks, and two people will bring dessert. How many ways can they be divided into these three groups?

We solve this using the multinomial partition formula:

$$\#\text{ partitions} = \binom{12}{6,4,2} = \frac{12!}{6!48} = \frac{(12)(11)(10)(9)(8)(7)}{48} = (22)(10)(9)(7) = 13,860.$$

Example 1.37
Suppose we have a lottery with numbers from 1 to 59. You are allowed to select five numbers, and you can choose the same number more than once. What is the probability of winning? Note that the numbers you select must be picked in the same order as the lottery to win.

What are the number of ways that five numbers can be chosen by the lottery? This is order-dependent sampling with replacement, so the formula is:

$$59^5 = 714,924,299.$$

Your chance of winning is one in 714,924,299.

If we only allowed selection of a number once (instead of putting the selected number back in the lottery urn), we would be doing order-dependent sampling without replacement, so the answer is

$$\frac{59!}{54!} = \frac{59 \times 58 \times 57 \times 56 \times 55}{=} 600,766,320$$

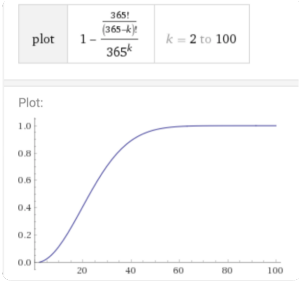
which increases your chances of winning a little bit.

Example 1.38
Assume you have a perfectly shuffled deck of cards. If you draw five cards, without replacement, what is the probability that exactly three of the five are kings? Note that this is not order-dependent. To answer this, we compute the number of five card hands with exactly three kings. The three kings must be chosen from four possible kings, and the other two cards chosen from 48 non-king cards. This gives the number of hands with three kings as $\binom{4}{3}\binom{48}{2}$. The total number of five card hands, with replacement, is $\binom{52}{5}$. Then,

$$\mathbb{P}[\{\text{Choose exactly 3 kings in five cards}\}] = \frac{\frac{4\cdot48\cdot47}{2}}{\frac{52\cdot51\cdot50\cdot49\cdot48}{5\cdot4\cdot3\cdot2}} = \frac{4512}{2598960} = 0.001736.$$

Here is another common example used to surprise a class.

Example 1.39
In a class with k students, assuming that each student was equally likely to be born in one of the 365 days in a regular calendar year, what is the probability that two or more students share a birthday?
Sometimes, it is easier to compute the probability of the complement of an event. In this case, we compute the probability that no student shares a common birthday. The number of ways to select k birthdays uniquely are $\frac{365!}{(365-k)!}$. The total number of ways to select k birthdays is 365^k. Hence, the probability that k students do not have a birthday in common among any two of them is



$$\mathbb{P}[\{\text{No common birthday in } k \text{ students}\}] = \frac{\frac{365!}{(365-k)!}}{365^k}.$$

Then, the probability that at least two students share a common birthday is

$$\mathbb{P}[\{\text{At least two students share a common birthday among } k \text{ students}\}] = 1 - \frac{\frac{365!}{(365-k)!}}{365^k}.$$

Note that this quickly approaches one! For $k = 64$ this is approximately 0.997. Our class size is bigger. For $k = 100$, the probability is 0.9999997, nearly 1. The curve of how the probability grows with k is shown in Figure 1.39.

1.4.4 Independent Trials

Up to now, we have assumed that every outcome was equally likely, and therefore we would compute probabilities by counting the number of outcomes. Jacob Bernoulli developed an extension for this were an experiment consisted of multiple identical, independent subexperiments, each with two possible outcomes (e.g. win or lose). However, the probability of an outcome in each subexperiment was unbalanced: the probability of winning was different than the probability of losing.

We refer to these subexperiments as Bernoulli trials. A Bernoulli trial is a random experiment with two outcomes, say "win" and "lose", where we define the probability of "win" as a number $p \in [0,1]$ and the probability of "lose" as $1-p$. For instance, a Bernoulli trial might be the outcome of a coin toss, where "heads" corresponds to "win".

If an experiment performs n independent Bernoulli trials, how many possible ways are there to get a total of k "win" outcomes? This is a combination problem, were we are going to select the k positions that have a "win" outcome among the n trials, and that is $\binom{n}{k}$. However, what is the probability of each of those combined outcomes? In each of those combined outcomes there are k subexperiments that resulted in "win" and $n-k$ subexperiments that resulted in "lose". Since these are independent subexperiments the probability of each of those outcomes is $p^k(1-p)^{n-k}$, as the probability for the combined outcome is the product of the probabilities of each subexperiment. Hence, the probability of the event $\{k$ "win" outcomes in n Bernoulli trials $\}$ is the sum of the probabilities of each outcome in the event, which is

$$\mathbb{P}[\{k \text{ "win" outcomes in } n \text{ Bernoulli trials }\}] = \binom{n}{k} p^k (1-p)^{n-k}$$

This distribution, discovered by Bernoulli, is termed the binomial distribution.

Example 1.40
You have a biased coin, such that "heads" occurs with probability 0.6. If you flip the coin 10 times, what is the probability that you have a total of 4 "heads" outcomes? Applying the above formula yields

$$\mathbb{P}[\{4 \text{ "heads" out of } 10\}] = \binom{10}{4} 0.6^4 0.4^6 \approx 0.1115.$$

Let's generalize the above result. Suppose that, instead of having two outcomes each subexperiment has r possible outcomes $a_1, \ldots, a_r$ with probabilities $p_1, \ldots, p_r$. We want to conduct n independent subexperiments and count the number of outcomes of each type. That is, we want to count the total number of outcomes n_1 of a_1, the number of outcomes n_2 of $a_2, \ldots$, and the total number of outcomes n_r of a_r. We know that the total number of partitions of n outcomes into r classes with n_k outcomes of class k is $\binom{n}{n_1, n_2, \ldots, n_r}$. What is important is that each of those partitions has the same probability because of the independence of the subexperiments: $p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$. This yields the following formula for the probability that, when we run n subexperiments, we will get n_1 outcomes of a_1, n_2 of a_2, $\ldots$, n_r of a_r:

$$\mathbb{P}\big[\{n_1 \text{ occurrences of } a_1, \ldots, n_r \text{ occurrences of } a_r\}\big] = \binom{n}{n_1, n_2, \ldots, n_r} p_1^{n_1} \cdots p_r^{n_r}$$

Example 1.41
I have a game with three outcomes: win, lose, draw. The probability of win is 0.4, lose 0.5, draw 0.1. If I play 10 times with independent outcomes, what is the probability of 4 wins, 4 lose and 2 draw?

$$\mathbb{P}[\text{win 4, lose 4, draw 2 of 10}] = \binom{10}{4,4,2} 0.4^4 0.5^4 0.1^2 = \frac{10!}{4!4!2!} \frac{1}{(2500)(25)} = \frac{(10)(9)(8)(7)(6)(5)}{(48)(2500)(25)} = \frac{(9)(7)}{(50)(25)} = 0.0504$$

Example 1.42
In a wireless channel, bits are flipped independently in error with probability 0.01. If you transmit 100 bits, what is the probability that 3 of them are flipped?

$$\mathbb{P}[\{\text{ 3 of 100 bits are flipped}\}] = \binom{100}{3}(0.01)^3(0.99)^{97} \approx 0.061.$$

For the same example, if we have an error correcting code that can correct up to three bits in error, what is the probability that all bits are recovered without error? This is the probability that no bits are flipped, plus the probability that one bit is flipped, plus the probability that two bits are flipped. That is,

$$\begin{aligned}\mathbb{P}[\{\text{ All bits recovered}\}] &= \binom{100}{0}(0.99)^{100} + \binom{100}{1}(0.01)(0.99)^{99} + \binom{100}{2}(0.01)^2(0.99)^{98} + \binom{100}{3}(0.01)^3(0.99)^{97} \\ &\approx 0.982.\end{aligned}$$

Chapter 2

Discrete Random Variables

2.1 Random Variables

A random variable is similar to a function; indeed, the most common definition of a random variable is a function which assigns a value in the space of real numbers $\Re$ to each outcome in Ω. Recall that functions can assign only one value to each outcome.

Definition 2.1
A random variable X in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ is a function $X : \Omega \to \Re$, such that, for any interval (a, b), the set $\{\omega \in \Omega : a < X(\omega) < b\}$ belongs to the event space $\mathcal{E}$.

By constraining random variables to functions where the inverse image of an interval (a, b) is an event in $\mathcal{E}$, we can compute $\mathbb{P}[\{\omega \in \Omega : a < X(\omega) < b\}]$. As we discussed earlier in 1.8, the smallest σ-field in $\Re$ that contains the open intervals (a, b) is known as the Borel σ-field $\mathcal{B}$. Using limits and the continuity of probability measures, we can then compute for any Borel set $A \in \mathcal{B}$, the probability $\mathbb{P}[\{\omega \in \Omega : X(\omega) \in A\}]$. That is, the inverse image using the function $X(\omega)$ of any Borel set A will be an event in $\mathcal{E}$. In a more formal mathematical definition, we such functions **measurable functions** from $(\Omega, \mathcal{E})$ into $(\Re, \mathcal{B})$. Figure 2.1 illustrates the concept of a random variable.

Random variables provide a useful abstraction in probability. First, by assigning numbers to outcomes, they allow us to map outcomes onto a quantitative scale, which will allow us to compute interesting statistics. More important, they allow us to recognize that many different experiments give rise to the same type of random variables, and thus can be analyzed by a common methodology without worrying about the individual details of the experiments. For instance, we discussed in the previous chapter the concept of Bernoulli trials as an experiment with two outcomes. That experiment can be a coin flip, a race between two people, a bet, a roll of a pair of dice to get a total of 7, a shot at a target, etc. By mapping one outcome to the number 1, and the other outcome to 0 we get a Bernoulli random variable. Thus, the analysis of Bernoulli random variables provides the tools for analysis in all the diverse experiments that give rise to such random variables. Similar abstractions will allow us to use a common set of random variables to analyze measurement errors that arise in acoustic, aerospace, electronic and biomedical measurements.

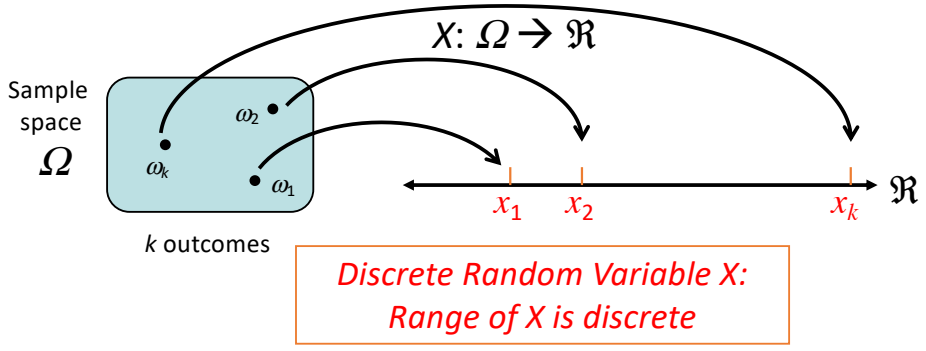


Figure 2.1: Discrete random variables map Ω into a discrete set of values in the real line.

We introduce some notation that we will use throughout the book: We use capital letters (e.g. X, Y, Z) to denote random variables, and we use lower case letters *(e.g. x, y, z) to denote the values that a random variable takes.

We denote by R_X the image of the sample space Ω (the **range**) as mapped by the random variable $X(\cdot)$. That is, $R_X \subset \Re$ is the set of possible values of $X(\omega), \omega \in \Omega$.

Definition 2.2
A **discrete random variable** in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ is a random variable X such that the range of X, denoted by $R_X = X(\Omega)$, has at most a countable number of elements.

We sometimes make a distinction to refer to a random variable as **finite** if R_X has a finite number of elements. Random variables that are not discrete can be either continuous or hybrid as described in the next chapter.

Example 2.1
Turn on a light source, and use a CCD detector to count the number of photons that hit the detector in an interval of one second. In this experiment, $\Omega = \{0, 1, 2, \ldots\}$. We define the random variable $X(\omega) = \omega$, as the outcomes are already numeric. The range R_X of this random variable is $R_X = \{0, 1, 2, \ldots\}$. X is a discrete random variable, as its range is discrete.

Example 2.2
Turn on a light source, and have a CCD detector that measures the time between the arrival of the first photon and the arrival of the second photon. In this experiment, $\Omega = [0, \infty)$. We define the random variable $X(\omega) = \omega$, as the outcomes are already numeric. The range R_X of this random variable is $R_X = [0, \infty)$, which is not countable. This random variable X is not a discrete random variable.

Suppose we define a different random variable $Y(\omega)$ as follows:

$$Y(\omega) = \begin{cases} 0 & \omega < 2ns, \\ 1 & \text{elsewhere.} \end{cases}$$

In this case, the range $R_Y = \{0, 1\}$, which is finite, so Y is a discrete random variable.

Most card experiments, dice experiments and coin flip experiments give rise to discrete random variables. We list some examples of discrete random variables below.

- The number of X-Ray photons detected in a pixel by an X-ray radiograph.
- The number of defective parts in a manufacturing process in 10 minutes.
- . The presence of a disease in a patient.
- The correctness of a software implementation of an algorithm.
- The number of parts that fail in an automobile in the course of a year.

Typical examples of random variables that are not discrete are the time until a part fails in an assembly plant, the error in location given by a GPS system, the error in measuring the distance to an obstacle using a LIDAR sensor and the time of arrival of customer at a service station.

A random variable X induces a probability measure $\mathbb{P}_X$ on $(\Re, \mathcal{B})$ using the function mapping. For any intervals $(a, b) \in \Re$, this probability is given by

$$\mathbb{P}_X((b, a)) = \mathbb{P}[\{\omega \in \Omega : b < X(\omega) < a\}]$$

and, more generally, for any set $B \in \mathcal{B}$, we have

$$\mathbb{P}_X(B) = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}].$$

Indeed, with this induced probability, we can show that $(\Re, \mathcal{B}, \mathbb{P}_X)$ is also a probability space. We call this space the sample space. The abstractions that random variables provide will allow us to use the same induced probability space for many different random experiments.

Example 2.3
Consider the experiment of tossing two unbiased coins. In the original space Ω, there are four outcomes: HH, HT, TH and TT, where H denotes a heads outcome and T denotes a tails outcome. We define a random variable X as follows:

$$X(\omega) = \begin{cases} -1 & \text{if } s \neq \text{HH, TT.} \\ 1 & \text{otherwise.} \end{cases}$$

In this experiment, $R_x = \{-1, 1\}$. The induced probability $\mathbb{P}_X$ can be defined on its atoms, so that $\mathbb{P}_X[\{1\}] = \mathbb{P}[\{\omega \in \Omega : X(\omega) = 1\}] = \mathbb{P}[\{HH, TT\}] = 0.5$. Similarly, $\mathbb{P}_X[\{-1\}] = 0.5$.

Now, consider a second experiment, consisting of tossing a single unbiased coin, with sample space $\Omega_1 = \{H, T\}$, and define variable Y as

$$Y(\omega) = \begin{cases} -1 & \text{if } \omega = H \\ 1 & \{textotherrwise. \end{cases}$$

The sample space and induced probability of this random experiment and random variable Y are the same as those of the first experiment and random variable X. Rather than treating these random variables a different, by using the sample space, we can treat them as identical random variables.

2.2 Discrete Random Variables

Consider a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, with a discrete random variable X defined on it, with values in $\{x_1, x_2, x_3, \ldots\}$. Since every set $\{x_i\}$ containing a singleton value is a Borel set, we can compute the probability $\mathbb{P}[\{\omega \in \Omega : X(\omega) = x_i\}]$. We can use this to define the induced probability measure $\mathbb{P}_X$ on R_X. We define this formally next.

2.2.1 Probability Mass Function

Definition 2.3
The **probability mass function** of a discrete random variable X defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, taking values in $\{x_1, x_2, x_3, \ldots\}$ is the function $P_X(x_i) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x_i\}]$.

To keep the notation simple, we refer to the set $\{X = x_1\} \equiv \{\omega \in \Omega : X(\omega) = x_i\}$. Thus, we will write equivalently the following forms for the probability mass function of a discrete random variable:

$$P_X(x) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] \equiv \mathbb{P}[\{X = x\}] \equiv \mathbb{P}[X = x].$$

In each case, it should be clear that this is computing the probability of an event $A \in \mathcal{E}$ defined all possible solutions of the equation $X(\omega) = x$. Figure 2.2 illustrates a probability mass function for a discrete random variable.

The probability mass function (PMF) of a random variable X satisfies the following basic properties:

1. **Non-negativity:** $P_X(x) \geq 0$ for all x.
2. **Normalization:** $\sum_{x \in R_X} P_X(x) = 1$.

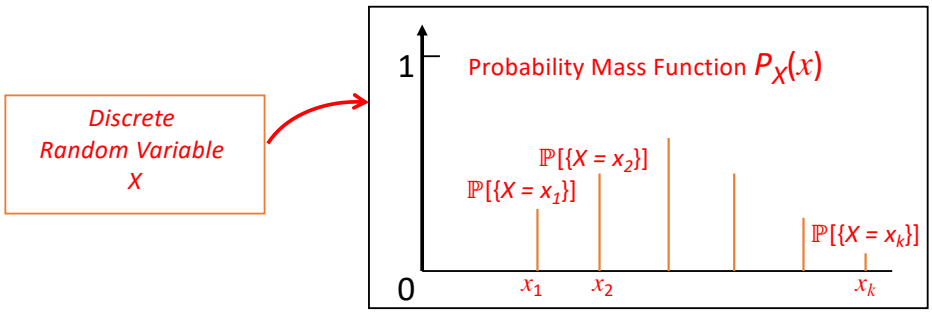


Figure 2.2: Illustration of a Probability Mass Function.

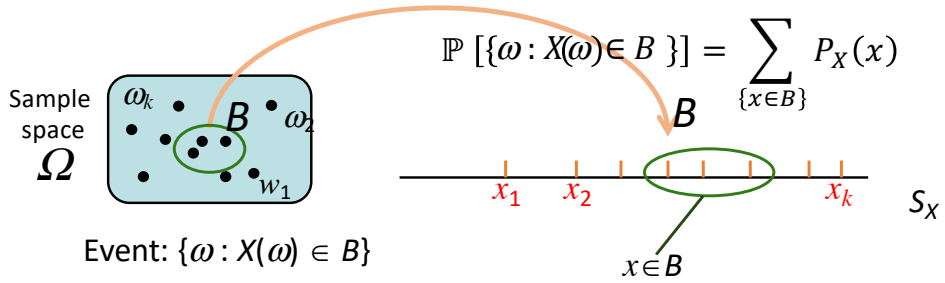


Figure 2.3: Computing the probability of events using the PMF.

3. **Additivity:** For any subset $B \subset R_X$, the probability that X falls in B is

$$\mathbb{P}_X[B] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}] = \sum_{x \in B} P_X(x).$$

Note that $\mathbb{P}_X[B]$ implicitly refers to the event $\mathbb{P}_X[B] = \mathbb{P}\big[\{X \in B\}\big]$.

The Additivity property follows because the event $\{\omega \in \Omega : X(\omega) \in B\}$ can be decomposed into disjoint events $\{\omega \in \Omega : X(\omega) = x_i\}$ for each $x_i \in B$. These events are disjoint because $X(\omega)$ is a function and thus can only assign a single value to each $\omega \in \Omega$. Then, the countable additivity property of the probability measure shows

$$\mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}] = \sum_{x \in B} \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] = \sum_{x \in B} \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] = \sum_{x \in B} P_X(x).$$

Figure 2.3 illustrates the approach at computing probabilities of events using the additivity property of the probability mass function. Any event in R_X will contain discrete elements x_i on which the probability mass function is defined. The induced probability of the event is the sum of the probability mass function on the elements that are in B.

Example 2.4
In this experiment, we roll two four-sided dice, with all outcomes on each dice being equally likely. Note that these dice are tetrahedral, so the number that a die rolls is the number at the bottom. We define the random variable X to be the sum of the numbers at the bottom of the dice.

The sample space is $\Omega = \{(i, j) : i, j = 1, 2, 3, 4\}$. The image $R_X = \{2, 3, 4, 5, 6, 7, 8\}$. Since this is a discrete set, we

compute the PMF as:

$$P_X(2) = \mathbb{P}[\{(1,1)\}] = \frac{1}{16}$$
$$P_X(3) = \mathbb{P}[\{(1,2),(2,1)\}] = \frac{1}{8}$$
$$P_X(4) = \mathbb{P}[\{(1,3),(2,2),(3,1)\}] = \frac{3}{16}$$
$$P_X(5) = \mathbb{P}[\{(1,4),(2,3),(3,2),(4,1)\}] = \frac{1}{4}$$
$$P_X(6) = \mathbb{P}[\{(2,4),(3,3),(4,2)\}] = \frac{3}{16}$$
$$P_X(7) = \mathbb{P}[\{(3,4),(4,3)\}] = \frac{1}{8}$$
$$P_X(8) = \mathbb{P}[\{(4,4)\}] = \frac{1}{16}$$

Using this, define the event $B = \{X \text{ is even }\}$. Then,

$$\mathbb{P}_X[B] = P_X(2) + P_X(4) + P_X(6) + P_X(8) = \frac{8}{16} = \frac{1}{2}.$$

Define the event $C = \{X \text{ is a multiple of 3 }\}$. Then,

$$\mathbb{P}_X[C] = P_X(3) + P_X(6) = \frac{5}{16}.$$

Example 2.5
In an experiment, we have a biased coin with two outcomes, H and T, with probability of H $= p > 0$. We are going to toss that coin an infinite number of times, so that an outcome of the experiment is an infinite sequence of Hs and Ts; e.g. HHHTHTTHTHHHTTTTTH.... The outcomes of each coin toss are independent, so this defines the outcomes in the original probability space as well as the underlying probabilities, e.g. we have $(\Omega, \mathcal{E}, \mathbb{P})$. On this probability space, we define a random variable $X(\omega)$ for an outcome $\omega \in \Omega$ as the position of the first H in ω. That is, $X(\text{TTHTHT}\dots) = 3, X(\text{THTTH}\dots) = 2$, etc. Note the possible values of X are discrete and countable. Find the probability mass function $P_X(x)$, and compute the induced probability of the event $B = \{X(\omega) \in [2,3]\}$.

We note that all outcomes for $X(k) = 3$ have to start with TTH, and the rest of the outcomes after the first toss result in the same $X(k)$. Using this reasoning, we can derive

$$P_X(x) = p(1-p)^{x-1}, x = 1, 2, \dots .$$

The induced probability $\mathbb{P}_X[B] = P(2) + P(3) = p(1-p) + p(1-p)^2 = p(1-p)(2-p)$.

Random variables of this type are called geometric random variables, because of the geometric decay of the PMF as x increases.

Example 2.6
This example shows we don't need to know anything about the underlying experiment if we know the probability mass function to compute probabilities for events defined in terms of the random variable. Assume $R_X = \{1,2,3,4\}$, and let the probability mass function be $P(x) = \frac{c}{x}$ for some $c > 0$. Find the value of c, and find the probabilities of the events $A = \{X \geq 2\}$ and $B = \{X < 3\}$.

We use the normalization property to compute c, since

$$P(1) + P(2) + P(3) + P(4) = c + \frac{c}{2} + \frac{c}{3} + \frac{c}{4} = \frac{25}{12}c = 1.$$

Hence, $c = \frac{12}{25}$. Next, we compute $\mathbb{P}_X[A] = P(2) + P(3) + P(4) = \frac{13}{25}$, and $\mathbb{P}_X[B] = P(1) + P(2) = \frac{18}{25}$.

2.2.2 Cumulative Distribution Function

The **cumulative distribution function (CDF)** of a random variable X in a probability space returns the probability that a random variable X is less than or equal to a value x:

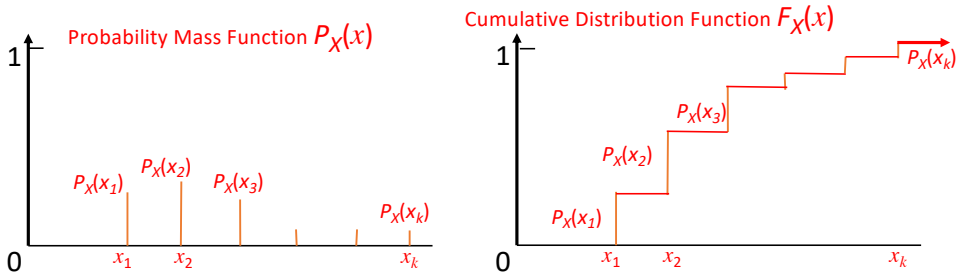


Figure 2.4: Relationship between the PMF and CDF of a random variable.

Definition 2.4 (Cumulative Distribution Function)
The **Cumulative Distribution Function** of the random variable X is defined as the function $F_X : \Re \to [0,1]$ which satisfies:

$$F_X(a) \equiv \mathbb{P}_X(\{X \in (-\infty, a]\}) = \mathbb{P}[\{\omega \in \Omega : X(\omega) \leq a\}].$$

We will sometimes use the notation $F(a)$ instead of $F_X(a)$ when it is clear which random variable we are referring to. In particular, for a generic argument, this is often written as $F_X(x)$ or just $F(x)$.

Figure 2.4 shows the relationship of the PMF and the CDF. In essence, the CDF is the sum of the PMF starting from the left at the smallest value of $x \in R_X$.

The CDF is a non-negative real-valued function $F_X(x) \in [0,1]$, defined for all real values of its argument $x \in \Re$. The CDF of any discrete random variable is a staircase function. If X takes on values $x_1, x_2, \ldots x_k$ with probabilities $P(x_1), P(x_2), \ldots, P(x_k)$, then the CDF has jumps at $x_1, x_2, \ldots x_k$ with heights $P(x_1), P(x_2), \ldots, P(x_k)$ and is flat in between the jumps.

Cumulative distribution functions have the following properties:

1. $F_X(\infty) = \lim_{x\to\infty} F_X(x) = 1, F_X(-\infty) = \lim_{x\to-\infty} = 0.$
2. $a \leq b$ implies that $F_X(a) \leq F_X(b)$, so $F(x)$ is non-decreasing in x.
3. $F_X(x)$ is piecewise constant and jumps at values of $x \in R_X \subset \Re$ such that $P(x) > 0$.
4. For all $b > a, \mathbb{P}_X[\{a < X \leq b\}] = F_X(b) - F_X(a)$.
5. $\lim_{\epsilon\to 0^+} F_X(a+\epsilon) = F_X(a)$ (continuity from the right)

Proof: The first properties follow from the continuity of probabilities. Define the events as $A_n = \{\omega \in \Omega : X(\omega) \leq n\}$. These form a non-decreasing sequence, so by Lemma 1.1

$$\lim_{n\to\infty} \mathbb{P}[A_n] = \lim_{n\to\infty} F(n) = \mathbb{P}[\cup_{k=1}^{\infty} A_n] = \mathbb{P}[\Omega] = 1$$

Similarly, the sequence $B_n = \{\omega \in \Omega : X(x) \leq -n\}$ forms a non-increasing sequence with an empty intersection, so

$$\lim_{n\to\infty} \mathbb{P}[B_n] = \lim_{n\to\infty} F(-n) = 0$$

The second property follows from the fact that $\{\omega \in \Omega : X(\omega) \leq a\} \subset \{\omega \in \Omega : X(\omega) \leq b\}$. The final property can be shown as follows: Define the sets $A_n = \{\omega \in \Omega | X(\omega) \leq a + 1/n\}$. Again, these sets are non-increasing, so

$$\lim_{n\to\infty} \mathbb{P}[A_n] = \lim_{n\to\infty} F(a+1/n) = \mathbb{P}[\cap_{n=1}^{\infty} A_n] = F(a)$$

Example 2.7
Consider the example 2.4 with two quadrilateral dice. The CDF of X is given by:

$$F_X(x) = \begin{cases} 0 & x < 2; \\ \frac{1}{16} & 2 \leq x < 3; \\ \frac{3}{16} & 3 \leq x < 4; \\ \frac{3}{8} & 4 \leq x < 5; \\ \frac{5}{8} & 5 \leq x < 6; \\ \frac{13}{16} & 6 \leq x < 7; \\ \frac{15}{16} & 7 \leq x < 8; \\ 1 & 8 \leq x. \end{cases}$$

Thus, the CDF is piecewise constant, and jumps at each integer value in R_X by the amount $P(x)$.

In general, cumulative distributions of a random variable are not very useful for computing statistics. Almost every computation uses the probability mass function instead. So why do we bother with defining CDFs and their properties? Recall that, in chapter 1, we constructed many cases where the probability of every atom is zero. For those cases, such as those involving continuous random variables, it is impossible to define a PMF. However, the concept of CDF applies to all random variables, continuous or discrete, and has nearly the same properties in all cases.

2.3 Statistics of Discrete Random Variables

We are used to seeing sample statistics in many different fields. In data science, samples are collected by repeating the same experiment independently many times, and generating the random variables associated with each of these experiments. Social statisticians work hard to select samples that correspond to the true population at large. Given a set of sample values for a random variable generated this way, a sample statistic maps these values into a single real number.

For instance, suppose the experiment is a student in EK 381 taking a midterm exam. The random variable maps the student answers into a number grade. If 80 students take the same exam, this can be viewed as repeating the experiment of selecting a student randomly 80 times and getting a value for the random variable. We assume the grading is done in whole numbers from 0 to 100, so the possible values for the random variables are discrete.

The first class after every exam, the professor is asked the same question: "What was the class average?" The class average is an example of a sample statistic. If $x_i, i = 1, \ldots, N$ are the values of the random variable X in N repetitions of the same experiment, the sample average or sample mean is defined as:

$$m_X = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Similarly, the sample variance is defined by $\mathsf{Var}[X] = \frac{1}{N} \sum_{i=1}^{N} (x_i - m_X)^2$, and the sample standard deviation is computed as $\sigma_X = \sqrt{\mathsf{Var}[X]}$. However, note that those statistics will change as N changes. In essence, they are random also, in a manner that will be made more precise later in the course. What we hope is that, as N grows, the statistics approach a limit and become constant, and thus represent an intrinsic property of a random variable.

There is another way to write the sample statistics, in terms of a sample probability mass function $\tilde{P}(x)$. In essence, compute a sample probability mass function as:

$$\tilde{P}_X(x) = \frac{1}{N} \sum_{i=1}^{N} I[x_i = x]$$

where $I[x_i = x]$ is the indicator function that is 1 if it is true, and 0 otherwise. This computes the relative frequency that the value x appears in the sample of size N. Then, using the sample probability mass function, the sample statistics can be written as:

$$m_X = \sum_{x \in R_X} \tilde{P}_X(x) x$$
$$\mathsf{Var}[X] = \sum_{x \in R_X} \tilde{P}_X(x)(x - m_X)^2$$

and $\sigma_X = \sqrt{\mathsf{Var}[X]}$.

This new form suggests that each random variable has a true statistic that can be defined in terms of its probability mass function. The sample statistics are random approximations of these true statistics.

Definition 2.5
A statistic of a discrete random variable is a map from its probability mass function to a real-valued quantity.

Below we define some of the most common statistics associated with discrete random variables.

2.3.1 Expected Value

The **expected value** of a discrete random variable X is defined as

$$\mathbb{E}[X] = \sum_{x \in R_X} x\, P_X(x).$$

This is also known as the **mean** or **average**. In these notes, we also sometimes use $\mu_X = \mathbb{E}[X]$.

The expected value has many interpretations: It is the weighted average of all possible values, using the PMF weights. It can be viewed as the center of "mass" of the PMF. Ideally, it would also be the sample average after one performs a large number repetitions of the experiment (to be substantiated later in this course): the sample mean should approach the true mean as number of samples increases!

Example 2.8
Consider the two quadrilateral dice example 2.4. Then,

$$\mathbb{E}[X] = sum_{x \in R_X} x\, P_X(x) = 2 \cdot \frac{1}{16} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{3}{16} + 5 \cdot \frac{1}{4} + 6 \cdot \frac{3}{16} + 7 \cdot \frac{1}{8} + 8 \cdot \frac{1}{16} = \frac{80}{16} = 5.$$

Note that, for some random variables where the range R_X is infinite, the expected value cannot be defined because the sum may not be finite! This is illustrated in the examples below:

Example 2.9
Assume we have a discrete random variable X with range $R_X = \{1, 2, \ldots\}$ and PMF given by $P_X(k) = \frac{6}{k^2 \pi^2}, k = 1, 2, \ldots$.

It is easy to verify that this is a valid PMF, as it is non-negative, and normalized properly because $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. This formula was derived by Leonard Euler in the early part of the 18th century. For this random variable, note that

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \frac{6}{\pi^2} \frac{k}{k^2} = \frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k} = \infty\ .$$

Thus, for statistics defined using expected values, it is possible that the statistics won't be defined if the required sums do not converge.

Example 2.10
We consider a signaling example where we want to transmit a single bit using a DC voltage. If the bit is 1, we transmit a voltage +1 volts. For the bit being 0, we transmit the voltage -1 volts. Assume that the bit is equally likely to be a 0 or a 1.

We construct the sample space for the experiment as $\Omega = \{0, 1\}$, the value of the bit. We define the random variable as the voltage $X(0) = -1, X(1) = 1$, so $R_X = \{-1, 1\}$. The probability measure in the original space is $\mathbb{P}[\{0\}] = \mathbb{P}[\{1\}] = 0.5$. The resulting PMF is given as

$$P_X(-1) = 0.5; \quad P_X(1) = 0.5.$$

Using this, we compute $\mathbb{E}[X] = 0.5 \cdot (-1) + 0.5 \cdot (1) = 0$.

Assume that we wanted to transmit two bits at a time. In this case, the sample space $\Omega = \{00, 01, 10, 11\}$, with each outcome having probability 0.25. Now, we define a new random variable Y corresponding to the voltage used for signaling, so that $Y(00) = -3, Y(01) = -1, Y(10) = 1, Y(11) = 3$.

The range space $R_Y = \{-3, -1, 1, 3\}$. The induced PMF is $P_Y(-3) = P_Y(-1) = P_Y(1) = P_Y(3) = 0.25$. Then,

$$\mathbb{E}[Y] = 0.25 \cdot (-3) + 0.25 \cdot (-1) + 0.25 \cdot (1) + 0.25 \cdot (3) = 0.$$

Thus, the two signaling schemes X, Y have the same expected value 0. However, they will differ in other statistics, such as average energy, where you can expect that the energy is proportional to X^2 or Y^2. To do this, we need to be able to compute averages of functions of random variables such as X^2.

2.4 Functions of a Random Variable

Consider a random variable X defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$. X is a function mapping outcomes in Ω into real numbers in $\Re$. Suppose we now define another function $g(\cdot)$ mapping a real number into another real number (e.g. $g : \Re \to \Re$.) Then, the composition of the two functions, $g(X(\omega))$ also maps outcomes in Ω into real numbers in $\Re$, so that each outcome is only mapped into a single real number. That is, the composition of the two functions is also a function. As long as the function $g(\cdot)$ is well behaved (measurable in the context discussed earlier), this composite function also defines a random variable in $(\Omega, \mathcal{E}, \mathbb{P})$! We denote this random variable as $Y = g(X)$ to indicate that the variable Y is derived by a function transformation of the random variable X, and the underlying random variable map $Y(\omega) \equiv g(X(\omega))$.

Note that this raises an interesting observation: we can define multiple random variables on the same probability space. We will explore this fully in later chapters. For the moment, let's focus on the case where $Y(\omega) = g(X(\omega))$. This case is often referred to as a **derived random variable**.

What is the range of Y as a random variable? It is derived from the range of X: $R_Y = \{g(x) : x \in R_X\}$. If X is a discrete random variable, then R_X is a countable, discrete range, and therefore R_Y will also be at most countable and discrete. Note that R_X is countably infinite does not imply R_Y will be, as the function $g(\cdot)$ may map many numbers in R_X into a single number in R_Y. For example, consider the function $g(\cdot)$ defined below that maps $\{1, 2, 3, \ldots, \}$ into $\{0, 1\}$:

$$g(x) = \begin{cases} 1 & \text{if } x \text{ is an odd positive integer} \\ 0 & \text{elsewhere} \end{cases}$$

If we know the function $g(\cdot)$ and the probability mass function of X, $P_X(x)$, we can compute the probability mass function for Y directly as $P_Y(y) = \sum_{x:g(x)=y} P_X(x)$, where the notation $\sum_{x:g(x)=y}$ means sum over each value of $x \in R_X$ such that $g(x) = y$ is satisfied. This is exactly the same approach we took to computing the probability mass function $P_X(x)$: The event $\{Y = y\}$ has an inverse image through the function g which is composed of a subset of R_X, which is $\{x \in R_X : g(x) = y\}$. Since R_X is discrete, this is a discrete set,

and

$$P_Y(y) = \mathbb{P}_X(\{x \in R_X : g(x) = y\}) = \sum_{x:g(x)=y} P_X(x).$$

As long as we have the properties of X, as summarized by its probability mass function P_X, we can compute all the properties of Y without having to refer to the original probability space $(\Omega, \mathcal{E}, \mathbb{P})$. We illustrate this with examples.

Example 2.11
Consider a discrete random variable X with values in $R_X = \{1, 2, 3, 4\}$ and probability mass function

$$P_X(x) = \begin{cases} 1/3 & x \leq 2 \\ 1/6 & x > 2 \end{cases}$$

Let $g(x) = x^3$ be a function, and define $Y = g(X)$ as a derived random variable. In this case, the range of Y is $R_Y = g(R_X) = \{1, 8, 27, 64\}$. The probability mass function of Y is now

$$P_Y(y) = \sum_{x:g(x)=y} P_X(x) = \begin{cases} 1/3 & y \leq 8 \\ 1/6 & y > 8 \end{cases}$$

Now, let's repeat the exercise for a different function: let $h(x) = 0$ for $x \leq 3$, and $h(x) = 1, x > 3$. Define $Z = h(X)$ be the resulting derived random variable. Then, $R_Z = h(R_X) = \{0, 1\}$, and the resulting probability mass function is

$$P_Z(z) = \sum_{x:h(x)=z} P_X(x) = \begin{cases} P_X(1) + P_X(2) + P_X(3) = 5/6 & z = 0 \\ P_X(4) = 1/6 & z = 1 \end{cases}$$

Example 2.12
Consider now the signaling example 2.10. Let $U = X^2$. Then, $U(-1) = 1, U(1) = 1$, so $R_U = \{1\}$. Hence, $P_U(1) = P_X(-1) + P_X(1) = 1$. Hence, $\mathbb{E}[U] = 1$.

Define $V = Y^2$. Then, $V(-3) = V(3) = 9, V(-1) = V(1) = 1$. Thus, $R_V = \{1, 9\}$, and $P_V(1) = P_Y(-1) + P_Y(1) = 0.5; P_V(9) = P_Y(-3) + P_Y(3) = 0.5$. The average is:

$$\mathbb{E}[V] = 1P_V(1) + 9P_V(9) = 5.$$

So, on average, signaling with two bits at a time in this scheme takes much more energy than signaling each bit separately.

For a derived random variable Y, we can compute all of its statistics using its probability mass function $P_Y(y)$. However, there is a simpler approach that avoids the need for computation of $P_Y(y)$. Consider computation of the expected value of Y (its mean). Using the approach in subsection 2.3, we compute $\mathbb{E}[Y]$ as

$$\mathbb{E}[Y] = \sum_{y \in R_Y} y P_Y(y)$$

However, note that, using the definition of $P_Y(y)$

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{y \in R_Y} y P_Y(y) = \sum_{y \in R_Y} y \sum_{x:g(x)=y} P_X(x) \\ &= \sum_{y \in R_Y} \sum_{x:g(x)=y} g(x) P_X(x) \quad \text{(since } y = g(x) \text{)} \\ &= \sum_{x \in R_X} g(x) P_X(x) \quad \text{(since } g \text{ is a function, and every } x \in R_X \text{ is mapped into some } y \in R_Y) \end{aligned}$$

Thus, we can compute $\mathbb{E}[Y]$ directly using the definition of the function $g(\cdot)$ and the probability mass function P_X without having to compute P_Y.

Example 2.13
Back to the signaling example 2.10, we compute directly:

$$\mathbb{E}[X^2] = (-1)^2 P_X(-1) + (1)^2 P_X(1) = 1.$$

$$\mathbb{E}[Y^2] = (-3)^2 P_X(-3) + (-1)^2 P_X(-1) + (1)^2 P_X(1) + (3)^2 P_X(3) = 5.$$

Let's focus now on a special class of functions: affine functions $g(x) = ax + b$. Let $Y = g(X) = aX + b$. Then,

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{x \in R_X} g(x) P_X(x) = \sum_{x \in R_X} (ax + b) P_X(x) \\ &= a \sum_{x \in R_X} x P_X(x) + b \sum_{x \in R_X} P_X(x) \\ &= a\mathbb{E}[X] + b \quad \text{(using the definition of } \mathbb{E}[X] \text{ and the normalization property of } P_X.)\end{aligned}$$

Thus, when a random variable Y is defined by an affine transformation of a random variable X, its expected value is computed by the same affine transformation of the expected value of X, avoiding having to do any summations over P_X.

An important statistic that we use to characterize the randomness in random variables is the variance. The **variance** measures how spread out a random variable is around its mean, and is defined by

$$\mathsf{Var}[X] = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \sum_{x \in R_X} (x - \mu_X)^2 P_X(x).$$

Note that $\mathbb{E}[X]$ is a number, not a random variable. Hence, $Z = (X - \mathbb{E}[X])^2$ is transformation of the variable X. The variance of X is often referred to as $\sigma_X^2 = \mathsf{Var}[X]$, where σ_X is the positive square root of the variance of X, and is known as the **standard deviation.**

Example 2.14
Let X be a random variable, with $R_X = \{1, 3, 5\}$ and PMF $P_X(1) = P_X(3) = P_X(5) = \frac{1}{3}$. Then,

$$\mathbb{E}[X] = (1)P_X(1) + (3)P_X(3) + (5)P_X(5) = 3.$$

$$\mathsf{Var}[X] = (1-3)^2 P_X(1) + (3-3)^2 P_X(3) + (5-3)^2 P_X(5) = \frac{8}{3}.$$

The standard deviation is $\sigma_X = \sqrt{\frac{8}{3}}$.

There is an alternative formula for computing the variance of a random variable which is $\mathsf{Var}[X] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2$. We can show this as follows:

$$\begin{aligned}\mathsf{Var}[X] &= \sum_{x \in R_X} (x - \mu_X)^2 P_X(x) \\ &= \sum_{x \in R_X} (x^2 - 2x\mu_X + \mu_X^2) P_X(x) \\ &= \sum_{x \in R_X} x^2 P_X(x) - 2 \sum_{x \in R_X} x\mu_X P_X(x) + \sum_{x \in R_X} \mu_X^2) P_X(x) \\ &= \sum_{x \in R_X} x^2 P_X(x) - 2\mu_X \sum_{x \in R_X} x P_X(x) + \mu_X^2 sum_{x \in R_X} P_X(x) \\ &= \sum_{x \in R_X} x^2 P_X(x) - 2\mu_X^2 + \mu_X^2 = \sum_{x \in R_X} x^2 P_X(x) - \mu_X^2\end{aligned}$$

where the last line follows from the definition $\mu_X = \sum_{x \in R_X} x P_X(x)$ and the normalization property of the probability mass function $sum_{x \in R_X} P_X(x) = 1$.

Assume again we have an affine transformation $Y = aX + b$. We know that $\mathbb{E}[Y] = a\mathbb{E}[X] + b$. Can we compute a relationship for the variance of Y in terms of the variance of X? Reasoning as above, we obtain

$$\begin{aligned}
\mathsf{Var}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] &= \sum_{y \in R_Y} (y - \mathbb{E}[Y])^2 P_Y(y) \\
&= \sum_{x \in R_X} (ax + b - \mathbb{E}[Y])^2 P_X(x) \\
&= \sum_{x \in R_X} (ax + b - a\mathbb{E}[X] - b)^2 P_X(x) \\
&= \sum_{x \in R_X} a^2 (x - \mathbb{E}[X])^2 P_X(x) \\
&= a^2 \mathsf{Var}[X]
\end{aligned}$$

Note that the constant b in the transformation $Y = aX + b$ affects the mean $\mathbb{E}[Y]$, but does not affect the variance, because the variance is a measure of the variation of Y about its mean. Notice also that the scaling factor a is squared in the variance, as the variance is a quadratic statistic. In terms of standard deviations, we have $\sigma_Y = |a|\sigma_X$.

To illustrate that the constant b does not affect the variance, consider the special transformation $Y = X - \mathbb{E}[X]$, where $a = 1$ and $b = -\mathbb{E}[X]$. In this special case,

$$\mathbb{E}[Y] = \mathbb{E}[X] - \mathbb{E}[X] = 0; \mathsf{Var}[Y] = \mathbb{E}[Y^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathsf{Var}[X]$$

which highlights that the variance of a random variable does not change when it is shifted by a constant.

These results provide a shortcut for computing statistics of derived random variables when the transformation $Y = aX + b$ is an affine transformation:

$$\mathbb{E}[Y] = a\mathbb{E}[X] + b; \quad \mathsf{Var}[Y] = a^2 \mathsf{Var}[X]$$

The above results also highlight an important property of expectations. Suppose the function $g(x) = g_1(x) + g_2(x)$, and we define $Y = g(X) = g_1(X) + g_2(X)$. In the above linear case, $g_1(x) = ax, g_2(x) = b$. Then,

$$\mathbb{E}[Y] = \sum_{x \in R_X} (g_1(x) + g_2(x)) P_X(x) = \mathbb{E}[g_1(X)] + \mathbb{E}[g_2(X)]$$

because the sum is a linear operation, and can be separated into two sums. Also, if $Y = ag_1(X) + bg_2(X)$, then

$$\mathbb{E}[Y] = \mathbb{E}[ag_1(X) + bg_2(x)] = a\mathbb{E}[g_1(X)] + b\mathbb{E}[g_2(x)].$$

Thus, the expectation operator is a *linear operator*. We will exploit this property throughout the rest of this course.

There are other useful statistics that can be computed for a random variable X. We list a few below:

- n^{th} **Moment:** $\mathbb{E}[X^n] = \sum_{x \in R_X} x^n P_X(x)$.
- n^{th} **Central Moment:** $\mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^n\Big] = \sum_{x \in R_X} (x - \mu_X)^n P_X(x)$.
- **Median:** The median is a number $x_{med} \in \Re$ such that $\mathbb{P}_X[\{X < x_{med}\}] = \mathbb{P}_X[\{X > x_{med}\}]$. Note that such a number may not exist and, if it existed it may not be unique. For instance, consider a

random variable with two possible values, 0 or 1, and $P_X(0) = 0.1, P_X(1) = 0.9$. There is no median for this random variable. Similarly, consider another random variable with four possible values 0, 1, 2, 3, with $P_X(0) = P_X(1) = P_X(2) = P_X(3) = 0.25$. In this case, any number strictly between 1 and 2 serves as a median.

- **Mode**: The mode of a random variable X is any number x_{mod} such that $P_X(x_{mod}) \geq P_X(x)$ for all $x \in R_X$. Unlike the median, the mode of a discrete random variable must exist, but it may not be unique. The last example in the previous bullet has four possible values for the mode.

2.5 Important Families of Discrete Random Variables

Many experiments in engineering problems have the same underlying probability structure and give rise to the same type of random variable. In this section, we discuss several classes of discrete random variables that arise in many engineering applications. These classes of random variables have probability mass functions that can be described by a few parameters. Hence, they provide useful models for physical processes, as those parameters can be readily estimated from available sample data. Learning the properties of these random variables helps us avoid repetitive calculations.

The classes of random variables we discuss are:

- Bernoulli
- Uniform
- Binomial
- Geometric
- Poisson

For each family, we compute its statistics, so that we can avoid tedious summations when we can recognize the type of random variable involved.

2.5.1 Bernoulli(p) Random Variables

Let A be an event related to the outcome of some random experiment, such as a toss of a biased coin. Define the random variable X as the *indicator* function of A as:

$$X(\omega) = I_A(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is not in } A \\ 1 & \text{if } \omega \text{ is in } A. \end{cases}$$

Thus, X is one if the event A occurs, and zero otherwise. X is a random variable, with discrete values in range $\{0, 1\}$, and with probability mass function given by:

$$P_X(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1 \ . \end{cases}$$

where $p = \mathbb{P}[A]$ in the original probability space. Such a random variable is called a *Bernoulli* random variable, since it identifies the outcome of a Bernoulli trial, which is 1 if the event A occurs.

The range of a Bernoulli random variable is $R_X = \{0, 1\}$. Its CDF is computed as:

$$F_X(x) = \begin{cases} 0 & x < 0, \\ (1-p) & x \in [0, 1) \\ 1 & x > 1. \end{cases} \ .$$

Note that $F_X(x)$ is defined for all real values of x.

The expected value and other statistics of Bernoulli random variables are easily computed, since R_X only has two entries:

$$\mathbb{E}[X] = \sum_{x=0}^{1} xP_X(x) = 0(1-p) + 1p = p.$$

$$\mathbb{E}[X^2] = \sum_{x=0}^{1} x^2 P_X(x) = 0^2(1-p) + 1^2 p = p.$$

Its variance is computed as

$$\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p).$$

Bernoulli random variables are characterized by a single parameter p, which is easy to estimate from sample outcomes of the experiment. A summary of their properties is given below.

- X is a **Bernoulli**(p) random variable if it has PMF

$$P_X(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1\ . \end{cases}$$

- Range: $R_X = \{0, 1\}$.
- Expected Value: $\mathcal{E}[X] = p$.
- Variance: $\mathsf{Var}[X] = p(1-p)$.
- Interpretation: single trial with success probability p.

2.5.2 Discrete Uniform(a, b) Random Variables

Suppose we have a discrete random variable X, with range in $R_X = \{a, a+1, a+2, \ldots, b\}$, where $a \leq b$ are integers, so it can take $b - a + 1$ values. We assume that the probability mass function $P_X(x)$ is the same for each value $x \in R_X$, so that each of the values is equally likely. In this case, $P_X(x) = \frac{1}{b-a+1}, x \in R_X$, as there are $b - a + 1$ possible values, and the normalization property requires $\sum_{x \in R_X} P_X(x) = 1$.

Discrete Uniform(a, b) random variables are used commonly in models of games of chance, such as coin tosses, roulette wheels, dice rolls, where there is no assumption of bias towards any of the outcomes. The outcomes in R_X are ordered in increasing order, and are separated by one unit.

We compute the statistics of a Discrete Uniform(a, b) random variable X as follows: Its CDF is given by

$$F_X(x) = \frac{\lfloor x \rfloor - a + 1}{b - a + 1}$$

where the notation $\lfloor x \rfloor$ refers to the largest integer less than or equal to x. The expected value of X is computed as:

$$\mathbb{E}[X] = \sum_{x \in R_X} xP_X(x) = \sum_{j=a}^{b} j \frac{1}{b-a+1}$$

To do this sum, it helps to remember some summation equalities:

$$\sum_{j=1}^{n} j = \frac{n(n+1)}{2}$$

Let's define the derived random variable $Y = X - a$. Note that $R_Y = \{0, 1, 2, \ldots, b-a\}$, and $P_Y(y) = \frac{1}{b-a+1}, y \in R_y$. Now,

$$\mathbb{E}[Y] = \sum_{y \in R_Y} y P_Y(y) = \sum_{k=0}^{b-a} k \frac{1}{b-a+1} = \frac{1}{b-a+1} \frac{(b-a)(b-a+1)}{2} = \frac{b-a}{2}$$

For the original variable X, we know $\mathbb{E}[Y] = \mathbb{E}[X] - a$, so $\mathbb{E}[x] = a + \frac{b-a}{2} = \frac{a+b}{2}$.

We can compute the variance of Y as follows: First, we compute $\mathbb{E}[Y^2]$ as

$$\mathbb{E}[Y^2] = \sum_{k=0}^{b-a} k^2 \frac{1}{b-a+1}$$

To sum this, we use another summation formula:

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

Since the $k = 0$ term does not contribute to the sum ($k^2 = 0$) , we get:

$$\mathbb{E}[Y^2] = \sum_{k=0}^{b-a} k^2 \frac{1}{b-a+1} = \frac{(b-a)(b-a+1)(2(b-a)+1)}{6(b-a+1)} = \frac{(b-a)(2(b-a)+1)}{6}.$$

We compute the variance $\mathsf{Var}[Y]$ as

$$\mathsf{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{(b-a)(2(b-a)+1)}{6} - \frac{(b-a)^2}{4} = \frac{4(b-a)^2 + 2(b-a) - 3(b-a)^2}{12} = \frac{(b-a)b - a + 2}{12}$$

Since $Y = X - k$, we know $\mathsf{Var}[Y] = \mathsf{Var}[X]$.

Uniform random variables are characterized by two parameters, k and n. Their properties are summarized below:

- X is a **Discrete Uniform**(a, b) random variable if it has PMF

$$P_X(x) = \begin{cases} \frac{1}{b-a+1} & x = a, a+1, \ldots, b, \\ 0 & \text{otherwise.} \end{cases}$$

- Range: $R_X = \{a, a+1, \ldots, b\}$.
- Expected Value: $\mathcal{E}[X] = \frac{a+b}{2}$.
- Variance: $\mathsf{Var}[X] = \frac{(b-a)(b-a+2)}{12}$.
- Interpretation: equally likely to take any integer value between a and b.

2.5.3 Binomial(n, p) Random Variables

Suppose that a random experiment with a binary outcome of success or failure is repeated n times. Let x denote the number of times that such an experiment was a success. In terms of the notation used above in the context of Bernoulli random variables, let A denote an event, and let x denote the number of times that such an event occurs out of n independent trials of the same experiment. Then, X is a random variable with

discrete range $\{0, 1, \ldots, n\}$. Define the parameter p to be the probability of success in a single trial of the experiment, as in Bernoulli random variables.

A simple representation of X is given by

$$X = I_1 + I_2 + \ldots + I_n, \tag{2.1}$$

where I_k is the indicator that event A occurs at the independent trial k.

We have seen this problem worked out in Section 1.4.4. The probability of any outcome with k successes out of n is $p^k(1-p)^{n-k}$. There are $\binom{n}{k}$ outcomes with k successes. Thus, the probability mass function of X is given by

$$P_X(k) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = k\}] = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

Thus, the CDF of X is given by

$$\sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{(n-k)},$$

where $\lfloor x \rfloor$ is the largest integer that is less than or equal to x.

Binomial(n, p) random variables arise in various applications where there are two types of outcomes, and we are interested in the number of outcomes of one type. Such applications include repeated coin tosses, correct/erroneous bits, good/defective items, active/silent stations, etc. The important statistics of binomial random variables are derived below:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} \binom{n}{k} k p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} k p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \text{ (factor } np \text{ from sum)} \\
&= np \sum_{k'=0}^{n-1} \frac{(n-1)!}{(k')!(n-1-k')!} p^{k-1} (1-p)^{n-1-k'} \text{ (substitute } k' = k-1) \\
&= np
\end{aligned}$$

because the terms in the sum are the PMF for a Binomial$(n-1, p)$ RV, which add to 1 by normalization.

Similarly, to compute the variance of X, compute first the following expectation:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=0}^{n} \binom{n}{k} k(k-1)p^k(1-p)^{n-k} \\
&= \sum_{k=2}^{n} \frac{n!}{k!(n-k)!} k(k-1)p^k(1-p)^{n-k} \\
&= \sum_{k=2}^{n} \frac{n!}{(k-2)!(n-k)!} p^k(1-p)^{n-k} \\
&= n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2}(1-p)^{n-k} \text{ (factor } n(n-1)p^2 \text{ from sum)} \\
&= n(n-1)p^2 \sum_{k'=0}^{n-2} \frac{(n-2)!}{(k')!(n-2-k'!} p^{k-2}(1-p)^{n-2-k'} \text{ substitute } k' = k-2 \\
&= n(n-1)p^2
\end{aligned}$$

because the last sum is again the sum of the PMF of a Binomial$(n-2, p)$ random variable, which is 1 by normalization.

Note now that $\mathbb{E}[X(X-1)] = \mathbb{E}[X^2] - \mathbb{E}[X]$, so $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = n(n-1)p^2 + np$. Now we use the identity

$$\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = n^2p^2 - np^2 + np - n^2p^2 = n(p-p^2) = np(1-p).$$

In the above derivations, we have used extensive knowledge of binomial distributions to recognize identities, and to figure out how to factor terms so we can compute the sums. There is an alternative way of deriving these formulas, as discussed below.

Note that we can write $X = I_1 + I_2 + \ldots + I_n$, where I_k is the Bernoulli random variable indicating success in the k-th attempt. Then, using the linearity property of expectations, we have

$$\mathbb{E}[X] = \mathbb{E}[I_1 + I_2 + \ldots + I_n] = \mathbb{E}[I_1] + \mathbb{E}[I_2] + \ldots + \mathbb{E}[I_n] = np\ .$$

Note that we have avoided computing a difficult sum by using the fact that expectation is a linear operation, and the fact that, for Bernoulli random variables, $\mathbb{E}[I_k] = p$. To compute the variance, we use a property that we will derive in Chapter 5, that shows that the variance of a sum of **independent** random variables is the sum of the variances:

$$\mathsf{Var}[X] = \mathsf{Var}[I_1 + I_2 + \ldots + I_n] = \mathsf{Var}[I_1] + \ldots + \mathsf{Var}[I_n] = np(1-p).$$

Binomial random variables are characterized by the two parameters n and p. Their statistics are summarized below:

- X is a **Binomial**(n, p) random variable if it has PMF

$$P_X(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & x = 0, 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

- Range: $R_X = \{0, 1, \ldots, n\}$.
- Expected Value: $\mathcal{E}[X] = np$.
- Variance: $\mathsf{Var}[X] = np(1-p)$.
- Interpretation: # of successes in n independent Bernoulli(p) trials.

2.5.4 Geometric(p) Random Variables

The binomial random variable is obtained by fixing the number of Bernoulli trials and counting the number of successes. A different random variable is obtained by counting the number of trials until the first success occurs. Denote this random variable as X; this is a *geometric* random variable, and it takes values in the discrete infinite set $\{1, 2, \ldots\}$.

Note that $X = 1$ if and only if the first Bernoulli trial is successful. Hence, $P_X(1) = p$, where p is the single trial probability of success For $X = 2$, the first Bernoulli trial must fail, but the second one must succeed. Since the trials are independent, $P_X(2) = (1-p)p$. Reasoning along the same lines, $X = k$ if and only if the first $k-1$ Bernoulli trials failed, but the k-th Bernoulli trial succeeded. Using the independence properties, we get

$$P_X(k) = (1-p)^{k-1}p, k = 1, 2, \ldots$$

The corresponding CDF is

$$F_X(x) = 1 - (1-p)^{\lfloor x \rfloor} \ .$$

Geometric(p) random variables arise in applications where one is interested in the time between occurrence of events in a sequence of independent experiments. Such random variables have broad applications in different aspects of queuing theory. The important statistics of geometric random variables are summarized below:

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kP_X(k) = \sum_{k=1}^{\infty} kp(1-p)^k - 1$$

To sum the above expression, we use the following summation for geometric series for $0 < q < 1$:

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \ .$$

Differentiating both sides with respect to q (which is justified by the summability of the series for $p < 1$) yields:

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$$

Using this formula, we get:

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kp(1-p)^k - 1 = p\frac{1}{p^2} = \frac{1}{p} \ .$$

To compute the variance, we take another derivative of the summation equality, to get

$$\begin{aligned}\frac{d}{dq}\sum_{k=1}^{\infty} kq^{k-1} &= \sum_{k=1}^{\infty} \frac{d}{dq} kq^{k-1} \\ &= \sum_{k=1}^{\infty} k(k-1)q^{k-2} = \sum_{k=1}^{\infty} k(k-1)q^{k-2} \\ &= \frac{d}{dq}\frac{1}{(1-q)^2} = \frac{2}{(1-q)^3}\end{aligned}$$

Substituting $1 - p = q$ yields $\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} = \frac{2}{p^3}$.

Using these formulas allows us to compute $\mathbb{E}[X^2]$ as

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=0}^{\infty} k^2 p(1-p)^{k-1} = \sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1} + \sum_{k=1}^{\infty} kp(1-p)^{k-1} \\ &= \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{(2-p)}{p^2}\end{aligned}$$

Hence,

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

The geometric distribution is specified by a single parameter, p. Its properties are summarized below:

- X is a **Geometric**(p) random variable if it has PMF

$$P_X(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

- Range: $R_X = \{1, 2, \ldots\}$.
- Expected Value: $\mathcal{E}[X] = \dfrac{1}{p}$.
- Variance: $\mathsf{Var}[X] = \dfrac{1-p}{p^2}$.
- Interpretation: # of independent Bernoulli(p) trials until first success.

2.5.5 Poisson(λ) Random Variables

In many applications, we are interested in counting the number of occurrences of an event in a certain time period or in a certain region of space. The Poisson random variable arises in situations where the events occur "completely at random" in time or space; that is, where the likelihood of an event occurring at a particular time is equal to and independent of the event occurring at a different time. For example, Poisson random variables arise in counts of emissions from radioactive substances, in the number of photons emitted as a function of light intensity, in counts of demands for telephone connections, and in counts of defects in a chip.

One of the applications of the Poisson random variable is as an approximation to the binomial probabilities when the number of trials is large. If the number of trials n is large, and if p is small, then, letting $\lambda = np$, Simeón Poisson established this limit:

$$\lim_{n\to\infty, np=\lambda} \frac{n|}{k!(n-k)!} p^k (1-p)^{n_t - k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

We briefly overview his proof below. Let K_n be the binomial random variable for n trials, each of which has probability λ/n of succeeding. The probability mass function of K_n is

$$P_{K_n}(k) = nchoosek(\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} (1 - \frac{\lambda}{n})^{n-k}$$

Note the following limits:

$$\lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} = 1 \text{(same highest order power in numerator, denominator)} .$$

$$\lim_{n\to\infty} (1 - \frac{\lambda}{n})^n = e^{-\lambda} \text{ (Definition of exponential).}$$

$$\lim_{n\to\infty} (1 - \frac{\lambda}{n})^k = 1 .$$

Thus,

$$\lim_{n\to\infty} P_{K_n}(k) = \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \lim_{n\to\infty} \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Poisson(λ) random variables have an infinite, countable sample space $R_X = \{0, 1, 2, \ldots\}$ with the probability mass function

$$P_X(k) = \frac{\lambda^k}{k!} e^{-\lambda} \ .$$

where λ is the average number of event occurrences in the specified time interval or region of space. The corresponding CDF of X is

$$F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \frac{\lambda^k e^{-\lambda}}{k!} \ .$$

To compute the mean and variance of a Poisson(λ) random variable, we use a well-known summation formula

$$\sum_{k=0}^{\infty} \frac{(\lambda)^k}{k!} = e^{\lambda}$$

Then,

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=0}^{\infty} k P_X(k) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \quad \text{since the last sum is equal to } e^{\lambda}. \end{aligned}$$

To compute the variance, we compute the second moment first:

$$\mathbb{E}[X^2] = \sum_{k=0}^{\infty} k^2 P_X(k) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda}$$

In order to get an expression for this sum, we differentiate the exponential summation twice with respect to λ, to obtain

$$\frac{d^2}{d\lambda^2} e^{\lambda} = e^{\lambda} = \sum_{k=0}^{\infty} \frac{d^2}{d\lambda^2} \frac{\lambda^k}{k!} = \sum_{k=2}^{\infty} (k^2 - k) \frac{\lambda^{k-2}}{k!}$$

Therefore,

$$\mathbb{E}[X^2] = \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{(k)!} e^{-\lambda} = \sum_{k=1}^{\infty} (k^2 - k) \frac{\lambda^k}{(k)!} e^{-\lambda} + \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k)!} e^{-\lambda} = \lambda^2 + \lambda$$

We now compute the variance of X as

$$\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda \ .$$

Poisson(λ) random variables are specified by a single parameter λ. Its properties are summarized below:

- X is a **Poisson**(λ) random variable if it has PMF

$$P_X(x) = \begin{cases} \dfrac{\lambda^x}{x!} e^{-\lambda} & x = 0, 1, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

- Range: $R_X = \{0, 1, \ldots\}$.
- Expected Value: $\mathcal{E}[X] = \lambda$.
- Variance: $\mathsf{Var}[X] = \lambda$.

- Interpretation: # of arrivals in a fixed time window.

Example 2.15
Suppose we are at a service facility, with a total number of five servers. Assume there are seven potential customers in the facility, and the probability that any of them will require service is p, where each customer will require service independent of any other customers' requirements. Let X be the random variable denoting the number of service requests. What type of random variable is X? What is the expected number of requests? What is the probability that there will be more requests than available servers?

First, we recognize that the random variable X is a binomial random variable, as the sum of independent Bernoulli random variables (0-1 requests), with parameters $n = 7$ and p. The expected number of requests is thus $7p$. The probability that there will be more requests than available servers is

$$\mathbb{P}_X[\{X > 5\}] = P_X(6) + P_X(7) = \binom{7}{6} p^6(1-p) + \binom{7}{7} p^7 = 7p^6(1 - [p) + p^7 \ .$$

Example 2.16
You are waiting for a taxi at the corner of St. Mary's street and Commonwealth Avenue. When a taxi goes by the corner, there is a 0.9 probability that the taxi is occupied, and will not stop to pick you up. Assume that whether a taxi is occupied or not is independent of whether other taxis are occupied. Let X denote the number of taxis that come by the corner until one of them picks you up. What type of random variable is X? What is the expected number of taxis that you will see until you are picked up?

We recognize that whether each taxi is occupied or not is a Bernoulli trial, and the probability of success is $p = 0.1$. The random variable X is thus a geometric random variable. The expected number of taxis that you should expect to see until being picked up is thus $\mathbb{E}[X] = 10$.

Example 2.17
. Assume you have an X-ray source generating an X-ray beam with intensity equal to 10^5 photons/second towards a detector. Let X denote the number of photons collected by the detector photons over a period of a millisecond. If X is a Poisson random variable, what are its mean and standard deviation?

We compute the parameter $\lambda = 10^5 \cdot 10^{-3} = 100$ for the Poisson distribution of X. In this case, $\mathbb{E}[X] = 100, \mathsf{Var}[X] = 100$. Thus, the standard deviation is $\sigma_X = \sqrt{\mathsf{Var}[X]} = 10$.

Example 2.18
Suppose each episode of Game of Thrones includes a death of a major character with probability $3/4$, independent of whether deaths happen in any other episode. Assume there are an infinite number of episods to watch (it felt that way sometimes...) Define X to be the number of episodes you watch until you see the death of a major character. What type of random variable is X?

X is a Geometric($\frac{3}{4}$) random variable, where we explicitly provide the value for the parameter. Then, we know its statistics:

$$\mathbb{E}[X] = \frac{1}{p} = \frac{4}{3}; \mathsf{Var}[X] = \frac{1-p}{p^2} = \frac{\frac{1}{4}}{\frac{9}{16}} = \frac{4}{9}.$$

What is the probability that $X \geq 3$? Sometimes it is easier to compute the probability of the complement: the probability that $X \leq 2$. We know

$$\mathbb{P}_X[\{X \leq 2\}] = P_X(1) + P_X(2) = p + p(1-p) = \frac{3}{4} + \frac{3}{4} \cdot \frac{1}{4} = \frac{15}{16}.$$

Hence, $\mathbb{P}_X[\{X \geq 3\}] = 1 - \mathbb{P}_X[\{X \leq 2\}] = \frac{1}{16}$.

Let Y denote the number of episodes out of the first six you watch that contain a major character death. What type of random variable is Y? Y is a Binomial($6, \frac{3}{4}$) random variable. Hence, its key statistics are:

$$\mathbb{E}[Y] = np = 6 \cdot \frac{3}{4} = \frac{9}{2}; \quad \mathsf{Var}[Y] = np(1-p) = 6 \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{9}{8}.$$

What is $E[Y^2]$? We know that $E[Y^2] = E[Y]^2 + \mathsf{Var}[Y] = \frac{81}{4} + \frac{9}{8} = \frac{171}{8}$.

What is the probability that less than half of the six episodes include a death? This is $\mathbb{P}_Y[\{Y < 3\}]$, which we compute as

$$\begin{aligned}\mathbb{P}_Y[\{Y<3\}] &= P_Y(0) + P_Y(1) + P_Y(2) = (\frac{1}{4})^6 + 6(\frac{1}{4})^5(\frac{3}{4}) + \binom{6}{2}(\frac{1}{4})^4(\frac{3}{4})^2 \\ &= \frac{1+18+135}{4^6} = \frac{77}{2048}\end{aligned}$$

Given that less than half the episodes contain a death, what is the probability that exactly two of the six episodes contain a death? This is a conditional probability question: Let's define this in terms of events in the original probability space. Define event $A = \{Y = 2\}, B = \{Y \leq 2\}$. Then,

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[\{Y=2\}]}{\mathbb{P}[B]} = \frac{\frac{135}{6^4}}{\frac{154}{6^4}} = \frac{135}{154}.$$

We specialize the techniques of computing conditional probabilities to handle events defined in terms of random variables in the next section.

2.6 Conditional Probability Models

In Chapter 1, given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and an event $B \in \mathcal{E}$, we defined the conditional probability of any other event $A \in E$, conditioned on observing B, as

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

provided $\mathbb{P}[B] > 0$. Otherwise, we left the conditional probability as undefined. A discrete random variable X in $(\Omega, \mathcal{E}, \mathbb{P})$ with range in $R_X = \{x_i, i \in 1, 2, \ldots, \}$ defines events $A_i = \{\omega \in \Omega : X(\omega) = x_i\} \in \mathcal{E}$. Those events were used to define the probability mass function $P_X(x_i) = \mathbb{P}(A_i)$.

Assume we observe an event $B \in E$. We can define the conditional probability mass function of X given B as

$$P_{X|B}(x_i) = \mathbb{P}[A_i|B] = \begin{cases} \frac{\mathbb{P}[\{\omega \in \Omega : X(\omega) = x_i\} \cap B]}{\mathbb{P}[B]} & \text{if } \mathbb{P}[B] > 0 \\ \text{undefined} & \text{otherwise.} \end{cases}$$

This conditional probability mass function will have all the properties of a probability mass function on R_X, satisfying the basic properties of non-negativity, normalization and additivity:

$$\begin{aligned} P_{X|B}(x) &> 0 \text{for all } x \in R_x \\ \sum_{x \in R_x} P_{X|B}(x) &= 1 \\ \sum_{x \in C} P_{X|B}(x) &= m_{X|B}[C] \text{ for all } C \subset R_X \end{aligned}$$

There is a special case of interest, where we observe the event that X takes it values in a set $B \subset R_X$, and the conditioning event is $B_1 = \{\omega \in \Omega : X(\omega) \in B_1\}$. We are guaranteed that $B_1 \in \mathcal{E}$ is an event because X is a random variable, and $\mathbb{P}[B_1] = \mathbb{P}_X[B]$. In this special case, the conditional probability mass function simplifies: Specifically, note that

$$\{\omega \in \Omega : X(\omega) = x\} \cap B_1 = \begin{cases} \{\omega \in \Omega : X(\omega) = x\} & \text{if } x \in B\} \\ \emptyset & \text{if } x \notin B \end{cases}$$

We write, with a small abuse of notation, the conditional PMF $P_{X|B}(x)$ as

$$P_{X|B}(x) = \begin{cases} \frac{\mathbb{P}[\{\omega\in\Omega: X(\omega)=x\}]}{\mathbb{P}_X[B]} = \frac{P_X(x)}{\mathbb{P}_X(B)} & \text{if } x \in B \text{ and } \mathbb{P}_X[B] > 0 \\ 0 & \text{if } x \notin B \text{ and } \mathbb{P}_X[B] > 0 \\ \text{undefined} & \text{if } \mathbb{P}_X[B] = 0. \end{cases}$$

Thus, the conditional PMF $P_{X|B}(x)$ is proportional to the unconditional PMF $P_X(x)$, restricted to $x = B$, and rescaled to satisfy the normalization property. It is zero for any values $x \notin B$.

The conditional probability mass function has a range $R_{X|B} \subset B$, and satisfies all the properties of probability mass functions.

- (Non-negativity) $P_{X|B}(x) \geq 0$.
- (Normalization) $\sum_{x\in B} P_{X|B}(x) = 1$.
- (Additivity) For any set $C \in R_X$, the conditional probability that $X \in C$ given B is

$$\mathbb{P}[\{\omega \in \Omega : X(\omega) \in C\}|\{X(\omega) \in B\}] \equiv \mathbb{P}_{X|B}[C] = \sum_{x\in C} P_{X|B}(x).$$

Note that $\mathbb{P}_X[B] = \sum_{x_k\in B} P_X(x_k)$. Thus, we can write the conditional probability mass function of X given B entirely in terms of the random variable X and its probability mass function, as

$$P_{X|B}(x) = \begin{cases} \frac{P_X(x)}{sum_{x_k\in B} P_X(x_k)} & \text{if } x \in B \text{ and } sum_{x_k\in B} P_X(x_k) > 0 \\ 0 & \text{if } x \notin B \text{ and } sum_{x_k\in B} P_X(x_k) > 0 \\ \text{undefined} & \text{if } sum_{x_k\in B} P_X(x_k) = 0. \end{cases}$$

Now that we have a conditional probability mass function, we can define conditional statistics for the random variable X. For instance, the conditional expected value of X given an event B is given as

$$\mathbb{E}[X|B] = \sum_{x\in R_X} xP_{X|B}(x)$$

and the conditional variance as

$$\mathsf{Var}[X|B] = \mathbb{E}[(X - \mathbb{E}[X|B_1])^2|B] = \mathbb{E}[X^2|B_1] - (E[X|B])^2$$

For any function $g(X)$ that defines a derived random variable $Y = g(X)$, we can define the conditional expectation as

$$\mathcal{E}\big[g(X)|B\big] = \sum_{x\in R_X} g(x)P_{X|B}(x)\ .$$

Example 2.19
Assume X is a Binomial$(5, \frac{1}{3})$ random variable. Define $B = \{X \leq 2\}$. Compute $P_{X|B}(x)$, $\mathbb{E}[X|B]$ and $\mathsf{Var}[X|B]$.

$$P_{X|B}(x) = \begin{cases} \frac{P_X(x)}{\mathbb{P}_X[B]} & x \in B \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{P}_X[B] = P_X(0) + P_X(1) + P_X(2) = (\frac{2}{3})^5 + 5(\frac{2}{3})^4(\frac{1}{3}) + 10(\frac{2}{3})^3(\frac{1}{3})^2 = \frac{32+80+80}{3^5} = \frac{64}{81}.$$

$$P_{X|B}(0) = \frac{P_X(0)}{\mathbb{P}_X[B]} = \frac{32}{192} = \frac{1}{6}; P_{X|B}(1) = P_{X|B}(2) = \frac{80}{192} = \frac{5}{12}$$

Thus,

$$P_{X|B}(x) = \begin{cases} \frac{1}{6} & x = 0, \\ \frac{5}{12} & x = 1, 2, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[X^2|B] = 0P_{X|B}(0) + 1P_{X|B}(1) + 2P_{X|B}(2) = \frac{15}{12} = \frac{5}{4}.$$

$$\mathbb{E}[X^2|B] = 0^2P_{X|B}(0) + 1^2P_{X|B}(1) + 2^2P_{X|B}(2) = \frac{25}{12}.$$
$$\text{Var}[X|B] = \mathbb{E}[X^2|B] - (\mathbb{E}[X^2|B])^2 = \frac{25}{12} - \frac{25}{16} = \frac{25}{48}.$$

Example 2.20
Consider a manufacturing station, where the random arrival time X of a part to be processed is uniformly distributed in $R_X = \{1, 2, 3, \ldots, 20\}$. Thus, the probability mass function of X is $P_X(x) = \frac{1}{20}, x \in R_X$. Assume that you wait for the first 6 slots and the part X has not arrived yet. Equivalently, you observe the event $B = \{X > 6\}$. Compute the conditional probability mass function of X given B and its conditional expected value and variance.

Since B is defined in terms of R_X, we can use the simpler formula, restricting and rescaling the original $P_X(x)$. Note that $\mathbb{P}_X[B] = \sum_{x \in B} P_X(x) = \frac{14}{20}$. Then,

$$P_{X|B}(x) = \begin{cases} 0 & x \leq 6, \\ \frac{\frac{1}{20}}{\frac{14}{20}} = \frac{1}{14} & x > 6. \end{cases}.$$

Note that this is now a uniform distribution from 7 to 20, so we can use formulas for uniform distribution to compute mean and variance. The conditional expected value is

$$\mathbb{E}[X|B] = \sum_{x \in B} xP_{X|B}(x) = \frac{7 + 20}{2} = 13.5$$

The conditional variance is

$$\text{Var}[X|B] = \frac{(20 - 7)(20 - 7 + 2)}{12} = \frac{(13)(15)}{12} = \frac{65}{4} = 16.25.$$

Example 2.21
One of the interesting properties of a geometric random variable X is that it is "memoryless". Let X be a geometric random variable with parameter p. Assume we observe the event $B = \{X > k\}$ for some value k. What is the conditional mass distribution of X given B? Recall that $R_X = \{1, 2, \ldots, \}$, and $B = \{k + 1, k + 2, \ldots, \}$.

We compute

$$\mathbb{P}_X[B] = \sum_{k=7}^{\infty} P_X(k) = \sum_{k=7}^{\infty} p(1-p)^{k-1} = (1-p)^6 \sum_{k=1}^{\infty} p(1-p)^{k-1} = (1-p)^6$$

because we know, from normalization, that $\sum_{k=1}^{\infty} p(1-p)^{k-1} = 1$. Hence,

$$P_{X|B}(x) = \begin{cases} \frac{P_X(x)}{\mathbb{P}_X[B]} & x \geq 7 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} p(1-p)^{x-7} & x \geq 7 \\ 0 & \text{otherwise} \end{cases}$$

Define the additional wait time random variable $T = X - 6$. Then, note that

$$P_{T|B}(t) = \begin{cases} p(1-p)^t & t \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus, conditioned on B, T is a geometric random variable with the same parameter p as the original random variable X.

In words, the above expression states that, if a success has not occurred in the first j trials, the probability of having to perform at least k more trials until a success is the same as the probability of initially having to perform at least k trials. Thus, the system "forgets" the past failures and begins anew as if it were performing the first trial.

Hence, if you are waiting for a bus that should arrive in 10 minutes, and you have already waited two hours, the expected arrival time of the bus is still 10 minutes from now...as long as the arrival time was a geometric random variable.

Note that conditional probability mass functions obey the usual laws that probability mass functions obey. For instance, for a random variable X defined in the probability space $(\Omega, \mathcal{E}, \mathbb{P})$, we have:

- **Multiplication Rule:** For a random variable X and event $B \in \mathcal{E}$,

$$\mathbb{P}\big[\{X = x\} \cap B\big] = P_{X|B}(x)\, \mathbb{P}[B] \ .$$

  If $B \subset R_X$, then

$$\mathbb{P}\big[\{X = x\} \cap \{X \in B\}\big] = P_{X|B}(x)\mathbb{P}_X[B] = \begin{cases} P_X(x) & x \in B \\ 0 & \text{otherwise.} \end{cases}$$

- **Law of Total Probability:** For a partition of R_X as $B_1, \ldots, B_n$, we can write the probability mass function as a weighted sum of conditional probability mass functions, as:

$$P_X(x) = \sum_{i=1}^{n} P_{X|B_i}(x)\mathbb{P}_X[B_i] \ .$$

- **Bayes' Rule:** We can "flip" the conditioning, as in Bayes' Rule, with some care. Let $B \subset R_X$. Then,

$$\mathbb{P}_X\big[B\big|\{X = x\}\big] = \frac{P_{X|B}(x)\mathbb{P}_X[B]}{P_X(x)} \ .$$

Chapter 3

Continuous Random Variables

3.1 Introduction

In the previous chapter, we described a random variable X as a measurable function from a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ to the real line $\Re$. We then focused on studying discrete random variables where the range of X, denoted by $R_X = X(\Omega)$, has a discrete, possibly countably infinite number of elements. However, what about random variables where the range of X, denoted by R_X has an uncountable number of elements? This is illustrated in Figure 3.1 where the range of X maps Ω into an interval $[a, b]$. Suppose we wanted to define a uniform probability on that interval $[a, b]$. In this case, it is impossible to assign a probability mass to any point $x \in [a, b]$, other than 0, because we could not satisfy both the additivity property (the probability of the union of disjoint sets is the sum of the probabilities of the individual sets) and the normalization property (the probability of that $X(\omega) \in R_X$ equals 1).

In cases where the range R_X is uncountable, it is common that one cannot associate a nonzero probability with any individual outcome. Since there are uncountably many values of the random variable $X(\omega), \omega \in S$, we focus on defining probabilities of events, and not individual outcomes. In terms of events, our focus will be on events generated by the random variable X taking values in Borel sets: sets generated by countable unions, complements and intersections of intervals. By restricting the random variable X to be measurable, we guarantee that the inverse image of such a Borel set B, $\{\omega \in \Omega : X(\omega) \in B\}$ is an event in the event space $\mathcal{E}$, and thus has a probability assigned to it by the measure $\mathbb{P}$.

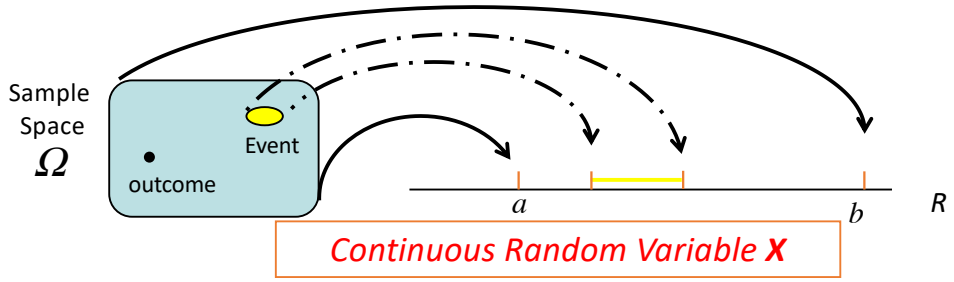


Figure 3.1: A continuous random variable has an uncountable range.

3.2 Continuous Random Variables

In Chapter **??**, section **??**, we defined the cumulative distribution function of a random variable X as:

$$F_X(a) \equiv \mathbb{P}_X(\{X \in (-\infty, a]\}) = \mathbb{P}[\{\omega \in \Omega : X(\omega) \leq a\}].$$

This definition is valid for all random variables, independent of whether the range R_X is discrete or not. The function $F_X(a)$ is defined for all $a \in \Re$.

This cumulative distribution function had the following properties:

1. (**Non-negativity**) $F_X(x) \geq 0$.

2. (**Normalization**) $F_X(\infty) = 1, F_X(-\infty) = 0$
3. (**Monotonicity**) $a \leq b$ implies that $F_X(a) \leq F_X(b)$
4. (**Right-continuity**) $\lim_{\epsilon \to 0^+} F_X(a + \epsilon) = F_X(a)$ (continuity from the right)

3.2.1 Cumulative Distribution Function

We will use the cumulative distribution of X to define a **continuous random variable**, although we will wait for a more precise definition later. Unlike discrete random variables, a **continuous random variable** must have a continuous cumulative distribution function (CDF) $F_X(x)$, as illustrated in Figure 3.2. Discontinuities in CDFs occur at values x which occur with positive probability, so that $\mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] > 0$. For continuous random variables, we want the probability $\mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] = 0$ for all $x \in \Re$. Hence, the CDF must be continuous.

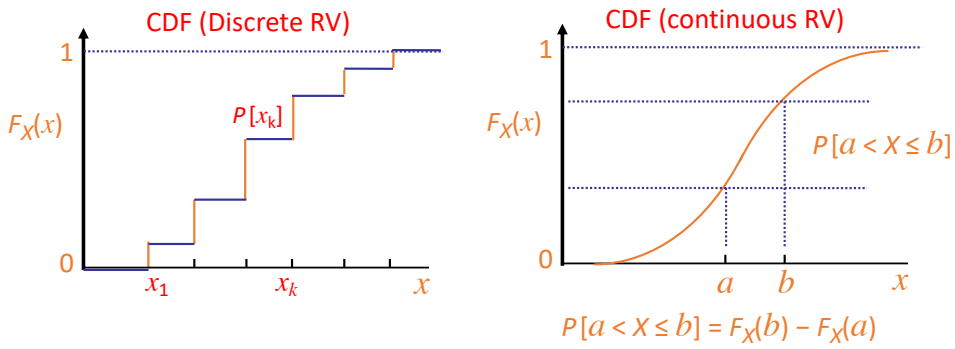


Figure 3.2: CDFs of discrete and continuous random variables.

The CDF of continuous random variables has the following additional properties:

1. **Continuity** $F_X(x)$ is a continuous function of x, i.e., $F_X(x) = \lim_{\epsilon \to 0} F_X(x + \epsilon)$.
2. $\mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}] = 0$ for all $x \in \Re$. Every atom in R_X has zero probability.
3. $\mathbb{P}[\{\omega \in \Omega : X(\omega) \leq x\}] = \mathbb{P}[\{s : X(\omega) < x\}]$.
4. For $a < b, \mathbb{P}[\{\omega \in \Omega : a < X(\omega) \leq b\}] = F_X(b) - F_X(a)$.
5. If y is any number in the range $0 < y < 1$, then there must be at least one number x such that $F_X(x) = y$. This is a consequence of the intermediate value theorem for continuous functions. Note that there could be multiple such numbers, as illustrated in Figure 3.3

Example 3.1
Suppose we want to choose a random number in the interval $(0, 1)$, with every number equally likely to be chosen. That is, $R_X = (0, 1)$. Intuitively, the meaning of random in this instance is that we do not favor any one number over others in the interval (0,1). One way of expressing the innate randomness of the choice is as follows: Given any subinterval of (0, 1), the probability that the chosen number lies in that subinterval is equal to the length of that interval. One way of capturing this is with the following CDF $F_X(x)$:

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ x & x \in (0, 1) \\ 1 & x \geq 1. \end{cases}$$

This cumulative distribution is illustrated in Figure 3.4. Note that the function is continuous; however, it is not differentiable at either $x = 0$ or $x = 1$, as the slopes from the left and right at those two points do not match.

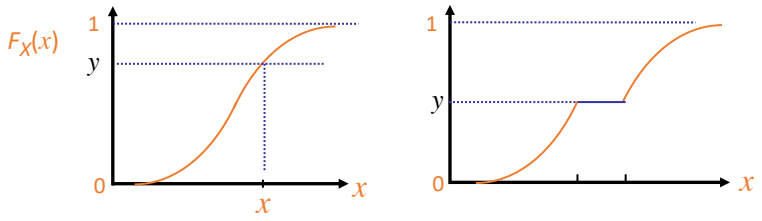


Figure 3.3: CDF where only one x satisfies $F_X(x) = y$, and where an interval of x satisfies $F_X(x) = y$.

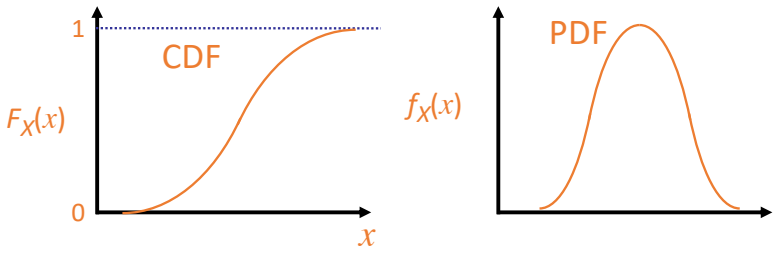


Figure 3.4: CDF and PDF for a continuous random variable.

For a random variable X to be continuous, it is not sufficient to have a continuous CDF. We want the CDF to be differentiable almost everywhere.[1] Formally, we define a continuous random variable below:

Definition 3.1
A random variable X is a **continuous random variable** if its cumulative distribution function $F_X(x)$ is continuous and differentiable almost everywhere. That is, its CDF can be written as an integral $F_X(x) = \int_{-\infty}^{x} f_X(x')d\,x'$ for some non-negative function $f_X(x')$. We refer to the function $f_X(x)$ as the **probability density function (PDF)**.

For a function to be differentiable almost everywhere, it must be differentiable everywhere except for a countable number of points $x_1, x_2, \ldots$, and there can only be a finite number of non-differentiable points in any finite-length interval. This means the CDF will have a probability density function:

$$f_X(x) = \begin{cases} \frac{d}{dx}F_X(x) & \text{if } F_X(x) \text{ is differentiable at } x, \\ \text{any non-negative number} & \text{otherwise.} \end{cases}$$

Figure 3.4 illustrates a cumulative distribution function for a continuous random variable and its corresponding probability density function (PDF). Note that this cumulative distribution function is differentiable everywhere, so the PDF is uniquely defined everywhere.

3.2.2 Probability Density Function

The PDF of a continuous random variable is not a probability and may take values greater than one, but it must be non-negative: It is a probability density. It is measured in units of probability per unit length. However, the **integral of a PDF over a region of** x is a probability, and must be a number in [0,1]. At this point, let's compare the concept of a PDF to the concept of a mass density for physical objects. Table 3.1 shows this comparison.

The probability density function for continuous random variables plays a similar role to the probability mass function for discrete random variables. The sum of the PMF of a discrete random variable over all the

[1] If you are curious, there are random variables with continuous CDF that are not differentiable almost everywhere. Look up references to Cantor distributions or the Devil's staircase function.

Physical mass in a system	probability in an experiment
Is non-negative	is non-negative
Density a function of space $\rho(x)$	probability density a function over the reals $\rho(x)$
Mass of region=	Probability of events =
Integral of density over region	Integral of density over outcomes in event

Table 3.1: Comparison of physical density and probability density

values in its domain R_X is equal to 1. Similarly, for a continuous random variable, the integral of its PDF over the entire real line is equal to 1. Although it is not a probability, if $f_X(a)$ is finite, then, for small ϵ, the probability that a sample value occurs in the interval $[a, a+\epsilon]$ is approximately $p_X(a)\epsilon$. Note that, as ϵ decreases, the probability that $X = a$ becomes zero.

The PDF satisfies the following basic properties:

1. **Non-negativity:** $f_X(x) \geq 0$.

   This follows from the monotonicity property of the cumulative distribution function, which is non-decreasing. Hence, its derivative, whenever it exists, is defined as

   $$f_X(x) = \lim_{\epsilon \to 0} \frac{F_X(x+\epsilon) - F_X(x)}{\epsilon} \ .$$

   For $\epsilon > 0$, the numerator inside the limit is always non-negative, and hence the limit, if it exists, must also be non-negative.

2. **Normalization:** $\displaystyle\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

   By definition, we know $F_X(x) = \int_{-\infty}^{x} f_X(u) d\, u$. We also know, by the normalization property of CDFs, that $\lim_{x\to\infty} F_X(x) = 1$. Thus,

   $$\lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} \int_{-\infty}^{x} f_X(u) d\ u = \int_{-\infty}^{x} \infty f_X(u) d\ u = 1 \ .$$

3. **Probability of an interval:** $\mathbb{P}_X[\{a < X \leq b\}] = \displaystyle\int_a^b f_X(x)\, dx$.

   Since $\mathbb{P}_X[\{X = x\}] = 0$ for any $x \in \Re$, we have

   $$\mathbb{P}_X[\{a < X \leq b\}] = \mathbb{P}_X[\{a \leq X \leq b\}] = \mathbb{P}_X[\{a < X < b\}]$$

   From the CDF properties, we know

   $$\mathbb{P}_X[\{a \leq X \leq b\}] = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(x) d\ x - \int_{-\infty}^{a} f_X(x) d\ x = \int_a^b f_X(x) d\ x \ .$$

4. $\lim_{x\to\infty} f_X(x) = 0; \lim_{x\to-\infty} f_X(x) = 0$.

   As the magnitude of x gets large, the PDF curve must decay to zero. Otherwise, the integral of the PDF would keep growing unbounded as $|x|$ increased. Furthermore, the slope of the pdf must also decay to zero as $|x|$ grows unbounded.

5. **PDF $\to$ CDF:** $\displaystyle\int_{-\infty}^{x} f_X(u)\, du = F_X(x)$.

   This is the definition of the PDF.

Example 3.2
Consider a continuous random variable X, with PDF specified as

$$f_X(x) = \begin{cases} 3x^2 & x \in [0,1], \\ 0 & \text{elsewhere..} \end{cases}$$

We note this satisfies the properties that we want in a PDF: It is non-negative, and it integrates to 1:

$$\int_{-\infty}^{\infty} f_X(x)d\,x = \int_0^1 3x^2 d\,x = 1\ .$$

Note that $f_X(1) = 3 > 1$ as it does not have to satisfy the bound for a probability. What this PDF indicates is that it is four times denser around $x = 1$ than around $x = 0.5$. If you generated independent samples of this random variable, the number of samples around 1 would be 4 times the number of samples around 0.5.

For discrete random variables X, the PMF provided the complete characterization of the probability properties of X. A similar property exists for continuous random variables X: The PDF provides the complete characterization its probability properties that we need for computing probabilities on the outcomes in $\Re$.

Example 3.3
A continuous random variable X has PDF

$$f_X(x) = \begin{cases} 0.75(1-x^2) & -1 \le x \le 1, \\ 0 & \text{otherwise..} \end{cases}$$

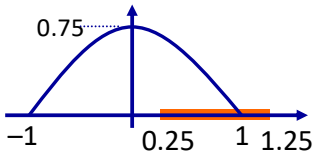


Figure 3.5: Figure for example 3.3.

This density is illustrated in Figure 3.2.2. Compute $\mathbb{P}_X[\{0.25 \le X \le 1.25\}]$. Using the basic properties of the PDF, we know $\mathbb{P}_X[\{0.25 \le X \le 1.25\}] = \int_{0.25}^{1.25} f_X(x)dx$. However, note that the region of integration involves two different pieces of the definition of f_X. Hence,

$$\mathbb{P}_X[\{0.25 \le X \le 1.25\}] = \int_{0.25}^{1.25} f_X(x)dx = \int_{0.25}^{1} 0.75(1-x^2)dx + \int_1^{1.25} 0dx = \frac{81}{256}$$

Example 3.4
A continuous random variable X has PDF

$$f_X(x) = \begin{cases} -2X & -1 \le x \le 0, \\ 0 & \text{otherwise..} \end{cases}$$

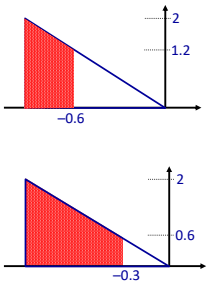


Figure 3.6: Figure for example 3.4.

This density is illustrated in Figure 3.2.2. Compute $F_X(-0.6)$ and $F_X(-0.3)$.

$$F_X(-0.6) = \int_{-\infty}^{-0.6} f_X(x)d\,x = \int_{-1}^{-0.6} (-2x)d\,x = 1 - 0.36 = 0.64\ .$$

$$F_X(-0.3) = \int_{-\infty}^{-0.3} f_X(x)d\,x = \int_{-1}^{-0.3} (-2x)d\,x = 1 - 0.09 = 0.91\ .$$

Example 3.5
Assume a continuous random variable has a PDF given by

$$f_X(x) = \begin{cases} 0.75(1-x^2) & -1 \le x \le 1, \\ 0 & \text{otherwise..} \end{cases}$$

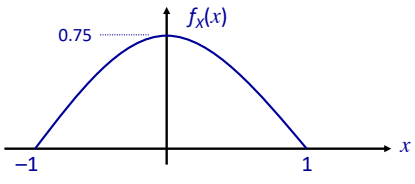


Figure 3.7: Figure for example 3.5.

This density is illustrated in Figure 3.7. Compute $F_X(0)$ and $F_X(0.5)$.

$$F_X(0) = \int_{-\infty}^{0} f_X(x)d\,x = \frac{1}{2} \text{by symmetry!}$$

$$F_X(0.5) = \int_{-\infty}^{0.5} 0.75(1-x^2)d\,x = \int_{-1}^{0.5} 0.75(1-x^2)d\,x$$
$$= 0.75(1.5) - 0.25x^3|_{-1}^{0.5} = \frac{9}{8} - \frac{1}{4}(1+\frac{1}{8}) = \frac{27}{32}\ .$$

3.3 Statistics of Continuous Random Variables

3.3.1 Expected Value

We have left the PDF undefined at points x where the CDF is not differentiable. At such points, the CDF has a different derivative when approached from the right as from the left. We are allowed to set the value of $f_X(x)$ arbitrarily to any nonnegative number at those few isolated points where the CDF is not differentiable. Note that this arbitrarily chosen value assigned to the pdf at these isolated points makes no difference whatsoever in any probability calculations, because the probability that this number occurs is zero. The probability that this number occurs is 0! In practice, we often choose either the derivative from the right or the derivative from the left as the value of the PDF at non-differentiable points of the CDF.

Example 3.6
Assume a continuous random variable has a CDF given by

$$F_X(x) = \begin{cases} 0 & x < -3, \\ \frac{1}{6}(x-3) & -3 \le x \le 3, \\ 0 & x > 3\ . \end{cases}$$

This density is illustrated in Figure 3.7. Compute $f_X(x)$, and define it for all $x \in \Re$.
Note that $F_X(x)$ is differentiable everywhere except at $x = \pm 3$. Then,

$$f_X(x) = \frac{d}{dx} F_X(x) = \begin{cases} 0 & x < -3, \\ \frac{1}{6} & -3 < x < 3, \\ 0 & x > 3\ . \end{cases}$$

This is also shown in Fig. 3.8. To complete the definition, we select $f_X(3) = 0 = f_X(-3)$, which matches the slope of one of the two line segments that meet at 3 and -3.

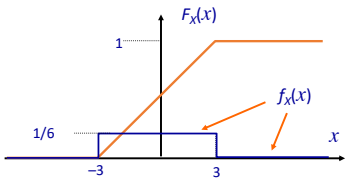


Figure 3.8: CDF and PDF for example 3.6.

As was the case for discrete random variables, we define the **expected value** of a continuous random variable X as

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

This is also known as the **mean** or **average.**. Similar to discrete random variables, this expected value can be viewed as the center of probability mass. If we repeat an experiment N times, add up all observed values of X, and divide by N to compute a sample average, the result should be pretty close to $\mathbb{E}[X]$. We sometimes use the notation $\mu_X = \mathbb{E}[X]$.

Note that, for the expectation to be defined, both of the integrals below must be finite.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_{-\infty}^{0} x\, f_X(x)\, dx + \int_{0}^{\infty} x\, f_X(x)\, dx\ .$$

This is not always the case, as shown in the next example.

Example 3.7
Let X be a continuous random variable with PDF given by: $f_X(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x \ge 0 \\ 0 & \text{otherwise.} \end{cases}$.

Note that this is a valid PDF, as it is nonnegative and properly normalized. It does decay to 0 slowly, in an inverse square law. For this random variable, its expected value does not exist:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{0}^{\infty} \frac{2x}{\pi(1+x^2)} d\ x = \ln(1+x^2)|_0^{\infty} = \infty\ .$$

This illustrates that some statistics of RVs may not be defined because the required expected values may not exist.

3.3.2 Variance

The **variance** measures how spread out a random variable is around its mean. For a continuous random variable X, it is defined using expectations in the same way as it was for discrete random variables:

$$\begin{aligned}\mathsf{Var}[X] &= \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 \\ &= \int_{-\infty}^{\infty} (x - \mu_X)^2 \, f_X(x)\, dx.\end{aligned}$$

We often refer to the variance of X as $\sigma_X^2 = \mathsf{Var}[X]$, where $\sigma_X \geq 0$ is the **standard deviation.**

3.3.3 Expected Value of a Function of a Random Variable

Let $g(\cdot)$ be a function mapping the range of a random variable X, R_X, into the real numbers $\Re$. Then, the variable $Y = g(X)$ is a random variable. We can compute the expected value of $Y = g(X)$ using the definition of the function and the PDF of X, as

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \ .$$

Note that this expression is valid no matter whether the random variable Y is discrete, continuous or of other types. It avoids the need for computing the detailed PDF or PMF of Y, by performing the averaging in terms of the PDF of the random variable X.

Example 3.8
Let X be a continuous random variable with PDF $f_X(x) = \begin{cases} 0.5 & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$.

We compute μ_X as

$$\mu_X = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) d\ x = \int_{-1}^{1} 0.5 x d\ x = 0 \ .$$

The variance σ_X^2 is given by:

$$\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \int_{-1}^{1} 0.5 x^2 d\ x = \frac{1}{3} \ .$$

Let $g(x) = |x|$, the absolute value function, and let $Y = g(X)$. Then,

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{-1}^{1} 0.5|x| d\ x = \int_{-1}^{0} 0.5(-x) d\ x + \int_{0}^{1} 0.5(x) d\ x = 0.5 \ .$$

An important class of functions are the affine functions $g(x) = ax + b$. For these classes of functions, we establish the same relations that were established in 2.4. Let $Y = g(X) = aX + b$. Then, $\mathbb{E}[Y] = \mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$. In addition, we can compute the variance as:

$$\mathsf{Var}[Y] = \mathbb{E}[(aX + b - a\mathbb{E}[x] - b)^2] = \mathbb{E}[(a(X - \mathbb{E}[X]))^2] = a^2 \mathsf{Var}[X] \ .$$

Thus, the variance of Y does not depend on the constant b, and is related to the variance of X as $\mathsf{Var}[Y] = a^2 \mathsf{Var}[X]$, as variance is a square statistic. Note that, in terms of standard deviation, $\sigma_Y = \sqrt{\mathsf{Var}[Y]} = |a|\sigma_X$, so that the standard deviation scales linearly with a.

Let $g(x) = ag_1(x) + bg_2(x)$, and let $Y = g(X)$. Then,

$$\mathbb{E}[Y] = \mathbb{E}[ag_1(X)] + \mathbb{E}[bg_2(X)] = a \int_{-\infty}^{\infty} g_1(x) d\ x + b \int_{-\infty}^{\infty} g_2(x) d\ x = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

emphasizing the fact that $\mathbb{E}[\cdot]$ is a linear operator.

3.3.4 Moments

Using the expectation operator, we define the following moments for continuous random variables, in exactly the same way they were defined for discrete random variables:

Definitions (same as for discrete random variables):

Mean of X	$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx.$
Variance of X	$\mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2\, f_X(x)\, dx.$
n^{th} moment of X	$\mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n\, f_X(x)\, dx.$
n^{th} central moment of X	$\mathbb{E}[(X - \mathbb{E}[X])^n] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^n\, f_X(x)\, dx.$

Example 3.9
Assume a continuous random variable X which has a PDF given by

$$f_X(x) = \begin{cases} \frac{3}{2}x^2 & -1 \le x \le 1, \\ 0 & \text{elsewhere.} \end{cases}$$

This density is illustrated in Figure 3.9. Compute the mean, second moment, third moment and fourth central moment.
First, note the symmetry of $f_X(x)$ about zero. This means that, for any odd function where $f(x) = -f(-x)$, we have $\mathbb{E}[f(X)] = 0$. In particular, the first and third moments are expectations of odd functions $f(x) = x$ and $f(x) = x^3$, so we have $\mathbb{E}[X] = 0, \mathbb{E}[X^3] = 0$.
The second moment is

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) d\, x = \int_{-1}^{1} x^2 \frac{3}{2} x^2 d\, x = \frac{3}{10} x^5 |_{-1}^{1} = \frac{3}{5}\ .$$

Since the mean is zero, the fourth central moment is equal to the fourth moment:

$$\mathbb{E}[X^4] = \int_{-\infty}^{\infty} x^4 f_X(x) d\, x = \int_{-1}^{1} x^4 \frac{3}{2} x^2 d\, x = \frac{3}{14} x^7 |_{-1}^{1} = \frac{3}{7}\ .$$

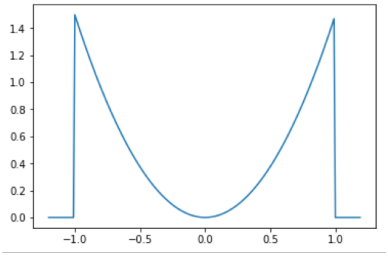


Figure 3.9: Figure for example 3.9.

3.4 Important Families of Continuous Random Variables

Although most experimental measurements are of limited precision, it is often easier to model their outcomes in terms of continuous-valued random variables because it facilitates the resulting analysis. Furthermore, the limiting form of many discrete-valued random variables result in continuous-valued random variables. Below, we describe some of the most useful continuous-valued random variables. Specifically, we overview the properties of the following families of continuous random variables:

- Uniform
- Exponential
- Gaussian (Normal)

These families of continuous RVs are used to model the outcomes of common experiments. Members of a given family differ only by the values of the few parameters of the family, which are easy to estimate

from sample data. We also discuss a few other families of continuous random variables that are used less frequently in engineering applications.

3.4.1 Uniform(a, b) Random Variables

The simplest continuous random variable is the Uniform(a, b) random variable X, where X is equally likely to achieve any value in an interval of the real line, $[a, b]$. The probability density function of X is given by:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

The corresponding cumulative distribution function is given by

$$F_X(x) = \begin{cases} 0 & x < a, \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b\ . \end{cases}$$

The PDF and CDF of uniform random variables are shown in Figure 3.10.

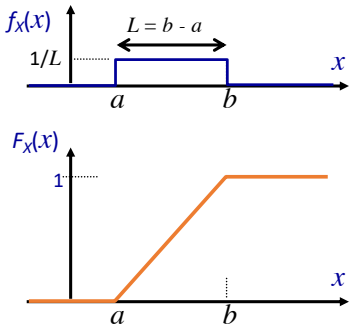


Figure 3.10: CDF and PDF for uniform RVs.

We use the notation $X \sim \text{Uniform}([a, b])$ to denote a random variable with continuous uniform distribution on the interval a, b. Using the PDF, we compute the statistics of $X \sim \text{Uniform}([a, b])$ as:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) d\ x = \int_a^b \frac{x}{b-a} d\ x = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2} \quad \text{Mean}$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) d\ x = \int_a^b \frac{x^2}{b-a} d\ x = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}$$

$$\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \frac{(b-a)^2}{12} \quad \text{Variance}$$

Example 3.10

Consider a random wave of known amplitude A is oscillating at frequency ω_0 radians per second, but with unknown phase. We model the unknown phase as a random variable Θ, uniformly distributed on the interval $[-\pi, \pi]$, so that the time history of the wave is represented as

$$x(t) = A cos(\omega_0 t + \Theta).$$

From the properties of uniform random variables, we know the average phase $\mathbb{E}[\Theta] = 0$, and the variance of the phase is $\text{Var}[\Theta] = \frac{(\pi-(-\pi))^2}{12} = \frac{\pi^2}{3}$.

The important statistics of uniform random variables are summarized below:

- PDF: $f_X(x) = \begin{cases} \dfrac{1}{b-a} & a\ lex \le b, \\ 0 & \text{otherwise.} \end{cases}$
- CDF: $F_X(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x < b \\ 1 & b \le x\ . \end{cases}$
- Expected Value: $\mathbb{E}[X] = \dfrac{a+b}{2}$.

- Variance: $\mathsf{Var}[X] = \dfrac{(b-a)^2}{12}$.

- Interpretation: Equally likely to take any value between a and b.

3.4.2 Exponential(λ) Random Variables

Exponential(λ) random variables arise in the modeling of the time between occurrence of events, such as the time between customer requests in service systems, the durations for call connections in phone systems, and the modeling of lifetimes of devices and systems. The exponential random variable X with parameter λ has a probability density function

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

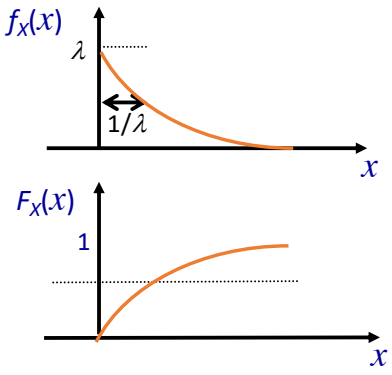


Figure 3.11: CDF and PDF for exponential RVs.

The parameter λ is denoted as the rate of the exponential random variable, and it is typically measured as units per time. An exponential random variable only takes values in the non-negative real line. The corresponding CDF is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

The PDF and CDF of exponential random variables are shown in Figure 3.11.

The exponential random variable is similar to the discrete geometric random variable, in that it is the limit of the geometric random variable, as the difference between values of a geometric random variable gets small. For example, assume that an interval of length T seconds was subdivided into subintervals of length T/n, and assume that, for each subinterval, there is a Bernoulli trial with probability of success $p = \frac{\lambda T}{n}$, where λ is the average number of events per second, so λT is the average number of events per T seconds. Then, the number of subintervals until the occurrence of the next event is a geometric random variable M. Let X denote the time until the next successful event. Then, for any t which is a multiple of T/n,

$$\mathbb{P}[\{X > t\})] = \mathbb{P}[\{M > \frac{nt}{T}\}] = (1-p)^{nt/T} = \left((1 - \frac{\lambda T}{n})^n\right)^{t/T}$$

In the limit, we get

$$\lim_{n\to\infty} \mathbb{P}[\{X > t\}] = e^{-\lambda t}$$

which is $1 - F_X(t)$ for an exponential random variable X with rate λ.

We use the notation $X \sim$ exponential(λ) to denote a random variable X with exponential distribution, parameter λ. The important expectations of an exponential random variable $X \sim$ exponential(λ) are computed as:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) d\, x = \int_0^{\infty} x\lambda e^{-\lambda x} d\, x \text{ (integrate by parts)} \\
&= -xe^{-\lambda x}|_0^{\infty} + \int_0^{\infty} \lambda e^{-\lambda x} d\, x \\
&= 0 - \frac{1}{\lambda} e^{-\lambda x}|_0^{\infty} = \frac{1}{\lambda} \\
\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x f_X(x) d\, x = \int_0^{\infty} x^2 \lambda e^{-\lambda x} d\, x = -\int_0^{\infty} x^2 d\, e^{-\lambda x} \text{ (integrate by parts twice)} \\
&= x^2 e^{-\lambda x}|_0^{\infty} - 2\int_0^{\infty} x e^{-\lambda x} d\, x = \frac{2}{\lambda}\int_0^{\infty} x d\, e^{-\lambda x} \\
&= \frac{2}{\lambda} x e^{-\lambda x}|_0^{\infty} - \frac{2}{\lambda}\int_0^{\infty} e^{-\lambda x} d\, x = \frac{2}{\lambda^2} \\
\mathsf{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}
\end{aligned}$$

Example 3.11
The duration of a service repair request for a broken appliance is modeled as an exponential random variable X with parameter $\lambda = 0.1$ repairs/minute. The repair person charges a fixed rate of \$5.00 for the first five minutes, then \$0.50 for each additional minute. Compute the expected time to repair the appliance, the variance of the repair time, and the expected cost of the repair.

Since X is an exponential random variable, the expected repair time and variance are computed as:

$$\mathbb{E}[X] = \frac{1}{\lambda} = 10\text{minutes. } \mathsf{Var}[X] = \frac{1}{\lambda^2} = 100\text{minutes}^2 \ .$$

The cost can be viewed as a function $g(X)$ defined by

$$g(x) = \begin{cases} 0 & x < 0, \\ 5 & 0 \le x \le 5, \\ 5 + 0.5(x-5) & x \ge 5 \ . \end{cases}$$

Then,

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int_0^{\infty} g(x) f_X(x) dx = \int_0^{\infty} 5 f_X(x) d\, x + \int_5^{\infty} 0.5(x-5) f_X(x) d\, x \\
&= 5 + \int_5^{\infty} 0.5(x-5) 0.1 e^{-0.1x} d\, x = 5\int_0^{\infty} 0.5(y) 0.1 e^{-0.1(y+5)} d\, y \text{ substitute } y = x-5 \\
&= 5 + 0.5e^{-0.5}\int_0^{\infty} 0.1 y e^{-0.1y} d\, y = 5 + 0.5e^{-0.5}\mathbb{E}[X] = 5 + 5e^{-0.5} \approx \$8.03 \ .
\end{aligned}$$

Note that the expected cost $\mathbb{E}[g(X)]$ is not equal to $g(\mathbb{E}[X]) = \$7.50$. This is because $g(\cdot)$ is not an affine function.

The properties of exponential(λ) random variables are summarized below:

- PDF: $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0, \\ 0 & x < 0. \end{cases}$
- CDF: $F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \ . \end{cases}$
- Expected Value: $\mathbb{E}[X] = \frac{1}{\lambda}$.
- Variance: $\mathsf{Var}[X] = \frac{1}{\lambda^2}$.

- Interpretation: Continuous waiting time. "Continuous version" of geometric random variables.
- Applications: Packet interarrival times, call durations, hard drive lifetimes.

3.4.3 Gaussian(μ, σ^2) Random Variables

Gaussian(μ, σ^2) random variables model many situations where the random event consists of the sum of a large number of small random variables. They are named after Karl Friedrich Gauss, who used this class of random variables to model errors in measurements for the least squares estimation of orbital parameters from telescope observations. Gaussian random variables are determined by two parameters: their mean μ and their variance σ^2.
The probability density function of a Gaussian random variable is given by

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

. Its corresponding CDF is given by

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} d\ u = \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} d\ y\ .$$

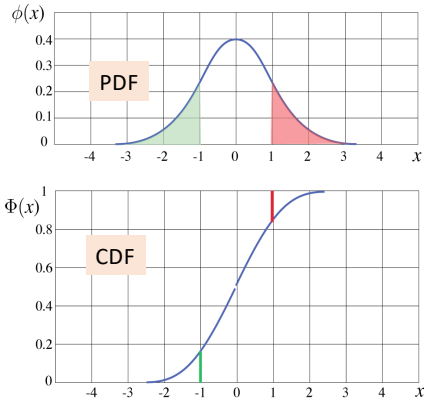


Figure 3.12: PDF and CDF for Gaussian RVs.

where the last equality follows by substituting $y = \frac{u-\mu}{\sigma}$. Note that the last integral corresponds to a Gaussian CDF win mean zero and variance 1. The PDF and CDF of Gaussian random variables are shown in Figure 3.12.

We refer to a Gaussian(0,1) random variable as a standard Gaussian random variable. Note that the CDF of any Gaussian(μ, σ^2) random variable can be computed in terms of the CDF of a standard Gaussian random variable. We formally define the CDF of a standard Gaussian as the function

$$\Phi(x) \equiv \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} d\ yn\ ,$$

and the standard normal complementary CDF as $Q(x) \equiv 1 - \Phi(x)$. Unfortunately, $\Phi(x)$ cannot be computed in closed form, but its values are tabulated in Appendix C.

Gaussian random variables are also known as Normal random variables because many sets of data gathered from a variety of physical phenomena seem to fit the Gaussian (or normal) distribution. In these sets of data, errors arise as the combination of many small effects; to develop the exact distribution of the sum of many random variables is unwieldy. Fortunately, the central limit theorem, which we study in Chapter 8, asserts that if many "small" random causes produce a net effect, then that effect can be approximately modeled as a normal or Gaussian random variable.

We often write $X \sim \mathcal{N}(\mu, \sigma^2)$, or use the phrase "$X$ is $\mathcal{N}(\mu, \sigma^2)$", to denote that X is a Gaussian(μ, σ^2) random variable with mean μ and variance σ^2. The statistics of a Gaussian random variable X are specified in its parameters:

$$\mathbb{E}[X] = \mu \qquad \mathsf{Var}[X] = \sigma^2\ .$$

We note that this notation varies across texts. Some texts will refer to a Gassian random variable as $\mathcal{N}(\mu, \sigma)$, using the standard deviation instead of the variance. We chose our notation because it generalizes to vectors in a natural way.

Normal distributions are used in many situations. In many classes, professors believe that the distribution of grades must be normally distributed with a given mean and variance. Thus, you see the phenomena that

exams are graded "on the curve," where the actual grades are mapped nonlinearly into the Normal bell-shaped PDF, and letter grades are assigned based on the percentile of the grade using the standard Normal CDF $\Phi(x)$. Similarly, SAT and GRE actual scores are nonlinearly mapped so that the final scores correspond to a $\mathcal{N}(500, 10000)$ distribution.

Gaussian random variables have an interesting property: an affine transformation of a Gaussian random variable is also a Gaussian random variable. That is, if X $\mathcal{N}(\mu, \sigma^2)$ is Gaussian, then $Y = aX + b$ is also Gaussian for any real scalars a, b. We will show this later in this chapter. Furthermore, we know $\mathbb{E}[Y] = a\mathbb{E}[X] + b, \mathsf{Var}[Y] = a^2\mathsf{Var}[X]$, so

$$X\ N(\mu, \sigma^2) \rightarrow Y = aX + b\ \mathcal{N}(a\mu + b, a^2\sigma^2)$$

This important property is another reason why Gaussian variables are often used in engineering models.

The Gaussian PDF is symmetric about its mean. This implies that all odd central moments are zero. Using integration by parts, we can compute all even central moments as a multiple of the variance σ^2, as

$$\mathbb{E}[(X - \mathbb{E}[X])^{2n}] = (2n-1)(2n-3)\cdots(1)\sigma^2$$

. Thus, $\mathbb{E}[(X - \mathbb{E}[X])^4] = 3\sigma^2, \mathbb{E}[(X - \mathbb{E}[X])^6] = 15\sigma^2$.

To perform computations about probabilities of Gaussians, we use the standard normal CDF function $\Phi(\cdot)$. Appendix C includes the detailed tabulated standard normal CDF. We note the following properties which are useful for computation:

$$\Phi(-x) = 1 - \Phi(x) \qquad .\Phi(x) - \Phi(-x) = 2\Phi(x) - 1$$

$$Q(x) = \Phi(-x) = 1 - \Phi(x)$$

The way we use the standard tables for computation is for computing probabilities for a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. Recall that

$$F_X(x) = \Phi(\frac{x-\mu}{\sigma})\ .$$

Note that the argument of the standard Gaussian function $\Phi(\cdot)$ is expressed as the difference between the value x and the mean of the random variable, expressed in units of standard deviations. That is, the statistic $z_x = \frac{x-\mu}{\sigma}$ used as the argument for Φ is the number of standard deviations away from the average. We illustrate this with the following example:

Example 3.12
Consider a Gaussian random variable $X \sim \mathcal{N}(1, 4)$. Determine the probability that X lies between -1 and 3.

From its definition, $\mathbb{P}[\{-1 < X \leq 3\}] = F_X(3) - F_X(1)$. The standard deviation of X is $\sigma_X = \sqrt{4} = 2$. Then,

$$\begin{aligned}
\mathbb{P}[\{-1 < X \leq 3\}] &= F_X(3) - F_X(-1) \\
z_3 = \frac{3-1}{2} = 1;\ z_{-1} &= \frac{-1-1}{2} = -1; \\
\mathbb{P}[\{-1 < X \leq 3\}] &= F_X(3) - F_X(-1) = \Phi(z_3) - \Phi(z_{-1}) = \Phi(1) - \Phi(-1) \\
&= \Phi(1) - (1 - \Phi(1)) = 2\Phi(1) - 1 = 2(0.8413)\text{–}1 = 0.6826
\end{aligned}$$

where the number for $\Phi(1)$ was obtained from the table in Appendix C.

Example 3.13
An underwater microphone is measuring the average acoustic pressure X to detect whether there is a submarine generating sounds in its neighborhood. If no submarine is present, the background acoustic pressure is modeled as a Gaussian, with $X \sim \mathcal{N}(2, 4)$. If the submarine is present, the measured acoustic pressure is modeled as $X \sim \mathcal{N}(3, 4)$.

The microphone uses a simple threshold $T \in (2, 3)$, and if the measured $X > T$, it declares that a submarine is present. A false alarm happens when there is no submarine present (so $X \sim \mathcal{N}(2, 4)$), yet $X > T$. What is the probability of a false alarm? Express the answer in terms of T and the standard complementary CDF $Q(\cdot)$.

When no submarine is present, $X \sim \mathcal{N}(2,4)$, so $\mu = \sigma = 2$. $P_F = \mathbb{P}[\{X > T\}] = 1 - F_X(T)$. Computing the z-statistic, $z_T = \frac{T-2}{2}$. Thus,

$$P_F = 1 - \Phi(z_T) = Q(z_T) = Q(\frac{T-2}{2}).$$

If the submarine is present, but $X < T$, the microphone will not declare a detection, and thus the detection will be missed. Express the probability of missed detection in terms of T using the complementary CDF $Q(\cdot)$.

When the submarine is present, $X \sim \mathcal{N}(3,4)$, so $\mu = 3, \sigma = 2$. Then, $P_{MD} = \mathbb{P}[\{X < T\} = F_X(T)$. The z-statistic is $z_T = \frac{T-3}{2}$, so

$$P_{MD} = F_X(T) = \Phi(z_T) = \Phi(\frac{T-3}{2}) = Q(\frac{3-T}{2})$$

A summary of the properties of a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is:

- PDF: $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- CDF: $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}\, d\, w$.
- $\Phi(z)$ is the standard normal CDF. $Q(z) = 1 - \Phi(z)$ is the standard normal complementary CDF.
- Expected Value: $\mathbb{E}[X] = \mu$.
- Variance: $\mathsf{Var}[X] = \sigma^2$.
- Interpretation: Sum (or average) of many small random effects.
- Applications: Noise modeling, linear systems, high-dimensional data.

3.4.4 Other families of continuous random variables

Below we quickly overview other classes of continuous random variables that are used less frequently in engineering. This section is primarily for reference, and won't be used much in the rest of this course.

Gamma and Erlang random variables Gamma random variable appear in may applications. For example, it is often used to model the time to service customers in queuing systems, the lifetime of devices in reliability studies, and the defect clustering behavior in VLSI chips. The probability density function of a gamma random variable X has two parameters $\rho > 0, \lambda > 0$, and is given by

$$f_X = \frac{\alpha(\alpha x)^{\rho-1} e^{-\lambda x}}{\Gamma(\rho)}$$

where $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} d\, x$. Note that, for z a positive integer, $\Gamma(z) = (z-1)!$. Other notable values are $\Gamma(0.5) = \sqrt{\pi}$.

The versatility of the gamma distribution is that, by properly choosing the two parameters, it can take a variety of shapes, which can be used to fit specific distributions. For instance, when $\rho = 1$, we obtain the exponential random variable. By letting $\rho = m$, where m is a positive integer, we obtain the m-stage Erlang distribution, which is the distribution of the sum of m independent and identical exponential random variables, each with rate λ.

The CDF of general Gamma distributions can only be expressed in terms of special functions and is seldom used for computations. The important expectations of a gamma random variable X with parameters ρ, λ are given by:

$$\mathbb{E}[X] = \frac{\rho}{\lambda}$$
$$\mathsf{Var}[X] = \frac{\rho}{\lambda^2}$$

Rayleigh random variables Rayleigh random variables are often used to model the random magnitude of a vector. As such the variables must be positive. The PDF of a Rayleigh random variable X with parameter α is given by: $f_X(x) = \frac{x}{\alpha^2} e^{-x^2/2\alpha^2}$.

The CDF of a Rayleigh random variable X with parameter α is $F_X(x) = 1 - e^{-\frac{x^2}{2\alpha^2}}$. Some of its important statistics are

$$\mathbb{E}[X] = \alpha\sqrt{\frac{\pi}{2}}$$
$$\mathsf{Var}[X] = \frac{4-\pi}{2}\alpha^2$$

Laplacian random variable: The Laplacian random variable models a two-sided exponential distribution with parameter λ. The probability density function of a Laplacian random variable X is given by

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda|x|} \ .$$

Its CDF is given by

$$F_X(x) = \begin{cases} \frac{1}{2}e^{\lambda x} & x < 0, \\ 1 - \frac{1}{2}e^{-\lambda x} & x \geq 0 \ . \end{cases}$$

with expectations:

$$\mathbb{E}[X] = 0$$
$$\mathsf{Var}[X] = \frac{2}{\lambda^2}$$

Cauchy random variable: The Cauchy random variable is often used as an example to illustrate distributions which do not decay fast enough as $x \to \infty$, so that no moments exist. We call those *heavy-tailed* distributions. The probability density function of a Cauchy random variable with parameter β is given by $f_X = \frac{\beta/\pi}{\beta^2 + x^2}$.

The CDF of a Cauchy random variable X with parameter β is $F_X(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(\frac{x}{\beta})$.

Due to its symmetry, the mean is often taken to be zero, though the formal expected value of the density does not have a unique value. It is easy to verify that the variance of this distribution does not exist either.

In Table 3.2 we summarize the characteristics of important random variables, where the more general (shifted) forms of the Laplacian and Cauchy distributions are given.

Example 3.14
Consider the following quick questions regarding continuous random variables:

Discrete-Valued X						
Name	Range R_X	Parameters	PMF $P_X(x)$	CDF $F_X(x)$	Mean	Variance z)
Bernoulli	$\{0,1\}$	$0 \le p \le 1$	$P(x) = \begin{cases} 1-p & x=0 \\ p & x=1 \end{cases}$	$\begin{cases} 0 & x<0, \\ (1-p) & x \in [0,1) \\ 1 & x>1. \end{cases}$	p	$p(1-p)$
Uniform	$\{k, k+1, \ldots, k+n\}$	$n>0, k$ integer	$\frac{1}{n+1}$	$\frac{\lfloor x \rfloor - k + +1}{n+1}$	$k+\frac{n}{2}$	$\frac{(n+1)^2-1}{12}$
Binomial	$\{0,\ldots,n\}$	$0 \le p \le 1$	$\binom{n}{x} p^x (1-p)^{(n-x)}$	$\sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{(n-k)}$	np	$np(1-p)$
Geometric	$\{1,\ldots\}$	$0<p<1$	$(1-p)^{x-1}p$	$(1-(1-p)^{\lfloor x \rfloor})$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$\{0,1,\ldots\}$	$0<\lambda$	$\frac{\lambda^x e^{-\lambda}}{x!}$	$\sum_{k=0}^{\lfloor x \rfloor} \frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ

Continuous-Valued X						
Name	Range	Parameters	PDF $f_X(x)$	CMF $F_X(x)$	Mean	Variance
Uniform	$[a,b]$	$a<b$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gaussian	$[-\infty,\infty]$	μ, σ^2	$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$	$1-Q((x-\mu)/\sigma)$	μ	σ^2
Exponential	$[0,\infty]$	$\lambda>0$	$\lambda e^{-\lambda x}$	$1-e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Erlang	$[0,\infty]$	$\lambda>0$, $n>0$	$\frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$	$1-e^{-\lambda x}\sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!}$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
Gamma	$[0,\infty]$	$\lambda, \rho>0$	$\frac{\lambda(\lambda x)^{(\rho-1)} e^{-\lambda x}}{\Gamma(\rho)}$	N/A	$\frac{\rho}{\lambda}$	$\frac{\rho}{\lambda^2}$
Rayleigh	$[0,\infty]$	α^2	$\frac{x}{\alpha^2} e^{-x^2/2\alpha^2}$	$1-e^{-x^2/2\alpha^2}$	$\alpha\sqrt{\frac{\pi}{2}}$	$(2-\frac{\pi}{2})\alpha^2$
Laplacian	$[-\infty,\infty]$	$\lambda>0, \mu$	$\frac{\lambda}{2} e^{-\lambda\lvert x-\mu \rvert}$	$\begin{cases} \frac{1}{2} e^{\lambda(x-\mu)} & x<\mu \\ 1-\frac{1}{2} e^{-\lambda(x-\mu)} & x \ge \mu \end{cases}$	μ	$\frac{2}{\lambda^2}$
Cauchy	$[-\infty,\infty]$	$\beta>0, \alpha$	$\frac{\beta/\pi}{\beta^2+(x-\alpha)^2}$	$\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\left(\frac{x-\alpha}{\beta}\right)$	Undef	Undef

Table 3.2: Important random variables. (N/A under the PDF column indicates that there is no simplified form.)

1. if $X \sim \text{Uniform}([0,1])$, and $Y = -2X + 1$, compute $\mathbb{E}[Y]$ and $\text{Var}[Y]$.
   Answer: $\mathbb{E}[X] = \frac{0+1}{2} = 0.5; \mathbb{E}[Y] = -2\mathbb{E}[X] + 1 = 0; \text{Var}[X] = \frac{1}{12}; \text{Var}[Y] = (-2)^2\text{Var}[X] = \frac{1}{3}$.
2. If $X \sim \text{Uniform}([a,b])$, and $\mathbb{E}[X] = 2, \text{Var}[X] = 4$, what are a, b?
   Answer: $\mathbb{E}[X] = 2 = \frac{a+b}{2}; \text{Var}[X] = 4 = \frac{(b-a)^2}{12}$ so $b - a = 4\sqrt{3}$. Thus, $b = 2 + 2\sqrt{3}, a = 2 - 2\sqrt{3}$.
3. If $X \sim \mathcal{N}(0, 0.5)$ and $Y = -2X + 1$, what is the probability that $Y > 5$, in terms of the standard Gaussian CDF $\Phi(x)$?
   Answer: $\mathbb{E}[Y] = 1, \text{Var}[Y] = (-2)^2(0.5) = 2$. Thus, $\mathbb{P}[\{Y > 5\}] = 1 - F_Y(5)$. Since Y has mean 1, standard deviation $\sqrt{2}$, the z statistic $z_5 = \frac{5-1}{\sqrt{2}} = 2\sqrt{2}$. Hence, $\mathbb{P}[\{Y > 5\}] = 1 - \Phi(2\sqrt{2}) = Q(2\sqrt{2})$.
4. If X is Gaussian with $\mathbb{E}[X] = 1, \mathbb{E}[X^2] = 5$, what is the variance of X?
   Answer: $\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 4$.

3.5 Conditional Probability for Continuous Random Variables

Consider a probability space $(\Omega, \mathcal{E}, \mathbb{P})$. For any events $A, B \in \mathcal{E}$ such that $\mathbb{P}[B] > 0$, we define the conditional probability of A given B as:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B - A] + \mathbb{P}[B \cap A]} \ .$$

Let X be a random variable defined on $(\Omega, \mathcal{E}, \mathbb{P})$. Then, $\{\omega \in \Omega : X(\omega) \leq a\} \equiv \{X \leq a\}$ defines an event in $\mathcal{E}$. Using this event, we can define the conditional cumulative distribution function $F_{X|B}(a)$ as follows:

$$F_{X|B}(a) = \frac{\mathbb{P}[\{\omega \in \Omega : X(\omega) \leq a\} \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[\{X \leq a\} \cap B]}{\mathbb{P}[B]} \ .$$

Note that this definition is valid for all random variables, not just discrete or continuous ones. For discrete random variables, we defined the conditional probability mass function in **??**, as $P_{X|B}(a)$ by exploiting the fact that R_X was discrete:

$$P_{X|B}(a) = \frac{\mathbb{P}[\{X = a\} \cap B]}{\mathbb{P}[B]} \ .$$

For the special case that $B \subset X$, so that the event is $\{\omega \in \Omega : X(\omega) \in B\}$, this simplified to

$$P_{X|B}(a) = \begin{cases} \frac{P(a)}{\mathbb{P}_X[B]} & a \in B_1, \\ 0 & \text{otherwise.} \end{cases}$$

We referred to this operation as restrict/rescale: restrict the probability mass functions to $a \in B$, and rescale so that $\sum_{x \in B} P_{X|B}(x) = 1$.

Suppose X is a continuous random variable, so it has a probability density function $f_X(x)$ defined almost everywhere. We can compute the conditional CDF $F_{X|B}(a)$ as indicated above. Then, we define the conditional probability density function $f_{X|B}(a)$ as the derivative of the conditional CDF:

$$f_{X|B}(A) = \frac{d}{da}F_{X|B}(a) = \frac{\frac{d}{da}\mathbb{P}[\{X \leq a\} \cap B]}{\mathbb{P}[B]} = \frac{\lim_{\epsilon \to 0} \frac{\mathbb{P}[\{X \leq a+\epsilon\} \cap B] - \mathbb{P}[\{X \leq a\} \cap B]}{\epsilon}}{\mathbb{P}[B]}$$

It should be clear that, if X has a CDF that is differentiable almost everywhere, the conditional CDF will also be differentiable almost everywhere, so the conditional PDF will exist as defined above.

We can simplify this when the conditioning event is $\{X \in B \subset \Re\}$. In this case, $\mathbb{P}_X[B] = \int_{x \in B} f_X(x) d\, x = \mathbb{P}[\{X \in B\}]$. For this case, we have the following:

$$\mathbb{P}[\{X \leq a + \epsilon\} \cap \{X \in B\}] - \mathbb{P}[\{X \leq a\} \cap \{X \in B\}] = \int_{x \in (-\infty, a+\epsilon] \cap B} f_X(x) d\, x - \int_{x \in (-\infty, a] \cap B} f_X(x) d\, x \ .$$

Thus, taking limits and using the fundamental theorem of calculus,

$$\lim_{\epsilon\to 0}\frac{\mathbb{P}[\{X\le a+\epsilon\}\cap\{X\in B\}]-\mathbb{P}[\{X\le a\}\cap\{X\in B\}]}{\epsilon}=\begin{cases}0 & \text{if } a\notin B,\\ \lim_{\epsilon\to 0}\frac{\mathbb{P}[\{X\le a+\epsilon\}]-\mathbb{P}[\{X\le a\}]}{\epsilon} & \text{if } a\in B.\end{cases}$$

Thus, when the conditioning event is $\{X\in B\}$, we have $f_{X|B}(x)=\begin{cases}\frac{f_X(x)}{\mathbb{P}_X[B]} & x\in B\\ 0 & \text{otherwise.}\end{cases}$

Example 3.15
Let $X\sim$ Exponential(2) be an exponential random variable with rate 2. Consider the event B generated by $\{X>1\}$. Then, $\mathbb{P}_X[B]=1-F_X(1)=e^{-2}$, and the conditional PDF of X given B is

$$f_{X|B}(x)=\begin{cases}\frac{2e^{-2x}}{e^{-2}}=2e^{-2(x-1)} & x\ge 1,\\ 0 & x<1\ .\end{cases}$$

Note that the conditional $f_{X|B}(x)$ is just the original $f_X(x)$ shifted to start at $x=1$! This is the memoryless property for exponential random variables that we showed earlier for geometric random variables. If we define the time to go as $Y=X-1$, then $f_{Y|B}(a)=f_X(a)$. Thus, if you have waited for one hour for an arrival, the time you have left to wait has the same distribution as the original arrival time.

With the conditional PDF, we can define conditional statistics: The conditional expected value of X is

$$\mathbb{E}[X|B]=\int_{-\infty}^{\infty}xf_{X|B}(x)d\ x\ .$$

The conditional expected value of a function $g(X)$ is

$$\mathbb{E}[g(X)|B]=\int_{-\infty}^{\infty}g(x)f_X(x)d\ x\ .$$

With these equations, we can now compute the conditional variance of X given observation of event B as $\mathsf{Var}[X|B]=\mathbb{E}[X^2|B]-(\mathbb{E}[X|B])^2$.

3.6 Functions of a Continuous Random Variable

Assume we have a random variable X defined on a probability space $(\Omega,\mathcal{E},\mathbb{P})$. Any measurable function $g:\Re\to\Re$ can be used to define a derived random variable $Y=g(X)$ on the same probability space. However, even if X is a continuous random variable, it is unclear as to whether the resulting random variable Y will be continuous, or discrete, or perhaps a mixed random variable, a type that we have not discussed yet. Even if X is a continuous random variable and $g(\cdot)$ is a continuous function, the resulting random variable Y is not guaranteed to be continuous.

We have described previously how to compute statistics of Y, such as $\mathbb{E}[Y]$ or $\mathbb{E}[Y^2]$, without having to compute the resulting PMF or PDF of Y; e.g. $\mathbb{E}[Y]=\int_{-\infty}^{\infty}g(x)f_X(x)d\ x$. What we are after in this section is computing the full PMF or PDF of Y, whenever it is appropriate to do so.

Example 3.16
Let $X\sim$ Uniform$([-1,1])$. Define $g(x)=\begin{cases}0 & x<0,\\ x & x\ge 0\ .\end{cases}$. Note that $g(x)$ is continuous, and X is a continuous random variable, but $Y=g(X)$ has the property that $\mathbb{P}[\{Y=0\}]=0.5$, so that there is mass at the point $Y=0$. Thus, Y is not a continuous random variable, because the CDF of Y is not continuous.

3.6.1 Transforming Continuous to Discrete

One case in which we can handle the transformation $Y = g(X)$ is whenever the function $g(\cdot)$ is piecewise constant. In this case, $Y = g(X)$ will be a discrete random variable, with range R_Y written as the list of discrete values that $g(x)$ can take. In this case, we can determine the PMF of Y as follows:

For each value $y \in R_Y$, determine the set of values of x such that $g(x) = y$. Formally, find $A_y = \{x : g(x) = y\}$ for each $y \in R_Y$. Then, compute $P_Y(y) = \int_{x \in A_y} f_X(x) d\ x$. The resulting random variable is discrete, and its statistics can be obtained for the PMF function computed above.

3.6.2 Transforming Continuous to Continuous

If X is a continuous random variable, the function $g(x)$ is continuous, differentiable almost everywhere, and its derivative $g'(x) = \frac{d}{dx} g(x)$ is not zero on any interval (but can be zero at specific values), then $Y = g(X)$ is a continuous random variable. Under these conditions, the set of values x such that $g(x) = y$ is discrete, and has probability zero. Thus, probability mass does not accumulate at any value of y.

For these cases, the PDF of Y can be determined from the PDF of X and knowledge of the function $g(\cdot)$ and its derivative. Given the range R_X, compute the range $R_Y = g(R_X)$. For each $y \in R_Y$, determine the set of all values of $x \in R_X$ such that $g(x) \leq y$. Formally, find $B_y = \{x \in R_X : g(x) \leq y\}$ for each $y \in R_Y$. Then, the CDF of Y is determined as

$$F_Y(y) = \mathbb{P}[\{Y \leq y\} = \int_{B_y} f_X(x) d\ x\ .$$

Once the CDF of Y is found, the PDF is obtained as the derivative $f_Y(y) = \frac{d}{dy} F_Y(y)$.

There is a special case of functions $g(\cdot)$ for which the computation of the PDF of Y is simpler, and can be done avoiding the need to compute the CDF of Y first. That is the case where $g(\cdot)$ is **strictly monotonic**: either strictly increasing ($g(x) > g(y)$ if $x > y$) or strictly decreasing ($g(x) < g(y)$ if $x > y$). In this case, the function $g(x)$ has an inverse function $h(y) = g^{-1}(y)$, and the PDF of $Y = g(X)$ is $f_Y(y) = f_X\big(h(y)\big)\left|\frac{d}{dy} h(y)\right|$.

A special case of monotone functions $g(\cdot)$ are affine functions $g(x) = ax + b$ where the slope a is non-zero. In this case, $h(y) = \frac{1}{a}(y - b)$, $\frac{d}{dy} h(y) = \frac{1}{a}$ and

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right)\ .$$

We illustrate this with several important examples:

Example 3.17
Let X be a Gaussian random variable such that $X \sim \mathcal{N}(\mu, \sigma^2)$. Let $g(x) = ax + b$, with $a \neq 0$. Then, $Y = g(X)$ is an affine transformation of a Gaussian random variable. By the above formula,

$$f_Y(y) = \frac{1}{|a|} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\frac{y-b}{a} - \mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi a^2 \sigma^2}} e^{-\frac{(y-b-\mu a)^2}{2a^2\sigma^2}} \sim \mathcal{N}(a\mu + b, a^2\sigma^2)\ .$$

This proves the important property that affine transformations of Gaussian random variables are Gaussian random variables.

Example 3.18
Let X be a uniform random variable, with $X \sim$ Uniform$((0, 1])$. Let $g(x) = -\frac{1}{\lambda} \ln(x)$ which is a monotone, strictly decreasing function with inverse $h(y) = e^{-\lambda y}$. Let $Y = g(X)$; then, $R_Y = (0, \infty)$; then,

$$f_Y(y) = f_X\big(h(y)\big)\left|\frac{d}{dy} h(y)\right| = |\frac{d}{dy} h(y)| = \lambda e^{-\lambda y}, y > 0\ .$$

This shows $Y \sim$ Exponential(λ), an exponential random variable.

Example 3.19
Consider a function $q(y)$ that is continuous, monotone non-decreasing, differentiable almost everywhere with values in $[0,1]$ such that $\lim_{y\to-\infty} q(y) = 0, \lim_{y\to\infty} q(y) = 1$. Assume that $q(y)$ is strictly monotone increasing over its range $R_Y = \{y \in \Re : 0 < q(y) < 1\}$. Let X be a uniform random variable on $[0,1]$. We want to find a transformation $Y = g(X)$ such that the derived random variable Y has CDF $F_Y(y) = q(y)$.

Let $r(y) = q^{-1}(y)$ be the inverse of q, so that $r : (0,1) \to R_X$, and define $Y = r(X)$. Then,

$$F_Y(y) = \mathbb{P}[\{Y \le y\}] = \mathbb{P}[\{X \le q(y)\}] = F_X(q(y)) = q(y)$$

Hence, we can transform uniform random variables on [0,1] to random variables Y with CDF $q(Y)$ as long as $q(Y)$ is strictly increasing over its effective range.

Example 3.20
Let $X \sim \mathcal{N}(0,1)$, and let $Y = X^2$. Note that $g(\cdot)$ is continuously differentiable, but not monotone. For any value $y \in [0,\infty)$, let $B_y = \{x \in \Re : x^2 \le y\} = .\{x \in \Re : -\sqrt{y}x \le \sqrt{y}\}$. Then,

$$F_Y(y) = \mathbb{P}_X[B_y] = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) \ .$$

Hence, its PDF for $y > 0$ is

$$f_Y(y) = \frac{d}{dy}\left(\Phi(\sqrt{y}) - \Phi(-\sqrt{y})\right) = \frac{1}{2\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}} + \frac{1}{2\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}} = \frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}} \ .$$

3.7 Mixed Random Variables

There are many random variables that are neither continuous nor discrete. For instance, consider the random variable X with CDF given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 0.5 + 0.5x & 0 \le x < 1 \\ 1 & x \ge 1 \ . \end{cases}$$

This CDF is not continuous, as it has a jump at $x = 0$. However, the range of X is $R_X = [0,1]$, an uncountable space. Random variables with a CDF that has a discrete set of discontinuities, but is almost surely differentiable elsewhere are mixtures of discrete and continuous random variables. We refer to such random variables as **mixed random variables.**

The difficulty with mixed random variables X is that we cannot compute either a probability mass function or a probability density function from the CDF $F_X(x)$. Hence, we don't have the basic information needed for computing statistics, or expectations of functions of X.

We will overcome this difficulty by defining a generalized version of a CDF using generalized derivatives of the CDF. Specifically, at points where the CDF has discontinuities, we represent the derivative using an impuse $\delta(\cdot)$ function. In engineering, the impulse function is defined by the following properties:

$$\delta(a) = 0 \text{if } a \ne 0$$

$$\int_b^c \delta(a)\, d\ a = \begin{cases} 0 & \text{if } b \le c < 0 \\ 1 & \text{if } b \le 0 \le c \\ 0 & \text{if } 0 < b \le c \ . \end{cases}$$

$$\int_{-\infty}^{\infty} \delta(a-s)g(\omega)\, d\ s = g(a) \quad \text{if } g \text{ is continuous at } a.$$

Using this concept, we define the PDF of a mixed random variable X as: $f_X(x) = \frac{d}{dx}F_X(x)$ where we use impulse functions to represent derivatives at points where the CDF is discontinuous. Note we can use this do define a PDF for discrete random variables also. For example,

$$f_X(x) = 0.5\delta(x+1) + 0.5\delta(x-1)$$

is the density of a random variable taking on the values $\{-1, 1\}$ each with equal probability.

The most important property of impulse functions is that we can integrate them. We use PDFs to compute probabilities of events by integrating the PDF over the range of values in the event. Thus, for the above variable X, we can compute the second moment directly, as

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) d\ x = \int_{-\infty}^{\infty} x^2(0.5\delta(x+1) + 0.5\delta(x-1))d\ x = 0.5(-1)^2 + 0.5(1)^2 = 1\ .$$

Similarly, assume that the CDF of a mixed random variable X is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 0.5 + 0.5x & 0 \leq x < 1 \\ 1 & x \geq 1\ . \end{cases}$$

The PDF can be computed as

$$f_X(x) = 0.5\delta(x) + 0.5I_{\{x\in(0,1(\}}$$

where the indicator function $I_{\{x\in(0,1(\}} = \begin{cases} 0 & x \notin A \\ 1 & x \in A \end{cases}$.

Note that we still maintain the fundamental relationships between CDF and PDF:

$$F_X(a) = \int_{-\infty}^{a} f_X(\omega) ds.$$

Furthermore, for random variables $Y = g(X)$, we still have

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) d\ x\ .$$

whenever the integrals are finite and well-defined.

Example 3.21
A service station has two servers that it can use to handle services: a robot that always completes its service in 10 seconds, and a human that completes its service in a random time, distributed uniformly between 5 and 15 seconds. When you request service, you will be assigned the robot with probability 0.6, and the human with probability 0.4. Let X denote the random variable representing the time at which your service request will be completed. Note that X can take values between 5 and 15 seconds, a continuous interval.

What is the CDF of X? Let's compute this using the Law of Total Probability. Let B_1 be the set of all outcomes where the robot performs your service, and B_2 be the set of all outcomes where the human performs the service. Then, B_1, B_2 is a partition of all the possible outcomes. Using the Law of Total Probability,

$$F_X(x) = \mathbb{P}[\{X \leq x\}] = \mathbb{P}[\{X \leq x\}|B_1]\mathbb{P}[B_1] + \mathbb{P}[\{X \leq x\}|B_2]\mathbb{P}[B_2]$$

From the information in the problem, $\mathbb{P}[B_1] = 0.6, \mathbb{P}[B_2] = 0.4$. We are also given

$$\mathbb{P}[\{X \leq x\}|B_1] = \begin{cases} 0 & x < 10 \\ 1 & x \geq 10\ . \end{cases}; \quad \mathbb{P}[\{X \leq x\}|B_2] = \begin{cases} 0 & x < 5 \\ \frac{x-5}{10} & 5 \geq x < 10 \\ 1 & x \geq 10\ . \end{cases}$$

The CDF of X is thus obtained by direct substitution into the formula above. It is clearly the CDF of a mixed random variable, as it has a discontinuity at $x = 10$.

We compute the PDF of X as the derivative of this CDF, as

$$f_X(x) = 0.6\delta(x-10) + 0.04 I_{\{x\in[5,15]\}} \ .$$

We can now compute the expected value of X and its variance as:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) d\ x = 0.6 \int_{-\infty}^{\infty} x\delta(x-10)dx + 0.04 \int_5^{15} x dx = 6 + 4 = 10 \\
\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x) d\ x = 0.6 \int_{-\infty}^{\infty} x^2\delta(x-10)dx + 0.04 \int_5^{15} x^2 dx \\
&= 60 + 0.04 \frac{15^3 - 5^3}{3} = 60 + 130/3 \\
\mathsf{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 60 + 130/3 - 100 = \frac{10}{3}
\end{aligned}$$

Chapter 4

Pairs of Random Variables

4.1 Multiple Random Variables

In the previous two chapters, we have seen one way to define multiple random variables on the same probability space $(\Omega, \mathcal{E}, \mathbb{P})$, by using a function $g(\cdot)$ to map a random variable $X(\omega)$ to a different random variable $Y(\omega) = g(X(\omega))$. However, it is natural in many experiments to generate more than one random variable for each outcome, and for the second random variable not to be derived from the value of the first random variable. Consider an experiment where one rolls two six-sided dice. One random variable, $X(\omega)$, is the value of the first die, and the other random variable, $Y(\omega)$, is the value of the second dice. In this case, notice that $Y(\omega)$ can have multiple values for each value of $X(\omega)$, which means that $Y(\omega)$ is not derived as a function of $X(\omega)$. In this experiment, we expect that the values that $Y(\omega)$ takes and $X(\omega)$ takes are not related, and appear uniformly in $\{1, 2, \ldots, 6\}$. We recognize that this experiment was simply the combination of two independent experiments, and that perhaps we can treat X and Y as random variables from different experiments. Thus, it would be sufficient to know the individual probability mass functions $P_X(x), P_Y(y)$ to conduct further analyses.

However, consider an experiment of rolling two dice, but generating two random variables as follows: the first, $X(\omega)$, is the sum of the dice outcomes, and the second, $Y(\omega)$ is the product of the dice outcomes. Now, $X(\omega)$ takes values in $\{2, 3, \ldots, 12\}$, and $Y(\omega)$ takes values in a very different discrete set. Furthermore, their values are related in unusual ways: if $X(\omega) = 2$, then $Y(\omega) = 1$. If $X(\omega) = 4$, then $Y(\omega) \in \{3, 4\}$. It is clear that the values of X, Y depend closely on the full outcome ω, and cannot be separated as two independent subexperiments. In essence, the random variables are now a two-dimensional function $\underline{g}(\omega) = (X(\omega), Y(\omega))$, with values in a discrete subset of $\Re^2$. The choice of function $\underline{g}(\cdot)$ defines the range $R_{X,Y}$ and will define a probability mass function in that range.

Note that both experiments use the same underlying probability space $(\Omega, \mathcal{E}, \mathbb{P})$, with the same outcomes Ω and the same discrete probability measure $\mathbb{P}$. However, we defined different random variables in the experiments. We could have generated more than two random variables for the same outcome. **Multiple random variables** are the result of a vector-valued function that assigns multiple real numbers to each outcome in the sample space. Intuitively, we can think of multiple random variables as the observations from an experiment that simultaneously produces two or more numbers for each outcome. The above discussion highlights that the relationship between multiple random variables is more general than what we saw in earlier chapters, where one random variable was derived from the other random variable by a function transformation.

In this chapter, we focus on generalizing the concepts we developed for scalar random variables in Chapter 2 and Chapter 3 to the experiments that generate two random variables $X(\omega), Y(\omega)$ for each outcome. In later chapters, we generalize this to experiments that generate random vectors of higher dimension for each outcome.

4.2 Pairs of Random Variables

Formally, a pair of random variables (X, Y) in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ consists of a vector-valued function from $\Omega \to \Re^2$. We also refer to such a pair (X, Y) of random variables as bivariate random variables, or

joint random variables.

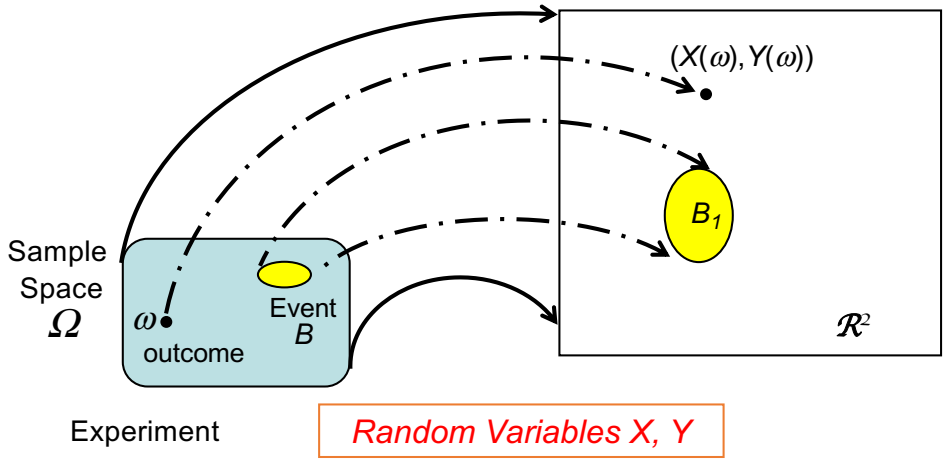


Figure 4.1: Bivariate random variables map single outcomes into two numerical values.

Figure 4.1 illustrates how pairs of random variables map individual outcomes $\omega \in \Omega$ into an ordered pair $(X(\omega), Y(\omega)) \in \Re^2$. We are interested in computing probabilities on the possible values of $X(\omega), Y(\omega)$, such as the probability that $(X(\omega), Y(\omega)) \in B_1 \subset \Re^2$ in Fig. 4.1. Thus, we restrict ourselves to functions where the *inverse image* of reasonable sets such as rectangular subsets of $\Re^2$ generate events $B \in \mathcal{E}$ for which $\mathbb{P}[B]$ is defined.[1] Then, we compute such probabilities as $\mathbb{P}[\{\omega \in \Omega : (X(\omega), Y($

For scalar random variables X, we defined the cumulative distribution function $F_X(x)$ as a function that summarized the probability of events defined in terms of intervals of values of X. For bivariate random variables X, Y, each random variable has its own CDF $F_X(x)$ and $F_Y(y)$, defined as in the previous chapters as $F_X(x) = \mathbb{P}[\{\omega \in \Omega : X(\omega) \leq x\}], F_Y(y) = \mathbb{P}[\{\omega \in \Omega : Y(\omega) \leq y\}]$. However, these CDF functions do not capture how the values of the random variables relate to each other.

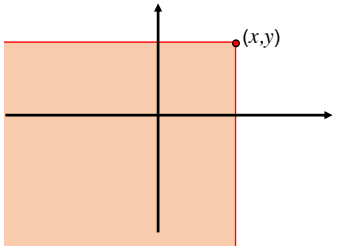


Figure 4.2: The CDF $F_{X,Y}(x, y)$ is the probability that the random variables take values in the shaded area .

To capture the probabilistic relationship between the two random variables, we define the **joint cumulative distribution function (CDF)** for values $(x, y) \in \Re^2$ as

$$F_{X,Y}(x, y) = \mathbb{P}[\{\omega \in \Omega : X(\omega) \leq x,\ Y(\omega) \leq y\}] = \mathbb{P}\big[\{\omega \in \Omega : X(\omega) \leq x\} \cap \{\omega \in \Omega : Y(\omega) \leq y\}\big].$$

That is, the joint CDF $F_{X,Y}(x, y)$ measures the probability of the event of outcomes where the random variables take values in the semi-infinite rectangle $(-\infty, x] \times (-\infty, y]$. This is illustrated in Figure 4.2. Note that this definition of CDF makes no distinction as to whether the joint random variables X, Y are discrete-valued or continuous-valued.

The joint CDF satisfies the following basic properties:

- **Non-negativity:** $0 \leq F_{X,Y}(x, y)$.
- **Normalization:** $\lim_{x,y\to\infty} F_{X,Y}(x, y) = 1$.
- **Non-decreasing:** For any $x \leq \tilde{x}$ and $y \leq \tilde{y}$, $F_{X,Y}(x, y) \leq F_{X,Y}(\tilde{x}, \tilde{y})$.

[1] Formally, we define the Borel σ-field in $\Re^2$ as that generated from two-dimensional intervals by countable unions, intersections and complementation, and we require the function $g(\omega) = (X(\omega), Y(\omega))$ to be measurable, so inverse images of Borel sets are events in $\mathcal{E}$.

- **Marginalization:** $\lim_{y\to\infty} F_{X,Y}(x,y) = F_X(x)$ and $\lim_{x\to\infty} F_{X,Y}(x,y) = F_Y(y)$.
- $\lim_{x,\to-\infty} F_{X,Y}(x,y) = 0$ and $\lim_{y\to-\infty} F_{X,Y}(x,y) = 0$.
- $F_{X,Y}(x,y)$ is a right-continuous function of x for each y and a right-continuous function of y for each x. That is,

$$\lim_{\epsilon>0,\epsilon\to 0} F_{X,Y}(x+\epsilon,y) = F_{X,Y}(x,y); \quad \lim_{\epsilon>0,\epsilon\to 0} F_{X,Y}(x,y+\epsilon) = F_{X,Y}(x,y).$$

Note that the marginal CDFs $F_X(x), F_X(y)$ can be derived from the joint CDF. The converse, however, is not true, as is clear from the dice discussion earlier.

Using the joint CDF, we can perform computations of the probabilities that the bivariate random variables take values in certain intervals, as illustrated in the following example.

Example 4.1
Compute the following probabilities using the joint CDF:

(a) $\mathbb{P}[\{X > x\} \cup \{Y > y\}]$ (b) $\mathbb{P}[\{X \le x\} \cup \{Y \le y\}]$
(c) $\mathbb{P}[\{X \le x\} \cup \{Y > y\}]$ (d) $\mathbb{P}[\{\omega \in \Omega : X(\omega) \in (x, x'], Y(\omega) \in (y, y']\}$

Answer: Figure 4.3 shows the areas in $\Re^2$ for the questions, with some ambiguity as to whether the red boundaries are part of the region of interest. Notice the specific choice of the questions, to determine the type of interval required, as that determines whether the boundary is included.

For (a), we see the answer is the complement of the joint CDF, as $\mathbb{P}[\{X > x\} \cup \{Y > y\}] = 1 - F_{X,Y}(x,y)$.

For (b), the answer is a little more complex:

$$\begin{aligned}\mathbb{P}[\{X \le x\} \cup \{Y \le y\}] &= \mathbb{P}[\{X \le x\}] + \mathbb{P}[\{Y \le y\}] - \mathbb{P}[[\{X \le x\} \cap \{Y \le y\}] \\ &= F_{X,Y}(x,\infty) + F_{X,Y}(\infty,y) - F_{X,Y}(x,y) = F_X(x) + F_Y(y) - F_{X,Y}(x,y)\end{aligned}$$

For (c), we see that $\{X \le x, y, Y \le y\} \cap \{Y > y\} = \emptyset$, so

$$\mathbb{P}[\{X \le x\} \cup \{Y > y\}] = F_{X,Y}(x,y) + (1 - F_{X,Y}(\infty,y)) = F_{X,Y}(x,y) + (1 - F_Y(y))$$

For (d), we have

$$\mathbb{P}[\{\omega \in \Omega : X(\omega) \in (x,x'], Y(x) \in (y,y']\} = F_{X,Y}(x',y') - F_{X,Y}(x,y') - F_{X,Y}(x',y) + F_{X,Y}(x,y)$$

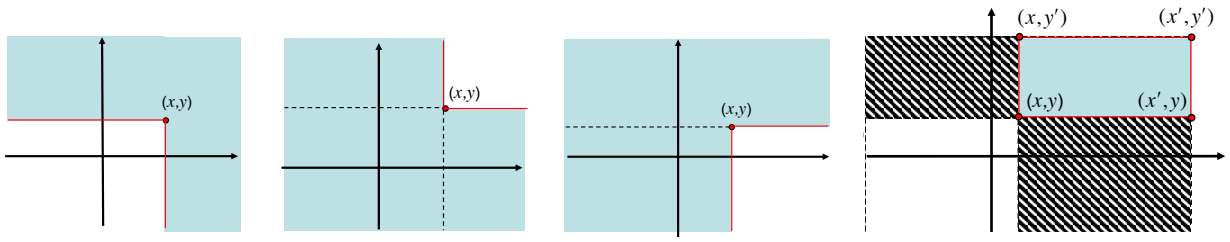


Figure 4.3: Regions of interest for the questions in 4.1.

4.3 Pairs of Discrete Random Variables

4.3.1 Joint Probability Mass Function

A pair of random variables X, Y is discrete if X and Y are discrete random variables. For discrete bivariate random variables, we define the joint probability mass function $P_{X,Y}(x,y)$ as follows:

Definition 4.1
The **joint probability mass function (PMF)** of a pair of discrete random variables X and Y is

$$P_{X,Y}(x,y) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x,\ Y(\omega) = y\}] = \mathbb{P}\big[\{X = x\} \cap \{Y = y\}\big].$$

The joint PMF is zero except at a discrete number of points in $\Re^2$, each of which has positive probability mass. The **range** $R_{X,Y}$ of a pair of discrete random variables is the set of all possible pairs of values,

$$R_{X,Y} = \{(x,y) : P_{X,Y}(x,y) > 0\}.$$

The joint PMF satisfies the following properties:

- **Non-negativity:** $P_{X,Y}(x,y) \geq 0$.
- **Normalization:** $\sum_{(x,y)\in R_{X,Y}} P_{X,Y}(x,y) = 1$.
- **Probability of an event:** Suppose we have a set $B \subset R_{x,y}$. Then,

$$\mathbb{P}[\{(x,y) \in B\}] = \mathbb{P}\big[\{(X,Y) \in B\}\big] = \sum_{(x,y)\in B} P_{X,Y}(x,y).$$

When the range sets R_X, R_Y of the two random variables are finite, we can visualize the joint PMF as an array of probability masses. Let $R_X = \{x_1, x_2, \ldots, x_n\}, R_Y = \{y_1, y_2, \ldots, y_m\}$, as illustrated in Table 4.1. Note that some of the numbers in the array can be zero, as the joint range $R_{X,Y}$ is often not equal to the cross product $R_X \times R_Y$.

$Y \backslash X$	x_1	x_2	$\cdots$	x_{n-1}	x_n
y_1	$P_{X,Y}(x_1,y_1)$	$P_{X,Y}(x_2,y_1)$	$\cdots$	$P_{X,Y}(x_{n-1},y_1)$	$P_{X,Y}(x_n,y_1)$
y_2	$P_{X,Y}(x_1,y_2)$	$P_{X,Y}(x_2,y_2)$	$\cdots$	$P_{X,Y}(x_{n-1},y_2)$	$P_{X,Y}(x_n,y_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{m-1}	$P_{X,Y}(x_1,y_{m-1})$	$P_{X,Y}(x_2,y_{m-1})$	$\cdots$	$P_{X,Y}(x_{n-1},y_{m-1})$	$P_{X,Y}(x_n,y_{m-1})$
y_m	$P_{X,Y}(x_1,y_m)$	$P_{X,Y}(x_2,y_m)$	$\cdots$	$P_{X,Y}(x_{n-1},y_m)$	$P_{X,Y}(x_n,y_m)$

Table 4.1: Visualization of joint PMF as a table of probability masses

From the joint PMF, we can obtain marginal PMFs for each random variable X, Y as follows: The **marginal PMF** $P_X(x)$ is just the PMF of X, and can be obtained from the joint PMF $P_{X,Y}(x,y)$ as:

$$P_X(x) = \sum_{y\in R_Y} P_{X,Y}(x,y).$$

Note we sum over all the possible values of the variable that we are trying to eliminate, Y. Similarly, the the marginal PMF $P_Y(y)$ is just the PMF of Y, obtained from the joint PMF as

$$P_Y(y) = \sum_{x\in R_X} P_{X,Y}(x,y).$$

In terms of the array representation in Table 4.1, $P_X(x)$ is obtained by summing the elements of the column corresponding to $X = x$, and $P_Y(y)$ is obtained by summing the elements of the row corresponding to $Y = y$.

Example 4.2
Given the details of an experiment, we can compute the joint PMF of a pair of discrete random variables X, Y by using the underlying probability measure $\mathbb{P}$ on the probability space $(\Omega, \mathcal{E}, \mathbb{P})$. Specifically,

$$P_{X,Y}(x,y) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x,\ Y(\omega) = y\}].$$

We illustrate this below.

Let the experiment consist of rolling two six-sided dice: an outcome ω is an ordered pair of numbers (a, b), with $a, b \in \{1, 2, \ldots, 6\}$. Each outcome in Ω is equally likely. We define the discrete random variables X, Y as follows:

$$X(\omega) = \begin{cases} 1 & \text{if the sum of dice rolls is odd,} \\ 0 & \text{otherwise;} \end{cases} \qquad Y(\omega) = \begin{cases} 1 & \text{if the product of dice rolls is odd,} \\ 0 & \text{otherwise.} \end{cases}$$

The range of X, Y is $R_{X,Y} = \{(0,0), (0,1), (1,0), (1,1)\}$. To compute the PMF, for each value $(x, y) \in R_{X,Y}$, we compute the set of outcomes $\omega \in \Omega$ that map to that value, and compute the probability of that set. For instance,

$$\{\omega \in \Omega : X(\omega) = 0, Y(\omega) = 0\} = \{(2,2), (2,4), (2,6), (4,2), (4,4), (4,6), (6,2), (6,4), (6,6)\}.$$

Thus,

$$P_{X,Y}(0,0) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = 0, Y(\omega) = 0\}] = \frac{9}{36} = \frac{1}{4}.$$

What about $P_{X,Y}(1,1)$? The set $\{\omega \in \Omega : X(\omega) = 1, Y(\omega) = 1\} = \emptyset$, because no pair of dice outcome can have an odd sum and an odd product! Thus, $P_{X,Y}(1,1) = 0$.

To complete the PMF, note that

$$\{\omega \in \Omega : X(\omega) = 0, Y(\omega) = 1\} = \{(1,1), (1,3), (1,5), (3,1), (3,3), (3,5), (5,1), (5,3), (5,5)\},$$

so $P_{X,Y}(0,1) = \frac{1}{4}$ also. Hence, by normalization, we must have $P_{X,Y}(1,0) = \frac{1}{2}$. We can verify this, as $\{\omega \in \Omega : X(\omega) = 1, Y(\omega) = 0\}$ will have the remaining 18 outcomes.

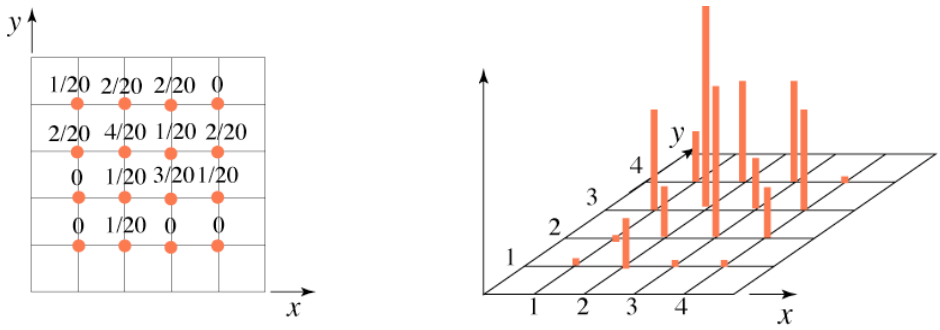


Figure 4.4: Figure for example 4.3.

Example 4.3
Consider a pair of random variables X, Y with joint PMF as illustrated in Figure 4.4, where the array representation of the joint PMF is shown on the left. Compute the probability that (X, Y) take values in the set $B = \{(x, y) : x \in [1, 2], y \in [2, 3]\}$. Also, compute the marginal PMF functions $P_X(x)$ and $P_X(y)$.

To compute $\mathbb{P}[\{(x, y) \in B\}]$, we use the joint PMF and add the probability over the masses at the points in B:

$$\mathbb{P}[\{(x, y) \in B\}] = P_{X,Y}(1,2) + P_{X,Y}(1,3) + P_{X,Y}(2,2) + P_{X,Y}(2,3) = 0 + \frac{2}{20} + \frac{1}{20} + \frac{4}{20} = \frac{7}{20}.$$

For the marginal PMFs, we first compute the PMF of X:

$$P_X(1) = \sum_{y \in R_Y} P_{X,Y}(1, y) = P_{X,Y}(1,3) + P_{X,Y}(1,4) = \frac{2}{20} + \frac{1}{20} = \frac{3}{20}$$

$$P_X(2) = \sum_{y \in R_Y} P_{X,Y}(2, y) = \frac{1}{20} + \frac{1}{20} + \frac{4}{20} + \frac{2}{20} = \frac{8}{20}$$

$$P_X(3) = \sum_{y \in R_Y} P_{X,Y}(3, y) = \frac{3}{20} + \frac{1}{20} + \frac{2}{20} = \frac{6}{20}$$

$$P_X(4) = \sum_{y \in R_Y} P_{X,Y}(4, y) = \frac{1}{20} + \frac{2}{20} = \frac{3}{20}$$

For the marginal PMF of Y, we compute:

$$P_Y(1) = \sum_{x \in R_X} P_{X,Y}(x,1) = P_{X,Y}(2,1) = \frac{1}{20}$$
$$P_Y(2) = \sum_{x \in R_X} P_{X,Y}(x,2) = P_{X,Y}(2,2) + P_{X,Y}(3,2) + P_{X,Y}(4,2) = \frac{5}{20}$$
$$P_Y(3) = \sum_{x \in R_X} P_{X,Y}(x,3) = P_{X,Y}(1,3) + P_{X,Y}(2,3) + P_{X,Y}(3,3) + P_{X,Y}(4,3) = \frac{9}{20}$$
$$P_Y(4) = \sum_{x \in R_X} P_{X,Y}(x,4) = P_{X,Y}(1,4) + P_{X,Y}(2,4) + P_{X,Y}(3,4) = \frac{5}{20}$$

4.3.2 Conditional PMF

For discrete random variables X in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, we defined the conditional probability mass function of X given observation of an event $B \in \mathcal{E}$ as

$$P_{X|B}(x) = \begin{cases} \dfrac{\mathbb{P}[\{s : X(\omega) = x\} \cap B]}{\mathbb{P}[B]}, & \mathbb{P}[B] > 0 \\ \text{undefined} & \text{otherwise.} \end{cases}$$

When we have a pair of random variables X, Y, the set B can be defined in terms of the random variable Y as $B = \{\omega \in \Omega : Y(\omega) = y\}$, which we write as an abbreviated $\{Y = y\}$. For this case, the following relationships hold:

$$\mathbb{P}[\{X = x\} \cap B] = \mathbb{P}[\{X = x\} \cap \{Y = y\}] = P_{X,Y}(x,y); \quad \mathbb{P}[B] = \mathbb{P}[\{Y = y\}] = P_Y(y)$$

Thus, we define the **conditional PMF** that $X = x$ given that $Y = y$ is observed as

$$P_{X|Y}(x|y) = \begin{cases} \dfrac{P_{X,Y}(x,y)}{P_Y(y)}, & P_Y(y) > 0 \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Note the following: if the numerator $P_{X,Y}(x,y) > 0$ for some x, y, then $P_Y(y) > 0$ and $P_X(x) > 0$ as obtained by marginalization. Hence, the reason for the *undefined* clause in the above equation is to handle the case when both numerator and denominator in the ratio are zero, in which case we don't define that conditional probability.

Similarly, we define the conditional PMF that $Y = y$ given that $X = x$ is observed as

$$P_{Y|X}(y|x) = \begin{cases} \dfrac{P_{X,Y}(x,y)}{P_X(x)}, & P_X(x) > 0 \\ \text{undefined} & \text{otherwise.} \end{cases}$$

In essence, the conditional probability mass function is the ratio of the joint PMF to the marginal PMF of the variable being observed. When both $P_X(x) > 0, P_Y(y) > 0$, we can also represent the joint CMF as the product of the conditional CMF and the marginal CMF, as

$$P_{X,Y}(x,y) = P_{X|Y}(x|y)P_Y(y) = P_{Y|X}(y|x)P(x).$$

We refer to this property as the **Multiplication Rule.**

The conditional PMF $P_{X|Y}(x|y)$ is a valid probability mass function on R_X, and thus satisfies the following basic properties of probability mass functions:

- **Non-negativity:** $P_{X|Y}(x|y) \geq 0$ and $P_{Y|X}(y|x) \geq 0$ for all $x \in R_X$, $y \in R_Y$.

- **Normalization:** $\sum_{x \in R_X} P_{X|Y}(x|y) = 1$ for any $y \in R_Y$ and $\sum_{y \in R_Y} P_{Y|X}(y|x) = 1$ for any $x \in R_X$.
- **Additivity:** For any event $B \subset R_X$, the probability that X falls in B given $Y = y$ is

$$\mathbb{P}[B|\{Y = y\}] = \sum_{x \in B} P_{X|Y}(x|y) \text{ for } y \in R_Y.$$

For any event $B \subset R_Y$, the probability that Y falls in B given $X = x$ is

$$\mathbb{P}[B|\{X = x\}] = \sum_{y \in B} P_{Y|X}(y|x) \text{ for } x \in R_X.$$

Example 4.4
Consider two random variables X, Y with the joint PMF function used in the previous example 4.3, illustrated in Figure 4.4. Compute $P_{X|Y}(x|y)$ for $y = 2, y = 3$. The table below has the joint PMF of X, Y for this example. To compute

$Y \backslash X$	1	2	3	4
1	0	$\frac{1}{20}$	0	0
2	0	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{1}{20}$
3	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	0

Table 4.2: Visualization of joint PMF as a table of probability masses

$P_{X|Y}(x|3)$, we **restrict** the the value of Y to the row $Y = 3$. We sum the probability masses in that row to get $P_Y(3) = \frac{9}{20}$. We use these to rescale the values in that row to get:

$$P_{X|Y}(1|3) = \frac{P_{X,Y}(1,3)}{P_Y(3)} = \frac{\frac{2}{20}}{\frac{9}{20}} = \frac{2}{9}$$
$$P_{X|Y}(2|3) = \frac{P_{X,Y}(2,3)}{P_Y(3)} = \frac{\frac{4}{20}}{\frac{9}{20}} = \frac{4}{9}$$
$$P_{X|Y}(3|3) = \frac{P_{X,Y}(3,3)}{P_Y(3)} = \frac{\frac{1}{20}}{\frac{9}{20}} = \frac{1}{9}$$
$$P_{X|Y}(4|3) = \frac{P_{X,Y}(4,3)}{P_Y(3)} = \frac{\frac{2}{20}}{\frac{9}{20}} = \frac{2}{9}$$

Notice that $P_{X|Y}(x|3)$ is proportional to the row $P_{X,Y}(x,3)$, **rescaled** by dividing by $P_Y(3)$ so that $\sum_{x \in R_X} P_{X|Y}(x|3) = 1$.

Similarly, $P_{X|Y}(x|2)$ is computed as follows: $P_Y(2) = \sum_{x \in R_X} P_{X,Y}(x,2) = \frac{5}{20}$. Then,

$$P_{X|Y}(1|2) = \frac{P_{X,Y}(1,2)}{P_Y(2)} = \frac{0}{\frac{5}{20}} = 0$$
$$P_{X|Y}(2|2) = \frac{P_{X,Y}(2,2)}{P_Y(2)} = \frac{\frac{1}{20}}{\frac{5}{20}} = \frac{1}{5}$$
$$P_{X|Y}(3|2) = \frac{P_{X,Y}(3,2)}{P_Y(2)} = \frac{\frac{3}{20}}{\frac{5}{20}} = \frac{3}{5}$$
$$P_{X|Y}(4|2) = \frac{P_{X,Y}(4,2)}{P_Y(2)} = \frac{\frac{1}{20}}{\frac{5}{20}} = \frac{1}{5}$$

The techniques we developed for conditional probabilities can be extended for conditional PMF functions, as follows. Let $R_Y = \{y_1, y_2, \ldots\}$ denote the discrete range of the random variable Y. Then, the events $\{\omega \in \Omega : Y(\omega) = y_1\}, \{\omega \in \Omega : Y(\omega) = y_2\}, \ldots$ are mutually disjoint if $y_1 \neq y_2$, because Y is a function, so there is only one value of y associated with an outcome $\omega \in \Omega$. Furthermore, they are collectively exhaustive,

because every $\omega \in \Omega$ must be mapped to some $y \in R_Y$. Thus, we can derive a version of the Law of Total Probability for pairs of discrete random variables X, Y , which is:

Law of Total Probability:

$$P_X(x) = \sum_{y \in R_Y} P_{X|Y}(x|y) P_Y(y)$$
$$P_Y(y) = \sum_{x \in R_X} P_{Y|X}(y|x) P_X(x) \ .$$

We can also develop a version of Bayes' Rule for pairs of discrete random variables, as:

Bayes' Rule:

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} = \frac{P_{Y|X}(y|x) P_X(x)}{P_Y(y)}$$
$$= \frac{P_{Y|X}(y|x) P_X(x)}{\sum_{x' \in R_X} P_{Y|X}(y|x') P_X(x')}$$

Example 4.5
Consider an X-ray source that generates photons with a specified rate λ photons per unit time. The emitted photons go through a mask that absorbs each photon with probability p, independently for each photon. For instance, in computed tomography machines, X-ray sources are typically modulated with masks to attenuate low-energy rays during X-ray imaging, as they contribute little to the quality of the image and get absorbed in body tissues.

Assume we operate the X-ray source for a single unit of time. The number of photons emitted is represented as a Poisson random variable with parameter λ, denoted by N. That is,

$$\mathbb{P}[\{N = n\}] \equiv P_N(n) = \frac{\lambda^n}{n!} e^{-\lambda}, \ n = 0, 1, 2, \ldots$$

We are interested in the number of photons that make it through the mask. That is a second random variable X. Note that if we know that $N = n$, then we can characterize the type of random variable that X is: There are n independent trials for photons to go through, and the success rate of each trial is $(1-p)$. Thus, conditioned on $N = n$, X is a binomial random variable with parameters n, $(1-p)$. That is,

$$\mathbb{P}[\{X = k\}|\{N = n\}] \equiv P_{X|N}(k|n) = \binom{n}{k} (1-p)^k p^{n-k}, \ k = 1, 2, \ldots, n.$$

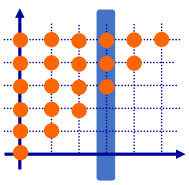
With these ideas, we can define the joint PMF of N, X as the product of a conditional probability and a marginal probability, as

$$P_{N,X}(n,x) = P_{X|N}(x|n) P_N(n) = \binom{n}{x} (1-p)^x p^{n-x} \frac{\lambda^n}{n!} e^{-\lambda}.$$

The range of values for N, X require that $X \leq N$, so it is

$$R_{N,X} = \{(n,x) : n \in \{0, 1, 2, \ldots\}, x \in \{0, 1, \ldots, n\}\}$$

We can now perform computations that would be of interest, such as finding the marginal probability of X, the number of photons that make it through the mask. We get the marginal probability of X from the joint probability of N, X by marginalization over the values of N. Notice that, for a particular value of $X = x$, we have $P_{N,X}(n,x) = 0, n < x$, as illustrated in the figure on the right.



Thus, the marginal probability of X is computed as:

$$\begin{aligned}
P_X(x) &= \sum_{n=0}^{\infty} P_{N,X}(n,x) = \sum_{n=x}^{\infty} \binom{n}{x}(1-p)^x p^{n-x}\frac{\lambda^n}{n!}e^{-\lambda} \text{ where the lower sum limit is } x \text{ because of the range } R_{N,X} \\
&= \sum_{n=x}^{\infty} \frac{n!}{x!(n-x)!}(1-p)^x p^{n-x}\frac{\lambda^n}{n!}e^{-\lambda} = \sum_{n=x}^{\infty} \frac{\lambda^n}{x!(n-x)!}(1-p)^x p^{n-x}e^{-\lambda} \text{ (cancel the } n! \text{ terms)} \\
&= \frac{\big(\lambda(1-p)\big)^x}{x!}e^{-\lambda}\sum_{n'=0}^{\infty}\frac{(\lambda p)^{n'}}{(n')!} \text{ (substitute } n-x=n') \\
&= \frac{\big(\lambda(1-p)\big)^x}{x!}e^{-\lambda}e^{\lambda p} = \frac{\big(\lambda(1-p)\big)^x}{x!}e^{-\lambda(1-p)} \text{ (recognizing the sum is an exponential.)}
\end{aligned}$$

Remarkably, we have just proven that the number of photons that make it through the mask is also a Poisson random variable, with parameter $\lambda(1-p)$, which is the product of the incoming photon intensity times the probability that each photon makes it through. Thus, we know that the expected number of photons that make it through the mask is $\lambda(1-p)$, and the variance of the number of photons that make it through the mask is also $\lambda(1-p)$.

The above result can be stated generally as: A Poisson random variable with intensity λ that undergoes independent sampling for each instance with probability p remains a Poisson random variable with a reduced intensity $p(1-\lambda)$. This result has many applications in engineering: For instance, consider a fork in a traffic road, where cars randomly choose with probability p to take the left fork and with probability $(1-p)$ to take the right fork. If the number of arrivals to the fork is modeled as a Poisson random variable with intensity λ, the number of departures on the left fork will be a Poisson random variable with intensity λp. Similarly, the number of departures on the right fork will be a Poisson random variable with intensity $\lambda(1-p)$.

Many sensor systems that count particles using physical mechanisms are modeled similarly. For instance, Geiger counters for radiation detection interact with α-particles, and detect each particle with a given probability. X-ray detector panels use scintillating materials that interact with incoming X-ray photons, and convert each photon to electrons with a given probability. If the arrival of particles is modeled as a Poisson random variable, the measured counts in these systems will also be Poisson random variables, albeit with reduced intensity.

Example 4.6
Consider the model of the previous example 4.5 for the pair of random variables N, X. Assume we observe that $X = 5$. What is the conditional probability distribution of N, given the information that $X = 5$?

To solve this, we apply Bayes' Rule for discrete random variables, as $P_{N|X}(n|x) = \dfrac{P_{N,X}(n,x)}{P_X(x)}$.

Fortunately, we have expressions for all of these from the previous problem:

$$P_{N,X}(n,x) = \binom{n}{x}(1-p)^x p^{n-x}\frac{\lambda^n}{n!}e^{-\lambda};\ P_X(x) = \frac{\big(\lambda(1-p)\big)^x}{x!}e^{-\lambda(1-p)}\ .$$

Hence, for observing $X = x$,

$$P_{N|X}(n|x) = \frac{P_{N,X}(n,x)}{P_X(x)} = \frac{\binom{n}{x}(1-p)^x p^{n-x}\frac{\lambda^n}{n!}e^{-\lambda}}{\frac{\big(\lambda(1-p)\big)^x}{x!}e^{-\lambda(1-p)}} = \frac{I_{\{n\geq x\}}\frac{n!}{x!(n-x)!}(1-p)^x p^{n-x}\frac{\lambda^n}{n!}e^{-\lambda}}{\frac{\big(\lambda(1-p)\big)^x}{x!}e^{-\lambda(1-p)}}$$

where the indicator function $I_{\{n\geq x\}}$ is 1 if the condition is true ($n \geq x$), and zero otherwise. Canceling the appropriate factors in the numerator and denominator, we get

$$P_{N|X}(n|x) = \frac{I_{\{n\geq x\}}\frac{\lambda^n}{(n-x)!}(1-p)^x p^{n-x}e^{-\lambda}}{\big(\lambda(1-p)\big)^x e^{-\lambda(1-p)}} = I_{\{n\geq x\}}\frac{\lambda^{n-x}}{(n-x)!}p^{n-x}e^{-\lambda p}$$

Substituting $x = 5$ above gives the desired conditional PMF for N.

We can recognize what type of conditional distribution is $P_{N|X}(n|x)$ by defining a derived random variable, conditioned on knowing $X = x$, as $N' = N - x$. Note that

$$P_{N|X}(n|x) = I_{\{n\geq x\}}\frac{(\lambda p)^{n-x}}{(n-x)!}e^{-\lambda p} = I_{\{n'\geq 0\}}\frac{(\lambda p)^{n'}}{(n')!}e^{-\lambda p} = P_{N'|X}(n'|x).$$

Thus, conditioned on $X = x$, N has the PMF of the sum of x and a Poisson random variable with intensity λp. Notice that the gap from $X = x$ to $N = n$ corresponds to absorbed photons, and the probability of absorption for each photon is p. As discussed in example 4.5, the number of absorbed photons is a Poisson random variable with intensity λp. Using this reasoning, we could have obtained the above answer directly with no computation.

4.4 Pairs of Continuous Random Variables

For discrete scalar random variables X, the PMF $P_X(x)$ described how probability mass accumulated in discrete points in the real line $\Re$. In contrast, a continuous random variable X spreads its probability over the real line $\Re$ so that there is no probability mass at any point, but instead we have a probability density function (PDF) $f_X(x)$, measured in probability per unit length, that describes how probability is accumulated. Indeed, for a random variable X to be continuous, its cumulative distribution function (CDF) $F_X(x)$ must be differentiable almost everywhere, and

$$f_X(x) = \begin{cases} \frac{d}{dx} F_X(x) & \text{if } F_X(x) \text{ is differentiable at } x, \\ \text{arbitrary} & \text{elsewhere.} \end{cases}$$

We want to extend these concepts to bivariate random variables (X, Y) defined on a common probability space $(\Omega, \mathcal{E}, \mathbb{P})$. In the previous section, we saw how we defined discrete bivariate random variables, and characterized their properties using the joint PMF function $P_{X,Y}(x, y)$. We define the concept of **jointly continuous** bivariate random variables as follows.

Definition 4.2
A pair of random variables X, Y are said to be **jointly continuous** if their joint CDF $F_{X,Y}(x, y)$ is continuous, and differentiable almost everywhere, so that there exists a joint probability density function $f_{X,Y}(x, y)$ such that

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x', y') d\, x' d\, y' \ .$$

An implication of this definition is that there is no region $B \subset \Re^2$ where the area of B is zero, and the probability $\mathbb{P}[\{\omega \in \Omega : (X(\omega), Y(\omega)) \in B\}] > 0$. Thus, there are no points which have positive probability masses, and there are no lines or curves with zero area that have positive probability of occurring.

4.4.1 Joint Probability Density Function

From the above definition, the joint probability density function (PDF) $f_{X,Y}(x, y)$ is computed as

$$f_{X,Y}(x, y) = \begin{cases} \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x, y) \partial x \partial y & \text{if } F_{X,Y}(x, y) \text{ is differentiable at } (x, y), \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

The **range** $R_{X,Y}$ of a pair of continuous random variables is the set of all possible pairs of values,

$$R_{X,Y} = \big\{(x, y) : f_{X,Y}(x, y) > 0\big\}.$$

The joint PDF has some structural properties that we highlight below:

- $f_{X,Y}(x, y) \geq 0$ for all $(x, y) \in \Re^2$. This follows from the fact that the joint CDF $F_{X,Y}(x, y)$ is monotone non-decreasing, and thus has non-negative derivatives.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x', y') d\, x' d\, y' = F_{X,Y}(\infty, \infty) = 1$. That is, the total volume between the surface map of $f_{X,Y}(x, y)$ and the x-y plane must be equal to 1.

Being a density and not a probability, the joint PDF can take positive values greater than 1.

Figure 4.5 illustrates the joint PDF of a pair of jointly continuous random variables X, Y. The top figure shows a small area ΔA around a point (x, y). the probability that (X, Y) take on values in ΔA is approximately computed as

$$\mathbb{P}[\{(X,Y) \in \Delta A\}] \approx f_{X,Y}(x,y)|\Delta A|,$$

where $|\Delta A|$ is the area of the region ΔA. This is approximately the volume of a column over ΔA, with height $f_{X,Y}(x, y)$, that is, the volume under the $f_{X,Y}$ graph that is above the area ΔA.
For any subset $A \in \Re^2$ with positive area, we compute the probability that (X, Y) take values in A using the joint PDF as

$$\mathbb{P}[\{(X,Y) \in A\}] = \iint_A f_{X,Y}(x,y)d\ x d\ y.$$

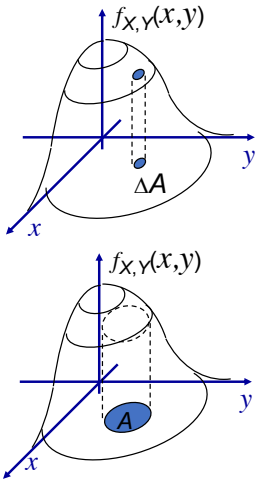


Figure 4.5: Illustration of joint PDF used for computation of probabilities.

Example 4.7
Consider a continuous random variable X defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$. For a continuous function $g : \Re \to \Re$, we define the random variable $Y = g(X)$. Is the pair (X, Y) a jointly continuous pair of random variables?

The answer is no. To make this discussion simpler, let $g(x) = x$, and define the region $B = \{(x, y) \in \Re : x = y\}$. Note that this region is a line in the x-y plane, and has no area: $|B| = 0$. However, it is clear that $\mathbb{P}[\{(X, Y) \in B\}] = 1$, so that there is probability mass for a set of zero area. Hence, the pair of random variables is not jointly continuous.

You can extend this argument for any continuous function $g(\cdot)$. Basically, the set $B = \{(x, y) \in \Re : x = y\}$ represents a continuous line in the x-y plane which has zero area, and the probability that (X, Y) take values in B is one. This argument can also be extended to discontinuous functions $g(\cdot)$.

Example 4.8
Consider a pair of random variables (X, Y) with joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} 1 & 0 \le x \le 1, 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the plot of this joint PDF is a cube of height 1 over the rectangle of area 1, and hence this join PDF satisfies the properties highlighted above: It is nonegative, and it integrates to 1, as the volume under the graph is 1.

What is the joint CDF of (X, Y)? By definition, this is $F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{\infty}^{y} f_{X,Y}(x,y)d\ x d\ y.$

Note that this integral is zero as long as either $x \le 0$ or $y \le 0$, as the integral takes place over a region where $f_{X,Y}(x, y) = 0$. Furthermore, if $x > 1$ and $y > 1$, then $F_{X,Y}(x, y) = 1$, because we integrate over the entire region where $f_{X,Y}(x, y) > 0$, namely the range $R_{X,Y}$. Elsewhere, we integrate to compute $F_{X,Y}(x, y)$. To make this easier, lets rewrite the joint PDF of X, Y using indicator functions, as: $f_{X,Y}(x, y) = 1I_{\{x\in[0,1]\}}I_{\{y\in[0,1]\}}$. Then, for $x > 0, y > 0$, we have

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} I_{\{x\in[0,1]\}}I_{\{y\in[0,1]\}}d\ x d\ y \\ &= (\int_{0}^{x} I_{\{x\in[0,1]\}}d\ x)(\int_{0}^{y} I_{\{y\in[0,1]\}}d\ y) = \min(x,1)\min(y,1). \end{aligned}$$

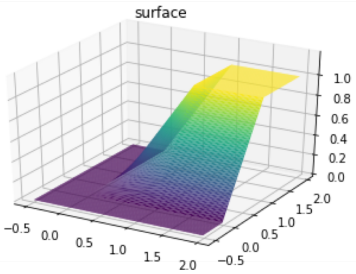


Figure 4.6: Joint CDF for Example 4.8.

Figure 4.6 shows a plot of the resulting CDF.

Putting all the equations together yields

$$F_{X,Y}(x,y) = \begin{cases} \min(x,1)\min(y,1), & x \geq 0, y \geq 0 \\ -0 & \text{otherwise.} \end{cases}$$

Example 4.9
Consider a pair of random variables (X,Y) with joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq x \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Compute the following probabilities: What is the probability that $X+Y>1$?. What is the probability that $(X-0.5)^2 + (Y-0.5)^2 < 0.25$?

A diagram is helpful to identify the sets involved in the answering the questions: First, the range $R_{X,Y}$ is a triangle in the plane, with corners (0,0), (1,1), (0,1). This is illustrated in Figure 4.7, which shows a plot of the joint PDF of (X,Y) with the range $R_{X,Y}$ outlined in the x-y plane. The intersection of the region $\{(x,y) : x+y \geq 1\}$ is highlighted in orange. Let $A = \{(x,y) : x+y \geq 1, 0 \leq x \leq y \leq 1\}$ denote that intersection region. The probability that $X+Y>1$ is the probability that (X,Y) take values in A, which is computed from the join PDF as $\mathbb{P}[\{(X,Y) \in A\}] = \iint_A f_{X,Y}(x,y)d\ xd\ y$.

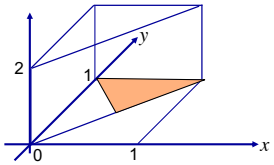


Figure 4.7: Joint PDF for Example 4.9.

Fortunately, the joint probability density function is constant ($=2$) in the region A, so we can compute the integral using simple geometric ideas: The volume of the region between the graph of the joint PDF and the area A is just the height times the area of the triangular base. The height is 2, and the triangle is seen to have a base of 1, height 0.5 so its area is 0.25. Hence, $\mathbb{P}[\{(X,Y) \in A\}] = 2 \times 0.25 = 0.5$.

Similarly, Let $B = \{(x,y) : (x-0.5)^2 + (y-0.5)^2 < 0.25\} \cap R_{X,Y}$. This area is highlighted in Figure 4.8. Then the probability that $(X-0.5)^2 + (Y-0.5)^2 < 0.25$ is the probability that (X,Y) takes values in B, which is $\mathbb{P}[\{(X,Y) \in B\}] = \iint_B f_{X,Y}(x,y)d\ xd\ y$. Reasoning as above, this is 2 times the area of B, which can be recognized from Figure 4.8 to be a half circle with radius 0.5. Hence,

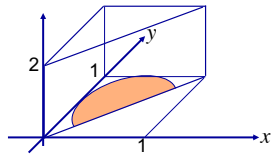


Figure 4.8: Joint PDF for Example 4.9.

$$\mathbb{P}[\{(X,Y) \in B\}] = 2 \times (0.5\pi(0.5)^2) = \frac{\pi}{4}\ .$$

4.4.2 Marginal PDF

If X,Y are jointly continuous random variables, then X and Y are continuous random variables individually, and have probability density functions $f_X(x)$ and $f_Y(y)$, called the marginal probability density functions. These can be computed from the joint CDF of (X,Y), by computing the marginal CDFs of X,Y and differentiating to obtain the marginal pdfs:

$$\begin{aligned} F_X(x) &= F_{X,Y}(x,\infty) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x',y)d\ ydx' \\ f_X(x) &= \frac{d}{dx}F_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\ F_Y(y) &= F_{X,Y}(\infty,y) = \int_{-\infty}^{y}\int_{-\infty}^{\infty} f_{X,Y}(x,y')d\ xdy' \\ f_Y(y) &= \frac{d}{dy}F_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x',y)d\ x'\ . \end{aligned}$$

Alternatively, we obtain the marginal PDF of X at $X=x$ directly from the joint PDF by integrating the joint PDF over all values y such that $(x,y) \in R_{X,Y}$. The result is still a density, not a probability.

Example 4.10
Let X, Y be jointly continuous random variables with PDF given by $f_{X,Y}(x,y) = \begin{cases} c(1-x-y) & 0 \le x \le (1-y) \le 1 \\ 0 & \text{otherwise,} \end{cases}$ where c is a constant that needs to be determined so that $\int_{-\infty}^{\infty} f_{X,Y}(x,y)d\ xd\ y = 1$. Find the value of c, and then the marginal PDF $f_X(x)$. Also, compute the probability that $X < Y$.

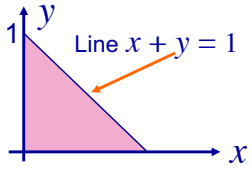


Figure 4.9: Range $R_{X,Y}$.

To begin with, it is always useful to visualize the range $R_{X,Y}$ where the joint PDF is non-zero. In this case, it is a triangle defined by the inequalities $0 \le x \le (1-y) \le 1$. Figure 4.9 shows this area, a triangle defined by the three inequalities $x \ge 0, y \ge 0, x+y \le 1$. This will help us evaluate the limits of integration for computing the marginal densities or the constant of integration. Let's compute c first. We can do this with geometry if we visualize the graph of the joint PDF as a pyramid, as shown in Figure 4.10 because the joint PDF is a linear function of x, y. Since we know the volume of a pyramid is $(1/3)\times$ base area $\times$ height, and the height is c, the volume under the joint PDF is

$$\iint_{R_{X,Y}} f_{X,Y}(x,y)dxdy = \frac{1}{3} \times \frac{1}{2} \times c = \frac{c}{6} = 1,$$

which implies that $c = 6$.

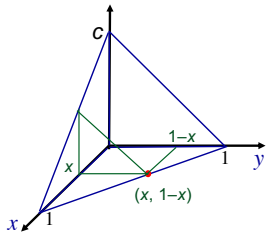


Figure 4.10: Joint PDF.

Alternatively, we compute this more generally from the double integral directly. Using the diagrams of Figures 4.9 and 4.10 to set the limits of integration, we obtain:

$$\iint_{R_{X,Y}} f_{X,Y}(x,y)dxdy = \int_0^1 \left(\int_0^{1-x} c(1-x-y)\, dy \right) dx$$

$$= \int_0^1 \frac{c(1-x)^2}{2} dx = \frac{c}{6}$$

and we get the same answer, $c = 6$.

To compute the marginal $f_X(x)$, we integrate the joint PDF over the range of possible values of Y with nonzero joint PDF for a given value $X = x$. Using the diagram of Figure 4.9 to set limits, we see that the range of values of Y for a given $X = x$ is $y \in [0, 1-x]$. Thus,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \begin{cases} \int_0^{1-x} 6(1-x-y)dy = 3(1-x)^2 & x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\int_{-\infty}^{\infty} f_X(x)dx = 1$, which is the normalization property of PDFs. By symmetry, we also get that the marginal PDF of Y is $f_Y(y) = \begin{cases} 3(1-y)^2 & y \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$

Finally, we compute the probability that $X \le Y$. If we are really clever, we see that the line $X = Y$ bisects $R_{X,Y}$ into two equal regions, so the volume under the joint PDF in the region $\{(x,y) : (x,y) \in R_{X,Y}, x \le y\}$ is exactly $1/2$. However, let's avoid cleverness and compute this as an integral, as one would have to do in a more general setting. The key is to visualize the region, and set the right limits for the integrals. We note that the maximum value for X such that $X \le Y$ is $1/2$, and that, for each value $X = x$, the region of values of y that we are interested is $y \in [x, 1-x]$. Then,

$$\mathbb{P}[\{X \le Y\}] = \int_0^{1/2} (\int_x^{1-x} 6(1-x-y)dy)dx = \int_0^{1/2} \frac{3}{2}(1-2x)^2 dx = \frac{1}{2}.$$

Example 4.11
Consider a pair of continuous random variables X, Y, uniformly distributed on the unit disk with radius 1, centered at (0,0). Thus, the joint PDF of X, Y is given by

$$f_{X,Y} = \begin{cases} \frac{1}{\pi} & 0 \le x^2 + y^2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The joint PDF of X, Y is illustrated in Figure 4.11.
For this problem, we want to compute $\mathbb{E}[X], \mathbb{E}[Y]$. We also want to compute the marginal PDFs $f_X(x), f_Y(y)$. Note that, by symmetry of the density, both $\mathbb{E}[X] = 0, \mathbb{E}[Y] = 0$.

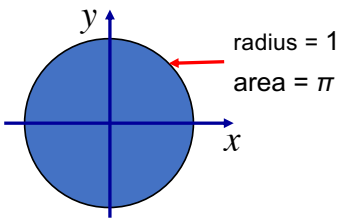


Figure 4.11: Example 4.11.

To compute the marginal density, it is useful to examine the range $R_{X,Y}$ illustrated in Fig. 4.11. For a given value of x, the values of y range from $-\sqrt{1-x^2}$ to $\sqrt{1-x^2}$. Then, the marginal density $f_X(x)$ is computed as:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\,dy = \frac{2\sqrt{1-x^2}}{\pi}.$$

By symmetry, it is clear that

$$f_Y(y) = \frac{2\sqrt{1-xy^2}}{\pi}.$$

Example 4.12
Consider joint continuous random variables X, Y with join PDF $f_{X,Y}(x,y) = \begin{cases} e^{-x} & 0 \le y \le x \le \infty \\ 0 & \text{otherwise} \end{cases}$.

Compute the marginal PDFs for X, Y, and compute the probability that $X + Y \le c$ for a constant $c \ge 0$.

It is useful to visualize the region $R_{X,Y}$ where the joint PDF is positive. This is an infinite triangle in the x-y plane, with origin at (0,0), bounded by the x-axis and the line $x = y$. Using this to compute limits, we see that, for a fixed x, the range of possible values of y is from 0 to x; for a fixed y, the range of possible x is from y to ∞. Thus,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \begin{cases} \int_0^x e^{-x}dy = xe^{-x} & x \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \begin{cases} \int_y^{\infty} e^{-x}dx = e^{-y} & y \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

To compute the probability that $X + Y \le c$, for $c \ge 0$, visualize the area in the x-y plane where $(x,y) \le c$ and $(x,y) \in R_{X,Y}$: This is a triangular area, where $y \in [0, c/2]$, and $x \in [y, c-y]$. Denote this area as B. This helps set the limits for the integrals to compute the probability as:

$$\begin{aligned}\mathbb{P}[\{X+Y \le c\}] &= \iint_B f_{X,Y}(x,y)dxdy = \int_0^{c/2}(\int_y^{c-y} e^{-x}dx)dy \\ &= \int_0^{c/2}(e^{-y} - e^{y-c})dy = 1 - e^{-c/2} - e^{-c/2} + e^{-c} = (1-e^{-c/2})^2.\end{aligned}$$

This last computation is useful if we wanted to define a derived random variable $Z = X + Y$. We have just computed

$$F_Z(z) = \mathbb{P}[\{Z \le z\}] = \mathbb{P}[\{X+Y \le z\}] = \begin{cases} (1-e^{-z/2})^2 & z \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

From this, we can compute the PDF of Z as

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \begin{cases} (1-e^{-z/2})e^{-z/2} & z \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

This is a useful technique for computing the PDFs of derived random variables: Compute the CDF first, then differentiate to get the PDF.

4.4.3 Conditional PDF

We want to extend the concept of conditional probability to jointly continuous random variables. Let X, Y be jointly continuous random variables, and define the set $A = \{(X,Y) \in B\}$ for some $B \subset R_{X,Y}$. Conditioned on observing that A has occurred, we define the conditional CDF of X, Y given A using the definition of conditional probability for events, as

$$F_{X,Y|A}(x,y|A) = \begin{cases} \frac{\mathbb{P}[\{X\le x, Y\le y\}\cap A]}{\mathbb{P}[A]}, & \mathbb{P}[A] > 0 \\ \text{undefined} & \mathbb{P}[A] = 0 \end{cases}.$$
$$= \begin{cases} \frac{\iint_{(-\infty,x]\times(-\infty,y]\cap B} f_{X,Y}(x,y)dxdy}{\iint_B f_{X,Y}(x,y)dxdy} & \mathbb{P}[A] > 0 \\ \text{undefined} & \mathbb{P}[A] = 0 \end{cases}.$$

That is, we restrict the probability to values $(x,y) \in B$, and rescale the probability so that it satisfies the normalization properties. From this, we can obtain the conditional density as

$$f_{X,Y|A}(x,y|A) = \frac{\partial}{\partial x}\frac{\partial}{\partial y} F_{X,Y|A}(x,y|A).$$

This yields the result:

$$f_{X,Y|A}(x,y|A) = \begin{cases} \frac{f(x,y)}{\mathbb{P}[A]}, & (x,y) \in A, \mathbb{P}[A] > 0 \\ 0, & (x,y) \notin A, \mathbb{P}[A] > 0 \\ \text{undefined} & \mathbb{P}[A] = 0 \end{cases}.$$

which has the same interpretation we saw previously: we restrict the range of the conditional density to values in the observed set A, and we rescale the conditional density to satisfy the normalization property.

We are also interested in the conditional probability of X given observations of values of Y. Consider first observing the event $A = \{Y \le y\}$. From the definition of conditional probability for events,

$$F_{X|A}(x|A) = \frac{\mathbb{P}[\{X \le x\} \cap A]}{\mathbb{P}[A]} = \frac{\int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x',y')dx'dy'}{F_Y(y)}.$$

for all y such that $F_Y(y) > 0$. From this conditional CDF, we compute the conditional PDF of X given A, as

$$f_{X|A}(x|A) = \frac{d}{dx}F_{X|A}(x|A) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,y')dy'}{F_Y(y)}.$$

Example 4.13
Let X, Y be jointly continuous random variables with joint PDF $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \le x \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$. Let $A = \{Y \le 0.5\}$. Compute the conditional density of X given that A is observed.

Note that $R_{X,Y}$, the range where the joint PDF is positive, is a triangle formed by the lines $x = 0, x = y, y = 1$, which helps us identify the limits of integration. This is shown in Figure 4.12. Proceeding as above,

$$\mathbb{P}[A] = F_Y(0.5) = \int_0^{0} .5(\int_0^{y} 2dx)dy = 0.25 \quad \text{(2 times the area of orange triangle)}$$
$$f_{X|A}(x|A) = \frac{\int_{-\infty}^{0.5} f_{X,Y}(x,y')dy'}{F_Y(0.5)} = \begin{cases} \frac{\int_x^{0.5} 2dy}{0.25} = 4 - 2x & x \in (0, 0.5) \\ 0 & \text{otherwise.} \end{cases}$$

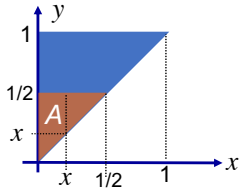


Figure 4.12: Range $R_{X,Y}$.

What if we observe the event $A = \{Y = y\}$? In this case, $\mathbb{P}[A] = 0$, so we cannot apply the definitions of conditional probability for events. We will use a limiting argument to define a conditional PDF of X given observation of the event $\{Y = y\}$, as follows.

Define the event $B = \{Y \in (y, y+\Delta]\}$ for some $\Delta > 0$. Then, $\mathbb{P}[B] = F_Y(y+\Delta) - F_Y(y)$. Assume we select y so that $\mathbb{P}[B] > 0$; that is, we select y in the interior of R_Y. Then, we define the conditional CDF

and PDF of X as:

$$F_{X|B}(x|B) = \frac{\mathbb{P}[\{X \le x\} \cap B]}{\mathbb{P}[B]} = \frac{\int_{-\infty}^{x}\int_{y}^{y+\Delta} f_{X,Y}(x',y')dx'dy'}{F_Y(y+\Delta) - F_Y(y)}.$$

From this conditional CDF,we get the density by differentiation:

$$f_{X|B}(x|B) = \frac{d}{dx}F_{X|B}(x|B) = \frac{\int_{y}^{y+\Delta} f_{X,Y}(x,y')dy'}{\int_{y}^{y+\Delta} f_Y(y')dy'}.$$

If we take limits as $\Delta \to 0$ in the above expression, both the numerator and denominator go to zero. However, using L'Hopital's rule, we can compute the limit as:

$$\lim_{\Delta\to 0} \frac{\int_{y}^{y+\Delta} f_{X,Y}(x,y')dy'}{\int_{y}^{y+\Delta} f_Y(y')dy'} = \lim_{\Delta\to 0} \frac{\frac{d}{d\Delta}\int_{y}^{y+\Delta} f_{X,Y}(x,y')dy'}{\frac{d}{d\Delta}\int_{y}^{y+\Delta} f_Y(y')dy'} = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

as long as $f_Y(y) > 0$. This allows us to define the conditional PDF of X when $Y = y$ as this limit:

Definition 4.3
Given two jointly continuous random variables X, Y, the **conditional PDF** of X given that $Y = y$ is given by

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & f_Y(y) > 0 \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Similarly, the conditional PDF of Y given $X = x$ is defined as

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & f_X(x) > 0 \\ \text{undefined} & \text{otherwise.} \end{cases}$$

The conditional PDF of X given $Y = y$ is a probability density for the continuous random variable X, and thus satisfies the following basic properties of probability densities:

- **Non-negativity:** $f_{X|Y}(x|y) \ge 0$ and $f_{Y|X}(y|x) \ge 0$ for all x and y where $f_X(x) > 0, f_Y(y) > 0$.
- **Normalization:** $\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,dx = 1$ for any y such that $f_Y(y) > 0$, and $\int_{-\infty}^{\infty} f_{Y|X}(y|x)\,dy = 1$ for any x such that $f_X(x) > 0$.
- **Additivity:** For any event $B \subset R_X$, the probability that X takes values in B given $Y = y$ is

  $$\mathbb{P}[\{X \in B\}|\{Y = y\}] = \int_B f_{X|Y}(x|y)\,dy.$$

  For any event $C \subset R_Y$, the probability that Y takes values in C given $X = x$ is

  $$\mathbb{P}[\{Y \in C\}|\{X = x\}] = \int_C f_{Y|X}(y|x)\,dx.$$

The techniques we developed for conditional probabilities also apply to conditional PDFs:

- **Multiplication Rule:** $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$.
- **Law of Total Probability:**

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x)\,f_X(x)\,dx$$

- **Bayes' Rule:**

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

Example 4.14
Consider two jointly continuous random variables X, Y, with joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} 6(1-x-y) & 0 \le x \le 1-y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

This is the same joint PDF considered in Example 4.10, illustrated in Figures 4.9 and 4.10. Let $\alpha \in (0,1)$. Compute $f_{X|Y}(x|\alpha)$.

Note that in Example 4.10, we computed the marginal PDF of Y as $f_Y(y) = \begin{cases} 3(1-y)^2 & y \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$

From the definition of conditional PDF, we have

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0, \\ \text{undefined} & \text{elsewhere.} \end{cases}$$

We need to be careful to account for limits in substituting in the numerator. Note that, if $Y = \alpha$, then $f_{X,Y}(x,\alpha) = 0$ if $x > 1-\alpha$. Thus,

$$f_{X|Y}(x|\alpha) = \begin{cases} \frac{6(1-x-\alpha)}{3(1-\alpha)^2} = 2\frac{(1-x-\alpha)}{(1-\alpha)^2} & \text{if } \alpha \in (0,1),\ x \in (0,1-\alpha), \\ 0 & \text{if } \alpha \in (0,1),\ x \notin (0,1-\alpha), \\ \text{undefined} & \alpha \notin (0,1). \end{cases}$$

4.5 Conditional Probability and Expectation

Given two discrete random variables X, Y, we have defined the conditional probability mass function $P_{X|Y}(x|y)$ as the probability that $X = x$ given that we have observed the event $Y = y$. Using this conditional PMF, we define the **conditional expected value of a function $g(X)$ given $Y = y$** as

$$\mathbb{E}[g(X)|Y=y] = \sum_{x \in R_X} g(x) P_{X|Y}(x|y).$$

Note that this expected value is a function of y, as we are averaging $g(X)$ over the conditional PMF of X given $Y = y$. Denote $h(y) = \mathbb{E}[g(X)|Y=y]$. Then, we can compute the expected value of $h(Y)$ over the PMF of Y, as

$$\mathbb{E}[h(Y)] = \sum_{y \in R_Y} h(y) P_Y(y).$$

Let's combine the last two equations, to get:

$$\begin{aligned} \mathbb{E}[h(Y)] = \mathbb{E}[\mathbb{E}[g(X)|Y]] &= \sum_{y \in R_Y} \mathbb{E}[g(X)|Y=y] P_Y(y) \\ &= \sum_{y \in R_Y} \sum_{x \in R_X} g(x) P_{X|Y}(x|y) P_Y(y) \\ &= \sum_{y \in R_Y} \sum_{x \in R_X} g(x) P_{X,Y}(x,y) \\ &= \mathbb{E}[g(X)] \end{aligned}$$

This last result is known as the smoothing property of conditional expectations. Basically, $\mathbb{E}\big[\mathbb{E}[g(X)|Y]\big] = \mathbb{E}[g(X)]$. In particular, this is true for the function $g(X) = X$, so that $\mathbb{E}\big[\mathbb{E}[X|Y]\big] = \mathbb{E}[X]$.

We can obtain a similar result for jointly continuous random variables X, Y. Given a function $g(x)$ we compute the conditional expected value of $g(X)$ given $Y = y$ using the conditional PDF of X given $Y = y$ as

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

Note that this will be a function of y, which we denote as $h(y)$. Proceeding as before,

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

which shows the smoothing property of conditional expectations also holds for jointly continuous random variables.

Example 4.15
Let X be a continuous random variable, uniformly selected in $(0, 1)$. Hence, $f_X(x) = \begin{cases} 1 & x \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$

Given that $X = x$, select Y to be a uniform random variable on $[0, x]$. That is,

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{x} & y \in [0, x] \\ .0 & \text{otherwise.} \end{cases}$$

Combining these two densities, we have

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x) = \begin{cases} \frac{1}{x} & 0 \le y \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Compute $\mathbb{E}[X|Y = y]$, and $\mathsf{Var}[X|Y = y]$.

We first compute the marginal density of y, given by

$$f_Y(y) = \int_y^1 \frac{1}{x}\, dx = -\ln(y).$$

Note that this integrates to 1 for $y \in [0, 1]$, as a density should. To compute the conditional density of X given $Y = y$, we need to compute the conditional density $f_{X|Y}(x|y)$, which we do using Bayes' Rule and the Law of Total Probability as

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \frac{\frac{1}{x}}{f_Y(y)},\ 0 \le y \le x \le 1 \\
&= \frac{\frac{1}{x}}{-\ln(y)} = \frac{-1}{x \ln(y)},\ 0 \le y \le x \le 1
\end{aligned}$$

Note how the limits of integration were evaluated for computing $f_Y(y)$, as we know that $x \in [y, 1]$. Using this conditional density, we get

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx = \int_y^1 x \frac{-1}{x \ln(y)} dx = \frac{y-1}{\ln(y)},\ y \in (0,1)$$

Let's now compute $\mathbb{E}[X]$ as:

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X|Y = y]\big] = \int_0^1 \frac{y-1}{\ln(y)} f_Y(y)\, dy = \int_0^1 \frac{y-1}{\ln(y)} (-\ln(y))\, dy = \int_0^1 (1-y)\, dy = \frac{1}{2}.$$

which is exactly what it should be, as X was a uniform random variable on $[0,1]$.

To compute the conditional variance of X given $Y=y$, we compute first $\mathbb{E}[X^2|Y=y]$:

$$\begin{aligned}\mathbb{E}[X^2|Y=y] &= \int_0^1 x^2 \frac{-1}{x\ln(y)}\,dx = \frac{-1}{\ln(y)}\int_0^1 x\,dx \\ &= \frac{-1}{2\ln(y)} \\ \mathsf{Var}[X|Y=y] &= \mathbb{E}[X^2|Y=y] - \Big(\mathbb{E}[X|Y=y]\Big)^2 = \frac{-1}{2\ln(y)} - \frac{(1-y)^2}{(\ln(y))^2}\end{aligned}$$

4.6 Independence of Pairs of Random Variables

In a probability space $(\Omega,\mathcal{E},\mathbb{P})$, two events $A,B\in\mathcal{E}$ are called independent if $\mathbb{P}[A\cap B]=\mathbb{P}[A]\mathbb{P}[B]$. For pairs of random variables X, Y, the concept of independence is stronger: we want events of the type $A=\{X\in C\subset R_X\}$ and $B=\{Y\in D\subset R_Y\}$ to be independent for any choice of $C\subset R_X, D\subset R_Y$. Fortunately, there is a simple way to check for independence without checking all such pairs of events. If the sets $C=(-\infty,x]$, and $D=(\infty,y]$, then $\mathbb{P}[\{X\in C\}\cap\{Y\in D\}]=F_{X,Y}(x,y)$. Furthermore, $\mathbb{P}[\{X\in C\}]=F_X(x)$ and $\mathbb{P}[\{Y\in D\}]=F_Y(y)$. Thus, independence requires that $F_{X,Y}(x,y)=F_X(x)F_Y(y)$ for all $(x,y)\in\Re^2$. It turns out that this condition is also sufficient to guarantee that $\mathbb{P}[\{X\in C\}\cap\{Y\in D\}]=\mathbb{P}[\{X\in X\}]\mathbb{P}[\{Y\in D\}]$ for any sets C and D defined by unions and intersections of intervals (Borel sets), because all those probabilities can be computed from the joint and marginal CDFs.

Definition 4.4
A pair of random variables X and Y are **independent** if and only if $F_{X,Y}(x,y)=F_X(x)F_Y(y)$.

For pairs of discrete random variables, the above condition leads to a characterization of independence in terms of the probability mass functions, as follows:

Lemma 4.1
A pair of discrete random variables X,Y are independent if and only if $P_{X,Y}(x,y)=P_X(x)P_Y(y)$.

Proof: To show the if part, assume $P_{X,Y}(x,y)=P(x)P(y)$. Note this means $R_{X,Y}=R_X\times R_Y$, because $P_{X,Y}(x,y)>0$ implies both $P_X(x)$ and $P_Y(y)$ are positive. Then,

$$\begin{aligned}F_{X,Y}(x,y) &= \sum_{\substack{(x_i,y_j)\in R_{X,Y}\\ x_i\le x, y_j\le y}} P(x_i)P(y_j) = \sum_{\substack{x_i\in R_X\\ x_i\le x}}\sum_{\substack{y_j\in R_Y\\ y_j\le y}} P(x_i)P(y_j) \\ &= \sum_{\substack{x_i\in R_X\\ x_i\le x}} P(x_i) \sum_{\substack{y_j\in R_Y\\ y_j\le y}} P(y_j) = F_X(x)F_Y(y)\end{aligned}$$

and hence the random variables X,Y are independent.

To show the only if part, assume X,Y are independent, so $F_{X,Y}(x,y)=F_X(x)F_Y(y)$. Again, this implies $R_{X,y}=R_X\times R_y$, because $F_{X,Y}(x,y)$ must change values everywhere $F_X(x)$ changes value and $F_Y(y)$ changes value. Let (x,y) be a point in $R_{X,Y}$. Since R_X,R_Y are discrete sets, there is an $\epsilon>0$ such that $F_X(x)-F_X(x-\epsilon)=P_X(x)$ and $F_Y(y)-F(y-\epsilon)=P_Y(y)$. We want to compute $\mathbb{P}[\{X\in(x-\epsilon,x]\}\cap\{Y\in(y-\epsilon,x]\}=P_{X,Y}(x,y)$. In Example 4.1 we showed that

$$\mathbb{P}[\{X\in(x-\epsilon,x]\}\cap\{Y\in(y-\epsilon,x]\}] = F_{X,Y}(x,y)-F_{X,Y}(x-\epsilon,y)-F_{X,Y}(x,y-\epsilon)+F_{X,Y}(x-\epsilon,y-\epsilon)$$

Since X, Y are independent, this means

$$\begin{aligned} P_{X,Y}(x,y) &= F_X(x)F_Y(y) - F_X(x-\epsilon)F_Y(y) - F_X(x)F_Y(y-\epsilon) + F_X(x-\epsilon)F_Y(y-\epsilon) \\ &= (F_X(x) - F_X(x-\epsilon))F_Y(y) - (F_X(x) - F_X(x-\epsilon))F_Y(y-\epsilon) \\ &= P_X(x)F_Y(y) - P(x)F_Y(y-\epsilon) = P_X(x)(F_Y(y) - F_Y(y-\epsilon)) \\ &= P_X(x)P_Y(y) \end{aligned}$$

For pairs of continuous random variables, we have a similar equivalent condition in terms of probability density functions:

Lemma 4.2
A pair of jointly continuous random variables X, Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

proof: The if direction is easy to prove, because

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x',y')dy'dx' = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x',y')dy'dx' = \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(x')f_Y(y')dy'dx' \\ &= (\int_{-\infty}^{x} f_X(x')dx')(\int_{-\infty}^{y} f_Y(y)dy')) = F_X(x)F_Y(y) \end{aligned}$$

and hence, X and Y are independent.

To show the only if direction, let X, Y be independent. Then, $F_{X,Y}(x,y) = F_X(x)F_y(y)$. Then,

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{X,Y}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_X(x)F_y(y) \\ &= (\frac{\partial}{\partial x}F_X(x))(\frac{\partial}{\partial y}F_Y(y)) = f_X(x)f_Y(y), \end{aligned}$$

establishing the result.

Independence is one of the most important properties used in modeling experiments with multiple random variables. By assuming independence, we can describe the two-dimensional joint PDF as a product of two one-dimensional PDFs.

Independence between a pair of random variables has implications on the conditional probability. For a pair of discrete random variables X, Y, we know that the conditional probability mass function of X given observations that $Y = y$ satisfies the following relationship: $P_{X,Y}(x,y) = P_{X|Y}(x|y)P_Y(y)$.. If X, Y are independent, then $P_{X,Y}(x,y) = P_X(x)P_Y(y)$. This means that, for independent X, Y, the conditional probability mass function is equal to the marginal, unconditional probability mass function:

$$P_{X|Y}(x|y) = P_X(x) \text{ for all } y \in R_Y \ .$$

A similar result applies to jointly continuous random variables X, Y that are independent. For jointly continuous X, Y, we know $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$. If X, Y are independent, then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Thus, for jointly continuous, independent X, Y, we have

$$f_{X|Y}(x|y) = f_X(x) \text{ for all } y \in R_Y \ .$$

Independence between pairs of random variables is often a property that is assumed. To prove that a pair of random variables are independent, one would have to verify the factorization property $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ or $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all values (x,y). In some cases, we can recognize that X, Y are dependent simply by looking at the range sets $R_X, R_Y, R_{X,Y}$. Specifically, if X, Y are independent random variables, then $R_{X,Y} = R_X \times R_Y$. For discrete random variables, this is because $P_{X,Y}(x,y) > 0$ only if both $P_X(x), P_Y(y) > 0$. Thus, to recognize two discrete random variables X, Y are dependent, we simply need

to find a pair (x, y) where $P_{X,Y}(x, y) = 0$, but $P_X(x) > 0, P_Y(y) > 0$. We can recognize this by finding a zero entry in the table representation of the joint PMF, where neither the entire row nor the entire column containing that entry is zero.

For jointly continuous, independent random variables X, Y, $R_{X,Y} = R_X \times R_Y$ follows because $f_{X,Y}(x, y) > 0$ only if both $f_X(x), f_Y(y) > 0$. Thus, to recognize that two random variables are dependent, we simply need to find a pair $(x, y) \in \Re^2$ where $f_{X,Y}(x, y) = 0$, but $f_X(x) > 0, f_Y(y) > 0$. We can recognize this by finding a point $(x', y') \in \Re^2$ where $f_{X,Y}(x', y') = 0$, but either the line $x = x'$ or the line $y = y'$ intersect $R_{X,Y}$. Thus, for jointly continuous, independent random variables, the range $R_{X,Y}$ must be of rectangular type with boundaries parallel to the edges.

Note that showing $R_{X,Y} = R_X \times R_Y$ is insufficient to show independence of X, Y. It is a necessary condition, so if it is not satisfied, then the random variables are not independent.

Example 4.16
Consider two discrete random variables X, Y with the joint PMF function used in examples 4.3 and 4.4, which is shown in the table below. We can quickly see that X, Y are not independent, as $P_{X,Y}(4, 4) = 0$, but the row corresponding to

$Y \backslash X$	1	2	3	4
1	0	$\frac{1}{20}$	0	0
2	0	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{1}{20}$
3	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	0

$Y = 4$ and the column corresponding to $X = 4$ are not identically zero. We could have picked several other zero entries to verify that X, Y are not independent.

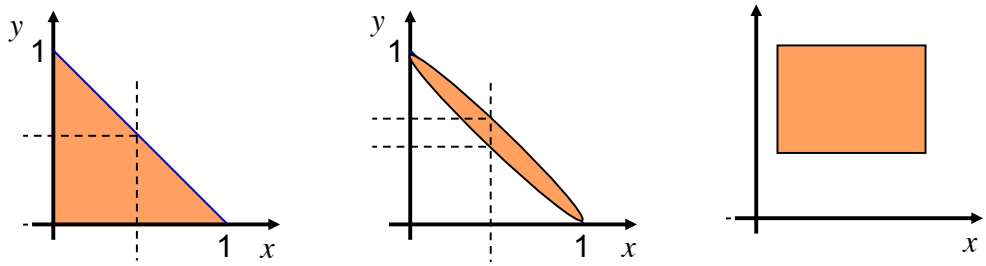


Figure 4.13: Figure for example 4.17.

Example 4.17
Assume X, Y are jointly continuous random variables with range $R_{X,Y}$ as one of the three ranges depicted in Figure 4.13. For which one of the three ranges can X, Y be independent random variables?

Consider the range on the left. We can select a point $(x, y) \notin R_{X,Y}$, such as $(0.6, 0.6)$, where $f_{X,Y}(x, y) = 0$. However, $f_X(0.6) > 0$, and $f_Y(0.6) > 0$, because the line $x = 0.6$ intersects the range $R_{X,Y}$ with a non-zero length, and the line $y = 0.6$ intersects the range $R_{X,Y}$ also with a non-zero length. Therefore, X and Y cannot be independent random variables.

Consider next the range in the center. Again, we can select a point $(x, y) = (0.3, 0.3) \notin R_{X,Y}$ so that the vertical and horizontal lines through this point intersect $R_{X,Y}$ with non-zero length. This implies that $f_{X,Y}(0.3, 0.3) = 0$ while $f_X(0.3) > 0, f_Y(0.3) > 0$, so X, Y cannot be independent.

On the other hand, if the range $R_{X,Y}$ is as depicted in the figure on the right, then we cannot find a point $(x', y') \notin R_{X,Y}$ where both the vertical line $x = x'$ and the horizontal line $y = y'$ have positive length intersection with $R_{X,Y}$. In this case, it is possible that X, Y are independent. To show independence, we need to verify that, for all $(x, y) \in R_{X,Y}$, we have $f_{X,Y}(x, y) = f_X(x) f_Y(y)$.

4.7 Expected Value of a Function of Two Random Variables

In the previous section, we have developed the concept of joint probability mass functions, and joint probability density functions, to characterize the properties of pairs of random variables X, Y on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$. Consider now a function $g : \Re^2 \to \Re$. This function defines a new random variable $W = g(X, Y)$. We can compute the **expected value**, or mean of W, using the joint PMF or the joint PDF of X, Y, as follows:

$$\begin{aligned} \text{Discrete:} \quad \mathbb{E}[W] &= \sum_{x \in R_X} \sum_{y \in R_Y} g(x,y) P_{X,Y}(x,y) \\ \text{Continuous:} \quad \mathbb{E}[W] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dx\, dy \end{aligned}$$

In either case, note that the expectation operation is a **linear operation:** For any functions $g_1(x, y), g_2(x, y)$, and constants a_1, a_2,

$$\mathbb{E}\big[a_1 g_1(X,Y) + a_2 g_2(X,Y)\big] = a_1 \mathbb{E}[g_1(X,Y)] + + a_2 \mathbb{E}[g_n(X,Y)],$$

This is because the expectation operation is based on summation and integration, both of which are linear operations. That is, for discrete random variables X, Y,

$$\begin{aligned} \mathbb{E}\big[a_1 g_1(X,Y) + a_2 g_2(X,Y)\big] &= \sum_{x \in R_X} \sum_{y \in R_Y} \big(a_1 g_1(x,y) + a_2 g_2(x,y)\big) P_{X,Y}(x,y) \\ &= \sum_{x \in R_X} \sum_{y \in R_Y} a_1 g_1(x,y) P_{X,Y}(x,y) + \sum_{x \in R_X} \sum_{y \in R_Y} a_2 g_2(x,y) P_{X,Y}(x,y) \\ &= a_1 \sum_{x \in R_X} \sum_{y \in R_Y} g_1(x,y) P_{X,Y}(x,y) + a_2 \sum_{x \in R_X} \sum_{y \in R_Y} g_2(x,y) P_{X,Y}(x,y) \\ &= a_1 \mathbb{E}\big[g_1(X,Y)\big] + a_2 \mathbb{E}\big[g_2(X,Y)\big] \end{aligned}$$

A similar argument shows the result for jointly continuous random variables using integrals instead of sums.

A useful special case is when the function $g(x, y)$ is an affine function, so that $g(x, y) = ax + by + c$ for some constants a, b, c. In this case,

$$\mathbb{E}[aX + bY + c] = \mathbb{E}[aX] + \mathbb{E}[bY] + \mathbb{E}[c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

Note that this is true regardless of whether X, Y are independent or not. It is strictly a consequence of the linearity of the expectation operator $\mathbb{E}[\cdot]$.

However, if X, Y were independent, and $g(x, y) = f_1(x) f_2(y)$ so that it can be written as a separable product of two functions, we have an interesting decomposition. Assume that X, Y were jointly continuous random variables with joint PDF $f_{X,Y}(x, y)$. Then,

$$\begin{aligned} \mathbb{E}[g(X,Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dx\, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(x) f_2(y) f_{X,Y}(x,y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(x) f_2(y) f_X(x) f_Y(y)\, dx\, dy \quad \text{because } X, Y \text{ are independent,} \\ &= \left(\int_{-\infty}^{\infty} f_1(x) f_X(x)\, dx \right) \left(\int_{-\infty}^{\infty} f_2(y) f_Y(y)\, dy \right) \\ &= \mathbb{E}[f_1(X)] \mathbb{E}[f_2(X)] \end{aligned}$$

The smoothing property of conditional expectation continues to apply to functions $g(X,Y)$. We show this for jointly continuous X,Y below, as

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

The above results allow us to compute the expected value of a random variable W that is derived from X,Y by a function $W=g(X,Y)$.

Example 4.18
Assume that the number of people in line at the bank when you arrive is N, where N is random, having a Poisson distribution with parameter α. The time T that it takes to serve each person ahead of you can be described by an exponential distribution with parameter λ, and is independent of N. The time to serve each person is thee same. How long do you expect to wait before someone starts to serve you?

Let W be the time you will wait. W is a function of N, T, as $W=NT$.

Since N,T are independent, $E[W]=E[T]E[N]=\frac{alpha}{\lambda}$.

4.7.1 Transformation of pairs of random variables

In some cases we want to compute the full probability mass function or probability density function of W, depending on whether W is discrete or continuous. If W is discrete with range R_W, then for each $w_i \in R_W$, we can define the inverse image of w as the set $g^{-1}(w)=A_w=\{(x,y)\in R_{X,Y} : g(x,y)=w\}$. As long as the function g is well-behaved, we compute the probability mass function of W as:

$$P_W(w)=\mathbb{P}[\{\omega : (X(\omega),Y(\omega))\in A_w\}]=\sum_{(x,y)\in A_w} F_{X,Y}(x,y).$$

Thus, we can readily derive the probability mass function of W from the joint probability mass function of X,Y, as long as we can readily compute the inverse image $g^{-1}(w)$.

If X,Y are jointly continuous, and the map $W=g(X,Y)$ results in a continuous random variable W, the above approach is limited because the probability that W takes on a particular value is zero. In this case,we can instead compute the cumulative distribution function $F_W(w)$. Let $B_w=\{(x,y)\in R_{X,Y} : g(x,y)\le w\}$ be the region in $R_{X,Y}$ that maps into values $g(x,y)\le w$. In this case, the CDF $F_W(w)$ can be computed as

$$F_W(w)=\iint_{(x,y)\in B_w} f_{X,Y}(x,y)\,dx\,dy.$$

From the CDF, we can get the PDF of W by differentiation, as $f_W(w)=\frac{d}{dw}F_W(w)$.

The above equations were derived for general functions $g(x,y)$, and require solving for the inverse maps of a region of values B_w when W is continuous W or A_w for discrete W. This can be challenging for complicated functions $g(x,y)$. However, there are cases of functions $g(x,y)$ where these inverse maps are straightforward to compute. For instance, let $g(x,y)=ax+by+c$ be a linear function, where $a,b\neq 0$. Then, the line $ax+by+c=w$ divides the x-y plane into two half planes, one of which is B_w. In particular, let's consider $W=X+Y$.

In this case, for discrete X, Y, the set $A_w = \{(x,y) \in \Re^2 : x+y = w\} = \{(x, w-x) : x \in \Re\}$. Therefore, the PMF of W can be computed as

$$P_W(w) = \sum_{x \in \Re} P_{X,Y}(x, w-x) = \sum_{x \in R_X} P_{X,Y}(x, w-x),$$

where the second equality follows because $P_{X,Y}(x, w-x) = 0$ unless $x \in R_X$.

This operation is illustrated in Figure 4.14. To get the probability mass $P_W(w)$, we sum up all the probability masses $P_{X,Y}(x,y)$ on the line $x+y = w$.

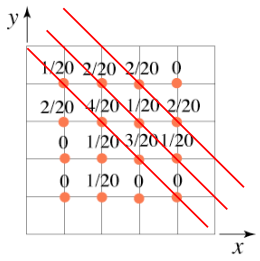


Figure 4.14: Projection to compute PMF of $X+Y$.

For jointly continuous X, Y, the set the set $B_w = \{(x,y) \in \Re^2 : x+y \leq w\} = \{(x,y) : x \in \Re, y \in (-\infty, w-x]\}$. Therefore, we compute the CDF $F_W(w)$ as

$$F_W(w) = \mathbb{P}[\{X+Y \leq w\}] = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{w-x} f_{X,Y}(x,y)\, dy\, dx.$$

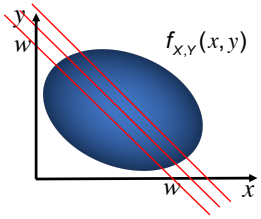


Figure 4.15: Projection to compute PDF of $X+Y$.

From this CDF, we compute the PDF of W by differentiating:

$$f_W(w) = \frac{d}{dw} \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{w-x} f_{X,Y}(x,y)\, dy\, dx = \int_{x=-\infty}^{\infty} \left(\frac{d}{dw} \int_{y=-\infty}^{w-x} f_{X,Y}(x,y)\, dy\right) dx = \int_{x=-\infty}^{\infty} f_{X,Y}(x, w-x)\, dx.$$

This operation is shown in Figure 4.15. In essence, one integrates the joint PDF along the line $x+y = w$. This is similar to computing a marginal distribution from a joint distribution, except we integrate along an inclined line instead of a vertical or horizontal line.

For the special case that X, Y are independent,

$$f_W(w) = \int_{x=-\infty}^{\infty} f_{X,Y}(x, w-x)\, dx = \int_{x=-\infty}^{\infty} f_X(x) f_Y(w-x)\, dx,$$

which shows that the probability density of the sum of independent random variables X and Y is the convolution of their probability densities.

Example 4.19

Assume we have a pair of continuous random variables X, Y with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} 6(1-x-y) & x \geq 0, y \geq 0, x+y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $Z = \max(X, Y)$. Find the probability density function $f_Z(z)$.

The joint PDF of X, Y is illustrated in Figure 4.16, where we have drawn also contours for equal values of z, illustrated by the red squares on the x-y plane. We first compute the cumulative distribution of Z for values $z \leq 0.5$. In this range, the region of integration B_z lies entirely in $R_{X,Y}$.

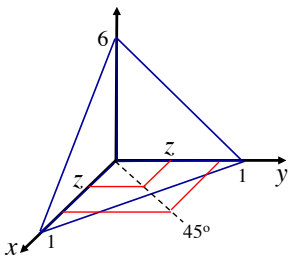


Figure 4.16: Figure for example 4.19.

Using the limits as indicated in Figure 4.16, we obtain for $0 \leq z \leq 0.5$,

$$\mathbb{P}[\{Z \leq z\}] = F_Z(z) = \mathbb{P}[\{X \leq z\} \cap \{Y \leq z\}] = \int_0^z \int_0^z f_{X,Y}(x,y)\, dx\, dy$$
$$= \int_0^z \int_0^z 6(1-x-y)\, dx\, dy = \int_0^z 6(1-y)z - 3z^2\, dy = 6z^2 - 6z^3, z \in [0, 0.5]$$

Note that $F_Z(z) = 0$ for $z \leq 0$. Furthermore, $F_Z(z) = 1$ for $z > 1$, as the region of integration expands to include all of the range $R_{X,Y}$.

For $z \in [0.5, 1]$, examine the diagram in Figure 4.16. The region of integration B_w now expands beyond $R_{X,Y}$. It is easier to compute this as follows:

$$\begin{aligned}
\mathbb{P}[\{Z \leq z\}] = F_Z(z) &= 1 - \mathbb{P}[\{X \geq z\}] - \mathbb{P}[\{Y \geq z\}] \\
&= 1 - \int_z^1 \int_0^{1-x} 6(1-x-y)\, dy\, dx - \int_z^1 \int_0^{1-y} 6(1-x-y)\, dx\, dy \\
&= 1 - \int_z^1 3(1-x)^2\, dx - \int_z^1 3(1-y)^2\, dy \\
&= 1 - 2(1-z)^3,\ z \in [0.5, 1]
\end{aligned}$$

At $z = 0.5$, $F_Z(z) = \frac{3}{4}$, which agrees with the value computed previously, as $F_Z(z)$ is a continuous function.

The density $f_Z(z)$ is now readily obtained by differentiating, to get

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \begin{cases} 0 & z \notin [0,1], \\ 12z - 18z^2 & z \in (0, 0.5), \\ 6(1-z)^2 & .z \in (0.5, 1). \end{cases}$$

Example 4.20
Let X, Y be independent, uniform$(0,1)$ random variables, and let $Z = X + Y$. Find the PDF of Z.

Note first that the range of Z will be $R_Z = [0, 2]$, the set of values that can have probability. Using the formula provided above for the sum of random variables,

$$f_Z(z) = \int_{x=-\infty}^{\infty} f_{X,Y}(x, z-x)\, dx = \int_{x=-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx \quad \text{by independence,}$$

We use the fact that $R_X = [0,1], R_Y = [0,1]$ to determine the limits of integration, as follows:

$$f_Z(z) = \begin{cases} 0 & z \notin [0,2], \\ \int_0^1 f_Y(z-x)\, dx = \int_0^z\ dx = z & z \in [0,1], \\ \int_0^1 f_Y(z-x)\, dx = \int_{z-1}^1\ dx = 2 - z & z \in [1,2] \end{cases}$$

Example 4.21
Let X, Y be independent, exponential(λ) random variables, and let $Z = X + Y$. Find the PDF of Z.

Note first that the range of Z will be $R_Z = [0, \infty)$, the set of values that can have probability. Using the formula provided above for the sum of random variables,

$$\begin{aligned}
f_Z(z) &= \int_{x=-\infty}^{\infty} f_{X,Y}(x, z-x)\, dx = \int_{x=-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx \quad \text{by independence,} \\
&= \begin{cases} 0 & z \leq 0, \\ \int_0^z \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)}\, dx & z \geq 0 \end{cases} = \begin{cases} 0 & z \leq 0, \\ \lambda^2 z e - \lambda z\, dx & z \geq 0 \end{cases}
\end{aligned}$$

The sum of two independent exponential random variables with the same rate parameter defines a random variable that has an Erlang$(2, \lambda)$ distribution. If we were to sum n independent exponential random variables with the same rate parameter λ, we obtain an Erlang(n, λ) random variable.

Example 4.22
Let X, Y be independent standard Gaussian random variables, so $X, Y \sim N(0,1)$, and let $Z = aX + bY + c$ for some constants $a \neq 0, b > 0, c$. Find the PDF of Z.

We start by finding the CDF of Z, exploiting the independence of X, Y and the positivity of b, as

$$F_Z(z) = \mathbb{P}[\{aX + bY + c \leq z\}] = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\frac{z-ax-c}{b}} f_X(x) f_Y(y)\, dy\, dx.$$

Differentiating, we get

$$f_Z(z) = \int_{x=-\infty}^{\infty} \left(\frac{d}{dz} \int_{y=-\infty}^{\frac{z-ax-c}{b}} f_X(x) f_Y(y)\, dy \right) dx$$
$$= \int_{x=-\infty}^{\infty} \frac{1}{b} f_X(x) f_Y(\frac{z-ax-c}{b})\, dx$$

Substitute the Gaussian PDF formulas for X, Y to get:

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-ax-c)^2}{2b^2}}\, dx$$

Let's manipulate the exponent in the integral to isolate the dependence on x as follows:

$$\begin{aligned}
-\frac{x^2}{2} - \frac{(z-ax-c)^2}{2b^2} &= -\frac{x^2}{2} - \frac{(ax)^2}{2b^2} + \frac{2ax(z-c)}{2b^2} - \frac{(z-c)^2}{2b^2} \\
&= -(1+\frac{a^2}{b^2})\frac{x^2}{2} + \frac{a(z-c)}{b^2}x - \frac{(z-c)^2}{2b^2} \\
&= \frac{(a^2+b^2)}{b^2}\left(-\frac{x^2}{2} + \frac{a(z-c)}{a^2+b^2}x - \frac{(z-c)^2}{2(a^2+b^2)} \right) \\
&= \frac{(a^2+b^2)}{b^2}\left(-\frac{(x-\frac{a(z-c)}{a^2+b^2})^2}{2} + \frac{a^2(z-c)^2}{2(a^2+b^2)^2} - \frac{(z-c)^2}{2(a^2+b^2)} \right) \\
&= -\frac{(x-\frac{a(z-c)}{a^2+b^2})^2}{2K^2} + \frac{a^2(z-c)^2}{2(a^2+b^2)b^2} - \frac{(z-c)^2}{2b^2} \quad \text{where } K^2 = \frac{b^2}{a^2+b^2} \\
&= -\frac{(x-\frac{a(z-c)}{a^2+b^2})^2}{2K^2} - \frac{(z-c)^2}{2(a^2+b^2)} \quad \text{where } K^2 = \frac{b^2}{a^2+b^2}
\end{aligned}$$

The reason for this transformation is to express the integral in terms of an integral for a Gaussian PDF. With this transformation, we have

$$f_Z(z) = \frac{1}{b}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\frac{a(z-c)}{a^2+b^2})^2}{2K^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-c)^2}{2(a^2+b^2)}}\, dx = \frac{1}{\sqrt{2\pi(a^2+b^2)}} e^{-\frac{(z-c)^2}{2(a^2+b^2)}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi K^2}} e^{-\frac{(x-\frac{a(z-c)}{a^2+b^2})^2}{2K^2}}\, dx$$
$$= \frac{1}{\sqrt{2\pi(a^2+b^2)}} e^{-\frac{(z-c)^2}{2(a^2+b^2)}}$$

because the last integral is the integral of a Gaussian PDF with variance K^2 and a given mean, which equals 1 because of the normalization property of PDFs. Furthermore, note that $f_Z(z)$ is also a Gaussian pdf, with mean c and variance a^2+b^2 (thus $Z \sim N(c, a^2+b^2)$.) We have just shown that an affine combination of two independent Gaussians will also be a Gaussian random variable. With a similar argument, we can show that any affine combination of Gaussian random variables will be a Gaussian random variable.

Example 4.23
Let X, Y be independent continuous random variables, and let $Z = \max(X, Y)$. Find the PDF of Z.

In contrast with Example 4.19, we don't specify the pdf of the random variables, but we specify that they are independent. We first derive the CDF of Z:

$$\mathbb{P}[\{Z \le z\}] = \mathbb{P}[\{X \le z\} \cap \{Y \le z\}] = \mathbb{P}[\{X \le z\}]\mathbb{P}[\{Y \le z\}] \quad \text{by independence.}$$

Hence,

$$F_Z(z) = F_X(z) F_Y(z)$$

and the PDF of Z can be obtained as

$$f_Z(z) = \frac{d}{dz} F_Z(z) = F_X(z) f_Y(z) + F_Y(z) f_X(z).$$

Given the CDF and PDF of the random variables X, Y, we can get the CDF and PDF of Z.

To illustrate this, consider the following pair of jointly continuous random variables X, Y, with joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} (1-\frac{x}{2})(1-\frac{y}{2}) & 0 \le x, y \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

Let $Z = \max(X, Y)$. Then,

$$f_X(x) = \begin{cases} (1 - \frac{x}{2}) & 0 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases} \qquad f_Y(y) = \begin{cases} (1 - \frac{y}{2}) & 0 \le y \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x < 0, \\ x - \frac{x^2}{4} & 0 \le x \le 2, \\ 1 & x > 2 \end{cases} \qquad F_Y(Y) = \begin{cases} 0 & y < 0, \\ y - \frac{y^2}{4} & 0 \le y \le 2, \\ 1 & y > 2 \end{cases}$$

Using the above formula,

$$f_Z(z) = \begin{cases} 0 & z < 0, \\ 2(z - \frac{z^2}{4})(1 - \frac{z}{2}) & 0 \le z \le 2, \\ 0 & z > 2. \end{cases}$$

Does the same idea work for the minimum of two random variables? Let $W = \min(X, Y)$. Then,

$$\mathbb{P}[\{W > w\}] = \mathbb{P}[\{X > w\} \cap \{Y > w\}] = \mathbb{P}[\{X > w\}]\mathbb{P}[\{Y > w\}] \quad \text{by independence.}$$

Hence, $1 - F_W(w) = (1 - F_X(w))(1 - F_Y(w))$ which leads to

$$F_W(w) = 1 - (1 - F_X(w))(1 - F_Y(w)) = F_X(w) + F_Y(w) - F_X(w)F_Y(w).$$

Differentiating with respect to w yields

$$f_W(w) = \frac{d}{dw} F_W(w) = (1 - F_Y(w))f_X(w) + (1 - F_X(w))f_Y(w)$$

We conclude this chapter with two examples from a mathematics competition. Questions like these often show up as interview questions for companies like Google. We state first the word problems, and then formulate the problem using pairs of random variables. These examples are difficult, but show how the techniques of this Chapter are used to formulate and solve problems.

Example 4.24
You have a stick of length 1. You pick a point along the stick, uniformly distributed, to break it into two pieces. You take the longer of the two pieces, you pick a point uniformly along that piece, and break the long piece into two pieces. You now have three pieces. What is the expected length of the shortest of the three pieces remaining?

Let X denote the length of the shorter piece remaining after the first break. Since the first break was uniform distributed, it is straightforward to compute the PDF of X as

$$f_X(x) = \begin{cases} 2 & 0 \le x \le 0.5 \\ .0 & \text{otherwise.} \end{cases}$$

The length of the longer piece is $1 - X$. Let Y denote the length of the shortest of the two pieces that remain after breaking the longer piece. Then, Y has conditional PDF

$$f_{Y|X}(y|x) = \begin{cases} \frac{2}{1-x}, & y \in [0, \frac{1-x}{2}] \\ 0 & \text{otherwise.} \end{cases}$$

and thus is distributed uniformly in $[0, \frac{1-X}{2}]$ using a similar argument as before. The joint PDF of X, Y is defined using using the multiplication rule as:

$$f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x) = \begin{cases} \frac{4}{1-x} & 0 \le x \le 0.5, 0 \le y \le \frac{1-x}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

We already know that Y is the shortest of the two pieces from the second break, and X is the length of the shortest piece after the first break. Hence, the length of the shortest of the three pieces is $\min(X, Y)$. We have now transformed the original problem into computing the expected value of a function of two random variables, where we know the joint PDF:

$$\mathbb{E}[\min(X, Y)] = \int_0^{\frac{1}{2}} \Big(\int_0^{\frac{1-x}{2}} \min(x, y) \frac{4}{1-x} \, dy \Big) \, dx$$

The rest is tedious calculus that is easy to do with a computer. We have completed the probability part of the problem, and written the correct integral. Nevertheless, let's show the calculus computation. The trick is to figure out the regions where we can write explicitly the minimum of x, y. First, assume $x \geq \frac{1-x}{2}$, which is equivalent to $x \geq \frac{1}{3}$. Then, $\min(x,y) = y$ for $y \in [0, \frac{1-x}{2}]$. Next, if $x < \frac{1}{3}$, then $\min(x,y) = x$ for $y \in [x, \frac{1-x}{2}]$, and $\min(x,y) = y$ for $y \in [0, x]$. We use this to rewrite the integral as:

$$\mathbb{E}[\min(X,Y)] = \int_0^{\frac{1}{3}} \Big(\int_0^{\frac{1-x}{2}} \min(x,y) \frac{4}{1-x}\, dy \Big)\, dx + \int_{\frac{1}{3}}^{\frac{1}{2}} \Big(\int_0^{\frac{1-x}{2}} \min(x,y) \frac{4}{1-x}\, dy \Big)\, dx$$

$$\int_{\frac{1}{3}}^{\frac{1}{2}} \Big(\int_0^{\frac{1-x}{2}} \min(x,y) \frac{4}{1-x}\, dy \Big)\, dx = \int_{\frac{1}{3}}^{\frac{1}{2}} \Big(\int_0^{\frac{1-x}{2}} y \frac{4}{1-x}\, dy \Big)\, dx$$

$$= \int_{\frac{1}{3}}^{\frac{1}{2}} \frac{4(1-x)^2)}{8(1-x)}\, dx = \int_{\frac{1}{3}}^{\frac{1}{2}} \frac{(1-x))}{2}\, dx = \frac{1}{9} - \frac{1}{16}$$

$$\int_0^{\frac{1}{3}} \Big(\int_0^{\frac{1-x}{2}} \min(x,y) \frac{4}{1-x}\, dy \Big)\, dx = \int_0^{\frac{1}{3}} \Big(\int_0^{x} y \frac{4}{1-x}\, dy + \int_x^{\frac{1-x}{2}} x \frac{4}{1-x}\, dy \Big)\, dx$$

$$= \int_0^{\frac{1}{3}} \Big(\frac{2x^2}{1-x} + \frac{2x(1-3x)}{1-x} \Big)\, dx$$

$$= \int_0^{\frac{1}{3}} \frac{2x - 4x^2}{1-x}\, dx \approx 0.078.$$

$$\mathbb{E}[\min(X,Y)] = 0.078 + \frac{1}{9} - \frac{1}{16} \approx 0.1266$$

Example 4.25
You have a stick of length 1. You pick two points along the stick, uniformly distributed, selected independently, You break the stick at the two points, resulting in three sticks. What is the expected length of the shortest stick?

The difference in this example from the previous example is that the points are selected independently, not sequentially. Let's propose a formulation using pairs of random variables. Let X be one of the points, and Y be the other. We know the joint PDF of X, Y, given by

$$f_{X,Y}(x,y) = \begin{cases} .1 & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

In terms of X, Y, what is the length of the shortest stick? Let $S(X,Y)$ be this length. If $X > Y$, then $S(X,Y) = \min(Y, X-Y, 1-X)$. Let B be the event that $X > Y$. By symmetry, $\mathbb{P}[B] = \frac{1}{2}$. Then, the conditional joint PDF of X, Y is given by

$$f_{X,Y|B}(x,y) = \begin{cases} \frac{f_{X,Y}(x,y)}{\mathbb{P}[B]} & (x,y) \in B, \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 2 & 0 \leq y \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Our answer is the conditional expected value of $S(X,Y)$ given the event B, because either X or Y has to be the smallest, so without loss of generality, we call X the smallest. Note that this introduces a factor of 2 to the conditional density, corresponding to mapping the original probability density from the unit square to the triangle $0 \leq y \leq x \leq 1$. Hence, our answer is

$$\mathbb{E}[S(X,Y)|X > Y] = \int_0^1 \Big(\int_0^x 2 \min(y, x-y, 1-x)\, dy \Big)\, dx$$

The rest is calculus...it does require breaking down the integral into regions where we can recognize which one of the terms is the minimum so we can do the integrals. A diagram will be most useful. We need to identify the regions in the triangle $0 \leq y \leq x \leq 1$ where $\min(y, x-y, 1-x) = y$, $\min(y, x-y, 1-x) = x-y$, and $\min(y, x-y, 1-x) = 1-x$ and compute the appropriate expected values in those regions.

The diagram is shown on the right. The three regions have a common point $(2/3, 1/3)$ where all three lengths are equal. Region 1 in the diagram is the region where the minimum is y, so $y < x-y, y < 1-x$. Hence, this region is $y < x/2, y < 1-x$. Region 2 is where the minimum is $x-y$, so $x-y < y, x-y < 1-x$ so $y > x/2, y > 2x-1$. Region 3 is where the minimum is $1-x$, so $1-x < y, 1-x < x-y$ and therefore $y > 1-x, y > 2x-1$. The answer we want is

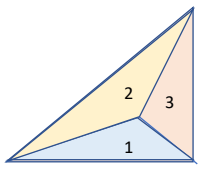


Figure 4.17: Example 4.25.

$$\mathbb{E}[S(X,Y)|X > Y] = \iint_{R_1} 2y\, dx\, dy + \iint_{R_2} 2(x-y)\, dx\, dy + \iint_{R_3} 2(1-x)\, dx\, dy$$

Computing each integral yields:

$$\iint_{R_1} 2y\,dx\,dy = \int_0^{\frac{1}{3}} \Big(\int_{2y}^{1-y} 2y\,dx \Big)\,dy = \int_0^{\frac{1}{3}} 2y(1-3y)\,dy = \frac{1}{9} - \frac{2}{27} = \frac{1}{27}$$

$$\begin{aligned}
\iint_{R_2} 2(x-y)\,dx\,dy &= \int_0^{\frac{2}{3}} \Big(\int_{\frac{x}{2}}^{x} 2(x-y)\,dy \Big)\,dx + \int_{\frac{2}{3}}^{1} \Big(\int_{2x-1}^{x} 2(x-y)\,dy \Big)\,dx \\
&= \int_0^{\frac{2}{3}} (x^2 - x^2 + \frac{x^2}{4})\,dx + \int_{\frac{2}{3}}^{1} (2x(1-x) - x^2 + (2x-1)^2)\,dx \\
&= \int_0^{\frac{2}{3}} \frac{x^2}{4}\,dx + \int_{\frac{2}{3}}^{1} (2x - 2x^2 - x^2 + 4x^2 - 4x + 1)\,dx \\
&= \frac{1}{12} \cdot \frac{8}{27} + \int_{\frac{2}{3}}^{1} (x^2 - 2x + 1)dx = \frac{2}{81} + \frac{1}{81} = \frac{1}{27}
\end{aligned}$$

$$\iint_{R_3} 2(1-x)\,dx\,dy = \int_{\frac{2}{3}}^{1} \Big(\int_{1-x}^{2x-1} (1-x)\,dy \Big)\,dx = \int_{\frac{2}{3}}^{1} (1-x)^2\,dx = \frac{1}{27}$$

Assembling the answer yields that the expected value of the shortest piece is $\frac{1}{9}$. Note that this is a little shorter than the answer to the previous example. The reason is that, in the previous problem, after we selected the first point, we broke the longer of the two pieces. Here, we select the second break randomly, so we can break the shorter of the two pieces, thereby resulting in shorter pieces. It is useful to check that your answers have common sense explanations.

Chapter 5

Second-Order Analysis of Random Vectors

5.1 Introduction

In Chapter 4, we developed a characterization of the properties of pairs of random variables X, Y defined on the same probability space $(\Omega, \mathcal{E}, \mathbb{P})$ by defining either a joint probability mass function (PMF) or a joint probability density function (PDF), which can be used to compute probabilities of joint events and expectations of functions of the random variables. Using the joint PMF or joint PDF, we computed statistics such as the expected value of a function $g(X, Y)$.

In this chapter, we focus on second order statistics of a pair of random variables X, Y. These statistics generalize the concepts of variance and standard deviation to pairs of random variables, and are easily computed from sample data. We describe how these statistics change for linear or affine transformations of the pair X, Y. We study the special case of jointly Gaussian random variables X, Y, where the joint PDF is entirely described in terms of its second order statistics, and show special properties of jointly Gaussian random variables that make them suitable models for problems in estimation and detection. We conclude the chapter with a generalization of second order statistics to random vectors involving 2 or more random variables.

5.2 Covariance and Correlation

Consider a pair of random variables X, Y defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$. If discrete, these random variables are characterized by a joint PMF $P_{X,Y}(x, y)$ and marginal PMFs $P_X(x), P_Y(y)$ derived from the joint PMF by

$$P_X(x) = \sum_{y \in R_Y} P_{X,Y}(x, y); \quad P_Y(y) = \sum_{x \in R_X} P_{X,Y}(x, y).$$

If X, Y are jointly continuous, the random variables are characterized by the joint PDF $f_{X,Y}(x, y)$, and marginal PDFs $f_X(x), f_Y(y)$ computed as:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy; \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx.$$

Using the marginal PMFs or PDFs, we can compute the means of X and Y, as $\mathbb{E}[X], \mathbb{E}[Y]$. We also compute the variance of each of the random variables X, Y as

$$\sigma_X^2 = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)^2\big] = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2,$$

$$\sigma_Y^2 = \mathbb{E}\big[\big(Y - \mathbb{E}[Y]\big)^2\big] = \mathbb{E}\big[Y^2\big] - \big(\mathbb{E}[Y]\big)^2.$$

These variances measure how much each of the random variables deviates from their average values. However, as statistics, they provide no information as to how the deviations of the random variables depend on each other.

To capture that information, we define several joint statistics for the random variables X, Y. First, we define the **cross-correlation** between X and Y as $\mathbb{E}[XY]$. An important property of the cross-correlation

is

$$\Big(\mathbb{E}[XY]\Big)^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

This follows from well-known Cauchy-Schwarz inequality, which states that, for functions $f(x), g(x)$ with finite square integrals,

$$\big|\int_{-\infty}^{\infty} f(x)g(x)\,dx\big| \leq \big(\int_{-\infty}^{\infty} f(x)^2\,dx\big)^{1/2}\big(\int_{-\infty}^{\infty} g(x)^2\,dx\big)^{1/2}.$$

Similarly, for square summable sequences x_n, y_n,

$$\big|\sum_{n=1}^{\infty} x_n y_n\big| \leq \big(\sum_{n=1}^{\infty} x_n^2\big)^{1/2}\big(\sum_{n=1}^{\infty} y_n^2\big)^{1/2}.$$

For continuous random variables, this implies

$$\begin{aligned}\Big|\mathbb{E}[XY]\Big| &= \Big|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,dx\,dy\Big| = \Big|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(x f_{X,Y}(x,y)^{\frac{1}{2}}\big)\big(y f_{X,Y}(x,y)^{\frac{1}{2}}\big)\,dx\,dy\Big| \\ &\leq \big(\int_{-\infty}^{\infty} x^2 f_{X,Y}(x,y)\,dx\,dy\big)^{1/2}\big(\int_{-\infty}^{\infty} y^2 f_{X,Y}(x,y)\,dx\,dy\big)^{1/2} = \big(\mathbb{E}[X^2]\big)^{1/2}\big(\mathbb{E}[Y^2]\big)^{1/2}\end{aligned}$$

The cross-correlation depends on the expected value of the individual random variables. To eliminate the dependence on the mean of the random variables, we define the **covariance** of random variables X and Y as

$$\mathsf{Cov}[X,Y] = \mathbb{E}\big[\big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big)\big].$$

Intuitively, this captures how X and Y vary together with respect to their expected values. Unlike variances, the covariance between two random variables can be negative. A negative covariance indicates that, when X is greater than its mean $\mathbb{E}[X]$, Y is likely to be less than its mean $\mathbb{E}[Y]$. The covariance will be an important part of how we can estimate the value of one variable (e.g. Y) based on measurements of the other variable (X).

Since X, Y are real-valued random variables, $\mathsf{Cov}[X,Y] = \mathsf{Cov}[Y,X]$. As is the case for variances, there is a useful formula for computing covariances from cross-correlations:

$$\begin{aligned}\mathsf{Cov}[X,Y] &= \mathbb{E}\big[\big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big)\big] = \mathbb{E}\big[XY - \mathbb{E}[X]Y - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]\big] \\ &= \mathbb{E}[XY] - \mathbb{E}\big[\mathbb{E}[X]Y\big] - \mathbb{E}\big[X\mathbb{E}[Y]\big] + \mathbb{E}\big[\mathbb{E}[X]\mathbb{E}[Y]\big] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Using the Cauchy-Schwarz inequality as before, we get the following:

$$|\mathsf{Cov}[X,Y]| \leq \sqrt{\mathsf{Var}[X]\mathsf{Var}[Y]}.$$

Using this inequality, we define the **correlation coefficient** $\rho_{X,Y}$ between two random variables X, Y as

$$\rho_{X,Y} = \frac{\mathsf{Cov}[X,Y]}{\sqrt{\mathsf{Var}[X]\mathsf{Var}[Y]}}$$

The correlation coefficient has magnitude less than or equal to 1, so its range is in $[-1,1]$.

Another way of interpreting the correlation coefficient is that it is the covariance of the normalized random variables $F = \frac{X}{\sqrt{\mathsf{Var}[X]}}$ and $G = \frac{Y}{\sqrt{\mathsf{Var}[Y]}}$. Normalizing each of the random variables by dividing by their standard deviation results in random variables F and G with variance 1. This normalization is used extensively in data science and statistics to reduce the effects of measurement units for feature values.

Example 5.1
Let X, Y be a pair of random variables, and define $Z = X + Y$. Then,

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}[X] + \mathbb{E}[Y] \qquad \text{(linearity of expectation)} \\
\mathbb{E}[Z^2] &= \mathbb{E}\big[(X+Y)^2\big] = \mathbb{E}[X^2 + 2XY + Y^2] = \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] \\
&= \mathbb{E}[X]^2 + \mathsf{Var}[X] + 2\Big(\mathbb{E}[X]\mathbb{E}[Y] + \mathsf{Cov}[X,Y]\Big) + \mathbb{E}[Y]^2 + \mathsf{Var}[Y] \quad \text{(definitions of variance, covariance)} \\
&= \Big(\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2\Big) + \mathsf{Var}[X] + 2\mathsf{Cov}[X,Y] + \mathsf{Var}[Y] \\
&= \mathbb{E}[Z]^2 + \mathsf{Var}[X] + 2\mathsf{Cov}[X,Y] + \mathsf{Var}[Y] \\
\mathsf{Var}[Z] &= \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \mathsf{Var}[X] + 2\mathsf{Cov}[X,Y] + \mathsf{Var}[Y]
\end{aligned}$$

This provides a quick way of calculating the covariance of a sum of random variables. The result does not depend on the mean of the random variables.

Example 5.2
Can the correlation coefficient have magnitude 1? Let X be a random variable, and let $Y = -3X + 1$. Then,

$$\begin{aligned}
\mathsf{Var}[Y] &= (-3)^2\mathsf{Var}[X] = 9\mathsf{Var}[X] \\
\mathsf{Cov}[X,Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[-3X^2 + X] - \mathbb{E}[X]\mathbb{E}[-3X+1] = -3\big(\mathbb{E}[X^2] - \mathbb{E}[X]^2\big) = -3\mathsf{Var}[X] \\
\rho_{X,Y} &= \frac{\mathsf{Cov}[X,Y]}{\sqrt{\mathsf{Var}[X]\mathsf{Var}[Y]}} = \frac{-3\mathsf{Var}X}{\sqrt{9\mathsf{Var}[X]^2}} = -1
\end{aligned}$$

When the magnitude of the correlation coefficient is either 1 or -1, it usually indicates a linear dependence between the two variables X, Y. Notice that, in this case, the correlation coefficient has a negative sign, suggesting a negative linear dependence.

Note also that the correlation coefficient is a scale-independent measure of how the random variables depend on each other. Thus, the scale factor of -3 between X and Y only affects the correlation coefficient by its sign, not its magnitude.

Example 5.3
Consider a pair of jointly continuous random variables X, Y with $f_{X,Y}(x,y)$ given as

$$f_{X,Y}(x,y) = \begin{cases} xy & 0 \le x \le 1, 0 \le y \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

The marginal distributions are given as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \begin{cases} \int_0^2 xy\,dy = 2x & x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \begin{cases} \int_0^1 xy\,dx = \frac{y}{2} & y \in [0,2] \\ 0 & \text{otherwise.} \end{cases}$$

Note that X, Y are independent, as the range $R_{X,Y} = R_X \times R_Y$ and thus $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Using these densities, we compute the first and second order statistics as follows:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 2x^2\,dx = \frac{2}{3} \\
\mathbb{E}[Y] &= \int_{-\infty}^{\infty} y f_Y(y)\,dx = \int_0^2 \frac{y^2}{2}\,dy = \frac{4}{3} \\
\mathsf{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_0^1 2x^3\,dx - \frac{4}{9} = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} \\
\mathsf{Var}[Y] &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \int_0^2 \frac{y^3}{2}\,dy - \frac{16}{9} = 2 - \frac{16}{9} = \frac{2}{9} \\
\mathbb{E}[XY] &= \mathbb{E}[X]\mathbb{E}[Y] = \frac{8}{9} \quad \text{(because of independence.)} \\
\mathsf{Cov}[X,Y] &= 0, \quad \rho_{X,Y} = 0
\end{aligned}$$

When two random variables are independent, their covariance is 0. The converse is not true, as we will see later.

Two random variables X and Y are **uncorrelated** if $\mathsf{Cov}[X,Y]=0$ (or $\rho_{X,Y}=0$).

- If X and Y are uncorrelated, we have that $\mathsf{Var}[X+Y]=\mathsf{Var}[X]+\mathsf{Var}[Y]$ and $\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$.
- Independence of X and Y implies that they are uncorrelated. However, uncorrelated X and Y need not be independent.

To clarify, if X,Y are independent, then, for bounded functions f,g, we have

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)g(y)f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)g(y)f_X(x)f_Y(y)\,dx\,dy \text{ (independence of PDF)} \\ &= \Big(\int_{-\infty}^{\infty} f(x)f_X(x)\,dx\Big)\Big(\int_{-\infty}^{\infty} g(y)f_Y(y)\,dy\Big) = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]\end{aligned}$$

The converse of this is also true: if $\mathbb{E}[f(X)g(Y)]=\mathbb{E}[f(X)]\mathbb{E}[g(Y)]$ for any bounded functions f,g, then X and Y are independent. However, X,Y are uncorrelated if and only if $\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$. Thus, the condition for uncorrelated random variables involves only linear functions of X,Y, whereas the condition for independence must hold for the broader class of bounded nonlinear functions of X, Y.

Furthermore, if X,Y are independent, then $f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\frac{f_X(x)f_Y(y)}{f_Y(y)}=f_X(x)$. Hence, $\mathbb{E}[X|Y=y]=\mathbb{E}[X]$ for all $y\in R_Y$. Similarly, $\mathbb{E}[Y|X=x]=\mathbb{E}[Y]$ for all $x\in R_X$. Independence is a *strong* property of the underlying densities of the random variables, while uncorrelatedness is only a property of second order statistics.

One of the interesting properties of uncorrelated random variables X,Y is that, if $Z=X+Y$, then $\mathsf{Var}[Z]=\mathsf{Var}[X]+\mathsf{Var}[Y]$. This is because, as derived in Example 5.1,

$$\mathsf{Var}[Z]=\mathsf{Var}[X]+\mathsf{Var}[Y]+2\mathsf{Cov}[X,Y]=\mathsf{Var}[X]+\mathsf{Var}[Y],$$

since $\mathsf{Cov}[X,Y]=0$ because X,Y are uncorrelated. This generalizes to arbitrary sums, so that the variance of a sum of uncorrelated random variables is the sum of the variances of the individual random variables.

Two random variables X and Y are **orthogonal** if and only if $\mathbb{E}[XY]=0$. If X and Y are orthogonal, $\mathbb{E}[(X+Y)^2]=\mathbb{E}[X^2]+\mathbb{E}[Y^2]$. Note that orthogonal and uncorrelated random variables are different concepts. If two random variables are both orthogonal and uncorrelated, then the mean of at least one must be zero. For zero mean random variables, orthogonality and uncorrelatedness are equivalent. For instance, the random variables X,Y in Example 5.3 are independent, and thus uncorrelated. However, they are not orthogonal, because neither X nor Y has zero mean.

Example 5.4

Consider a pair of discrete random variables X,Y with joint PMF given by the table on the right. Are X,Y independent? Are X,Y uncorrelated? What is the covariance of X,Y?

$P_{XY}(x,y)$		y = 0	y = 1	y = 2
x	0	0.01	0.00	0.00
	1	0.09	0.09	0.00
	2	0.00	0.00	0.81

With respect to independence, the answer is clearly not. Note that $P_{X,Y}(0,1)=0$, but $P_X(0)=0.01$ and $P_Y(1)=0.09$.

Are X,Y uncorrelated? We compute $\mathbb{E}[X]=0.18+2\cdot 0.81=1.80$, and $\mathbb{E}[Y]=0.09+2\cdot 0.81=1.71$. We then compute

$$\mathbb{E}[XY]=0.09\cdot 1\cdot 1+0.81\cdot 2\cdot 2=3.33\neq(1.71)\cdot(1.89)$$

Hence, they are not uncorrelated.

The covariance $\mathsf{Cov}[X,Y]=3.33-(1.71)\cdot(1.89)\approx 0.252$.

Example 5.5
Consider a pair of discrete random variables X, Y with joint PMF given by the table on the right. What are the means and variances of X, Y? Are X, Y independent?
We compute the marginal PMFs by doing column and row sums to get

$P_{XY}(x,y)$		y			
		0	1	2	3
x	0	0.06	0.18	0.24	0.12
	1	0.04	0.12	0.16	0.08

$$P_X(0) = 0.6, P_X(1) = 0.4.$$

$$P_Y(0) = 0.1, P_Y(1) = 0.3, P_Y(2) = 0.4, P_Y(3) = 0.2.$$

With this, we compute $\mathbb{E}[X] = 0.6 \cdot 0 + 0.4 \cdot 1 = 0.4$; similarly, $\mathbb{E}[X^2] = 0.6 \cdot 0^2 + 0.4 \cdot 1^2 = 0.4$. Thus, $\mathsf{Var}[X] = 0.4 - (0.4)^2 = 0.24$.

For Y, $\mathbb{E}[Y] = 0 \cdot 0.1 + 1 \cdot 0.3 + 2 \cdot 0.4 + 3 \cdot 0.2 = 1.7$. Similarly, $\mathbb{E}[Y^2] = 0^2 \cdot 0.1 + 1^2 \cdot 0.3 + 2^2 \cdot 0.4 + 3^2 \cdot 0.2 = 4.1$l. Hence, $\mathsf{Var}[Y] = 4.10 - (1.7)^2 = 4.10 - 2.89 = 1.21$.

With respect to independence, note that there are no zeros in the table, so $R_{X,Y} = R_X \times R_Y$. We now have to check that $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ for all $(x,y) \in R_{X,Y}$. We quickly verify that this is indeed the case, so X, Y are independent. Therefore, $\mathsf{Cov}[X,Y] = 0$.

Example 5.6
Consider a pair of continuous random variables X, Y, uniformly distributed on the unit disk with radius 1, centered at (0,0). Thus, the joint PDF of X, Y is given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & 0 \le x^2 + y^2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The joint PDF of X, Y is illustrated in Figure 5.1. We saw this example in the previous chapter, as Example 4.11. Are X, Y independent? Are X, Y uncorrelated? What are the means, variances and covariances of X, Y?

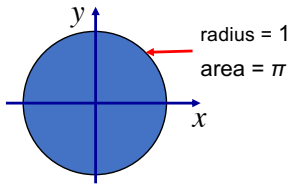


Figure 5.1: Example 5.6.

With respect to independence, consider the point $(x,y) = (0.9, 0.9)$. This point is outside the unit circle, so $f_{X,Y}(0.9, 0.9) = 0$. However, it is clear that a vertical line through that point intersects the unit circle, and so does a horizontal line. This means that $f_X(0.9) > 0, f_Y(0.9) > 0$, and therefore, $f_{X,Y}(0.9, 0.9) = 0 \neq f_X(0.9)f_Y(0.9)$. Hence, X, Y are not independent.

By symmetry, we note that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. We can also verify these using the results of Example 4.11, where we showed that $f_X(x) = \frac{2\sqrt{1-x^2}}{\pi}$, $f_Y(y) = \frac{2\sqrt{1-y^2}}{\pi}$. Both of these functions are even functions, so $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. By symmetry, we can also show that $E[XY] = 0$. We will show that directly by computation:

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,dx\,dy = \int_{-1}^{1}\Big(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} y\,dy\Big)\frac{x}{\pi}\,dx$$

The inner integral evaluates as

$$\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} y\,dy = \frac{y^2}{2}\Big|_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} = 0.$$

Thus, $\mathbb{E}[XY] = 0$, and $\mathsf{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$, so X and Y are uncorrelated. In this case, X, Y are also orthogonal.

It is also clear that $\mathsf{Var}[X] = \mathsf{Var}[Y]$ by symmetry. To compute $\mathsf{Var}[X]$, since X has zero mean, we get

$$\mathsf{Var}[X] = \mathbb{E}[X^2] = \int_{-1}^{1} x^2 \frac{2\sqrt{1-x^2}}{\pi}\,dx = \frac{1}{4} = \mathsf{Var}[Y],$$

where the integral can be evaluated using a trigonometric substitution $x = \sin(\theta)$.

5.3 Algebra of Covariances

Assume we have two random variables X, Y, for which we know their means $\mathbb{E}[X], \mathbb{E}[Y]$, their variances $\mathsf{Var}[X], \mathsf{Var}[Y]$, and their covariance $\mathsf{Cov}[X, Y]$. Define new random variables, linearly related to these, as

$$U = aX + bY + e; \qquad V = cX + dY + f$$

We want to compute the means and variances of U, V and their covariance. To answer this, we exploit the properties of the linearity of the expectation operation, as

$$\mathbb{E}[U] = \mathbb{E}[aX + bY + e] = a\mathbb{E}[X] + b\mathbb{E}[Y] + e\mathbb{E}[1] = a\mathbb{E}[X] + b\mathbb{E}[Y] + e.$$
$$\mathbb{E}[V] = \mathbb{E}[cX + dY + f] = c\mathbb{E}[X] + d\mathbb{E}[Y] + f.$$

What about the variance of U? Since variance is a quadratic statistic, we have to expand a quadratic to compute this. Suppose we compute this as $\mathsf{Var}[U] = \mathbb{E}[U^2] - (\mathbb{E}[U])^2$. Then,

$$\begin{aligned}
\mathbb{E}[U^2] &= \mathbb{E}[(aX + bY + e)^2] = \mathbb{E}[a^2X^2 + 2abXY + b^2Y^2 + 2aeX + 2beY + e^2] \\
&= a^2\ \mathsf{Var}[X] + a^2(\mathbb{E}[X])^2 + 2ab\ \mathsf{Cov}[X, Y] + 2ab\ \mathbb{E}[X]\mathbb{E}[Y] + b^2\ \mathsf{Var}[Y] + b^2(\mathbb{E}[Y])^2 \\
&\quad + 2ae\ \mathbb{E}[X] + 2be\ \mathbb{E}[Y] + e^2 \\
&= \Big(a^2\ \mathsf{Var}[X] + 2ab\ \mathsf{Cov}[X, Y] + b^2\ \mathsf{Var}[Y]\Big) + \Big(a^2(\mathbb{E}[X])^2 + 2ab\ \mathbb{E}[X]\mathbb{E}[Y] + b^2(\mathbb{E}[Y])^2 \\
&\quad + 2ae\ \mathbb{E}[X] + 2be\ \mathbb{E}[Y] + e^2\Big) \\
&= \Big(a^2\ \mathsf{Var}[X] + 2ab\ \mathsf{Cov}[X, Y] + b^2\ \mathsf{Var}[Y]\Big) + (\mathbb{E}[U])^2 \\
\mathsf{Var}[U] &= E[U^2) - (\mathbb{E}[U])^2 = a^2\ \mathsf{Var}[X] + 2ab\ \mathsf{Cov}[X, Y] + b^2\ \mathsf{Var}[Y]
\end{aligned}$$

However, we know that variances do not depend on the mean of the variables. That is, $\mathsf{Var}[U] = \mathsf{Var}\Big[U - \mathbb{E}[U]\Big]$. Indeed, we should have been able to compute the variance of U by assuming all the variables had zero mean. This leads to a much simple computation, as

$$\begin{aligned}
\mathsf{Var}[U] = \mathsf{Var}[U - a\mathbb{E}[X] - b\mathbb{E}[Y] - e] &= \mathbb{E}\Big[(a\tilde{X} + b\tilde{Y})^2\Big] \\
&= \mathbb{E}[a^2(\tilde{X})^2] + 2\mathbb{E}[ab\tilde{X}\tilde{Y}] + \mathbb{E}[b^2(\tilde{Y})^2] \\
&= a^2\ \mathsf{Var}[X] + 2ab\ \mathsf{Cov}[X, Y] + b^2\ \mathsf{Var}[Y].
\end{aligned}$$

where $\tilde{X} = X - \mathbb{E}[X], \tilde{Y} = Y - \mathbb{E}[Y]$, and thus $\mathsf{Var}[X] = \mathbb{E}[(\tilde{X})^2], \mathsf{Var}[Y] = \mathbb{E}[(\tilde{Y})^2]$, and $\mathsf{Cov}[X, Y] = \mathbb{E}[\tilde{X}\tilde{Y}]$. By considering only the zero-mean random variables, we are able to get to a simpler formula for variances without having to consider the extra terms associated with the means. This avoids unnecessary algebraic errors that arise when including all the terms involving the means of the random variables.

Similarly, we compute the variance of V as

$$\mathsf{Var}[V] = \mathsf{Var}\Big[V - \mathbb{E}[V]\Big] = \mathbb{E}\Big[(c\tilde{X} + d\tilde{Y})^2\Big] = c^2\ \mathsf{Var}[X] + 2cd\ \mathsf{Cov}[X, Y] + d^2\ \mathsf{Var}[Y].$$

Furthermore, the covariance of U, V is given by

$$\begin{aligned}
\mathsf{Cov}[U, V] = \mathsf{Cov}\Big[U - \mathbb{E}[U], V - \mathbb{E}[V]\Big] &= \mathbb{E}\Big[(a\tilde{X} + b\tilde{Y})(c\tilde{X} + d\tilde{Y})\Big] \\
&= ac\ \mathsf{Var}[X] + (ad + bc)\ \mathsf{Cov}[X, Y] + bd\ \mathsf{Var}[Y].
\end{aligned}$$

Example 5.7
Consider X, Y as defined in Example 5.6. We know that $\mathbb{E}[X] = \mathbb{E}[Y] = 0, \mathsf{Var}[X] = \mathsf{Var}[Y] = \frac{1}{4}, \mathsf{Cov}[X, Y] = 0$. Thus, X, Y are uncorrelated and orthogonal.

Let $U = 3X + 2Y + 1, V = 2X - 3Y - 1$. Compute the means, variances and covariance of U, V.

The means are easy: Using linearity of expectation, we get

$$\mathbb{E}[U] = 3\mathbb{E}[X] + 2\mathbb{E}[Y] + 1 = 1; \qquad \mathbb{E}[V] = 2\mathbb{E}[X] - 3\mathbb{E}[Y] - 1 = -1.$$

For variances, using the approach that we deal only with zero-mean variables, we get

$$\mathsf{Var}[U] = 9\mathsf{Var}[X] + 12\mathsf{Cov}[X,Y] + 4\mathsf{Var}[Y] = 13\mathsf{Var}[X] = \frac{13}{4}.$$

$$\mathsf{Var}[V] = 4\mathsf{Var}[X] - 12\mathsf{Cov}[X,Y] + 9\mathsf{Var}[Y] = 13\mathsf{Var}[X] = \frac{13}{4}.$$

$$\mathsf{Cov}[U,V] = 6\mathsf{Var}[X] - 9\mathsf{Cov}[X,Y] + 4\mathsf{Cov}[X,Y] - 6\mathsf{Var}[Y] = 0.$$

Our transformations resulted in U, V that are also uncorrelated, but no longer orthogonal, because neither has zero-mean. Why? If we write the transformation as a matrix:

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

you will notice that the first and second rows of the transformation matrix for X, Y are perpendicular vectors. We will explore this further when we discuss random vectors.

5.4 Jointly Gaussian Random Variables:

There is a class of jointly continuous random variables whose joint PDF is entirely specified by its second order statistics. Recall that Gaussian random variables had PDFs specified entirely in terms of their means and variances. In this section, we define the concept of pairs of jointly Gaussian random variables, where the joint PDFs are specified entirely by first- and second-order statistics, and explore their properties.

We begin by constructing a pair of independent, standard Gaussian random variables. Let U, V be standard Gaussian random variables defined on the same probability space. That is, $U \sim N(0,1), V \sim N(0,1)$ both have zero mean and unit variance. To merge them into joint random variables, we assume that U, V are independent, resulting in a pair of **independent unit Gaussian random variables**. In this case, the joint PDF is

$$f_{U,V}(u,v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} = \frac{1}{2\pi} e^{-\frac{u^2+v^2}{2}}.$$

The joint probability density of a pair of unit Gaussian random variables is shown in Figure 5.2. The density is centered at (0,0), and has a circular symmetry, decaying to 0 as $u^2 + v^2$ approaches infinity. Consider now a pair of random variables X, Y defined in terms of U, V as

$$X = \sigma_X U + \mu_X; \qquad Y = \sigma_Y U + \mu_Y$$

where $\sigma_X, \sigma_Y > 0$, and μ_X, μ_Y are constants. Since X depends only on U and Y depends only on V, X and Y are also independent random variables.

Note that $\mathbb{E}[X] = \sigma_X \mathbb{E}[U] + \mu_X = \mu_X$, $\mathsf{Var}[X] = \sigma_X^2 \mathsf{Var}[U] = \sigma_X^2$. Similarly, $\mathbb{E}[Y] = \mu_Y, \mathsf{Var}[Y] = \sigma_Y^2$. Since X is a linear transformation of the variable U, we can obtain the density of X using the methods of Chapter 3 as

$$f_X(x) = \frac{1}{|\sigma_X|} f_U\left(\frac{X - \mu_X}{\sigma_X}\right) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(X-\mu_X)^2}{2\sigma_X^2}}.$$

which also follows because a linear transformation of a Gaussian random variable results in another Gaussian random variable. Similarly,

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(Y-\mu_Y)^2}{2\sigma_Y^2}}.$$

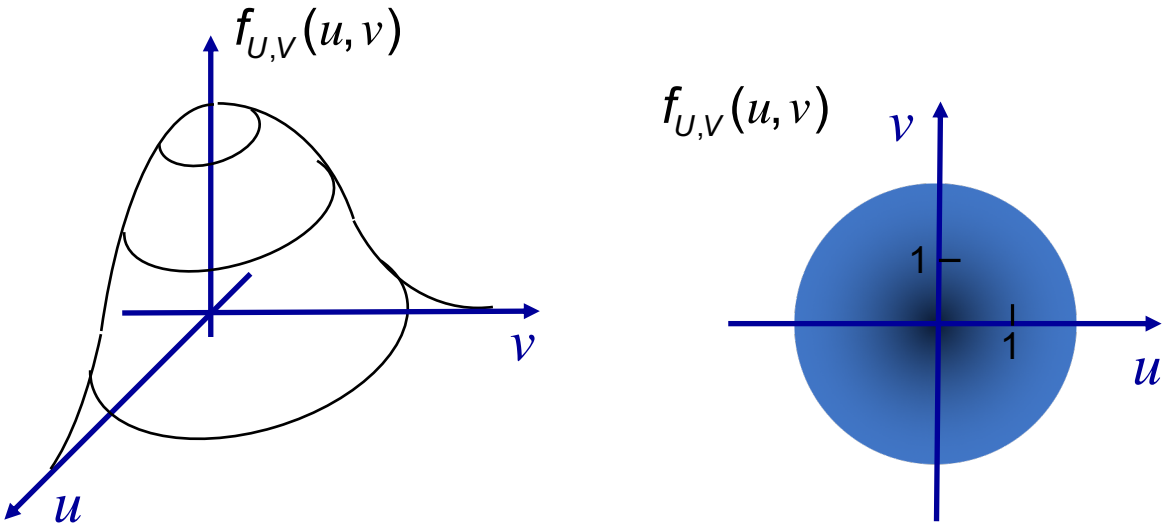


Figure 5.2: Illustration of the density of a pair of independent unit Gaussian random variables.

and, because X, Y are independent, their joint PDF is given by

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(X-\mu_X)^2}{2\sigma_X^2}} \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(Y-\mu_Y)^2}{2\sigma_Y^2}} = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{(X-\mu_X)^2}{2\sigma_X^2} + \frac{(Y-\mu_Y)^2}{2\sigma_Y^2}}$$

An illustration of the joint PDF of X, Y is shown in the figure on the right. Note that the level sets of the probability density function (curves where $f_{X,Y}(x,y) = K$ for some constant K) are now ellipses, and the center of the PDF has shifted to the mean (μ_X, μ_Y). The individual standard deviations are measures of the relative elongation of the ellipses along each axis. The major axes of the ellipses are aligned with the x and y axes, because X and Y are still independent.

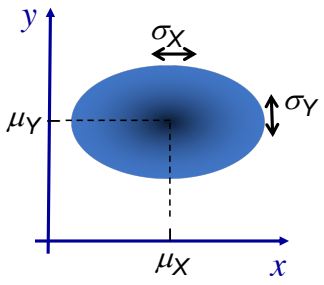


Figure 5.3: Gaussian PDF with unequal variances.

Consider now $Z = X + Y$ to be the sum of two independent jointly Gaussian random variables. We want to show that this is also a Gaussian random variable. If we know this, then the PDF of Z can be computed trivially by knowing $\mathbb{E}[Z] = \mu_X + \mu_Y$, and $\mathsf{Var}[Z] = \sigma_X^2 + \sigma_Y^2$, since the variances add when X, Y are uncorrelated and hence independent. We show this for the case where the means $\mu_X = \mu_Y = 0$, as we can always add a constant to shift the means. We refer to Section 4.7.1 for determining the density of a sum

of two jointly continuous random variables, as

$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\,dx = C\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma_X^2}-\frac{(z-x)^2}{2\sigma_Y^2}}\,dx \\
&= Ce^{-\frac{z^2}{2\sigma_Y^2}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}(\frac{1}{\sigma_X^2}+\frac{1}{\sigma_Y^2})+\frac{xz}{\sigma_Y^2}}\,dx \\
&= Ce^{-\frac{z^2}{2\sigma_Y^2}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}(\frac{1}{\sigma_X^2}+\frac{1}{\sigma_Y^2})+\frac{xz}{\sigma_Y^2}-\frac{z^2}{2}\frac{\sigma_X^2}{\sigma_Y^2(\sigma_X^2+\sigma_Y^2)}+\frac{z^2}{2}\frac{\sigma_X^2}{\sigma_Y^2(\sigma_X^2+\sigma_Y^2)}}\,dx \quad \text{(add and subtract same term)} \\
&= Ce^{-\frac{z^2}{2\sigma_Y^2}+\frac{z^2}{2}\frac{\sigma_X^2}{\sigma_Y^2(\sigma_X^2+\sigma_Y^2)}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}(\frac{1}{\sigma_X^2}+\frac{1}{\sigma_Y}{}^2)+\frac{xz}{\sigma_Y^2}-\frac{z^2}{2}\frac{\sigma_X^2}{\sigma_Y^2(\sigma_X^2+\sigma_Y^2)}}\,dx \\
&= Ce^{-\frac{z^2}{2(\sigma_Y^2+\sigma_X^2)}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{1}{\sigma_X^2}+\frac{1}{\sigma_Y^2}\right)\left(x^2-\frac{\sigma_X^2\sigma_Y^2}{\sigma_Y^2+\sigma_X^2}\frac{2xz}{\sigma_Y^2}+z^2\frac{\sigma_X^2}{\sigma_Y^2(\sigma_X^2+\sigma_Y^2)}\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)}\,dx \\
&= Ce^{-\frac{z^2}{2(\sigma_Y^2+\sigma_X^2)}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{1}{\sigma_X^2}+\frac{1}{\sigma_Y^2}\right)\left(x-\frac{\sigma_X^2}{\sigma_X^2+\sigma_Y^2}\right)^2}\,dx \\
&= C_1 e^{-\frac{z^2}{2(\sigma_Y^2+\sigma_X^2)}}
\end{aligned}$$

where the constant is chosen C_1 to satisfy the normalization property $\int_{-\infty}^{\infty} f_Z(z)\,dz = 1$. The result shows that Z is a Gaussian random variable with zero mean, and variance $\sigma_X^2+\sigma_Y^2$.

Using the above argument, we can show that a random variable $X = aU+bV+\mu_X$ will be Gaussian, with mean $\mathbb{E}[X] = \mathbb{E}[aU+bV+\mu_X] = \mu_X$, and variance $\mathsf{Var}[X] = \mathsf{Var}[aU+bV] = a^2\mathsf{Var}[U]+b^2\mathsf{Var}[V] = a^2+b^2$. Similarly, a random variable $Y = cU+dV+\mu_Y$ will be Gaussian, with mean $\mathbb{E}[Y] = \mu_Y$, and variance $\mathsf{Var}[Y] = c^2+d^2$.

We formally define jointly Gaussian random variables as follows: A pair of random variables X and Y are **jointly Gaussian random variables** if they are linear functions of independent unit Gaussian random variables U and V:

$$X = aU+bV+\mu_X \qquad Y = cU+dV+\mu_Y\ .$$

We now compute the covariance of X, Y as

$$\mathsf{Cov}[X,Y] = \mathbb{E}[(aU+bV)(cU+dV])] = ac\mathbb{E}[U^2]+(ad+bc)\mathbb{E}[UV]+bd\mathbb{E}[V^2] = ac+bd,$$

since U, V are zero-mean, independent, unit variance random variables. The resulting correlation coefficient is

$$\rho_{X,Y} = \frac{ac+bd}{\sqrt{(a^2+b^2)(c^2+d^2)}}.$$

When the correlation coefficient of X, Y has magnitude less than 1, we can write the joint PDF of X, Y as

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}}e^{-\frac{1}{2(1-\rho_{X,Y}^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}$$

Thus, the joint PDF is fully specified by the first- and second-order statistics: the means μ_X, μ_Y, the variances σ_X^2, σ_Y^2, and the correlation coefficient $\rho_{X,Y}$.

This is a difficult formula to remember, and it does not generalize to more than two Gaussian random variables. However, we can write this in terms of vectors and matrices as follows:

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{\det\left(2\pi\begin{bmatrix}\sigma_X^2 & \mathsf{Cov}[X,Y]\\ \mathsf{Cov}[X,Y] & \sigma_Y^2\end{bmatrix}\right)}}e^{-\frac{1}{2}\begin{bmatrix}x-\mu_x & y-\mu_y\end{bmatrix}\begin{bmatrix}\sigma_X^2 & \mathsf{Cov}[X,Y]\\ \mathsf{Cov}[X,Y] & \sigma_Y^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_x\\ y-\mu_y\end{bmatrix}}.$$

This form that uses the inverse of a matrix formed from the individual covariances generalizes well to three or more Gaussian random variables.

An illustration of the joint PDF of X, Y in this general case is shown in the figure on the right. Note that the level sets of the probability density function (curves where $f_{X,Y}(x,y) = K$ for some constant K) are still ellipses, and the center of the PDF has shifted to the mean (μ_X, μ_Y). The individual standard deviations are measures of the relative elongation of the ellipses along each axis. However, note that the major axes of the ellipses are no longer aligned with the x and y axes, because X and Y are now correlated and not independent. That is seen in the joint PDF by the presence of xy terms in the exponent of the density. Note that, if $\rho_{X,Y} = 0$, these terms vanish.

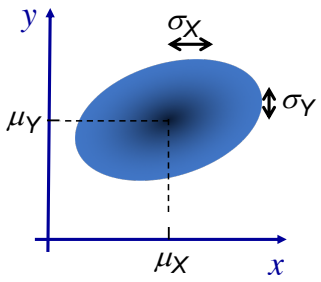


Figure 5.4: Correlated Gaussian PDF.

Jointly Gaussian random variables satisfy the following properties:

- **Any linear function of X and Y plus a constant is Gaussian:** If $Z = \alpha X + \beta Y + \gamma$, then Z is Gaussian with $\mathbb{E}[Z] = \mu_Z, \mathsf{Var}[Z] = \sigma_Z^2$ where

$$\mu_Z = \alpha\mu_X + \beta\mu_Y + \gamma, \qquad \sigma_Z^2 = \alpha^2\sigma_X^2 + \beta^2\sigma_Y^2 + 2\alpha\beta\mathsf{Cov}[X,Y]$$

  The reason for this is that, since X, Y are linear combinations of independent, unit Gaussian random variables U, V plus a constant, we can substitute for X, Y and write Z as a linear combination of U, V plus a constant. We have already shown this is a Gaussian random variable.

- **Marginal PDFs are Gaussian:** X is Gaussian with $\mathbb{E}[X] = \mu_X, \mathsf{Var}[X] = \sigma_X^2$ and Y is Gaussian with $\mathbb{E}[Y] = \mu_Y, \mathsf{Var}[Y] = \sigma_Y^2$.

  The function $Z = 1 \cdot X + 0 \cdot Y$ is a linear combination, and hence it is Gaussian. We know its mean and variance by computation as above.

- **Uncorrelated $\implies$ Independence:** X and Y are uncorrelated ($\mathsf{Cov}[X,Y] = 0$ or $\rho_{X,Y} = 0$) if and only if X and Y are independent.

  This follows by examining the form of the joint density function described above. If $\rho_{X,Y} = 0$, then we can separate $F_{X,Y} = f_X(x)f_Y(y)$. In general, uncorrelated random variables are not independent. However, for jointly Gaussian random variables, uncorrelated Gaussian random variables are independent. This means we can verify independence strictly using second-order statistics.

- $|\rho_{X,Y}| = 1$ if and only if Y is a deterministic linear function of X (and vice versa). In this case, we can write Y as $Y = \rho_{X,Y}\dfrac{\sigma_Y}{\sigma_X}(X - \mu_X) + \mu_Y$.

- **Conditional PDF of X given $Y = y$ is Gaussian:** The conditional PDF $f_{X|Y}(x|y)$ of X given $Y = y$ is Gaussian with mean $\mathbb{E}[X|Y=y]$ and variance $\mathsf{Var}[X|Y=y]$ to be computed as:

$$\mathbb{E}[X|Y=y] = \mu_X + \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y - \mu_Y) = \mu_X + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(y - \mu_Y)$$

$$\mathsf{Var}[X|Y=y] = (1 - \rho_{X,Y}^2)\sigma_X^2 = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}$$

Let's derive this result. We know the following:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} e^{-\frac{1}{2(1-\rho_{X,Y}^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} e^{Q(x-\mu_X, y-\mu_Y)}, \quad \text{where}$$

$$Q(x,y) = -\frac{1}{2(1-\rho_{X,Y}^2)}\left((\frac{x}{\sigma_X})^2 - 2\rho_{X,Y}(\frac{x}{\sigma_X})(\frac{y}{\sigma_Y}) + (\frac{y}{\sigma_Y})^2\right)$$

$$= -\frac{1}{2(1-\rho_{X,Y}^2)}\left((\frac{x}{\sigma_X})^2 - 2\rho_{X,Y}(\frac{x}{\sigma_X})(\frac{y}{\sigma_Y}) + (\rho_{X,Y}\frac{y}{\sigma_Y})^2 + (1-\rho_{X,Y}^2)(\frac{y}{\sigma_Y})^2\right)$$

$$= -\frac{1}{2(1-\rho_{X,Y}^2)\sigma_X^2}\left(x - \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}y\right)^2 - \frac{y^2}{2\sigma_Y^2}. \quad \text{Hence,}$$

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} e^{-\frac{1}{2(1-\rho_{X,Y}^2)\sigma_X^2}\left((x-\mu_X) - \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)^2 - \frac{(y-\mu_Y)^2}{2\sigma_Y^2}}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}$$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}}(\sqrt{2\pi}\sigma_Y) e^{-\frac{1}{2(1-\rho_{X,Y}^2)\sigma_X^2}\left((x-\mu_X) - \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)^2}$$

$$= \frac{1}{\sigma_X\sqrt{2\pi(1-\rho_{X,Y}^2)}} e^{-\frac{1}{2(1-\rho_{X,Y}^2)\sigma_X^2}\left((x-\mu_X) - \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)^2}$$

We recognize the above expression as a Gaussian density, with statistics

$$\mathbb{E}[X|Y=y] = \mu_X + \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y) = \mu_X + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(y-\mu_Y)$$

$$\mathsf{Var}[X|Y=y] = (1-\rho_{X,Y}^2)\sigma_X^2 = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}$$

Notice that the conditional covariance does not depend on the actual observed value $Y = y$; it only depends on the second order statistics of X, Y. Notice also that the conditional covariance $\mathsf{Var}[X|Y=y]$ is no larger than the unconditional covariance $\mathsf{Var}[X]$, as we are subtracting a nonnegative term.

The above formulas for the conditional mean and variance are very important in estimation, as we will illustrate in a subsequent chapter. Specifically, $\mathbb{E}[X|Y=y]$ is an estimate of the random variable X based on measuring that the random variable Y has value y. Define

$$e = X - \mathbb{E}[X|Y] = X - \mu_X - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y-\mu_Y).$$

Then, this is the error in the estimate of X given observation Y. In this case, e is a linear function of X and Y plus a constant.

Note some important properties of the estimation error:

- $\mathbb{E}[e(y)] = 0$. This follows directly by noting that $e = \tilde{X} - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\tilde{Y}$, and thus it is a linear combination of zero-mean random variables.

- $\mathbb{E}[e^2] = \mathsf{Var}[X|Y=y]$. Note that $\mathsf{Var}[X|Y=y]$ is a constant that does not depend on y. This follows because

$$\begin{aligned}\mathbb{E}[e^2] &= \mathbb{E}[\tilde{X}^2] - 2\frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}[\tilde{X}\tilde{Y}] + (\frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]})^2\mathbb{E}[\tilde{Y}^2] \\ &= \mathsf{Var}[X] - 2\frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathsf{Cov}[X,Y] + (\frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]})^2\mathsf{Var}[Y] \\ &= \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}\end{aligned}$$

  ß A more subtle proof of the above uses iterated expectations, as

$$\mathbb{E}[e^2] = \mathbb{E}\big[\mathbb{E}[e^2|Y]\big] = \mathbb{E}\big[\mathsf{Var}[X|Y]\big] = \mathsf{Var}[X|Y],$$

  which follows because $\mathsf{Var}[X|Y]$ is a constant that does not depend on Y.

- $\mathsf{Cov}[e,Y] = 0$. This states that the estimation error is uncorrelated with the measurement Y. We compute this directly as

$$\mathsf{Cov}[e,Y] = \mathbb{E}[(\tilde{X} - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\tilde{Y})\tilde{Y}] = \mathsf{Cov}[X,Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathsf{Var}[Y] = 0.$$

- $\mathbb{E}[eY] = 0$, so that the estimation error is orthogonal to the measurement Y. This is because $\mathbb{E}[eY] = \mathsf{Cov}[e,Y] + \mathbb{E}[e]\mathbb{E}[Y] = 0$ because $\mathbb{E}[e] = 0$ and $\mathsf{Cov}[e,Y] = 0$.

- e, Y are jointly Gaussian, since e is a linear transformation of X, Y, and $\mathsf{Cov}[e,Y] = 0$, then e, Y are independent!

Example 5.8
Let X, Y be zero-mean, unit variance Gaussian random variables with correlation coefficient $\rho_{X,Y} = 0.5$. Compute the covariance of X and Y. Compute the conditional probability density of X given $Y = 2$.

From the correlation coefficient definition,

$$\rho_{X,Y} = \frac{\mathsf{Cov}[X,Y]}{\sqrt{\mathsf{Var}[X]\mathsf{Var}[Y]}} = \mathsf{Cov}[X,Y] = 0.5.$$

For the conditional density, we know it is Gaussian, so we compute the conditional mean and the conditional covariance.

$$\mathbb{E}[X|Y=2] = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(2 - \mathbb{E}[Y]) = 0.5 \cdot 2 = 1.$$

$$\mathsf{Var}[X|Y=2] = \mathsf{Var}[X] - \frac{cov[X,Y]^2}{\mathsf{Var}[Y]} = 1 - 0.25 = 0.75.$$

The conditional density is Gaussian with mean 1, variance 0.75.

Example 5.9
Assume that X, Y are *correlated,* jointly Gaussian random variables, such that $\mathbb{E}[X] = \mathbb{E}[Y] = 1, \mathsf{Var}[X] = 1, \mathsf{Var}[Y] = 1$ and $\mathsf{Cov}[X,Y] = 0.5$. Define derived random variables $A = 2X - 3, B = X - 2Y$.

1. Are A, B Gaussian?
   Yes. Linear combinations of joint Gaussians are Gaussian.
2. What are $\mathbb{E}[A], \mathbb{E}[B]$?
   Using the linearity of expectations, $\mathbb{E}[A] = 2\mathbb{E}[X] - 3 = -1$. $\mathbb{E}[B] = \mathbb{E}[X] - 2\mathbb{E}[Y] = -1$.
3. Compute $\mathsf{Var}[A], \mathsf{Var}[B]$.
   Since A is a scaled version of X, translated, we have $\mathsf{Var}[A] = (2)^2\mathsf{Var}[X] = 4$. For B, we use the method for representing the zero-mean random variables $\tilde{B}, \tilde{X}, \tilde{Y}$, so that

$$\mathsf{Var}[B] = \mathbb{E}[\tilde{B}^2] = \mathbb{E}[(\tilde{X} - 2\tilde{Y})^2] = \mathbb{E}[\tilde{X}^2] - 4\mathbb{E}[\tilde{X}\tilde{Y}] + 4\mathbb{E}[\tilde{Y}^2] = \mathsf{Var}[X] - 4\mathsf{Cov}[X,Y] + 4\mathsf{Var}[Y] = 3.$$

4. Compute $\mathsf{Cov}(A,B)$.
   Proceeding as before with the zero-mean representations,
$$\mathsf{Cov}[A,B] = 2\mathbb{E}[\tilde{X}^2] - 4\mathbb{E}[\tilde{X}\tilde{Y}] = 2\mathsf{Var}[X] - 4\mathsf{Cov}[X,Y] = 2-2 = 0.$$
5. Are X,Y independent? Explain.
   They are clearly not independent, since the covariance is non-zero.
6. Are A,B independent? Explain.
   Yes, they are independent, because they are uncorrelated and Gaussian.
7. Compute $\mathbb{E}[Y|A=a]$.
   We know $\mathbb{E}[Y|A=a] = \mathbb{E}[Y] + \frac{\mathsf{Cov}[A,Y]}{\mathsf{Var}[A]}(a-\mathbb{E}[A])$. We have most of those terms computed, except for $\mathsf{Cov}[A,Y]$, which is $\mathsf{Cov}[A,Y] = \mathbb{E}[\tilde{A}\tilde{Y}] = \mathbb{E}[2\tilde{X}\tilde{Y}] = 1$. Hence, $\mathbb{E}[Y|A=a] = 1 + \frac{1}{4}(a+1)$.
8. Let $e = Y - \mathbb{E}[Y|A=a]$. Compute $\mathbb{E}[e^2]$.
   Since e is the conditional estimation error, this is asking for the conditional variance $\mathsf{Var}[Y|A=a] = \mathsf{Var}[Y] - \frac{\mathsf{Cov}[Y,A]^2}{\mathsf{Var}[A]} = 1 - \frac{1}{4} = \frac{3}{4}$.
9. Compute the covariance between B and Y.
   By now, we know how to do this with the zero-mean versions:
$$\mathsf{Cov}[B,Y] = \mathbb{E}[\tilde{B}\tilde{Y}] = \mathbb{E}[(\tilde{X}-2\tilde{Y})\tilde{Y}] = \mathsf{Cov}[X,Y] - 2\mathsf{Var}[Y] = -\frac{3}{2}.$$

Example 5.10
Suppose we have two jointly continuous random variables X,Y, with marginal probability densities $f_X(x), f_Y(y)$ that are Gaussian. Must the pair X,Y be jointly Gaussian random variables?
Surprisingly, the answer to this is no. Consider the following jointly continuous random variables X,Y with joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi}e^{-\frac{x^2+y^2}{2}} & 0 \le xy, \\ 0 & \text{otherwise.} \end{cases}$$

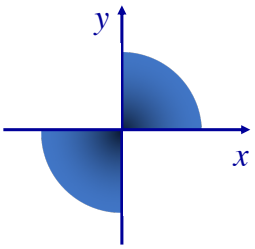


Figure 5.5: Non Gaussian PDF with Gaussian marginals.

This density is illustrated in the figure on the right. As you can see, it is definitely not a Gaussian, since the range of (X,Y) is not all of $\Re^2$. The marginal density of X is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \begin{cases} \frac{1}{\pi}e^{-\frac{x^2}{2}}\int_0^{\infty} e^{-\frac{y^2}{2}}\,dy = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} & x \ge 0, \\ \frac{1}{\pi}e^{-\frac{x^2}{2}}\int_{-\infty}^{0} e^{-\frac{y^2}{2}}\,dy = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} & x < 0. \end{cases}$$

which is Gaussian. Similarly, the marginal density of Y is Gaussian. This shows that having Gaussian marginal densities does not guarantee that the joint density is Gaussian.

5.5 Random Vectors

So far, we have focused our analysis on pairs of random variables X,Y. Nevertheless, the theory that we introduced for pairs of random variables extends easily to higher dimensional vectors. Given a probability space $(\Omega,\mathcal{E},\mathbb{P})$, we can define a **random vector** as a function that maps outcomes $\omega\in\Omega$ to vectors $\underline{x}(\omega)$ that take values in $\Re^n$, an n-dimensional Euclidean space. The theory of random vectors parallels the development we have presented for pairs of random variables. We can define the cumulative distribution function $F_{\underline{X}}(\underline{x})$ for general random vectors. If the random vectors are discrete, one defines the joint Probability Mass Function $P_{\underline{X}}(\underline{x})$ in a similar manner as we did for pairs of random variables. Random vectors are jointly continuous if there is a density $f_{\underline{X}}(\underline{x})$ such that the joint CDF can be written as

$$F_{\underline{X}}(\underline{x}) = \int_{\underline{a}\le\underline{x}} \cdots \int f_{\underline{X}}(a_1,\cdots,a_n)\,da_1\cdots da_n.$$

While all of this is formally interesting, one seldom has enough information to compute the full multidimensional joint probability density of random vectors, unless one has extra structure. For instance, if the components of the random vector $\underline{X} = \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}^T$ are independent, then $f_{\underline{X}}(\underline{x}) = \prod_{k=1}^{n} f_{X_k}(x_k)$. However, it is much easier to compute statistics such as means, variances and covariances.

In this section, we focus on defining first- and second-order statistics for random vectors, and describe how they change as the random vectors undergo linear transformations. We show subsequently how to extend our analysis of pairs of jointly Gaussian random variables to Gaussian random vectors, where the full joint PDF can be defined in terms of first- and second order statistics.

Let $\underline{X}$ be a random vector with values in $\Re^n$. We assume random vectors are column vectors, so $\underline{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$. We define the mean of $\underline{X}$, or its expected value, as $\mathbb{E}[\underline{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$.

Since expectation is a linear operation, this is simply the vector of expected values, one for each random variable in the random vector $\underline{X}$. For pairs of random variables X, Y, this corresponds to stacking the individual expected values into a vector, as

$$\underline{X} = \begin{bmatrix} X \\ Y \end{bmatrix}; \qquad \mathbb{E}[\underline{X}] = \begin{bmatrix} \mathbb{E}[X] \\ \mathbb{E}[Y] \end{bmatrix}.$$

For pairs of random variables $\underline{X}$, we define the covariance matrix $\Sigma_{\underline{X}}$ as

$$\Sigma_{\underline{X}} = \begin{bmatrix} \mathsf{Var}[X] & \mathsf{Cov}[X,Y] \\ \mathsf{Cov}[X,Y] & \mathsf{Var}[Y] \end{bmatrix}.$$

Note that this is a symmetric matrix. We can write this covariance matrix as:

$$\begin{aligned} \Sigma_{\underline{X}} &= \begin{bmatrix} \mathbb{E}[(X-\mathbb{E}[X])^2] & \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] & \mathbb{E}[(Y-\mathbb{E}[Y])^2] \end{bmatrix} \\ &= \mathbb{E}\Big[\begin{bmatrix} (X-\mathbb{E}[X])^2 & (X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ (X-\mathbb{E}[X])(Y-\mathbb{E}[Y]) & (Y-\mathbb{E}[Y])^2 \end{bmatrix} \Big], \\ &= \mathbb{E}\Big[(\underline{X}-\mathbb{E}[\underline{X}])(\underline{X}-\mathbb{E}[\underline{X}])^T \Big] \end{aligned}$$

where $\underline{X}^T$ is the transpose of the column vector, resulting in a row vector. Hence, the covariance matrix is the expected value of the outer product between a column vector of dimension 2, and a row vector of dimension 2, resulting in a 2×2 matrix. Note that this is simply arranging the scalar statistics $\mathsf{Var}[X], \mathsf{Var}[Y], \mathsf{Cov}[X,Y]$ in a matrix form. We can generalize this to n-dimensional random vectors.

For an n-dimensional random vector $\underline{X}$, the **covariance matrix** is an $n \times n$ matrix defined as

$$\Sigma_{\underline{X}} = \mathbb{E}\Big[(\underline{X}-\mathbb{E}[\underline{X}])(\underline{X}-\mathbb{E}[\underline{X}])^T \Big]$$

Using the linearity property of expectations, and multiplying the matrix, we get

$$\begin{aligned} \Sigma_{\underline{X}} &= \mathbb{E}\Big[\underline{X}\underline{X}^T - \mathbb{E}[\underline{X}]\underline{X}^T - \underline{X}\mathbb{E}[\underline{X}]^T + \mathbb{E}[\underline{X}]\mathbb{E}[\underline{X}]^T \Big] \\ &= \mathbb{E}\Big[\underline{X}\underline{X}^T \Big] - \mathbb{E}\Big[\mathbb{E}[\underline{X}]\underline{X}^T \Big] - \mathbb{E}\Big[\underline{X}\mathbb{E}[\underline{X}]^T \Big] + \mathbb{E}\Big[\mathbb{E}[\underline{X}]\mathbb{E}[\underline{X}]^T \Big] \\ &= \mathbb{E}\Big[\underline{X}\underline{X}^T \Big] - \mathbb{E}[\underline{X}]\mathbb{E}\Big[\underline{X}^T \Big] - \mathbb{E}\Big[\underline{X} \Big]\mathbb{E}[\underline{X}]^T + \mathbb{E}[\underline{X}]\mathbb{E}[\underline{X}]^T \quad \text{(Take out constants from expectations)} \\ &= \mathbb{E}\Big[\underline{X}\underline{X}^T \Big] - \mathbb{E}[\underline{X}]\mathbb{E}[\underline{X}]^T \qquad \text{(add the last 3 terms, which are the same.)} \end{aligned}$$

This is the generalization of the scalar identity $\mathsf{Cov}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ to the vector case. As in the scalar case, the covariance matrix can be computed as the difference between the second moment matrix $\mathbb{E}\left[\underline{X}\underline{X}^T\right]$ and the outer product of the mean vectors $\mathbb{E}[\underline{X}]\mathbb{E}[\underline{X}]^T$.

Note that every element in the covariance matrix is either a variance of a random variable, or a covariance between two random variables. Specifically,

$$\Sigma_{\underline{X}} = . \begin{bmatrix} \mathsf{Var}[X_1] & \mathsf{Cov}[X_1, X_2] & \cdots & \mathsf{Cov}[X_1, X_n] \\ \mathsf{Cov}[X_1, X_2] & \mathsf{Var}[X_2] & \cdots & \mathsf{Cov}[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}[X_1, X_n] & \mathsf{Cov}[X_2, X_n] & \cdots & \mathsf{Var}[X_n] \end{bmatrix}$$

Thus, the covariance matrix is a compilation of the second order statistics for the scalar components of the random vector $\underline{X}$.

The covariance matrix $\Sigma_{\underline{X}}$ has the following properties:

- It is a symmetric matrix.
- It is a positive semidefinite matrix: for any non-zero n-dimensional vector $\underline{a}$, the scalar defined by the matrix vector product $\underline{a}^T \Sigma_{\underline{X}} \underline{a} \geq 0$. See the appendix on linear algebra for details on what positive semi-definite means.
- The matrix $\Sigma_{\underline{X}}$ has all of its eigenvalues on the real line, and they are non-negative.
- The matrix $\Sigma_{\underline{X}}$ has n distinct eigenvectors, and each eigenvector is perpendicular to the others.

These properties will be useful in later chapters when we discuss problems of feature aggregation in data science problems. We briefly justify the most important property, that states that the covariance matrix must be positive semidefinite. Note the following: Given a random n-dimensional vector $\underline{X}$ and an n-dimensional constant vector $\underline{a}$, the random variable $Z = \underline{a}^T \underline{X}$ is a linear combination of the elements of X. If X were zero-mean, then $\mathbb{E}[Z] = \mathbb{E}[\underline{a}^T \underline{X}] = \underline{a}^T \mathbb{E}[X] = 0$. Thus, Z is also zero mean, with variance

$$\begin{aligned} \mathsf{Var}[Z] = \mathbb{E}[Z^2] &= \mathbb{E}[\underline{a}^T \underline{X}\underline{X}^T \underline{a}] \quad \text{since } \underline{a}^T \underline{X} = \underline{X}^T \underline{a}, \\ &= \underline{a}^T \mathbb{E}[\underline{X}\underline{X}^T] \underline{a} = \underline{a}^T \Sigma_{\underline{X}} \underline{a} \geq 0 \end{aligned}$$

Thus, the positive semidefinite property follows because covariances of random variables are non-negative. Note how we carefully moved the constants $\underline{a}$ from the correct side of the expectation to keep the dimensions matching for the vector-matrix products.

Example 5.11
Suppose we have jointly continuous random variables $\underline{X} = [X_1, X_2, X_3]^T$, with joint probability density function

$$f_{\underline{X}}(\underline{x}) = \begin{cases} 6 & 0 \leq x_1 \leq x_2 \leq x_3 \leq 1, \\ 0 & \text{elsewhere.} \end{cases}$$

Compute the covariance matrix $\Sigma_{\underline{X}}$.

The range $R_{\underline{X}}$ of the density is shown on the right. We can see that it is an inverted triangular pyramid with base area 0.5 and height 1, so its volume is $\frac{1}{6}$, hence we use the constant 6 as the density in the range.

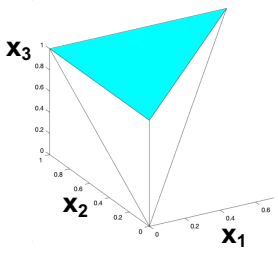


Figure 5.6: Range for Example.

We begin by computed the expected values:

$$\mathbb{E}[X_1] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_1 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_1\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 x_3^3\,dx_3 = \frac{1}{4}$$

$$\mathbb{E}[X_2] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_2 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_2\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 2x_3^3\,dx_3 = \frac{1}{2}$$

$$\mathbb{E}[X_3] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_3 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_3\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 3x_3^3\,dx_3 = \frac{3}{4}$$

Next, we compute the second moments:

$$\mathbb{E}[X_1^2] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_1^2 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_1^2\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 \frac{x_3^4}{2}\,dx_3 = \frac{1}{10}$$

$$\mathbb{E}[X_2^2] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_2^2 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_2^2\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 \frac{3x_3^4}{2}\,dx_3 = \frac{3}{10}$$

$$\mathbb{E}[X_3^2] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_3^2 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_3^2\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 3x_3^4\,dx_3 = \frac{3}{5}$$

Finally, we compute the covariances between the components of $\underline{X}$ as

$$\mathbb{E}[X_1X_2] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_1x_2 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_1x_2\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 \frac{3x_3^4}{4}\,dx_3 = \frac{3}{20}$$

$$\mathbb{E}[X_1X_3] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_1x_3 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_1x_3\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 x_3^4\,dx_3 = \frac{1}{5}$$

$$\mathbb{E}[X_2X_3] = \iiint\limits_{\underline{x}\in R_{\underline{X}}} x_2x_3 f_{\underline{X}}(\underline{x})\,d\underline{x} = \int_0^1 \Big(\int_0^{x_3}(\int_0^{x_2} 6x_2x_3\,dx_1)\,dx_2\Big)dx_3 = \int_0^1 2x_3^4\,dx_3 = \frac{2}{5}$$

Thus, the variances and covariances are given by:

$$\begin{aligned}
\mathsf{Var}[X_1] &= \mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2 = \frac{1}{10} - \frac{1}{16} = \frac{3}{80}\\
\mathsf{Var}[X_2] &= \mathbb{E}[X_2^2] - (\mathbb{E}[X_2])^2 = \frac{3}{10} - \frac{1}{4} = \frac{1}{20}\\
\mathsf{Var}[X_3] &= \mathbb{E}[X_3^2] - (\mathbb{E}[X_3])^2 = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}\\
\mathsf{Cov}[X_1,X_2] &= \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] = \frac{3}{20} - \frac{1}{8} = \frac{1}{40}\\
\mathsf{Cov}[X_1,X_3] &= \mathbb{E}[X_1X_3] - \mathbb{E}[X_1]\mathbb{E}[X_3] = \frac{1}{5} - \frac{3}{16} = \frac{1}{80}\\
\mathsf{Cov}[X_2,X_3] &= \mathbb{E}[X_2X_3] - \mathbb{E}[X_2]\mathbb{E}[X_3] = \frac{2}{5} - \frac{3}{8} = \frac{1}{40}
\end{aligned}$$

The full covariance matrix is

$$\Sigma_{\underline{X}} = \begin{bmatrix} 0.1 & 0.15 & 0.2 \\ 0.15 & 0.3 & 0.4 \\ 0.2 & 0.3 & 0.6 \end{bmatrix} - \begin{bmatrix} 0.25 \\ 0.5 \\ 0.75 \end{bmatrix} \begin{bmatrix} 0.25 & 0.5 & 0.75 \end{bmatrix} = \begin{bmatrix} 0.0375 & 0.0250 & 0.0125 \\ 0.0250 & 0.0500 & 0.0125 \\ 0.0125 & 0.0125 & 0.0375 \end{bmatrix}$$

Assume we have a random n-dimensional vector $\underline{X}$ with mean $\underline{m}_{\underline{X}}$ and covariance $\Sigma_{\underline{X}}$. Define an affine transformation of $\underline{X}$ as follows: Let $\mathbf{A}$ be an $m \times n$ matrix, and $\underline{d}$ be an m-dimensional vector. The m-dimensional random vector $\underline{Y}$ is given by:

$$\underline{Y} = \mathbf{A}\underline{X} + \underline{d}.$$

We want to compute the first- and second-order statistics of $\underline{Y}$ based on knowing the statistics of $\underline{X}$.

It is easy to compute the mean using linearity of expectation:

$$\mathbb{E}[\underline{Y}] = \mathbb{E}[\mathbf{A}\underline{X} + \underline{d}] = \mathbb{E}[\mathbf{A}\underline{X}] + \mathbb{E}[\underline{d}] = \mathbf{A}\mathbb{E}[\underline{X}] + \underline{d} = \mathbf{A}\underline{m}_{\underline{X}} + \underline{d}$$

where we have pulled out constants from the expectations. Note that, since we are dealing with matrices and vectors, we move the constant matrix $\mathbf{A}$ out on the left side of the expectation, so that the dimensions of the matrices agree when doing matrix-vector multiplication.

To compute the covariance matrix of $\underline{Y}$, we subtract the mean from both sides, to get:

$$\underline{Y} - \mathbb{E}[\underline{Y}] = \mathbf{A}\underline{X} + \underline{d} - \mathbf{A}\underline{m}_{\underline{X}} - \underline{d} = \mathbf{A}(\underline{X} - \underline{m}_{\underline{X}})$$

Using the definition of covariance, we compute it as follows:

$$\begin{aligned}
\Sigma_{\underline{Y}} &= \mathbb{E}\left[(\underline{Y} - \mathbb{E}[\underline{Y}])(\underline{Y} - \mathbb{E}[\underline{Y}])^T\right] = \mathbb{E}\left[\mathbf{A}(\underline{X} - \underline{m}_{\underline{X}})\left(\mathbf{A}(\underline{X} - \underline{m}_{\underline{X}})\right)^T\right] \\
&= \mathbb{E}\left[\mathbf{A}(\underline{X} - \underline{m}_{\underline{X}})(\underline{X} - \underline{m}_{\underline{X}})^T\mathbf{A}^T\right] \\
&= \mathbf{A}\mathbb{E}\left[(\underline{X} - \underline{m}_{\underline{X}})(\underline{X} - \underline{m}_{\underline{X}})^T\right]\mathbf{A}^T \\
&= \mathbf{A}\Sigma_{\underline{X}}\mathbf{A}^T
\end{aligned}$$

This is the generalization of the scalar scaling law for covariances, where if $Y = aX$, then $\mathsf{Var}[Y] = a^2\mathsf{Var}[X]$. The extension to vectors is careful to keep the order of the scaling by $\mathbf{A}$ and $\mathbf{A}^T$ to keep the dimensions of the resulting matrix correct.

Example 5.12

Let's revisit the example of 5.7. We have a pair of random variables X, Y with first- and second-order statistics $\mathbb{E}[X] = \mathbb{E}[Y] = 0, \mathsf{Var}[X] = \mathsf{Var}[Y] = \frac{1}{4}, \mathsf{Cov}[X, Y] = 0$.

Let's form this into a vector $\underline{X} = \begin{bmatrix} X \\ Y \end{bmatrix}$. The mean vector $\underline{m}_{\underline{X}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and the resulting covariance matrix is

$$\Sigma_{\underline{X}} = \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{bmatrix}.$$

Define two new variables defined as $U = 3X + 2Y + 1, V = 2X - 3Y - 1$. Define the vector $\underline{W} = \begin{bmatrix} U \\ V \end{bmatrix}$. We can write the transformation from $\underline{X}$ to $\underline{W}$ as:

$$\underline{W} = \begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix}\underline{X} + \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Then, the first- and second-order statistics of $\underline{W}$ are:

$$\mathbb{E}[\underline{W}] = \begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix}\mathbb{E}[\underline{X}] + \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

$$\begin{aligned}
\Sigma_{\underline{W}} &= \begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix}\Sigma_{\underline{X}}\begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix}^T = \begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix}\begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{bmatrix}\begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix}\begin{bmatrix} \frac{3}{4} & \frac{1}{2} \\ \frac{1}{2} & -\frac{3}{4} \end{bmatrix} = \begin{bmatrix} \frac{13}{4} & 0 \\ 0 & \frac{13}{4} \end{bmatrix}
\end{aligned}$$

which says that $\mathsf{Var}[U] = \frac{13}{4}, \mathsf{Var}[V] = \frac{13}{4}, \mathsf{Cov}[U, V] = 0$. These are the same answers we saw in Example 5.7.

Example 5.13
Let's revisit Example 5.9. Assume that X, Y are correlated random variables, such that $\mathbb{E}[X] = \mathbb{E}[Y] = 1, \text{Var}[X] = 1, \text{Var}[Y] = 1$ and $\text{Cov}[X, Y] = 0.5$. Let $\underline{X} = \begin{bmatrix} X \\ Y \end{bmatrix}$. Then,

$$\mathbb{E}[\underline{X}] = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \Sigma_{\underline{X}} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

Define derived random variables $A = 2X - 3, B = X - 2Y$. Let $\underline{W} = \begin{bmatrix} A \\ B \end{bmatrix}$. Then,

$$\underline{W} = \begin{bmatrix} 2 & 0 \\ 1 & -2 \end{bmatrix} \underline{X} + \begin{bmatrix} -3 & 0 \end{bmatrix}.$$

Using this equation, we obtain

$$\begin{aligned} \mathbb{E}[\underline{W}] &= \begin{bmatrix} 2 & 0 \\ 1 & -2 \end{bmatrix} \mathbb{E}[\underline{X}] + \begin{bmatrix} -3 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -3 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \\ \Sigma_{\underline{W}} &= \begin{bmatrix} 2 & 0 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 1 & -2 \end{bmatrix}^T = \begin{bmatrix} 2 & 0 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & -2 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 1 & -1.5 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix} \end{aligned}$$

and thus $\text{Var}[A] = 4, \text{Var}[B] = 3$, and $\text{Cov}[A, B] = 0$, agreeing with the results from Example 5.9.

As the examples illustrate, the use of random vectors enables us to recover the same first- and second-order statistics for the random variables when we analyze them individually as pairs of random variables. The advantage of the vector notation is that it scales nicely to compute statistics for random vectors of dimension greater than 2, exploiting simple results from linear algebra.

5.5.1 Gaussian random vectors

A special case of random vectors is what are termed **Gaussian random vectors**. For pairs of jointly Gaussian random variables X, Y, their joint PDF is completely characterized by the first- and second-order statistics. Extending this to random vectors of dimension greater than two is straightforward, as we will show below.

We define a jointly Gaussian random vector as a generalization of what we did with pairs of random variables. First, we define n independent standard Gaussian random variables $Z_i \sim \mathcal{N}(0, 1)$. We define the vector

$$\underline{Z} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

Then, an n-dimensional random vector $\underline{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ is defined to be a *Gaussian random vector* (or equivalently, $\{X_1, \ldots, X_n\}$ are defined to be a set of jointly Gaussian random variables) if

$$\underline{X} = \mathbf{A}\underline{Z} + \underline{b}$$

for some $n \times n$ matrix $\mathbf{A}$ and some n-dimensional vector $\underline{b}$.

Note that $\underline{Z}$ has mean $\underline{0}$, with covariance matrix as the $n \times n$ identity matrix $\mathbf{I}_n$. Hence, $\mathbb{E}[\underline{X}] = \mathbf{A}\underline{0} + \underline{b} = \underline{b}$. Furthermore, the covariance matrix of $\underline{X}$ is

$$\Sigma_{\underline{X}} = \mathbf{A}\Sigma_{\underline{Z}}\mathbf{A}^T = \mathbf{A}\mathbf{A}^T.$$

For a Gaussian random vector $\underline{Z}$ to be jointly continuous, the transformation $\mathbf{A}$ must be invertible. This means that the resulting covariance $\Sigma_{\underline{X}}$ is invertible. We focus only on jointly continuous random Gaussian random variables in this text.

An equivalent definition is that $\underline{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ is a Gaussian random vector if, for all constant vectors $\underline{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$, the random variable $Z = \sum_{k=1}^n a_k X_k$ is a Gaussian random variable. Note that $Z = \underline{a}^T \underline{X}$ in vector notation. As noted before, it is *not enough* that each entry X_i is marginally a Gaussian random variable for the vector to be a Gaussian random vector! *All* linear combinations of the entries must also be Gaussian. The converse, however is true: the entries of a Gaussian random vector are individually Gaussian random variables.

If $\underline{X}$ had mean $\underline{m}_{\underline{X}}$ and covariance $\Sigma_{\underline{X}}$, then $Z = \underline{a}^T \underline{X}$ is a scalar Gaussian random variable with mean $\mathbb{E}[Z] = \underline{a}^T \underline{m}_{\underline{X}}$ and variance $\underline{a}^T \Sigma_{\underline{X}} \underline{a}$.

A jointly continuous Gaussian random vector $\underline{X}$ have a probability density function that is completely described by its mean $\underline{m}_{\underline{X}}$ and covariance $\Sigma_{\underline{X}}$. We use the notation $\underline{X} \sim N(\underline{m}_{\underline{X}}, \Sigma_{\underline{X}})$ to denote this density. We can write the joint PDF of $\underline{X}$ as

$$f_{\underline{X}}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_{\underline{X}})}} e^{-\frac{1}{2}(\underline{x} - \underline{m}_{\underline{X}})^T (\Sigma_{\underline{X}})^{-1} (\underline{x} - \underline{m}_{\underline{X}})}.$$

An important property of pairs of jointly Gaussian random variables X, Y is that they are independent if and only if $\mathsf{Cov}[X, Y] = 0$. For Gaussian random vectors, the components $X_1, X_2, \ldots, X_n$ are mutually independent if and only if $\mathsf{Cov}[X_i, X_j] = 0$ for all $i, j \in 1, \ldots, n, i \neq j$. What this means is that the covariance matrix $\Sigma_{\underline{X}}$ is diagonal, with zeros in all the non-diagonal entries. For independent random vectors, the covariance matrix is

$$\Sigma_{\underline{X}} = \begin{bmatrix} \mathsf{Var}[X_1] & 0 & \cdots & 0 \\ 0 & \mathsf{Var}[X_2] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathsf{Var}[X_n] \end{bmatrix}$$

In this special case,

$$\Sigma_{\underline{X}}^{-1} = \begin{bmatrix} \frac{1}{\mathsf{Var}[X_1]} & 0 & \cdots & 0 \\ 0 & \frac{1}{\mathsf{Var}[X_2]} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\mathsf{Var}[X_n]} \end{bmatrix}$$

and the joint probability density factors as

$$f_{\underline{X}}(\underline{x}) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi \mathsf{Var}[X_k]}} e^{-\frac{(x_k - m_k)^2}{2\mathsf{Var}[X_k]}},$$

which shows the equivalence between independence and having a diagonal covariance matrix.

Example 5.14
Let's revisit Example 5.12 where X, Y are jointly Gaussian random variables with first- and second-order statistics $\mathbb{E}[X] = \mathbb{E}[Y] = 1, \text{Var}[X] = 1, \text{Var}[Y] = 1$ and $\text{Cov}[X, Y] = 0.5$. Let $\underline{X} = \begin{bmatrix} X \\ Y \end{bmatrix}$. Then,

$$\mathbb{E}[\underline{X}] = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \Sigma_{\underline{X}} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

Define $\underline{W}$ as

$$\underline{W} = \begin{bmatrix} 2 & 0 \\ 1 & -2 \end{bmatrix} \underline{X} + \begin{bmatrix} -3 & 0 \end{bmatrix}.$$

Then, from Example 5.12, we know

$$\mathbb{E}[\underline{W}] = \underline{m}_{\underline{W}} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}; \qquad \Sigma_{\underline{W}} = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}.$$

which implies that the two components of $\underline{W}$ are uncorrelated, and hence, mutually independent. The joint density of $\underline{W}$ is Gaussian, and given by

$$f_{\underline{W}}(\underline{w}) = \left(\frac{1}{\sqrt{8\pi}} e^{-\frac{(w_1+1)^2}{8}} \right) \left(\frac{1}{\sqrt{6\pi}} e^{-\frac{(w_2+2)^2}{6}} \right),$$

which shows the factored form.

Chapter 6

Detection Theory

In this chapter we start our investigation of statistical detection theory, also referred to as hypothesis testing or sometimes decision theory. The fundamental problem in statistical detection theory is summarized as follows: In a probability experiment, one and only one of several possible events has happened. After collecting observations with distributions that depend on which event happened, make a decision as to which one of the events actually happened. To illustrate this, consider the following example:

Example 6.1
A sonar system transmits pressure pulses into the water in a given direction, hoping to determine whether a submarine is present in that direction or not. The pulses propagate through the water, and interact with background as well as with a submarine if it is present. The sonar receiver listens for echoes, which may come from the submarine, as well as from background such as ocean floor features, large sea mammals, school of fish, etc. The receiver collects the echoes, and must decide whether there is a submarine present or not based on the received signal.

Note the key components of this problem. There are two possible events, corresponding to many different outcomes in the sample space: the event where a submarine is present in the direction of the sonar pulses, and the event where the submarine is absent. These events are disjoint, and in the terminology of probability events, collectively exhaustive: one of the two events must happen. We collect a measurement, which is a random variable that is a function of the outcome in the experiment. Based on the observed measurement, we must make a decision as to which one of the two possible events is "best to choose."

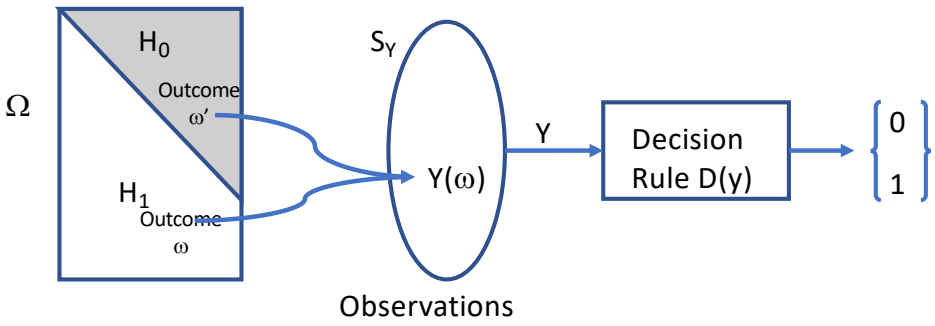


Figure 6.1: Detection problem components.

A general model of this process is shown in Figure 6.1. There are two possible events in the sample space Ω, each of which represents many outcomes. Each of these events is called a *hypothesis*. We use a measurement instrument that collects a random variable Y. Based on the measurement observation $Y = y$, we must design a rule to decide which is the correct hypothesis.

From Figure 6.1 we see that we will need three components in our model:

1. A model of generation processes that creates H_0, H_1.
2. A model of the observation process that generates the observation $Y = y$.
3. A decision rule $D(y)$ that maps each possible observation value y to an associated decision.

In general, the first two elements are set by the experiment or the restrictions of the physical data gathering situation, and we need to model them, but we don't control their design. For example, if we are trying to decide whether an area in a breast cancer mammogram is cancerous or not, the true state of that area (cancerous or not) is selected by processes outside of our control. The measurement instrument (the X-Ray imager) is a physical sensor that generates noisy images depending on whether the area is cancerous or not.

We want to avoid generating a complete description of the probability space $(\Omega, \mathcal{E}, \mathbb{P})$ to model the relationship of the observations Y and the event hypotheses H_0, H_1. Assume Y is a discrete random variable. Using the Law of Total Probability yields

$$\begin{aligned}\mathbb{P}[\{Y=y\}] &= \mathbb{P}[\{Y=y\} \cap H_0] + \mathbb{P}[\{Y=y\} \cap H_1] = \mathbb{P}[\{Y=y\}|H_0]\mathbb{P}[H_0] + \mathbb{P}[\{Y=y\}|H_1]\mathbb{P}[H_1] \\ &= P_{Y|H_0}(y)\mathbb{P}[H_0] + P_{Y|H_1}(y)\mathbb{P}[H_1]\end{aligned}$$

This indicates the components of how we model the detection problem:

1. A model of generation processes that creates H_0, H_1: $\mathbb{P}[H_1], \mathbb{P}[H_0]$.
2. A model of the observation process that generates the observation $Y = y$: $\mathbb{P}[\{Y = y\}|H_0], \mathbb{P}[\{Y = y\}|H_1]$.

This is a compact, probabilistic description that represents the detection problem. Based on this model, we design a decision rule that maps the possible measurement values into a decision. When there are only two possible hypotheses H_0, H_1, this decision rule corresponds to a partition of the space of possible observations into two regions: the region where the decision will be H_1, and the region where the decision will be H_0, as illustrated in Figure 6.2.

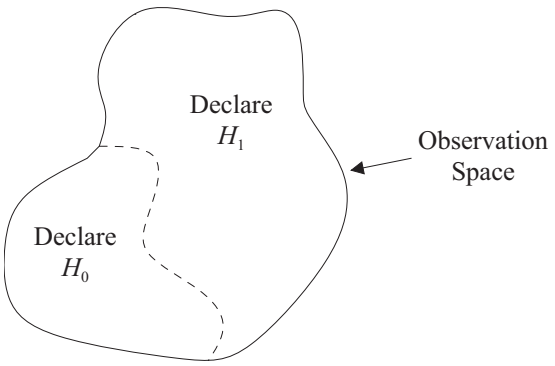


Figure 6.2: Illustration of a decision rule as a partition of the observation space into disjoint regions, illustrated here for the case of two possibilities.

We first discuss in detail the case that arises when there are only two possible hypotheses, termed binary hypothesis testing. Subsequently, we discuss the more general case of M hypotheses, for $M > 2$.

6.1 Binary Hypothesis Testing

In this section we consider the simplest case when there are only two possible states of nature or hypotheses, which by convention we label as H_0 and H_1. This situation is termed "binary hypothesis testing" and the H_0 hypothesis is usually termed the "null hypothesis," due to its typical association with the absence of some quantity of interest.

The binary case is of considerable practical importance, as well as having a long and rich history. Let's examine a few motivating applications before proceeding to more detailed developments.

Example 6.2 (Communications)
Consider the following simplified version of a communication system, where a source broadcasts one bit, (either 0 or 1) The transmitter encodes this bit by a voltage, which is either 0 or E, depending on the bit. The receiver observes a noisy version of the transmitted signal, where the noise is additive, and is represented by a random variable w with zero-mean, variance σ^2, and Gaussian distribution. The receiver knows the nature of the signal E, the statistics of the noise σ^2, and the apriori probability $p(k)$ that the bit sent was k, where $k = 0, 1$. The receiver must take the received signal, y, and map this using a rule $D(y)$ into either 0 or 1, depending on the value of r. The problem is to determine the decision rule for which the probability of receiver error is minimized.

Example 6.3 (Radar)
A simple radar system makes a scalar observation y to determine the absence or presence of a target at a given range and heading. If a target is present (hypothesis H_1), the observed signal is $y = E + w$, where E is a known signal level, and $w \sim N(0, \sigma^2)$. If no target is present (hypothesis H_0), then only noise is received $y = w$. Find the decision rule for maximizing the probability of detecting the target, given a bound on the probability of false alarm.

Example 6.4 (Quality Control)
At a factory, an automatic quality control device is used to determine whether a manufactured unit is satisfactory (hypothesis H_0) or defective (hypothesis H_1), by measuring a simple quality factor q. Past statistics indicate that one out of every 10 units is defective. For satisfactory units, $q \sim N(2, \sigma^2)$, whereas for defective units, $q \sim N(1, \sigma^2)$. The quality control device is set to remove all units for which $q < t$, where t is a threshold to be designed. The problem is to determine the optimal threshold setting in order to maximize the probability of detecting a defect, subject to the constraint that the probability of removing a satisfactory unit is at most 0.005.

All of the above examples illustrate the problem of binary hypothesis testing. We will develop the relevant theory next.

6.1.1 Detection model

The detection problem is set in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, which we model in a very abbreviated way. Assume there are only two hypothesis, denoted as H_0 and H_1, which are events in $\mathcal{E}$ are events in the which are mutually disjoint, and collectively exhaustive ($H_0 \cup H_1 = \Omega$). We know $\mathbb{P}[H_0], \mathbb{P}[H_1]$. The figure on the right illustrates the events H_0, H_1 in the sample space, representing a partition of Ω.

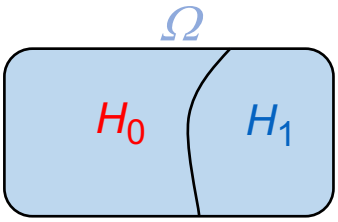


Figure 6.3: Events H_0, H_1.

Observation model: The measurement is a random variable Y defined on $(\Omega, \mathcal{E}, \mathbb{P})$. Y can be either discrete or continuous. For discrete Y, we model the measurement using a pair of conditional probability mass functions $P_{Y|H_1}(y), P_{Y|H_0}(y)$. For continuous Y, we model the measurement in terms of a pair of conditional probability density functions $f_{Y|H_1}(y), f_{Y|H_0}(y)$. These conditional probability functions are known, and are referred to as the **likelihoods** of the measurement $Y = y$ given the different hypotheses.

The figure on the right illustrates the observation model. Note that outcomes in H_0 and outcomes in H_1 can map to the same observation $Y = y$. However, it may be more likely to occur under one of those two hypotheses, as determined by the likelihoods $P_{Y|H_1}(y), P_{Y|H_0}(y)$ or $f_{Y|H_1}(y), f_{Y|H_0}(y)$. These likelihoods will influence which decisions to make.

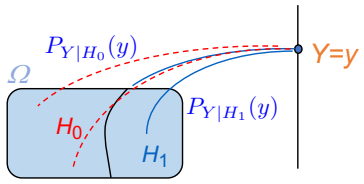


Figure 6.4: Likelihoods $P_{Y|H_1}(y), P_{Y|H_0}(y)$.

Decision rule: A decision rule is a function $U = D(Y)$ of the random variable Y, that maps Y into a decision $U \in \{0, 1\}$. The decision $D(y) = 0$ corresponds to deciding that H_0 is the selected hypothesis when the observation is $Y = y$, and $D(y) = 1$ indicates that H_1 is the selected hypothesis for $Y = y$. $U = D(Y)$ is a discrete random variable, mapping the range R_Y into two possible values. The sets $\{y \in R_Y : D(y) = 0\}$ and $\{y \in R_Y : D(y) = 1\}$ form a partition of R_Y, because $D(\cdot)$ is a function defined everywhere on R_Y. This is illustrated in Figure 6.2.

The decision rule is the solution we design for the detection problem. To do proper design, we select the decision rule on the basis of how good its performance will be.

One way to measure performance is in terms of the errors made by the decision rule. Specifically, when H_0 is true, and generates a measurement $Y = y$ such that $D(y) = 1$, the decision rule has made an error. The figure on the right illustrates the two types of error that the decision rule can make. When H_0 is the event that generates. measurement $Y = y$, and the decision rule selects $D(y) = 1$, we call this a **false alarm**. This terminology dates back to early detection problems such as detecting aircraft using radar, where H_0 was the hypothesis that no airplanes were present. Similarly, when the measurement y is generated by H_1, and $D(y)$ is such that $D(y) = 0$, we refer to this as a **missed detection.**

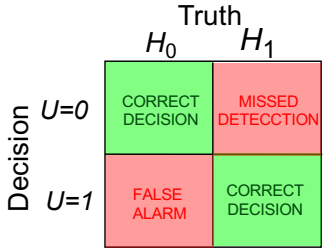


Figure 6.5: Types of Detection Errors.

Given a detection rule $U = D(Y)$, we can compute the probability of a missed detection using the likelihood $P_{Y|H_1}(y)$ if Y is discrete or $f_{Y|H_1}(h)$ if Y is continuous. Denote by A_0 the subset of the range of Y where $D(y) = 0$: $A_0 = \{y \in R_Y : D(y) = 0\}$. Then, the probability of a missed detection is

$$P_{MD} \equiv \mathbb{P}[y \in A_0 | H_1] = \begin{cases} \sum_{y \in A_0} P_{Y|H_1}(y) & Y \text{ is a discrete random variable,} \\ \int_{y \in A_0} f_{Y|H_1}(y)\, dy & Y \text{ is a continuous random variable.} \end{cases}$$

Thus, P_{MD} is the probability of making an erroneous decision when H_0 is true.

Similarly, let $A_1 = \{y \in R_Y : D(y) = 1\}$. Then, $A_0 \cup A_1 = R_Y$, the range of possible values of Y. The probability of a false alarm is computed using the likelihood $P_{Y|H_0}(y)$ if Y is discrete or $f_{Y|H_0}(h)$ if Y is continuous as follows:

$$P_{FA} \equiv \mathbb{P}[y \in A_1 | H_0] = \begin{cases} \sum_{y \in A_1} P_{Y|H_0}(y) & Y \text{ is a discrete random variable,} \\ \int_{y \in A_1} f_{Y|H_0}(y)\, dy & Y \text{ is a continuous random variable.} \end{cases}$$

P_{FA} is the probability of making an erroneous decision when H_1 is true.

Note that P_{FA}, P_{MD} are conditional statistics. If we know $\mathbb{P}[H_0], \mathbb{P}[H_1]$, we can compute unconditional statistics such as the average probability of error using the Law of Total Probability, as:

$$P_e \equiv \mathbb{P}[\text{Error}] = \mathbb{P}[\text{Error}|H_0]\mathbb{P}[H_0] + \mathbb{P}[\text{Error}|H_1]\mathbb{P}[H_1] = P_{FA}\mathbb{P}[H_0] + P_{MD}\mathbb{P}[H_1].$$

We can now use these performance measures to define criteria for selecting a decision rule. We describe different approaches for designing decision rules next.

6.2 Maximum Likelihood Detection

The most common approach for designing a decision rule is known as **maximum likelihood detection.** Assume that Y is a discrete random variable. Given a measurement y, we compute the likelihood of this measurement under each hypothesis, using $P_{Y|H_0}(y)$ and $P_{Y|H_1}(y)$. The maximum likelihood (ML) decision selects the hypothesis that has the largest likelihood for that measurement. That is,

$$D^{ML}(y) = \begin{cases} 1, & P_{Y|H_1}(y) \geq P_{Y|H_0}(y), \\ 0, & P_{Y|H_1}(y) < P_{Y|H_0}(y). \end{cases}$$

We break ties arbitrarily, so we assign a tie to 1.

The maximum likelihood method for detection and estimation was developed by the statistician R. A. Fisher in the early 20th century, although some limited results appeared earlier.

Example 6.5
Assume we have a coin, which may be biased so that the probability of obtaining heads is 0.6. Hypothesis H_1 is that the coin has probability of heads = 0.6. Hypothesis H_0 is that the coin is unbiased, so the probability of heads = 0.5. To detect whether the coin is biased or not, we conduct an experiment, where we flip the coin independently 5 times, and count the number of heads that appear in the experiment. Thus, the measurement in the experiment, Y, is the number of heads in five coin flips.

Y is a discrete random variable, with $R_Y = \{0, 1, 2, 3, 4, 5\}$. The above description lets us describe the likelihood functions: $P_{Y|H_0}(y)$ is the probability mass function of a Binomial(5,0.5) random variable, and $P_{Y|H_1}(y)$ is the probability mass function of a Binomial(5,0.6) random variable. In this case, the range R_Y is small, so we can enumerate the two probability mass functions, and compare their values for each $y \in R_Y$, as shown in the table below.

Y:	0	1	2	3	4	5
$P_{Y\|H_1}$	0.01024	0.0768	0.2304	**0.3456**	**0.2592**	**0.07776**
$P_{Y\|H_0}$	**0.03125**	**0.15625**	**0.3125**	0.3125	0.15625	0.03125

To compute the maximum likelihood decision, we compare the numbers in each column, and pick the larger of the two numbers. In the table above, we have highlighted the larger number in bold and magenta color. Thus we see that the maximum likelihood decision rule becomes:

$$D^{ML}(y) = \begin{cases} 1, & y = 3, 4, 5, \\ 0, & y = 0, 1, 2. \end{cases}$$

The decision agrees with intuition: a larger count of heads suggests the coin is more likely to be unbalanced, whereas a smaller count of heads indicates the coin is more likely to be balanced.

What is the performance of the maximum likelihood decision rule? Let's compute the probability of missed detection. As discussed above, this is the probability that, when H_1 is the correct hypothesis, we get a value y where the decision $D^{ML}(y)$ is 0. Therefore,

$$P_{MD} = \mathbb{P}[\{y = 0, 1, 2\}|H_1] = P_{Y|H_1}(0) + P_{Y|H_1}(1) + P_{Y|H_1}(2) = 0.31744.$$

Similarly, the probability of false alarm is the probability that, when H_0 is the correct hypothesis, we get a measurement $Y = y$ where $D^{ML}(y) = 1$. Then,

$$P_{FA} = \mathbb{P}[\{y = 3, 4, 5\}|H_0] = P_{Y|H_0}(3) + P_{Y|H_0}(4) + P_{Y|H_0}(5) = 0.5$$

Assuming that $\mathbb{P}[H_0] = \mathbb{P}[H_1] = 0.5$, we can compute the probability of error as

$$P_e = \mathbb{P}[H_0]P_{FA} + \mathbb{P}[H_1]P_{MD} = 0.40872.$$

We can rewrite the maximum likelihood decision rule in terms of a ratio. Define the **likelihood ratio** as a function of the measurement value $Y = y$, as

$$\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)}.$$

The maximum likelihood decision rule can be written in terms of the likelihood ratio as

$$D^{ML}(y) = \begin{cases} 1, & \mathcal{L}(y) \geq 1, \\ 0, & \mathcal{L}(y) < 1. \end{cases}$$

We abbreviate this decision using this notation: $D^{ML}(y) = \{\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} 1\}$. This indicates that, when the inequality is in the "greater than" direction, the decision selected is that of hypothesis H_1, and when the inequality is reversed, the decision selected is that of hypothesis H_0.

We can often compute the maximum likelihood decision rule analytically using the expressions for the probability mass functions and the likelihood ratio. For Example 6.5, the likelihood ratio is

$$\mathcal{L}(y) = \frac{\binom{5}{y}(0.4)^{5-y}(0.6)^y}{\binom{5}{y}(0.5)^{5-y}(0.5)^y} = \frac{(0.4)^{5-y}(0.6)^y}{(0.5)^5} = 2^5(0.4)^{5-y}(0.6)^y = (0.8)^5(1.5)^y$$

We want to compare $\mathcal{L}(y)$ to 1. Therefore, the maximum likelihood detection rule is $\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} 1$.

To compute the performance of the maximum likelihood detector, we need to identify the values of $Y = y$ for which $D^{ML}(y) = 0$ and for which $D^{ML}(y) = 1$. When we enumerate the likelihoods for all values of $Y = y$ as in Example 6.5, this is straightforward. For larger R_Y, enumeration is impractical, so we need to further simplify the maximum likelihood decision rule to determine these regions.

To simplify this, we make the following observation: $\mathcal{L}(y) > 1 \iff \ln\big(\mathcal{L}(y)\big) > 0$. Computing the logarithm of the likelihood ratio $\mathcal{L}(y)$ yields $\ln\big(\mathcal{L}(y)\big) = 5\ln(0.8) + y\ln(1.5)$. Then,

$$\ln\big(\mathcal{L}(y)\big) > 0 \iff y > \frac{5\ln(1.25)}{\ln(1.5)} \approx 2.751.$$

Thus, for $y = 3, 4, 5$, the likelihood ratio $\mathcal{L}(y)$ is greater than 1, and for $y = 0, 1, 2$, the likelihood ratio is less than 1. This is the same maximum likelihood decision rule derived in Example 6.5.

Using logarithms often makes it easier to identify the decision rule in terms of a region of values of y, as we saw above. We can write the maximum likelihood decision rule in terms of the **log-likelihood ratio**, the logarithm of the likelihood ratio, as $D^{ML}(y) = \{\ln\left(\frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)}\right) \underset{H_0}{\overset{H_1}{\gtrless}} 0\}$.

Example 6.6

Radar systems usually send trains of pulses to detect the presence of aircraft in the direction the radar is aimed at. Each of these pulses potentially generates a reflection; for each pulse, a decision as to whether an aircraft is present or not can be made based on the received pulse signal strength, comparing it to a threshold. The final decision for detecting the presence of aircraft is based on the total number of pulses received that had sufficient signal strength. The detections on each pulse are assumed to be independent, conditioned on whether an aircraft is present or not.

Assume that the probability of detecting an aircraft in a single pulse, assuming the aircraft is present, is p_1. If the aircraft is not present, the probability of having enough background signal strength to generate a detection is p_0. Assume that n pulses get transmitted, and $p_1 > p_0$. What is the maximum likelihood detector?

The problem is stated in terms of two hypotheses: H_1 is where the aircraft is present, and H_0 is where there is no aircraft present. From the problem description, the observation Y consists of the number of pulses that generate a detection, which can take values in $\{0, 1, \ldots, n\}$. The likelihood $P_{Y|H_1}(y)$ is a Binomial(n, p_1) distribution, and the likelihood $P_{Y|H_0}(y)$ is a Binomial(n, p_0) distribution.

Since n, p_1, p_0 are left as variables, we cannot simply enumerate the possible values of Y in a table and find the best decision for each value of y. Nevertheless, we can analyze this using log-likelihood ratios, as:

$$\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} = \frac{\binom{n}{y}p_1^y(1-p_1)^{n-y}}{\binom{n}{y}p_0^y(1-p_0)^{n-y}} = \Big(\frac{1-p_1}{1-p_0}\Big)^n \Big(\frac{(p_1(1-p_0)}{p_0(1-p_1)}\Big)^y$$

$$\ln(\mathcal{L}(y)) = n\ln(\frac{1-p_1}{1-p_0}) + y\ln(\frac{(p_1(1-p_0)}{p_0(1-p_1)})$$

We see that the log-likelihood ratio is increasing in y (because $p_1 > p_0$, so $1 - p_1 < 1 - p_0$.) Furthermore for $y = 0$, the log-likelihood ratio is negative. Hence, there is a value of y for which the log-likelihood ratio equals 1. That value is

$$y^* = \frac{n\ln(1-p_0) - n\ln(1-p_1)}{\ln(p_1(1-p_0)) - \ln(p_0(1-p_1))}$$

For instance, if $p_1 = 0.7, p_2 = 0.2, n = 20$, we get $y^* \approx 8.78$, so the maximum likelihood detector declares a detection if 9 or more pulses are detected. Hence, $D^{ML}(y) = \{y \underset{H_0}{\overset{H_1}{\gtrless}} 8.78\}$, which is a simple detector to implement.

We can now compute the probabilities of missed detection and false alarm as sums, as

$$P_{MD} = \mathbb{P}[\{Y < y^*\}|H_1] = \sum_{y<y^*} \binom{n}{y} (p_1)^y (1-p_1)^{n-y}.$$

$$P_{FA} = \mathbb{P}[\{Y > y^*\}|H_0] = \sum_{y>y*} \binom{n}{y} (p_0)^y (1-p_0)^{n-y}.$$

For the values $p_1 = 0.7, p_2 = 0.2, n = 20$, we get $P_{MD}] \approx 0.005, P_{FA} \approx 0.010$, which shows that, even though single pulse detection is not very accurate, by sending 20 pulses we increase our performance to near-perfect detection.

For continuous observations Y, the maximum likelihood rule is expressed in terms of the likelihood ratio using the conditional probability densities $f_{Y|H_1}(y), f_{Y|H_0}(y)$. In this case, $\mathcal{L}(y) = \frac{f_{Y|H_1}(y)}{f_{Y|H_0}(y)}$, and the maximum likelihood decision rule is given as $D^{ML}(y) = \{\frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} 1\}$. For continuous random variables, enumerating the likelihood values for each y is no longer possible; to find the regions $A_1 = \{y \in R_Y : D^{ML}(y) = 1\}$ and $A_0 = \{y \in R_Y : D^{ML}(y) = 0\}$, we use the log-likelihood ratio to solve for the region.

Example 6.7
You are interested in diagnosing whether a person has a fever associated with a particular disease based on measuring their temperature. If the person does not have a disease, the measured temperature is expected to be a Gaussian random variable with mean 98.1 degrees Fahrenheit and standard deviation 1 degree Fahrenheit. If the person has the disease, the average temperature is 101 degrees Fahrenheit and standard deviation 1 degree Fahrenheit. What is the maximum likelihood detector? For the maximum likelihood detector, what are the probabilities of missed detection and false alarm?

Let H_1 be the event where the person has the disease, and H_0 the event where the person does not have the disease. The maximum likelihood detector is readily written in terms of the likelihood ratio as:

$$D^{ML}(y) = \left\{ \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-101)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-98.1)^2}{2}}} \underset{H_0}{\overset{H_1}{\gtrless}} 1. \right\}$$

To evaluate the performance, we use the log-likelihood ratio, which is

$$\ln \mathcal{L}(y) = \ln \left(\frac{e^{-\frac{(y-101)^2}{2}}}{e^{-\frac{(y-98.1)^2}{2}}} \right) = -\frac{(y-101)^2}{2} + \frac{(y-98.1)^2}{2}$$
$$= (101 - 98.1)y - \frac{101^2}{2} + \frac{98.1^2}{2} = (101 - 98.1)y - \frac{(101 - 98.1)(101 + 98.1)}{2}.$$

Equating this to 0, we get that $y^* = \frac{(101+98.1)}{2} = 99.55$, the average of the two expected values. If $y > y^*$, then $D^{ML}(y) = 1$, and if $y < y^*, D^{ML} = 0$. This is illustrated in the figure on the right, where the vertical blue line shows the value of y^*. To the right of that blue line, we have $f_{Y|H_1}(y) > f_{Y|H_0}(y)$. To the left, the inequality is reversed. We can now compute the performance as follows:

$$P_{FA} = \mathbb{P}[\{y \geq 99.55\}|H_0] = 1 - \Phi(1.45) = Q(1.45).$$

where the threshold $y^* = 99.55$ is 1.45 standard deviations higher than the average 98.1. Similarly,

$$P_{MD} = \mathbb{P}[\{y \geq 99.55\}|H_1] = \Phi(-1.45) = Q(1.45).$$

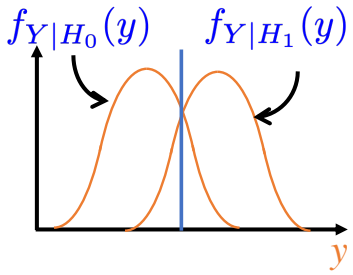


Figure 6.6: Example 6.7.

6.3 Maximum A Posteriori (MAP) Detection

In maximum likelihood detection, we designed the detection rule independent of the prior probabilities of each event hypothesis, $\mathbb{P}[H_0]$ and $\mathbb{P}[H_1]$. However, in many cases, the probabilities $\mathbb{P}[H_0]$ and $\mathbb{P}[H_1]$ can be very different. For instance, when testing for the presence of measles in a college-age student, the probability that the observed symptoms actually come from measles is small, as most college-age students have received an immunization vaccine. In this section, we show how to design detection algorithms that integrate this type of information.

Assume the measurements Y are discrete-valued, and we know $\mathbb{P}[H_0], \mathbb{P}[H_1]$. We refer to $\mathbb{P}[H_0], \mathbb{P}[H_1]$ as the prior probabilities, as they are known before measuring Y. After measuring Y, we compute the a posteriori or conditional probabilities of H_0 and H_1 given $Y = y$ using Bayes' Rule, as

$$\mathbb{P}[H_0|\{Y=y\}] = \frac{\mathbb{P}[H_0 \cap \{Y=y\}]}{\mathbb{P}[\{Y=y\}]} = \frac{\mathbb{P}[\{Y=y\}|H_0]\mathbb{P}[H_0]}{\mathbb{P}[\{Y=y\}]} = \frac{P_{Y|H_0}(y)\mathbb{P}[H_0]}{\mathbb{P}[\{Y=y\}]},$$

where the denominator is computed using the Law of Total Probability, as

$$\mathbb{P}[\{Y=y\}] = \mathbb{P}[\{Y=y\}|H_0]\mathbb{P}[H_0] + \mathbb{P}[\{Y=y\}|H_1]\mathbb{P}[H_1] = P_{Y|H_0}(y)\mathbb{P}[H_0] + P_{Y|H_1}(y)\mathbb{P}[H_1].$$

Similarly,

$$\mathbb{P}[H_1|\{Y=y\}] = \frac{\mathbb{P}[\{Y=y\}|H_1]\mathbb{P}[H_1]}{\mathbb{P}[\{Y=y\}]} = \frac{P_{Y|H_1}(y)\mathbb{P}[H_1]}{\mathbb{P}[\{Y=y\}]}.$$

The **maximum a posteriori (MAP)** decision rule is defined as follows:

$$D^{MAP}(y) = \begin{cases} 1, & \mathbb{P}[H_1|\{Y=y\}] \geq \mathbb{P}[H_0|\{Y=y\}], \\ 0, & \mathbb{P}[H_0|\{Y=y\}] > \mathbb{P}[H_1|\{Y=y\}]. \end{cases}$$

where we arbitrarily assign ties to 1. Since the denominator in Bayes' Rule is the same for $\mathbb{P}[H_1|\{Y=y\}]$ and $\mathbb{P}[H_0|\{Y=y\}]$, this rule is the same as

$$D^{MAP}(y) = \begin{cases} 1, & P_{Y|H_1}(y)\mathbb{P}[H_1] \geq P_{Y|H_0}(y)\mathbb{P}[H_0], \\ 0, & P_{Y|H_0}(y)\mathbb{P}[H_0]] > P_{Y|H_1}(y)\mathbb{P}[H_1]. \end{cases}$$

This allows us to rewrite the MAP decision rule in terms of the likelihood ratio, as

$$D^{MAP}(y) = \left\{ \mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \right\}.$$

Note that the data-dependent computation in the MAP decision rule is to compute the likelihood ratio, just as in the ML decision rule. What changes is the threshold that one compares the maximum likelihood to. In the ML case, the threshold is 1. This is also true in the MAP case if $\mathbb{P}[H_0] = \mathbb{P}[H_1]$. However, if $\mathbb{P}[H_1] > \mathbb{P}[H_0]$, the threshold is lower than 1, and the number of y for which the decision equals 1 is possibly increased. If $\mathbb{P}[H_0]$ is larger, then the threshold is larger than 1, and the number of y for which the decision equals 1 may be decreased.

Example 6.8

Assume we have the same problem as Example 6.5, but the prior probability that the coin is biased is only $\mathbb{P}[H_1] = 0.4$, so $\mathbb{P}[H_0] = 0.6$ because H_0, H_1 form a partition of Ω. From Example 6.5, we know the likelihoods of Y, the number of heads observed in 6 trials, are shown in the table below.

Y:	0	1	2	3	4	5
$P_{Y\|H_1}$	0.01024	0.0768	0.2304	0.3456	0.2592	**0.07776**
$P_{Y\|H_0}$	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125
$\mathcal{L}(y)$	0.3277	0.4915	0.7373	1.1059	**1.6589**	**2.4883**

We have added to the table a row computing the likelihood ratio for each value of Y. The threshold in the MAP decision rule is $\frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} = 1.5$. The values of $Y = y$ for which the likelihood ratio exceeds the threshold are highlighted in bold magenta above. We see that increasing the threshold has decreased the number of y for which the MAP decision is 1. The MAP decision rule and the ML decision rule from Example 6.5 are shown below:

To compute the maximum likelihood decision, we compare the numbers in each column, and pick the larger of the two numbers. In the table above, we have highlighted the larger number in bold and magenta color. Thus we see that the maximum likelihood decision rule becomes:

$$D^{MAP}(y) = \begin{cases} 1, & y = 4, 5, \\ 0, & y = 0, 1, 2, 3. \end{cases} \qquad D^{ML}(y) = \begin{cases} 1, & y = 3, 4, 5, \\ 0, & y = 0, 1, 2. \end{cases}$$

The decision agrees with intuition: a larger count of heads suggests the coin is more likely to be unbalanced, whereas a smaller count of heads indicates the coin is more likely to be balanced.

Since the ML and MAP decision rules are different, they have different performance. The probability of false alarm for the MAP decision rule is

$$P_{FA}^{MAP} = \mathbb{P}[\{Y = 4, 5\}|H_0] = P_{Y|H_0}(4) + P_{Y|H_0}(4) = 0.1875.$$

The probability of missed detection for the MAP decision rule is

$$P_{MD}^{MAP} = \mathbb{P}[\{Y = 0, 1, 2, 3\}|H_1] = P_{Y|H_1}(0) + P_{Y|H_1}(1) + P_{Y|H_1}(2) + P_{Y|H_1}(3) = 0.633.$$

In contrast, for the ML decision rule, $P_{FA}^{ML} = 0.5, P_{MD}^{ML} = 0.3174$. Thus, increasing the threshold reduced the probability of false alarm, and increased the probability of missed detection. The probability of error for each of the detectors is

$$P_e^{MAP} = \mathbb{P}[H_0]P_{FA}^{MAP} + \mathbb{P}[H_1]P_{MD}^{MAP} = 0.6 \cdot 0.1875 + 0.4 \cdot 0.633 \approx 0.3777.$$

$$P_e^{ML} = \mathbb{P}[H_0]P_{FA}^{ML} + \mathbb{P}[H_1]P_{MD}^{ML} = 0.6 \cdot 0.5 + 0.4 \cdot 0.3174 \approx 0.4270.$$

We will show later that the MAP decision rule achieves the minimum probability of error among all possible decision rules.

The MAP decision rule for continuous-valued measurements Y is a straightforward extension of the MAP decision rule for discrete-valued measurements Y. We have to be a bit careful to define $\mathbb{P}[H_0|\{Y = y\}]$ and $\mathbb{P}[H_1|\{Y = y\}]$ using a limiting argument, as in Chapter 4.4.3, because $\mathbb{P}[\{Y = y\}] = 0$. Specifically,

$$\begin{aligned}\mathbb{P}[H_0|\{Y \in (y, y+\Delta]\}] &= \frac{\mathbb{P}[H_0 \cap \{Y \in (y, y+\Delta]\}]}{\mathbb{P}[\{Y \in (y, y+\Delta)\}]} = \frac{\mathbb{P}[\{Y \in (y, y+\Delta]\}|H_0]\mathbb{P}[H_0]}{\mathbb{P}[\{Y \in (y, y+\Delta)\}]} \\ &= \frac{(F_{Y|H_0}(y+\Delta) - F_{Y|H_0}(y))\mathbb{P}[H_0]}{F_Y(y+\Delta) - F_Y(y)}\end{aligned}$$

As $\Delta \to 0$, both numerator and denominator approach 0. We use L'Hopital's rule to evaluate the limit, as

$$\lim_{\Delta\to 0} \mathbb{P}[H_0|\{Y \in (y, y+\Delta]\}] = \lim_{\Delta\to 0} \frac{\frac{d}{d\Delta}(F_{Y|H_0}(y+\Delta) - F_{Y|H_0}(y))\mathbb{P}[H_0]}{\frac{d}{d\Delta}(F_Y(y+\Delta) - F_Y(y))} = \frac{f_{Y|H_0}(y)\mathbb{P}[H_0]}{f_Y(y)} = \mathbb{P}[H_0|Y = y].$$

Similarly, $\mathbb{P}[H_1|Y = y] = \frac{f_{Y|H_1}(y)\mathbb{P}[H_1]}{f_Y(y)}$, and the marginal density is obtained by the Law of Total Probability as

$$f_Y(y) = f_{Y|H_0}(y)\mathbb{P}[H_0] + f_{Y|H_1}(y)\mathbb{P}[H_1].$$

This leads to the MAP decision rule in terms of the likelihood ratio

$$D^{MAP}(y) = \left\{ \mathcal{L}(y) = \frac{f_{Y|H_1}(y)}{f_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \right\}.$$

Example 6.9
The delay Y in arrival of an on-line order is modeled as an exponential random variable, but the rate of that random variable iis one of two possible rates. Under hypothesis H_1, the rate is 0.2/day, and under hypothesis H_0, the rate is 0.1/day. The prior probability that hypothesis H_0 is correct is $\mathbb{P}[H_0] = 0.6$. Assume we observe $Y = y$. What is the MAP decision rule, and what is its probability of error?

The threshold for the MAP decision rule for the probability of error is $T = \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} = \frac{3}{2}$. The likelihood ratio for the exponential random variables is

$$\mathcal{L}(y) = \frac{0.2e^{-0.2y}}{0.1e^{-0.1y}} = 2e^{-0.1y},$$

which is decreasing as y increases. Thus, longer observed delays y make hypothesis H_0 more likely, as its rate of arrival is smaller.

The boundary for the decision region in terms of y can be found by solving $\mathcal{L}(y) = 2e^{-0.1y} = \frac{3}{2}$. Taking logarithms,

$$-0.1y = \ln(3) - \ln(4) \Rightarrow y = 10(\ln(4) - \ln(3)) \approx 2.877.$$

Thus, if $y < 2.877$, select $D^{MAP}(y) = 1$; else, select $D^{MAP}(y) = 0$. With these regions, we have

$$P_{FA} = \int_0^{2.877} f_{Y|H_0}(y)\, dy = F_{Y|H_0}(2.877) = 1 - e^{-0.2877} = 0.25,$$

$$P_{MD} = \int_{2.877}^{\infty} f_{Y|H_1}(y)\,dy = e^{-0.2*2.877} = \frac{9}{16} = 0.5625.$$

The probability of error is

$$P_e = \mathbb{P}[H_0]P_{FA} + \mathbb{P}[H_1]P_{MD} = 0.6 * 0.25 + 0.4 * 0.5625 = 0.375.$$

We conclude this section by showing that the MAP decision rule minimizes the probability of error among all decision rules. For any decision rule $D(y)$, the probability of error conditioned on $Y = y$ is given as follows: Since H_0, H_1 are a partition of Ω,

$$\mathbb{P}[\text{Error}|Y = y] = \mathbb{P}[\text{Error} \cap H_1|Y = y] + \mathbb{P}[\text{Error} \cap H_0|Y = y].$$

$$\mathbb{P}[\text{Error} \cap H_1|Y = y] = \mathbb{P}[\text{Error}|Y = y, H_1]\mathbb{P}[H_1|Y = y];$$
$$\mathbb{P}[\text{Error} \cap H_0|Y = y] = \mathbb{P}[\text{Error}|Y = y, H_0]\mathbb{P}[H_0|Y = y],$$

which follows from the definition of conditional probability. Note that $\mathbb{P}[\text{Error}|Y = y, H_0] = I_{D(y)=1}$, where I_A is the indicator function that is 1 if A is true, and 0 elsewhere. Similarly, $\mathbb{P}[\text{Error}|Y = y, H_1] = I_{D(y)=0}$. Therefore,

$$\mathbb{P}[\text{Error}|Y = y] = I_{D(y)=1}\mathbb{P}[H_0|Y = y] + I_{D(y)=0}\mathbb{P}[H_1|Y = y].$$

Note that $D^{MAP}(y)$ selects the smallest of the two terms for each $Y = y$, and hence has the smallest probability of error for each $Y = y$. The unconditional probability of error is, assuming Y is discrete, as

$$P_e = \sum_{y \in R_Y} \Big(I_{D(y)=1}\mathbb{P}[H_0|Y = y] + I_{D(y)=0}\mathbb{P}[H_1|Y = y] \Big) P_Y(y),$$

which $D^{MAP}(y)$ will minimize because it minimizes each term in the sum.

For continuous Y, we get

$$P_e = \int_{y \in R_Y} \Big(I_{D(y)=1}\mathbb{P}[H_0|Y = y] + I_{D(y)=0}\mathbb{P}[H_1|Y = y] \Big) f_Y(y)\,dy,$$

which is minimized by $D^{MAP}(y)$ because $D^{MAP}(y)$ minimizes the integrand for every value of y, and hence it minimizes the integral.

6.4 Minimum Bayes Risk Detection

In many important situations, there is a different cost associated with the different types of errors. For instance, in luggage inspection, a false alarm can result in an unnecessary opening of a suitcase to check its contents. However, a missed detection can result in an explosive entering the airplane. In breast cancer diagnosis, a false alarm can lead to an unneeded biopsy, whereas a missed detection can be life-threatening.

To properly evaluate this tradeoff, we assign different costs to the different types of errors, and design a decision rule to minimize the total expected cost. Formally, let C_{ij} denote the cost of deciding U_i when H_j is true. We typically select $C_{11} = 0, C_{00} = 0$, so that correct decisions involve no cost; while this is not essential, it is wasted space to consider the full generality, as it is never used in practice. The key tradeoff is the relative cost of a missed detection C_{01} and a false alarm C_{10}. The Figure on the right illustrates the indexing as to what the costs mean for different values of decision and true hypothesis.

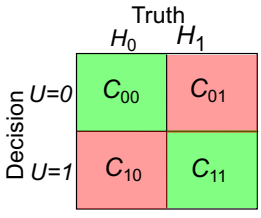


Figure 6.7: Bayes' Costs.

We follow closely the development in the previous section where we showed the MAP decision rule minimized P_e, the probability of making an error. For an arbitrary decision rule $D(y)$, let R denote the cost

of the decision rule. R is a random variable defined on the experiment, which depends on the outcome s and the observation y, as $R(s,y) = C_{01}I_{\{\omega\in H_1\}\cap\{D(y(\omega))=0\}} + C_{10}I_{\{\omega\in H_0\}\cap\{D(y(\omega))=1\}}$. Then, the conditional probability mass function of R given $Y=y$ and $D(y)=0$ is

$$P_{R|Y,\{D=0\}}(r|y) = \begin{cases}\mathbb{P}[H_1|Y=y], & r=C_{01}\\ \mathbb{P}[H_0|Y=y], & r=0.\end{cases} \quad P_{R|Y,\{D=1\}}(r|y) = \begin{cases}\mathbb{P}[H_0|Y=y], & r=C_{10}\\ \mathbb{P}[H_1|Y=y], & r=0.\end{cases}$$

Then,

$$\mathbb{E}[R|Y=y] = C_{01}\mathbb{P}[H_1|Y=y]I_{D(y)=0} + C_{10}\mathbb{P}[H_0|Y=y]I_{D(y)=1}.$$

The decision that minimizes this conditional expected risk given measurement $Y=y$ is

$$D^{MBR}(y) = \begin{cases}1, & C_{01}\mathbb{P}[H_1|Y=y] \geq C_{10}\mathbb{P}[H_0|Y=y]\\ 0, & C_{01}\mathbb{P}[H_1|Y=y] < C_{10}\mathbb{P}[H_0|Y=y].\end{cases}$$

For discrete random variables Y, the expected risk for any decision $D(y)$ is written as:

$$\mathbb{E}[R] = \sum_{y\in R_Y} \mathbb{E}[R|Y=y]P_Y(y).$$

Since the minimum Bayes risk (MBR) minimizes each term of the sum among all decision rules, it is the optimal decision rule for minimizing the expected Bayes risk. For continuous random variables Y, the expected Bayes risk of any decision rule is

$$\mathbb{E}[R] = \int_{y\in R_Y} \mathbb{E}[R|Y=y]f_Y(y)\,dy.$$

The MBR decision rule $D^{MBR}(y)$ minimizes the integrand for each y, and hence minimizes the expectation.

We can write D^{MBR} in terms of the likelihood ratio $\mathcal{L}(y)$ using Bayes' Rule: for discrete Y,

$$\mathbb{P}[H_1|Y=y] = \frac{P_{Y|H_1}(y)\mathbb{P}[H_1]}{P_Y(y)}; \quad \mathbb{P}[H_0|Y=y] = \frac{P_{Y|H_0}(y)\mathbb{P}[H_0]}{P_Y(y)}.$$

Recall that $\iff$ means "if and only if"; then,

$$\begin{aligned} C_{01}\mathbb{P}[H_1|Y=y] \geq C_{10}\mathbb{P}[H_0|Y=y] &\iff C_{01}P_{Y|H_1}(y)\mathbb{P}[H_1] \geq C_{10}P_{Y|H_0}(y)\mathbb{P}[H_0]\\ &\iff \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \geq \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]}\end{aligned}$$

Thus, the minimum Bayes risk decision rule is

$$D^{MBR}(y) = \left\{\frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]}\right\}.$$

For continuous measurements $Y=y$, the minimum Bayes risk decision rule is

$$D^{MBR}(y) = \left\{\frac{f_{Y|H_1}(y)}{f_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]}\right\}.$$

Note the following: The MAP decision rule is a special case of the MBR decision rule when $C_{10}=C_{01}$. The ML decision rule is another case of the MBR decision rule when $C_{10}=C_{01}, \mathbb{P}[H_1]=\mathbb{P}[H_0]$. In general, all MBR decision rules are based on comparing the likelihood ratio value for $Y=y$ to a threshold, where the threshold is computed from the relative costs and the prior probabilities of H_0, H_1.

The threshold varies with the relative cost of false alarms and missed detections in an intuitive manner. If missed detection are more expensive than false alarms, then the threshold for the likelihood ratio is set lower, so that one is more likely to decide that H_1 is the correct hypothesis.

Example 6.10
Consider Example 6.7, which sought to diagnose the presence of a disease by measuring the temperature. Assume a priori that the probability of having the disease ($\mathbb{P}[H_1]$) is 0.4, and thus the probability of not having the disease ($\mathbb{P}[H_0]$) is 0.6. However, the cost of a missed detection is 10 (C_{01}), whereas the cost of a false alarm (C_{10}) is 1. What is the minimum Bayes risk decision rule, and what are the resulting probabilities of false alarm and missed detection?

The MBR decision rule is

$$D^{MBR}(y) = \left\{ \mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]} = \frac{3}{20} \right\}.$$

From the results of Example 6.7, the likelihood ratio is

$$\mathcal{L}(y) = \frac{e^{-\frac{(y-101)^2}{2}}}{e^{-\frac{(y-98.1)^2}{2}}}$$

Then,

$$\mathcal{L}(y) \leq \frac{3}{20} \iff \ln\Big(\mathcal{L}(y)\Big) \leq \ln(3) - \ln(20)$$

$$\ln\Big(\mathcal{L}(y)\Big) = (101 - 98.1)y - \frac{(101 - 98.1)(101 + 98.1)}{2} = 2.9y - (2.9) \cdot (99.55)$$

$$D^{MBR}(y) = \left\{ y \underset{H_0}{\overset{H_1}{\gtrless}} 99.55 + \frac{1}{2.9}(\ln(3) - \ln(20)) = 99.2159 \right\}.$$

Thus, we see that the threshold for the decision rule has been lowered as compared to the ML decision rule of Example 6.7. This means that the probability of missed detection decreases, and the probability of false alarm increases. Since the mean under H_0 is 98.1, the threshold is 1.1159 standard deviations higher than the mean, so $P_{FA} = Q(1.1159)$. Since the mean under H_0 is 101, the threshold is 1.7841 standard deviations lower than the mean, so $P_{MD} = \Phi(-1.7841) = Q(1.7841)$.

The minimum expected Bayes risk is given in terms of these measures, as

$$\mathbb{E}[R] = \mathbb{P}[H_0]C_{10}P_{FA} + \mathbb{P}[H_1]C_{01}P_{MD} = 0.6P_{FA} + 4P_{MD} = 0.6Q(1.1159) + 4Q(1.841)$$

for the MBR decision rule, and

$$\mathbb{E}[R] = 0.6Q(1.45) + 4Q(1.45)$$

for the ML decision rule, which is higher than the MBR expected Bayes risk.

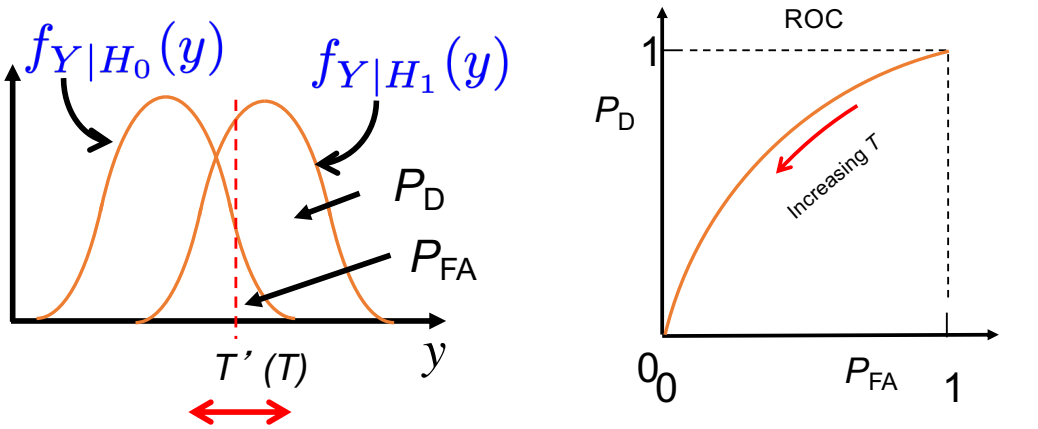
6.5 Performance and the Receiver Operating Characteristic

In the discussion so far, we have found that the optimal decision rule for binary hypotheses is a likelihood ratio test, where we compute a function of the measured data (the likelihood ratio) and compare it to a threshold. The choice of threshold depends on the prior probabilities of each hypotheses, plus the costs of making a missed detection. These four parameters are summarized in a single threshold T; to design an optimal decision rule, we simply select this threshold T, and the decision rule is

$$D(y) = \left\{ \mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} T \right\}.$$

The choice of threshold T controls the tradeoff between the conditional performance statistics P_{MD} and P_{FA}. As T increases, the decision rule selects H_1 less often, which increases P_{MD} and decreases P_{FA}.

Define the probability of detection $P_D = 1 - P_{MD}$. As the threshold T decreases to 0, the region of measurements $Y = y$ for which the decision is 1 increases, eventually becoming the entire range R_Y. When the threshold is 0, the performance statistics are $P_D = 1, P_{FA} = 1$, since the decision is always 1. Similarly, as the threshold increases to ∞, the region of measurements $Y = y$ for which the decision is 1 decreases, eventually becoming empty. For a threshold of ∞, the performance statistics are $P_D = 0, P_{FA} = 0$. As the threshold T is varied from 0 to ∞, we can trace a locus of performance of $P_D(T)$ versus $P_{FA}(T)$, which is called the **Receiver Operating Characteristic** or ROC for the detection problem. The design of an optimal decision rule based on likelihood ratios reduces to selecting a point on the ROC that trades off P_D versus P_{FA}. An illustration of a ROC is given in Figure 6.8.



Densities indicating decision regions Receiver Operating Characteristic (ROC)

Figure 6.8: Illustration of ROC for detection involving two Gaussian Distributions.

Let us emphasize some features of the ROC. First, note that the threshold T is a parameter along the curve. Thus any one point on the ROC corresponds to a particular choice of threshold (and vice versa). The ROC itself does not depend on the costs C_{ij} or the apriori probabilities $\mathbb{P}[H_i]$. These terms can be used to determine a particular threshold, and thus a particular operating point corresponding to the optimal Bayes risk detector. A couple of important properties of the ROC are:

- The ROC is monotone non-decreasing. Increasing P_{FA} results in increasing P_D.

- The ROC is a concave curve, with the graph above the $P_D = P_{FA}$ line. Performance on the line $P_D = P_{FA}$ correspond to detectors that that randomly guess $D(y) = 1$ with probability p, independent of the measured value y. The optimal detectors achieve better performance by using the information in y. This argument can be extended to show the ROC is a concave curve.

Determining the ROC requires computing the region $A_1(T) = \{y \in R_Y : \mathcal{L}(y) \geq T\}$ where the likelihood ratio decision rule results in decision 1 for threshold T. If we know that region, then $P_D(T) = \mathbb{P}[\{y \in A_1(T)\}|H_1], P_{FA}(T) = \mathbb{P}[\{y \in A_1(T)\}|H_0]$. By varying T, we obtain the points on the ROC. We discuss examples to show how this is done.

Example 6.11
We have a coin that may be biased so that the probability of Heads is 0.8 (Hypothesis H_1.) If the coin is unbiased, the probability of Heads is 0.5 (Hypothesis H_0.) We conduct three independent flips and count the number of heads as our measurement Y. The likelihoods and the likelihood ratio are shown in the table below:

Y:	0	1	2	3
$P_{Y\|H_1}$	0.0080	0.0960	0.3840	0.5120
$P_{Y\|H_0}$	0.1250	0.3750	0.3750	0.1250
$\mathcal{L}(y)$	0.0640	0.2560	1.0240	4.0960

We see that, for thresholds T above 4.1, the decision is always $D(y) = 0$, and so $P_D = 0, P_{FA} = 0$. For thresholds around $T = 1.1$, $D(y) = 1$ if $y = 3$, and 0 otherwise. Thus, $P_D = 0.5120, P_{FA} = 0.1250$. As we lower the threshold to between 0.26 and 1.02, $D(y) = 1$ for $y = 2, 3$ and 0 for $y = 0, 1$. Then, $P_D = 0.896, P_{FA} = 0.5$. We continue this and obtain the various points, plotted on the ROC figure on the right.

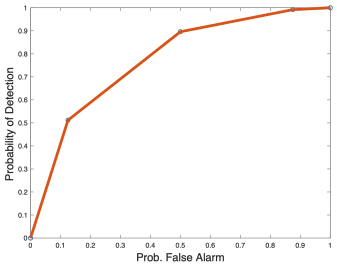


Figure 6.9: ROC for example.

Note that we have connected the discrete points in the ROC with straight lines. One can achieve performance on those straight lines by randomly switching between the thresholds corresponding to the two endpoints on the line. That type of random decision rule can be used to achieve a desired P_{FA} that is different from the finite ones obtained by the discrete breakpoints in the likelihood ratio table above.

Example 6.12
We have a light source that can either have an intensity of 100 photons/second, or 200 photons/second. We measure the number of photons emitted over a 1 second period, and have to decide which is the correct intensity for the light source. Let H_1 correspond to intensity of 120 photons/second, and H_0 correspond to intensity of 100 photons/second. If H_1 is correct, the number of photons measured is a Poisson(120) random variable; if H_0 is correct, the number is a Poisson(100) random variable.

The likelihood ratio for this problem is

$$\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} = \frac{\frac{120^y}{y!}e^{-120}}{\frac{100^y}{y!}e^{-100}} = (1.2)^y e^{-20}$$

An optimal likelihood ratio test is $\left\{\mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} T\right\}$ for a threshold T. Taking logarithms, of both sides, this reduces to

$$\ln\left(\mathcal{L}(y)\right) = y\ln(1.2) - 20 > \ln(T) \iff y > \frac{\ln(T) + 20}{\ln 1.2}.$$

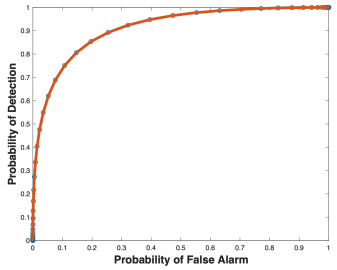


Figure 6.10: ROC for example.

For instance, for the ML decision rule, $T = 1$, and so the ML decision rule is $\{y \underset{H_0}{\overset{H_1}{\gtrless}} 109.7\}$. The ROC is shown in Figure 6.10 , where we have connected the discrete points in the ROC with straight lines.

Example 6.13 (Scalar Gaussian Detection)
Consider again the problem of determining which of two Gaussian densities of scalar observation comes from. In particular, suppose y is scalar and distributed $N(0, \sigma^2)$ under H_0 and distributed $N(m, \sigma^2)$ under H_1. The likelihood ratio is

$$\mathcal{L}(y) = \frac{\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-m)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y)^2}{2\sigma^2}}} = e^{-\frac{(y-m)^2}{2\sigma^2}+\frac{(y)^2}{2\sigma^2}}.$$

and the log-likelihood ratio is

$$\ln\Big(\mathcal{L}(y)\Big) = \frac{1}{2\sigma^2}(2my - m^2).$$

Hence, comparing the log-likelihood ratio to the log of a threshold T yields the decision rule

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{m}{2} + \frac{\sigma^2 \ln(T)}{m} = \Gamma.$$

From this, we can use the Gaussian likelihood formulas to obtain P_D and P_{FA} as:

$$P_D = 1 - \Phi(\frac{\Gamma - m}{\sigma}) = Q(\frac{\Gamma - m}{\sigma}); \quad P_{FA} = 1 - \Phi(\frac{\Gamma}{\sigma}) = Q(\frac{\Gamma}{\sigma}).$$

These calculations of P_D and P_F are illustrated in Figure 6.11.

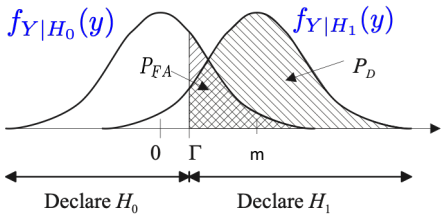


Figure 6.11: Illustration of P_D and P_{FA} calculation.

Example 6.14 (Gaussian detection with different variances)
Suppose y is scalar and distributed $N(0, \sigma_1^2)$ under H_0 and distributed $N(0, \sigma_0^2)$ under H_1. Assume $\sigma_1 < \sigma_0$. Thus, the Gaussians have the same mean, but different variances.

$$\mathcal{L}(y) = \frac{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(y)^2}{2\sigma_1^2}}}{\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y)^2}{2\sigma_0^2}}} = \frac{\sigma_1}{\sigma_0} e^{-\frac{y^2}{2\sigma_1^2} + \frac{y^2}{2\sigma_0^2}}.$$

The log-likelihood ratio is

$$\ln\Big(\mathcal{L}(y)\Big) = -\frac{y^2}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) + \ln(\sigma_1) - \ln(\sigma_0).$$

Hence, comparing the log-likelihood ratio to the log of a threshold T yields the decision rule

$$-y^2 \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\sigma_1^2\sigma_0^2}{\sigma_0^2 - \sigma_1^2}\Big(\ln(\sigma_0) - \ln(\sigma_1) + \ln(T)\Big) = \Gamma.$$

Note we were careful in dividing by numbers that are positive, so the sign of the inequalities was preserved. Unlike the case where the means were different, the detector is quadratic in the measurement. Since the density of Y under H_0 has larger variance, higher magnitudes of the measured y provide more support for hypothesis H_0. We can simplify the decision rule: In terms of y, we select H_1 if $|y| \leq \sqrt{\Gamma}$, otherwise, we select H_0. From this, we can use the Gaussian likelihood formulas to obtain P_D and P_{FA} as:

$$P_D = \mathbb{P}[\{|Y| \leq \sqrt{\Gamma}\}|H_1] = \Phi\left(\frac{\sqrt{\Gamma}}{\sigma_1}\right) - \Phi\left(-\frac{\sqrt{\Gamma}}{\sigma_1}\right).$$

$$P_{FA} = \mathbb{P}[\{|Y| \leq \sqrt{\Gamma}\}|H_1] = \Phi\left(\frac{\sqrt{\Gamma}}{\sigma_0}\right) - \Phi\left(-\frac{\sqrt{\Gamma}}{\sigma_0}\right)$$

The ROC can now be obtained by varying Γ from 0 to ∞. The ROC is shown in Figure 6.12.

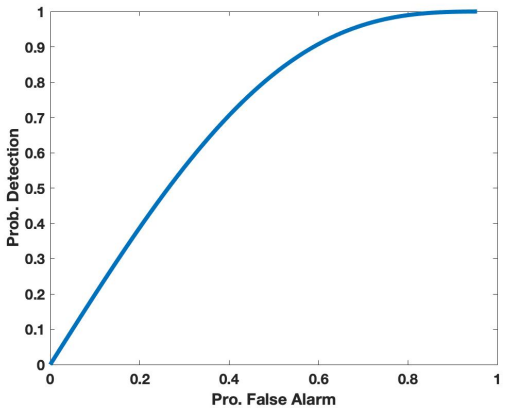


Figure 6.12: ROC for Gaussian hypotheses with different varlances.

6.6 Binary Hypothesis Testing with Vector Observations

The previous sections have assumed that the measurement Y is a scalar measurement, either discrete-valued or continuous valued. The previous theories developed extend completely to the case of pairs of measurements X, Y, or vector-valued measurements $\underline{Y}$. We briefly overview these extensions for pairs of measurements X, Y.

As before, we assume that the two hypotheses H_0, H_1 are events in $\mathcal{E}$ that form a partition of the sample space Ω. We assume that we observe a pair of random variables X, Y. If X, Y are discrete random variables, we assume that we are given the conditional joint probability mass functions $P_{X,Y|H_0}(x,y)$ and $P_{X,Y|H_1}(x,y)$. With this information, the likelihood ratio can be defined as a function of values $(x,y) \in R_{X,Y}$ by

$$\mathcal{L}(x,y) = \frac{P_{X,Y|H_1}(x,y)}{P_{X,Y|H_0}(x,y)}.$$

For jointly continuous measurements, the likelihoods are given by the conditional joint probability density functions $f_{X,Y|H_0}(x,y)$ and $f_{X,Y|H_1}(x,y)$. The likelihood ratio is defined as

$$\mathcal{L}(x,y) = \frac{f_{X,Y|H_1}(x,y)}{f_{X,Y|H_0}(x,y)}.$$

Once we have the likelihood ratios, the ML, MAP and MBR detectors are defined in identical manner to the scalar case:

$$D^{ML}(x,y) = \left\{ \mathcal{L}(x,y) \underset{H_0}{\overset{H_1}{\gtrless}} 1 \right\}.$$

$$D^{MAP}(x,y) = \left\{ \mathcal{L}(x,y) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \right\}.$$

$$D^{ML}(x,y) = \left\{ \mathcal{L}(x,y) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]} \right\}.$$

What is unique about the vector case is that the optimal decision rule depends only on a scalar function of the vector of observations X, Y. This holds true for higher-dimensional vectors: there is always a scalar function of the measurement vector $\underline{Y}$ that serves as a **sufficient statistic** to make an optimal decision.

The hard part of detection with vector observations is finding the decision regions so that we can compute performance metrics such as the probability of false alarm or the probability of missed detection. For pairs of random variables, we need to find the regions $\{(x,y) \in R_{X,Y} : D(x,y) = 0\}$ and $\{(x,y) \in R_{X,Y} : D(x,y) = 1\}$. For scalar measurements, we did this by analyzing the likelihood ratio test, and simplifying the equations to identify the regions. This is significantly harder for vector measurements, but there are special cases where we can do this.

We illustrate these extensions to vector observations with examples below.

Example 6.15

We are going to extend the diagnosis problem discussed in Example 6.7. The patient believes he has the flu. The hypothesis H_1 is the patient has the flu versus H_0 that the patient only has a cold. Let X be the measured temperature, and let Y be the results of a rapid influenza diagnostic test (RIDT) done on a mucus sample. We model the likelihood of X as a conditional Gaussian random variable with mean 98 degrees and standard deviation 2 degrees under H_0, and mean 102 degrees with standard deviation 2 degrees under H_1. The RIDT test is a color test, so we model the likelihood of Y in a very simple manner as a conditional Gaussian random variable (in the visible color spectrum) with mean wavelength 500 nm and standard deviation 100 nm under H_0, and mean wavelength 650 nm and standard deviation 100 nm under H_1. We assume that X, Y are conditionally independent given H_0, and also conditionally independent under H_1.

With the above information, we can now write the conditional joint probability density of (X,Y) given H_0 and H_1 as

$$f_{X,Y|H_0}(x,y) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-98)^2}{8}} \frac{1}{100\sqrt{2\pi}} e^{-\frac{(y-500)^2}{20000}}.$$

$$f_{X,Y|H_1}(x,y) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-102)^2}{8}} \frac{1}{100\sqrt{2\pi}} e^{-\frac{(y-650)^2}{20000}}.$$

The likelihood ratio is:

$$\begin{aligned}\mathcal{L}(x,y) &= \frac{f_{X,Y|H_1}(x,y)}{f_{X,Y|H_0}(x,y)} \\ &= \frac{\frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-102)^2}{8}} \frac{1}{100\sqrt{2\pi}} e^{-\frac{(y-650)^2}{20000}}}{\frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-98)^2}{8}} \frac{1}{100\sqrt{2\pi}} e^{-\frac{(y-500)^2}{20000}}} \\ &= e^{-\frac{(x-102)^2}{8}+\frac{(x-98)^2}{8}} e^{-\frac{(y-650)^2}{20000}+\frac{(y-500)^2}{20000}}\end{aligned}$$

Taking logarithms yields the log-likelihood ratio:

$$\begin{aligned}\ln\Big(\mathcal{L}(x,y)\Big) &= -\frac{(x-102)^2}{8} + \frac{(x-98)^2}{8} - \frac{(y-650)^2}{20000} + \frac{(y-500)^2}{20000} \\ &= \frac{(102-98)(2x-200)}{8} + \frac{150(2y-1150)}{20000} \\ &= x - \frac{4\cdot 200}{8} + \frac{3}{200}y - \frac{3\cdot 23}{8} \\ &= x + \frac{3}{200}y - \frac{869}{8}.\end{aligned}$$

The maximum likelihood detector compares the log-likelihood ratio to the threshold 0. This test becomes:

$$D^{ML}(x,y) = \Big\{x + \frac{3}{200}y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{869}{8}\Big\}.$$

The decision rule reduces to comparing a scalar statistic $x + \frac{3}{20}y$ to a threshold. This defines a region in x-y space where the decision is 0, and another region where the decision is 1, separated by the line $x + \frac{3}{200}y = \frac{869}{8}$. With this definition of decision regions, we can now do compute P_{FA} and P_{MD} as two-dimensionl integrals.

In this case, there is a simpler method for computing performance. Define the statistic $Z = X + \frac{3}{200}Y$ as a linear combination of X, Y. Z is a **sufficient statistic** for this problem, because the max-likelihood detector depends only on Z: $D^{ML}(x,y) = \Big\{z \underset{H_0}{\overset{H_1}{\gtrless}} \frac{869}{8}\Big\}$.

Since X, Y are jointly Gaussian conditioned on H_0, Z is a Gaussian random variable conditioned on H_0. Its conditional mean is $\mathbb{E}[Z|H_0] = \mathbb{E}[X|H_0] + \frac{3}{200}\mathbb{E}[Y|H_0] = 98 + 7.5 = 105.5$. Since X and Y are conditionally independent given H_0, we get

$$\mathsf{Var}[Z|H_0] = \mathsf{Var}[X|H_0] + \left(\frac{3}{200}\right)^2 \mathsf{Var}[Y|H_0] = 4 + \frac{9}{40000}10000 = 6.25.$$

Similarly, Z is Gaussian conditioned on H_1 with $\mathbb{E}[Z|H_1] = 102 + \frac{3}{200}650 = 111.75$, and $\mathsf{Var}[Z|H_1] = 6.25$. We write the ML detector in terms of Z as

$$D^{ML}(z) = \Big\{z \underset{H_0}{\overset{H_1}{\gtrless}} 108.625\Big\}.$$

and now we can analyze its performance the same way we did for a scalar Gaussian random variable decision rule. Thus,

$$P_{FA} = Q\Big(\frac{108.625 - 105.5}{\sqrt{6.25}}\Big); \qquad P_{MD} = Q\Big(\frac{111.75 - 105.5}{\sqrt{6.25}}\Big).$$

Example 6.16
Consider the radar detection example, where N independent pulses are sent out. However, instead of making a detection on each pulse return and counting the number of detections, we measure the signal strength of each return, so that a vector of signal strength measurements is collected. We assume that each pulse provides a measurement Y_i, where

$$Y_i = \begin{cases} W_i & \text{if hypothesis } H_0 \text{ is true (no target present)} \\ E + W_i & \text{if hypothesis } H_1 \text{ is true (target present).}\end{cases}$$

where E is a known constant, $W_i, i = 1, \ldots, N$ are independent, zero-mean Gaussian random variables with variance σ^2.

The above model results in a vector of observations $\underline{Y}$, where the components Y_i are jointly Gaussian and independent. Under hypothesis H_1, each Y_i has mean E and variance σ^2, whereas under hypothesis H_0, each Y_i has mean 0 and variance σ^2. In this case, the likelihood ratio is given by

$$\mathcal{L}(\underline{y}) = \frac{f_{\underline{Y}|H_1}(\underline{y})}{f_{\underline{Y}|H_0}(\underline{y})} = \prod_{i=1}^{N} \frac{e^{\frac{-(y_i-E)^2}{2\sigma^2}}}{e^{\frac{-(y_i)^2}{2\sigma^2}}} = \prod_{i=1}^{N} e^{\frac{2Ey_i - E^2}{2\sigma^2}} = e^{\frac{2E\left(\sum_{i=1}^{N} y_i\right) - NE^2}{2\sigma^2}}$$

Taking logs of both sides the decision rule can be reduced to:

$$\frac{1}{N}\sum_{i=1}^{N} y_i \underset{H_0}{\overset{H_1}{\gtrless}} \frac{E}{2} + \frac{\sigma^2 \ln(T)}{NE}$$

where T is the threshold used in the likelihood ratio test (e.g. 1 for the maximum likelihood detector.) In this case, note that a scalar sufficient statistic is $Z = \frac{1}{N}\sum_{i=1}^{N} Y_i$, which is a linear combination of $\underline{Y}$ and hence Gaussian conditioned on H_0 and on H_1. The mean of Z under H_1 is

$$\mathbb{E}[Z|H_1] = \mathbb{E}[\sum_{i=1}^{N} \frac{Y_i}{N}|H_1] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[Y_i|H_1 = \frac{1}{N}\sum_{i=1}^{N} E = E.$$

Similarly, the mean of Z under H_0 is 0. The variance of Z under both H_1 and H_0 is

$$\mathsf{Var}[Z|H_1] = \mathsf{Var}[\sum_{i=1}^{N} \frac{Y_i}{N}|H_1] = \sum_{i=1}^{N} \mathsf{Var}[\frac{Y_i}{N}|H_1] = \sum_{i=1}^{N} \frac{\sigma^2}{N^2} = \frac{\sigma^2}{N},$$

because the Y_i components are independent (and thus uncorrelated), so the variance of the sum is the sum of the variances of the individual components.

Thus, the effect of using N measurements is equivalent to using one measurement with variance reduced by a factor of $1/N$, thereby increasing the effective signal-to-noise ratio in the detector. Denote by $\Gamma = \frac{E}{2} + \frac{\sigma^2 \ln(T)}{NE}$ as the threshold used in the log-likelihood ratio test for Z. We can now compute the performance statistics as a function of this threshold using the Gaussian properties of Z, as

$$P_{FA} = \mathbb{P}[Z > \Gamma|H_0] = Q(\frac{\Gamma N^{\frac{1}{2}}}{\sigma}).$$

$$P_{MD} = \mathbb{P}[Z < \Gamma|H_1] = Q(\frac{(E-\Gamma)N^{\frac{1}{2}}}{\sigma}).$$

The effect of increasing N is to get a more accurate measurement. This means the performance of the detector, as captured in the ROC curve, improves. As $N \to \infty$, both P_{FA} and P_{MD} decrease to zero. The ROC for different values of N is illustrated in Figure 6.13.

6.7 M-ary Hypothesis Testing

The exposition so far has focused on binary hypothesis testing problems. When there are M possibilities or hypotheses, we term the problem an *M-ary detection or hypothesis testing problem.* We have M events in $(\Omega, \mathcal{E}, \mathbb{P})$, denoted as $H_i, i = 0, \ldots, M-1$, which are mutually exclusive and collectively exhaustive, so they form a partition of Ω. We assume there are measurements $\underline{Y}$ which are random vectors that provide the information for detection. If $\underline{Y}$ is discrete-valued, we are provided the conditional probability mass functions $P_{\underline{Y}|H_i}(\underline{y})$ for $i = 0, 1, \ldots, M-1$. If $\underline{Y}$ is a jointly continuous random vector, we are provided the conditional probability density functions $f_{\underline{Y}|H_i}(\underline{y})$ for $i = 0, 1, \ldots, M-1$.

A decision rule $D(\underline{y})$ is. function that maps each observed value $\underline{y}$ into $\{0, 1, \ldots, M-1\}$ where decision k means that hypothesis H_k is the selected hypothesis. The concepts for designing decision rules that we presented previously for binary hypothesis testing extend naturally to this case. For the maximum likelihood decision rule, we want to select $D(\underline{y}) = k$ whenever

$$P_{\underline{Y}|H_k}(\underline{y}) \geq P_{\underline{Y}|H_j}(\underline{y}), \text{ for all } j \neq i(\ \underline{y} \text{ discrete}).$$

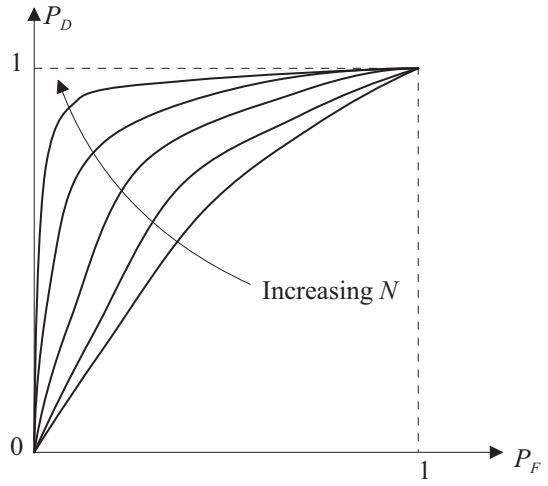


Figure 6.13: Illustration ROC behavior as we obtain more independent observations.

$$f_{\underline{Y}|H_k}(\underline{y}) \geq f_{\underline{Y}|H_j}(\underline{y}), \text{ for all } j \neq i(\ \underline{y} \text{ continuous}).$$

For the maximum a posteriori decision rule, we want to select $D(\underline{y}) = k$ whenever

$$\mathbb{P}[H_k|\underline{Y} = \underline{y}] \geq \mathbb{P}[H_j|\underline{Y} = \underline{y}], \text{ for all } j \neq i.$$

Equivalently, select $D(\underline{y}) = k$ whenever

$$P_{\underline{Y}|H_k}(\underline{y})\mathbb{P}[H_k] \geq P_{\underline{Y}|H_j}(\underline{y})\mathbb{P}[H_j], \text{ for all } j \neq i(\ \underline{y} \text{ discrete}).$$

$$f_{\underline{Y}|H_k}(\underline{y})\mathbb{P}[H_k] \geq f_{\underline{Y}|H_j}(\underline{y})\mathbb{P}[H_j], \text{ for all } j \neq i(\ \underline{y} \text{ continuous}).$$

As before, the MAP decision rule will minimize the average probability of error. If we defined costs C_{ij} associated with the cost of selecting decision U_i when hypothesis H_j is true, we can also define an equivalent theory for the minimum Bayes risk decision rule as in the binary hypothesis testing problems.

The biggest difference in the m-ary detection case is that there is no longer a sufficient scalar statistic like the likelihood ratio that we can compare to a threshold for optimal decision rules. Instead, the optimal decision rules must compute the M likelihoods, scale them appropriately, and pick the best decision on the basis of the resulting scaled values.

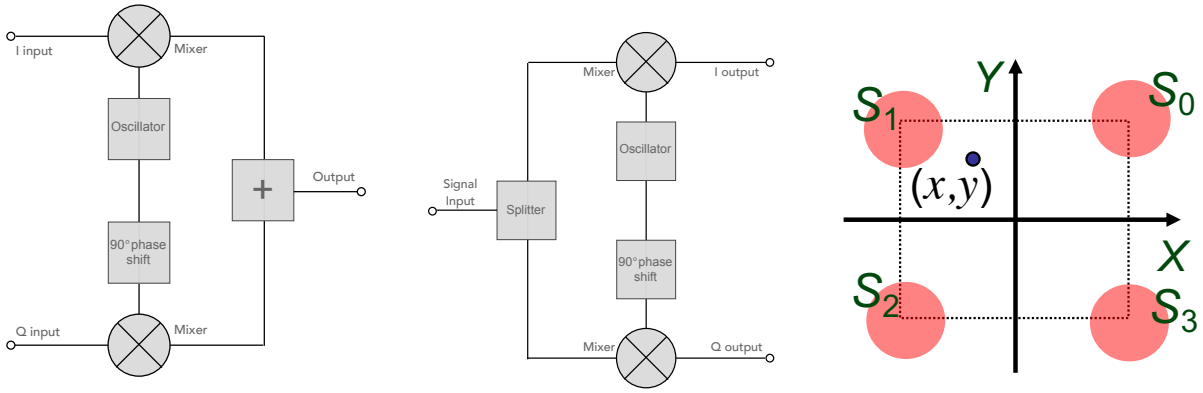
We illustrate m-ary detection problems with a couple of examples.

Example 6.17
Consider a communications problem where we try to communicate two bits at a time. We denote our two bits as pairs $A, B \in \{-1, 1\}$. We have four basic signals we are sending (1,1), (-1,1),(-1,-1),(1,-1), corresponding to hypotheses H_0, H_1, H_2, H_3 correspond to the transmitted symbols in this order.

To send the symbols, we use a variation of quadrature amplitude modulation, using short pulses of the form $s(t) = A\cos(\omega t) + B\sin(\omega t), t \in [0, T]$. A typical QAM modulation scheme is show in Figure 6.14, where the input I is the in-phase component, corresponding to the symbol A, and the input Q is the quadrature component, corresponding to the symbol B. The resulting transmitted pulse is $s(t) = A\cos(\omega t) + B\sin(\omega t), t \in [0, T]$

The signals propagate through the environment to a receiver, that demodulates the signal using a quadrature demodulation scheme, as shown in Figure 6.14. In the demodulator, the received signal is split, and multiplied each by $\cos(\omega t)$ and $\sin(\omega t)$. The in-phase output of the demodulator, $I(t)$, corresponds to the signal $s(t)\cos(\omega t)$, and the quadrature output $Q(t)$ corresponds to the signal $s(t)\sin(\omega t)$.

Note that $I(t) = A\cos^2(\omega t) + B\cos(\omega t)\sin(\omega t)$. Thus, averaging $I(t)$ over an interval of a few periods yields the output $A/2$, as the second term averages to 0. Similarly, $Q(t) = A\cos(\omega t)\sin(\omega t) + B\sin^2(\omega t)$, which averages to $B/2$. This



QAM modulator QAM demodulator Signal Constellation

Figure 6.14: Figures for Example **??**.

averaging is a low-pass filter and generates measurements of the transmitted A, B bits. Of course, these measurements are corrupted by background noise in the detector, signal corruption in the transmission channel, and small phase errors in the oscillators between the modulator and the demodulator.

The mathematical model of this detection problem is as follows: we measure two continuous random variables (X, Y), corresponding to the averages of the I and Q outputs of the demodulator. We model the statistics of these random variables as follows: We assume that X, Y are conditionally independent and Gaussian given any of the four hypotheses $H_i, i = 0, 1, 2, 3$. Furthermore, under each hypothesis, the variance of X is σ^2, and the variance of Y is also σ^2. However, the means change between hypotheses:

- Under H_0, $\mathbb{E}[X|H_0] \equiv m_x^0 = 1, \mathbb{E}[Y|H_0] \equiv m_y^0 = 1$.
- Under H_1, $\mathbb{E}[X|H_1] \equiv m_x^1 = -1, \mathbb{E}[Y|H_1] \equiv m_y^1 = 1$.
- Under H_2, $\mathbb{E}[X|H_2] \equiv m_x^2 = -1, \mathbb{E}[Y|H_2] \equiv m_y^2 = -1$.
- Under H_3, $\mathbb{E}[X|H_3] \equiv m_x^3 = 1, \mathbb{E}[Y|H_3] \equiv m_y^3 = -1$.

The signals are illustrated in Figure 6.14.

The likelihood under H_i is thus $f_{X,Y|H_i}(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-m_x^i)^2+(y-m_y)^2}{2\sigma^2}}$. To pick the largest one, we can compare the logarithms of the likelihoods, and subtract a common constant from all of them, to get a different comparison function $c^i(x, y)$ as

$$c^i(x, y) = \ln\left(f_{X,Y|H_i}(x, y)\right) - \ln\left(\frac{1}{2\pi\sigma^2}\right) = -\frac{(x - m_x^i)^2 + (y - m_y)^2}{2\sigma^2}.$$

We can scale $c^i(x, y)$ and subtract the same term to all i, to get

$$d^i(x, y) = 2\sigma^2 c^i(x, y) + \frac{x^2}{2} + \frac{y^2}{2} = m_x^i x + m_y^i y - \frac{(m_i^x)^2 + (m_i^y)^2}{2} = m_x^i x + m_y^i y - 1.$$

Every transformation we did above preserved the order of the likelihoods $f_{X,Y|H_i}(x, y)$. Hence, the maximum likelihood decision is

$$D^{ML}(x, y) = U_{i^*}, \text{ where } i^* \in \arg\max_{i=0,1,2,3} m_x^i x + m_y^i y - 1.$$

Note the -1 is not important. Then, we decide 0 when: $x + y > x - y; x + y > -x - y; x + y > -x + y$. Combine these inequalities, we get the region $x > 0, y > 0$. Thus, we decide 0 if we measure x, y in the first quadrant. Similarly, we decide 1 if the measurement (x, y) is in $x < 0, y > 0$, U_2 if the measurement is in $x, y < 0$, and U_3 if $x > 0, y < 0$.

We've simplified the decision rule so we could identify the decision regions in terms of the regions of the measurement range $R_{X,Y}$. We can use this to analyze the performance. Note the following: we can compute $\mathbb{P}[D^{ML}(X, Y) = 0|H_0] = \mathbb{P}[X \geq 0, Y \geq 0|H_0] = \mathbb{P}[X \geq 0|H_0]\mathbb{P}[Y \geq 0|H_0]$ because of the conditional independence of X, Y. Thus,

$$\mathbb{P}[D^{ML}(X, Y) = 0|H_0] = \Phi(\frac{1}{\sigma})\Phi(\frac{1}{\sigma}) = \Phi(\frac{1}{\sigma})^2.$$

This is the probability that we don't make an error when H_0 is the correct hypothesis. By symmetry, this is also $\mathbb{P}[D^{ML}(X,Y) = U_i|H_i], i = 1,2,3$. If all the hypotheses had equal prior probability $\mathbb{P}[H_i]$, the expected probability of correct decoding is $\Phi(\frac{1}{\sigma})^2$.

Example 6.18
Suppose we want to detect which of three possible N-dimensional signals $\underline{m}^k, k = 0,1,2$ is being received in the presence of noise. Under hypothesis H_k the observation $\underline{Y}$ is given by the vector Gaussian density:

$$f_{\underline{Y}|H_k}(\underline{y}) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_j - m_j^k)^2}{2}}.$$

This means each of the components of the observation $\underline{Y}$ is conditionally independent, Gaussian, with variance 1 and conditional expectation given by the components of $\underline{m}^k$.

Assume that we want a minimum probability of error decision rule, namely the MAP decision rule. Let $P_k = \mathbb{P}[H_k]$, the prior probabilities. The MAP decision rule picks

$$D^{MAP}(\underline{y}) = U_{i^*}, \text{ where } i^* \in \arg\min_{k=0,1,2} P_k \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_j - m_j^k)^2}{2}}.$$

We now have a valid decision rule, but the decision regions are hard to figure out, and this requires a lot of computation. We simplify the decision rules by taking transformations that preserve the order: we first compute log-likelihoods, and subtract common constants, to define

$$c^k(\underline{y}) = \ln\left(f_{\underline{Y}|H_k}(\underline{y})\right) + \ln(P_k) - N\ln(\frac{1}{\sqrt{2\pi}}) = \ln(P_k) - \sum_{j=1}^{n} \frac{(y_j - m_j^k)^2}{2}.$$

We can further simplify this by adding the same term to all the $c^k(\underline{y})$, as

$$d^k(\underline{y}) = c^k(\underline{y}) + \sum_{j=1}^{N} \frac{y_j^2}{2} = \ln(P_k) + (\underline{m}^k)^T\underline{y} - \frac{1}{2}(\underline{m}^k)^T\underline{m}^k,$$

where we have used vector notation for transposes. The terms $d^k(\underline{y})$ are referred to as discriminant functions; in this case, they are linear functions of $\underline{y}$, which help establish the regions.

Thus, the decision 0 is made whenever

$$\ln(P_0) + (\underline{m}^0)^T\underline{y} - \frac{1}{2}(\underline{m}^0)^T\underline{m}^0 > \ln(P_1) + (\underline{m}^1)^T\underline{y} - \frac{1}{2}(\underline{m}^1)^T\underline{m}^1,$$

$$\ln(P_0) + (\underline{m}^0)^T\underline{y} - \frac{1}{2}(\underline{m}^0)^T\underline{m}^0 > \ln(P_2) + (\underline{m}^2)^T\underline{y} - \frac{1}{2}(\underline{m}^2)^T\underline{m}^2.$$

Combining the $\underline{y}$ terms on the left side of the first equation, we get:

$$(\underline{m}^0 - \underline{m}^1)^T\underline{y} > \ln(P_1) - \ln(P_0) + \frac{1}{2}\left((\underline{m}^0)^T\underline{m}^0 - (\underline{m}^1)^T\underline{m}^1\right)$$

which defines a half-plane perpendicular to the line connecting $\underline{m}^0$ and $\underline{m}^1$. Working with the second equation yields

$$(\underline{m}^0 - \underline{m}^2)^T\underline{y} > \ln(P_2) - \ln(P_0) + \frac{1}{2}\left((\underline{m}^0)^T\underline{m}^0 - (\underline{m}^2)^T\underline{m}^2\right)$$

which is another half plane perpendicular to the line connecting $\underline{m}^0$ and $\underline{m}^2$. The intersection of the two half-planes is the region of $\underline{y}$ where we decide 0. A similar analysis can be done to determine the regions for 1 and U_2.

It is worth noting that, if the prior probabilities are all equal to 1/3, then the half-plane separating $\underline{m}^0$ and $\underline{m}^1$ goes through the midpoint of the line connecting $\underline{m}^0$ and $\underline{m}^1$. This is because, setting $\underline{y} = \frac{\underline{m}^0 + \underline{m}^1}{2}$, we get

$$(\underline{m}^0)^T\underline{y} - \frac{1}{2}(\underline{m}^0)^T\underline{m}^0 = (\underline{m}^0)^T\underline{m}^1.$$

$$(\underline{m}^1)^T\underline{y} - \frac{1}{2}(\underline{m}^1)^T\underline{m}^1 = (\underline{m}^0)^T\underline{m}^1.$$

Thus, this value of $\underline{y}$ is on the boundary of the decision regions between $0,1$. The resulting decision regions are illustrated in Figure 6.15 for a two-dimensional case. The decision boundaries are the bisectors of the lines connecting the means

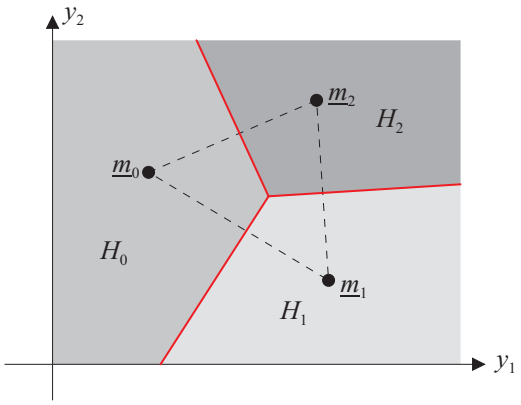


Figure 6.15: Illustration of the ML decision rule in the observation space.

under the different hypotheses. In general, this type of decision strategy is called a *nearest neighbor classifier* or a *minimum distance receiver* in the literature. Given the decision regions, we can now calculate performances, albeit with complicated integrals even in the case where we have conditionally independent measurements, because the decision regions are not parallel to the y_1, y_2 axes.

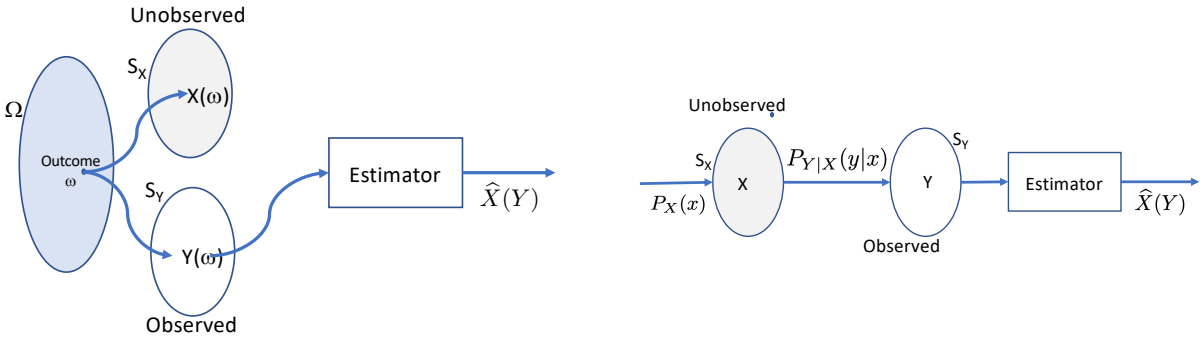
As a final comment in this Chapter, techniques such as nearest neighbor classifiers and linear discriminants are used extensively in data science and machine learning without much theoretical justification. In this Chapter, we have learned classes of statistical models for which nearest average classifiers and linear discriminants lead to optimal decision rules. We will use this to understand the hidden assumptions behind many classification methods in data science.

Chapter 7

Estimation

7.1 Introduction

In this chapter we consider the problem of estimating or inferring the values of unknown variables based on observation of a related set of random variables. A simple model of the estimation situation we are considering is depicted in Figure 7.1(a). An experiment generates pairs of random variables X, Y. We observe one of the two random variables, Y. Based on the observed value $Y = y$, we want to estimate the unobserved variable X by using an estimation rule $\widehat{x}(y)$. This model can be extended to cases where $\underline{X}, \underline{Y}$ are random vectors, so that several random variables are observed, and several unknown random variables are to be estimated.



(a) Jointly distributed random variables

(b) Sequential experiment

Figure 7.1: Different Views of Estimation Problem.

Assuming X, Y are discrete random variables, the probabilistic description of the variables X, Y is summarized by the joint probability mass function $P_{X,Y}(x, y)$, which we factor using the product rule as $P_X(x)P_{Y|X}(y|x)$. The second term in this factorization is the likelihood function, which captures the statistical relationship of how Y varies depending on the value of X. We can view this experiment as a sequential experiment, where the unobserved variable X is generated first, with probability law $P_X(x)$. Depending on the value of X, the observed variable Y is generated with probability law $P_{Y|X}(y|x)$. Figure 7.1(b) shows this sequential model, which is the one that we use in discussing estimation problems in this chapter.

This model has two components:

1. A model of the experiment that generates the unobserved random variable X, described by either $P_X(x)$ if X is discrete, or $f_X(x)$ if X is continuous.
2. A model of the observed random variable Y, represented by the conditional probability mass function $P_{Y|X}(y|x)$, if Y is discrete, or the conditional probability density function $f_{Y|X}(y|x)$ if y is continuous.

This model captures the essential elements of many problems in engineering and science, including: finding the location of a target based on radar observations, estimating the heart rate of a patient from electrical measurements, discerning O^+ density in the atmosphere from brightness measurements, and estimating depth in a scene from apparent motion in video.

The goal of estimation is to obtain an estimation rule that maps each observed value $y \in R_Y$ to a corresponding estimated value $\widehat{x}(y) \in \Re$. In some cases, we restrict the estimated value to be in R_X. The rule $\widehat{x}(y)$ is a function from R_Y into $\Re$. This is similar to the decision rules of the previous chapter. In hypothesis testing, a decision rule $D(y)$ mapped observations $Y = y$ into a discrete choice of hypothesis; the choice of decision rule depended on which criteria was used to design that decision rule $D(y)$. We will follow similar approaches for designing the estimation rule.

An important random variable in estimation is the estimation error $X - \widehat{x}(Y)$. This error is a random variable that depends on both X and Y. An estimator $\widehat{x}(y)$ is called **unbiased** in the Bayesian sense (or simply unbiased in the rest of this chapter) if

$$\mathbb{E}[X - \widehat{x}(Y)] = 0.$$

This implies that the error $X - \widehat{x}(Y)$ is an orthogonal random variable to the constant random variable 1. The bias of an estimator is known as $B = \mathbb{E}[X - \widehat{x}(Y)]$.

We note that there is a different concept of unbiased estimator in statistics, where X is not viewed as a random variable, but instead as an unknown constant. In statistics, an estimator $\widehat{x}(y)$ is called **unbiased** if

$$\mathbb{E}[\widehat{x}(Y)|X] = X \text{ for all values of } X.$$

This is a stronger requirement for unbiased estimation. In the remainder of this chapter we use the weaker concept of unbiased estimator in the Bayesian sense.

Another important statistic of an estimator is its mean-square error. The **mean-square error (MSE)** of an estimator is $\text{MSE} = \mathbb{E}\big[(X - \widehat{x}(Y))^2\big]$. We will use these statistics to design and characterize the performance of estimation rules.

7.2 Maximum Likelihood and Maximum A Posteriori Estimation

As was the case for hypothesis testing problems, we refer to the conditional distributions $P_{Y|X}(y|x)$ as likelihoods, because they are probability mass functions over Y, but they are general functions of $X = x$. Since we observe the value of $Y = y$, we are more interested in the properties of $P_{Y|X}(y|x)$ as functions of x, hence we use the term likelihood. For continuous random variables Y, the same applies to $f_{Y|X}(y|x)$, which are densities over Y, but general functions over $X = x$.

We define a **maximum likelihood estimator (ML)** $\widehat{x}_{ML}(y)$ as follows:

$$\widehat{x}_{ML}(y) \in \underset{x \in R_X}{\text{argmax}}\, P_{Y|X}(y|x), \quad Y \text{ discrete},$$
$$\widehat{x}_{ML}(y) \in \underset{x \in R_X}{\text{argmax}}\, f_{Y|X}(y|x), \quad Y \text{ continuous},$$

Since it is possible that the likelihood functions have multiple global maxima as a function of x, we use the set notation above to indicate that the ML estimator selects one of the global maxima of the likelihood functions.

The maximum likelihood estimator selects a value of $x \in R_X$ that maximizes the likelihood that the observation $Y = y$ was obtained, hence its name. It is similar to the maximum likelihood decision rule for hypothesis testing. The main difference is that, in binary hypothesis testing, selecting the maximum of two numbers is a straightforward operation. In contrast, selecting the maximum of a continuum of numbers (in case X is continuous) requires the use of optimization techniques involving calculus.

Similar to the ML estimator, we define a **maximum a posteriori (MAP)** estimator $\widehat{x}_{MAP}(y)$ as follows:

$$\begin{aligned}
\widehat{x}_{MAP}(y) &\in \operatorname*{argmax}_{x\in R_X} P(x)P_{Y|X}(y|x), \quad X,Y \text{ discrete},\\
\widehat{x}_{MAP}(y) &\in \operatorname*{argmax}_{x\in R_X} f(x)f_{Y|X}(y|x), \quad X,Y \text{ jointly continuous},\\
\widehat{x}_{MAP}(y) &\in \operatorname*{argmax}_{x\in R_X} f(x)P_{Y|X}(y|x), \quad Y \text{ discrete, } X \text{ continuous},\\
\widehat{x}_{MAP}(y) &\in \operatorname*{argmax}_{x\in R_X} P(x)f_{Y|X}(y|x), \quad Y \text{ continuous, } X \text{ discrete},
\end{aligned}$$

where we have added some extra cases to allow for the possibility that one of X, Y is discrete, while the other is continuous. For instance, in speech recognition, the features of a sound we hear (Y) have continuous values, but the set of possible phonemes (X) that generate that sound is discrete (around 64 phonemes in English). Similarly, in digital communications decoding, the transmitted symbols X are discrete, but the measured waveforms Y are continuous.

Example 7.1

The number of customers arriving at a service station when they open in the morning is modeled as a Binomial(4,0.5) random variable. No other customers arrive that day. Given that X customers arrive, the time Y hours to service their requests is modeled as an exponential random variable with parameter $\lambda(X) = \frac{1}{1+X}$. We come in the next day, and observe that $Y = 2$ for the previous day. We want to estimate the actual number of customers that arrived the previous day.

Note that this is a problem with continuous-valued measurements Y but discrete unknown X. Let's find both the ML and the MAP estimate of X given $Y = 2$. From the problem description, we know that $R_X = \{0, 1, 2, 3, 4\}$, and the likelihood function is known from the properties of exponential random variables, as

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1+x}e^{-\frac{y}{1+x}} & y \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

For $Y = 2$, the ML estimator will be

$$\widehat{x}_{ML}(y) \in \operatorname*{argmax}_{x\in\{0,1,2,3,4\}} f_{Y|X}(2|x) = \frac{1}{1+x}e^{-\frac{2}{1+x}}.$$

To find the maximum, we enumerate the values for each x :

x :	0	1	2	3	4
	0.135	0.184	0.171	0.152	0.134

Based on these numbers, $\widehat{x}_{ML}(2) = 1$.

From the problem description, we know $P_X(x) = \binom{4}{x}0.5^4$. Hence, the MAP estimator is

$$\widehat{x}_{MAP}(y) \in \operatorname*{argmax}_{x\in\{0,1,2,3,4\}} P_X(x)f_{Y|X}(2|x) = \binom{4}{x}0.5^4\frac{1}{1+x}e^{-\frac{2}{1+x}}.$$

To find the maximum, we enumerate the values for each x :

x :	0	1	2	3	4
	0.008	0.046	0.064	0.038	0.008

and we get that $\widehat{x}_{MAP}(2) = 2$. The difference arises because the prior probability that $X = 2$ is higher than that of $X = 1$.

Could we find the form of the estimators for arbitrary measurements Y? We do this for the ML estimator. In this case, it is possible, since all X does is decrease the service rate as X increases. Thus, for small Y, the best estimate is likely to be $X = 0$, and for large Y, it will be $X = 4$. The plot of the different densities $f_{Y|X}(y|x)$ is show in Figure 7.2 below. The figure illustrates when the different curves are maximal, and we can find those intervals by solving for the intersection points of the curves.

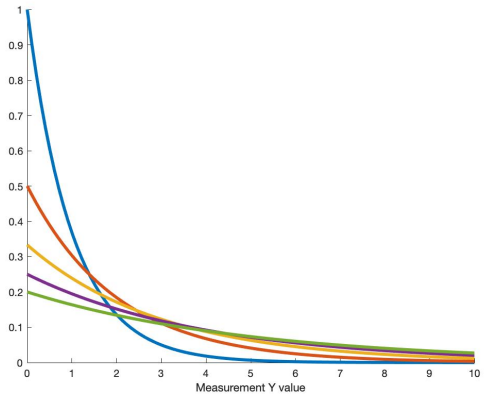


Figure 7.2: Plots of the different densities for different values of X.

We need to find the value of y where the likelihood function $f_{Y|X}(y|0) = f_{Y|X}(y|1)$, which is

$$e^{-y} = \frac{1}{2}e^{-y/2} \iff e^{y/2} = 2 \iff y = 2\ln(2).$$

Similarly, the value of y where the likelihood function $f_{Y|X}(y|1) = f_{Y|X}(y|2)$ is

$$\frac{1}{2}e^{-y/2} = \frac{1}{3}e^{-y/3} \iff e^{y/6} = \frac{3}{2} \iff y = 6\ln(\frac{3}{2}).$$

Similarly, the value of y where the likelihood function $f_{Y|X}(y|2) = f_{Y|X}(y|3)$ is

$$\frac{1}{3}e^{-y/3} = \frac{1}{4}e^{-y/4} \iff e^{y/12} = \frac{4}{3} \iff y = 12\ln(\frac{4}{3}).$$

Finally, the value of y where the likelihood function $f_{Y|X}(y|3) = f_{Y|X}(y|4)$ is

$$\frac{1}{4}e^{-y/4} = \frac{1}{5}e^{-y/5} \iff e^{y/20} = \frac{5}{4} \iff y = 20\ln(\frac{4}{3}).$$

Hence, the ML estimator is

$$\widehat{x}_{ML}(y) = \begin{cases} 0, & y \in [0, 2\ln(2)], \\ 1, & y \in (2\ln(2), 6\ln(3/2)], \\ 2, & y \in (6\ln(3/2), 12\ln(4/3)], \\ 3, & y \in (12\ln(3/4)], 20\ln(5/4)] \\ 4, & y > 20\ln(5/4). \end{cases}$$

Example 7.2
One of the most useful applications of estimation is in estimating the parameters of an unknown probability distribution from observed samples. Let X be a random variable in [0,1], with density $f_X(x) = 2x, x \in [0,1]; 0$ otherwise. Given $X = x$, let Y be a Binomial(N, x) random variable. This corresponds to the following scenario. We have a coin with unknown probability of heads X, as a number between 0 and 1. To estimate X, we flip this coin N times and count the number of heads (Y). Now we want to estimate the original unknown probability X given the number of heads observed (Y out of N).

A quick estimator might be the fraction of heads: $\widehat{x}(y) = \frac{y}{N}$. What is the maximum likelihood estimator? From the problem description,

$$P_{Y|X}(y|x) = \binom{N}{y}(x)^y(1-x)^{N-y}.$$

$$\widehat{x}_{ML}(y) = \underset{x\in[0,1]}{\operatorname{argmax}} \binom{N}{y}(x)^y(1-x)^{N-y}.$$

To simplify this, we maximize the log-likelihood, which has the maximum in the same locations as the likelihood, because the logarithm is a monotone increasing function for positive numbers (e.g. likelihoods).

$$\widehat{x}_{ML}(y) = \underset{x\in[0,1]}{\operatorname{argmax}} \ln(\binom{N}{y}) + y\ln(x) + (N-y)\ln(1-x).$$

To maximize, take the derivative with respect to x and set it equal to 0, as

$$\frac{d}{dx}\Big(\ln(\binom{N}{y}) + y\ln(x) + (N-y)\ln(1-x)\Big) = \frac{y}{x} - \frac{N-y}{1-x} = 0.$$

Solving for x will give us the estimator $\widehat{x}_{ML}(y) = \frac{y}{N}$, which can be verified by substituting into the above equation and checking it solves it. Thus, the ML estimator is the fraction of heads out of N trials.

What is the MAP estimator?

$$\widehat{x}_{MAP}(y) = \operatorname*{argmax}_{x\in[0,1]} 2x\binom{N}{y}(x)^y(1-x)^{N-y}.$$

Taking logarithms,

$$\widehat{x}_{MAP}(y) = \operatorname*{argmax}_{x\in[0,1]} \ln(2) + \ln(x) + \ln(\binom{N}{y}) + y\ln(x) + (N-y)\ln(1-x).$$

Differentiating with respect to x:

$$\frac{d}{dx}\Big(\ln(2) + \ln(x) + \ln(\binom{N}{y}) + y\ln(x) + (N-y)\ln(1-x).\Big) = \frac{y+1}{x} - \frac{N-y}{1-x} = 0.$$

Solving,

$$(y+1)(1-x) - (N-y)x = 0 \iff y+1-x-Nx = 0 \iff x = \frac{y+1}{N+1},$$

so $\widehat{x}_{MAP}(y) = \frac{y+1}{N+1}$, which is a little larger than the ML estimator.

Are these estimates unbiased? Note the following:

$$\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|X]\big] = \mathbb{E}[NX] = N\mathbb{E}[X],$$

because Y is distributed as a Binomial(N, X) random variable. Then, for the ML estimator,

$$\mathbb{E}[X - \frac{Y}{N}] = \mathbb{E}[X] - \frac{\mathbb{E}[Y]}{N} = \mathbb{E}[X] - \mathbb{E}[X] = 0,$$

so the ML estimator is unbiased. For the MAP estimator,

$$\mathbb{E}[X - \frac{Y+1}{N+1}] = \mathbb{E}[X] - \frac{N\mathbb{E}[X]+1}{N+1} = \frac{E[X]-1}{N+1} \neq 0,$$

so the MAP estimator is biased.

Example 7.3
We have a receiver, at a distance X meters from a transmitter. The transmitter transmits a signal with power 100, and the signal decays as $\frac{1}{X^2}$ to reach the receiver so the nominal received signal $S = \frac{100}{X^2}$. The receiver measures the signal strength in decibels, and the signal in decibels has some noise in it. The measured signal Y is given as

$$Y = 40 - 40\log_{10}(X) + W$$

where W is a Gaussian random variable with mean 0 and variance 4, independent of X. The prior distribution of X is

$$f_X(x) = \begin{cases} \frac{2x}{10^6} & 0 < x \leq 10^3, \\ 0 & \text{otherwise.} \end{cases}$$

Compute the ML and MAP estimators of X given observations $Y = y$.

From the problem description,

$$f_{Y|X}(y|x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(y+40\log_{10}(x))^2}{8}}.$$

The ML estimator is

$$\widehat{x}_{ML}(y) = \operatorname*{argmax}_{x\in(0,1000]} \frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}.$$

As was the case with detection problems, it is often easier to maximize the log-likelihood, since the maximum of the log-likelihood is in the same location as the maximum of the likelihood. Thus,

$$\widehat{x}_{ML}(y) = \underset{x\in(0,1000]}{\operatorname{argmax}} \ln\Big(\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\Big) = C - \frac{(y-40+40\log_{10}(x))^2}{8},$$

where C is a constant that does not depend on x, and so it won't affect the location of the maximizing x.

We try to find the maximum by differentiating and setting the derivative equal to 0:

$$\frac{d}{dx}(y-40+40\log_{10}(x))^2 = 2((y-40+40\log_{10}(x))\frac{40}{x\ln(10)} = 0.$$

Eliminating constants yields the following equation:

$$(y-40+40\log_{10}(x)) = 0 \iff x = 10^{\frac{40-y}{40}}.$$

Note that this is not the ML estimator yet, because it is possible that this value of X is greater than 1000. If it is, the best estimate is to set $x = 1000$. The ML estimator is thus

$$\widehat{x}_{ML}(y) = \begin{cases} 10^{\frac{40-y}{40}}, & y >= -80, \\ 10^3 & y < -80. \end{cases}$$

What about the MAP estimator? It is

$$\widehat{x}_{MAP}(y) = \underset{x\in(0,1000]}{\operatorname{argmax}} \frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}.$$

Taking logarithms as before yields

$$\widehat{x}_{MAP}(y) = \underset{x\in(0,1000]}{\operatorname{argmax}} \ln\Big(\frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\Big) = C_1 - \frac{(y-40+40\log_{10}(x))^2}{8} + \ln(x).$$

Differentiating with respect to x, multiplying by -1 and setting it to 0 yields

$$2((y-40+40\log_{10}(x))\frac{40}{x\ln(10)} - \frac{1}{x} = 0 \iff y-40+40\log_{10}(x) - \frac{\ln(10)}{80} = 0.$$

We see the effect of the a priori information on the MAP estimator. It increases the estimated distance. The ML estimator assumes that X is uniformly distributed in $(0, 10^3)$ with a density that does not depend on X. The MAP estimator has more probability for larger values of X. Thus,

$$\widehat{x}_{MAP}(y) = \begin{cases} 10^{\frac{40+\frac{\ln(10)}{80}-y}{40}}, & y >= -80 + \frac{\ln(10)}{80}, \\ 10^3 & \text{elsewhere.} \end{cases}$$

Example 7.4
Assume that X, Y are joint Gaussian random variables, with means μ_X, μ_Y and variances σ_X^2, σ_Y^2 respectively, and with given covariance $\mathrm{Cov}[X,Y]$. Using the results of Chapter 5.4, we know that we can write the joint density of X, Y as

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

where the conditional density $f_{X|Y}(x|y)$ is Gaussian, with mean $\mu_X + \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[Y]}(y-\mu_Y)$, and variance $\sigma_X^2 - \frac{\mathrm{Cov}[X,Y]^2}{\sigma_Y^2}$.

Then, the MAP estimator of X given Y is

$$\widehat{x}_{MAP}(y) \in \arg\max_{x\in(-\infty,\infty)} f_{X|Y}(x|y) = \mu_X + \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[Y]}(y-\mu_Y),$$

because the maximum of a Gaussian density is at its mean. Hence, in this Gaussian case,

$$\widehat{x}_{MAP}(y) = \mathbb{E}[X \mid Y = y].$$

Let us close this section by summarizing what we have learned about ML and MAP estimates estimates:

- ML estimates are equal to MAP estimates when the marginal density or distribution of the unobserved variable X does not depend on x.
- The ML estimate is the maximizing value for the likelihood: $\text{argmax}_x f_{Y|X}(y \mid x)$.
- The MAP estimate is the conditional mode: $\text{argmax}_x f_{X|Y}(x|y)$.
- The MAP and ML estimates may be biased.
- The MAP and ML estimates may not be unique.
- In general, the MAP and ML estimates are nonlinear functions of the observation.
- For jointly Gaussian problems, the MAP estimate is the same as the conditional mean, and in this case is a linear estimate and MMSE.
- In general, finding the MAP or ML estimate requires finding the maximum of the conditional density or likelihood, which may be a difficult problem.

7.3 Minimum Mean Square Error Estimation

For any estimator $\widehat{x}(y)$, the error $X - \widehat{x}(Y)$ is a random variable. The mean square error is MSE $= \mathbb{E}[(X-\widehat{x}(Y))^2]$. We want to find the estimator $\widehat{x}(y)$ that results in the minimum mean-square error (MMSE). We refer to this estimator as $\widehat{x}_{MMSE}(y)$. As before, assume we are given either a PMF $P_X(x)$ or PDF $f_X(x)$, depending on whether X is discrete or continuous as a random variable. Furthermore, assume we know the likelihoods $P_{Y|X}(y|x)$ or $f_{Y|X}(y|x)$, depending on whether Y is discrete or continuous.

Consider first the special case where Y is discrete and $P_{Y|X}(0|x) = 1$ for all $x \in R_X$. In simple words, the measurement $Y = 0$ always happen, no matter what X is. In this case, any estimator must be a constant: $\widehat{x}(0) = a$. What is the best choice of constant a to minimize the mean square error? Define $\mu_X = \mathbb{E}[X]$. Consider the following identity:

$$\begin{aligned}\mathbb{E}[(X-a)^2] &= \mathbb{E}[(X-\mu_X+\mu_X-a)^2] = \mathbb{E}[(X-\mu_X)^2 + 2(X-\mu_X)(\mu_X-a) + (\mu_X-a)^2] \\ &= \mathbb{E}[(X-\mu_X)^2] + 2\mathbb{E}[(X-\mu_X)](\mu_X-a) + (\mu_X-a)^2 \\ &= \mathsf{Var}[X] + (\mu_X-a)^2\end{aligned}$$

Since the last term is non-negative, and is zero when $a = \mu_X$, this means that $\widehat{x}_{MMSE}(0) = \mu_X = \mathbb{E}[X]$.

Let's now derive the MMSE estimator for general forms of likelihood. We will use the Law of Total Expectation, with iterated expectations, as follows:

$$\mathbb{E}[(X-\widehat{x}(Y))^2] = \mathbb{E}\Big[\mathbb{E}[(X-\widehat{x}(Y))^2|Y]\Big]$$

Let's focus on the inner expectation:

$$\begin{aligned}\mathbb{E}[(X-\widehat{x}(Y))^2|Y] &= \mathbb{E}[X^2 - 2X\widehat{x}(Y) + (\widehat{x}(Y))^2|Y] \\ &= \mathbb{E}[X^2|Y] - 2\mathbb{E}[X|Y]\widehat{x}(Y) + (\widehat{x}(Y))^2 \quad \text{[functions of } Y \text{ are conditionally constant]} \\ &= \mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2 + (\mathbb{E}[X|Y])^2 - 2\mathbb{E}[X|Y]\widehat{x}(Y) + (\widehat{x}(Y))^2 \quad \text{[add, subtract same]} \\ &= \mathsf{Var}[X|Y] + (\mathbb{E}[X|Y] - \widehat{x}(Y))^2 \quad \text{[factor into a square]}\end{aligned}$$

The last term is non-negative, and is zero only when $\widehat{x}(Y) = \mathbb{E}[X|Y]$. Any other estimator will have a larger conditional mean square error, and thus will also have a larger unconditional MSE. Thus, the MMSE estimator is

$$\widehat{x}_{MMSE}(y) = \mathbb{E}[X|Y = y].$$

Note that our derivation of this did not depend on whether X or Y was continuous or discrete. This is a strong result: the estimator that results in the smallest mean square error is the conditional mean of X given observation of Y. Note also that, when X is discrete, $\widehat{x}_{MMSE}(y)$ is a real number, and may not belong to R_X.

The MMSE estimator has some interesting properties, discussed below:

- The MMSE estimator is unbiased. That is, $\mathbb{E}\big[X - \mathbb{E}[X|Y]\big] = 0$. This follows from the Law of Total Expectation, because

$$\mathbb{E}\Big[X - \mathbb{E}[X|Y]\Big] = \mathbb{E}[X] - \mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[X] - \mathbb{E}[X] = 0.$$

- The error $X - \mathbb{E}[X|Y]$ is orthogonal to any random variable $Z = g(Y)$, for any bounded function Y. Again, this is a function of the Law of Total Expectation,(as

$$\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y)] = \mathbb{E}\Big[\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y) \mid Y]\Big] = \mathbb{E}\Big[\mathbb{E}[X - \mathbb{E}[X|Y] \mid Y]g(Y)\Big],$$

  because $g(Y)$ is known if Y is observed, Then, $\mathbb{E}[X - \mathbb{E}[X|Y] \mid Y] = \mathbb{E}[X - \mathbb{E}[X|Y]|Y] = \mathbb{E}[X|Y] - \mathbb{E}[X|Y] = 0$. Substituting into the above equation yields the orthogonality property.

- The estimator $\widehat{x}_{MMSE}(Y)$ is orthogonal to the error $X - \mathbb{E}[X|Y]$, by the above property, because it is a function of Y.

The main limitation in computing the MMSE estimator is that one needs to compute the expected value of X given $Y = y$. This can be hard to do. Furthermore, for discrete X, the MMSE estimate may not be in the range R_X. We revisit a couple of our earlier examples to illustrate this.

Example 7.5
Consider Example 7.2, where

$$P_X(x) = \binom{4}{x} 0.5^4,$$

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1+x} e^{-\frac{y}{1+x}} & y \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Then, using Bayes' Rule, we get

$$P_{X|Y}(x|y) = \frac{\frac{1}{1+x} e^{-\frac{y}{1+x}} \binom{4}{x} 0.5^4}{f_Y(y)}$$

where $f_Y(y) = \sum_{x=0}^{4} \frac{1}{1+x} e^{-\frac{y}{1+x}} \binom{4}{x} 0.5^4$. Thus,

$$\begin{aligned} \widehat{x}_{MMSE}(y) = \mathbb{E}[X|Y=y] &= \frac{1}{f_Y(y)} \sum_{x=0}^{4} \frac{x}{1+x} e^{-\frac{y}{1+x}} \binom{4}{x} 0.5^4 \\ &= \frac{\frac{1}{2}\binom{4}{1}e^{-y/2} + \frac{2}{3}\binom{4}{2}e^{-y/3} + \frac{3}{4}\binom{4}{3}e^{-y/4} + \frac{4}{5}\binom{4}{4}e^{-y/5}}{e^{-y} + \frac{1}{2}\binom{4}{1}e^{-y/2} + \frac{1}{3}\binom{4}{2}e^{-y/3} + \frac{1}{4}\binom{4}{3}e^{-y/4} + \frac{1}{5}\binom{4}{4}e^{-y/5}}. \end{aligned}$$

Note the complexity of the MMSE estimator. Fortunately, since X only takes five possible values, we are able to write the terms in each sum. For the observation $Y = 2$, the MMSE estimator is $\widehat{x}_{MMSE}(2) \approx 1.95$, which is different from $\widehat{x}_{ML}(2) = 1$ and $\widehat{x}_{MP}(2) = 2$. this also highlights that the MMSE estimate of a discrete random variable can be a real number, whereas the MAP and ML estimates are restricted to elements in R_X, which are integers.

Example 7.6
Consider Example 7.3, where

$$f_X(x) = \begin{cases} \frac{2x}{10^6} & 0 < x \leq 10^3, \\ 0 & \text{otherwise.} \end{cases}$$

$$f_{Y|X}(y|x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}.$$

We use Bayes' Rule to compute:

$$f_{X|Y}(x|y) = \begin{cases} \frac{\frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}}e^{-\frac{(y+40\log_{10}(x))^2}{8}}}{f_Y(y)} & 0 < x \le 10^3, \\ 0 & \text{otherwise.} \end{cases}$$

where

$$f_Y(y) = \int_0^{1000} \frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\, dx$$

Then,

$$\widehat{x}_{MMSE}(y) = \frac{\int_0^{1000} x\frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\, dx}{\int_0^{1000} \frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\, dx}.$$

Although we can write the integrals for computing $\widehat{x}_{MMSE}(y)$, these integrals are hard to compute exactly, illustrating the computational complexity of computing the MMSE estimators. In contrast, the MAP and ML estimators were easy to compute in closed form.

Example 7.7

Here we consider a simple example of a joint density where we can compute the conditional expected value needed. Let X, Y be jointly continuous random variables with joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} 2(x+y) & 0 \le x \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx = \begin{cases} \int_0^y 2(x+y)\, dx = 3y^2, & y \in [0,1] \\ 0 & \text{otherwise,} \end{cases}$$

and the conditional density is

$$f_{X,Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{2(x+y)}{3y^2} & 0 \le x \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\widehat{x}_{MMSE}(y) = \mathbb{E}[X|Y=y] = \frac{1}{3y^2}\int_0^y x2(x+y)\, dx = \frac{1}{3y^2}(\frac{2}{3}y^3 + y^3) = \frac{5}{9}y.$$

In this case, the integrals involved simple polynomials, so we could compute the needed expectations. The MMSE estimator turns out to be a linear function of y. We can also compute the MAP estimator as $\widehat{x}_{MAP}(y) \in \operatorname{argmax}_{x\in[0,y]} f_{X,Y}(x,y) = \operatorname{argmax}_{x\in[0,y]} 2(x+y)$. Hence, $\widehat{x}_{MAP}(y) = y$ for all $y \in [0,1]$. This illustrates how biased the MAP estimator can be.

When the estimator takes on this simple form, we can compute the mean square error exactly. In this case, the error is $X - \frac{5}{9}Y$. Then,

$$\mathbb{E}[Y] = \int_0^1 y(3y^2)\, dy = \frac{3}{4}; \qquad f_X(x) = \begin{cases} \int_x^1 2(x+y)\, dy = 2(x)(1-x) + 1 - x^2 = 1 + 2x - 3x^2 & x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[X] = \int_0^1 (x + 2x^2 - 3x^3)\, dx = \frac{1}{2} + \frac{2}{3} - \frac{3}{4} = \frac{5}{12},$$

$$\mathbb{E}[X - \frac{5}{9}Y] = \mathbb{E}[X] - \frac{5}{9}\mathbb{E}[Y] = \frac{5}{12} - \frac{5}{9}\cdot\frac{3}{4} = 0;$$

which we expected because the MMSE estimator is unbiased. To compute the MMSE, we need to compute some additional expectations:

$$\mathbb{E}[Y^2] = \int_0^1 y^2(3y^2)\, dy = \frac{3}{5}; \quad \mathbb{E}[X^2] = \int_0^1 x^2(1 + 2x - 3x^2)\, dx = \frac{1}{3} + \frac{1}{2} - \frac{3}{5} = \frac{7}{30}$$

$$\mathbb{E}[XY] = \int_0^1 \Big(\int_0^y xy2(x+y)\, dx\Big)\, dy = \int_0^1 \frac{5}{3}y^4\, dy = \frac{1}{3}$$

$$\begin{aligned} \text{MMSE} &= \mathbb{E}[(X - \frac{5}{9}Y)^2] = \mathbb{E}[X^2] - \frac{10}{9}\mathbb{E}[XY] + \frac{25}{81}\mathbb{E}[Y^2] \\ &= \frac{7}{30} - \frac{10}{27} + \frac{5}{27} = \frac{13}{270} \end{aligned}$$

It is useful to compare this to the MSE of the MAP estimator, which is

$$MSE_{MAP} = \mathbb{E}[(X-Y)^2] = \mathbb{E}[X^2] - 2\mathbb{E}[XY] + \mathbb{E}[Y^2] = \frac{3}{5} + \frac{7}{30} - 2\frac{1}{3} = \frac{5}{30} = \frac{1}{6},$$

which is much larger than the MMSE.

There is one special case where the MMSE is easy to compute: the case where X, Y are joint Gaussian random variables. In this case, the conditional expected value of X given Y was derived in Chapter 5.4 as

$$\widehat{x}_{MMSE}(y) = \mathbb{E}[X|Y=y] = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(y - \mathbb{E}[Y]),$$

with mean square error given by

$$\mathbb{E}\left[(X - \mathbb{E}[X|Y])^2\right] = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}.$$

In this case, both the MMSE estimator and the minimum mean square error are given in terms of the first- and second-order statistics of X, Y, so no complex integrals need to be computed.

Example 7.8

We measure distance using acoustic echoes by measuring the travel time of a sound pulse. Let X be the unknown distance to an object, which we model as a Gaussian random variable with mean 10^3 meters and standard deviation 100 meters. Assuming the speed of sound is 300 meters/second (to simplify computation), the round trip time of a pulse from a sensor to the object is $2X/300$ seconds. However, our cheap timer is not perfectly accurate, so we model the measurement Y as

$$Y = \frac{1}{150}X + W$$

where W is independent of X, Gaussian, with zero-mean and standard deviation 0.2 seconds.

From the above discussion, we can easily compute the first- and second-order statistics of X, Y as follows:

$$\begin{aligned}
\mathbb{E}[X] &= 1000; \qquad \mathsf{Var}[X] = 10,000; \\
\mathbb{E}[Y] &= \frac{1}{150}\mathbb{E}[X] + \mathbb{E}[W] = \frac{1000}{150} = \frac{20}{3}; \\
\mathsf{Var}[Y] &= \frac{1}{(150)^2}\mathsf{Var}[X] + \mathsf{Var}[W] = \frac{10000}{22500} + \frac{1}{25} = \frac{4}{9} + \frac{1}{25} = \frac{109}{225} \quad X, W \text{ are uncorrelated.} \\
\mathsf{Cov}[X,Y] &= \mathsf{Cov}[X, \frac{1}{150}X + W] = \frac{1}{150}\mathsf{Var}[X] + \mathsf{Cov}[X,W] = \frac{1}{150}\mathsf{Var}[X] = \frac{10000}{150} = \frac{200}{3}.
\end{aligned}$$

Then, the MMSE estimator of the distance X given the measurement $Y = y$ is

$$\widehat{x}_{MMSE}(y) = 1000 + \frac{\frac{200}{3}}{\frac{109}{225}}(y - \frac{20}{3}) = 1000 + \frac{15000}{109}(y - \frac{20}{3}),$$

with MMSE given by

$$MMSE = 10000 - \frac{\frac{40000}{9}}{\frac{109}{225}} = 10000 - \frac{1,000,000}{109} \approx 826 \text{ meters}^2.$$

Thus, the measurement cut down the standard deviation of the location to under 30 meters.

Example 7.9

In this example we wish to estimate X by observing a related random variable Y, where the random variables X and Y are jointly distributed with the density shown in Figure 7.3. This density is uniform over the depicted diamond shaped region. Note that this characterization provides all the information to find both a prior model for X (i.e. the marginal distribution $f_X(x)$) as well as the relationship between Y and X as given by $f_{Y|X}(y \mid x)$.

To find the MMSE estimate for this problem the quantity we need to find is the conditional density $f_{X|Y}(x \mid y)$. We find $f_{X|Y}(x \mid y)$ almost by inspection: restrict the joint density to the slice $Y = y$, and rescale so it normalizes to a probability. Recall that $f_{X|Y}(x \mid y)$ will be a slice of the joint density $f_{X,Y}(x,y)$ parallel to the x-axis and scaled to have unit area. This conditional density is shown on the right in Figure 7.3 for any nontrivial value of y. Since the original density is "flat," each slice will be flat, so all we really need to determine are the edges. The height follows from the constraint that the density has unit area.

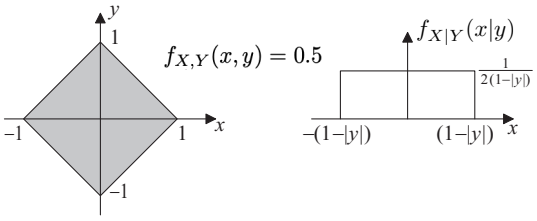


Figure 7.3: MMSE Example

Now, given this density it is easy to see that $\widehat{x}_{MMSE}(y) = \mathbb{E}[x \mid y] = 0$. In this case, the MSE is

$$MSE = E[X^2] = \mathsf{Var}[X] = 4\int_0^1 \left(\int_0^{1-x} x^2 \frac{1}{2}\, dy\right) dx = 2\int_0^1 (x^2 - x^3)\, dx = \frac{1}{6}.$$

Example 7.10
Suppose X and Y are related by the following joint density function:

$$f_{X,Y}(x,y) = \begin{cases} 10x & 0 \le x \le y^2,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

To find the MMSE estimate for this problem we need to find the conditional density $f_{X|Y}(x \mid y)$. By integrating $f_{X,Y}(x,y)$ with respect to y we can find the marginal density for y:

$$f_Y(y) = \begin{cases} \int_0^{y^2} 10x\, dx = 5y^4, 0 \le y \le 1, \\ 0 \qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

Now we can use Bayes' Rule to obtain the conditional density:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{10x}{5y^4} & 0 \le x \le y^2 \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The mean of the conditional density is now found as:

$$\mathbb{E}[X \mid Y = y] = \int_0^{y^2} \frac{2x^2}{y^4}\, dx = \frac{2}{3}y^2.$$

Thus $\widehat{x}_{MMSE}(y) = \frac{2}{3}y^2$. Note that this estimate is a *nonlinear* function of y in this case.

Next let us find the conditional variance $\mathsf{Var}[X|Y = y]$.

$$\mathbb{E}[X^2|Y = y] = \int_0^{y^2} \frac{2x^3}{y^4}\, dx = \frac{1}{2}y^4.$$

$$\mathsf{Var}[X|Y = y] = \mathbb{E}[X^2|Y = y] - (\mathbb{E}[X|Y = y])^2 = \frac{1}{18}y^4.$$

Finally, the minimum mean square error is obtained as:

$$\mathsf{MMSE} = \mathbb{E}\Big[\mathsf{Var}[X|Y]\Big] = \int_0^1 \frac{1}{18}y^4 \cdot 5y^4\, dy = \frac{5}{162}.$$

Let us close by summarizing the properties of MMSE estimators.

- The MMSE estimator is the conditional mean $E[X \mid Y]$.
- The MMSE estimator is always unbiased.
- The MMSE estimator error is orthogonal to any random variable that is a function of the observation Y.

- In general, the MMSE estimator is a nonlinear function of the observation Y, and can be hard to compute.
- For jointly Gaussian problems only, the MMSE estimator is linear in Y and the conditional variance is independent of the observation Y.

7.4 Linear Least Squares Estimation

As noted in the previous section, the MMSE estimator $\mathbb{E}[X|Y]$ is often a complex nonlinear function that is hard to compute. To find a simpler estimator, we want to restrict the estimators to have a restricted functional form. In linear least squares estimation, we restrict the estimator to be of the form $\widehat{x}(y) = ay + b$, for some constants a, b. The linear least squares estimator (LLSE) is the estimator of this form that minimizes the mean square error $\mathbb{E}[(X - aY - b)^2]$. For estimators of this form,

$$\begin{aligned}\mathbb{E}[(X-aY-b)^2] &= \mathbb{E}[(X^2 + a^2Y^2 + b^2 - 2aXY - 2bX + 2abY)] \\ &= \mathbb{E}[X^2] + a^2\mathbb{E}[Y^2] + b^2 - 2a\mathbb{E}[X,Y] - 2b\mathbb{E}[X] + 2ab\mathbb{E}[Y] \\ &= (\mathbb{E}[X])^2 + \mathsf{Var}[X] + a^2(\mathbb{E}[Y])^2 + a^2\mathsf{Var}[Y] + b^2 - 2a\mathsf{Cov}[X,Y] - 2a\mathbb{E}[X]\mathbb{E}[Y] - 2b\mathbb{E}[X] + 2ab\mathbb{E}[Y]\end{aligned}$$

We are going to manipulate this expression by adding and subtracting some terms to complete squares. This will help us identify the values of a, b that result in minimum mean square error. We highlight in red terms that we add and subtract to help us complete the squares.

$$\begin{aligned}\mathbb{E}[(X-aY-b)^2] &= (\mathbb{E}[X])^2 + \mathsf{Var}[X] + a^2(\mathbb{E}[Y])^2 + a^2\mathsf{Var}[Y] + b^2 \\ &\quad - 2a\mathsf{Cov}[X,Y] - 2a\mathbb{E}[X]\mathbb{E}[Y] - 2b\mathbb{E}[X] + 2ab\mathbb{E}[Y] \\ &= (b - \mathbb{E}[X] + a\mathbb{E}[Y])^2 + \mathsf{Var}[X] + a^2\mathsf{Var}[Y] - 2a\mathsf{Cov}[X,Y] \\ &= (b - \mathbb{E}[X] + a\mathbb{E}[Y])^2 + \mathsf{Var}[X] + a^2\mathsf{Var}[Y] - 2a\mathsf{Cov}[X,Y] + \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]} - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]} \\ &= (b - \mathbb{E}[X] + a\mathbb{E}[Y])^2 + \mathsf{Var}[X] + \mathsf{Var}[Y]\Big(a - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\Big)^2 - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}\end{aligned}$$

The values of a and b that minimize the mean square error are now obvious. Note that a, b are only present in the two quadratic terms in the right hand side of the equation. Those quadratic terms are non-negative, and are zero only when $b^* = \mathbb{E}[X] - a\mathbb{E}[Y]$ and $a^* = \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}$. With these choices, we get the minimum mean square error for the linear estimator as

$$\mathbb{E}[(X - a^*Y - b^*)^2] = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}.$$

The linear estimator that achieves this error is the LLSE estimator, given by:

$$\widehat{x}_{LLSE}(Y) = \mathbb{E}[X] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}[Y] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}Y = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y - \mathbb{E}[Y]).$$

This estimator is easy to compute, as it only depends on the first- and second-order statistics of X, Y. It is not necessary to know the joint distributions of X, Y.

The LLSE estimator has several nice properties that we discuss next:

- The LLSE estimator $\widehat{x}_{LLSE}(Y)$ is unbiased, as can be seen from:

$$\mathbb{E}[X - \widehat{x}_{LLSE}(Y)] = \mathbb{E}\left[X - \mathbb{E}[X] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y - \mathbb{E}[Y])\right] = \mathbb{E}[X] - \mathbb{E}[X] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(\mathbb{E}[Y] - \mathbb{E}[Y]) = 0.$$

- The error in the LLSE estimator is orthogonal to the observations Y. Again, we show this by direct computation, highlighting in red where we add and subtract equal terms. We also highlight in blue terms that evaluate to zero, so that the logic is clear for how the equations simplify.

$$\begin{aligned}
\mathbb{E}[(X-\widehat{x}_{LLSE}(Y))Y] &= \mathbb{E}\left[\left(X-\mathbb{E}[X]-\frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y-\mathbb{E}[Y])\right)Y\right] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])Y\Big] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])Y\Big] - \mathbb{E}\Big[(X-\mathbb{E}[X])\mathbb{E}[Y]\Big] + \mathbb{E}\Big[(X-\mathbb{E}[X])\mathbb{E}[Y]\Big] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\Big] + \mathbb{E}\Big[(X-\mathbb{E}[X])\Big]\mathbb{E}[Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y\Big] \\
&= \mathsf{Cov}[X,Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y - (Y-\mathbb{E}[Y])\mathbb{E}[Y] + (Y-\mathbb{E}[Y])\mathbb{E}[Y]\Big] \\
&= \mathsf{Cov}[X,Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])(Y-\mathbb{E}[Y])\Big] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])\Big]\mathbb{E}[Y] \\
&= \mathsf{Cov}[X,Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathsf{Var}[Y] = \mathsf{Cov}[X,Y] - \mathsf{Cov}[X,Y] = 0
\end{aligned}$$

  The terms in blue above are 0 because $\mathbb{E}\big[(X-\mathbb{E}[X])\big] = 0, \mathbb{E}\big[(Y-\mathbb{E}[Y])\big] = 0$. As a consequence, the LLSE error $X-\widehat{x}_{LLSE}(Y)$ is also orthogonal to any linear function of Y. In this derivation, we have also shown that $\mathbb{E}\Big[(X-\mathbb{E}[X])Y\Big] = \mathsf{Cov}[X,Y]$, and $\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y\Big] = \mathsf{Var}[Y]$.

- The mean-square error of the LLSE estimator is no smaller than the mean square error for the MMSE estimator. The MMSE estimator mean-square error is the smallest among all the nonlinear estimators, whereas the LLSE estimator mean-square error is the smallest among all the linear estimators only. However, if the MMSE estimator is a linear function of Y, then the mean square errors of the LLSE and MMSE estimators are the same, and the estimators are also equal. Thus, for jointly Gaussian X, Y, the MMSE and LLSE and MAP are equal and have the same mean-square error.
- One interpretation for the LLSE estimator is that, given the first- and second-order statistics for X, Y, it approximates the joint density of X, Y as a Gaussian density with these statistics. The MMSE estimator for this Gaussian problem is the same as the LLSE estimator.

This orthogonality property is depicted in Figure 7.4. The idea is that the optimal estimate is that linear function of the data which has no correlation with the error. Intuitively, if correlation remained between the error and the estimate, there would remain information in the error of help in estimating X that we should have extracted. Note that this geometric condition implies that the error is orthogonal (i.e. uncorrelated with) both the data itself (which is obviously a trivial function of the data) as well as the LLSE estimate (which is clearly a linear function of the data).

We can derive the LLSE estimator from the following properties: We want an unbiased estimator, that is orthogonal to the observations. These properties can be derived directly from the properties of projections, which we have not emphasized in our development. Given these two properties, the unbiased property implies

$$\mathbb{E}[(X-aY-b)] = 0 \iff b = \mathbb{E}[X] - a\mathbb{E}[Y].$$

Furthermore, the orthogonality property implies

$$\mathbb{E}[(X-aY-b)Y] = 0 \iff E[(X-E[X]) - a[Y-E[Y])Y] = 0 \text{ (substitute } b \text{ in.)}$$

$$\mathbb{E}[(X-E[X]) - a[Y-E[Y])Y] = \mathsf{Cov}[X,Y] - a\mathsf{Var}[Y] = 0 \iff a = \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}$$

Hence,

$$\widehat{x}_{LLSE}(y) = \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}Y + \mathbb{E}[X] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}[Y] = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y-\mathbb{E}[Y]).$$

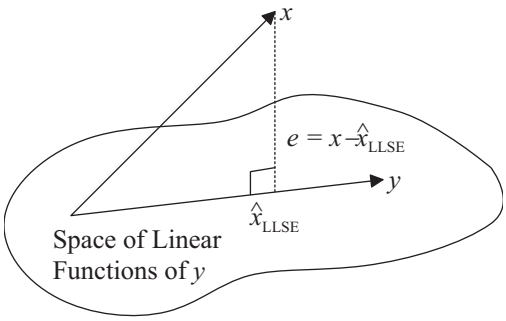


Figure 7.4: Illustration of the projection theorem for LLSE.

We can use the orthogonality property also to derive an expression for the variance of the error. Since the error $X - \widehat{x}_{LLSE}(Y)$ is orthogonal to any linear function of Y, it is also orthogonal to the estimate $\widehat{x}_{LLSE}(y)$. Hence,

$$\mathsf{Var}[X] = \mathsf{Var}[X - \widehat{x}_{LLSE}(Y)] + \mathsf{Var}[\widehat{x}_{LLSE}(Y)].$$

Since $\widehat{x}_{LLSE}(Y) = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y - \mathbb{E}[Y])$ is a scaled and shifted version of Y, its variance is

$$\mathsf{Var}[\widehat{x}_{LLSE}(Y)] = \Big(\frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\Big)^2 \mathsf{Var}[Y] = \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}.$$

Hence,

$$\mathsf{Var}[X - \widehat{x}_{LLSE}(Y)] = E[(X - \widehat{x}_{LLSE}(Y))^2] = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}.$$

Example 7.11
For this example let us revisit the problem of Example 7.10. We need the second order quantities $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathsf{Cov}[X,Y[$, $\mathsf{Var}[X]$, and $\mathsf{Var}[Y]$, which we compute as:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \int_0^1 5y^5\, dy = \frac{5}{6}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dx\, dy = \int_0^1 \int_0^{y^2} x 10x\, dx\, dy == \frac{10}{21}$$

$$\mathsf{Var}[Y] = \int_{-\infty}^{\infty} y^2 f_Y(y)\, dy - \mathbb{E}[Y]^2 = \int_0^1 5y^6\, dy - \Big(\frac{5}{6}\Big)^2 = \frac{5}{252}$$

$$\mathsf{Var}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f_X(x)\, dx\, dy - \mathbb{E}[X]^2 = \int_0^1\int_0^{y^2} x^2\, 10x\, dx\, dy - \Big(\frac{10}{21}\Big)^2 = \frac{5}{18} - \Big(\frac{10}{21}\Big)^2 = \frac{5}{98}$$

$$\mathsf{Cov}[X,Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\, dx\, dy - \mathbb{E}[X]\mathbb{E}[Y] = \int_0^1\int_0^{y^2} 10x^2 y\; dx\, dy - \frac{5}{6}\frac{10}{21} = .\frac{5}{12} - \frac{5}{6}\frac{10}{21} = \frac{5}{252}$$

Thus we obtain for the LLSE:

$$\widehat{x}_{\mathsf{LLSE}}(y) = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\,(y - \mathbb{E}[Y]) = \frac{10}{21} + \frac{5/252}{5/252}\left(y - \frac{5}{6}\right) = y - \frac{5}{14}$$

as before. Using the formula for the MSE we obtain

$$\mathsf{MSE} = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]} = \frac{5}{98} - \frac{(5/252)^2}{5/252} = \frac{55}{1764} = 0.0312$$

Note that this MSE is worse than that obtained by the optimal MMSE estimator of Example 7.10 – but not much worse.

Example 7.12
Consider a simple example where a temperature sensor measures the true temperature T, which is assumed to be a Gaussian random variable, with mean 27 and variance 9. The measurement observed is modeled as

$$Y = T + V$$

where V is measurement error, independent of T, corresponding to quantization noise. V is assumed to be uniformly distributed on [-3,3]. Hence, $\mathbb{E}[V] = 0, \mathsf{Var}[V] = 3$. This is a common model for measurements, where part of the measurement models the relationship between the unobserved variable T and the measurement Y, and the other part represents the errors in the measurement process.

The goal is to generate the LLSE estimate of T based on observation of Y. We compute the needed statistics below:

$$\begin{aligned}\mathbb{E}[T] = 27;\ \mathbb{E}[Y] = \mathbb{E}[T] + \mathbb{E}[V] = 27;\ \mathsf{Var}[T] = 9;\\ \mathsf{Var}[Y] = \mathsf{Var}[T] + \mathsf{Var}[V] + 2\mathsf{Cov}[T,V] = \mathsf{Var}[T] + \mathsf{Var}[V] = 9+3 = 12\ (T,V \text{ independent.})\\ \mathsf{Cov}[T,Y] = \mathsf{Cov}[T,T+V] = \mathsf{Var}[T] + \mathsf{Cov}[T,V] = \mathsf{Var}[T] = 9;\end{aligned}$$

With these statistics, we have

$$\widehat{T}_{LLSE}(y) = \mathbb{E}[T] + \frac{\mathsf{Cov}[T,Y]}{\mathsf{Var}[Y]}(y - \mathbb{E}[Y]) = 27 + \frac{9}{12}(y-27) = \frac{1}{4}\cdot 27 + \frac{3}{4}\cdot y.$$

The MSE error is

$$MSE = \mathsf{Var}[T] - \frac{\mathsf{Cov}[T,Y]^2}{\mathsf{Var}[T]} = 9 - \frac{81}{12} = \frac{9}{4}.$$

Example 7.13
Let's revisit the example of estimating probabilities using multiple trials. Let P be a uniform random variable on [0,1]. Let measurement Y be a Binomial(N,P) random variable, corresponding to the total number of successes in N independent trials with probability of success P for each trial. From previous examples, we know

$$\widehat{x}_{ML}(y) = \widehat{x}_{MAP}(y) = \frac{y}{N}.$$

As discussed previously, this is an unbiased estimate.

Can we compute the MMSE estimate for this simple case? The conditional density of P given observation $Y = y$ is given by

$$f_{P|Y}(p|y) = \frac{P_{Y|P}(y|p)f_P(p)}{P_Y(y)} = \frac{\binom{N}{y}p^y(1-p)^{N-y}}{\binom{N}{y}\int_0^1 q^y(1-q)^{N-y}\,dq}.$$

Evaluating the denominator is not easy, but we can do the integral as follows, using repeated integration by parts:

$$\begin{aligned}&\int_0^1 q^y(1-q)^{N-y}\,dq = \frac{N-y}{y+1}\int_0^1 q^{y+1}(1-q)^{N-y-1}\,dq\\ &\Rightarrow \int_0^1 q^y(1-q)^{N-y}\,dq = \frac{(N-y)!k!}{N!}\int_0^1 q^N\,dq = \frac{1}{N+1}\cdot\frac{1}{\binom{N}{y}}\end{aligned}$$

Thus, $P_Y(y) = \frac{1}{N+1}, y \in \{0,1,\ldots,N\}$, which is a discrete uniform distribution. Intuitively, this makes sense. Since we have no information about P, every value of Y is equally likely when averaged over all the possible P.

The conditional density of P given $Y = y$ is

$$f_{P|Y}(p|y) = (N+1)\binom{N}{y}p^y(1-p)^{N-y}.$$

The MMSE estimator is

$$\widehat{P}_{MMSE}(y) = \int_0^1 p f_{P|Y}(p|y)\,dp = \int_0^1 (N+1)p\binom{N}{y}p^y(1-p)^{N-y}\,dp.$$

This has another difficult integral, but we can again evaluate it using repeated integration by parts as:

$$\begin{aligned}&\int_0^1 p^{y+1}(1-p)^{N-y}\,dp = \frac{N-y}{y+2}\int_0^1 p^{y+2}(1-p)^{N-y-1}\,dp\\ &\Rightarrow \int_0^1 p^{y+1}(1-p)^{N-y}\,dp = \frac{(N-y)!(y+1)!}{(N+1)!}\int_0^1 p^{N+1}\,dp = \frac{1}{N+2}\cdot\frac{1}{\binom{N+1}{y+1}}\end{aligned}$$

Thus,

$$\widehat{P}_{MMSE}(y) = (N+1)\cdot\binom{N}{y}\cdot\frac{1}{N+2}\cdot\frac{1}{\binom{N+1}{y+1}} = \frac{y+1}{N+2}.$$

Note this is different from the MAP estimator, but it is also unbiased. Since it is linear, this is also the LLSE estimator! We can verify this through computation, as

$$\mathbb{E}[P] = 0.5;\ \mathbb{E}[Y] = 0.5N;\ \mathsf{Var}[P] = \frac{1}{12};\ \mathsf{Var}[Y] = \frac{N(N+2)}{12}$$
$$\mathsf{Cov}[Y,P] = \mathbb{E}\big[\mathbb{E}[(Y-0.5N)(P-0.5)|P]\big] = \mathbb{E}\big[\mathbb{E}[(Y-0.5N)|P](P-0.5)\big]$$
$$= \mathbb{E}\big[N(P-0.5)^2\big] = N\int_0^1 (p-0.5)^2\,dp = \frac{N}{12}$$

Then,

$$\widehat{P}_{LLSE}(y) = (N+1)\cdot\binom{N}{y}\cdot\frac{1}{N+2}\cdot\frac{1}{\binom{N+1}{y+1}} = \frac{y+1}{N+2} = \widehat{P}_{MMSE}(y).$$

The resulting MSE for the LLSE and MMSE estimator is

$$MSE = Var[P] - \frac{Cov[Y,P]^2}{Var[Y]} = \frac{1}{12} - \frac{N^2}{144}\cdot\frac{12}{N(N+2)} = \frac{1}{12}(1-\frac{N}{N+2}) = \frac{1}{6(N+2)}.$$

The MSE for the MAP estimator is

$$MSE = E[(P-\frac{Y}{N})^2 = \mathsf{Var}[P] - 2\frac{\mathsf{Cov}[P,Y]}{N} + \frac{\mathsf{Var}[Y]}{N^2} = \frac{1}{12} - \frac{2}{12} + \frac{N+2}{12N} = \frac{1}{6N}.$$

Note the MAP MSE is slightly larger than the MMSE and LLSE MSE.

Let us close by summarizing the properties of LLSE estimates:

- The LLSE estimate is the minimum MSE estimate over all *linear* functions of the data.
- The LLSE estimate is always unbiased.
- The associated error covariances satisfy is at least as large as the error covariance of the MMSE estimate.
- The LLSE estimate equals the MMSE estimate for the jointly Gaussian case.
- The LLSE estimate only requires knowledge of second-order properties.

7.5 Estimation for Random Vectors

We conclude this chapter by discussing how the estimation concepts extend to random vectors. This extension is critical for many applications, including statistics and data science. In this case, the unobserved variables and the observed variables can both be vectors. The model is as follows. We suppose that we have a random vector $\underline{Z}$ that can be partitioned into two subvectors as $\underline{Z} = \begin{bmatrix}\underline{X}\\ \underline{Y}\end{bmatrix}$, where $\underline{Y}$ is the observation vector and $\underline{X}$ is the unobserved vector. Let $X \in \Re^n, Y \in \Re^m$. The objective in estimation is now to construct an estimator $\underline{\widehat{x}}(\underline{y})$ that will estimate the unobserved vector $\underline{X}$ based on the observed values $\underline{Y} = \underline{y}$. Note that the vector estimate is a vector composed of the estimates for each of the components of X, so that

$$\underline{\widehat{x}}(\underline{y}) = \begin{bmatrix}\widehat{x}_1(\underline{y})\\ \vdots \\ \widehat{x}_n(\underline{y})\end{bmatrix}.$$

We begin with the statistical description of the random vectors. Assuming the random vectors are continuous valued, with joint densities, we will have a joint density

$$f_{\underline{X},\underline{Y}}(\underline{x},\underline{y}) = f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})f_{\underline{X}}(\underline{x})$$

where we have shown the factorization of the joint density into a conditional density for the observed variables, given the unobserved variables, and a marginal density for the unobserved variables.

The mean square error of an estimator of the vector $\underline{X}$ given $\underline{Y}$ is defined as

$$MSE = \mathbb{E}[\sum_{k=1}^{n}(X_i - \widehat{x}_i(\underline{Y}))^2] = \mathbb{E}[(\underline{X} - \underline{\widehat{x}}(\underline{y}))^T(\underline{X} - \underline{\widehat{x}}(\underline{y}))].$$

This is a common metric that will be used for evaluating the quality of estimators.

7.5.1 ML and MAP estimation for random vectors

Given this statistical description, we extend our estimation concepts to random vectors. The **maximum likelihood estimate** (ML) is given by

$$\underline{\widehat{x}}_{ML}(\underline{y}) \in \arg\max_{\underline{x}} f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})$$

where the maximization is now over vector arguments, and often requires iterative search algorithms. The main difficulty in this estimation is that the likelihood function $f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})$ can have many local maxima, which makes the search for a global maximum a difficult combinatorial problem. However, for some special cases, it will be possible to find global maxima, as we will illustrate with examples.

Similarly, the **maximum a posteriori** (MAP) estimate is given by

$$\underline{\widehat{x}}_{MAP}(\underline{y}) \in \arg\max_{\underline{x}} f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})f_{\underline{X}}(\underline{x}).$$

The MAP estimator includes the additional information on the prior density of the unobserved variables $\underline{X}$. The optimization still has difficulties with local maxima. Nevertheless, specifying a prior distribution for $\underline{X}$ serves as a regularization term that guides the optimization towards specific regions in the search space space where the maxima are expected to be, and makes the solution less sensitive to measurement errors.

Example 7.14

One of the most interesting applications of vector estimation is for estimating the parameters of the distribution of a random variable, given many observation samples. For instance, assume you are measuring the delays in your favorite transportation mechanism, the Green Line. This delay T is modeled as an exponential random variable with parameter λ, but the parameter λ is unknown. To estimate λ, we observe the sample delays over many days. We assume the delays in different days are independent, with the same underlying distribution for T. Let T_k denote the delay measured at day k, where $k = 1, 2, \ldots, N$.

To formulate this in the form of a vector estimation problem, let $X = \lambda$ be the unobserved variable. Our observed vector $\underline{Y}$ is given as

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}.$$

With this choice of variables, we need the statistical description. Note that, given $\lambda = x$, the conditional density of Y_k is an exponential density:

$$f_{Y_k|X}(y_k|x) = \begin{cases} xe^{-xy_k} & y_k \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, the fact that each of the Y_k represent an independent sample of T yields the following expression for the joint conditional density of $\underline{Y}$ given X:

$$f_{\underline{Y}|X}(\underline{y}|x) = \prod_{k=1}^{N} f_{Y_k|X}(y_k|x).$$

With this information, we can now obtain the ML estimate of X, given observation of the vector $\underline{Y}$, as

$$\widehat{x}_{ML}(\underline{y}) \in \arg\max_{x\geq 0} f_{\underline{Y}|X}(\underline{y}|x) = \arg\max_{x\geq 0} \ln\left(f_{\underline{Y}|X}(\underline{y}|x)\right)$$

$$\ln\left(f_{\underline{Y}|X}(\underline{y}|x)\right) = \sum_{k=1}^{N} \ln\left(f_{Y_k|x}(y_k|x)\right) = N\ln(x) - \sum_{k=1}^{N} y_k x$$

To solve the maximization problem, we differentiate with respect to x and find where the derivative is zero: as long as that happens for positive x, it will be the maximum.

$$\frac{d}{dx}\left(N\ln(x) - \sum_{k=1}^{N} y_k x\right) = \frac{N}{x} - \sum_{k=1}^{N} y_k = 0 \iff x = \frac{N}{\sum_{k=1}^{N} y_k}.$$

Note that the value of x where the derivative vanishes is unique and non-negative. Therefore, the ML estimator is $\widehat{x}_{ML}(\underline{y}) = \frac{N}{\sum_{k=1}^{N} y_k}$. This is an intuitive estimator, as it says that the estimate of the rate λ is the inverse of the average delay.

What if we had some prior information on X, and we wanted to generate the MAP estimator? Assume that we know that the rate is distributed uniformly in $[0.1, 1.1]$, corresponding to average delays between 0.909 and 10 minutes. Thus,

$$f_X(x) = \begin{cases} 1 & x \in [0.1, 1.1], \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the MAP estimate is given by

$$\widehat{x}_{MAP}(\underline{y}) \in \arg\max_{x\geq 0} f_{\underline{Y}|X}(\underline{y}|x) f_X(x) = \arg\max_{x\in[0.1,1.1]} \ln\left(f_{\underline{Y}|X}(\underline{y}|x)\right) = \arg\max_{x\in[0.1,1.1]} N\ln(x) - \sum_{k=1}^{N} y_k x.$$

The prior information resulted in restricting the search to the interval $x \in [0.1, 1.1]$, because the product of the two densities is zero outside of this interval, and cannot be a maximum. Proceeding as above, the point where the derivative with respect to zero vanishes is $x = \frac{N}{\sum_{k=1}^{N} y_k}$. However, this value of x may not be in $[0.1, 1.1]$, in which case we find the closest value in the interval where the function is maximized (the log-likelihood has a unique maximum and is continuous), so that the MAP estimator is

$$\widehat{x}_{MAP}(\underline{y}) = \begin{cases} 0.1 & \frac{N}{\sum_{k=1}^{N} y_k} < 0.1, \\ 1.1 & \frac{N}{\sum_{k=1}^{N} y_k} > 1.1, \\ \frac{N}{\sum_{k=1}^{N} y_k} & \text{otherwise.} \end{cases}$$

Example 7.15
Let U be a Gaussian random variable, with unknown mean m and variance v. We are interested in estimating the mean and variance of U. We collect N independent samples of U, where sample k is denoted as Y_k.

When posed as an estimation problem, our unobserved variables are m and v. Let $\underline{X} = \begin{bmatrix} m \\ v \end{bmatrix} \equiv \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. The observed vector is

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}.$$

As before, we need the likelihood of the observed vector given the unobserved vector. From our Gaussian assumptions, we have

$$f_{Y_k|\underline{X}}(y_k|\underline{x}) = \frac{1}{\sqrt{2\pi x_2}} e^{-\frac{(y_k - x_1)^2}{2x_2}}.$$

Furthermore, under the assumption of independent sampling, we have

$$f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) = \prod_{k=1}^{N} f_{Y_k|\underline{X}}(y_k|\underline{x}).$$

Thus, the log-likelihood is given as

$$\ln\left(f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})\right) = \sum_{k=1}^{N} \ln\left(f_{Y_k|x}(y_k|\underline{x})\right) = -\frac{N}{2}\ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2}.$$

We wish to maximize the above over any real-valued x_1, and for $x_2 \geq 0$, as x_2 represents the unknown covariance. Hence, the ML estimate is

$$\widehat{\underline{x}}_{ML}(\underline{y}) \in \arg \max_{x_2 \geq 0, x_1} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right).$$

To maximize, take the partial derivative of the above with respect to x_1 and x_2 and set both equal to zero:

$$\frac{\partial}{\partial x_1} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right) = \sum_{k=1}^{N} \frac{(y_k - x_1)}{x_2} = 0 \iff x_1 = \frac{\sum_{k=1}^{N} y_k}{N}.$$

$$\frac{\partial}{\partial x_2} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right) = -\frac{N}{2x_2} + \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2^2} = 0 \iff x_2 = \frac{\sum_{k=1}^{N} (y_k - x_1)^2}{N}.$$

Note that the solution for x_2 is always non-negative, satisfying the constraints. The ML estimate is thus

$$\widehat{m}_{ML} = \frac{\sum_{k=1}^{N} y_k}{N}; \quad \widehat{v}_{ML} = \frac{\sum_{k=1}^{N} (y_k - \widehat{m}_{ML})^2}{N}.$$

What if we had some prior information on m and v? Assume that, apriori, we knew that m was Gaussian, with mean 0, variance v, and v was uniform in [1,5]. This implies

$$f_{\underline{X}}(\underline{x}) = f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2) = \begin{cases} \frac{1}{\sqrt{2\pi x_2}} e^{-\frac{x_1^2}{2x_2}} \cdot \frac{1}{4} & x_2 \in [1,5], x_1 \in \Re, \\ 0 & \text{otherwise..} \end{cases}$$

The MAP estimator is now obtained as

$$\widehat{\underline{x}}_{MAP}(\underline{y}) \in \arg \max_{x_2 > 0, x_1} \left(f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) f_{\underline{X}}(\underline{x}) \right) = \arg \max_{x_2 \in [1,5], x_1} \left(f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) f_{\underline{X}}(\underline{x}) \right).$$

because the joint density is zero when $x_2 \notin [1,5]$ and hence cannot be maximal there. Taking logarithms, we get

$$\ln \left(f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) f_{\underline{X}}(\underline{x}) \right) = \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right) - \frac{1}{2} \ln(2\pi x_2) - \ln(4) - \frac{x_1^2}{2x_2}.$$

Differentiating with respect to x_1 and x_2 yields:

$$\frac{\partial}{\partial x_1} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} - \frac{1}{2} \ln(2\pi) - \ln(4) - \frac{x_1^2}{2x_2} \right) = \sum_{k=1}^{N} \frac{(y_k - x_1)}{x_2} - \frac{x_1}{x_2} = 0.$$

$$\frac{\partial}{\partial x_2} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right) = -\frac{N}{2x_2} + \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2^2} - \frac{1}{2x_2} = 0.$$

The equations are easily solved as:

$$x_1 = \frac{\sum_{k=1}^{N} y_k}{N+1}.$$

For the second equation, we have

$$-(N+1)x_2 + \sum_{k=1}^{N} (y_k - x_1)^2 = 0 \iff x_2 = \frac{\sum_{k=1}^{N} (y_k - x_1)^2}{N+1}.$$

Taking into account the constraints on the optimization, the MAP estimator is

$$\widehat{m}_{MAP} = \frac{\sum_{k=1}^{N} y_k}{N+1}; \quad \widehat{v}_{MAP} = \begin{cases} 1 & \frac{\sum_{k=1}^{N} (y_k - \widehat{m}_{ML})^2}{N+1} < 1, \\ 5 & \frac{\sum_{k=1}^{N} (y_k - \widehat{m}_{ML})^2}{N+1} > 5, \\ \frac{\sum_{k=1}^{N} (y_k - \widehat{m}_{ML})^2}{N+1} & \text{otherwise.} \end{cases}$$

7.5.2 MMSE and LLSE estimation for random vectors

The MMSE estimator of $\underline{X}$, based on observations of $\underline{Y}$, is given by

$$\widehat{\underline{x}}_{MMSE}(\underline{y}) = \begin{bmatrix} \mathbb{E}[X_1|\underline{Y}] \\ \vdots \\ \mathbb{E}[X_n|\underline{Y}] \end{bmatrix} = \begin{bmatrix} \int \cdots \int x_1 f_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y})\, dx_1 \\ \vdots \\ \int \cdots \int x_n f_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y})\, dx_n \end{bmatrix} = \mathbb{E}[\underline{X}|\underline{Y}].$$

Computation of this estimate requires the conditional density $f_{\underline{X}|\underline{Y}}((\underline{x}|\underline{y})$, obtained by Bayes' Rule as

$$f_{\underline{X}|\underline{Y}}((\underline{x}|\underline{y}) = \frac{f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) f_{\underline{X}}(\underline{x})}{f_{\underline{Y}}(\underline{y})}.$$

Computing the denominator in Bayes' Rule requires a multidimensional integral that is usually very difficult to evaluate; this limits our ability to compute MMSE estimators for general distributions.

There is a special case where the MMSE solution can be computed efficiently: when $\underline{X}$ and $\underline{Y}$ are jointly Gaussian random vectors. As we derived in 5, the conditional expected value $\mathbb{E}[\underline{X}|\underline{Y}]$ is a linear function of the observed value $\underline{Y} = \underline{y}$, and the MMSE estimator will be the same as the LLSE estimator. We will discuss the LLSE estimator for random vectors below.

We assume the random vector $\underline{Z} = \begin{bmatrix} \underline{X} \\ \underline{Y} \end{bmatrix}$ has first-order statistics given by

$$\mathbb{E}[\underline{Z}] = \begin{bmatrix} \underline{\mu}_{\underline{X}} \\ \underline{\mu}_{\underline{Y}} \end{bmatrix}.$$

Let n_X, n_Y be the dimensions of the random vectors $\underline{X}, \underline{Y}$. The random vector $\underline{Z}$ has covariance matrix $\Sigma_{\underline{Z}}$, which can be partitioned along the dimensions of $\underline{X}, \underline{Y}$ as follows:

$$\Sigma_{\underline{Z}} = \begin{bmatrix} \Sigma_{\underline{X}} & \Sigma_{\underline{X},\underline{Y}} \\ \Sigma_{\underline{Y},\underline{X}} & \Sigma_{\underline{Y}} \end{bmatrix}.$$

Note that $\Sigma_{\underline{X}}$ is an $n_X \times n_X$ matrix, which is the covariance matrix for the unobserved vector $\underline{X}$. $\Sigma_{\underline{Y}}$ is an $n_Y \times n_Y$ matrix, which is the covariance matrix for the observed vector $\underline{Y}$. The matrix $\Sigma_{\underline{X},\underline{Y}}$ is an $n_X \times n_Y$ matrix known as the cross-covariance between $\underline{X}$ and $\underline{Y}$, and is defined as

$$\Sigma_{\underline{X},\underline{Y}} = \mathbb{E}\Big[(\underline{X} - \mathbb{E}[\underline{X}])(\underline{Y} - \mathbb{E}[\underline{Y}])^T\Big].$$

A linear estimator of $\underline{X}$ based on $\underline{Y}$ is an estimator of the form $\widehat{\underline{x}}(\underline{y}) = \mathbf{A}\underline{y} + \underline{b}$ for a constant matrix $\mathbf{A}$ of dimension $n_X \times n_Y$, and $\underline{b}$ is a constant vector of dimension n_X.

We want to find the best linear estimator of $\underline{X}$ given observation $\underline{Y}$, where best is the estimator that yields the smallest least squares error. Rather than posing this as an optimization problem, we will derive this estimator using the orthogonality and unbiased properties of the LLSE estimator, which can be established using the principles of best approximation. Specifically, we seek a linear estimator that is unbiased, and where the estimation error is orthogonal to the observations.

The first condition is an unbiased estimator, which requires

$$\mathbb{E}[\underline{X} - \mathbf{A}\underline{Y} - \underline{b}] = 0 = \underline{\mu}_{\underline{X}} - \mathbf{A}\underline{\mu}_{\underline{Y}} - \underline{b} \iff \underline{b} = \underline{\mu}_{\underline{X}} - \mathbf{A}\underline{\mu}_{\underline{Y}}.$$

The next condition is that the estimation error must be orthogonal to the measurements. The error is a vector, $\underline{e} = \underline{X} - \mathbf{A}\underline{Y} - \underline{b}$. Orthogonality requires that every component of the error is orthogonal to every

component of the measurement vector:

$$\begin{aligned}\mathbb{E}\big[(\underline{X}-\mathbf{A}\underline{Y}-\underline{b})\underline{Y}^T\big] = 0 &= \mathbb{E}\big[\underline{X}\underline{Y}^T-\mathbf{A}\underline{Y}\underline{Y}^T-\underline{b}\underline{Y}^T\big]\\ &= \mathbb{E}\big[(\underline{X}-\mathbb{E}[\underline{X}])\underline{Y}^T-\mathbf{A}(\underline{Y}-\mathbb{E}[\underline{Y}])\underline{Y}^T\big] \quad \text{(substituting for } b,)\\ &= \mathbb{E}\big[(\underline{X}-\mathbb{E}[\underline{X}])\underline{Y}^T\big]-\mathbf{A}\mathbb{E}\big[(\underline{Y}-\mathbb{E}[\underline{Y}])\underline{Y}^T\big]\\ &= \Sigma_{\underline{X},\underline{Y}}-\mathbf{A}\Sigma_{\underline{Y}} = 0 \iff \mathbf{A}=\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}\end{aligned}$$

Thus, we have our LLSE estimator for random vectors:

$$\widehat{\underline{x}}_{LLSE}(\underline{y}) = \mathbb{E}[\underline{X}]+\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}(\underline{Y}-\mathbb{E}[\underline{Y}]).$$

This is very similar to the structure of the scalar LLSE estimator we discussed in the previous section. The main difference is that, in dealing with vectors, we must take matrix inverses and preserve the order of the operations when we take constants out of the expectations.

We can also use orthogonality to derive an expression for the covariance of the estimation error. We know the estimator is orthogonal to the error vector, because the estimator is a linear function of the measurement Y. Hence, $\underline{X} = \underline{e}+\widehat{\underline{x}}_{LLSE}(\underline{y})$ is the sum of two uncorrelated vectors. Thus,

$$\Sigma_{\underline{X}} = \Sigma_{\underline{e}}+\Sigma_{\widehat{\underline{x}}_{LLSE}(\underline{Y})}.$$

We also know that $\widehat{\underline{x}}_{LLSE}(\underline{Y})$ is a linear transformation of $\underline{Y}$, so

$$\Sigma_{\widehat{\underline{x}}_{LLSE}(\underline{Y})} = \Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}\Sigma_{\underline{Y}}\Sigma_{\underline{Y}}^{-1}\Sigma_{\underline{X},\underline{Y}}^T = \Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}\Sigma_{\underline{X},\underline{Y}}^T.$$

Therefore,

$$\Sigma_{\underline{e}} = \Sigma_{\underline{X}}-\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}\Sigma_{\underline{X},\underline{Y}}^T.$$

We illustrate these results with an example:

Example 7.16
Let $\underline{X}$ be a random, two-dimensional vector with statistics $\mathbb{E}[\underline{X}] = \underline{0}, \Sigma_{\underline{X}} = \begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}$.

Let $\underline{W}$ also be a two-dimensional vector, uncorrelated with $\underline{X}$, with statistics $\mathbb{E}[\underline{W}] = 0, \Sigma_{\underline{W}} = \begin{bmatrix}0.1 & 0\\ 0 & 0.1\end{bmatrix}$.

Define the observation vector $\underline{Y} = \underline{X}+\underline{W}$. Then, the first and second order statistics of $\underline{X},\underline{Y}$ are:

$$\mathbb{E}[Y] = 0;\ \Sigma_{\underline{Y}} = \Sigma_{\underline{X}}+\Sigma_{ullW} = \begin{bmatrix}1.1 & -0.9\\ -0.9 & 1.1\end{bmatrix}$$
$$\Sigma_{\underline{X},\underline{Y}} = \mathsf{Cov}[\underline{X},\underline{Y}] = \mathsf{Cov}[\underline{X},\underline{X}]+\mathsf{Cov}[\underline{X},\underline{W}] = \mathsf{Cov}[\underline{X},\underline{X}] = \Sigma_{\underline{X}}$$

Note that $\Sigma_{\underline{Y}}^{-1} = \begin{bmatrix}2.75 & 2.25\\ 2.25 & 2.75\end{bmatrix}$. With this, the LLSE estimator is

$$\widehat{\underline{x}}_{LLSE}(\underline{y}) = \underline{\mu}_{\underline{X}}+\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}(\underline{y}-\underline{\mu}_{\underline{Y}}) = \begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}\begin{bmatrix}2.75 & 2.25\\ 2.25 & 2.75\end{bmatrix}\underline{y} = \begin{bmatrix}0.725 & -0.225\\ -0.225 & 0.725\end{bmatrix}\underline{y}$$

The covariance of the estimation error $\underline{e}$ is

$$\Sigma_{\underline{e}} = \begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}-\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}\begin{bmatrix}2.75 & 2.25\\ 2.25 & 2.75\end{bmatrix}\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix} = \begin{bmatrix}0.0725 & -0.0225\\ -0.0225 & 0.0725\end{bmatrix}$$

Chapter 8

Sums of Random Variables: Bounds and Limits

So far in this course, we have focused mostly on pairs of random variables X and Y. Many experiments of interest generate more than two random variables for each outcome. When we consider we consider $n \geq 2$ random variables $X_1, \ldots, X_n$, we describe their probabilistic behavior using a joint Cumulative distribution function (CDF) of the form

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}[\{X_1 \leq x_1, \ldots, X_n \leq x_n\}],$$

which is the natural extension of the joint CDF for pairs of random variables. When the joint random variables are discrete, we define the joint probability mass function (PMF) as

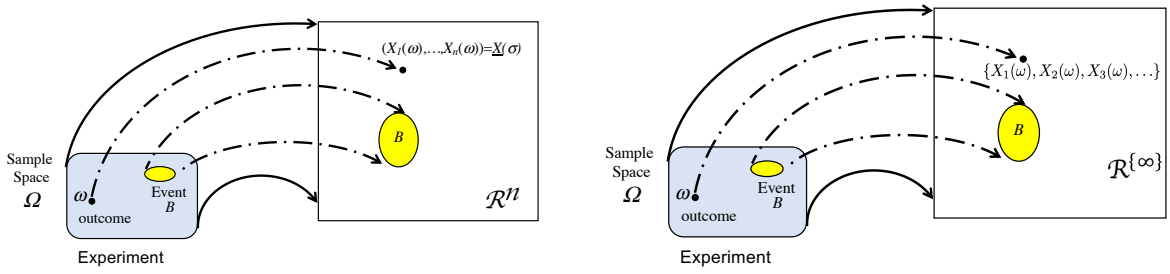
$$P_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}[\{X_1 = x_1, \ldots, X_n = x_n\}].$$

When the random variables are continuous, we define the joint probability density function (PDF) as

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{X_1,\ldots,X_n}(x_1,\ldots,x_n).$$

All of the basic properties that we established for CDFs, PMFs and PDFs for pairs of random variables extend naturally to CDFs, PMFs and PDFs of n random variables.

In this chapter, we study experiments that generate a countably infinite collection of random variables. Such collections are often called discrete time random processes, as the index of the random variables can be mapped to the countable natural numbers. Figure 8.1 compares experiments that generate random vectors, which we have discussed previously, to ones that generate a countable collection of random variables. Formally, each element $X_k(\omega)$ of the collection $\{X_1, X_2, X_3, \ldots\}$ is a random variable, a measurable function from the sample space Ω to the real numbers. Such collections are often called random processes or stochastic processes. A random process is an indexed collection $\{X_t, t \in T\}$ of random variables generated by a single experiment. When the index T is countable and can be mapped to the natural numbers $\mathcal{N}$, we refer to such processes as discrete-time or discrete-index random processes. Such processes are generalizations of the concept of random vectors introduced in earlier chapters, as shown in Figure 8.1.



Random vectors | Infinite sequence of random variables

Figure 8.1: Experiments generate infinite sequences of random variables.

The study of general random processes is a subject of a more advanced course, and requires tools that we will not introduce in this course. There are special cases which we can address with simple extensions of

the methodology we have described in previous chapters. Like random vectors, for general random processes one would have to define the joint probability mass functions for finite set of distinct indices $k_1, k_2, \ldots, k_n$, of the form $P_{X_{k_1},X_{k_2},\ldots,X_{k_n}}(x_1, x_2, \ldots, x_n)$ if the random variables are discrete. Equivalently, one would need to define joint probability density functions $f_{X_{k_1},X_{k_2},\ldots,X_{k_n}}(x_1, x_2, \ldots, x_n)$. Computing such joint densities is cumbersome and hard to describe.

In this chapter, we focus on the special case where the collection of random variables $\{X_1, X_2, X_3, \ldots\}$ are mutually independent. This implies that, for any finite set of distinct indices $k_1, k_2, \ldots, k_n$, and subsets $A_1, A_2, \ldots A_n \subset \Re$, we have

$$\mathbb{P}[\{X_{k_1} \in A_1\}, \{X_{k_2} \in A_2\}, \ldots, \{X_{k_n} \in A_n\}] = \mathbb{P}[\{X_{k_1} \in A_1\}]\mathbb{P}[\{X_{k_1} \in A_1\}] \cdots \mathbb{P}[\{X_{k_n} \in A_n\}].$$

When the random variables are discrete, the joint probability mass functions factor as

$$P_{X_{k_1},X_{k_2},\ldots,X_{k_n}}(x_1, x_2, \ldots, x_n) = P_{X_{k_1}}(x_1)P_{X_{k_2}}(x_2) \cdot P_{X_{k_n}}(x_n).$$

For continuous random variables, the joint densities factor as

$$f_{X_{k_1},X_{k_2},\ldots,X_{k_n}}(x_1, x_2, \ldots, x_n) = f_{X_{k_1}}(x_1)f_{X_{k_2}}(x_2) \cdot f_{X_{k_n}}(x_n).$$

This independence property will allow us to analyze properties of the collection of random variables using the tools we have developed for the analysis of pairs of random variables in earlier chapters.

Of particular interest is the case where the collection $\{X_1, X_2, X_3, \ldots\}$ corresponds to outputs of repeating an experiment independently, with an infinite number of trials. For instance, let X_i correspond to the output of a Bernoulli trial, with parameter p that represents the probability that $X_i = 1$. The empirical theory of probability suggests that p should be the fraction of experiments that result in an outcome $X_i = 1$. From the results of the last chapter, the maximum likelihood estimate of p given observations of the outcomes of the first N experiments is $\frac{\sum_{i=1}^{N} X_i}{N}$. What happens as the number of experiments N increases to infinity? In the limit, we would expect that this estimate, which is a derived random variable, would converge in some sense to the correct value p. We will analyze the behavior of such sequences of random variables and make precise in what manner do such sequences converge.

8.1 Independent, Identically Distributed Random Variables

A collection of random variables $\{X_n, n \in \mathcal{N}\}$ is referred to as an independent, identically distributed collection of random variables if the random variables $X_1, X_2, \ldots$ are mutually independent, and the marginal cumulative distribution function of each random variable is the same for each random variable. That is, $F_{X_k}(x) = F_{X_j}(x)$ for any $j, k \in \mathcal{N}$. We use the short-hand notation **i.i.d.** to represent independent and identically distributed in the rest of this chapter.

Let $\{X_n, n \in \mathcal{N}\}$ be an i.i.d. collection of random variables, each of which has finite mean μ and finite variance σ^2. Define a sequence of dependent random variables S_n using partial sums as:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Using linearity of expectation and the i.i.d. property, we establish the following:

$$\mathbb{E}[S_n] = \sum_{j=1}^{n} E[X_j] = n\mu.$$

What about the covariance of S_n? This is also computed readily, as

$$\begin{aligned}\mathsf{Var}S_n &= \mathbb{E}[(S_n - \mathbb{E}[S_n])^2] = \mathbb{E}\big[\big(\sum_{j=1}^{n}(X_j-\mu)\big)^2\big] \\ &= \mathbb{E}\big[\sum_{j=1}^{n}\sum_{k=1}^{n}(X_j-\mu)(X_i-\mu)\big] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}\big[(X_j-\mu)(X_i-\mu)\big] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n}\mathsf{Cov}[X_j,X_k] = \sum_{j=1}^{n}\mathsf{Var}[X_j] = n\sigma^2\end{aligned}$$

where the last equality follows because the X_j are i.i.d., hence $\mathsf{Cov}[X_j, X_k] = 0$ if $k \neq j$, and $\mathsf{Cov}[X_j, X_k] = \mathsf{Var}[X_j] = \sigma^2$ if $k = j$.

Notice that, as n grows, $\mathbb{E}[S_n]$ and $\mathsf{Var}[S_n]$ both grow linearly with n. Thus, we don't expect any type of convergence for the sequence S_n. Let's define instead the variables $M_n = \frac{S_n}{n}$, the average of the first n random variables X_k. Then,

$$\mathbb{E}[M_n] = \frac{\mathbb{E}[S_n]}{n} = \mu,$$

and

$$\mathsf{Var}[M_n] = \big(\frac{1}{n}\big)^2\mathsf{Var}[S_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Note now that, as n increases, the variables M_n have the same mean, and the variance of the random variables decreases. The distribution of M_n becomes more concentrated about its average μ.

Example 8.1
Let X be an exponential random variable with $\lambda = 1$. Thus, $\mathbb{E}[X] = \frac{1}{\lambda} = 1$. Let M_n denote the sample mean of n independent samples of X. How many samples are needed so that the variance of the sample mean is less than or equal to 0.01?

From the properties of exponential random variables, $\mathsf{Var}[X] = \frac{1}{\lambda^2} = 1$. Hence, for the average of n samples, $\mathsf{Var}[M_n] = \frac{\mathsf{Var}X}{n}$. This means that we need at least 100 samples for the variance to be 0.01 or less.

At this point, we don't know much about the probability distribution of M_n. Indeed, since M_n is a sum of independent random variables, its probability density function is an n-fold convolution of the densities of the scaled random variables X_j/n. In order to make statements concerning the probability of events related to M_n, we discuss next some estimates of such probabilities based on only mean and variance information.

8.2 Useful inequalities for Random Variables

In order to analyze notions of convergence of random variables, it is useful to bound the errors between the limit random variable and elements of the sequence using inequalities that do not require knowledge of the full distribution of the random variables. Below, we present a few useful inequalities:

8.2.1 Markov inequality

Suppose that X is a non-negative random variable with known finite mean, and we want to obtain some bounds on the probability distribution function of X. The **Markov Inequality** is given by

$$\mathbb{P}[\{X \geq a\}] = \int_a^\infty f_X(x)\,dx \leq \frac{\mathbb{E}[X]}{a}.$$

How do show the Markov inequality is true? The steps below illustrate the argument when X is a continuous random variable with finite expected value. Since X is non-negative, the density $f_X(x)$ is zero for $x < 0$.

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^{\infty} x f_X(x)\, dx \\ &= \int_0^{a} x f_X(x)\, dx + \int_a^{\infty} x f_X(x)\, dx \\ &\geq \int_a^{\infty} x f_X(x)\, dx \qquad \text{(Drop the first term, non-negative)} \\ &\geq a \int_a^{\infty} f_X(x)\, dx = a\mathbb{P}[\{X \geq a\}] \qquad (x \geq c \text{ in the integrand}) \end{aligned}$$

The Markov inequality follows by dividing both sides by a.

The above argument can be generalized as follows: Let $g(x) \geq 0$ everywhere, and let $g(x) > a > 0$ for all $x \in A$, for a subset A of the real line $\Re$. Then,

$$\begin{aligned} E[g(X)] &= \int_{x\in A} g(x) f_X(x)\, dx + \int_{x\notin A} g(x) f_X(x)\, dx \\ &\geq \int_{x\in A} g(x) f_X(x)\, dx \geq a \int_{x\in A} f_X(x)\, dx \\ &= a\mathbb{P}[\{X \in A\}]. \end{aligned}$$

Hence, $\mathbb{P}[\{X \in A\}] \leq \frac{E[g(X)]}{a}$.

8.2.2 Chebyshev inequality

Suppose that the mean μ and variance σ^2 of a random variable X are known, and we would like to bound the probability that the variable is far from its mean. The Chebyshev inequality states that

$$\mathbb{P}[\{|X - \mu| \geq a\}] \leq \frac{\sigma^2}{a^2}.$$

The Chebyshev inequality can be derived from the Markov Inequality, by defining the non-negative random variable $Y = (X - \mu)^2$. Since $\mathbb{E}[Y] = \mathsf{Var}[X]$ is finite, the Markov inequality states that

$$\mathbb{P}[\{Y \geq a^2\}] \leq \frac{\mathbb{E}[Y]}{a^2} = \frac{\sigma^2}{a^2}.$$

In terms of equivalent events,

$$\mathbb{P}[\{|X - \mu| \geq a\}] = \mathbb{P}[\{Y \geq a^2\}] \leq \frac{\sigma^2}{a^2},$$

which shows the Chebyshev Inequality.

A different way of writing the Chebyshev inequality is as follows: Let $a = a'\sigma$. Then,

$$\mathbb{P}[\{|X - \mu| \geq a\}]\mathbb{P}[\{|X - \mu| \geq a'\sigma\}] =\leq \frac{\sigma^2}{a'^2\sigma^2} = \frac{1}{a'^2}.$$

This can be interpreted as in terms of number of standard deviations away from the mean. The probability that X is more than a' standard deviations away from its mean is less than $\frac{1}{a'^2}$.

The above can be generalized for any random variable X such that $\mathbb{E}[(X - \mu)^n]$ is finite for some even number n, as

$$\mathbb{P}[\{|X - m| \geq a\}] = \mathbb{P}[\{|X - m|^n \geq a^n\}] \leq \frac{\mathbb{E}[|X - m|^n]}{a^n}$$

or, more generally, for any real, nonnegative, even function $g(x)$ which is non-decreasing for $x > 0$, and has finite expectation. Then,

$$\mathbb{P}[\{g(X) \geq g(a)\}] \leq \frac{\mathbb{E}[g(X)]}{g(a)}.$$

Example 8.2
A random variable W, which represents the waiting time to be served at a restaurant, is uniformly distributed in the interval from 0 to 10 minutes. Estimate a bound on the probability that the wait is at least 8 minutes.

Note that, in this case, we know the exact probability of the event $\{W \geq 8\}$, because we have the density of W: Hence, $\mathbb{P}[\{W \geq 8\}] = 0.2$. What if we estimated this using either the Markov inequality or the Chebyshev inequality? We know that $\mathbb{E}[W] = 5$, and $\mathsf{Var}[W] = \frac{100}{12} = \frac{25}{3}$. We also know that $W \geq 0$. Hence, the Markov inequality indicates that

$$\mathbb{P}[\{W \geq 8\}] \leq \frac{\mathbb{E}[W]}{8} = \frac{5}{8},$$

which is much larger than 0.2. It shows that the bound can be loose.

What about the Chebyshev inequality? It states:

$$\mathbb{P}[\{|W - 5| \geq 3\}] \leq \frac{\frac{25}{3}}{9} = \frac{25}{27}.$$

If we divide by 2 to represent the one-sided probability that $W > 8$, we have

$$\mathbb{P}[\{W \geq 8\}] \leq \frac{25}{54},$$

which is closer to 0.2, but still a loose bound.

Example 8.3
Assume X is Gaussian, with mean 0 and variance 1. Then, $\mathbb{P}[\{|X| > a\}] = 2Q(a)$, where $Q(\cdot)$ is the standard Gaussian complementary cumulative distribution function. We can compare, as a function of a, the estimate generated by the Chebyshev inequality and the true value $2Q(a)$, as:

Value of a	Chebyshev Inequality	2Q(a)
a = 2	0.25	0.0455
a = 3	0.111	0.0027
a = 4	0.0625	0.0000633
a = 5	0.04	0.0000006

The values illustrate the conservative nature of the Chebyshev inequality.

Example 8.4
Chebyshev's Inequality can provide a tight bound for some distributions. Consider the discrete random variable X with range in $R_X = \{-1, 1\}$ such that $P(1) = 0.5, P(-1) = 0.5$. Then, $\mathbb{E}[X] = 0, \mathsf{Var}[X] = 1$. Therefore, Chebyshev's Inequality states that

$$\mathbb{P}[\{|X - \mathbb{E}[X]| \geq 1\}] \leq 1.$$

However, we know that $\mathbb{P}[\{|X - \mathbb{E}[X]| \geq 1\}] = 1$ in this example, so the bound is equal to the actual probability.

8.2.3 Chernoff and Jensen Inequalities

There are other bounds on probabilities of random variables that are useful to know. We discuss them briefly here without proof.

Given a random variable X, define a new random variable Y_ϵ as:

$$Y_\epsilon = \begin{cases} 1 & X \geq \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

That is, Y is the indicator random variable that $X \geq \epsilon$.

Then, for all $t \geq 0$, the following inequality holds:

$$e^{tX} \geq e^{t\epsilon} Y.$$

Thus,

$$\mathbb{E}\left[e^{tX}\right] \geq \mathbb{E}\left[e^{t\epsilon}Y\right] = e^{t\epsilon}\mathbb{P}[\{X \geq \epsilon\}],$$

which implies that

$$\mathbb{P}[\{X \geq \epsilon\}] \leq e^{-t\epsilon}\mathbb{E}\left[e^{tX}\right], \quad t \geq 0.$$

This bound can be tightened through the choice of t, as follows:

$$\mathbb{P}[\{X \geq \epsilon\}] \leq \min_{t \geq 0} e^{-t\epsilon}\mathbb{E}\left[e^{tX}\right], \quad t \geq 0.$$

Note that this bound requires computation of $\mathbb{E}\left[e^{tX}\right]$, which is equivalent to computing the characteristic function (or moment-generating function) of X! Thus, this bound requires extensive knowledge of the full probability density function of X, and not just its mean and variance.

Another useful inequality is Jensen's inequality. A *convex* function $g(x)$ of a continuous variable x in an interval I is a function such that, for any $\alpha \in [0,1]$, any x, $y \in I$, the following is true:

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

Let X denote a random variable with probability density or probability mass function distributed over I, and let μ denote its mean, which must be in I. Then, for any convex function g, we have

$$g(\mu) \leq \mathbb{E}[g(X)].$$

One way to recognize that this is true is to note that, if X were a discrete random variable with $P_X(x) = \alpha, P_X(y) = 1-\alpha$, then the definition of g as convex implies

$$g(\alpha x + (1-\alpha)y) = g(\mathbb{E}[X]) \leq \alpha g(x) + (1-\alpha)g(y) = \mathbb{E}[g(X)].$$

This can be extended to other discrete probability mass functions, and in a limiting argument to continuous random variables X.

Jensen's inequality can be used to derive many inequalities concerning moments of random variables, such as the Cauchy-Schwartz inequality that we used to prove that the correlation coefficient between two random variables X, Y is a number with magnitude less than or equal to 1.

Example 8.5
Assume X is Binomial(n,p). Then, using the Chernoff bound, we have

$$\mathbb{P}[\{X \geq \epsilon\}] \leq \min_{t \geq 0} e^{-t\epsilon}\mathbb{E}\left[e^{tX}\right], \quad t \geq 0.$$

We can compute $\mathbb{E}\left[e^{tX}\right]$ in this case, as X is the sum of n independent Bernoulli(p) random variables $Y_1, \ldots, Y_n$. Hence,

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^{t\sum_{k=1}^n Y_k}\right] = \prod_{k=1}^n \mathbb{E}[e^{tY_k}] = \prod_{k=1}^n (1-p+pe^t) = (1-p+pe^t)^n.$$

Let's compute a bound on $\mathbb{P}[\{X \geq \alpha n\}]$ for $1 > \alpha > p$. Then,

$$\min_{t \geq 0} e^{-t\alpha n}\mathbb{E}\left[e^{tX}\right] = \min_{t \geq 0} e^{-t\alpha n}(1-p+pe^t)^n$$

Taking the derivative with respect to t and setting it equal to 0 yields

$$\frac{d}{dt}\left(e^{-t\alpha n}(1-p+pe^t)^n\right) = -\alpha n e^{-t\alpha n}(1-p+pe^t)^n + npe^t e^{-t\alpha n}(1-p+pe^t)^{n-1} = 0.$$

Dividing by common factors yields the solution at the minimum value:

$$\alpha n(1-p+pe^t) = npe^t \Rightarrow e^t = \frac{\alpha(1-p)}{p(1-\alpha)}$$

Substituting into the bound, we get:

$$\min_{t\geq 0} e^{-t\alpha n}\mathbb{E}\left[e^{tX}\right] = \left(\frac{p(1-\alpha)}{\alpha(1-p)}\right)^{\alpha n}\left(1-p+p\frac{\alpha(1-p)}{p(1-\alpha)}\right) = \left(\frac{p(1-\alpha)}{\alpha(1-p)}\right)^{\alpha n}\frac{(1-p)}{(1-\alpha)} = \left(\frac{p}{\alpha}\right)^{\alpha n}\left(\frac{1-\alpha}{1-p}\right)^{\alpha n-1}$$

For $p = 0.5, \alpha = 0.75$ the above bound is $\mathbb{P}[\{X \geq \alpha n\}] = 2\left(\frac{1}{3}\right)^{0.75n}$, which decays fast as n increases.

8.2.4 Hoeffding's Inequality

Hoeffding's inequality provides bounds on probabilities of the averages of random variables. Let $X_1, \ldots, X_n$ be independent random variables whose range $R_{X_k} \subset [a_k, b_k]$, where $-\infty < a_k < b_k < \infty$. That is, with probability 1, $a_k \geq X_k \leq b_k$ for $k = 1, \ldots, n$. We define the sample mean of these variables by

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Then,

$$\mathbb{P}[\{M_n - \mathbb{E}[M_n] \geq \epsilon\}] \leq e^{-\frac{2n^2\epsilon^2}{\sum_{k=1}^n (b_k-a_k)^2}},$$

$$\mathbb{P}[\{M_n - \mathbb{E}[M_n] \leq -\epsilon\}] \leq e^{-\frac{2n^2\epsilon^2}{\sum_{k=1}^n (b_k-a_k)^2}}.$$

We can combine the two bounds to get a bound that is similar to the Chebyshev bound, as

$$\mathbb{P}[\{|M_n - \mathbb{E}[M_n]| \geq \epsilon\}] \leq 2e^{-\frac{2n^2\epsilon^2}{\sum_{k=1}^n (b_k-a_k)^2}}.$$

For the special case that X_k are independent, identically distributed Bernoulli(p) random variables, $a_k = 0, b_k = 1$, and thus $\sum_{k=1}^n (b_k - a_k)^2 = n$. In this case, Hoeffding's inequality yields

$$\mathbb{P}[\{|M_n - p| \geq \epsilon\}] \leq 2e^{-2n\epsilon^2}.$$

Example 8.6

Let's apply Hoeffding's inequality to the previous example, where X is Binomial(n, p), so that X is the sum of n independent Bernoulli(p) random variables $Y_1, \ldots, Y_n$. We want to compute $\mathbb{P}[\{X \geq \alpha n\}]$ for $1 > \alpha > p$. Note that $M_n = \frac{X}{n}$, so $\mathbb{P}[\{X \geq \alpha n\}] = \mathbb{P}[\{M_n \geq \alpha\}] = \mathbb{P}[\{M_n - p \geq \alpha - p\}$. Using Hoeffding's inequality, we have

$$\mathbb{P}[\{M_n - p \geq \alpha - p\} \leq e^{-2n(\alpha-p)^2}.$$

For $\alpha = 0.75, p = 0.5$, this bound becomes $\mathbb{P}[\{M_n - p \geq 0.25\} \leq e^{-\frac{n}{8}}$.

8.3 The Law of Large Numbers

The law of large numbers has a central role in probability and statistics. It states that if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value. It is consistent with the frequency interpretation of the concept of probability, where the probability of an event is the fraction of times when the event occurs if the experiment were repeated independently an infinite number of times. There are two main versions of the Law of Large Numbers: the weak law of large numbers and the strong law of large numbers. The differences are subtle, and we will highlight some of the

We state the weak law of large numbers first, and then prove it.

Theorem 8.1 (Weak Law of Large Numbers)
Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite means $\mathbb{E}[X_n] = \mu$, and define the sequence of sample means $\{M_n\}$ as

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[\{\, |M_n - \mu| > \varepsilon \,\}] = 0.$$

The proof of this theorem in its general case is subtle, and requires a truncation argument. We will instead show this using the additional assumption that $\mathsf{Var}[X_n] = \sigma^2 < \infty$. In this case, we have already shown in Section 8.1 that $\mathbb{E}[M_n] = \mu, \mathsf{Var}[M_n] = \frac{\sigma^2}{n}$. Using Chebyshev's inequality, we have that,

$$\mathbb{P}[\{|M_n - \mu| \geq \varepsilon\}] \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Taking the limit of this as $n \to \infty$ establishes the weak law of large numbers.

As mentioned earlier, the weak law applies in the case of i.i.d. random variables, but it also applies in some other cases. For instance, if the X_n have finite bounded variances, and are uncorrelated, the law still holds. Even if the variances grow unbounded with n, as long as the variance of the averages M_n goes to zero as $n \to \infty$, the same argument can be applied to establish the weak law of large numbers.

The type of convergence used in the weak law of large numbers is convergence in probability. A sequence of random variables $\{M_n\}$ converges to a limiting random variable M in probability if and only if

$$\lim_{n\to\infty} \mathbb{P}[\{|M_n - M| > \varepsilon\}] = 0.$$

When the random variables X_n have finite variance, bounded by σ^2, we can show the averages M_n converge to their limit in mean square also, which means

$$\lim_{n\to\infty} \mathbb{E}[(M_n - \mu)^2] = 0.$$

This is trivial to show as we know the variance of M_n goes to zero, and the mean is μ.

Example 8.7
Assume X is a Bernoulli random variable, with probability p that $X = 1$. Let X_k be a repetition of the same experiment, for $k = 1, 2, \ldots$. From our results in estimation, we know that the maximum likelihood estimate of p given n observations X_k is given by

$$\widehat{p}_{ML}(\{X_k, k = 1, 2, \ldots, n\}) = \frac{\sum_{k=1}^{n} X_k}{n}.$$

which is the sample average discussed above. By the weak Law of Large Numbers,

$$\mathbb{P}[\{|\frac{\sum_{k=1}^{n} X_k}{n} - \widehat{p}_{ML}(\{X_k, k = 1, 2, \ldots, n\})| \geq \varepsilon\}] \leq \frac{\sigma^2}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2},$$

which converges to zero as $n \to \infty$.

Example 8.8
One of the problems with MMSE estimation that we discussed in Section 7.3 is that the integrals are hard to compute. For instance, in Example 7.6, we had to compute the following integral to get the conditional density of X given $Y = y$: $\int_0^{1000} \frac{2x}{10^6} \frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\, dx$.

In general, suppose we have a function $g(x)$, and we wanted to compute $\int_a^b g(x)\, dx$, but $g(x)$ was a continuous function that was hard to integrate. We can compute the integral approximately using the weak Law of Large Numbers as follows: Let $\{X_n\}$ be an i.i.d. sequence of random variables, uniformly distributed in $[a, b]$. Let $Y_n = g(X_n)$. Then, $\{Y_n\}$ is also an i.i.d. sequence, and $\mathbb{E}[Y] = \int_a^b \frac{g(x)}{b-a}\, dx$. Given that $[a, b]$ is a bounded interval and $g(x)$ is continuous, we can show that $\mathsf{Var}[Y_n] = \sigma_Y^2 < \infty$.

By the weak Law of Large Numbers, the average $\frac{Y_1+Y_2+\ldots+Y_n}{n}$ is close to $\mathbb{E}[Y]$. Hence, an approximation for the integral is

$$\int_a^b g(x)\,dx \approx (b-a)\frac{Y_1+Y_2+\ldots+Y_n}{n}.$$

Furthermore, we can compute the probability that the error is significant using the Chebyshev inequality. This probabilistic technique is known as the Monte Carlo method of integration.

The statement of the weak law of large numbers is a statement about probabilities, averaged over all the outcomes in the experiment. It does not guarantee that, for any outcome $\omega \in \Omega$ that generates a sequence of realizations of random variables $X_1(\omega), X_2(\omega), \ldots$, the average of those random variables will be close to $\mathbb{E}[X] = \mu$. It does not even guarantee that the set of outcomes for which the average does not converge to μ has zero probability of occurring. For that, we need the Strong Law of Large Numbers, stated next:

Theorem 8.2 (Strong Law of Large Numbers)
Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite mean μ. Define the sequence of sample means $\{M_n\}$ as

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

then,

$$\mathbb{P}\left[\{\omega \in \Omega : \lim_{n\to\infty} M_n(\omega) = \mu\}\right] = 1.$$

The type of convergence in the strong law of large numbers is known as almost sure convergence. It states that the probability of an outcome where the sequence does not converge is zero. The proof is more complex than that of the weak law and is beyond the scope of our course. The strong law requires independence of the random variables X_k, whereas the weak law can be established using uncorrelated assumptions.

The main difference between the strong law of large numbers and the weak law of large numbers is where the limit is placed in the statement: The weak law states:

$$\lim_{n\to\infty} \mathbb{P}[\{\, |M_n - \mu| > \varepsilon \,\}] = 0,$$

whereas the strong law states:

$$\mathbb{P}\left[\{\omega \in \Omega : \lim_{n\to\infty} M_n = \mu\}\right] = 1.$$

Thus, the strong law states that, for any $\varepsilon > 0$, the probability of the event $\{|M_n - \mu| > \varepsilon$ for at most a finite $n\}$ is equal to 1.

8.4 The Central Limit Theorem

The law of large numbers characterizes that the sample averages M_n converge to a deterministic quantity, the mean $\mathbb{E}[X] = \mu$. Basically, it states that the cumulative distribution function $F_{M_n}(z)$ converges to a unit step function:

$$F_{M_n}(z) = \begin{cases} 0 & z < \mu \\ 1 & z \geq \mu. \end{cases}$$

It is often of interest to characterize the error $M_n - \mu$. We know from our previous analysis that, if the sequence $\{X_k\}$ is i.i.d., with finite mean μ and finite variance σ^2, the error $M_n - \mu$ has 0 mean, and variance $\frac{\sigma^2}{n}$. If we define a scaled variable $Y_n = \frac{\sqrt{n}}{\sigma}(M_n - \mu)$, the variables Y_n have zero mean and variance 1 for all n. We can express Y_n in terms of the partial sums $S_n = nM_n$ as

$$Y_n = \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

The Central Limit Theorem states that, as n increases, the cumulative distribution functions of Y_n converge to a special form, as stated below:

Theorem 8.3 (Central Limit Theorem)
Consider a sequence of independent, identically distributed random variables $\{X_n\}$ with finite mean μ and finite variance σ^2. Denote the partial sum S_n and the partial average M_n as

$$S_n = \sum_{i=1}^{n} X_i; \; M_n = \frac{1}{n} S_n.$$

Define the new random sequence $\{Y_n\}$ as

$$Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Then, for any real. number y, the sequence of cumulative distribution functions $F_{Y_n}(y)$ converges to $\Phi(y)$, the cumulative distribution function of a standard Gaussian random variable with mean 0 and variance 1.

The surprising part of the Central Limit Theorem (CLT) is that the distribution of the individual random variables can be arbitrary. This is why Gaussian random variables are used so often in probabilistic analysis, since they approximately model sums of many independent effects. Note also the scaling used in the Central Limit Theorem: S_n has mean $n\mu$ and variance $n\sigma^2$. Hence, Y_n is measured in terms of units of standard deviation away from the mean, a similar scaling that we used when computing probabilities of Gaussian random variables.

We sketch a brief proof the the CLT by computing what are known as characteristic functions, which are the Fourier transform of the probability density functions of continuous random variables, or equivalently the Fourier transform of the of the generalized probability mass functions (expressed as the sum of $\delta(\cdot)$ functions) for discrete random variables. Since density functions integrate to 1 and probability mass functions sum to 1, the characteristic function transform will be well-defined for all $j\omega$, with $j = \sqrt{-1}$.

The characteristic function of a random variable X is

$$\Psi_W(\omega) = \mathbb{E}[e^{j\omega X}] = \begin{cases} \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)\, dx & X \text{ continuous,} \\ \sum_{x_k \in R_X} e^{j\omega x_k} P(x_k) & X \text{ discrete.} \end{cases}$$

Note that

$$Y_n = \frac{1}{\sigma_X \sqrt{n}} \sum_{k=1}^{n} (X_k - \mu_x)$$

is a sum of independent, zero-mean random variables. There is a convergence result in probability called Lévy's continuity theorem, which states that, if the characteristic functions of a sequence of random variables Y_n converge pointwise as $n \to \infty$ to a function $\psi(\omega)$ which is continuous at $\omega = 0$, then the CDFs of Y_n converge pointwise to the CDF of a random variable Y with characteristic function $\psi(\omega)$. We will use this result to prove the CLT using characteristic functions.

The characteristic function of Y_n is given by:

$$\begin{aligned} \Psi_{Y_n}(\omega) = \mathbb{E}[e^{j\omega Y_n}] &= \mathbb{E}[e^{j\omega \frac{1}{\sigma_X \sqrt{n}} \sum_{k=1}^{n} (X_k - \mu)}] \\ &= \mathbb{E}\left[\prod_{k=1}^{n} e^{j\omega \frac{1}{\sigma_X \sqrt{n}} (X_k - \mu)}\right] \\ &= \prod_{k=1}^{n} \mathbb{E}\left[e^{j\omega \frac{1}{\sigma_X \sqrt{n}} (X_k - \mu)}\right] \quad \text{(independence)} \\ &= \left(\mathbb{E}\left[e^{j\omega \frac{1}{\sigma_X \sqrt{n}} (X_1 - \mu)}\right]\right)^n \quad \text{(identically distributed)} \end{aligned}$$

where the last equalities follows from the independent, identically distributed assumption. We expand the exponential in the expression using a Taylor series as:

$$e^{j\omega \frac{X_1 - \mu}{\sigma\sqrt{n}}} = 1 + \frac{j\omega(X_1 - \mu)}{\sigma\sqrt{n}} - \frac{\omega^2 (X_1 - \mu)^2}{2\sigma^2 n} + \ldots$$

For large n, we neglect terms beyond the first three terms to get the approximation:

$$\mathbb{E}[e^{j\omega\frac{X_1-\mu}{\sigma\sqrt{n}}}] \approx 1 + \frac{j\omega\mathbb{E}[X_1-\mu]}{\sigma\sqrt{n}} - \frac{\omega^2\mathbb{E}[(X_1-\mu)^2]}{2\sigma^2 n}$$
$$\approx 1 - \frac{\omega^2}{2n}$$

because $\mathbb{E}[X_1-\mu]=0, \mathbb{E}[(X_1-\mu)^2]=\sigma^2$. Thus,

$$\Psi_{Y_n}(\omega) \approx (1-\frac{\omega^2}{2n})^n$$

and, taking limits as $n\to\infty$, we get

$$\lim_{n\to\infty}\Psi_{Y_n}(s) = e^{-\omega^2/2}$$

Let Z be a zero mean, unit variance Gaussian random variable. Then,

$$\Psi_Z(\omega) = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}+j\omega z}dz = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}+j\omega z+\frac{\omega^2}{2}-\frac{\omega^2}{2}}dz = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(z-j\omega z)^2}{2}-\frac{\omega^2}{2}}dz = e^{\frac{-\omega^2}{2}}.$$

Thus, the characteristic functions of Y_n converge for each ω to the characteristic function of a zero-mean, unit variance Gaussian random variable for all values. By Lévy's continuity theorem, this implies that the CDF of Y_n converges to the CDF of a Gaussian$(0,1)$ random variable.

The CLT implies that, given any i.i.d. sequence of random variables, we can compute probabilities of events relating to the sum of the random variables approximately using a Gaussian distribution. That is,

$$\mathbb{P}[\{(X_1+X_2+\ldots+X_n)\le a\}] \approx \Phi\left(\frac{a-n\mu}{\sqrt{n}\sigma}\right)$$

and

$$\mathbb{P}[\{\frac{1}{n}(X_1+X_2+\ldots+X_n)\le b\}] \approx \Phi\left(\frac{b-\mu}{\frac{\sigma}{\sqrt{n}}}\right).$$

As a rule of thumb, these approximation are very accurate as long as $\frac{|a-n\mu_X|}{\sqrt{n}\sigma}$ is less than 3.

Example 8.9

Assume we have a disk drive that takes X milliseconds for each disk access time, where X is a random variable, uniformly distributed in [0,12]. Assume one must access disk 12 times independently, and define the total access time $T = X_1 + \ldots + X_{12}$. Then, $\mathbb{E}[T]12\mathbb{E}[X] = 72$ msec, and $\mathsf{Var}[T] = 12\mathsf{Var}[X] = 12\cdot\frac{12^2}{12} = 144$. Therefore, the standard deviation of the sum is 12. We want to compute the probability that the total wait time is greater than 75 seconds.

We approximate this with the CLT, since T is the sum of i.i.d. random variables.

$$\mathbb{P}[T>75] = 1 - F_T(75) \approx 1 - \Phi\left(\frac{75-72}{12}\right) = 1-\Phi(0.25) = Q(0.25).$$

What about the probability that $T<48$? This is

$$F_T(48) \approx \Phi(\frac{48-72}{12}) = \Phi(-2) = Q(2).$$

Note that, to compute this exactly, we would need the probability density of T, which would require performing 12 convolutions.

Example 8.10

A Modem transmits 10^4 bits, where each bit is i.i.d. with probability $p=0.5$. We would like to estimate the probability that we get more that 5100 one bits. We also want to estimate the probability that the number of one bits we receive is in the interval (4900,5100].

The total number of one bits received, T is the sum of 10^4 independent Bernoulli random variables. We know this is a Binomial $(10^4, 0.5)$ random variable, but computing the quantities asked involve summing between 100 and 200 binomial terms. We approximate this using. the CLT as follows:

$$\mathbb{E}[T] = 10^4 p = 5000; \quad \mathsf{Var}[T] = 10^4 p(1-p) = 2500; \quad \sigma_T = 50.$$

With this approximation, we quickly estimate

$$\mathbb{P}[\{T > 5100\}] = 1 - F_T(5100) \approx 1 - \Phi(\frac{5100 - 5000}{50}) = 1 - \Phi(2) = Q(2).$$

$$\mathbb{P}[\{T \in (4900, 5100]\}] = F_T(5100) - F_T(4900) \approx \Phi(2) - \Phi(-2) = \Phi(2) - Q(2).$$

Chapter 9

Sample Statistics

Suppose we have a random variable X, and we collect n independent samples $X_1, X_2, \ldots, X_n$ of this random variable. The probability model is that the samples are random variables $X_1, X_2, \ldots, X_n$ are mutually independent and identically distributed with the same distribution as X. As we discussed in Chapter **??**, the sample mean $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is an approximation to the expected value $\mathbb{E}[X]$ that converges with probability 1 to the true expected value $\mathbb{E}[X]$, by the Strong Law of Large Numbers.

For any finite n, the sample mean M_n is a random variable. This random variable is the sum of n independent random variables, so describing statistical properties such as its PDF if X were a continuous random variable would require computing n-fold convolutions of the PDF $f_X(x)$.

Nevertheless, we know

$$\mathbb{E}\left[M_n - \mathbb{E}[X]\right] = \mathbb{E}[\frac{1}{n}\sum_{k=1}^{n} X_k] - \mathbb{E}[X] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k] - \mathbb{E}[X] = 0$$

from the property that all the X_k are identically distributed. Under the assumption that the random variable X has finite variance σ^2, we can also compute

$$\mathsf{Var}[M_n] = \frac{1}{n^2}\sum_{k=1}^{n}\mathsf{Var}[X_k] = \frac{\sigma^2}{n},$$

because the random variables X_k are independent.

The Central Limit Theorem states that a scaled version of $M_n - \mathbb{E}[X]$ has a CDF that converges to that of a standard Gaussian random variable with mean 0 and variance 1. Specifically, we define $Z_n = \sqrt{n}\frac{M_n - \mathbb{E}[X]}{\sigma}$ as the scaled random variable. Then,

$$\lim_{n\to\infty}\mathbb{P}[\{Z_n \le x\}] = \Phi(x), \quad \text{for all } x \in \Re.$$

In this chapter, we are concerned with finite collections of independent samples of a random variable, using these samples to estimate properties of the random variable X. Unlike the limit results of **??**, we want to estimate the accuracy we can obtain from a fixed finite number of samples n. We consider problems in both estimation and detection. For instance, we want to estimate the average height of women in the Boston are by measuring the height of 100 women, uniformly selected from Boston's population. How accurate will our estimate be? As another instance, consider conducting a trial for a new vaccine trial with a test group of 100 subjects and a control group of another 100 subjects. Do the results indicate that the vaccine makes a significant difference, and what confidence do we have in that conclusion?

9.1 Estimation of Mean and Variance

If we don't know the true mean, but can collect independent samples of X, the sample mean M_n is often a reasonable estimator for the true mean $\mathbb{E}[X]$. The sample mean is computed by generating n independent, identically distributed $X_k, k = 1, \ldots, n$, each of which is identically distributed as X. In this case,

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

has mean $\mathbb{E}[X]$, and M_n converges to $\mathbb{E}[X]$ by the Strong Law of Large Numbers. If X has finite variance σ^2, M_n is the sum of independent, identically distributed random variables, and thus has variance $\frac{\sigma^2}{n}$.

Suppose that we would like to estimate the variance σ^2 of X. Assuming the variance is finite, it is obtained as

$$\mathsf{Var}[X] = \sigma^2 = \mathbb{E}[\Big(X - \mathbb{E}[X]\Big)^2].$$

Given knowledge of $\mathbb{E}[X]$, and samples $X_1, \ldots, X_n$, an estimate of the variance can be obtained as

$$\widehat{V}_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mathbb{E}[X])^2.$$

Since the X_k are independent and identically distributed as X, we have

$$\mathbb{E}[\widehat{V}_n] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[(X_k - \mathbb{E}[X])^2] = \frac{1}{n}\sum_{k=1}^{n}\mathsf{Var}[X] = \mathsf{Var}[X].$$

By the Strong Law of Large Numbers, we know $\lim_{n\to\infty}\widehat{V}_n = \mathsf{Var}[X]$ with probability 1.

What if we did not know the mean $\mathbb{E}[X]$, but had only the sample values $X_k, k = 1, \ldots, n$ to estimate the variance? In this case, we may consider estimating the variance by using the sample mean M_n. That is,

$$\overline{V}_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - M_n)^2.$$

This can be simplified as

$$\begin{aligned}
\overline{V}_n &= \frac{1}{n}\sum_{k=1}^{n}(X_k^2 - 2X_kM_n + M_n^2) \\
&= \frac{1}{n}\sum_{k=1}^{n}X_k^2 - 2\big(\frac{1}{n}\sum_{k=1}^{n}X_k\big)M_n + M_n^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}X_k^2 - 2M_n^2 + M_n^2 = \frac{1}{n}\sum_{k=1}^{n}X_k^2 - M_n^2
\end{aligned}$$

Using the previous equation, we can compute the expected value of this estimate as

$$\mathbb{E}[\overline{V}_n] = \mathbb{E}[X^2] - \mathbb{E}[M_n^2] = \mathbb{E}[X]^2 + \mathsf{Var}[X] - \mathbb{E}[X]^2 - \frac{\mathsf{Var}[X]}{n} = \frac{n-1}{n}\mathsf{Var}[X].$$

This shows that the estimate $\overline{V}_n$ is a biased estimate of $\mathsf{Var}[X]$, and underestimates it by a small amount. To compensate for this, one can use the unbiased estimate:

$$V_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2.$$

This sample variance is an unbiased estimate of the true variance of X based on the samples $X_1, \ldots, X_n$. Most computer packages compute the sample variance as V_n.

What about an estimate of the standard deviation? While we can generate different estimates for the standard deviation directly, the common definition the sample standard deviation is

$$\widehat{\sigma}_n = \sqrt{V_n}.$$

This guarantees the consistent interpretation that the sample standard deviation is the square root of the sample variance,

9.2 Confidence Intervals for Sample Means

In the press, we read reports that quote statistics such as $57\% \pm 3\%$ of responders prefer brand A to brand B, with confidence interval 95%. How were such numbers calculated? We discuss this in this section.

Assume we have a random variable X, and we collect n independent samples X_k of X. Assume X has finite mean μ and variance σ^2. The sample mean of X, given n samples X_i is $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, which is a random variable. From the previous analysis, we know

$$\mathbb{E}[M_n] = \mathbb{E}[X] = \mu;\ \mathsf{Var}[M_n] = \frac{\mathsf{Var}[X]}{n} = \frac{\sigma^2}{n}.$$

M_n is an estimate of $\mathbb{E}[X]$. Given a small constant α, we want to find an interval $[A, B]$ such that

$$\mathbb{P}[\{A \leq \mathbb{E}[X] \leq B\}|M_n] = 1 - \alpha.$$

If we find such numbers, $B - A$ is called the confidence interval and $1 - \alpha$ the confidence.

Often, we select the interval to be centered about the sample mean M_n, in order to determine how close M_n is to the true mean $\mathbb{E}[X]$. Specifically, consider the event $\{|M_n - \mu| < \epsilon\}$ for some $\epsilon > 0$. Given the statistical properties of M_n, we can compute $\mathbb{P}[\{|M_n - \mu| < \epsilon\}] = q$. We say that the true mean is in the interval $[M_n - \epsilon, M_n + \epsilon]$ with confidence q.

We can use several of the limit theorems from Chapter 8 to estimate these confidence intervals. The variance of M_n is $\frac{\sigma^2}{n}$, which is small for large values numbers of samples n. If we know σ^2, the Chebyshev inequality yields

$$\mathbb{P}[\{|M_n - \mu| \geq \epsilon\}] \leq \frac{\sigma^2}{n\epsilon^2}.$$

Thus, $\mathbb{P}[\{|M_n - \mu| < \epsilon\}] \geq 1 - \frac{\sigma^2}{n\epsilon^2}$, yielding an estimate of the confidence level $q = 1 - \frac{\sigma^2}{n\epsilon^2}$ for fixed values of n, ϵ.

If the random variables X_k are bounded with values in $[a, b]$, we can use Hoeffding's inequality to get an improved confidence level:

$$\mathbb{P}[\{|M_n - \mu| \geq \epsilon\}] \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Thus, the true mean is in the interval $[M_n - \epsilon, M_n + \epsilon]$ with confidence level $q = 1 - 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}$.

For large n, we can approximate this probability using the Central Limit Theorem (CLT). The CLT states that the random variable $Z = \frac{\sqrt{n}(M_n - \mu)}{\sigma}$ has the distribution of a standard Gaussian random variable, so $Z \sim \mathcal{N}(0, 1)$. Then,

$$\mathbb{P}[\{|M_n - \mu| \geq \epsilon\}] = \mathbb{P}[\{|Z| \geq \frac{\sqrt{n}\epsilon}{\sigma}\} = 2(1 - \Phi(\frac{\epsilon\sqrt{n}}{\sigma})) = 2Q(\frac{\epsilon\sqrt{n}}{\sigma}).$$

Thus, the true mean is in the interval $[M_n - \epsilon, M_n + \epsilon]$ with confidence level $q = 1 - 2Q(\frac{\epsilon\sqrt{n}}{\sigma})$.

If we know ϵ and n, we can estimate the confidence level that $|M_n - \mu| < \epsilon$ using the above results. What if we knew ϵ and the desired confidence level $1 - \alpha$, and wanted to know how large n had to be to get that confidence level for that interval?

To answer this, if we are using the CLT, we determine a value T such that $Q(T) = \alpha/2$ using the standard Gaussian CDF table in Appendix C. Then,

$$\mathbb{P}[\{|\frac{\sqrt{n}(M_n - \mu)}{\sigma_X} \leq T\}] \approx 1 - \alpha,$$

or equivalently,

$$\mathbb{P}[|M_n - \mu| \leq \frac{T\sigma_X}{\sqrt{n}}] \approx 1 - \alpha.$$

This translates to the following statement: with confidence level $1-\alpha$, the true expected value μ lies in the interval $[M_n - \frac{T\sigma_X}{\sqrt{n}}, M_n + \frac{T\sigma}{\sqrt{n}}]$. The length of the confidence interval is $2\frac{T\sigma}{\sqrt{n}}$. To determine the number of samples n required for an interval of length 2ϵ, we solve

$$\frac{T\sigma_X}{\sqrt{n}} = \epsilon \Longrightarrow n = (\frac{T^2\sigma^2}{\epsilon^2}).$$

Similarly, we can determine how large n needs to be using other bounds such as the Chebyshev bound or Hoeffding's Inequality. We illustrate this process in the examples below.

Example 9.1
Suppose X is a Bernoulli(0.25) random variable. Assume we collect 100 independent, identically distributed samples of X, denoted as $X_k, k = 1, \ldots, 100$. Define $M_{100} = \frac{1}{100}\sum_{k=1}^{100} X_i$. Estimate the probability $\mathbb{P}[\{|M_{100} - 0.25| > 0.01\}]$.

The variance of a Bernoulli(p) random variable is $p(1-p)$. Thus, the variance of X is $\frac{3}{16}$, and the standard deviation is $\frac{\sqrt{3}}{4}$. Using the Chebyshev Inequality, we obtain

$$\mathbb{P}[\{|M_{100} - 0.25| > 0.01\}] \leq \frac{\frac{3}{16}}{100 \cdot 0.01^2} \leq \frac{3}{0.16} = 18.75$$

which is a useless estimate, as we know probabilities are less than 1. This means we don't have enough samples to estimate the mean of X accurately.

Since Bernoulli random variables take values in $[0, 1]$, Hoeffding's inequality yields

$$\mathbb{P}[\{|M_{100} - \mu| \geq 0.01\}] \leq 2e^{-\frac{200(0.01)^2}{1^2}} = 2e^{-0.02}$$

which is also a number greater than one, so it is not a useful bound.

What about the estimate from the Central Limit Theorem? In this case, M_n is approximated by a Gaussian with mean 0.25 and variance $\frac{3}{16\cdot 100} = \frac{3}{1600}$. The transformation $Z = \frac{M_n - 0.25}{\sqrt{\frac{3}{1600}}} = \frac{40(M_n - 0.25)}{\sqrt{3}}$ makes Z a standard Gaussian random variable. The event $\{|M_n - 0.25| > 0.01\}$ is equivalent to the event $\{|Z| > \frac{0.4}{\sqrt{3}}\}$. Thus, we can estimate the desired probability as

$$\mathbb{P}[\{|M_n - 0.25| > 0.01\}] \approx \mathbb{P}[\{|Z| > \frac{0.4}{\sqrt{3}}\}] = 2Q(\frac{0.4}{\sqrt{3}}) \approx 0.8174.$$

Example 9.2
Continuing the example 9.1, we would like to estimate the required number of samples n so that the sample mean $M_n \in [\mu - 0.01, \mu + 0.01]$ with confidence 0.95.

Using the Chebyshev inequality, we want

$$\mathbb{P}[\{|M_n - 0.25| > 0.01\}] \leq \frac{\frac{3}{16}}{n \cdot 0.01^2} \leq 0.05$$

Combining the last two equations, we get $n \geq \frac{\frac{3}{16}}{5\cdot(0.01)^3} = \frac{300{,}000}{8} = 37{,}500$. It is clear why 100 samples were inadequate in the previous example.

Using Hoeffding's inequality yields

$$\mathbb{P}[\{|M_n - \mu| \geq 0.01\}] \leq 2e^{-\frac{2n(0.01)^2}{1^2}} = 2e^{-\frac{2n}{10000}}.$$

Let n be such that

$$2e^{-\frac{2n}{10000}} = 0.05 \iff -\frac{2n}{10000} = \ln(0.025) \iff n = 5000\ln(40) \approx 18{,}444.$$

Using the Central Limit Theorem, we get the following estimate:

$$\mathbb{P}[\{|M_n - 0.25| > 0.01\}] \approx 2Q(\frac{0.01}{\sqrt{\frac{3}{16n}}}) \leq 0.05.$$

We use the standard Gaussian table in Appendix C to find the value of z^* such that $Q(z) = 0.025$, or equivalently, $\Phi(z) = 0.975$. Looking at the table, we find $z^* = 1.96$. Hence, as long as $\frac{0.01}{\sqrt{\frac{3}{16n}}} > z^*$, we have $Q(\frac{0.01}{\sqrt{\frac{3}{16n}}}) \leq 0.025$. Simplifying the above inequality, we get

$$\frac{0.01\sqrt{16n}}{\sqrt{3}} > 1.96 \Longrightarrow n > (1.96\sqrt{3})^2 \cdot (25)^2 \approx 7203.$$

We see that the estimate obtained from the Central Limit Theorem can give us the required confidence for $M_n \in [\mu - 0.01, \mu + 0.01]$ with a smaller number of samples than the estimate from the Chebyshev Inequality or Hoeffding's inequality.

Example 9.3
Let's ask a different question related to example 9.1: Given that you have collected 1000 samples $X_k, k = 1, \ldots, 1000$, what is the 95% confidence interval around μ for the estimate M_{1000}?

Using the Chebyshev Inequality, we have

$$\mathbb{P}[\{|M_{1000} - 0.25| > \epsilon\}] \leq \frac{\frac{3}{16}}{1000 \cdot \epsilon^2} \leq 0.05$$

This implies

$$\epsilon^2 \geq \frac{6}{1600} \Longrightarrow \epsilon \geq \frac{\sqrt{6}}{40} \approx 0.3873.$$

Using the CLT, we get $Q(\frac{\epsilon}{\sqrt{\frac{3}{16000}}}) = 0.025$, which means that

$$\frac{\epsilon}{\sqrt{\frac{3}{16000}}} = 1.96 \Longrightarrow \epsilon = 1.96\sqrt{3/16000} \approx 0.0268.$$

Using Hoeffding's inequality, we get $e^{-\frac{2000(\epsilon)^2}{1^2}} = 0.025$ which means that $\epsilon^2 = \frac{\ln(40)}{2000} \Longrightarrow \epsilon = 0.0429$.

Example 9.4
We are taking measurements of an unknown distance d, and the measurements are noisy. Hence, we assume that a measurement $X = d + W$, where W is a zero-mean random variable with variance σ^2. Hence, $\mathbb{E}[X] = d$, $\text{Var}[X] = \sigma^2$. We can repeat this measurement n times, resulting in n independent, identically distributed measurements $X_k, k = 1, \ldots, n$. We will estimate d as the sample mean of these measurements, as

$$\hat{d} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Suppose we want to determine how many measurements are needed to obtain 99% confidence interval that the error $|\hat{d} - d| \leq 0.1$? Assuming n is large, so that we use the Central Limit Theorem approximation, so that the random variable $\frac{\sqrt{n}(\hat{d}-d)}{\sigma}$ is approximated by a standard Gaussian random variable with mean 0, variance 1. Using the standard Gaussian table in Appendix C, we determine a value z^* such that $Q(z) = 0.005$, or equivalently, $\Phi(z) = 0.995$. Looking at the table, we find $z^* = 2.575$. This implies that $\mathbb{P}[\{\frac{\sqrt{n}(\hat{d}-d)}{\sigma}| \leq 2.575\}] = 0.99$, or equivalently, $\mathbb{P}[\{|\hat{d} - d| \leq 2.575\frac{\sigma}{\sqrt{n}}\}] = 0.99$.

We want to find n so that the 99% confidence interval is $|\hat{d} - d| \leq 0.1$. Hence, we must select n such that $2.575\frac{\sigma}{\sqrt{n}} \leq 0.1$. This requires $n \geq (25.75)^2\sigma^2$. For $\sigma = 1$, this is approximately 663 samples.

Example 9.5
Suppose we measure the response time X of a service system, and are interested in estimating the mean response time. The 10 measurements we collect are listed in the observation vector $\underline{Y}$ below:

$$\underline{Y} = \begin{bmatrix}41.6 & 41.48 & 42.34 & 41.95 & 41.86 & 42.18 & 41.72 & 42.26 & 41.81 & 42.04\end{bmatrix}^T$$

The sample mean is $M_{10} = 41.924$, which is an approximation of $\mathbb{E}[X]$. Suppose we know $\sigma_X = 0.1$. We want to find a 95% confidence interval for $\mathbb{E}[X]$.

Going to our table for $Q(\cdot)$, we try to find a value T such that $Q(T) = 0.025$. We find that $T \approx 1.96$. Then,

$$\mathbb{P}\left[\{|\mathbb{E}[X] - M_{10}| \leq \frac{1.96(0.1)}{\sqrt{10}} \approx 0.062\right] \approx 0.95.$$

Thus, we say that $\mathbb{E}[M] \in [41.862, 41.986]$ with confidence 95%.

In the previous examples, we assume that we know the variance of X, denoted by σ^2. In many practical situations, we don't know the variance, but have only the observed sample values. We have two approaches for this: one is to use an upper bound on the standard deviation, computed from the properties of the random variable in question. For instance, the variance of a Bernoulli(p) random variable is $p(1-p)$. For any value of p, this number is less than or equal to 0.25, so that the standard deviation is bounded above by 0.5. We can use similar approaches with other types of random variables, provided we have bounds on their parameters.

If the random variable X is bounded with range $R_X \subset [a, b]$, we can use Hoeffding's inequality, which does not require knowledge of the variance, but instead uses knowledge of the bounds on the range of X. Alternatively, we can bound the variance by $\frac{(b-a)^2}{4}$, the largest variance any random variable can have with range $R_X \subset [a, b]$.

A second approach is to use the sample standard deviation as a substitute for the true standard deviation. We illustrate this with examples below.

Example 9.6
We are interested in estimating the probability p that people like bananas. We want a confidence interval of length 0.06 around our estimate, with confidence level 95.5%, corresponding to $T = 2$ standard deviations. How many people do we need to poll, assuming that the opinions of people are independent?

Note that the answer any one person gives is a Bernoulli random variable, which is 1 if they like bananas, and 0 if they don't. We don't allow "I don't know" responses...Thus, if we knew p, the variance in the random variable X corresponding to a response would be $p(1-p)$, which is a number less than 0.25. Let's use this as a bound for the true variance which we don't know. Let the response of person k be X_k, and let $Z_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Then,

$$\mathbb{P}\left[\{|Z_n - p| \leq \frac{2\sqrt{0.25}}{\sqrt{n}}\}\right] \geq 0.955.$$

To get the confidence interval we want, we must have $\frac{2\sqrt{0.25}}{\sqrt{n}} = \frac{1}{\sqrt{n}} = \frac{0.06}{2} = 0.03$. Hence, $\sqrt{n} \approx \frac{100}{3}$, so $n \approx 1112$ persons need to be interviewed. We could reduce this number somewhat by estimating the variance of the specific responses adaptively. By using a bound, we get a conservative number to interview that does not depend on the actual responses.

Note that another bound on the variance is used in Hoeffding's inequality: when the range of X is bounded by $[a, b]$, the variance is bounded by $(b-a)^2/4$.

Example 9.7
Let's return to Example 9.5. Assume we did not know the variance of X. Let M_n denote the mean response time given n observed data. We can estimate the variance using the estimator $V_n = \frac{1}{n-1}\sum_{k=1}^{n}(Y_k - M_n)^2$.

For the data provided in Example 9.5, with $n = 10$ samples, the variance estimate is $V_{10} = 0.081$. Taking the square root yields a sample standard deviation of 0.284.

Now, with only 10 measurements, the 95% confidence interval would be

$$\mathbb{P}\left[\{|\mathbb{E}[X] - M_{10}| \leq \frac{1.96(0.284)}{\sqrt{10}} \approx 0.175\}\right] \approx 0.95.$$

Thus, our confidence interval increases almost by a factor of 3: $\mathbb{E}[X]$ is in the interval [41.75, 42.1] with confidence 95%.

Example 9.8
Here is an example we use in many engineering applications. You are trying to estimate the reliability of a system by using a simulation program that introduces the various random effects that can cause system failures. Note that, in each simulation, the system either fails or not, and hence the outcome of each simulation is a Bernoulli random variable X_k, where $X_k = 1$ indicates success. The reliability we are trying to estimate is $\mathbb{E}[X]$. If we conduct 100 simulations, and the estimated reliability $\hat{p} = M_{100} = 0.95$, and the sample variance V_n is 0.05, what can we say regarding the confidence interval and the level of confidence for this estimate?

Let's look for the 0.955 confidence level interval, corresponding to a threshold of two standard deviations. With $V_n = 0.05$, the sample standard deviation is 0.223. With 100 simulations, the length of the confidence interval is

$$\mathbb{P}[\{|M_{100} - \mathbb{E}[X]| \leq \frac{2 \cdot (0.223)}{\sqrt{100}}\}] \geq 0.955.$$

Thus, our true reliability $\mathbb{E}[X] \in [0.905, 0.995]$ with confidence 95.5%. Note that this is an estimate, because the sample variance V_n was random, and not a bound on the true variance.

What if we increased the number of simulations to 2500? Then our confidence interval tightens significantly, so $\mathbb{E}[X] \in [0.936, 0.964]$ with confidence 95.5%. The important relationship is that the width of the confidence interval is inversely proportional to the square root of the number of simulations.

We conclude this section by referencing some examples illustrating how confidence intervals are used. In 2008, a Gallup survey (https://news.gallup.com/poll/105850/ownership-may-good-wellbeing.aspx) was conducted to determine whether TV ownership was good for wellbeing. People questioned were asked to rate their life on a scale of 0 to 10. Specifically, they were asked: "Please imagine a ladder with 11 steps, numbered zero to 10, where the top represents the best possible life for you, and the bottom represents the worst possible life for you, which step comes closer to the way you feel about your life?" The responses were sorted into those that came from households with TVs, and those that came from households without TVs.

Note that the answers are integers from 0 to 10. Just like we did for Bernoulli replies, we can bound the variance of the responses by the variance of a discrete uniform distribution on $\{0, 1, \ldots, 10\}$, which is 10. Hence, the standard deviation is $\sqrt{10}$. For a population of 810, a 95% confidence level would result in a confidence interval of ± 0.24.

Typical outcomes of this poll are statements such as: "For the European data, one can say with 95% confidence that the true population for wellbeing among those without TVs is between 4.88 and 5.26." This estimate resulted from a sample of 810 persons that did not have TVs in their home. Note that this confidence interval (± 0.19) is narrower than the worst-case interval above, indicating that the Gallup survey used a standard deviation estimate based on the responses that was smaller than the worst-case estimate. Similarly, another statement in the poll was "For those with TVs the 95% confidence interval for well-being is much narrower – between 5.78 and 5.82 – because of the larger sample size." In this case, the poll included 40,267 households with TVs in their home. An increase in the number of samples by a factor of nearly 50 reduced the confidence interval by a factor of close to 7. The ratio is not exactly $\sqrt{n}$ because the estimate of the standard deviation also changed.

Given that 2020 is the year of the U.S. Census, one should note that the U.S. Census Bureau routinely uses confidence levels of 90% in their surveys, which is about 1.645 standard deviations. One survey of the number of people in poverty in 1995 stated a confidence level of 90% for the statistics: "The number of people in poverty in the United States is 35,534,124 to 37,315,094." That means if the Census Bureau repeated the survey using the same techniques, 90 percent of the time the results would fall between 35,534,124 and 37,315,094 people in poverty. The stated figure (35,534,124 to 37,315,094) is the confidence interval. Now you know a little more as to how to interpret such statistical statements that appear in our news reports.

9.3 Sampling Gaussian Random Variables

In the previous sections, we did not assume that the variable X that had n independent, identically distributed samples was Gaussian. For large n, we were able to use properties like the Central Limit Theorem

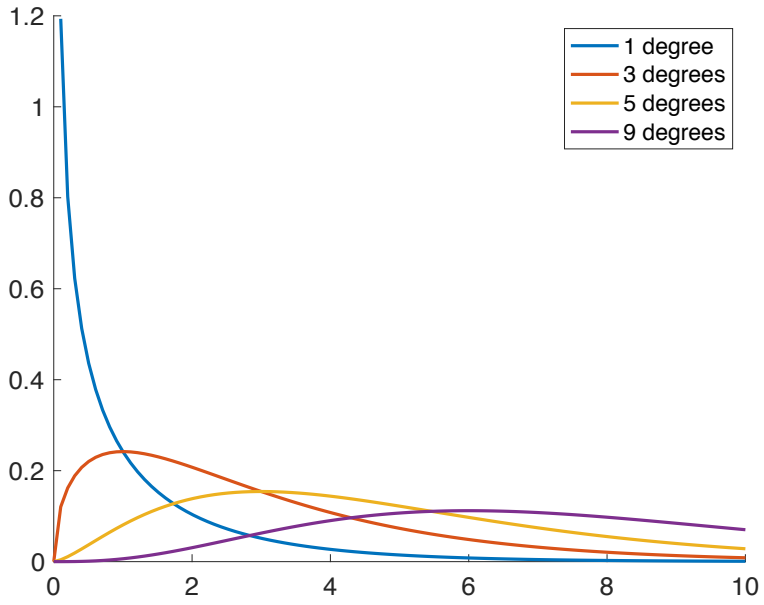


Figure 9.1: PDF of *chi*-squared random variables with different degrees of freedom.

to approximate the distribution of the sample mean as Gaussian, and get confidence intervals for estimates of the sample mean. However, we could not do the same for estimates of the sample variance, or for small number of samples n.

When $X_k, k = 1, \ldots, n$ are Gaussian with mean μ and variance σ^2, the sample mean M_n is Gaussian, and we can use Gaussian properties to get confidence intervals for small values of n. We have

$$\mathbb{P}[\{|M_n - \mu| \geq \epsilon\}] = 2(1 - \Phi(\frac{\epsilon\sqrt{n}}{\sigma})) = 2Q(\frac{\epsilon\sqrt{n}}{\sigma}).$$

What about the sample variance? The estimate of the sample variance is $V_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2$. This random variable is now the sum of squares of random variables. We introduce two new classes of continuous random variables which will be used to analyze properties of the random variance.

Definition 9.1
Let $X_1, \ldots, X_n$ be independent, standard Gaussian random variables with mean 0, variance 1. Define the random variable $Y = X_1^2 + \ldots + X_n^2$. Then, Y is said to be a chi-squared random variable with n degrees of freedom. We write this as $Y \sim \chi^2(n)$.

Figure 9.1 shows the probability density function for Student's t random variables with different degrees of freedom.

We can derive the following properties for $Y \sim \chi^2(n)$:

- $\mathbb{E}[Y] = \sum_{k=1}^{n} \mathbb{E}[X_k^2] = n$.
- $\mathbb{E}[Y^2] = \sum_{j=1}^{n}\sum_{k=1}^{n} \mathbb{E}[X_j^2 X_k^2]$. We can compute each term in the sum as

$$\mathbb{E}[X_j^2 X_k^2] = \begin{cases} \mathbb{E}[X_j^2]\mathbb{E}[X_k]^2 = 1 & j \neq k \\ \mathbb{E}[X_k^4] = 3 & j = k. \end{cases}$$

Thus, $\mathbb{E}[Y^2] = 3n + n^2 - n = 2n + n^2$.

- $\mathsf{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = 2n$.
- Let $U \sim \chi^2(n)$ and $V \sim \chi^2(m)$ be independent random variables. Then, $Y = U + V \sim \chi^2(m+n)$.

Like standard Gaussians, the CDF of chi-squared random variables is tabulated and used to compute probabilities of intervals. An important property of chi-squared random variables is to analyze the statistics of estimates of the sample variance, when the underlying random variables X_k are Gaussian.

Let $X_1, \ldots, X_n$ be independent, identically distributed Gaussian random variables with $X_k \sim \mathcal{N}(\mu, \sigma^2)$. The sample mean and variance are:

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k; \quad V_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2.$$

Then, we will show that the random variable $Y = \frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k - M_n)^2$ is a $\chi^2(n-1)$ random variable. Moreover, Y and M_n are independent random variables. Note that Y is proportional to the sample variance, as $Y = \frac{n-1}{\sigma^2}V_n$.

Let's first show that Y and M_n are independent random variables. Write $\sigma^2 Y$ as

$$\sigma^2 Y = \sum_{k=1}^{n}(X_k - M_n)^2 = (X_1 - M_n)^2 + \sum_{k=2}^{n}(X_k - M_n)^2 = \Big(\sum_{k=2}^{n}(X_k - M_n)\Big)^2 + \sum_{k=2}^{n}(X_k - M_n)^2$$

where the last equality follows because $\sum_{k=1}^{n}(X_k - M_n) = 0$. We know X_k are i.i.d. and Gaussian. Let's define a linear variable transformation as follows: $W_1 = M_n$; $W_2 = X_2 - M_n$; $W_3 = X_3 - M_n$; $\cdots W_n = X_n - M_n$. This is a linear transformation, so the variables W_k are Gaussian, and zero-mean. Furthermore, the inverse of the transformation is

$$X_2 = W_2 - W_1;\ X_3 = W_3 + W_1; \cdots X_n = W_n + W_1; X_1 = W_1 - W_2 - \cdots - W_n.$$

As a matrix, we write this as

$$\underline{X} = \mathbf{A}\underline{W} = \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \underline{W}.$$

Note that $\det[\mathbf{A}] = n$. Since the X_k are independent, we have

$$f_{\underline{X}}(x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\sum_{k=1}^{n}\frac{(x_k-\mu)^2}{2\sigma^2}}.$$

Using the linear transformation, the joint PDF of $\underline{W}$ is Gaussian, and given by

$$f_{\underline{W}}(w_1, \ldots, w_n) = \frac{n}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{(w_1 - \sum_{k=2}^{n} w_k - \mu)^2}{2\sigma^2}} e^{-\sum_{k=2}^{n}\frac{(w_k w_1 - \mu)^2}{2\sigma^2}}.$$

Let's expand and regroup the quadratic in the exponent, as

$$\begin{aligned}(w_1 - \sum_{k=2}^{n} w_k - \mu)^2 + \sum_{k=2}^{n}(w_k - w_1 - \mu)^2 &= w_1^2 - 2w_1\sum_{k=2}^{n}(w_k - \mu) + \Big(\sum_{k=2}^{n} w_k - \mu\Big)^2 + \\ &\sum_{k=2}^{n}(w_k - \mu)^2 + 2w_1\sum_{k=2}^{n}(w_k - \mu) + \sum_{k=2}^{n} w_1^2 \\ &= n w_1^2 + \sum_{k=2}^{n}(w_k - \mu)^2 + \Big(\sum_{k=2}^{n} w_k - \mu\Big)^2\end{aligned}$$

Hence,

$$f_{\underline{W}}(w_1,\ldots,w_n) = \frac{n}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{nw_1^2}{2\sigma^2}} e^{frac\sum_{k=2}^n (w_k-\mu)^2 + \left(\sum_{k=2}^n w_k - \mu\right)^2 2\sigma^2},$$

which shows that W_1 is independent of $W_2, W_3, \ldots, W_n$.

Observe that $M_n = W_1$, and $Y = \frac{1}{\sigma^2}\Big(\sum_{k=2}^n W_k^2 + \big(\sum_{k=2}^n W_k\big)^2\Big)$. Hence M_n and Y are independent.

To show that Y is a *chi*-squared random variable with $n-1$ degrees of freedom, note the following:

$$U = \sum_{k=1}^n \frac{(X_k-\mu)^2}{\sigma^2}$$

is a *chi*-squared random variable with n degrees of freedom. Then,

$$\begin{aligned} U &= \sum_{k=1}^n \frac{(X_k - M_n + M_n - \mu)^2}{\sigma^2} = \sum_{k=1}^n \frac{(X_k-M_n)^2}{\sigma^2} + 2\sum_{k=1}^n \frac{(X_k-M_n)(M_n-\mu)}{\sigma^2} + \sum_{k=1}^n \frac{(M_n-\mu)^2}{\sigma^2} \\ &= Y + 2(M_n-\mu)\sum_{k=1}^n \frac{(X_k-M_n)}{\sigma^2} + n\frac{(M_n-\mu)^2}{\sigma^2} \\ &= Y + n\frac{(M_n-\mu)^2}{\sigma^2} \end{aligned}$$

where the middle term vanishes because M_n is the sample mean of the X_k. The last term is the square of a standard Gaussian random variable also, because $\mathbb{E}[M_n] = \mu$, $\mathsf{Var}[M_n] = \frac{\sigma^2}{n}$. So, we have $V = Y + Z$, where $V \sim \chi^2(n), Z \sim \chi^2(1)$ and Z is independent of Y. This means that Y is a *chi*-squared random variable with $n-1$ degrees of freedom.

Another standard distribution that is used in statistics is the Student's t-distribution. The CDF of this distribution is also tabulated. Let Z be a standard Gaussian random variable, and let Y be a *chi*-squared distributed random variable with n degrees of freedom, that is independent of Z. Then, the random variable

$$W = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

has a Student's t-distribution with n degrees of freedom, abbreviated as $W \sim T(n)$. Figure 9.2 shows the PDF of a Student's t-distribution with different degrees of freedom, as well as a standard Gaussian PDF. The plots illustrate that the Student's t-distribution approaches a standard Gaussian PDF as the number of degrees of freedom increases.

The following properties of $W \sim T(n)$ are stated without proof:

- For $n > 1$, $\mathbb{E}[W] = 0$. For $n = 1$, $\mathbb{E}[W]$ is undefined.
- For $n > 2$, $\mathsf{Var}[W] = \frac{n}{n-2}$. For $n = 1, 2$, $\mathsf{Var}[W]$ is undefined (infinite).
- For large n, the density of W approaches $\mathcal{N}(0,1)$.
- The PDF of W is an even function, symmetric about 0.

Why are Students' t-distributions important? Given $X_1, X_2, \ldots, X_n$ i.i.d. Gaussian random variables with mean μ and variance σ^2, and let M_n, V_n denote the sample mean and variance of these variables. Let $\widehat{\sigma} = \sqrt{V_n}$ denote the sample standard deviation. Then,

$$W = \frac{\sqrt{n}(M_n-\mu)}{\widehat{\sigma}} = \frac{\sqrt{n}(M_n-\mu)}{\sqrt{V_n}}$$

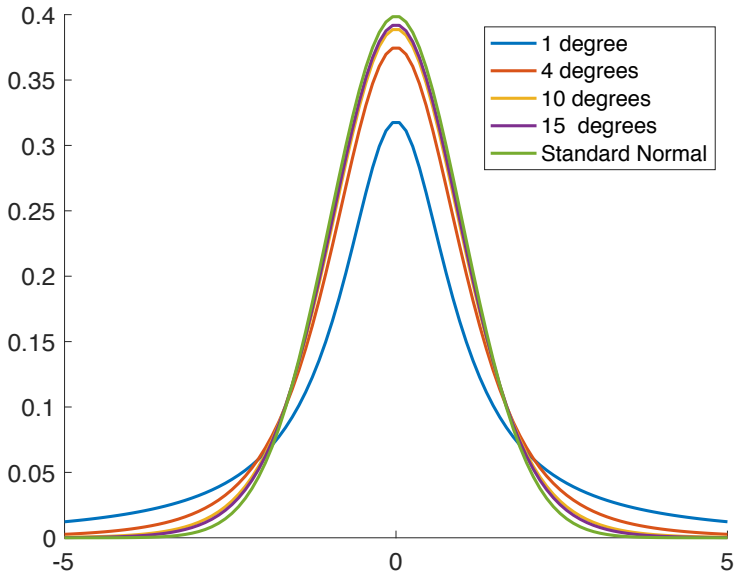


Figure 9.2: PDF of *chi*-squared random variables with different degrees of freedom.

has a Student's t-distribution with $n-1$ degrees of freedom ($W \sim T(n-1)$.)

To see this, note the following:

$$W = \frac{\sqrt{n}(M_n - \mu)}{\sqrt{V_n}} = \frac{\sqrt{n}(M_n - \mu)}{\sigma} \frac{\sigma}{\sqrt{V_n}}.$$

The variable $Z = \frac{\sqrt{n}(M_n-\mu)}{\sigma}$ is a standard Gaussian random variable. The variable $\frac{V_n}{\sigma^2}$ can be written as $\frac{V_n}{\sigma^2} = \frac{1}{n-1}Y$, where Y is a *chi*-squared random variable with $n-1$ degrees of freedom. Hence, the ratio is a Student's t-distribution with $n-1$ degrees of freedom.

We can use this to compute confidence intervals for samples of Gaussian random variables without specifying either the mean or variance of the distribution, as shown in the example below.

Example 9.9

Let's return to the problem of example 9.5, with the additional assumption that response time X of a service system is Gaussian with unknown mean μ and variance σ^2. We collect 10 independent measurements of X, listed in the observation vector $\underline{Y}$ below:

$$\underline{Y} = \begin{bmatrix} 41.6 & 41.48 & 42.34 & 41.95 & 41.86 & 42.18 & 41.72 & 42.26 & 41.81 & 42.04 \end{bmatrix}^T$$

The sample mean is $M_{10} = 41.924$, which is an approximation of $\mathbb{E}[X]$. The sample variance is 0.0807, and the sample standard deviation $\widehat{\sigma}$ is 0.284.

We want to find a 95% confidence interval for $\mathbb{E}[X]$. We have 10 samples, so $\frac{\sqrt{10}(M_{10}-\mu)}{\sqrt{V_{10}}}$ $T(9)$. We use Microsoft Excel or MATLAB to find the value for which the CDF of a $T(9)$ random variable has value 0.975, which is approximately 2.262.

Then,

$$\mathbb{P}\left[\{|\frac{\sqrt{10}(M_{10}-\mu)}{\sqrt{V_{10}}}| \leq 2.262\}\right] = \mathbb{P}\left[\{|M_{10}-\mu| \leq \frac{2.262 \cdot 0.284}{\sqrt{10}} \approx 0.236\}\right] = 0.95.$$

Thus, we say that $\mathbb{E}[M] \in [41.698, 42.160]$ with confidence 95%. The increase in the width of the confidence interval, when compared with the estimate of 9.7, is due to the uncertainty in the estimate of the standard deviation.

We can also get confidence intervals on the sample variance of a normal distribution. The sample variance,

based on the sample random variables $X_1, \ldots, X_n$, is

$$V_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2,$$

where M_n is the sample mean. We know that the random variable $Y = \frac{(n-1)V_n}{\sigma^2}$ has a *chi*-squared distribution with $n-1$ degrees of freedom. To find a $1-\alpha$ confidence interval, we look at the CDF of $Y \sim \chi(n-1)$ to determine thresholds $t_1, t_2 \geq 0$ such that

$$\mathbb{P}[\{Y \leq t_1\}] = \alpha/2]; \quad \mathbb{P}[\{Y \leq t_2\}] = 1 - \alpha/2.$$

This guarantees that

$$\mathbb{P}[\{t_1 \leq Y \leq t_2\}] = \mathbb{P}[\{t_1 \leq \frac{(n-1)\widehat{\sigma^2}}{\sigma^2} \leq t_2\}] = 1 - \alpha.$$

We can take inverses to obtain

$$\mathbb{P}[\{\frac{1}{t_1} \geq \frac{\sigma^2}{(n-1)\widehat{\sigma^2}} \geq \frac{1}{t_2}\}] = \mathbb{P}[\{\frac{(n-1)\widehat{\sigma^2}}{t_1} \geq \sigma^2 \geq \frac{(n-1)\widehat{\sigma^2}}{t_2}\}].$$

This gives a $(1-\alpha)$ confidence interval for the true variance σ^2 as $[\frac{(n-1)\widehat{\sigma^2}}{t_2}, \frac{(n-1)\widehat{\sigma^2}}{t_1}]$.

Example 9.10
For the problem of example 9.5, $n = 10$ and the sample variance is 0.0807. To obtain a 95% confidence interval, we compute the thresholds t_1, t_2 for $\alpha = 0.05$ using Microsoft Excel or MATLAB, and obtain $t_1 = 2.700, t_2 = 19.023$. This yields a 95% confidence interval that the true variance $\sigma^2 \in [0.038, 0.269]$. Our sample variance is in this interval, but the interval is large, as n is small.

9.4 Significance Testing based on Sample Statistics

In significance testing, we are interested in determining whether a set of observations show effects that differ significantly from those expected from a nominal model. The nominal model is our null hypothesis H_0, which describes the nominal probability distribution function of the observations. For simplicity, assume Y to be a continuous random variable described by a probability density function $f_{Y|H_0}(y)$. We observe a sample of that random variable, and we are interested in determining whether the sample of the random variable is consistent with the assumed distribution $f_{Y|H_0}(y)$. In contrast to binary hypothesis testing, there is no alternative hypothesis H_1 with a similar probability model for Y. Instead, the alternative is that H_0 is not the correct hypothesis. The question answered by significance testing is whether the observed value of Y is consistent with the hypothesis H_0, or whether the value is inconsistent, so that the hypothesis that Y was generated according to H_0 should be rejected.

The types of error that one makes in significance testing are denoted as Type I and Type II errors. A Type I error occurs when we reject the null hypothesis, declaring that the observed value of Y is inconsistent with the null hypothesis, even though the data was generated according to H_0. This error is a false positive, or a false alarm, using our nomenclature from Chapter **??**. A Type II error occurs when we declare that the observed value is consistent with the null hypothesis, even though it was not generated by a density corresponding to the null hypothesis. This type of error is a false negative, or a missed detection.

To design a test of significance for the null hypothesis, we start with a value of α, called the *level of significance*. We want to design a test such that the probability of false alarm is less than or equal to α. To do this, we select a set $R_0 \subset \Re$ of values such that $\mathbb{P}[\{Y \in R_0 | H_0\}] = \alpha$. The significance test declares the value is inconsistent and rejects H_0 if $Y \in R_0$, and fails to reject H_0 if the observed value $Y \notin R_0$.

There are many ways of selecting the set R_0 that satisfy $\mathbb{P}[\{Y \in R_0 | H_0\}] = \alpha$. The two most common ways are one-sided tests and two-sided tests. For a typical **one-sided test**, let $F_{Y|H_0}(y)$ denote the cumulative

distribution function of the observation Y conditioned on the null hypotheses. We select a value t_α such that $F_{Y|H_0}(t_\alpha) = 1-\alpha$, and select $R_0 = \{y > t_\alpha\}$. One-sided tests are appropriate for evaluating when the observed value of Y is too large to be consistent with the null hypothesis. In this case, α is the probability of a Type I error, namely, a false alarm. Although these tests appear to focus only on rejection of H_0 for values that are too high, we can extend this to values that are too low by considering the observation $-Y$ instead of Y.

A **two-sided test** is designed to test whether the observed random variable is either too high or too low to be consistent with the null hypothesis. Define t_l, t_h as follows:

$$F_{Y|H_0}(t_l) = \frac{\alpha}{2}; \qquad F_{Y|H_0}(t_h) = 1 - \frac{\alpha}{2}.$$

Define the reject set as $R_0 = \{y : y < t_l \text{ or } y > t_h\}$. Then, $\mathbb{P}[\{Y \in R_0\}|H_0] = \alpha$.

Given an observation $Y = y_0$, the **p-value** of y_0 is defined as the probability, under the null hypothesis, that you will observe a value as extreme or more extreme that y_0. For a one-sided test, the p-value is defined as $1 - F_{Y|H_0}(y_0)$. For a two-sided test, the definition is more nuanced, and depends on the specific nature of the CDF $F_{Y|H_0}(y_0)$; one definition is $2\min\left(F_{Y|H_0}(y_0), 1 - F_{Y|H_0}(y_0)\right)$. If the p-value is smaller than α, the null hypothesis is rejected. This is a different way of implementing the hypothesis test that does not require computing the inverse of the CDF $F_{Y|H_0}(y)$ to obtain a threshold.

Example 9.11

Our probability model for how late the Green Line is in arriving at its scheduled stop on St. Mary's street is that Y, the delay time in minutes, is an exponential random variable with rate parameter $\lambda = 0.5$, so that the expected delay time is 2 minutes. This is our null hypothesis. We are going to measure the observed delay time Y, and we want to design a significance test for hypothesis H_0 at a confidence level of $1 - \alpha = 0.95$, looking for evidence that the null hypothesis is inconsistent with the observed data if the measured delay time is too large.

The appropriate test is a one-sided test, as we are looking for delays that are too large to be consistent with the null hypothesis. Using the properties of exponential random variables, the probability distribution function of Y is

$$F_{Y|H_0}(y) = \begin{cases} 0 & y \le 0 \\ 1 - e^{-0.5y} & y > 0. \end{cases}$$

We want to define the reject set $R_0 = \{y > t_{0.05}\}$ for some threshold value $t_{0.05}$ that gives a confidence level of 0.95. Hence, we want $F_{Y|H_0}(t_{0.05}) = 1 - \alpha = 0.95$. Thus,

$$e^{-0.5t_{0.05}} = 0.05 \Rightarrow t_{0.05} = 5.9915.$$

Hence, our test of significance is $Y > 5.9915$, defining the region of measurements for which the null hypothesis is rejected.

For any measured value $Y = y$, its p-value is computed as $1 - F_{Y|H_0}(y) = e^{-0.5y}$. If the p-value of the measurement is less than the desired significance level $\alpha = 0.05$, the null hypothesis is rejected.

As the above example illustrates, the key to designing a test of significance is to identify the conditional probability distribution of the test statistic Y under the null hypothesis H_0. Using this conditional PDF $F_{Y|H_0}(y)$, we can compute thresholds for the appropriate significance level, and determine the p-values of measured test values $Y = y$.

9.4.1 The One Sample Z-Test

Consider the null hypothesis that the random variable X is a Gaussian random variable with known mean μ and variance σ^2. As an observation, we collect n independent observations of X, and want to accept or reject the hypothesis that the measurements were generated according to the null hypothesis. The one-sample Z test consists of determining whether the batch of n measurements is consistent with null hypothesis.

To make the decision, we use the sample mean of the observations as the statistic Y for the test. Thus, we test the hypothesis that the sample mean of the n observations $X_1, \ldots, X_n$ is consistent with the assumption that the observations were generated according to the null hypothesis.

This type of hypothesis test is known as a one-sample Z-test. Under the null hypothesis H_0, the sample mean $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is a Gaussian random variable with mean μ and variance $\frac{\sigma^2}{n}$. We want to design a test with a level of significance α that the sample mean is different from μ. The appropriate test is a two-sided test, as the sample mean can be either too large or too small.

The random variable $Z = \frac{\sqrt{n}(M_n - \mu)}{\sigma}$, referred to as the Z-statistic, is known to be a standard Gaussian random variable with mean 0 and variance 1. Given the value $M_n = \hat{\mu}_n$, the resulting Z-statistic is $z = \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma}$. The hypothesis test can be expressed in terms of the Z-statistic, as we want to find a threshold $T_{\alpha/2}$ so that $\mathbb{P}[\{|Z| > T_{\alpha/2}\}] = \alpha$, which is the same problem as finding a $1-\alpha$ confidence interval for the estimate M_n. The threshold is computed the same way: we find the value $T_{\alpha/2}$ so that $\Phi(-T_{\alpha/2}) = Q(T_{\alpha/2}) = \alpha/2$, or equivalently $\Phi(T_{\alpha/2}) = 1 - \alpha/2$. For instance, if $\alpha = 0.05$, then $T_{\alpha/2} = 1.96$. Then, if $|z| > T_{\alpha/2}$, the observations do not support the null hypothesis at a level of significance α.

An equivalent way of implementing a Z-test is to compute the p-value of the sample mean M_n, or equivalently, the Z- statistic. The p-value of a measurement is the probability of getting a measurement value that is more extreme than the current measurement. With a two-sided test and a Gaussian null hypothesis, the p-value of $Z = z$ is $\Phi(-z) + (1 - \Phi(z)) = 2\Phi(-|z|)$. If the p-value is less than the level of significance α, then the evidence indicates that the null hypothesis can be rejected at that level of significance. The advantage of this approach is that we don't have to compute the inverse of the standard Gaussian CDF Φ to compute a threshold.

Example 9.12
Assume that a probabilistic model for the weight of a randomly selected male person in the US is a Gaussian random variable measured in pounds, with mean 195, and standard deviation 30. We believe that Canadians have the same weight distribution, so we designed an experiment to weigh 100 randomly selected Canadian males, and compute their average weight, denoted as W_{ave}. Design a statistical test with significance level 0.01 to determine whether the measured W_{ave} supports the null hypothesis that the weight of Canadian males has the same probability model as the weight of US males.

The measured random variable is W_{ave}, which is the average of 100 independent samples of Canadian male weights. To answer the question, we need to compute the probability distribution of W_{ave} under the null hypothesis, given that

$$W_{ave} = \frac{1}{100}(W_1 + W_2 + \ldots + W_{100}).$$

Under the null hypothesis, the W_i's are independent Gaussian random variables, with mean 195 and standard deviation 30.

The Z statistic for this problem is $Z = \frac{10(W_{ave} - 195}{30}$. We want to define a two-sided test to accept or reject the null hypothesis with significance level 0.01, we are looking for a threshold $T_{0.005}$ such that, if $|Z| > T_{0.005}$, we will reject the hypothesis with significance level 0.01.

Thus, we need to select $T_{0.005}$ such that $Q(T_0.005) = 1 - \Phi(T_{0.005}) = 0.005$; this implies $T_0.005 = 2.576$.

We reject the null hypothesis with significance level 0.01 whenever $|Z| > 2.576$, or equivalently the average weight difference $|W_{ave} - 195| > 7.728$ pounds. Note the effect of selecting a sample size of 100 persons had in reducing the standard deviation of the test statistic W_{ave}. If we had weighed 9,000 Canadian males, the threshold would be much smaller, as the standard deviation of the sample mean would be 1, and now a smaller difference in average weight would be significant.

For this problem, we can compute the p-value of a measured $W_{ave} = W$, by computing $\mathbb{P}[\{|W_{ave} - 195| > |195 - W|\}|H_0]$ as the probability that the null hypothesis would yield a measurement more extreme than W. This yields a p-value for W of $2Q(\frac{|W-195|}{3})$.

Example 9.13
The lifetime of a certain cell type has been determined to be distributed according to a Gaussian distribution with mean 1570 hours and a standard deviation of 120 hours. You perform an experiment and measure the lifetime of 100 cells, and compute a sample mean lifetime of 1600 hours. Is the sample mean you measure significantly different from the population mean at a significance level of 0.05?

The Z statistic is $z = \frac{\sqrt{100}(1600-1570)}{120} = 2.5$. The p-value of z can be computed from Appendix C as $2\Phi(-2.5) = 0.0124$. Since the p-value is less than the significance level, we can reject the null hypothesis that the experiment lifetimes were sampled from a $\mathcal{N}(1570, 14400)$ distribution.

When the underlying null hypothesis is not a normal distribution, we can still use Z-tests provided that the number of samples n is sufficiently large (e.g. greater than 30). This is because the Z statistic will have an approximately Gaussian distribution, according to the Central Limit Theorem in Chapter **??**.

9.4.2 The One Sample T-Test

In the One Sample Z-Test, the null hypothesis assumed that both the mean and the standard deviation were known. In many applications, these parameters are rarely known. We discuss a different test, where we know the mean but not the standard deviation of the null hypothesis.

As in the Z-Test, we collect n observations of a random variable X, which is assumed under the null hypothesis H_0 to be Gaussian, with known mean μ, but with unknown variance σ^2. We would like to test the hypothesis that the sample mean $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is consistent with the null hypothesis at a level of significance α.

Note that we don't have a well-specified PDF for the sample mean. We know that $\mathbb{E}[M_n|H_0] = \mu$, and $f_{M_n|H_0}(x)$ is Gaussian, but we don't know its variance. Let's compute the sample variance V_n, and the sample standard deviation $\widehat{\sigma} = \sqrt{V_n}$ as described earlier. Then, transform M_n to a new random variable known as the T-statistic, as

$$T = \frac{\sqrt{n}(M_n - \mu)}{\widehat{\sigma}}.$$

If the null hypothesis is true, T is distributed according to a Student's t-distribution with $n-1$ degrees of freedom, as shown in Section 9.3. Thus, we know $f_{T|H_0}(t)$, and can perform a test of the null hypothesis with level of significance α.

The Student's t-distribution PDF is symmetric about 0. We use a two-sided test, so we compute threshold $t_{\alpha/2}$ so that $F_{T|H_0}(-t_{\alpha_2}) = \alpha/2$. Then, our decision rule is: if $|T| > t_{\alpha_2}$, we reject hypothesis H_0 at a level of significance α. Otherwise, we don't reject hypothesis H_0.

Equivalently, we compute the p-value of the computed T-statistic $T = t$, as $p = 2 * F(-|t|)$. If $p < \alpha$, we can reject hypothesis H_0 at a level of significance α.

Example 9.14
Consider the problem of example 9.13, except that we don't know the true standard deviation σ^2 of the lifetime of the cells. You perform an experiment and measure the lifetime of 100 cells, and compute a sample mean lifetime of 1600 hours and a sample standard deviation of 120 hours. Is the sample mean you measure significantly different from the population mean at a significance level of 0.05?

Compute the T-statistic:

$$t = \frac{\sqrt{n}(M_n - 1570)}{\widehat{\sigma}} = \frac{10(1600 - 1570)}{120} = 2.5.$$

The distribution of the T statistic is a Student's t-distribution with 99 degrees of freedom. Looking up the p-value for 2.5 in either MATLAB or Microsoft Excel, it is $2 \cdot 0.00703 = 0.01406$, which is less than 0.05, so the results support rejecting hypothesis H_0 with a level of significance 0.05.

Similarly, the threshold $t_{0.025}$ is 1.984. Since 2.5 is greater than that threshold, the results support rejecting hypothesis H_0.

Suppose we approximated the T- statistic distribution by a standard Gaussian distribution. What would be the corresponding threshold $t_{0.025}$? We have computed this to be 1.96. We see that the threshold using the correct distribution is slightly larger.

9.4.3 Two Samples T- and Z-tests

In one sample tests, we want to evaluate the null hypothesis that a collection of observations is consistent with a prior probability model. In two sample tests, we are interested in evaluating the null hypothesis that two sets of observations are consistent with a common probability model. We begin with the two-sample Z-tests.

Assume we have two Gaussian random variables X, Y, where $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Assume we collect a set of n_1 independent samples $X_1, \ldots, X_{n_1}$ of X, and n_2 independent samples $Y_1, \ldots, Y_{n_2}$ of Y. We want to test the null hypothesis that $\mu_1 = \mu_2$ with a level of significance α.

The sample mean of the first set, $M_{n_1}^{(1)} = \frac{1}{n_1}\sum_{k=1}^{n_1} X_k$, is a Gaussian random variable with mean μ_1 and variance $\frac{\sigma_1^2}{n_1}$. Similarly, the sample mean of the second set, $M_{n_2}^{(2)} = \frac{1}{n_2}\sum_{k=1}^{n_2} Y_k$ is a Gaussian random variable with mean μ_2 and variance $\frac{\sigma_2^2}{n_2}$. Random variables $M_{n_1}^{(1)}, M_{n_2}^{(2)}$ are independent.

Under hypothesis H_0, the difference $M_{n_1}^{(1)} - M_{n_2}^{(2)}$ is a Gaussian random variable with mean 0 and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, as it is the difference of two independent Gaussian random variables. We define the Z-statistic as

$$Z = \frac{M_{n_1}^{(1)} - M_{n_2}^{(2)}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Under H_0, Z is a Gaussian random variable with mean 0, variance 1. To evaluate H_0 with a level of significance α, we perform the same test as before: Compute the test statistic $Z = z$ based on the data. Then, compute threshold $t_{\alpha/2}$ such that $\Phi(-t_{\alpha/2}) = \alpha/2$, and determine whether $|z| > t_{\alpha/2}$. If it is, reject the null hypothesis H_0 with level of significance α. Equivalent, compute the p-value $p = 2\Phi(-|t|)$ and reject the null hypothesis with significance level α if $p < \alpha$.

Note that we don't need to know the values of $\mu_1 = \mu_2$ to conduct this Z-test. However, we do need to know the standard deviations of the two sets σ_1 and σ_2.

What if the variances σ_1^2, σ_2^2 were not known? We can use a simple generalization of the one-sample T-test when the unknown variances are assumed to be the same. We know $\frac{M_{n_1}^{(1)} - M_{n_2}^{(2)}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$ is a standard Gaussian random variable. We also know that $(n_1 - 1)\frac{V_{n_1}^{(1)}}{\sigma^2} + (n_2 - 1)\frac{V_{n_2}^{(2)}}{\sigma^2}$ is a *chi*-squared random variable with $n_1 + n_2 - 2$ degrees of freedom. Then, the T-statistic can be defined as

$$T = \frac{M_{n_1}^{(1)} - M_{n_2}^{(2)}}{\widehat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$\widehat{\sigma} = \sqrt{\frac{(n_1 - 1)V_{n_1}^{(1)} + (n_2 - 1)V_{n_2}^{(2)}}{n_1 + n_2 - 2}}$$

is the pooled variance.

The T- statistic has a Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom, and can now be used to accept or reject the null hypothesis with a desired level of significance.

When the variances are unequal and unknown, one can derive a more complex test with approximate numbers of degrees of freedom, known as Welch's t-test. This results in T-statistics that have fractional degrees of freedom. The details can be found in statistics books or in Wikipedia.

Example 9.15
To investigate the effect of a new hay fever drug on driving skills, a researcher studies 24 individuals with hay fever: 12 who have been taking the drug and 12 who have not. All participants then entered a simulator and were given a driving test which assigned a score to each driver as summarized in the table below:

Control	23	15	16	25	20	17	18	14	12	19	21	22
Drug	16	21	16	11	24	21	18	15	19	22	13	24

We want to test the null hypothesis that the drug has no adverse effects in decreasing the average score of the drivers with a lever of significance 0.05. We compute the sample mean and variance for the two groups as $M^{(1)} = 18.5, M^{(2)} = 18.33$, $V^{(1)} = 15.18, V^{(2)} = 17.88$. We assume the variances are the same, since the sampled variances are similar, and compute the pooled variance as $\widehat{\sigma}^2 = 16.53$. Given the mean values, the resulting pooled variance, and the number of samples n_1, n_2, the value of the T-statistic is 0.1004. The one-sided p-value of this T-statistic with 22 degrees of freedom is 0.46, which is much higher than the desired level of significance of 0.05. Thus, we fail to reject the null hypothesis and are 95% confident that any difference between the two groups is due to chance variations.

The two-sample T-tests and Z-tests depend on the assumption that the distribution of the underlying random variables from which the samples are generated is Gaussian. When that assumption is violated, we can still apply the T-tests and Z-tests as appropriate when the number of samples in each group n_1, n_2 are sufficiently large (greater than 30) so that the Central Limit Theorem allows us to use Gaussian distribution approximations for the sample means $M_{n_1}^{(1)}, M_{n_2}^{(2)}$.

Chapter 10

Machine Learning and Data Science

10.1 Introduction

In the previous chapters, when discussing problems of detection or estimation, we have assumed that we know the joint distribution (PMF or PDF) that describes the relationships between the random variables. In detection, we used the conditional distributions of the observed data given various hypotheses, as well as the prior distributions on the hypotheses, to design algorithms that selected the best among the hypotheses based on the observed data. In estimation, we used the joint distribution of the observed and unobserved data to predict the value of the unobserved variables, based on the observed variables.

We have also covered discussed foundational concepts in statistics, based on estimating parameters from observed data. We focused on the sample mean of a random variable, based on a set of independent, identically distributed samples of that random variable. We showed that the sample mean is a good estimator of the true mean in this case, and developed interesting bounds on how many samples we need to get a "good enough" estimate of the true mean with high probability.

In this chapter, we provide a short introduction to the field of machine learning. Roughly speaking, machine learning is about making inferences such as detection or estimation, without knowing the joint distribution of random variables involved. Instead of knowledge of the distributions, we have training data that we claim represents independent, identically distributed samples generated by the same experiment. Using this data, machine learning algorithms design detection and estimation algorithms. The "learning" occurs by using the provided data samples to design detection or estimation algorithms.

In machine learning, we speak of two types of learning: Supervised learning and unsupervised learning. In supervised learning, the data provided includes samples of both the observations and the hypothesis labels. The machine learning algorithm can use the labels to learn an appropriate decision or estimation rule. In unsupervised learning, the data consists only of samples of the observations, without labels. The machine learning algorithm will have to uncover important features of the data and appropriate classes on its own. In this chapter, we will focus primarily on supervised learning, although we will discuss techniques such as clustering and principal component analysis that are used effectively in unsupervised learning.

The problem of classification in machine learning is equivalent to the problem of hypothesis testing in chapter 6. In these problems, the observations X can be discrete, continuous or a mixture of discrete and continuous variables, and the labels Y are a discrete hypothesis label. Similarly, the problem of regression is estimation 7 in a machine learning context. Here the observations are again random vectors of different types, and the label variable Y is a real number or vector that we are trying to estimate.

10.2 Learning probabilities from data

Suppose we were interested in "learning" a classifier for binary hypothesis testing problem. The design theory of Chapter 6 suggests that what we need to know are the likelihoods for the observation data $\underline{X}$ under both hypotheses. Assuming the observations are continuous random vectors, we need knowledge of $f_{\underline{X}|H_0}(\underline{x})$ and $f_{\underline{X}|H_1}(\underline{x})$. Knowing these, the decision rule consists of forming the likelihood ratio $\mathcal{L}(\underline{x})$ and comparing this ratio to a threshold, a design parameter that we choose.

In machine learning, the likelihood functions are not known. Instead, we are given training data for each hypotheses, of the form

$$(\underline{x}_1, H_0), \ldots, (\underline{x}_{n_0}, H_0); \quad (\underline{x}_{n_0+1}, H_1), \ldots, (\underline{x}_n, H_1),$$

where each data observation $\underline{x}_i$ is associated with a label that indicates whether this data was sampled from the likelihood $f_{\underline{X}|H_0}(\underline{x})$ or $f_{\underline{X}|H_1}(\underline{x})$. The data $\underline{X}$ are the observed features, and are assumed to be d-dimensional vectors in $\Re^d$. Our goal is to learn approximations to the likelihoods $f_{\underline{X}|H_0}(\underline{x})$ and $f_{\underline{X}|H_1}(\underline{x})$.

There are two types of approaches for estimating densities from data: parametric and non-parametric, which we describe below.

10.2.1 Parametric models

The parametric approach for machine learning assumes that we know the family that the densities $f_{\underline{X}|H_0}(\underline{x})$ or $f_{\underline{X}|H_1}(\underline{x})$ belong to. For instance, we assume that the densities are jointly Gaussian, with unknown means $\underline{\mu}_0, \underline{\mu}_1$ and unknown covariances Σ_0, Σ_1 respectively. The idea behind parametric density estimation is to use the training data to estimate the unknown parameters of the likelihood functions. The parameter estimation problems are straightforward, and use the maximum likelihood estimation algorithms described in Chapter 7.

Specifically, assume we are given n independent observations of a random vector $\underline{X} \in \Re^d$, denoted as $\underline{x}_1, \ldots, \underline{x}_n$. We assume that the joint PDF of $\underline{X}$ is one of a family of PDFs, parametrized by unknown parameters $\underline{\theta}$, so that $f_{\underline{X}}(\underline{x}) = g(\underline{x}, \underline{\theta})$. For instance, the function $g(\cdot)$ can be the joint density of a d-dimensional Gaussian random vector, with unknown mean and covariance matrix, which form the vector of unknown parameters $\underline{\theta}$. Based on the data, we estimate these parameters using an estimation algorithm such as maximum-likelihood, so that

$$\widehat{\underline{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}} \prod_{k=1}^{n} g(\underline{x}_k, \underline{\theta}).$$

The resulting approximate PDF of $\underline{X}$ is then $f_{\underline{X}}(\underline{x}) = g(\underline{x}, \widehat{\underline{\theta}})$. We illustrate this with examples below.

Example 10.1

We have two hypotheses we are trying to detect, based on a scalar Gaussian observation. Specifically, under hypothesis H_0, the observation X is a Gaussian random variable with mean 0, variance 4. Under H_1, it is a Gaussian random variable with mean 5, variance 4. Using the theory of Chapter 6, the maximum likelihood decision rule for this problem is

$$D_{ML}(x) = \begin{cases} H_1 & x \geq 2.5 \\ H_0 & x < 2.5 \end{cases}$$

and the probability of error in this detector is $Q(1.25) = 0.106$.

Now, assume we did not know the parameters of the Gaussian distribution, but instead are given 1000 samples (x_i, y_i), where $y_i \in \{H_0, H_1\}$, from each of the two distributions, $f_{X|H_0}(x), f_{X|H_1}(x)$. We assume the first 1000 samples come from H_0, and the second from H_1. We refer to the values x_i as the data, and the values y_i as the labels.

The vector of unknown parameters in the densities is $\underline{\theta} = [m_0, m_1, \sigma^2]$, corresponding to the unknown means and the common variance of the conditional densities $f_{X|H_0}(x), f_{X|H_1}(x)$. The parametric form of the likelihood function for each hypothesis is:

$$f_{X|H_k}(x) \equiv g_k(x, \underline{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m_k)^2}{2\sigma^2}}, \quad k = 0, 1$$

Using this and the fact that each of the data samples is independent, we can write the estimation problem as

$$\widehat{\underline{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}} \prod_{j=1}^{1000} g_0(x_j, \underline{\theta}) \prod_{j=1001}^{2000} g_1(x_j, \underline{\theta})$$

We can maximize the logarithm of the right-hand side as a simplification, to obtain

$$\widehat{\underline{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}} \sum_{j=1}^{1000} \left(-\frac{\ln(2\pi\sigma^2)}{2} - \frac{(x_j - m_0)^2}{2\sigma^2} \right) + \sum_{j=1001}^{2000} \left(-\frac{\ln(2\pi\sigma^2)}{2} - \frac{(x_j - m_1)^2}{2\sigma^2} \right)$$

Differentiating with respect to m_0, m_1 and σ^2 and setting the results to zero yields the following values for the maximum likelihood estimates:

$$\widehat{m}_0 = \frac{1}{1000}\sum_{i=1}^{000} x_i; \quad \widehat{m}_1 = \frac{1}{1000}\sum_{i=1001}^{2000} x_i$$

Since we assume the variances of the two densities are the same, we estimate the joint variance as

$$\widehat{\sigma}^2 = \frac{1}{2000}\big(\sum_{i=j}^{1000}(x_j - \widehat{m}_0)^2 + \sum_{j=1001}^{2000}(x_i - \widehat{m}_1)^2\big).$$

With these estimates, we can approximate the conditional densities $f_{X|H_0}(x), f_{X|H_1}(x)$, and implement an optimal maximum likelihood decision rule, which results in:

$$D_{ML}(x) = \begin{cases} H_1 & x \geq \frac{\widehat{m}_1 + \widehat{m}_0}{2} \\ H_0 & \text{otherwise.} \end{cases}$$

and we predict our probability of error to be $Q(\frac{\widehat{m}_1 - \widehat{m}_0}{2\sqrt{\widehat{\sigma^2}}})$.

We implemented the above problem in a Python script described below:

```
import numpy as np
X0 = 2*np.random.randn(1000,)  #generate H0 train data
X1 = 2*np.random.randn(1000,1).ravel()+5 # H1 train data
test0 = 2*np.random.randn(1000,1).ravel()  #H0 test data
test1 = 2*np.random.randn(1000,1).ravel()+5  #H1 test data
m0 = sum(X0)/1000.
m1 = sum(X1)/1000.0
var = (sum((X0 - m0)**2)+sum((X1 - m1)**2)) /2000.0
T = 0.5*(m1 + m0)
Pe = (sum(test0 > T) + sum(test1 < T))/2000
print(m0,m1,var)
print(T,Pe,T/np.sqrt(var) )
```

and obtained the following estimates:

$$\widehat{m}_0 = -0.018; \quad \widehat{m}_1 = 4.906; \quad \widehat{\sigma}^2 = 4.135.$$

Based on these estimates, the threshold in the decision rule is 2.462, and the predicted performance is $Q(1.21) \approx 0.101$.

We generated 2000 additional samples as test data, and evaluated the empirical performance of our detector on the test data, as the script indicates. The empirical probability of error was $P_e = 0.105$, which is a bit higher than predicted, because the estimated models from the training data are different than the actual models used to generate the data.

In this example, we generated lots of data to estimate 3 unknown parameters: 2000 independent samples. We will see that this complicates the problem when we deal with larger numbers of unknown parameters.

10.2.2 Nonparametric Density Estimation

Can we estimate the likelihoods $f_{\underline{X}|H_0}(\underline{x}), f_{\underline{X}|H_1}(\underline{x})$ without assuming a parametric form? There are numerous techniques that attempt to do this. The problem of non-parametric density estimation can be summarized as follows: given N independent samples $\underline{x}_k, k = 1, \ldots, N$ of an n-dimensional random vector $\underline{X}$ with PDF $f_{\underline{X}}(\underline{x})$, generate an estimate of the probability density function for all values $\underline{x} \in \Re^n$.

One approach is to construct an empirical probability mass function based on the observed samples, as follows:

$$\widehat{P}_{\underline{X}}(\underline{x}) = \begin{cases} 0 & \underline{x} \neq \underline{x}_k \text{ for some } k \in \{1, \ldots, N\} \\ \frac{1}{N} & \underline{x} = \underline{x}_k \text{ for some } k \in \{1, \ldots, N\}. \end{cases}$$

Unfortunately, this is a discrete probability mass function, and not a density. Furthermore, it assigns zero probability to obtaining any data values that are not in the training set $\underline{x}_k, k = 1, \ldots, N$.

The approach we propose for generating a better density estimate is **kernel density estimation** (KDE), which is a version of the Parzen's window estimator. KDE is a nonparametric density estimator, requiring no assumption that the underlying density function is from a parametric family. KDE will learn the shape of the density from the data automatically. The idea behind KDE is to blur the empirical probability mass function using a smooth kernel so that the probability mass function is extrapolated to a density that has range on values that were not included in a training set. Kernel density estimators smooth out the contribution of each observed data point over a local neighborhood of that data point.

Define a kernel $K(\underline{x})$ to be a smooth non-negative function with a peak at $\underline{x} = \underline{0}$, such that $\int_{\underline{x}} K(\underline{x})\, d\underline{x} = 1$. We can interpret the kernel as a probability density function, as it satisfies the non-negativity and normalization properties of probability densities. For most applications, we use standard Gaussian kernels, of the form

$$K_h(\underline{x}) = \frac{1}{(\sqrt{2\pi}h)^n} e^{-\frac{\underline{x}^T \underline{x}}{2h^2}},$$

which is the product of independent Gaussian densities with zero-mean and standard deviation h in each of the n dimensions of $\underline{X}$. The parameter h is known as the kernel width.

With this kernel, the KDE approximation is

$$\widehat{f}_{\underline{X}}(\underline{x}) = \frac{1}{N} \sum_{k=1}^{N} K_h(\underline{x} - \underline{x}_k)$$

Note that, for an arbitrary point $\underline{x} \in \Re^n$, the density will be determined primarily by the data points $\underline{x}_k$ that are within a distance of $3h$ of $\underline{x}$. The quality of a kernel estimate depends strongly on the value of its bandwidth h. It's important to choose the most appropriate bandwidth as a value that is too small or too large is not useful. Small values of h lead to very spiky estimates (not much smoothing) while larger h values lead to oversmoothing.

Example 10.2
Let's approximate the density of a Gaussian random variable X with mean 0, variance 1 using KDE. We select 100 independent, randomly generated samples of X, and we show the resulting KDE densities using different values for the width h in the figure below. We see that for values $h = 0.1$, the approximate density is too rough. For $h = 0.4, 0.5$, the density approximation is accurate. For $h = 0.6, 0.7$ we see the density is oversmoothed.

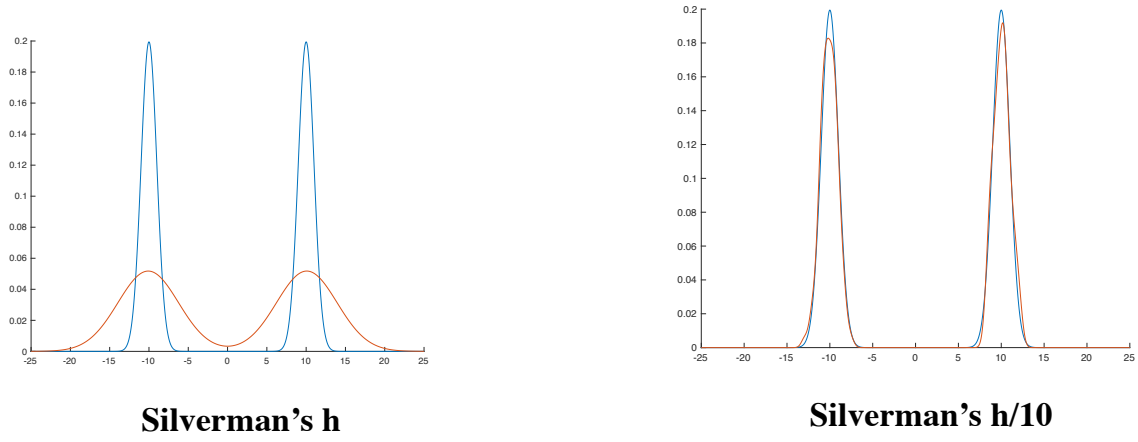
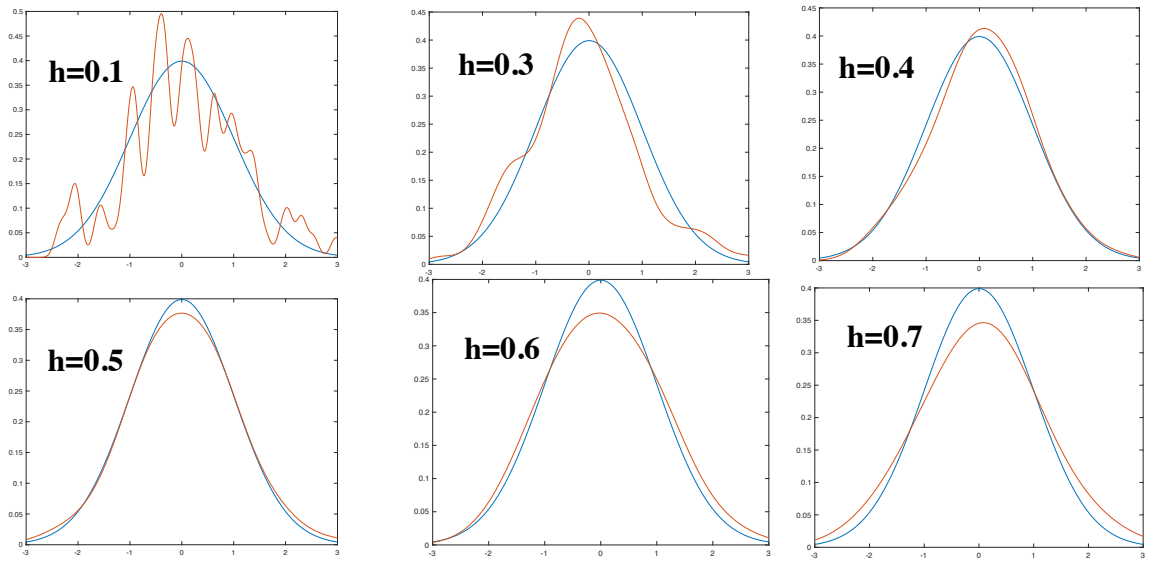


Figure 10.1: A multi-modal density where the standard deviation is not representative of the curvature. Red indicates the KDE approximation to the true density in blue.



There are many approaches to selecting the right value of h. A rule of thumb, known as Silverman's rule, is based on approximating Gaussian densities, and is given by $h = \left(\frac{4\widehat{\sigma}^5}{3N}\right)^{\frac{1}{5}} \approx 1.06\widehat{\sigma}N^{-1/5}$. If we evaluate this for the example above with $N = 100$ and $\sigma = 1$, we get $h \approx 0.42$. However, this rule can often be wrong if the density has multiple peaks. For instance, consider the density in Figure 10.1 below. Silverman's approximation yields an h that oversmooths the density, as the estimated standard deviation is more reflective of the spread in the two peaks than the curvature of the density. A different rule that reduces h by an order of magnitude yields an excellent approximation. In general, one selects h using a search that involves some form of cross-validation, where part of the data is used to estimate the density and another part is used to validate that the estimated density is accurate.

KDE can be integrated into classification problems with supervised learning. Given samples of $\underline{X}$ generated independently by $f_{\underline{X}|H_0}(\underline{x})$, we can construct a KDE estimate $\widehat{f}_{\underline{X}|H_0}(\underline{x})$ of this likelihood function. We can similarly construct a KDE estimate $\widehat{f}_{\underline{X}|H_1}(\underline{x})$. With these two estimates, the maximum likelihood decision rule is

$$\widehat{D}_{ML}(\underline{x}) = \begin{cases} H_1 & \widehat{f}_{\underline{X}|H_1}(\underline{x}) \geq \widehat{f}_{\underline{X}|H_0}(\underline{x}) \\ 0 & \text{otherwise.} \end{cases}$$

The main limitations of KDE estimators are two-fold. First, for high-dimensional data, it is hard to interpolate accurately. The number of data points required grows exponentially with the number of dimensions. Second, the complexity of the classifier is large: one has to compute the kernel distance to all the training points, which can be slow as the number of training points increases. We will show the performance we can achieve with KDE estimation on a real classification problem, described in the next section.

10.3 The IRIS data set

As a motivating application, we use a particular data set that was used by the father of modern statistics, Ronald Fisher. In addition to his work on statistics, Fisher was a mathematical biologist and applied statistics to maximum likelihood classification problems. This particular data set was reported in a paper in 1936, "The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis." The data was collected to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected from the same pasture, picked on the same day and measured with the same apparatus. The IRIS data set is widely studied, and copies of it are easily downloaded on the internet.

The data set of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. This discriminant is still referred to as Fisher's linear discriminant. The three types of flowers are shown in Figure 10.2.

Figure 10.2: The three types of iris flowers in the IRIS data set.

The four features in this data correspond to the sepal length in centimeters, the sepal width in centimeters, the petal length in centimeters and the petal width in centimeters. The petal and sepal leaves of an iris flower are shown in Figure 10.3.

To illustrate the properties of this data set, we performed a statistical analysis using the Python Seaborn package, that shows a kernel density estimate for the marginal densities for each of the four features as well as plots showing pairs of features, color coded to each of the iris varieties. This output is shown in Figure 10.4. This exploratory analysis shows that it will be relatively easy to separate the Setosa variety from the other two varieties, because of its smaller petals, but it will be harder to separate the other two varieties.

Example 10.3

We use the IRIS data set to evaluate the performance of parametric max-likelihood classifiers and explore some of their limitations. The observations in the IRIS data set are 4-dimensional vectors. There are three hypotheses, which we denote as H_0 (Setosa), H_1 (Versicolor) and H_2 (Virginica). We only have 50 observation samples per hypotheses. We assume that the densities $f_{\underline{X}|H_0}(\underline{x}), f_{\underline{X}|H_1}(\underline{x}), f_{\underline{X}|H_2}(\underline{x})$ are joint Gaussian densities, with means $\underline{\mu}_0, \underline{\mu}_1, \underline{\mu}_2$ respectively and a common covariance matrix $\mathbf{\Sigma}$. Note that the number of parameters required to estimate the densities is three four-dimensional means and a four-by-four symmetric covariance matrix, leading to $4 \times 3 + 10 = 22$ unknown parameters to be estimated,

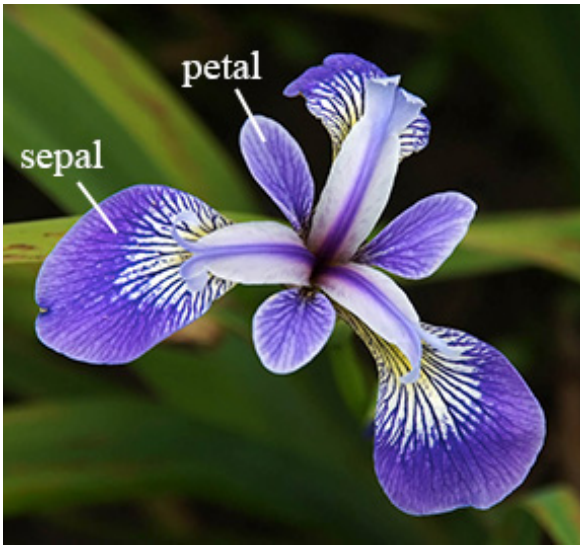


Figure 10.3: Names of leaves of iris flowers in the IRIS data set.

with 150 four-dimensional measurements. This is limited data for a problem with that many unknown parameters, and we can expect difficulty in estimating the densities accurately.

We assume the data provided is in the form $\{(\underline{x}_k, y_k)\}, k = 1, \ldots, 150$, where $y_k \in \{H_0, H_1, H_2\}$. The estimate of each of the class means is obtained using maximum likelihood estimation as

$$\widehat{\underline{\mu}}_j = \frac{1}{50} \sum_{k=1}^{150} \underline{x}_k I_{y_k = H_j}, \ j = 0, 1, 2,$$

where I_z is the indicator function which is 1 when z is true, and 0 otherwise. The covariance estimate is generated by

$$\widehat{\Sigma} = \frac{1}{150} \sum_{k=1}^{150} \left(\sum_{j=0}^{2} I_{y_k = H_j} (\underline{x}_k - \widehat{\underline{\mu}}_j)(\underline{x}_k - \widehat{\underline{\mu}}_j)^T \right).$$

We implemented the above equations and obtained the following estimates:

$$\widehat{\underline{\mu}}_0 = \begin{bmatrix} 5.006 \\ 3.428 \\ 1.462 \\ 0.246 \end{bmatrix} ; \ \widehat{\underline{\mu}}_1 = \begin{bmatrix} 5.936 \\ 2.77 \\ 4.26 \\ 1.326 \end{bmatrix} ; \ \widehat{\underline{\mu}}_2 = \begin{bmatrix} 6.588 \\ 2.974 \\ 5.552 \\ 2.026 \end{bmatrix} ; \ \widehat{\Sigma} = \begin{bmatrix} 0.261 & 0.091 & 0.165 & 0.038 \\ 0.091 & 0.114 & 0.055 & 0.032 \\ 0.165 & 0.05 & 0.183 & 0.042 \\ 0.038 & 0.032 & 0.042 & 0.041 \end{bmatrix}.$$

With these density estimates, we implemented a maximum likelihood classifier. On the training data, the performance of the classifier had a probability of error of 0.02. We subsequently broke the training data so that 70% of the data was used for training, and 30% was used for testing. The probability of error went up to 0.044, which is still very good given the limited amount of training data.

Example 10.4
We return to the IRIS data set, but using the KDE nonparametric estimators discussed previously to approximate the likelihood functions for each of the three species of iris flowers. For this problem, we can always get zero probability of error on the training data by selecting a small value of h, since this will put all the weight on the actual data point being tested, which is part of the training set. Thus, we divided the data into 70% training, 30% testing and evaluated the performance of the KDE maximum likelihood classifier described above. On the test data, our probability of error was 0.044, which is the same as the parametric estimator discussed in Example 10.3.

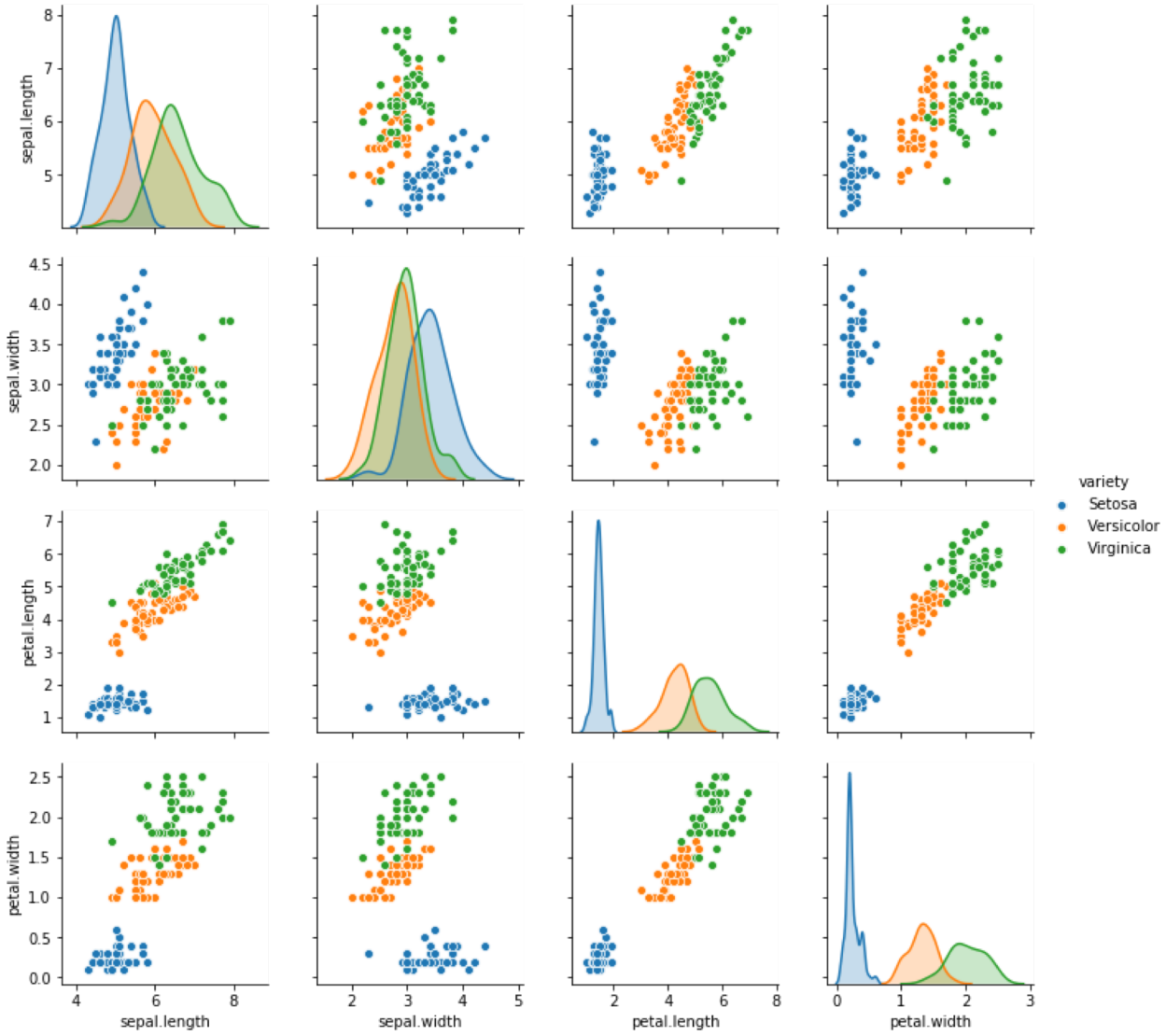


Figure 10.4: Seaborn pairs analysis of IRIS data.

10.4 Binary Classification

In this section, we focus on describing various machine learning approaches to the problem of selecting one of two hypotheses. This problem is a generalization of the binary hypothesis testing problem, to the case where we don't have probability models for the observations conditioned on the hypotheses. Instead, we are provided sample observation values that are obtained under each of the two hypotheses. Our goal is to design a decision rule that maps new observations into a selection of which hypothesis is best.

The binary classification problem has two hypotheses, H_0 and H_1. To simplify notation, we associate the decision value -1 with H_0 and $+1$ with H_1. We observe a vector of features $\underline{X}$ with values in $\Re^d$, and we want to design a decision rule $D(\underline{x})$ that maps the observed vector $\underline{X} = \underline{x}$ into one of the two decisions $\{-1, +1\}$.

Assume we are given training data of the form $(\underline{X}_1, Y_1), \ldots, (\underline{X}_n, Y_n)\}$, where $\underline{X}_k \in \Re^d$ is a sample observation value, and $Y_k \in \{-1, +1\}$ is the label that indicates which of the two hypotheses generated

that observation. The decision rule $D(\underline{x})$ assigns a label of -1 or 1 to each possible observed vector $\underline{x}$. We typically measure the performance of a decision rule by the error rate, which is the probability that a point is misclassified. Since we don't have probability models readily available, we compute this as the fraction of sample values that are misclassified, as we will show later.

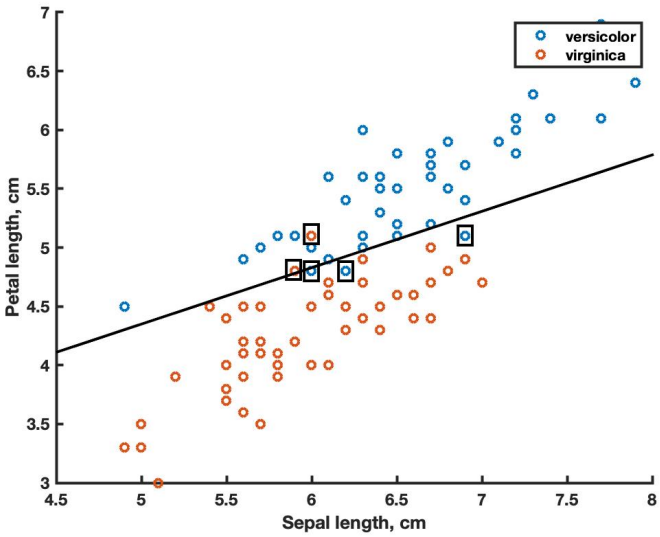


Figure 10.5: Sepal length versus petal length for two types of Iris flowers

Figure 10.5 plots a pair of features, sepal length and petal length, for two types of iris flowers, Versicolor and Virginica. The figure also shows a potential decision rule, indicated by a straight line. For data points above the line, the decision rule declares that flower type is Versicolor, whereas for data points below the line, the decision rule declares that the flower type is Virginica. The figure highlights the five data points where the decision rule makes errors, by including a black rectangle among the points.

In this section, we describe various approaches for designing binary classifiers using labeled training data and supervised training.

10.4.1 Clustering Classifiers

Assume we are given training data of the form $(\underline{X}_1, Y_1), \ldots, (\underline{X}_Y, Y_n)\}$, where $Y_k \in \{H_0, H_1\}$ is a categorical label. Given this labeled data, it is straightforward to compute the average observation under each hypothesis as:

$$\widehat{\underline{\mu}}_0 = \frac{\sum_{k=1}^{n} \underline{X}_k I_{Y_k=H_0}}{\sum_{k=1}^{n} I_{Y_k=H_0}}; \quad \widehat{\underline{\mu}}_1 = \frac{\sum_{k=1}^{n} \underline{X}_k I_{Y_k=H_1}}{\sum_{k=1}^{n} I_{Y_k=H_1}}$$

The clustering classifier assigns to an input value $\underline{x}$ the decision that has its average closest to the input value. That is,

$$D(\underline{x}) = \begin{cases} H_0, & \| \underline{X} - \widehat{\underline{\mu}}_0 \| < \| \underline{X} - \widehat{\underline{\mu}}_1 \| \\ H_1, & \| \underline{X} - \widehat{\underline{\mu}}_0 \| \geq \| \underline{X} - \widehat{\underline{\mu}}_1 \| \end{cases}$$

Note that the clustering classifier is easy to extend to K hypotheses, as long as training data is provided for each. Furthermore, the classifier can be extended to unsupervised classification, by analyzing the data using a clustering algorithm and discovering the clusters in the data.

The clustering classifier is computed in Figure 10.6 for the two length features for the IRIS data set, for the classes Versicolor and Virginica. The two class centers are shown as filled diamonds in each of the

classes. Looking at the results, we see that 15 out of the 100 data samples are classified incorrectly, for an error rate of 15% when evaluated using the training data.

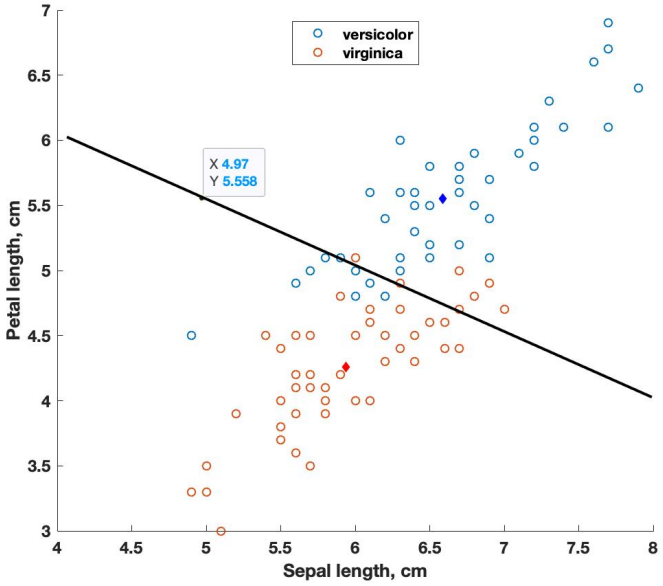


Figure 10.6: Illustration of clustering classifier

10.4.2 Nearest Neighbor and K-Nearest Neighbor Classifiers

Nearest neighbor classifiers are similar to clustering classifiers, but involve more computation. In clustering classifiers, the trained classifier only needs to remember the average values estimated for each class. In contrast, the classifier for nearest neighbor classifiers needs to remember all the training data provided.

To classify an input $\underline{x}$, the nearest neighbor classifier finds the training data point $\underline{X}_k$ that is closest to $\underline{x}$, and then assigns the label of that data point to the data $\underline{X}$. That is, one finds j^* such that

$$\| \underline{x} - \underline{X}_{j^*} \| \leq \| \underline{x} - X_j \| \text{ for all } j = 1, \ldots, n$$

The nearest neighbor classifier is then $D(\underline{x}) = Y_{j^*}$.

A simple extension of a nearest neighbor classifiers is to find the K nearest neighbors of the input data $\underline{x}$, denoted as $\underline{X}_{j_1}, \ldots, \underline{X}_{j_K}$. The label assigned to $\underline{x}$ would be determined by the labels $Y_{j_1}, \ldots, Y_{j_K}$, usually in majority voting. Finding the K nearest neighbors efficiently usually requires advanced data structures, particularly when the dimension d of the data is large.

10.4.3 Discriminant Analysis

Discriminant analysis classifiers are based on Gaussian parametric approximations of the likelihood functions $f_{\underline{X}|H_0}(\underline{x}), f_{\underline{X}|H_1}(\underline{x})$. The most common type of discriminant analysis is **Linear Discriminant Analysis (LDA)**, which assumes that the likelihood functions $f_{\underline{X}|H_0}(\underline{x}), f_{\underline{X}|H_1}(\underline{x})$ are jointly Gaussian with different means $\underline{\mu}_0, \underline{\mu}_1$ but with the same covariance matrix Σ. The decision rule is the maximum likelihood decision rule with the approximate densities.

Given training data of the form $(\underline{X}_1, Y_1), \ldots, (\underline{X}_Y, Y_n)\}$, where $Y_k \in \{H_0, H_1\}$ is a categorical label, compute the number of samples of type H_0, H_1 as $n_0 = sum_{k=1}^n \underline{X}_k I_{Y_k=H_0}, n_1 = sum_{k=1}^n \underline{X}_k I_{Y_k=H_1}$. Given this labeled data, it is straightforward to estimate the means of each of the two densities, in the same manner as we estimated the centers of the clustering classifiers earlier:

$$\widehat{\underline{\mu}}_0 = \frac{\sum_{k=1}^n \underline{X}_k I_{Y_k=H_0}}{n_0}; \quad \widehat{\underline{\mu}}_1 = \frac{\sum_{k=1}^n \underline{X}_k I_{Y_k=H_1}}{n_1}.$$

Estimating the common covariance matrix Σ is more involved. A common estimator is

$$\widehat{\Sigma} = \frac{1}{n-2} \sum_{k=1}^n \Big((\underline{X}_k - \widehat{\underline{\mu}}_0)^2 I_{Y_k=H_0} + (\underline{X}_k - \widehat{\underline{\mu}}_1)^2 I_{Y_k=H_1} \Big)$$

A different way of computing the same estimate is to first estimate $\widehat{\Sigma}_0, \widehat{\Sigma}_1$ as the covariances based on data labeled as H_0 or H_1 respectively, as:

$$\widehat{\Sigma}_0 = \frac{1}{n_0 - 1} \sum_{k=1}^n (\underline{X}_k - \widehat{\underline{\mu}}_0)^2 I_{Y_k=H_0}$$

$$\widehat{\Sigma}_1 = \frac{1}{n_1 - 1} \sum_{k=1}^n (\underline{X}_k - \widehat{\underline{\mu}}_1)^2 I_{Y_k=H_1}$$

We can subsequently combine these into a single estimate, as

$$\widehat{\Sigma} = \frac{1}{n-2} \big((n_0 - 1)\widehat{\Sigma}_0 + (n_1 - 1)\widehat{\Sigma}_1 \big).$$

With these estimates, we approximate the likelihoods of $\underline{X}$ given H_0, H_1 as

$$\widehat{f}_{\underline{X}|H_0}(\underline{x}) = \frac{1}{\sqrt{\det(2\pi\widehat{\Sigma})}} e^{-\frac{1}{2}(\underline{x} - \widehat{\underline{\mu}}_0)^T \widehat{\Sigma}^{-1} (\underline{x} - \widehat{\underline{\mu}}_0)}$$

$$\widehat{f}_{\underline{X}|H_1}(\underline{x}) = \frac{1}{\sqrt{\det(2\pi\widehat{\Sigma})}} e^{-\frac{1}{2}(\underline{x} - \widehat{\underline{\mu}}_1)^T \widehat{\Sigma}^{-1} (\underline{x} - \widehat{\underline{\mu}}_1)}$$

To derive the max-likelihood classifier using these likelihood estimates, we compare the log-likelihood ratio to 0, as

$$\begin{aligned} \ln\big(\mathcal{L}(\underline{x})\big) &= \frac{1}{2}(\underline{x} - \widehat{\underline{\mu}}_0)^T \widehat{\Sigma}^{-1} (\underline{x} - \widehat{\underline{\mu}}_0) - \frac{1}{2}(\underline{x} - \widehat{\underline{\mu}}_1)^T \widehat{\Sigma}^{-1} (\underline{x} - \widehat{\underline{\mu}}_1) \\ &= (\widehat{\underline{\mu}}_1 - \widehat{\underline{\mu}}_0)^T \widehat{\Sigma}^{-1} \underline{x} - \frac{1}{2}(\widehat{\underline{\mu}}_1^T \widehat{\Sigma}^{-1} \widehat{\underline{\mu}}_1 - \widehat{\underline{\mu}}_0^T \widehat{\Sigma}^{-1} \widehat{\underline{\mu}}_0) \underset{H_0}{\overset{H_1}{\gtrless}} 0. \end{aligned}$$

This decision rule is a linear decision rule of the form

$$D(\underline{x}) = \begin{cases} H_1 & \text{if } \underline{a}^T \underline{x} \geq b, \\ H_0 & \text{elsewhere,} \end{cases}$$

where $\underline{a} = \widehat{\Sigma}^{-1}(\widehat{\underline{\mu}}_1 - \widehat{\underline{\mu}}_0)$, and $b = \frac{1}{2}(\widehat{\underline{\mu}}_1^T \widehat{\Sigma}^{-1} \widehat{\underline{\mu}}_1 - \widehat{\underline{\mu}}_0^T \widehat{\Sigma}^{-1} \widehat{\underline{\mu}}_0)$. The LDA decision rule is similar to the closest average decision rule, except that the distances between data and the centers are modified by the estimate of the inverse covariance. If $\widehat{\Sigma} = \mathbf{I}_d$, the d-dimensional identity matrix, then the LDA decision rule is equivalent to the closest average decision rule.

We show the LDA decision on the same two-dimensional data for the IRIS data set as before. Compared with the clustering classifier, the decision rule has shifted the orientation of the separation line based on the estimated covariance, and has reduced the number of errors to 5, leading to a 5% error on the training data.

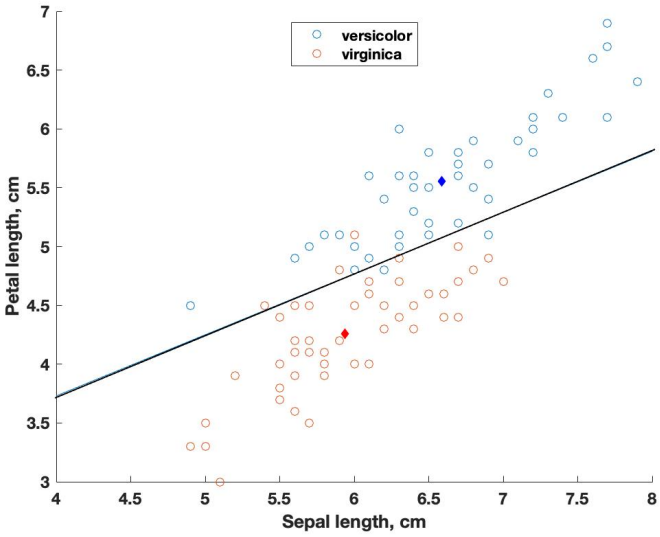


Figure 10.7: LDA decision rule for selecting between Versicolor and Virginica Iris

Instead of Linear Discriminant Analysis, we can do **Quadratic Discriminant Analysis (QDA)**, which is based on parametric modeling of the likelihood functions $f_{\underline{X}|H_0}(\underline{x}), f_{\underline{X}|H_1}(\underline{x})$ are jointly Gaussian with different means $\underline{\mu}_0, \underline{\mu}_1$ and different covariances Σ_1 and Σ_2. Given the same training data, the estimates can be obtained as

$$\widehat{\underline{\mu}}_0 = \frac{\sum_{k=1}^n \underline{X}_k I_{Y_k=H_0}}{n_0}; \quad \widehat{\underline{\mu}}_1 = \frac{\sum_{k=1}^n \underline{X}_k I_{Y_k=H_1}}{n_1}.$$

$$\widehat{\Sigma}_0 = \frac{1}{n_0-1}\sum_{k=1}^n (\underline{X}_k - \widehat{\underline{\mu}}_0)^2 I_{Y_k=H_0}; \quad \widehat{\Sigma}_1 = \frac{1}{n_1-1}\sum_{k=1}^n (\underline{X}_k - \widehat{\underline{\mu}}_1)^2 I_{Y_k=H_1}.$$

Computing the log-likelihood ratio results in

$$\begin{aligned}\ln\left(\mathcal{L}(\underline{x})\right) &= \frac{1}{2}(\underline{x}-\widehat{\underline{\mu}}_0)^T\widehat{\Sigma}_0^{-1}(\underline{x}-\widehat{\underline{\mu}}_0) - \frac{1}{2}(\underline{x}-\widehat{\underline{\mu}}_1)^T\widehat{\Sigma}_1^{-1}(\underline{x}-\widehat{\underline{\mu}}_1) + \frac{\ln(\det[\widehat{\Sigma}_0])}{2} - \frac{\ln(\det[\widehat{\Sigma}_1])}{2} \\ &= \frac{1}{2}\underline{x}^T(\widehat{\Sigma}_0^{-1} - \widehat{\Sigma}_1^{-1})\underline{x} + (\widehat{\underline{\mu}}_1^T\widehat{\Sigma}_1^{-1} - \widehat{\underline{\mu}}_0^T\widehat{\Sigma}_0^{-1})\underline{x} + \frac{\widehat{\underline{\mu}}_0^T\widehat{\Sigma}_0^{-1}\widehat{\underline{\mu}}_0 - \widehat{\underline{\mu}}_1^T\widehat{\Sigma}_1^{-1}\widehat{\underline{\mu}}_1}{2} + \frac{1}{2}\ln\left(\frac{\det[\widehat{\Sigma}_0]}{\det[\widehat{\Sigma}_1]}\right) \underset{H_0}{\overset{H_1}{\gtrless}} 0\end{aligned}$$

Figure 10.8 shows the boundary of the quadratic decision rule obtained by QDA for the two feature IRIS data set used in the previous examples. Although the empirical error rate on the training data is the same as the LDA empirical error rate 5%, the curvature of the region is likely to improve the error rate on test data.

10.4.4 Perceptron Classifier

Many modern approaches to binary classification use complex parametric decision functions, and use large-scale optimization techniques to select the parameters of such decision functions. Examples of such decision functions are neural networks, which use interconnected layers of linear and nonlinear elements with weights to map an observed input $\underline{x}$ to a decision $D(\underline{x})$. The training of such networks is beyond the scope of this course. However, we will describe a simple nonlinear neural network model for classification, proposed by Rosenblatt in 1958. It is known as Rosenblatt's perceptron network, and is illustrated in Figure **??**. Rosenblatt's perceptron was an attempt to model the processing of a neuron, based on earlier work (1943)

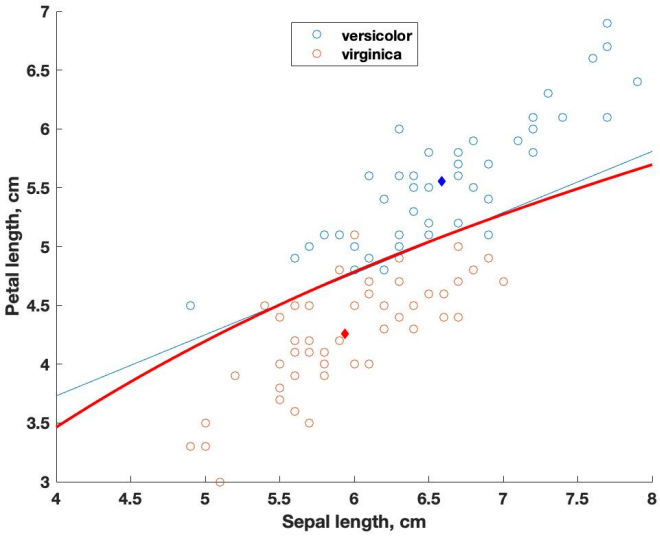


Figure 10.8: LDA decision rule for selecting between Versicolor and Virginica Iris

by McCulloch and Pitts. The figure shows how a vector of inputs, combined with weights is added and passed to a nonlinear function that examines the sign and selects the hypothesis.

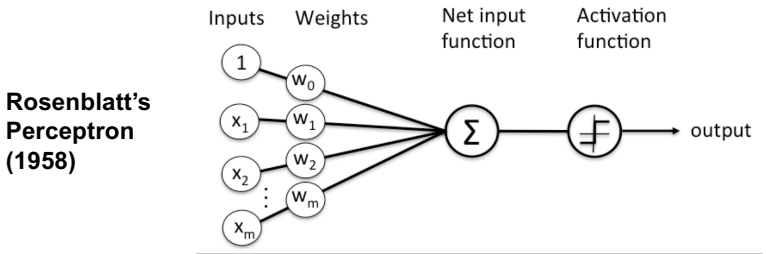


Figure 10.9: Rosenblatt's Perceptron Model.

Training a perceptron classifier can be accomplished without the use of iterative optimization techniques. The training is based on principles of regression: given data $(\underline{X}_k, Y_k), k = 1, \ldots, n$ with labels $Y_k \in \{H_0, H_1\}$, change the labels to numbers, where $Y_k = H_0$ is changed to $Y_k = -1$, and $Y_k = H_1$ is changed to $Y_k = 1$. We now have numerical data $\{(\underline{X}_1, Y_1), \ldots, (\underline{X}_n, Y_n)\}$.

Let the vector $\underline{a}$ correspond to the weights $w_1, \ldots, w_m$ that multiply the data $\underline{x}$, and the scalar b corresponds to the weight w_0 that is added as a bias to the nonlinear function in the perceptron classifier. We want to find the best linear predictor of Y from $\underline{X}$ to minimize the least-squares error:

$$\min_{\underline{a},b} \frac{1}{n} \sum_{k=1}^{n} (Y_k - \underline{a}^T \underline{X}_k - b)^2.$$

This optimization is known as linear regression, and is very similar to the problem of linear least-squares estimation, except that instead of PDFs to compute means and variances, we have sample data. Linear regression with minimum mean square error objectives was developed by Gauss in the early 19th century to estimate planetary orbits.

We can solve the linear regression problem using the LLSE solution of Chapter 7. Specifically, we compute

approximately

$$\mathbb{E}[\underline{X}] = \frac{1}{n}\sum_{k=1}^{n} \underline{X}_k; \quad \mathbb{E}[Y] = \frac{1}{n}\sum_{k=1}^{n} Y_k.$$

$$\mathsf{Var}[\underline{X}] = \Sigma_{\underline{X}} = \frac{1}{n-1}\sum_{k=1}^{n}(\underline{X}_k - \mathbb{E}[\underline{X}])(\underline{X}_k - \mathbb{E}[\underline{X}])^T; \mathsf{Cov}[Y, \underline{X}] = \Sigma_{Y,\underline{X}} = \frac{1}{n-1}\sum_{k=1}^{n}(\underline{X}_k - \mathbb{E}[\underline{X}])^T(Y_k - \mathbb{E}[Y]).$$

The LLSE estimator is

$$\widehat{y}_{LLSE}(\underline{x}) = \mathbb{E}[Y] + \Sigma_{Y,\underline{X}}\Sigma_{\underline{X}}^{-1}(\underline{x} - \mathbb{E}[X]),$$

which makes the optimal regression vector $\underline{a}^T = \Sigma_{Y,\underline{X}}\Sigma_{\underline{X}}^{-1}$ and the constat $b = \mathbb{E}[Y] - \underline{a}^T\mathbb{E}[x]$. This solution yields the optimal weights to use in the perceptron classifier.

As an alternative, we can convert the weight optimization problem by optimizing for the weights $\underline{w}$ directly, where $\underline{w}$ is a $(d+1)$-dimensional vector. We do this by forming the data matrix

We often convert this to a homogeneous formulation by grouping the constant d into the estimation vector $\underline{c}$, defining the vector $\underline{b}^T = \begin{bmatrix}\underline{c}^T & d\end{bmatrix}$, so that the estimate is of the form

$$\widehat{y}(\underline{x}) = \begin{bmatrix}\underline{c}^T & d\end{bmatrix}\begin{bmatrix}\underline{X} \\ 1\end{bmatrix} = \underline{b}^T\begin{bmatrix}\underline{X} \\ 1\end{bmatrix}.$$

This gets rid of the bias term d by merging it into the unknown vector $\underline{b}$. Note that this requires adding an extra dimension to the observations $\underline{X}$.

$$\mathbf{X} = \begin{bmatrix}1 & \underline{X}_1^T \\ 1 & \underline{X}_2^T \\ \vdots & \\ 1 & \underline{X}_n^T\end{bmatrix} \qquad \underline{y} = \begin{bmatrix}y_1 \\ y_2 \\ \vdots \\ y_n\end{bmatrix}.$$

Note that we have added an extra column of all 1 to the data, to account for the coefficient w_0 in the perceptron network. With this notation, the mean square error becomes

$$MSE = \frac{1}{n}\big(\underline{y} - \mathbf{X}\underline{w}\big)^T(\underline{y} - \mathbf{X}\underline{w}).$$

This is now a quadratic vector optimization problem. We solve this by taking the gradient with respect to $\underline{b}$ and setting this equal to zero:

$$\begin{aligned}\nabla_{\underline{b}} MSE = \nabla_{\underline{b}}\left(\frac{1}{n}\big(\underline{y} - \mathbf{X}\underline{w}\big)^T(\underline{y} - \mathbf{X}\underline{w})\right) = \frac{2}{n}\mathbf{X}^T\left(\mathbf{X}\underline{w} - \underline{y}\right) = 0 \\ \iff \mathbf{X}^T\mathbf{X}\underline{w}^* = \mathbf{X}^T\underline{y}\end{aligned}$$

The last set of equations are known as the **Normal Equations** for least squares estimation. As long as the dimension d is not very large, we can invert the matrix $\mathbf{X}^T\mathbf{X}$ and obtain the regression solution $\underline{w}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\underline{y}$. For large d, different approaches are used, as matrix inversion can be expensive. When d is larger than the number of data points N, this matrix is not invertible, and one must use a different approach, such as computing its pseudo-inverse.

Example 10.5

Consider a 1-dimensional example of regression. The points $X_1, \ldots, X_{100}$ are uniformly spaced in the interval [-1,1]. We generate the measurements y_k as follows:

$$y_k = 2X_k + w_k$$

where w_k are independent samples of a Gaussian random variable with mean 0 and variance 1. Note that we have intentionally selected X to have average 0, and y to also have average 0, so that we can assume $w_0 = 0$. In this case, the data matrix is a vector:

$$\mathbf{X} = \begin{bmatrix}X_1 \\ X_2 \\ \vdots \\ X_n\end{bmatrix}.$$

Based on the 100 values of (X_k, y_k), we generate an estimator as follows:

$$\widehat{y}(x) = w^* x, \qquad w^* = \frac{\sum_{k=1}^N x_k y_k}{\sum_{k=1}^N x_k^2}$$

The results are show in Figure 10.10.

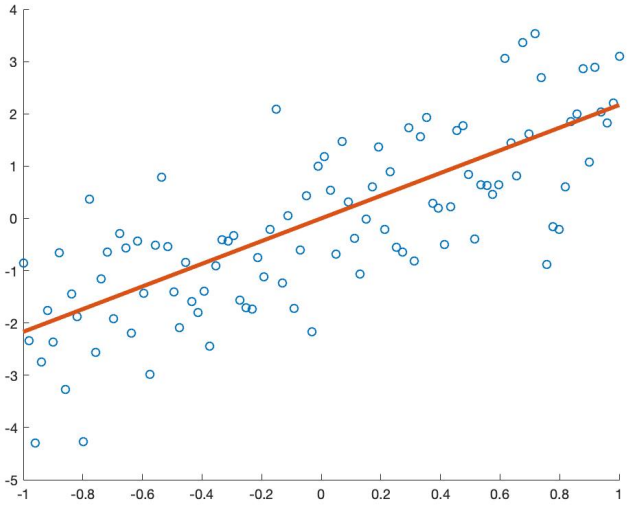


Figure 10.10: One-dimensional regression.

Example 10.6
Consider a 2-dimensional regression. We generate 400 sample points X_k uniformly spaced in the unit square $[-1, 1]^2$. Each of the points X_k is a 2-dimensional vector (X_{k1}, X_{k2}). At each of these points, we observe the function value

$$y_k = 2X_{k,1} + 3X_{k,2} + w_k$$

where w_k are independent samples of a Gaussian random variable with mean 0, variance 1.

The results of our 2-dimensional regression for this case are shown in Figure 10.11. The linear regression estimate is $(\underline{w}^*)^T = \begin{bmatrix} 2.028 & 3.033 \end{bmatrix}$, which are close to the true values used to simulate the data.

10.5 Dimensionality Reduction

In machine learning classification problems, the observations often involve large numbers of variables. For high-dimensional observations, it is hard to visualize the data and design the classifiers using either parametric techniques or optimization techniques. Dimensionality reduction algorithms reduce the number of random variables under consideration, by transforming the data to a smaller set of important features.

For making inferences and other decisions, not all dimensions of the observed data are critical. Consider images, expressed as long vectors. Certain "features" of the images are informative, but not all pixels contain relevant information; most images have similar borders, and the variability in certain regions is similar. Essentially, we wish to discover "features" where images differ the most and discard "features" with little variability.

There are many methods for dimensionality reduction. In this section, we focus on one of the simplest and most popular methods: **Principal Component Analysis (PCA)**. PCA finds linear combinations of the data that are uncorrelated and represent the best approximations to the variability in the data. It is

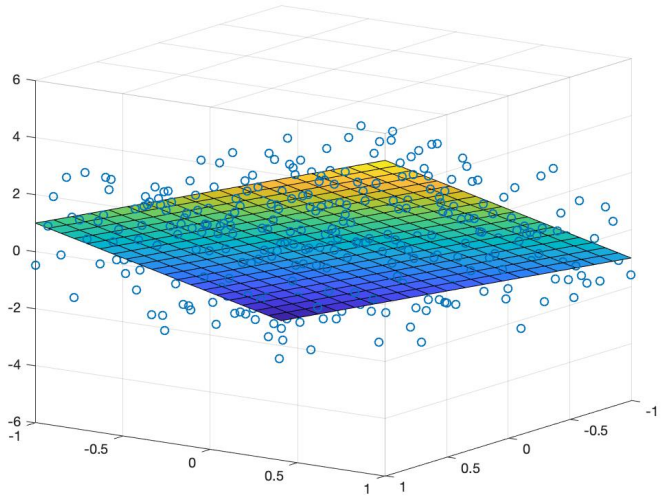


Figure 10.11: Two-dimensional regression.

based on second order statistics (means, variances, covariances), estimated from the training data. PCA is an unsupervised learning algorithm, as it ignores the labels in the data. It is simply trying to find low-dimensional approximations to the total data collected. PCA was invented in 1901 by Karl Pearson. It is also known as the Karhunen–Loève transform (KLT) in signal processing, and as factor analysis in statistics. PCA can be thought of as fitting a low-dimensional ellipsoid in a subspace to data. Each axis of the ellipsoid is a principal component. .

PCA has three main steps:

- Compute the sample covariance matrix of the full-dimensional data.
- Compute eigenvalues and eigenvectors of this covariance matrix.
- Select the eigenvectors associated with the largest k eigenvalues and transform your data into k-dimensional projections onto those eigenvectors.

We describe the mathematics of these steps below, illustrating how they can be computed. The first step is straightforward, and based on estimation results. Assume we are given data consisting of n independent samples of a random vector $\underline{X}$, denoted as $\underline{X}_k, k = 1, \ldots, n$. We assume that $\underline{X}$ takes values in $\Re^d$, where d can be a large number. In many applications where the data is an image, d is often larger than n. The sample mean of this data, which is an estimate of the true mean $\underline{\mu}_{\underline{X}}$, is computed as:

$$\widehat{\underline{\mu}}_{\underline{X}} = \frac{1}{n} \sum_{k=1}^{n} \underline{X}_K$$

Similarly, an unbiased estimate true covariance $\Sigma_{\underline{X}}$ is computed as

$$\widehat{\Sigma}_{\underline{X}} = \frac{1}{n-1} \sum_{k=1}^{n} (\underline{X}_K - \widehat{\mu}_{\underline{X}})(\underline{X}_K - \widehat{\mu}_{\underline{X}})^T$$

Now that we have the sample covariance, let's discuss the motivation for the second step: Computing the eigenvalues and eigenvectors of this covariance matrix. To do this, we will solve an approximation problem. Assume we wanted to project the data $\underline{X}_k$ onto a one-dimensional subspace, defined by the unit vector $\underline{v}_1$.

We do this by computing the inner product between $\underline{v}_1$ and each of the $\underline{X}_k$. Define these projection as $Z_k = \underline{v}_1^T \underline{X}_k, k = 1, \ldots, n$, which are n independent samples of a scalar random variable $Z = \underline{v}_1^T \underline{X}$.

The sample mean of these projections Z_k is

$$\widehat{\mu}_Z = \frac{1}{n} \sum_{k=1}^{n} \underline{v_1}^T \underline{X}_k = \underline{v}_1^T \widehat{\underline{\mu}}_{\underline{X}}.$$

Similarly, the sample covariance of Z is given by the unbiased estimate

$$\begin{aligned}\widehat{\sigma}_Z^2 &= \frac{1}{n-1} \sum_{k=1}^{n} (Z_k - \widehat{\mu}_Z)(Z_k - \widehat{\mu}_Z) = \frac{1}{n-1} \sum_{k=1}^{n} (\underline{v}_1^T \underline{X}_k - \underline{v}_1^T \widehat{\underline{\mu}}_{\underline{X}})(\underline{X}_K^T \underline{v}_1 - \widehat{\underline{\mu}}_{\underline{X}}^T \underline{v}_1) \\ &= \underline{v}_1^T \left(\frac{1}{n-1} \sum_{k=1}^{n} (\underline{X}_k - \widehat{\underline{\mu}}_{\underline{X}})(\underline{X}_k^T - \widehat{\underline{\mu}}_{\underline{X}}^T) \right) \underline{v}_1 = \underline{v}_1^T \left(\widehat{\underline{\Sigma}}_{\underline{X}} \right) \underline{v}_1\end{aligned}$$

In the original d-dimensional space, the approximation to each $\underline{X}_k$ is given by $\underline{V}_k = \widehat{\underline{\mu}}_{\underline{X}} + Z_k \underline{v}_1$. Ideally, we want to select the unit vector $\underline{v}_1$ to best approximate the data on this one-dimensional affine space, by minimizing the mean-square error, defined as $MSE = \frac{1}{n-1} \sum_{k=1}^{n} (\underline{X}_k - \underline{V}_k)^T (\underline{X}_k - \underline{V}_k)$. Define $\underline{\tilde{X}}_k = \underline{X}_k - \widehat{\underline{\mu}}_{\underline{X}}$. Then, the mean square error can be stated in terms the scalars Z_k as

$$\begin{aligned}MSE &= \frac{1}{n-1} \sum_{k=1}^{n} (\underline{\tilde{X}}_k - Z_k \underline{v}_1)^T (\underline{\tilde{X}}_k - Z_k \underline{v}_1) = \frac{1}{n-1} \sum_{k=1}^{n} \left(\underline{\tilde{X}}_k^T \underline{\tilde{X}}_k - 2 \underline{\tilde{X}}_k^T \underline{v}_1 Z_k + \underline{v}_1^T \underline{v}_1 Z_k^2 \right) \\ &= \frac{1}{n-1} \sum_{k=1}^{n} \left(\underline{\tilde{X}}_k^T \underline{\tilde{X}}_k - 2 \underline{\tilde{X}}_k^T \underline{v}_1 Z_k + Z_k^2 \right)\end{aligned}$$

because Z_k is a scalar, and $\underline{v}_1$ is a unit vector. However, recall that $Z_k = \underline{\tilde{X}}_k^T \underline{v}_1$. Hence,

$$\begin{aligned}MSE &= \frac{1}{n-1} \sum_{k=1}^{n} \underline{\tilde{X}}_k^T \underline{\tilde{X}}_k - \frac{1}{n-1} \sum_{k=1}^{n} Z_k^2 = \frac{1}{n-1} \sum_{k=1}^{n} \underline{\tilde{X}}_k^T \underline{\tilde{X}}_k - \frac{1}{n-1} \sum_{k=1}^{n} \underline{v}_1^T \underline{\tilde{X}}_k \underline{\tilde{X}}_k^T \underline{v}_1 \\ &= \frac{1}{n-1} \sum_{k=1}^{n} \underline{\tilde{X}}_k^T \underline{\tilde{X}}_k - \underline{v}_1^T \widehat{\underline{\Sigma}}_{\underline{X}} \underline{v}_1 = \frac{1}{n-1} \sum_{k=1}^{N} \underline{\tilde{X}}_k^T \underline{\tilde{X}}_k - \widehat{\sigma}_Z^2\end{aligned}$$

Thus, to minimize the mean square error in the approximation, we need to select the direction $\underline{v}_1$ to maximize the covariance $\widehat{\sigma}_Z^2$. This results in the following optimization problem:

$$\max_{\underline{v}_1 : \|\underline{v}_1\|^2 = 1} \underline{v}_1^T \left(\widehat{\underline{\Sigma}}_{\underline{X}} \right) \underline{v}_1$$

This is a well-studied optimization problem. The sample covariance matrix is a positive-semidefinite matrix with all eigenvalues real and non-negative, and a full set of orthonormal eigenvectors. Ordering the eigenvalues in decreasing order, so $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, we select $\underline{v}_1$ to be the eigenvector corresponding to the largest eigenvalue λ_1. This eigenvector is called the first principal component.

Note that $\widehat{\underline{\Sigma}}_{\underline{X}} \underline{v}_1 = \lambda_1 \underline{v}_1$, so the mean square error is reduced by λ_1. To obtain the next principal component, we find the unit vector that is orthogonal to $\underline{v}_1$ and maximizes the sample projection covariance $\underline{v}_2^T \left(\widehat{\underline{\Sigma}}_{\underline{X}} \right) \underline{v}_2$. The solution of this problem is the normalized eigenvector corresponding to λ_2, which is the second principal component. When using the first two principal components as approximations, the mean square error is reduced by $\lambda_1 + \lambda_2$.

We can continue this process until we get all the d eigenvectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_d$ of $\widehat{\underline{\Sigma}}_{\underline{X}}$. Let $\mathbf{V} = \begin{bmatrix} \underline{v}_1 & \underline{v}_2 & \ldots & \underline{v}_p \end{bmatrix}$. Since the eigenvectors are chosen to be orthogonal and normalized, we have the following relationships:

$$\widehat{\underline{\Sigma}}_{\underline{X}} \mathbf{V} = \mathbf{V} \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p); \quad \mathbf{V}^T \mathbf{V} = \mathbf{I}_d$$

where $\mathbf{I}_d$ is the d-dimensional identity matrix. For dimensionality reduction, we choose the k largest principal components, to create a projection matrix $\mathbf{V}_k = \begin{bmatrix} \underline{v}_1 & \underline{v}_2 & \dots & \underline{v}_k \end{bmatrix}$.

We can write all of the above operations in matrix notation. Let's define a couple of matrices: an $N \times p$ data matrix $\mathbf{X}$ that stacks all the data as columns in a matrix, with each row representing one sample $\underline{X}_k^T$. Let's also define the $n \times 1$ vector of all ones as $\underline{\mathbf{1}}_n$. Then,

$$\mathbf{X} = \begin{bmatrix} \underline{X}_1^T \\ \underline{X}_2^T \\ \vdots \\ \underline{X}_n^T \end{bmatrix}; \qquad \underline{\mathbf{1}}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

With this notation, the sample mean and covariance of $\underline{X}$ are easily computed as:

$$\widehat{\underline{\mu}}_{\underline{X}} = \frac{1}{n}\mathbf{X}^T\underline{\mathbf{1}}_n; \quad \widetilde{\mathbf{X}} = \mathbf{X} - \underline{\mathbf{1}}_n\widehat{\underline{\mu}}_{\underline{X}}^T; \quad \widehat{\Sigma}_{\underline{X}} = \frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$$

PCA then computes the eigenvector-eigenvalue decomposition of the sample covariance $\widehat{\Sigma}_{\underline{X}}$ as

$$\widehat{\Sigma}_{\underline{X}} = \mathbf{V}\text{diag}(\lambda_1, \dots, \lambda_d)\mathbf{V}^T = \begin{bmatrix} \underline{v}_1 & \dots & \underline{v}_d \end{bmatrix} \text{diag}(\lambda_1, \dots, \lambda_p) \begin{bmatrix} \underline{v}_1^T \\ \dots \\ \underline{v}_d^T \end{bmatrix}$$

Usually, this decomposition is computed using a singular value decomposition algorithm, but there are many other ways of computing this. The final step is to pick a projection matrix $\mathbf{V}_k$ of dimension $d \times k$ with the eigenvectors corresponding to the k largest eigenvalues, as $\mathbf{V}_k = \begin{bmatrix} \underline{v}_1 & \underline{v}_2 & \dots & \underline{v}_k \end{bmatrix}$. We use this matrix to project the data vectors $\underline{X}_j, j = 1, \dots, N$ to k-dimensional "feature" vectors, as

$$\mathbf{X}_{reduce} = (\mathbf{X} - \underline{\mathbf{1}}_n\widehat{\underline{\mu}}_{\underline{X}}^T)]\mathbf{V}_k$$

The matrix $\mathbf{X}_{reduce}$ is of dimension $n \times k$, and each row j is a k- dimensional feature vector corresponding to the original data d-dimensional sample $\underline{X}_j$.

How do we select k? One way is to look at the reduction in mean square error. If $k = d$, the resulting mean square error is zero, so we have reduced the mean square error by a fraction of 1. For smaller k, the fraction reduction in means square error is $\frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_d}$. Selecting this fraction so we reduce the approximation error to under 1% usually yields a good value for k.

To represent the approximation in the original space, we can convert k-dimensional feature vectors to d-dimensional estimates $\widehat{\underline{X}}_j$ of the original data points using the following expression:

$$\widehat{\mathbf{X}} = \mathbf{X}_{reduce}\mathbf{V}_k^T + \underline{\mathbf{1}}_n\widehat{\underline{\mu}}_{\underline{X}}^T$$

In coordinates, let $\underline{Z}_j$ be the transpose of the j-th row of $\mathbf{X}_{reduce}$, which is k-dimensional feature vector approximation of the observation $\underline{X}_j$. Then,

$$\widehat{\underline{X}}_j = \mathbf{V}_k\underline{Z}_j + \widehat{\underline{\mu}}_{\underline{X}}.$$

Example 10.7

To illustrate how PCA works, we show how to approximate a 3-dimensional Gaussian with a 2-dimensional projection. The random vector $\underline{X}$ is assumed to be three-dimensional, with mean $\underline{\mu}_{\underline{X}} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and covariance matrix

$$\Sigma_{\underline{X}} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{bmatrix}^T.$$

The following MATLAB script implements a 2-dimensional PCA approximation of X.

```
mu=[1; 2; 3];       %mean % Generate Gaussian points
A = [1 2 3; 3 1 2; 2 3 1];
Cov = A * A';

 [V,D]=eig(Cov);   %eigen-decomposition
% generate an nx3 matrix of samples points and plot them
n=300;
X=mvnrnd(transpose(mu),Cov,n);
scatter3(X(:,1),X(:,2),X(:,3)); xlabel('x'); ylabel('y'); zlabel('z');
sample_mu=transpose(X)*ones(n,1)/n;
[V_s,Coeff,d_s]=pca(X); hold on;
%plot eigenvectors
quiver3(sample_mu(1),sample_mu(2),sample_mu(3),V_s(1,1),V_s(2,1),V_s(3,1),15,'r');
quiver3(sample_mu(1),sample_mu(2),sample_mu(3),V_s(1,2),V_s(2,2),V_s(3,2),8,'g');
quiver3(sample_mu(1),sample_mu(2),sample_mu(3),V_s(1,3),V_s(2,3),V_s(3,3),8,'b');
% compute reduced 2D representation and plot these points
X_red=Coeff(:,1:2)*transpose(V_s(:,1:2))+ones(n,1)*transpose(sample_mu);
scatter3(X_red(:,1),X_red(:,2),X_red(:,3),'red');
%compute approximation error per element
MSerror=(norm(X-X_red,'fro'))^2/n
```

The results are shown in the figure below, where we plotted both the sample points and their approximations. The figure includes two 3-D views. The view on the right shows that the approximations all lie in a 2-dimensional plane. The resulting mean square error is 2.84 in the approximation.

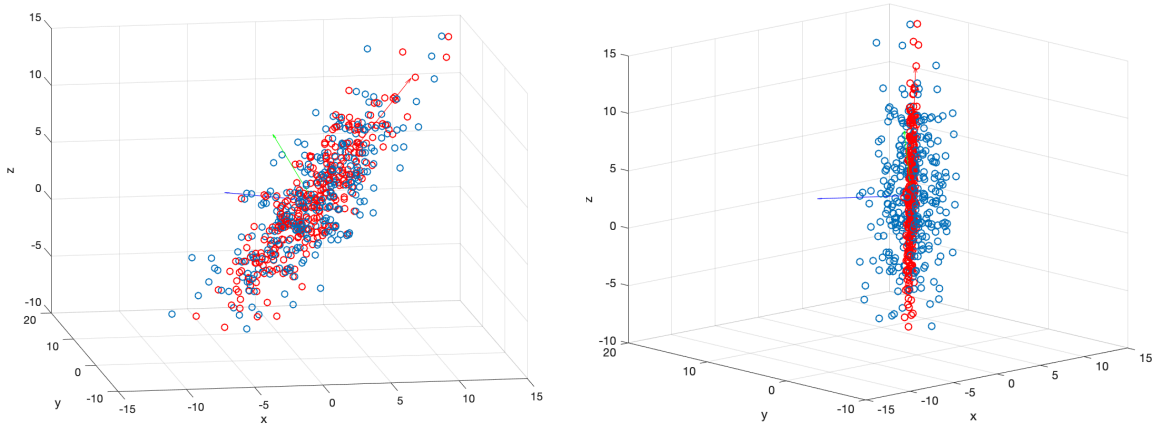


Figure 10.12: Two views of the approximation for Example 10.7.

Example 10.8
For our second example, we consider applying PCA to the IRIS data set, with two and three principal components. The results are shown in Figure 10.13 below. In this case, we cannot visualize the four-dimensional IRIS data and its approximation, so we show the two-dimensional features using the first two principal components, and the three-dimensional features using the first three principal components.

As the figures indicate, with two principal components, PCA captures 97.76% of the total variance in the data, and with three principal components, PCA captures 99.48% of the total variance in the data.

PCA is a powerful tool when combined with parametric classifiers such as LDA and QDA. By reducing the number of features in the data, one reduces the number of parameters which must be estimated in the likelihood functions, and can therefore generate better estimates with a smaller number of samples.

There are several important limitations of PCA. First, PCA ignores the labels in the data, and as such, may not find features that are best for classification. Instead, PCA find features that approximate the

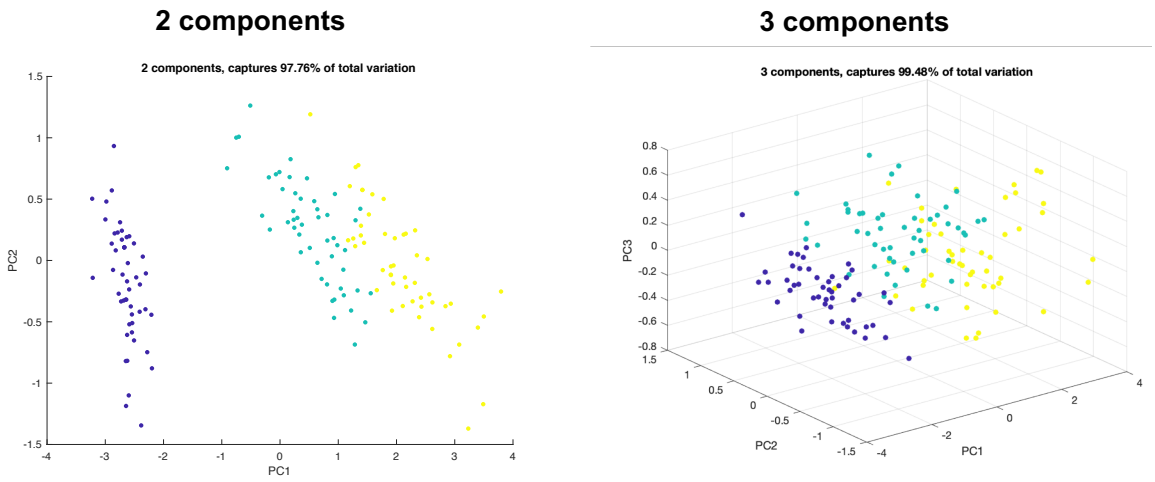


Figure 10.13: Two- and three-dimensional features for the IRIS data set.

data in lower dimensional spaces. Other dimensionality aggregation techniques such as Linear Discriminant Analysis are focused on supervised dimensionality reduction, and find features that separate the different hypotheses.

A second limitation of PCA is that the results depend on the scaling of the different variables. This is because the intrinsic distance used to approximate points is the Euclidean distance. Thus, in practice, one often normalizes each dimension of the data by an estimate of the standard deviation, so that all marginal distributions have variance 1.

The third limitation is that PCA is focused on finding features that are linear combinations of the data. In many feature sets, the best features may be nonlinear combinations. Extensions of PCA such as kernel PCA attempt to address this, by finding features using nonlinear kernels instead of inner product projections.

A common criticism of the features obtained by PCA is that they are linear combinations of all the dimensions d in the data, and hence do not provide insight into what the features mean. This is often expressed when PCA is used on medical data. Techniques such as sparse PCA and regularized PCA have been introduced to generate features that use only a fraction of the dimensions d.

In spite of the above limitations, PCA is broadly used as the first approach to dimensionality reduction because of its computational simplicity and its robust performance for high-dimensional data.

10.6 Summary

Machine learning addresses problems of classification and regression, in cases where we don't know the probabilistic relationship between observed values $\underline{X}$ and the labels or numeric values Y that we are trying to predict. Instead, we are provided with labeled training samples $(\underline{X}_1, Y_1), \ldots, (\underline{X}_N, Y_N)$ which form the basis for the design of classification and regression algorithms.

The algorithms discussed in this chapter comprise a small sample of the available techniques in the field. We have focused on algorithms where the learning is simple: there are no complex training procedures required to design the decision and estimation algorithms from training data. Thus, we avoided algorithms such as deep neural networks and support vector machines, where finding the parameters of the algorithm involve the solution of large-dimensional, complex optimization problems.

Chapter 11

Markov Chains

In the chapter on limit theorems, we saw sequences of random variables indexed by the natural numbers. The underlying experiments generated sequences of independent, identically distributed random variables, from which we constructed derived sequences such as the partial sum or the incremental average of the random variables.

Collections of random variables $\{X_t\}$ indexed by the natural numbers are known as discrete-time stochastic processes, or discrete-time random processes. The index is used to represent time. Such models are often used to represent random time signals that arise in dynamical systems, and have many interesting applications in engineering.

In this chapter, we focus on a class of discrete time random processes known as Markov chains. Markov chains are a special class of discrete time random processes because of two properties. First, the range of the individual variables X_t is discrete, which will will allow us to develop a rich connection between the probability models and concepts from graph theory. Second, the joint probability distribution functions of Markov chains will satisfy the Markov property, which we discuss later in this chapter. Markov chains can also be defined as random processes that are indexed in continuous time, but those extensions are outside the scope of this course.

Markov chains were introduced by Andrey Markov in 1906 to study extensions of the Law of Large Numbers and the Central Limit Theorem to sequences where the random variables were not independent and identically distributed. Such models form the basis for many interesting applications such as speech recognition, communications networks analysis and stochastic automata. Markov chains provide the foundation for many of today's leading technologies. Google's page rank algorithm was based on a Markov chain model of how websites are visited. Viterbi decoding, named after one of Qualcomm's founder, is based on Hidden Markov Model techniques, and is used extensively in modern communications. Markov chain models play fundamental roles in speech and natural language recognition. Markov models are used extensively in mathematical finance to analyze expected returns of different investment mixtures. They also provide the foundation for the analysis and design of network systems for handling random traffic demands.

In the remainder of this chapter, we discuss the foundations of discrete-time Markov Chains and explore their properties. First, we will introduce discrete-time, discrete-space Markov processes, and define the Markov property that characterizes such processes. Following this, we develop tools for computing probabilities in Markov chains. We develop approaches for characterizing how the marginal probability of the Markov chain evolves with the time index, and explore the limiting behavior of such systems. We also introduce tools for analysis of the transient behavior of Markov chains.

11.1 Definition of Markov Chains

Let $\mathcal{R}_X$ be a finite, or countably infinite set of possible values, which we call the **state space**. This set is a subset $\mathcal{R}_X \subset \Re$. Define a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ that generates a countably infinite sequence of random variables $X_0, X_1, X_2, \ldots$, each of which takes values in the state space $\mathcal{R}_X$. An outcome $\omega \in \Omega$ generates a sequence of numbers $X_0(\omega), X_1(\omega), \ldots$ with values in $\mathcal{R}_X$. For each outcome, we refer to this sequence as a trajectory of the Markov chain.

Given a finite subsets of these random variables $X_{t_1}, \ldots, X_{t_n}$, where $t_1, \ldots, t_n \in \{0, 1, 2, \ldots\}$, we can

compute joint probability mass functions of the form $P_{X_{t_1},\ldots,X_{t_m}}(x_{t_1},\ldots,x_{t_m})$. These joint probability mass functions (PMF) can be used to generate conditional probability mass functions as well as marginal probability mass functions. We refer to the indices t as times, so we think of $X_t(\omega), t = 0, 1, \ldots$ as a trajectory over time.

Without loss of generality, assume the indices $t_1 < t_2 < \ldots < t_m$ are ordered linearly in time. Using conditional probabilities and the product rule for probability mass functions, we can write the joint PMF of the random variables with those indices as

$$P_{X_{t_1},\ldots,X_{t_m}}(x_{t_1},\ldots,x_{t_m}) = P_{X_{t_m}|X_{t_1},\ldots,X_{t_{m-1}}}(x_{t_m}|x_{t_1},\ldots,x_{t_{m-1}})$$
$$P_{X_{t_{m-1}}|X_{t_1},\ldots,X_{t_{m-2}}}(x_{t_{m-1}}|x_{t_1},\ldots,x_{t_{m-2}})\cdots P_{X_{t_2}|X_{t_1}}(x_{t_2}|x_{t_1})P_{X_{t_1}}(x_{t_1})$$

We say that the sequence of random variables $X_0, X_1, X_2 \ldots$, satisfies the **Markov Property** if and only if, for any set of times $t > t_m > \ldots > t_1$, we have

$$P_{X_t|X_{t_1},\ldots,X_{t_m}}(x_t|x_{t_1},\ldots,x_{t_m}) = P_{X_t|X_{t_m}}(x_t|x_{t_m}).$$

That is, the conditional probability mass function of the random variable at time t, X_t, given values of random variables at different previous times $t_1, t_2, \ldots, t_m$, depends only on the value of the most recent random variable in its past. This simplifies how we write the joint probability mass function, as

$$P_{X_{t_1},\ldots,X_{t_m}}(x_{t_1},\ldots,x_{t_m}) = P_{X_{t_m}|X_{t_{m-1}}}(x_{t_m}|x_{t_{m-1}})P_{X_{t_{m-1}}|X_{t_{m-2}}}(x_{t_{m-1}}|x_{t_{m-2}})\cdots P_{X_{t_2}|X_{t_1}}(x_{t_2}|x_{t_1})P_{X_{t_1}}(x_{t_1})$$

Thus, we can specify the joint probability mass function (PMF) of a collection of random variables in terms of a product of pairwise conditional PMFs times the marginal PMF of the random variable with the earliest time index. This economical description is very useful in obtaining an economical probabilistic description of the Markov chain.

Of particular interest is the one-step conditional probability $P_{X_{t+1}|X_t}(x_{t+1}|xt)$. In general, this conditional probability depends on time. Assume that the state space is given as $\mathcal{R}_X = \{a_1, a_2, \ldots, a_n, \ldots\}$. Then, $P_{X_{t+1}|X_t}(x_{t+1} = a_k|x_t = a_j)$ depends on a_k, a_j, and t. When this conditional probability does not depend on t, we say the Markov chain is **homogeneous** or **time-invariant**. Homogeneous Markov chains have the nice property that the conditional probability mass function $P_{X_{t+1}|X_t}(x_{t+1} = a_k|x_t = a_j)$ is the same for all $t = 0, 1, 2, \ldots$. Hence, the full probability description of the Markov chain can be obtained from the marginal PMF $P_{X_0}(x_0)$ and the one-step conditional probability $P_{X_{t+1}|X_t}(x_{t+1} = a_k|x_t = a_j)$. As shorthand notation, we define the **transition probability kernel** of the Markov chain as a matrix $\mathbf{P}$ with elements defined as:

$$\mathbf{P}_{jk} = P_{X_{t+1}|X_t}(x_{t+1} = a_k|x_t = a_j),\ j, k \in \{1, 2, \ldots\}.$$

Thus, $\mathbf{P}_{jk}$ is the probability that, if the random variable X_t has value a_j, then the random variable X_{t+1} will take value a_k. The transition probability kernel has the following properties:

- $1 \geq \mathbf{P}_{jk} \geq 0$ for all $j, k \in \{1, 2, \ldots\}$. This follows because it was defined as a conditional probability, which is a probability.
- $\sum_k \mathbf{P}_{jk} = 1$. This property is the normalization property for conditional PMFs.

We refer to the random variable X_t as the **state** at time t. The Markov chain provides a probabilistic description of how the state X_t evolves over time.

Example 11.1
Consider the following Markov chain, where the state space is $\mathcal{R}_X = \{1, 2, 3, 4, 5\}$. We assume initially that $X_0 = 3$; that is, uniform; that is, $P_{X_0}(3) = 1, P_{X_0}(x) = 0$ if $x \neq 3$. Thus, we have defined the marginal PDF at time 0. We now describe the transition probability kernel, as follows: if $j \neq 1, j \neq 5$, then

$$P_{X_{t+1}|X_t}(x_{t+1} = k|x_t = j) = \begin{cases} 0.5 & k = j + 1 \\ 0.5 & k = j - 1 \\ 0 & \text{elsewhere.} \end{cases}$$

For $j = 1$, the transition probability kernel is

$$P_{X_{t+1}|X_t}(x_{t+1} = k|x_t = 1) = \begin{cases} 0.5 & k = 2 \\ 0.5 & k = 1 \\ 0 & \text{elsewhere.} \end{cases}$$

For $j = 5$, the transition probability kernel is

$$P_{X_{t+1}|X_t}(x_{t+1} = k|x_t = 5) = \begin{cases} 0.5 & k = 5 \\ 0.5 & k = 4 \\ 0 & \text{elsewhere.} \end{cases}$$

Note that we can represent this transition probability kernel as a matrix $\mathbf{P}$, where

$$\mathbf{P} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

We now have a complete description of the probabilistic structure of the Markov chain. We can answer questions such as: What is the probability that $X_0 = 3, X_1 = 2, X_2 = 1$? Note that this will be $P_{X_0}(3)\mathbf{P}_{32}\mathbf{P}_{21} = 0.25$. Another question might be what is the probability that $X_3 = 3$? Although we don't have an easy way of computing this yet, we see that there are two ways that $X_3 = 3$, which is with $X_2 = 2$ and $X_2 = 4$. Each of those two paths will have probability 0.25, so the probability that $X_3 = 3$ will be 0.5.

More rigorously, we would compute the joint probability of $X_1 = 3, X_2 = k, X_3 = 3$ as $\mathbf{P}_{3k}\mathbf{P}_{k3}$. To get the probability that $X_3 = 3$, we would sum over k this joint probability, thereby marginalizing the intermediate random variable $X_2 = k$. It so happens that this product is nonzero only for $k = 2$ and $k = 4$, so the sum is again 0.5.

In the special case that the state space $\mathcal{R}_X$ is finite, the set of possible states is $\{a_1, a_2, \ldots, a_K\}$, and the transition probability kernel $\mathbf{P}_{jk}$ can be represented as a $K \times K$ matrix $\mathbf{P}$ with $(j,k)-$element $\mathbf{P}_{jk}$. In this case we denote $\mathbf{P}$ as the **state transition matrix** or the **transition probability matrix** of the Markov chain. We study the special case of homogeneous, finite state Markov chains next.

11.2 Finite State Markov Chains

11.2.1 Graphical representation of the Markov chain

Consider a finite state Markov chain, with state space $\mathcal{R}_X = \{a_1, a_2, \ldots, a_K\}$. To simplify notation, we assume $\mathcal{R}_X = \{1, 2, \ldots, K\}$. For a homogeneous, finite state Markov chain, the transition probability kernel is represented by a state transition matrix $\mathbf{P}$, with properties

- $\mathbf{P}_{jk} \in [0, 1],\ j, k \in \{1, \ldots, K\}$.
- $\sum_{k=1}^{n} \mathbf{P}_{jk} = 1$ for $j = 1, \ldots, K$.

That is, all of the elements of $\mathbf{P}$ are nonnegative numbers less than or equal to 1, and the sum of every row equals one. Matrices that satisfy these two properties are known as **stochastic matrices.** Later in this section, we will describe some useful properties of stochastic matrices that help us understand the behavior of Markov chains.

The state transition matrix $\mathbf{P}$ is often sparse, containing many zeros. In Example 11.1, over half the matrix was composed of zeros. We can represent the contents of the matrix $\mathbf{P}$ in graphical form, where nodes indicate possible values of the state, and directed arcs between nodes represent transition probabilities. Thus,

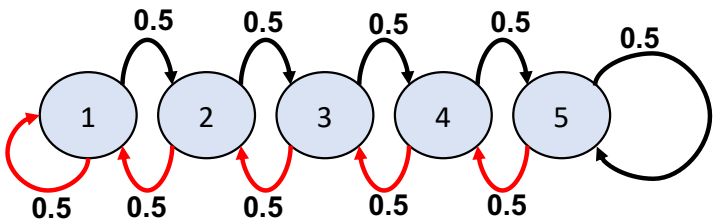


Figure 11.1: Graph of Markov chain state transition matrix for Example 11.1.

the graph contains K nodes (the cardinality of $\mathcal{R}_X$) and a number of directed arcs equal to the number of positive elements in $\mathbf{P}$. Figure 11.1 shows the graph that represents the state transition matrix $\mathbf{P}$ in Example 11.1. Note the directed arcs, with weights that correspond to the non-zero entries of $\mathbf{P}$. The condition that the rows of $\mathbf{P}$ must each add up to 1 implies that the sum of the probabilities of the arcs that leave each node must equal 1. This includes self-loop arcs where the transition is from a particular state to itself.

Example 11.2
Consider a four state Markov chain, with state transition matrix shows the graph for a four state Markov matrix

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & 0 & 0 \\ 0 & 0 & P_{23} & P_{24} \\ P_{31} & 0 & 0 & 0 \\ 0 & 0 & P_{43} & P_{44} \end{bmatrix}$$

What is the graph of the Markov chain?

The graph is shown in the figure below. The graph has 7 directed arcs, corresponding to the 7 non-zero elements of $\mathbf{P}$.

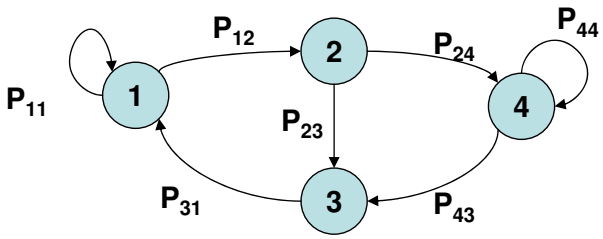


Figure 11.2: Graph of Markov Chain transition probabilities.

Example 11.3
One of the simplest Markov chain models has two states, $\mathcal{R}_X = \{1, 2\}$, corresponding to an on-off system. This model is often used for failure-repair processes. When the model is in state 1, the "on" state, there is a probability of failure p at each time. Eventually, a failure happens, and the state of the system transitions to state 2, the "off" state. In this state, there is a probability of repair q at each time. Eventually, the state transitions back to the "on" state 1. The state transition diagram is shown in Figure 11.3.

The graph representation of the state transition matrix helps us understand how the Markov chain behaves as a function of time. One view of the Markov chain is that it is a collection $\{X_t, t = 0, 1, \ldots\}$ of random variables with joint probability mass functions that satisfy the Markov property. A different view is to consider the sequence of values $\{X_0(s), X_1(s), \ldots\}$ that would occur from a single realization s of the experiment that generated the chain. We refer to such a sequence as a **trajectory** of the Markov chain. A trajectory is a time sequence of state values, and can be viewed as a trajectory on the graph, where transitions between states that are adjacent in time can only happen if there is a directed arc from the previous state to the next state. With this perspective, the Markov chain generates a probability distribution

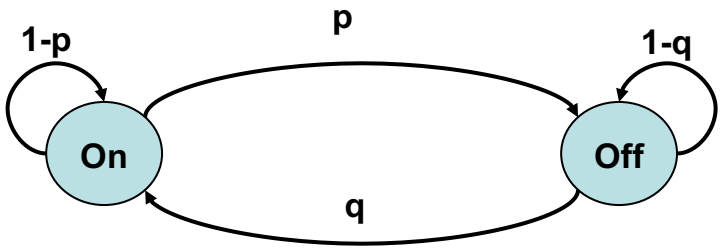


Figure 11.3: Graph of Markov Chain transition probabilities.

over possible state trajectories on the Markov chain graph. The Markov property establishes that, given knowledge that the chain is in state $X_t = k$ at time t, the probability distribution on the future trajectory of the state depends only on $X_t = k$, and not on any values $X_s, s < t$. Thus, $X_t = k$ has all the information needed to predict the future state values $X_\tau, \tau > t$.

Example 11.3 illustrates an important property of the state trajectories of Markov chains. We know the system remains in the same state for a random number of time steps before transitioning to another state. A possible state trajectory for the first 30 steps is 000000011000000000011110000011, where we see the trajectory start at state 0, stay there for 7 times before transitioning to state 1 in time 8. The next time the system visits state 0, it transitions to state 1 in 11 times. Because of the Markov property, the amount of time it takes to transition out of state 0 has the same distribution for every visit in the trajectory. For each state k, define the random variable $H_k(\omega)$ as follows:

$$H_k(\omega) = \min_{t>0}\{t : X_0(\omega) = k, X_t(\omega) \neq k\}.$$

Note we have included the explicit dependence on the realization of the trajectory ω. $H_k(\omega)$ is the first exit time that the Markov chain trajectory would leave state k, given that it started at time 0 in state k. .

H_k is a discrete random variable, with values in $\{1, 2, \ldots\}$. It can even take an arbitrarily large value, albeit with decreasing probability. The following result characterizes the PMF of H_k.

Lemma 11.1
For a homogeneous Markov chain with state transition matrix $\mathbf{P}$, the first exit time from state k, H_k is a geometric random variable with success probability $1 - \mathbf{P}_{kk}$.

To show this, note that, if the Markov chain is in state $X_0 = k$ at time 0, the probability that it exits at the next time is $1 - \mathbf{P}_{kk}$. Thus, $\mathbb{P}[H_k = 1] = 1 - \mathbf{P}_{kk}$. If it does not exit, then the chain remains in state $X_1 = k$ with probability $\mathbf{P}_{kk}$. The event that the chain exits at time 2 is independent of the prior history of the Markov chain, because of the Markov property, and has probability q_k of occurring. Hence, $\mathbb{P}[H_k = 2] = (1-\mathbf{P}_{kk})\mathbf{P}_{kk}$, and the probability that $X_2 = k$ is P^2_{kk}. Continuing by induction, we can establish that $\mathbb{P}[H_k = \ell] = (1 - \mathbf{P}_{kk})(\mathbf{P}_{kk})^{\ell-1}$, which is the PMF of a geometric random variable.

11.2.2 Evolution of marginal probabilities

Let $X_t, t = 0, 1, \ldots$ be a discrete time, finite-valued Markov chain with values in $\mathcal{R}_X = \{1, 2, \ldots, K\}$. The Markov chain has a marginal distribution at $t = 0$ as $P_{X_0}(x_0)$. We can represent this distribution as a vector, as illustrated below.

$$\underline{p}(0) = \begin{bmatrix} P_{X_0}(1) \\ P_{X_0}(2) \\ \vdots \\ P_{X_0}(K) \end{bmatrix}$$

Similarly, we denote the marginal PMF of X_t as a vector $\underline{p}(t)$, defined as

$$\underline{p}(t) = \begin{bmatrix} P_{X_t}(1) \\ P_{X_t}(2) \\ \vdots \\ P_{X_t}(K) \end{bmatrix}$$

The state transition matrix $\mathbf{P}$ can be used to compute the evolution of the marginal probability vectors $\underline{p}(t)$ over time, as follows: Note that, at time 1,

$$P_{X_0,X_1}(j,k) = P_{X_1|X_0}(k|j)P_{X_0}(j) = \mathbf{P}_{jk}P_{X_0}(j)$$

Hence, the marginal probability at time 1 is given by summing over the possible values j of X_0, as

$$P_{X_1}(k) = \sum_{j=1}^{K} \mathbf{P}_{jk}P_{X_0}(j)$$

which can be written in terms of matrix operations as

$$\underline{p}(1) = \mathbf{P}^T\underline{p}(0)$$

Extending the above argument inductively yields the following recursion:

$$\underline{p}(t) = \mathbf{P}(t)^T\underline{p}(0),$$

where $\mathbf{P}(m) \equiv \mathbf{P}^m$ for $m \geq 0$ is the m-step transition probability matrix. The multistep transition matrix satisfies the Chapman-Kolmogorov equation

$$\mathbf{P}(n+m) = \mathbf{P}(m)\mathbf{P}(n) = \mathbf{P}(n)\mathbf{P}(m) \quad \text{for } n, m \geq 0.$$

Note that $\mathbf{P}(0)$ is the K-dimensional identity matrix $\mathbf{I}_K$.

Note that the state transition matrix $\mathbf{P}$ and the multi-step transition matrix $\mathbf{P}(m)$ must satisfy the laws of conservation of probability. That is, for any row k, we must have

$$\sum_{j=1}^{\infty} \mathbf{P}_{kj} = 1; \quad \sum_{j=1}^{\infty} \mathbf{P}(m)_{kj} = 1;$$

Example 11.4

Assume a person starts in the middle of a room. At each time, with probability $p = 0.5$, they take a step to the right. With probability 0.5, they take a step to the left. However, if they are at the wall, and they try to take a step into the wall, they stay in place. Assume the walls on the left and right are five steps away from the center of the room. What is the probability that the person will be next to the right wall at time 10?

The figure below illustrates the Markov chain for this problem, under the assumption that $p = 0.5$. The starting position is in state 6, so that $P_{X_0}(6) = 1$. The state transition matrix is given by

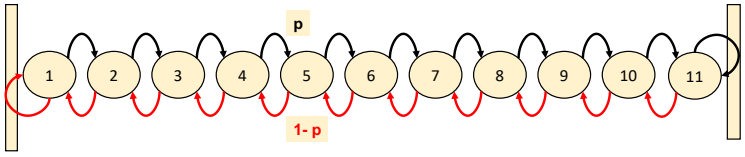


Figure 11.4: Random walk in a closed room.

$$\mathbf{P} = \begin{bmatrix} 1-p & p & 0 & 0 & \cdots & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 \\ 0 & 1-p & 0 & p & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1-p & 0 & p \\ 0 & 0 & \cdots & 0 & 1-p & p \end{bmatrix}$$

We are interested in computing $\mathbf{P}(10)_{6,11}$, the probability that, starting at state 6 at time 0, we are in state 11 at time 10. By direct computation, we get $\mathbf{P}(10)_{6,11} = 0.0439$. How would this change if we increased the time to 20? The probability of being next to the wall increases to 0.0741. If we consider the same question at time 100, the probability increases to 0.0905. After 200 steps, the marginal probability vector is

$$\underline{p}(200) = \begin{bmatrix} 0.0909 \\ 0.0909 \\ 0.0909 \\ 0.0909 \\ 0.0909 \\ 0.0909 \\ 0.0909 \\ 0.0909 \\ 0.0909 \\ 0.0909 \\ 0.0909 \end{bmatrix},$$

having reached a steady state.

Example 11.5
In Example 11.4, we saw the marginal probability vector $\underline{p}(t)$ approach a limit as $t \to \infty$. Do we see similar behavior in other examples? Consider the on-off system of Example 11.3. Let $p = 0.1, q = 0.2$. In this case, the state transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}.$$

Assuming we start in the "on" state 1, we compute the marginal probability vector after 5, 10, 20, and 40 times. The results are shown below:

$$\underline{p}(5) = \begin{bmatrix} 0.7227 \\ 0.2773 \end{bmatrix}; \ \underline{p}(10) = \begin{bmatrix} 0.6761 \\ 0.3239 \end{bmatrix}; \ \underline{p}(20) = \begin{bmatrix} 0.6669 \\ 0.3331 \end{bmatrix}; \ \underline{p}(40) = \begin{bmatrix} 0.6667 \\ 0.3333 \end{bmatrix}.$$

Again, we see the marginal probability vector approach a steady state with increasing t.

Assume that the marginal distribution vectors converge to a steady state marginal distribution $\underline{\pi}$. In this case, this steady state distribution must satisfy $\mathbf{P}^T\underline{\pi} = \underline{\pi}$. That implies that $\underline{\pi}$ is an eigenvector of the matrix $\mathbf{P}^T$, corresponding to an eigenvalue of 1. We know that $\mathbf{P}$ has an eigenvalue of 1, with eigenvector corresponding to the K-dimensional vector of all ones, because the sum of every row of $\mathbf{P}$ equals one. That is,

$$\mathbf{P}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} P_{1k} \\ \sum_{k=2}^{K} P_{1k} \\ \vdots \\ \sum_{k=1}^{K} P_{Kk} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Since the eigenvalues of $\mathbf{P}$ and $\mathbf{P}^T$ are the same, $\mathbf{P}^T$ also has an eigenvalue of 1, with corresponding eigenvector. Note also that $\underline{\pi}$ is the limit of a sequence of marginal probability mass functions, and hence the limit will also be a valid probability mass function: $\pi_k \in [0,1], \sum_{k=1}^{K} \pi_k = 1$.

To better understand the limit behavior of Markov chains, we discuss the properties of stochastic matrices that control the evolution of the marginal distributions.

11.2.3 Stochastic matrices

When the number of states is finite and equal to K, the state transition matrix will be an $K \times K$ matrix $\mathbf{P}$, where $\mathbf{P}$ is such that all of its entries are nonnegative and the rows sum up to 1. Nonnegative matrices with the property that the rows sum up to 1 are known as **stochastic matrices**.

We first quote a theorem for linear algebra that relates the locations of the eigenvalues of matrices to the elements of its rows.

Theorem 11.1 (Gershgoren's Theorem)
Consider a square matrix $\mathbf{A}$ of dimension $K \times K$. Define distances $d_i = \sum_{j=1, j\neq i}^{n} |\mathbf{A}_{ij}|$. Define the set of complex numbers

$$L = \{\lambda \in \mathbb{C} : |\lambda - \mathbf{A}_{ii}| \leq d_i \text{ for some } i \in \{1, \ldots, K\}\}.$$

Then, all of the eigenvalues of $\mathbf{A}$ are contained in the set L.

The distance d_i is the sum of the magnitude of the off-diagonal elements in row i. The set L consists of the union of circles of radius d_i centered around each of the diagonal elements $\mathbf{A}_{ii}$. Figure 11.5 illustrates the implications of Gershgoren's theorem for the matrix $A = \begin{bmatrix} 3 & 2 \\ 1 & 1 \end{bmatrix}$. The eigenvalues must lie in the union of two circles in the complex plane, centered at the diagonal elements (3,0) and (1,0), with radii 2 and 1, respectively. By direct computation, the eigenvalues are 3.7321 and 0.2679, which are in the union of the two circles.

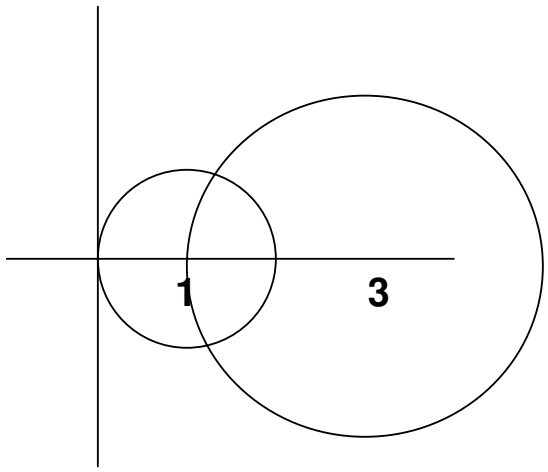


Figure 11.5: Illustration of Gershgoren's Theorem.

For stochastic matrices $\mathbf{A}$, the rows add up to 1, and all the elements are non-negative. This means that $d_i + \mathbf{A}_{ii} = 1$, and the center of the circle is on the non-negative real line. Hence, all of the eigenvalues of a stochastic matrix must be on or inside the unit circle of radius 1, centered at 0. Furthermore, since every row adds to 1, we know that the vector $\underline{1} = [1, 1, \ldots, 1]^T$ satisfies $\mathbf{A}\underline{1} = \underline{1}$, and is thus an eigenvector of the matrix $\mathbf{A}$ with eigenvalue equal to 1.

Figure 11.6 illustrates Gershgoren's theorem for stochastic matrices. Note that all of the eigenvalues λ of $\mathbf{A}$ must satisfy $|\lambda| \leq 1$.

The other theorem from mathematics that relates to the eigenvalues and eigenvectors of stochastic matrices is the Perron-Frobenius theorem, stated below:

Theorem 11.2 (Perron-Frobenius Theorem)
Consider a square matrix $\mathbf{A}$ of dimension $K \times K$ with non-negative elements. Then, there exists a non-negative real eigenvalue λ_{PF} with associated non-negative eigenvector, such that $|\lambda| \leq \lambda_{PF}$ for any other eigenvalue λ of $\mathbf{A}$. Furthermore, if $\mathbf{A}$ is such that $\mathbf{A}^k$ is strictly positive for some k, then $|\lambda| < \lambda_{PF}$ and the associated eigenvector with λ_{PF} can be chosen as strictly positive.

The Perron-Frobenius theorem establishes that $\lambda_{PF} = 1$ and that the associated eigenvector $\underline{\pi}$ can be chosen so that $\underline{\pi}$ is non-negative. Furthermore, it establishes the condition that is needed to ensure that $\underline{\pi} > 0$ and is a unique stationary density: if there exists k such that every element of $\mathbf{P}^k = \mathbf{P}(k)$ is positive. We will provide graphical conditions that are necessary and sufficient for this to be true.

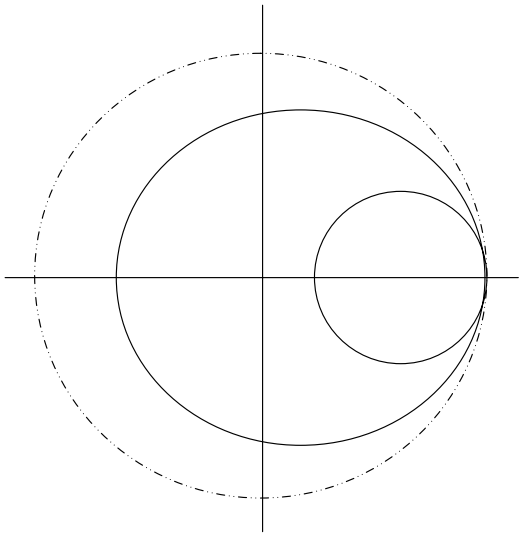


Figure 11.6: Illustration of Gershgoren's Theorem for stochastic matrices.

Example 11.6
Consider the "on"-"off" example in Example 11.5, with state transition matrix

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

The eigenvalues of this matrix are the solution of the quadratic equation

$$(s-p)(s-q) - (1-p)(1-q) = s^2 - (p+q)s + pq - 1 + (p+q) - pq = (s-1)(s-(p+q-1)) = 0$$

which are $1, p+q-1$. The magnitude of the second eigenvalue is strictly less than 1, unless both p, q are either 0 or 1. Note that, if $p, q \in (0,1)$, then $\mathbf{P} > 0$ and, by the Perron-Frobenius Theorem, there is at most one eigenvalue with magnitude 1, and the limit eigenvector can be chosen to be strictly positive. The eigenvector of $\mathbf{P}^T$ corresponding to the eigenvalue 1 satisfies:

$$\mathbf{P} = \begin{bmatrix} 1-p & q \\ p & 1-q \end{bmatrix} \underline{\pi} = \underline{\pi}.$$

This results in the equations

$$\begin{aligned} (1-p)\pi_1 + q\pi_2 = \pi_1 &\iff -p\pi_1 + q\pi_2 = 0 \\ p\pi_1 + (1-q)\pi_2 = \pi_2 &\iff p\pi_1 - q\pi_2 = 0 \end{aligned}$$

which reduce to $\pi_2 = \frac{p}{q}\pi_1$. To find π_1, we use the normalization property of PMFs, which says that $\pi_1 + \pi_2 = \pi_1(1+\frac{p}{q}) = 1$. This implies that $\pi_1 = \frac{q}{p+q}, \pi_2 = \frac{p}{p+q}$.

11.2.4 Steady-state behavior of Markov chains

As discussed previously, the marginal probability mass function $\underline{p}(t)$ evolves according to a linear system:

$$\underline{p}(t+1) = \mathbf{P}^T \underline{p}(t)$$

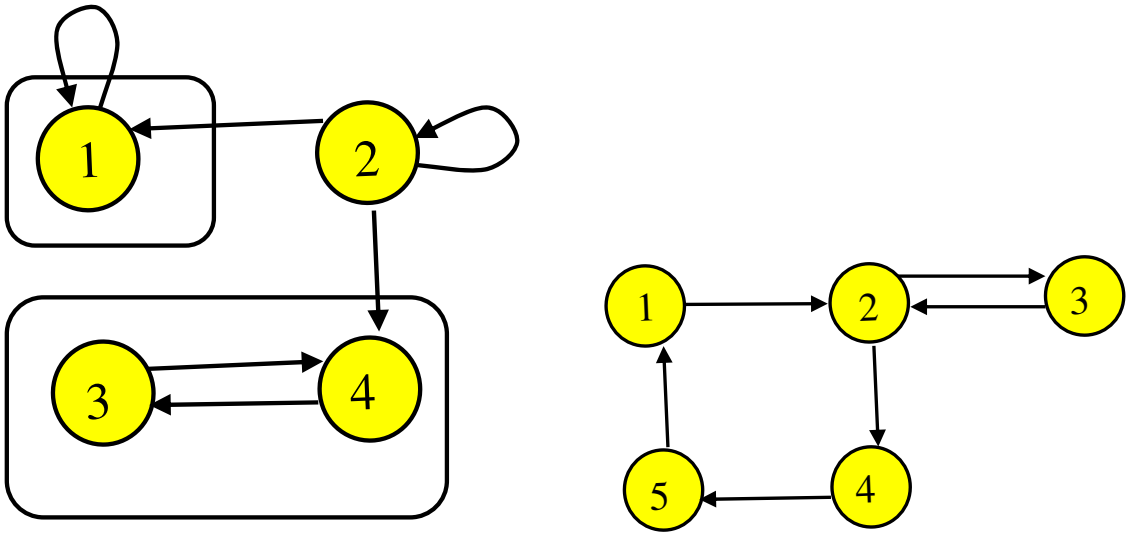
For homogeneous Markov chains in discrete time, this equation may have a limit as $t \to \infty$, as all the eigenvalues of $\mathbf{P}$ will have magnitude less than or equal to 1. We are interested in providing conditions where

$$\lim_{t\to\infty} \mathbf{P}^t = \mathbf{P}_\infty$$

and

$$\lim_{t\to\infty} \underline{p}(t) = \lim_{t\to\infty} (\mathbf{P}^t)^T \underline{p}(0) = \mathbf{P}_\infty^T \underline{p}(0) = \underline{\pi}$$

To illustrate issues that can arise, consider the two graphs illustrated in Figures 11.7(a) and 11.7(b). The first graph shows that, after starting in state 2, one can either go to state 1 or to states 3 and 4. Depending on which transition is used, the limit will be different. It is clear that this Markov chain may have multiple limiting distributions. The second figure illustrates a more complex case. If one starts in state 1 at time 0, note that one can only be in an odd-valued state at even times! This Markov chain will not approach a limit, but rather will oscillate between two limits!



(a) Markov chain with two limits (b) Period 2 Markov chain

Figure 11.7: Illustration of Markov chains with difficult limit behavior.

For finite state Markov chains, one can define regularity conditions that guarantee that there is a unique eigenvalue of P with magnitude 1, so that there are unique limits. Furthermore, these conditions can be established from the transition diagram of the Markov chain! We discuss these next.

Consider two states i, j of the Markov chain. State j is said to be **accessible** from state i if there exists a time n such that $(P^n)_{ij} > 0$. An equivalent graphical condition is that there exists a *directed path* with positive probability arcs from node i to node j in the Markov chain graph. In the reflected random walk diagram of Figure 11.4 in Example 11.4, every state is accessible from every other state. However, consider the minor variation shown in Figure 11.8, where one of the feasible arcs has been removed. In this case, state 7 is accessible from state 6, but state 6 is not accessible from state 7.

Two states i, j are said to **communicate** if i is accessible from j and j is accessible from i; by convention, every state is said to communicate with itself. Communication is a transitive, symmetric and reflexive binary relationship, hence it is an equivalence relationship. A **communicating class** is a non-empty set of states that communicate with each other, and no state in the class communicates with any state outside the class. The set of possible states of a finite-valued Markov Chain can be partitioned into disjoint communicating classes. For instance, the Markov Chain illustrated in Figure 11.8 has 2 communicating classes: $\{1, 2, 3, 4, 5, 6\}$ and $\{7, 8, 9, 10\}$.

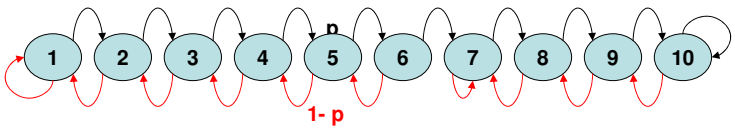


Figure 11.8: Example of Markov Chain with inaccessible states

When a Markov Chain has only one communicating class, it is said to be **irreducible**. In irreducible Markov Chains, every state communicates with every other state, as in Fig. 11.4.

A state i in a homogeneous Markov Chain is said to be **transient** if, given that the Markov Chain starts at state i, there is a non-zero probability that the state never returns to state i. Formally, assume $X_0 = i$, and define the random time $T = \min[t > 0 : X_t = i]$. Then, $\mathbb{P}[\{T = \infty\}] > 0$. Thus, there is positive probability that, when the trajectory of the Markov chain leaves a transient state, it will never return to it.

For finite-state Markov Chains, there is graphical way of identifying a transient state: A state i is transient if and only if there is a second state j such that j is accessible from i, but i is not accessible from j. In Figure 11.8, states 1, 2, 3, 4, 5 and 6 are transient states, and they can each access state 7, but cannot be accessed from state 7. Note that, if a state i is transient, every other state k in its communicating class is also transient, because that state k can communicates with state i and therefore can access a state j not in its communicating class.

When a state is not transient, it is called **recurrent**: recurrent states have the property that the expected time to return to the state, given that the Markov Chain starts in that state, is finite. In terms of the random time T defined previously, $\mathbb{E}[T] < \infty$ for recurrent states. In Fig. 11.8, states 7, 8, 9 and 10 are recurrent states. Note that, for finite state Markov Chains, we can label each communicating class as either recurrent or transient.

The meaning of transient states is that, as time grows, the probability of being in a transient state decays to zero. If there is a limiting probability distribution $\underline{\pi}$ and state i is transient, then $\underline{\pi}_i = 0$.

Note the following: If a finite state Markov chain has more than one recurrent communicating class, there will be more than one limiting distribution for $\underline{p}(t)$, and the limit will depend on the initial distribution $\underline{p}(0)$. The matrix P will have more than one eigenvalue equal to 1. This is the case in the Markov Chain in Fig. 11.7, where state 1 is one recurrent communicating class, and states $3, 4$ are the other recurrent communicating class.

When there is only one recurrent communicating class, there is a unique stationary probability distribution $\underline{\pi}$ such that

$$\mathbf{P}^T \underline{\pi} = \underline{\pi} \tag{11.1}$$

Specifically, the matrix P will have a single eigenvalue with value 1. However, this condition is insufficient to guarantee that this stationary probability distribution will be the limit distribution for arbitrary initial probability distributions.

Specifically, consider Fig. 11.7(b). It is easy to verify that all states belong to a single communicating class, which is recurrent. However, we have already established that, starting from the initial condition $X_0 = 1$, the probabilities $\underline{p}(t)$ do not approach a limit! Indeed, they will approach a limit cycle where they will shift among two different limits for odd and even values of n. In this case, there is a second eigenvalue of $\mathbf{P}$ on the unit circle, with value -1.

For a finite state Markov Chain, we define the **period** of state j as the greatest common divisor of the lengths of all the cycles from state j to itself in the graph of the Markov Chain. A more mathematical definition is that the period d is the largest integer d such that $(\mathbf{P}^n)_{jj} = 0$ unless n is divisible by d. A state with period 1 is said to be *aperiodic*.

Note that the *period of all the states in the same communicating class must be the same.* This follows because of the cycles for a state k in this communicating class must consist of states in that communicating class. The proof of this is a bit involved but straightforward from the definition.

A communicating class is periodic with period d if every state has period d greater than 1. There is a simple condition to recognize whether a communicating class is aperiodic: As long as one of the states in the communicating class has a self-loop (e.g. $\mathbf{P}_{ii} > 0$ for some i), the period of that state is 1, and the communicating class must be aperiodic.

We can now give conditions for a finite state Markov Chain to have a unique limiting probability distribution $\underline{\pi}$, which is approached from any initial probability distribution $\underline{p}(0)$. We state this below as a theorem.

Theorem 11.3
Assume that X_t is a finite state homogeneous Markov chain with state transition matrix $\mathbf{P}$. If the Markov chain has a single recurrent communicating class, and the class is *aperiodic*, then there exists a unique limit distribution $\underline{\pi}$.

Note that a Markov chain with transient states can approach a unique limit distribution $\underline{\pi}$ as long as there is only one recurrent, aperiodic communicating class. This limit distribution will have $\pi_k = 0$ for all transient states k.

There is a stronger result for the special case of irreducible Markov chains which have a single communicating class.

Theorem 11.4
Assume that X_t is a finite state homogeneous Markov chain with state transition matrix $\mathbf{P}$. If the Markov chain is *irreducible* and *aperiodic*, then there exists a unique limit distribution $\underline{\pi}$. Furthermore, this limit has the property that $\underline{\pi}_j > 0$ for all states j. Such a Markov Chain is called **ergodic**.

The combination of the irreducible and aperiodic conditions imply that there exists $k > 0$ such that $\mathbf{P}^k > 0$, that is, a matrix with strictly positive entries. In this case, the Perron-Frobenius theorem establishes the existence of a unique eigenvector of $\mathbf{P}^T$ for the eigenvalue 1 with strictly positive elements. The limit distribution $\underline{\pi}$ is this unique positive eigenvector of the matrix $\mathbf{P}^T$ corresponding to the eigenvalue 1, normalized so that its entries that sum up to 1.

11.2.5 Computing stationary probability distributions

An important problem in the analysis of Markov chains is computing the stationary probability distribution $\underline{\pi}$. The algebraic characterization is $\mathbf{P}^T\underline{\pi} = \underline{\pi}$, where $\mathbf{P}$ is the state transition matrix. This can be a cumbersome set of equations to solve. There is another set of equations based on the graphical representation of the Markov chain transitions that can be easier to analyze. A **cut C** of a directed graph is a set of arcs such that, when the arcs are removed from the graph, the graph is divided into two disjoint set of nodes with no arcs between them.

The useful property of cuts is that, given any cut of the Markov chain graph, the probability flow across that cut must equal zero once the system reaches the stationary distribution. A cut C specifies a subset $A \subset \mathcal{R}_X$ and its complement A^c in $\mathcal{R}_X$, and consists of the arcs going from A to A^c, and from A^c to A. Given a distribution $\underline{\pi}$ on the states of the Markov Chain, the net probability flow on a cut C is defined as

$$F(A, A^c) = \sum_{i\in A}\sum_{j\in A^c} \mathbf{P}_{ij}\pi_i - \sum_{j\in A^c}\sum_{i\in A} \mathbf{P}_{ji}\pi_j$$

The main result is that, if $\underline{\pi}$ is a stationary distribution of a Markov chain, then the net probability flow along any cut must be zero! This is summarized in the theorem below:

Theorem 11.5
$\underline{\pi}$ is a stationary distribution of a Markov chain if and only if $\sum_i \pi_i = 1$ and the net probability flow on any cut in the Markov chain graph is zero. That is, for any $A \subset \mathcal{R}_X$, we have

$$\sum_{i\in A}\sum_{j\in A^c} \mathbf{P}_{ij}\pi_i - \sum_{j\in A^c}\sum_{i\in A} \mathbf{P}_{ji}\pi_j = 0$$

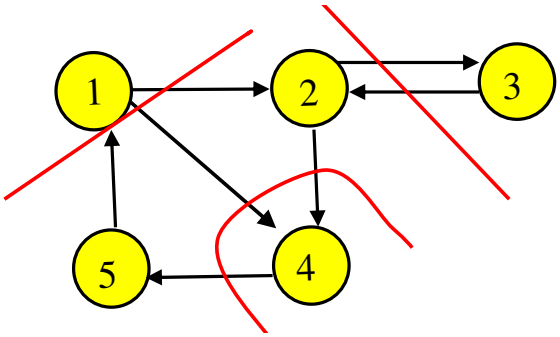


Figure 11.9: Illustration of probability balance

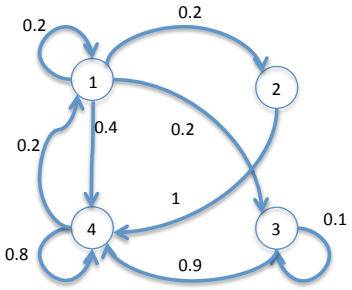


Figure 11.10: Diagram of the Markov Chain for the example

This property is referred to as **probability balance**.

To see that the theorem is equivalent to stationarity, note that if we select $A = \{i\}$, you get exactly the balance equations for the eigenvector:

$$\sum_{j \neq i} \mathbf{P}_{ij}\pi_i = \sum_{j \in \mathcal{R}_X, j \neq i} \mathbf{P}_{ji}\pi_j.$$

If we add $\mathbf{P}_{ii}\pi_i$ to both sides, we have

$$\left(\sum_{j=1}^{K} \mathbf{P}_{ij}\right)\pi_i = \pi_i = \sum_{j=1}^{K} \mathbf{P}_{ji}\pi_j.$$

This is the i-th equation of $\mathbf{P}^T\underline{\pi} = \underline{\pi}$. It is also easy to show the converse, so that starting from balance equations, one can show flow in and out of any group of states is zero for stationary distributions.

Why is this useful? Sometimes, it is easy to identify cuts that yield equations that are simpler than the eigenvector equations. To illustrate how to use probability balance to compute stationary distributions, consider the example in Figure 11.9. The example shows three different cuts, that separate the graph into two disconnected sets of nodes with no arcs across them. Applying fflow balance to each of these cuts yields the equations:

$$\begin{aligned} \mathbf{P}_{14}\pi_1 + \mathbf{P}_{12}\pi_1 - \mathbf{P}_{51}\pi_5 &= 0 \\ \mathbf{P}_{23}\pi_2 - \mathbf{P}_{32}\pi_3 &= 0 \\ \mathbf{P}_{24}\pi_2 + \mathbf{P}_{14}\pi_1 - \mathbf{P}_{45}\pi_4 &= 0. \end{aligned}$$

The above yields three equations in five unknowns, so it is insufficient to find a solution. We can add another cut, isolating state 5, to obtain the following equation: $\mathbf{P}_{51}\pi_5 = \mathbf{P}_{45}\pi_4$. Other cuts are possible, but will be redundant with these equations. Notice that none of those equations include a constant, so the solution $\pi_i = 0$ satisfies the equations. Just like we had to do in the eigenvector method, we must add a normalization equation:

$$\sum_{j=1}^{5} \pi_j = 1.$$

With that as a fifth equation, we now have a unique solution which will yield a positive, normalized $\underline{\pi}$.

Example 11.7
Consider a 4-state discrete time Markov chain, with transition probability matrix described below:

$$\mathbf{P} = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.4 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0.1 & 0.9 \\ 0.2 & 0 & 0 & 0.8 \end{pmatrix}$$

The graph illustrating the transitions of this Markov chain is shown in Fig. 11.10:

Looking at the diagram, it is easy to see that all 4 states are recurrent, as there are directed paths from any one state to any other state. Thus, the chain has a single recurrent communicating class, and thus is irreducible. One can also determine that the Markov chain is aperiodic, because there are some self-loops of length 1. Thus, the Markov chan has a unique steady state distribution, which can be computed as follows: To compute the steady state distribution, we need 4 equations. One of them is:

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

To find 3 others, cut node 2 away from the graph. The flow on that cut yields:

$$0.2\pi_1 = \pi_2$$

Cut node 3 away from the graph, to get:

$$0.2\pi_1 = 0.9\pi_3$$

To get the last equation, we can cut around node 1 to get:

$$0.8\pi_1 = 0.2\pi_4$$

Using the last 3 equations, we get:

$$\pi_2 = \pi_1/5;\ \pi_3 = 2\pi_1/9; \pi_4 = 4\pi_1$$

Substituting into the first equation yields:

$$\pi_1(1 + 1/5 + 2/9 + 4) = 1 \Rightarrow \pi_1 = \frac{45}{244}$$

$$\pi_2 = \frac{9}{244}; \quad \pi_3 = \frac{10}{244}; \quad \pi_r = \frac{180}{244}$$

Example 11.8
We want to model a counter that behaves as follows: The counter has three states: $\mathcal{R}_X = \{1, 2, 3\}$. When the counter is in state 3, it shifts to state 2 at the next time. When it is in state 2, it shifts to state 1 at the next time. When it is in state 1, it shifts to states 1, 2, or 3 at the next time, each with probability $\frac{1}{3}$.

The state transition matrix of this Markov chain is $\mathbf{P} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. The state transition diagram for the Markov chain is show in Figure 11.11.

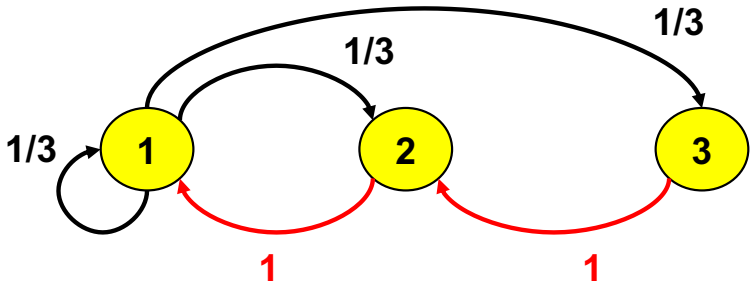


Figure 11.11: Diagram of the Markov chain for Example 11.8

A quick analysis of the graph shows that there is a single recurrent class, and that there are no transient states. Cuts around nodes 3 and 1 plus the normalization equations yields the following equations:

$$\frac{\pi_1}{3} = \pi_3$$

$$\frac{2\pi_1}{3} = \pi_2$$

$$\pi_1 + \pi_2 + \pi_3 = 1$$

Solving this yields the stationary distribution: $\underline{\pi} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{6} \end{bmatrix}$.

Example 11.9
Consider a model of a reflected random walk in a contained space. The state space is $\mathcal{R}_X = \{1, 2, \ldots, 10\}$. At each time t, if the state k is in $\{2, \ldots, 9\}$, the next state is $k+1$ with probability p and $k-1$ with probability $1-p$. If the current state is $k = 1$, then the next state is 1 with probability $1-p$ and 2 with probability p. If the current state is $k = 10$, the next state is 10 with probability p and 9 with probability $1-p$. The diagram of the Markov chain is displayed in Figure 11.12.

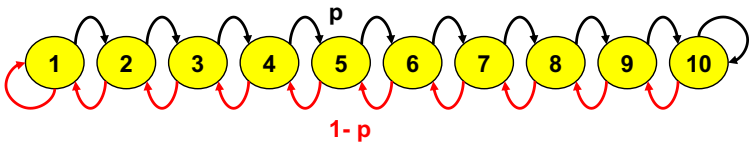


Figure 11.12: Diagram of the Markov chain for Example 11.9.

It is clear that the graph of this Markov chain is irreducible, and the presence of two self-loops makes it aperiodic. Hence, there is a unique stationary distribution. The linear structure of the Markov chain graph makes it easy to find 9 cuts, separating states $k, k+1$, for $k = 1, 2, \ldots, 9$. These cuts yield the following equations:

$$\begin{aligned} &p\pi_1 = (1-p)\pi_2; \quad && p\pi_2 = (1-p)\pi_3; \quad && p\pi_3 = (1-p)\pi_4; \\ &p\pi_4 = (1-p)\pi_5; \quad && p\pi_5 = (1-p)\pi_6; \quad && p\pi_6 = (1-p)\pi_7; \\ &p\pi_7 = (1-p)\pi_8; \quad && p\pi_8 = (1-p)\pi_9; \quad && p\pi_9 = (1-p)\pi_{10}; \end{aligned}$$

Solving, we get the following relationships:

$$\pi_k = \left(\frac{p}{1-p}\right)^{k-1} \pi_1, \ k = 2, \ldots, 10.$$

The tenth equation needed is the normalization equation:

$$\sum_{k=1}^{10} \pi_k = 1 \iff \sum_{k=1}^{10} \left(\frac{p}{1-p}\right)^{k-1} \pi_1 = 1.$$

Fortunately, we can sum this term:

$$\sum_{k=1}^{10} \left(\frac{p}{1-p}\right)^{k-1} = \frac{1 - \left(\frac{p}{1-p}\right)^{10}}{1 - \frac{p}{1-p}}.$$

Hence, $\pi_1 = \frac{1-\frac{p}{1-p}}{1-\left(\frac{p}{1-p}\right)^{10}}$, and $\pi_k = \left(\frac{p}{1-p}\right)^{k-1} \pi_1, \ k = 2, \ldots, 10$. This expression is valid as long as $p \neq 1-p$. Thus, if $p = 0.4$, we obtain $\pi_1 = 0.3392$, and $\pi_{10} = 0.0088$.

If we have symmetry, and $p = 1-p = 0.5$, the balance equations indicate that $\pi_j = \pi_k$ for all $j, k \in 1, \ldots, 10$ so the steady-state distribution is $\pi_k = 0.1$.

Although we have focused our analysis on ergodic Markov chains so far, it is often possible to analyze the limiting behavior of non-ergodic Markov chains. We illustrate this with two different examples.

Example 11.10
Consider the Markov chain with state transition diagram shown in Figure 11.13. The Markov chain has a single recurrent class, but has period 2. The state transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The state transition matrix still has an eigenvalue of 1, and there is a stationary distribution $\underline{\pi}$, which we can find using probability balance, as:

$$\pi_5 = \pi_1; \quad \pi_4 = \pi_5; \quad \pi_3 = 0.5\pi_2; \quad \pi_4 = 0.5\pi_2$$

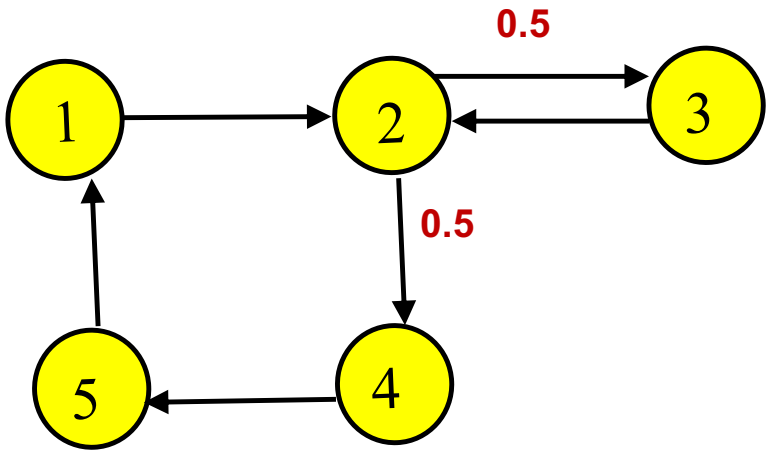


Figure 11.13: Diagram of the Markov chain for Example 11.10.

This means $\pi_1 = \pi_3 = \pi_4 = \pi_5$, and $\pi_2 = 2\pi_1$. Coupled with the normalization constraint, this yields $6\pi_1 = 1$, so $\pi_1 = \pi_3 = \pi_4 = \pi_5 = \frac{1}{6}, \pi_2 = \frac{1}{3}$. If the Markov chain starts with this distribution, it will stay in this distribution.

However, for different initial conditions, the limiting behavior will oscillate between two distributions, depending on the initial condition, and it won't converge to the stationary distribution. For instance, if $X_0 = 1$, the two distributions in the limit are $\begin{bmatrix} 0 \\ \frac{2}{3} \\ 0 \\ 0 \\ \frac{1}{3} \end{bmatrix}$ and $\begin{bmatrix} \frac{1}{3} \\ 0 \\ \frac{1}{3} \\ \frac{1}{3} \\ 0 \end{bmatrix}$.

Example 11.11
Consider the Markov chain with state transition diagram shown in Figure 11.13. The Markov chain has two communicating classes (states 1, 2, 3, and states 4, 5), but it has a single recurrent class (1, 2, 3). The state transition matrix is

$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & p & 0 & 1-p \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Since we know there is no steady state probability in the two transient states (4, 5), we can simply restrict our analysis

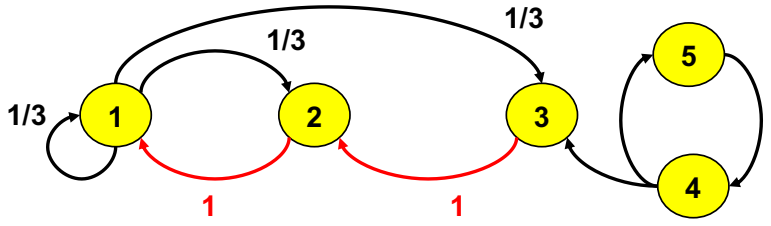


Figure 11.14: Diagram of the Markov chain for Example 11.11.

to the recurrent class, and analyze the steady state behavior of a 3 state model, with transition probability matrix

$$\mathbf{P}_r = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

This is the same Markov chain we analyzed in Example 11.6. Thus, the steady state probability in the original Markov

chain is

$$\underline{\pi} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{6} \\ 0 \\ 0 \end{bmatrix}$$

11.3 Markov chains with infinite state spaces

The above discussion focused on finite state Markov chains, where the state space $\mathcal{R}_X$ has a finite number of states. What changes when the state space is infinite? We can no longer use linear algebra to establish our results, as the transition probability function $\mathbf{P}_{ij}$ does not have a convenient representation as a finite matrix. We highlight some of the key issues and differences below.

Example 11.12
Consider a random walk with probability 0.5 of going forward or back at each time. For this Markov chain, the state space is the space of integers: $\mathcal{R}_X = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. It is easy to see that every state communicates with every other state. This Markov chain has period 2, and has a single communicating class. However, there cannot be an equilibrium distribution: we are no longer guaranteed that there is a positive "eigenvector" with eigenvalue 1. If there were, note that, by symmetry, every equilibrium state should have the same probability. However, since there are an infinite number of states, there is no way to select such a probability to satisfy the normalization condition $\sum_{k \in \mathcal{R}_X} \pi_k = 1$.

Example 11.13
Consider a Markov chain defined on the non-negative numbers as follows: $\mathbf{P}_{00} = 1/2, \mathbf{P}_{01} = 1/2$. For $k > 0$, $\mathbf{P}_{(k-1)k} = \mathbf{P}_{k(k+1)} = 1/2$. All other $\mathbf{P}_{ij} = 0, |i - j| \geq 2$. This chain is aperiodic (state 0 has a self-transition, so it has period 1) and has a single communicating class. However, this chain will not have an equilibrium distribution. Looking at balance equations, cutting between states i and j, we the relation:

$$\pi_k = \pi_{k+1}, k = 0, 1, \ldots$$

Hence, every state would have the same steady state probability, but with an infinite number of states, they would all be zero, a contradiction!

One way of seeing this is to look at the expected time to reach state 0 from state n. As we will show later with our transient analysis, no matter what state you start in, the expected number of steps it takes to reach state 0 is infinite!

The first important difference when the Markov chain has an infinite number of states is in the concept of recurrence. When the state transition graph was irreducible and the state space was finite, we could guarantee that $\mathbf{P}^n_{ij} > 0$ for every pair of states i, j; thus, with probability 1, we would visit state j when we start in state i in finite expected time. When the state space is infinite, this condition of irreducibility is no longer sufficient.

Let X_t be a time-homogeneous Markov chain with transition probability $\mathbf{P}$. Note that the state space $\mathcal{R}_X$ may be infinite. Define the following quantities:

$$T_i = \inf\{t \geq 1 : X_t = i\} = \textbf{ first passage time } \text{for state } \ i$$

When $X_0 = i$, then T_i is the revisit time for state i. We can now define some useful quantities relating how X_t visits a particular state i. Let $\mathcal{I}\{X_t = i\}$ denote the indicator function which is 1 when the event $X_t = i$ is true, and zero otherwise. Then,

$$\begin{aligned} V_i &= \sum_{t=0}^{\infty} \mathcal{I}\{X_t = i\} \ \text{ is the number of visits to state } \ i \\ f_i &= \mathbb{P}[T_i < \infty | X_0 = i] \text{ is the probability that the chain revisits state } i \\ m_i &= \mathbb{E}[T_i | X_0 = i] \text{ is the expected return time to state } i \end{aligned}$$

Consider the case of a finite-state aperiodic Markov chain with a single recurrent communicating class, but with some transient states. Let i be a transient state. Then, V_i is finite, and $f_i < 1$. However, if i is a recurrent state, we get that $V_i = \infty$ with probability 1, $f_i = 1$ and $m_i < \infty$, so that the chain continues to revisit state i. We use these concepts to extend the definition of recurrence to infinite state Markov chains:

Definition 11.1
A state i of a homogeneous Markov chain $\{X_t, t = 0, 1, \ldots\}$ is recurrent if

$$\mathbb{P}[V_i = \infty | X_0 = i] = 1.$$

A recurrent state is one that you return to an infinite number of times. Indeed, we can characterize a recurrent state as one for which $f_i = 1$, and a transient state as one for which $f_i < 1$. When the state space is infinite, we don't have simple graphical characterizations of what recurrent and transient states are. However, we can use the transition probabilities to get equivalent definitions:

Theorem 11.6
State i in a homogeneous Markov chain is recurrent if and only if

$$\sum_{n=0}^{\infty} (\mathbf{P}^n)_{ii} = \infty$$

To show this, note that for recurrent i, one has $\mathbb{P}[V_i = \infty | X_0 = i] = 1$. Note also the following interpretation:

$$(\mathbf{P}^n)_{ii} = \mathbb{P}[X_n = i | X_0 = i]$$

where $\mathbf{P}^n$ is the n-step transition probability kernel $\mathbf{P}(X_n = j | X_0 = i)$, which can be obtained through direct application of the one-step kernel n times. Thus,

$$\sum_{n=0}^{\infty} (\mathbf{P}^n)_{ii} = \sum_{n=0}^{\infty} \mathbb{E}[\mathcal{I}\{X_n = i\} | X_0 = i] = \mathbb{E}[\sum_{n=0}^{\infty} \mathcal{I}\{X_n = i\} | X_0 = i] = E[V_i | X_0 = i] = \infty$$

If i is a transient state, then $f_i < 1$. We can view the return process as a geometric random variable because of the Markov nature of the process X_t. The first return occurs with probability f_i, the second return with probability f_i^2, etc. Thus, the expected number of returns is $\frac{1}{1-f_i}$, which is finite. By the above argument, for transient states i, $\sum_{n=0}^{\infty} (\mathbf{P}^n)_{ii} < \infty$.

We can now use the same definitions we had previously for communicating classes. State i communicates with state j if $(\mathbf{P}^n)_{ij} > 0$ for some $n \geq 1$ and $(\mathbf{P}^m)_{ji} > 0$ for some $m \geq 1$. A communicating class C is a set of states such that, if $i, j \in C$, then i communicates with j. Furthermore, there are no states $k \notin C$ such that a state $j \in C$ communicates with state k.

Theorem 11.7
Let C be a communicating class in the homogeneous Markov chain X_t. Then, either all states in C are recurrent or all states in C are transient.

To see this, take any pair of states $i, j \in C$ and suppose that i is a transient state. Since i, j communicate, there exists $n, m \geq 0$ with $(\mathbf{P}^n)_{ij} > 0, (\mathbf{P}^m)_{ji} > 0$. Then, for any $r \geq 0$,

$$(\mathbf{P}^{n+m+r})_{ii} \geq (\mathbf{P}^n)_{ij} (\mathbf{P}^r)_{jj} (\mathbf{P}^m)_{ji}$$

So,

$$(\mathbf{P}^r)_{jj} \leq \frac{1}{(\mathbf{P}^n)_{ij} (\mathbf{P}^m)_{ji}} (\mathbf{P}^{n+m+r})_{ii}$$

Summing over all $r \geq 0$ yields

$$\sum_{r=0}^{\infty} (\mathbf{P}^r)_{jj} \leq \frac{1}{(\mathbf{P}^n)_{ij} (\mathbf{P}^m)_{ji}} \sum_{r=0}^{\infty} (\mathbf{P}^{n+m+r})_{ii}$$

The last sum is finite since i is transient, so the left hand side is also finite, indicating that j is also transient.

As was the case for finite state Markov chains, every recurrent communicating class will be closed: once a Markov chain enters a state in a recurrent class, the future states in the chain must belong to the same recurrent class. Otherwise, there would be a state i in the recurrent class that communicates with a transient state j (so $(\mathbf{P}^n)_{ij} > 0$ for some $n \geq 1$) but j does not communicate with i. We can thus show that this contradicts $\mathbb{P}[V_i = \infty] = 1$, so that i won't get revisited infinitely.

However, the converse is not true. If we have a closed communicating class, it may not be recurrent! We do have the following result: if a closed communicating class has a finite number of states, it must be recurrent. However, there will be examples of closed communicating classes that won't be recurrent. Examples 11.12 and 11.13 show closed communicating classes that are not recurrent.

Recurrence is the key property for extending our previous results to infinite Markov chains. The implications of recurrence are summarized below:

Theorem 11.8
Suppose $\mathbf{P}$ has a single communicating class C, which is recurrent. Then, for every state $j \in C$, $\mathbb{P}[T_j < \infty] = 1$.

We now focus on the steady state behavior. Does a steady state distribution exist? Can there be more than one? How can one calculate it? We define a couple of useful variables to help understand this behavior. Remember that T_k is the first return time for state k. Let

$$\begin{aligned} V_i^k &= \sum_{n=0}^{T_k} \mathcal{I}\{X_t = i\} = \text{ number of visits to state } i \text{ before visiting state } k. \\ \gamma_i^k &= \mathbb{E}[V_i^k | X_0 = k] \text{ expected number of visits to } i \text{ before revisiting } k \\ V_i(n) &= \sum_{k=0}^{n} \mathcal{I}\{X_t = i\} \text{ number of visits to state } i \text{ before time } n \end{aligned}$$

If there were an invariant distribution $\pi_i, i \in \mathcal{R}_X$, then one would like to show

$$\mathbb{E}[T_i | X_0 = i] = \frac{1}{\pi_i}, \quad \gamma_i^k = \frac{\pi_i}{\pi_k}$$

and

$$\lim_{n\to\infty} \frac{V_i(n)}{n} = \pi_i$$

almost everywhere.

The main result for existence and uniqueness of steady state distributions for general Markov chains requires two items: First, one must have recurrent states. Second, one must have the property that, for a recurrent state, the expected return time is finite. We call a state i *positive recurrent* if it is recurrent and $m_i = \mathbb{E}[T_i | X_0 = i] < \infty$. When a recurrent state has infinite expected return time, we call it *null recurrent*.

Theorem 11.9
Let $\mathbf{P}$ be the state transition kernel of an irreducible Markov chain. Then, the Markov chain has a positive recurrent state i if and only if it has an invariant distribution π. Furthermore, if it has an invariant distribution, then all states are positive recurrent, and $\mathbb{E}[T_i | X_0 = i] = \frac{1}{\pi_i}$ for all states i.

Note that this does not guarantee that all initial distributions approach the invariant distribution π. The problem is that we can still have periodic chains! Here is the final extension that we need:

Theorem 11.10
Let $\mathbf{P}$ be the transition probability kernel of an irreducible, aperiodic, positive recurrent Markov chain (also called ergodic), with invariant distribution π. Then, for any initial distribution, the marginal probabilities converge: $P_{X_n}(j) \to \pi_j$ as $n \to \infty$ for all j. In particular,

$$\lim_{n\to\infty} (\mathbf{P}^n)_{ij} = \pi_j$$

The difficulty in applying this theorem is that computing whether the Markov chain is positive recurrent is equivalent to finding the stationary probability distribution. In practice, we simply try to compute the stationary distribution using the properties of probability balance, and either we can find it, or we find a contradiction that shows such a stationary probability distribution cannot exist.

Computing the stationary distribution of ergodic Markov chains when the state space is infinite can be done using the balance equations $\underline{\pi} = \mathbf{P}^T \underline{\pi}$, where the vector notation is extended to infinite dimensions. This will now require solution of an infinite number of linear equations. The use of cuts is helpful in getting these equations into simple form, as illustrated below.

Example 11.14
Consider a Markov chain defined on the non-negative numbers, which is a model for a single-server queue. The state value represents the number of elements in a queue. The transition probabilities are $\mathbf{P}_{00} = 1-p; \mathbf{P}_{k(k+1)} = p, k = 0, 1, 2, \ldots; \mathbf{P}_{k(k-1)} = 1-p, k = 1, 2, 3, \ldots$. A state transition diagram of this Markov chain is shown in Figure 11.15. It

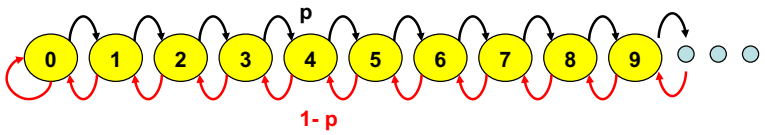


Figure 11.15: Diagram of the Markov chain for Example 11.14.

is clear that there is a single communicating class in this chain, and that the chain is irreducible, as there are no transient states. Furthermore, the chain is aperiodic because of the self-transition present in state 0, that makes the period of the chain equal to 1. Note that, if $p = 1 - p = 0.5$, this is the Markov chain we discussed in Example 11.13.

Assume $p < q$. Since this chain is linear, we can find cuts between any pair of consecutive states. For a cut between states k and $k+1$, probability balance yields the following equation:

$$p\pi_k = (1-p)\pi_{k+1} \iff \pi_{k+1} = \frac{p}{1-p}\pi_k = \left(\frac{p}{1-p}\right)^{k+1} \pi_0, \ k = 0, 1, \ldots$$

Define the utilization factor $\alpha = \frac{p}{1-p}$. Then, we have $\pi_{k+1} = \alpha^{k+1}\pi_0$. Substituting this into the normalization equation yields

$$\sum_{t=0}^{\infty} \alpha^t \pi_0 = 1 \iff \frac{\pi_0}{1-\alpha} = 1 \iff \pi_0 = 1 - \alpha,$$

where we have used the formula for summing a geometric series. Note that this sum exists only for $\alpha < 0$, which means $p < q$.

Thus, the steady state probability distribution is $\pi_k = (1-\alpha)\alpha^k$. This means the Markov chain is positive recurrent when $\alpha < 1$, and is ergodic.

This chain is aperiodic (state 0 has a self-transition, so it has period 1) and has a single communicating class. It is also easy to see that the mean revisit time for state 0 is finite, so the states are positive recurrent, and the chain will be ergodic.

Note that the probability balance equations are the same when $p = q$. However, in this case, we have

$$\pi_k = \pi_0, \ k = 1, 2, \ldots.$$

For this case, the normalization property yields

$$\sum_{k=0}^{\infty} \pi_0 = 1.$$

This equation has no solution, and thus the Markov chain is not positive recurrent and is not ergodic. Similar contradictions can be found for $p > 1 - p$.

Assume $\alpha < 1$. Then, the ergodic distribution is $\pi_k = (1-\alpha)\alpha^k, \ k = 0, 1, \ldots$. Can we compute $\mathbb{E}[X_\infty]$, the expected value of the state of the Markov chain in the limit?

Since X_∞ is a discrete random variable with PMF $P_{X_\infty}(k) = (1-\alpha)\alpha^k$, the expectation is

$$\begin{aligned}\mathbb{E}[X_\infty] &= \sum_{k=0}^{\infty} k(1-\alpha)\alpha^k = \alpha(1-\alpha)\sum_{k=1}^{\infty} k\alpha^{k-1} \\ &= \alpha(1-\alpha)\sum_{k=1}^{\infty}\frac{d}{d\alpha}\alpha^k = \alpha(1-\alpha)\frac{d}{d\alpha}\left(\sum_{k=1}^{\infty}\alpha^k\right) \\ &= \alpha(1-\alpha)\frac{d}{d\alpha}\left(\frac{1}{1-\alpha}-1\right) \\ &= \frac{\alpha}{1-\alpha}\end{aligned}$$

where we have interchanged differentiation and summation because of the convergence of the geometric series when $\alpha < 1$. This implies that, as $\alpha \to 1$, the expected value of the state (the length of the queue) blows up and approaches ∞.

11.4 Ergodicity and the Strong Law of Large Numbers

Markov chains were introduced by Andrey Markov and were named after him. He developed Markov chains to create correlated sequences of random variables, to study extensions of the strong Law of Large Numbers and the Central Limit Theorem for such sequences. In his first paper, in 1906, he proved that, for a Markov chain with positive transition probabilities, the average of the state values along a trajectory converges to the expected value of the limiting distribution (the fixed vector). This was an extension of the weak Law of Large Numbers. In later papers, he proved the Central Limit Theorem for such chains. Subsequently, he established that ergodic Markov chains have properties that generalize the Strong Law of Large Numbers.

Assuming a Markov chain $\{X_t\}$ is ergodic, the marginal distribution $P_{X_t}(x)$ converges to a limit distribution $\underline{\pi}$, where $\pi_i = \mathbb{P}[X_\infty = i]$. Then, for any bounded real-valued function $f : \mathcal{R}_X \to \Re$, we have

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{j=0}^{n} f(X_j) = \sum_{k\in\mathcal{R}_X} f(k)\pi_k = \mathbb{E}[f(X_\infty)]$$

almost surely. If we choose the function $f(k) = 1, f(j) = 0$ if $j \neq k$, we get the following statement:

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{j=0}^{n}\mathcal{I}\{X_j = k\} = \pi_k.$$

Hence, π_k is the fraction of time, on average, that the Markov chain spends in state k. If we choose the function $f(k) = k$, we get exactly the strong Law of Large Numbers, although we have to show this using a limiting argument when the number of states is infinite.

What is the key insight behind Markov's results? The Markov property of Markov chains established that the evolution of the process starting from a particular state k was independent of the past trajectory of the process. If state k was positive recurrent, the trajectories of states visited between visits to state k represented an independent sample of such possible trajectories. Defining as a random variable the sum of the function $f(X_j)$ over the number of states visited starting from state k before the next return to k (including the state k), every revisit provided independent, identically distributed random samples for $f(X_j)$. There is a subtle argument needed to handle the fact that each of those restarts might take different times in returning to state k, but again those random times are identically distributed. The results then follow from the strong Law of Large Numbers.

11.5 Transient Analysis of Markov Chains

Let $\{X_t\}$ be a homogeneous, discrete-time Markov chain with transition probability kernel $\mathbf{P}$, taking values in a discrete state space $\mathcal{R}_X$. Suppose we have a subset of states $A \subset \mathcal{R}_X$. Denote the trajectory of the

Markov chain for a specific outcome as $\{X_t(\omega)\}$. The *first hitting time* of the subset A starting from a state $X_0(s) = i$ is a random variable defined as:

$$H_i^A(\omega) = \inf\{n \geq 0 : X_n(\omega) \in A | X_0(\omega) = i\}.$$

H_i^A is a random variable, although we must allow for the possibility that it takes on an infinite value. Thus, it is a random variable taking values in $\Re \cup \{\infty\}$, a generalization of our earlier definitions. If $H_i^A(\omega) = \infty$, it means the process trajectory, for the experiment outcome ω, never reaches any of the states in A. The probability that the process hits A at all when it starts at state $X_0(\omega) = i$ is given by:

$$h_i^A = \mathbb{P}[\{H_i^A(\omega) < \infty\}].$$

In many problems of interest, we want to compute expected hitting times and hitting probabilities given a particular initial state X_0. Such hitting times can indicate successful completion of events and reaching of milestones. What is surprising is that we will be able to do these computations using simple linear algebra techniques, as described below.

Example 11.15
Let's first consider an example. Suppose we have a four state Markov chain, with transition probability matrix P given by:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note that this system has three communicating classes: 1, 4 and $\{2, 3\}$. However, only 1 and 4 are recurrent classes. Once the state reaches states 1 or 4, the state trajectory stays in those states for all future times.

Suppose we start in state 2. We would like to compute the expected number of steps required to reach states 1 or 4. We can compute this as follows: Let k_i denote the expected time to reach states 1 or 4 starting from from state i. Then, observe the following relationships:

$$k_1 = 0; k_4 = 0$$

What about k_2 and k_3? By the Markov nature of the process, the expected time to reach from state 2 is 1 plus the expected time to reach from whatever the next state is, weighted by the probability of transitioning to that state. In mathematical terms, this yields

$$k_2 = 1 + 0.5k_1 + 0.5k_3; \quad k_3 = 1 + 0.5k_2 + 0.5k_4$$

Basically, any trajectory that starts at i and hits the set $A = \{0, 4\}$ has to take the first step to a state that is connected to i. From that next state, by time invariance, the expected hitting time is the same as that of trajectories that start at that state.

These last two equations are easily solved once we substitute $k_1 = 0, k_4 = 0$ to get $k_2 = k_3 = 2$.

What about a hitting probability? Let the set $A = \{4\}$. Then, reasoning along the same lines, the probability of hitting A from a particular state k is the weightednsum of the probabilities of hitting A from whatever states k transitions, 2, weighted by the transition probabilities. In mathematical terms,

$$h_4^A = 1; \; h_3^A = 0.5h_2^A + 0.5h_4^A; \; h_2^A = 0.5h_1^A + 0.5h_4^A; \; h_1^A = 0$$

Solving these, we get $h_1^A = 0, h_2^A = 1/3, h_3^A = 2/3, h_4^A = 1$.

Can we generalize the insights from this example to arbitrary Markov chains? Let's first focus on Markov chains with fininte state space $\mathcal{R}_X$. The result below characterizes the general solution:

Theorem 11.11
Let $\underline{h}^A$ denote the vector of hitting probabilities for a subset A of the finite state space $\mathcal{R}_X$. Then, $\underline{h}^A$ is the smallest non-negative solution of the following set of linear equations:

$$\begin{cases} h_i^A = 1 & i \in A \\ h_i^A = \sum_j \mathbf{P}_{ij} h_j^A & i \notin A \end{cases}$$

In vector form,

$$\underline{h}^A = \hat{\mathbf{P}}\underline{h}^A; \quad h_i^A = 1, i \in A,$$

where $\hat{\mathbf{P}}$ is a reduced version of matrix $\mathbf{P}$ with the rows corresponding to $i \in A$ deleted. By smallest solution we mean that, if $\underline{x}$ is another non-negative solution, then $x_i \geq h_i^A$.

Note that we have the same number of equations and unknowns, as there is one equation for each $i \notin A$.

Let's prove the above. First, let's show that $\underline{h}^A$ satisfies the equations. Assume $x(0) = i \in A$. Then, the hitting time $H_i^A = 0$, and the hitting probability $h_i^A = 1$, which the theorem guarantees by construction. Now, assume that $x(0) = i \notin A$. Then, $H_i^A > 0$, as it will take at least one step to reach a state in A. By the Markov property of the process,

$$h_i^A = \mathbb{P}[H_i^A < \infty | X_0 = i] = \sum_j \mathbb{P}[H_i^A < \infty, X_1 = j | X_0 = i] = \sum_j \mathbb{P}[H_i^A < \infty | X_1 = j]\mathbf{P}_i j = \sum_j h_j^A P_{ij}$$

which shows that $\underline{h}^A$ satisfies theorem 11.11.

Now, suppose we have a non-negative solution $\underline{g}$ to the equations in theorem 11.11. We want to show that these must be greater than or equal to the expected hitting times. We know that $h_i^A = g_i$ for $i \in A$, as they are set to 1. Suppose $i \notin A$. Then,

$$g_i = \sum_j \mathbf{P}_{ij} g_j = \sum_{j \in A} \mathbf{P}_{ij} g_j + \sum_{j \notin A} \mathbf{P}_{ij} g_j = \sum_{j \in A} \mathbf{P}_{ij} + \sum_{j \notin A} \mathbf{P}_{ij} g_j$$

Now, substitute for g_j in the last term, to get:

$$g_i = \sum_{j \in A} \mathbf{P}_{ij} + \sum_{j \notin A} \mathbf{P}_{ij} (\sum_{k \in A} \mathbf{P}_{jk} + \sum_{k \notin A} \mathbf{P}_{jk} g_k) = \mathbb{P}[X_1 \in A] + \mathbb{P}[X_1 \notin A, X_2 \in A] + \sum_{j,k \notin A} \mathbf{P}_{ij}\mathbf{P}_{jk} g_k.$$

By repeated substitution, we get

$$\begin{aligned} g_i = \mathbb{P}[X_1 \in A] + \mathbb{P}[X_1 \notin A, X_2 \in A] + \mathbb{P}[X_1, X_2 \notin A, X_3 \in A] + \ldots + \\ \mathbb{P}[X_1, \ldots, X_{n-1} \notin A, X_n) \in A] + \sum_{j_1, \ldots, j_n \notin A} \mathbf{P}_{ij_1}\mathbf{P}_{j_1 j_2} \cdots \mathbf{P}_{j_{n-1} j_n} g_{j_n} \end{aligned}$$

Note that the sum of all but the last term are $\mathbb{P}[H_i^A \leq n]$. Thus, $g_i \geq \mathbb{P}[H_i^A \leq n]$ for any n, because the last term is non-negative ($g_k \geq 0$ for all k). Thus,

$$g_i \geq \lim_{n \to \infty} \mathbb{P}[H_i^A \leq n] = \mathbb{P}[H_i^A < \infty] = h_i$$

which shows that $\underline{h}$ is the smallest nonnegative solution.

Can there be multiple solutions? Consider the Markov chain with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix}$$

and let $A = \{1\}$. Clearly, $h_1^A = 1$, and $h_2^A = h_3^A = 0$, because starting at either state 2 or 3, one cannot reach state 1 at all. Note, however that the equations in theorem 11.11 can be solved by any vector $\underline{g} = (1, k, k)^T$. Of course, the smallest nonnegative solution among this is $(1, 0, 0)^T$, which are the hitting times.

The above theorem is true even if the state space $\mathcal{S}$ is infinite. However, we now have an infinite number of equations to consider, which makes numerical computation harder.

Example 11.16

Consider a random walk on $\{0, 1, 2, \ldots\}$, where $\mathbf{P}_{00} = 1$, and $\mathbf{P}_{i(i+1)} = \mathbf{P}_{i(i-1)} = 1/2$ for $i \geq 1$, $\mathbf{P}_{ij} = 0, |i - j| \geq 2$. This corresponds to an infinite gambler's ruin problem where the gambler never leaves until he is broke. We would like to

compute the hitting probability for the set $A = \{0\}$, corresponding to the gambler leaving broke. Here are the relevant equations for the hitting probability h_i^A:

$$\begin{aligned} h_0^A &= 1 \\ h_1^A &= 0.5h_0^A + 0.5h_2^A \\ &\vdots \\ h_n^A &= 0.5h_{n-1}^A + 0.5h_{n+1}^A \qquad \vdots \end{aligned}$$

We can solve this via z-transforms, as follows: the characteristic equation of the recursion is

$$0.5z^2 - z + 0.5 = 0$$

. By inspection, this has a repeated root at $z = 0$. Thus, this admits solutions of the form $h_n^A = C + Dn$ for some constants C, D. To match the initial condition $h_0^A = 1$, we get $C = 1$. The second equation yields $C + D = 0.5 + 0.5(C + 2D)$, which is true for all D. Thus, any value of $D \geq 0$ will yield a valid nonnegative solution! However, h_n^A is a probability, and as such, it must be less than 1. Indeed, the only solution that will yield a probability is $D = 0$, so $h_n^A = 1$ for all n! This means that you will always go broke, no matter where you start!

What if we change the problem so that $P_{i(i+1)} = 3/4, P_{i(i-1)} = 1/4$? This is a very nice game, with odds in the players' favor. In this case, the main recursion yields

$$h_n^A = 0.25h_{n-1}^A + 0.75h_{n+1}^A$$

with characteristic equation $1 - 4z + 3z^2 = 0$, which yields solution of the form $h_n^A = C + D(1/3)^n$. To fit the initial condition $h_0^A = 1$, we have $C + D = 1$, or $D = 1 - C$. Thus, the general form of the solution is

$$h_n^A = (1 - C)(\frac{1}{3})^n + C = (\frac{1}{3})^n + C(1 - (\frac{1}{3})^n)$$

Note that, for any $C \geq 0$, this remains nonnegative.

Thus, we don't have an easy way to select C. Here is where the choice of smallest non-negative solution gives an answer: the smallest non-negative solution is given by $C = 0$, which is $h_n^A = (\frac{1}{3})^n$. In this case, the probability of going broke decreases exponentially with increasing initial condition.

Theorem 11.11 deals with hitting probabilities. We can develop a similar result for hitting times.

Theorem 11.12
Let $\underline{k}^A$ denote the vector of expected hitting times for a subset A of the state space $\mathcal{R}_X$, where these values could be infinite. Then, $\underline{k}^A$ is the smallest non-negative solution of the following set of linear equations:

$$\begin{cases} k_i^A = 0 & i \in A \\ k_i^A = 1 + \sum_{j \in \mathcal{R}_X} \mathbf{P}_{ij} k_j^A & i \notin A \end{cases}$$

In vector form, $\underline{k}^A = \underline{1} + \hat{\mathbf{P}}\underline{k}^A; k_i^A = 0, i \in A$, where $\hat{\mathbf{P}}$ is the state transition matrix $\mathbf{P}$ with the rows for $i \in A$ removed.

To show this, we proceed as before. We show that $\underline{k}^A$ satisfies the equations in theorem 11.12. If $X_0 = i \in A$, then $H_i^A = 0$, so $k_i^A = 0$. If $X_0 = i \notin A$, then $H_i^A \geq 1$. By the Markov property, when $i \notin A$,

$$\mathbb{P}[H_i^A = n | X_0 = i] = \sum_{j \in \mathcal{R}_X} \mathbb{P}[H_i^A = n, X_1 = j | X_0 = i] = \sum_{j \in \mathcal{R}_X} \mathbb{P}[H_i^A = n | X_1 = j]\mathbf{P}_{ij}$$

Thus,

$$\begin{aligned} k_i^A &= \sum_{n=1}^{\infty} n\mathbb{P}[H_i^A = n] + \infty\mathbb{P}[H_i^A = \infty] = \sum_{n=1}^{\infty} \mathbb{P}[H_i^A \geq n] \\ &= \sum_{n=1}^{\infty} \sum_{j\in\mathcal{S}} \mathbb{P}[H_i^A \geq n, X_1 = j) = \sum_{n=1}^{\infty} \sum_{j\in\mathcal{R}_X} \mathbb{P}[H_i^A \geq n | X_1 = j]\mathbf{P}_{ij} \\ &= \sum_{j\in\mathcal{R}_X} \mathbf{P}_{ij} \sum_{n=1}^{\infty} \mathbb{P}[H_i^A \geq n | x(1) = j] \\ &= \sum_{j\in\mathcal{R}_X} \mathbf{P}_{ij}(1 + \mathbb{E}[H_j^A]) = 1 + \sum_{j\in\mathcal{R}_X} \mathbf{P}_{ij} k_j^A \end{aligned}$$

which shows that the expected hitting times satisfy the equations of theorem 11.12, even when they have infinite value!

Let g be any solution of the linear equations in the Theorem. Then, $g_i = k_i^A = 0$ for $i \in A$. Suppose $i \notin A$. then,

$$\begin{aligned} g_i &= 1 + \sum_{j\notin A} \mathbf{P}_{ij} g_j \\ &= 1 + \sum_{j\notin A} \mathbf{P}_{ij} \Big(1 + \sum_{k\notin A} \mathbf{P}_{jk} g_k\Big) \\ &= \mathbb{P}[H_i^A \geq 1] + \mathbb{P}[H_i^A \geq 2] + \sum_{j,k\notin A} \mathbf{P}_{ij}\mathbf{P}_{jk} g_k \end{aligned}$$

Continuing the substitutions, we get

$$g_i = \mathbb{P}[H_i^A \geq 1] + \mathbb{P}[H_i^A \geq 2] + \cdots + \mathbb{P}[H_i^A \geq n] + \sum_{j_1,\ldots,j_n\notin A} \mathbf{P}_{ij_1}\mathbf{P}_{j_1 j_2}\cdots\mathbf{P}_{j_{n-1}j_n} g_{j_n}$$

Noting that $g_j \geq 0$, we have

$$g_i \geq \lim_{n\to\infty} \big(\mathbb{P}[H_i^A \geq 1] + \mathbb{P}[H_i^A \geq 2] + \cdots + \mathbb{P}[H_i^A \geq n]\big) = \mathbb{E}[H_i^A] = k_i^A$$

which shows that $\underline{k}^A$ is the smallest nonnegative solution.

Example 11.17

Consider the previous example 11.16, where we set $\mathbf{P}_{i(i+1)} = 1/4, \mathbf{P}_{i(i-1)} = 3/4$. Note that, in average, we are headed towards 0. We want to compute the expected time to reach state 0 from any state n. The relevant equations from theorem 11.12 are:

$$\begin{aligned} k_0^0 &= 0; \\ k_1^0 &= 1 + 0.75k_0^0 + 0.25k_2^0 \\ &\vdots \\ k_n^0 &= 1 + 0.75k_{n-1}^0 + 0.25k_{n+1}^0 \\ &\vdots \end{aligned}$$

Note that this set of linear equations has an input which is a constant on the right hand side, corresponding to a pole at $z = 1$. Furthermore, the characteristic equation for this system is $(z-1)(z-3) = 0$, so the pole at $z = 1$ is repeated. This means the solution is of the form

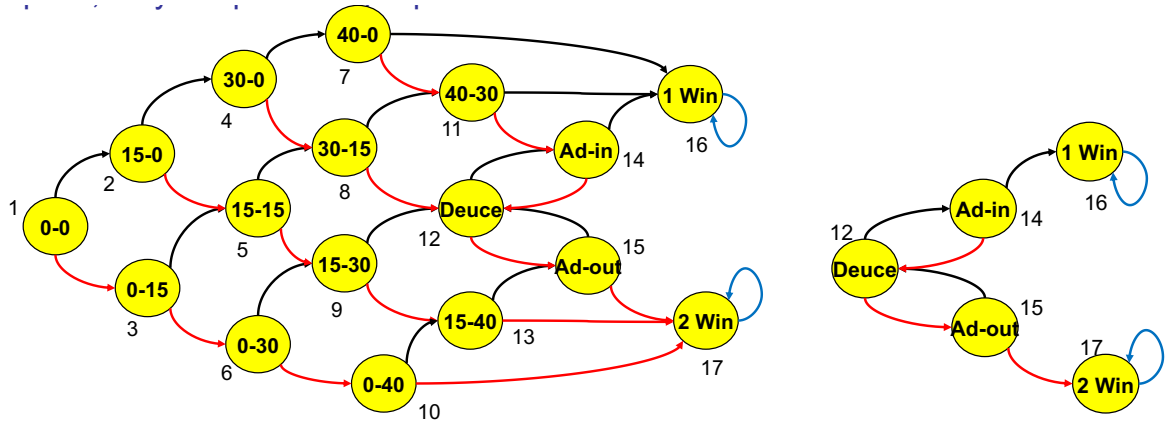
$$k_n^0 = Kn + A + B3^n.$$

Substituting into the above equations yields $K = 2$. The initial condition $k_0^0 = 0$ means $A = -B$. Note that $B \geq 0$ is required for the solution to stay non-negative. The smallest non-negative solution is $B = 0$, which yields $k_n^0 = 2n$.

Example 11.18
Here is a much more complex example. Consider a Tennis game, where the server's probability of winning a point is p, and the receiver's probability of winning a point is $1-p$. We assume that each point is an independent event, and that the probability of winning the point by Player 1 is the same no matter what the score. We can view the evolution of the score of the game as a Markov chain, which eventually ends in either Player 1, the server, winning the game, or Player 2, the receiver, winning the game. A state transition diagram using 17 states is show in Figure 11.16(a), where the tennis score is shown in the circle, and the state number is outside. The red transitions indicate points won by Player 2, and the black transitions indicate points won by Player 1. We have shortened the states somewhat by matching the 30-30 score and Deuce into the same state node, requiring two consecutive points to win by any player.

Analysis of this Markov chain shows that there are only two recurrent communicating classes: state 17 where Player 1 wins, and state

Can we compute the expected duration of the game as a function of p, the probability that Player 1 wins a point? At first, that seems like a daunting task given the size of the network. However, we can solve for this in stages. Conditioned on starting in the Deuce state, corresponding to state number 12, what is the expected number of games? We can solve this by analyzing the much smaller chain in Figure 11.16(b). Indeed, the expected exit time equations for the exit states $16, 17$ are:



(a) The full Tennis Markov chain. (b) The chain starting at Deuce.

Figure 11.16: Diagram of the Markov chain for Example 11.18.

$$k_{12} = 1 + pk_{14} + (1-p)k_{15}; \quad k_{14} = 1 + (1-p)k_{12}; \quad k_{15} = 1 + pk_{12};$$

Substituting the last two equations into the first one yields the solution:

$$\begin{aligned} k_{12} &= 1 + p(1 + (1-p)k_{12}) + (1-p)(1 + pk_{12}) \\ &= 2 + 2p(1-p)k_{12} \\ \Rightarrow k_{12} &= \frac{2}{1-2p(1-p)}, \; k_{14} = \frac{1+2(1-p)^2}{1-2p(1-p)}, \; k_{15} = \frac{1+2p^2}{1-2p(1-p)} \end{aligned}$$

Let's ask a second question: what is the probability that Player 1 wins, given we have reached Deuce? This is an exit probability question on the same Markov chain, where we want the probability that the Markov chain will reach state 16. The relevant equations are:

$$\begin{aligned} &h_{17} = 0; \; h_{16} = 1; \; h_{14} = ph_{16} + (1-p)h_{12} \\ &h_{12} = ph_{14} + (1-p)h_{15}; \; h_{15} = ph_{12} + (1-p)h_{17} \end{aligned}$$

Solving these yields the following:

$$h_{12} = p(p + (1-p)h_{12}) + (1-p)ph_{12} \Rightarrow h_{12} = \frac{p^2}{1-2p(1-p)}.$$

Now that we have solved for these, note that we can compute the exit times for any of the other states in the full Markov chain by back substitution! For instance, the following equations propagate the solution two layers back.

$$k_{11} = 1 + (1-p)k_{14}; \qquad k_{13} = 1 + pk_{15}$$
$$k_7 = 1 + (1-p)k_{11};\ k_8 = 1 + pk_{11} + (1-p)k_{12};$$
$$k_9 = 1 + pk_{12} + (1-p)k_{13};\ k_{10} = 1 + pk_{13}$$

It is straightforward to write the remaining equations, until we compute $k_1 = pk_2 + (1-p)k_3$, yielding the expected number of games to play.

11.6 Applications

In this section, we discuss two popular applications of the theory of Markov chains.

11.6.1 Google PageRank algorithm

Larry Page and Sergey Brin developed PageRank at Stanford University in 1996 as part of a research project about a new kind of search engine. Sergey Brin had the idea that information on the web could be ordered in a hierarchy by "link popularity": a page ranks higher as there are more links to it. Shortly after, Page and Brin founded Google Inc., the company behind the Google search engine.

The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. Google recalculates PageRank scores each time it crawls the Web and rebuilds its index. The formula uses a model of a random surfer who reaches their target site after several clicks, then switches to a random page. The PageRank value of a page reflects the chance that the random surfer will land on that page by clicking on a link.

The PageRank algorithm can best be modeled as a Markov chain in which the states are pages. Let j denote the state corresponding to a random surfer being in page k. The probability of transitioning to another page k is zero, unless there is a link on page j to page k. Then, the probability is uniform among the number of outgoing links to different pages out of j:

$$\mathbf{P}_{jk} = \begin{cases} 0 & \text{if there is no link to page } k \text{ on page } j, \\ \frac{1}{\text{Number distinct page links on page } j} & \text{elsewhere.} \end{cases}$$

What types of Markov chain does such a construction yield? First of all, it is a large one, with nearly a billion states. Second, the Markov chain is sparse, so that the number of transitions out of every row is a very small fraction of the number of nodes: hence, it is ideally viewed in terms of a graph. However, it is unclear that the resulting chain is irreducible. If a page has no links to other pages, it becomes a sink and there are no transitions out of it. Hence, there can be many transient states that have transitions that lead to such sinks. Thus, the resulting Markov chain is not ergodic.

One idea for making it ergodic is to allow the chain to transition to a random page, uniformly over all pages, when it reaches a page with no links. That would guarantee that the Markov chain would be not have any absorbing states, and that it would be aperiodic, as self-transitions would be possible, and it would even guarantee irreducibility. However, this would yield a hard Markov chain to analyze, as it would lose all the sparsity that was present in the original chain.

What Google's founders did was simpler and more clever: In addition to having probability of transitioning to any of the outgoing links in a page, they added a probability that they would transition from any page to any other page, uniformly. That is, let N be the total number of pages. Let $\alpha \in (0,1)$ be a relaxation factor. Then, the new transition probability was

$$\mathbf{P}_{ij}^{new} = \alpha * \mathbf{P}_{ij} + \frac{1-\alpha}{N}$$

Note that this guarantees that

$$\sum_{j=1}^{N} \mathbf{P}_{ij}^{new} = \alpha \sum_{j=1}^{N} \mathbf{P}_{ij} + \sum_{j=1}^{N} \frac{1-\alpha}{N} = \alpha + 1 - \alpha = 1.$$

Since $\mathbf{P}_{ij}^{new} \in (0,1)$, then $\mathbf{P}_{ij}^{new}$ is a valid stochastic matrix, and a state transition matrix for the new Markov chain. Furthermore, since $\mathbf{P}_{ij}^{new} > 0$, this chain is ergodic.

Google PageRank computes the stationary distribution of this Markov chain $\underline{\pi}$, and ranks pages in order of decreasing π_i. In principle, π_i is equal to the fraction of time that a random web surfer would spend on particular pages. However, solving for the eigenvector of a matrix of size $10^9 \times 10^9$ seems like a daunting task.

In this regard, the idea of adding the uniform transition probability makes this computation easier. Specifically, we can start with $\underline{p}(0) = \begin{bmatrix} \frac{1}{N} \\ \vdots \\ \frac{1}{N} \end{bmatrix}$. Then, we can compute

$$\underline{p}(t+1) = (\mathbf{P}^{new})^T \underline{p}(t).$$

In coordinates, this update is

$$\pi_j(t+1) = \sum_{i=1}^{N} \alpha \mathbf{P}_{ij} \pi_i(t) + \frac{1-\alpha)}{N} \pi_j(t)$$

Note that this is a very sparse update, so that computing an update iteration is of order $O(N)$, linear in the number of nodes. However, how many iterations are required? The rate of convergence of the iteration to steady state depends on the magnitude of the second largest eigenvalue of $\mathbf{P}^{new}$. Fortunately, that magnitude is no larger than α, so by selecting α, one can control the number of iterations. In practice, α is selected to be around 0.85, and the number of iterations required to converge is around 60.

11.6.2 Consensus Algorithms

Consider the following situation: a group of persons in a room generate estimates of a quantity X. Each person generates an estimate X_i. Each person shares their estimate with their immediate neighbors; each person then revised their estimate using a weighted linear combination of their own estimate and the estimate of their neighbors. Following this, another round of communication and averaging takes place. If we repeat this for many rounds, will ever person's estimate converge to the same estimate? Furthermore, if they converge, what estimate will they converge to?

While this problem seems a bit artificial in its description, the problem is at the heart of many applications: distributed training of deep neural networks where each agent only has part of the training data, formation flight of aircraft or birds, distributed control of robots, and similar problems.

Let's formulate this as a Markov chain problem. Assume there are K persons, and each person is represented by a node i. We assume that person i has n_i neighbors, denoted by a set N_i. For every node i and node $j \in N_i$, we assume there is an arc from i to j, and an arc from j to i. We assume the graph is connected, so that there is a path between every pair of nodes.

Let's define the update algorithm for node i. Denote by $X_i(n)$ the estimate of person i after the $n-th$ round of exchanges is complete. $X_i(0)$ is the initial estimate. Then,

$$X_i(n+1) = a_i X_i(n) + \frac{1-a_i}{n_i} \sum_{j \in N_i} X_j(n),$$

where $a_i \in (0,1)$. Each person i can have their own weight for their own estimate relative to that of their neighbors.

Writing this as a vector recursion, this is stated as:

$$\underline{X}(n+1) = \mathbf{P}\underline{X}(n)$$

where $\mathbf{P}_{ij} \in [0,1], \sum_{j=1}^{K} \mathbf{P}_{ij} = 1$. Hence, $\mathbf{P}$ is a stochastic matrix, and thus is the state transition matrix of a Markov chain. Furthermore, since we assume there is a path in the graph between every pair of nodes i, j, the Markov chain is irreducible, and it is aperiodic because there are self-loops, since $a_i > 0$. Thus, the Markov chain is ergodic, and $\mathbf{P}$ has a unique eigenvector corresponding to the eigenvalue 1. Indeed, since all the rows sum up to 1, we know that the eigenvector corresponding to the eigenvalue 1 is the vector of all ones.

This means that the estimates will converge:

$$\lim_{n\to\infty} \underline{X}(n+1) = C \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

for some constant C. Note that the convergence to consensus will happen independent of the numerical choices used to average the neighbors' estimates. Convergence is inevitable because of the ergodicity of the underlying Markov chain.

However, what will be the limit of the estimates that the persons converge to? That depends on the averaging parameters we choose. Denote the stationary distribution of the ergodic Markov chain with state transition matrix $\mathbf{P}$ as $\underline{\pi}$. Then, the consensus algorithm will converge to $\underline{\pi}^t \underline{X}(0)$, the average of the initial estimates using the stationary probability distribution of the Markov chain.

To establish this, define $\underline{1}$ to be the K-dimensional vector of all ones. Then, since $\underline{X}(n)$ converges to the consensus value $C\underline{1}$, we have that

$$\lim_{n\to\infty} \frac{1}{K}\underline{1}^T \underline{X}(n) = \lim_{n\to\infty} \frac{1}{K}\underline{1}^T \mathbf{P}^n \underline{X}(0) = C$$

Note also that, since the Markov chain is ergodic and $\frac{1}{N}\underline{1}$ is a probability distribution,

$$l \lim_{n\to\infty} \frac{1}{K}\underline{1}^T \mathbf{P}^n = \underline{\pi}^T.$$

This establishes that the average value $C = \underline{\pi}^T \underline{X}(0)$.

Appendix A

Summary of Linear Algebra

Linear algebra is concerned with the solution of sets of simultaneous systems of linear equations. The linear nature of these sets of equations leads naturally to both a convenient notation and a deep connections with the properties of vectors and matrices. These notes are intended to provide a summary and review of the important concepts and notation that arise.

A.1 Vectors

A **column vector** of dimension n (which we will often simply call a vector) is a vertical array of n numbers,

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} . \tag{A.1}$$

where $x_1, x_2, \ldots, x_n$ are real numbers. We often denote such column vectors in lowercase letters with vector notation, as shown in (A.1). The set of all n-dimensional vectors of real numbers is usually denoted by $\Re^n$.

For n-dimensional vectors, we define some elementary operations as follows:

- **Transpose:** The *transpose* of a column vector $\underline{x}$ is a row vector: $\underline{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ .. Note that the transpose of a row vector is a column vector, and vice versa.
- **Addition:** The sum of two n-dimensional vectors $\underline{x}$ and $\underline{y}$ is defined on a component-by-component basis as
$$\underline{x} + \underline{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}$$
- **Subtraction** The difference of two n-dimensional vectors $\underline{x}$ and $\underline{y}$ is defined on a component-by-component basis as
$$\underline{x} - \underline{y} = \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \vdots \\ x_n - y_n \end{bmatrix}$$
- **Scalar multiplication** The product of a vector $\underline{x}$ and a real number (a scalar) α is a vector $\underline{y}$ defined component wise as:
$$\underline{y} = \alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix} .$$

- **Inner product or Dot product** Given two vectors of the same length, $\underline{x}$, $\underline{y} \in \Re^n$, we can define the *dot* or *inner product* between the vectors as $\underline{x}^T\underline{y} = \sum_{i=1}^n x_i y_i$. Two vectors $\underline{x}$ and $\underline{y}$ are termed *orthogonal*, denoted $\underline{x} \perp \underline{y}$ if $\underline{x}^T\underline{y} = 0$.
- **Euclidean Norm:** The Euclidean norm of a vector $\underline{x}$ is defined as $\| \underline{x} \| = \sqrt{\underline{x}^T\underline{x}} = \sqrt{\sum_{i=1}^n x_i^2}$. This is the length of a vector.
- **Angle between vectors** The inner product provides information about the angle $\angle(\underline{x}, \underline{y})$ between two vectors $\underline{x}$ and $\underline{y}$. Specifically, $\underline{x}^T\underline{y} = ||\underline{x}||\,||\underline{y}|| \cos(\angle(\underline{x}, \underline{y}))$.
- **Zero vector** We define the zero vector in $\Re^n$ as the vector $\underline{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

A.1.1 Linear Independence

A set of vectors $\{\underline{x}_1, \ldots, \underline{x}_m\}$ in $\Re^n$ is termed **linearly dependent** if there exists scalars $\alpha_1, \alpha_2, \ldots, \alpha_m$, not all zero, such that

$$\alpha_1\underline{x}_1 + \alpha_2\underline{x}_2 + \cdots + \alpha_m\underline{x}_m = \underline{0} \; .$$

If there are no such scalars (except $\alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$), then we say the vectors are **linearly independent**.

If a set of vectors $\{\underline{x}_1, \ldots, \underline{x}_m\}$ are linearly dependent, then one of the vectors $\underline{x}_i$ can be written as a *linear combination* of the others.

Note the following facts:

- In $\Re^n$ we can have at most n linearly independent vectors in any given collection of n-dimensional vectors.
- Given any set of n linearly independent vectors $\{\underline{x}_1, \ldots, \underline{x}_n\}$ in $\Re^n$, any other vector can be written as a linear combination of the vectors $\underline{x}_1, \ldots, \underline{x}_n$. Any such set of n linearly independent vectors is termed a *basis* for $\Re^n$.
- An orthogonal basis $\{\underline{x}_1, \ldots, \underline{x}_n\}$ in $\Re^n$ is a basis where every two vectors are mutually orthogonal; that is, $\underline{x}_j^T\underline{x}_j = 0$ if $j \neq k$.
- An **orthonormal basis** is an orthogonal basis where every vector has norm or length 1. That is, $||\underline{x}_k|| = 1$ for all $k \in \{1, \ldots, n\}$.

A.2 Matrices

As in the case of vectors, *matrices* are simply arrays of real numbers in a regular grid. An $m \times n$ **matrix** is a rectangular array of numbers with m rows and n columns, defined as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

where $a_{11}, \ldots, a_{mn}$ are real numbers. We refer to $a_{k\ell}$ as the number in the k-th row, ℓ-th column. We will also refer to the number in the k-th row, ℓ-th column of a matrix A as $[\mathbf{A}]_{ij}$, or $(\mathbf{A})_{ij}$, depending on the situation.

The set of all $m \times n$ real-valued matrices is denoted $\Re^{m \times n}$. If $m = n$, we call the matrix $\mathbf{A}$ a **square** matrix.

Note that we can view vectors as special types of matrices. A column vector of dimension n can be viewed as an $n \times 1$ matrix, whereas a row vector of dimension n can be viewed as a $1 \times n$ matrix.

A.2.1 Matrix Operations

- **Transpose:** The transpose of an $m \times n$ matrix $\mathbf{A}$ is an $n \times m$ matrix deneoted by

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ & & & \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}.$$

- **Symmetric matrices:** A square matrix $\mathbf{A}$ is said to be **symmetric** if $\mathbf{A}^T = \mathbf{A}$.

- **Diagonal matrices:** A *diagonal* matrix is a square matrix that only has nonzero entries along its diagonal, and is of the form

$$\mathbf{A} = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_n \end{bmatrix}$$

  This is sometimes denoted as $\mathsf{diag}(a_1, \ldots, a_n)$.

- **Identity Matrix:** The *identity matrix* in n dimensions is a square $n \times n$ matrix, denoted by $\mathbf{I}_n$ and is the diagonal matrix with ones along its diagonal: $\mathbf{I}_n = \mathsf{diag}(1, ..., 1)$..

- **Matrix addition** Two matrices $\mathbf{A}, \mathbf{B}$ of the same dimensions $m \times n$ can be added, to obtain a new $m \times n$ matrix

$$\mathbf{A} + \mathbf{B} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix}.$$

- **Matrix Subtraction** Two matrices $\mathbf{A}, \mathbf{B}$ of the same dimensions $m \times n$ can be subtracted, to obtain a new $m \times n$ matrix

$$\mathbf{A} - \mathbf{B} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} & \cdots & a_{1n} - b_{1n} \\ a_{21} - b_{21} & a_{22} - b_{22} & \cdots & a_{2n} - b_{2n} \\ & & & \\ a_{m1} - b_{m1} & a_{m2} - b_{m2} & \cdots & a_{mn} - b_{mn} \end{bmatrix}.$$

- **Scalar Multiplication:** For any scalar α, $\alpha\mathbf{A} = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \cdots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \cdots & \alpha a_{2n} \\ & & & \\ \alpha a_{m1} & \alpha a_{m2} & \cdots & \alpha a_{mn} \end{bmatrix}$.

- **Matrix Multiplication:** Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ an $n \times p$ matrix. Then the *matrix product* of $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{C} = \mathbf{AB}$ where $\mathbf{C}$ is an $m \times p$ matrix whose elements are given by

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ & & & \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{bmatrix} \qquad \text{where } c_{k\ell} = \sum_{i=1}^{n} a_{ki} b_{i\ell}\ .$$

Note that $\mathbf{A}$ and $\mathbf{B}$ must satisfy some dimensional constraints for the above expression to make any sense. In particular, the number of columns of $\mathbf{A}$ must equal the number of rows of $\mathbf{B}$; otherwise, $\mathbf{AB}$ is undefined.

- **Properties of matrix multiplication:**
  - For $\mathbf{A} \in \Re^{m \times n}, \mathbf{I}_m\mathbf{A} = \mathbf{A}$ and $\mathbf{A}\mathbf{I}_n = \mathbf{A}$.
  - For $\mathbf{A} \in \Re^{m \times n}, \mathbf{B} \in \Re^{n \times p}$, $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.
  - For square matrices $\mathbf{A}, \mathbf{B} \in \Re^{n \times n}$, $\mathbf{AB}$ may not be equal to $\mathbf{BA}$, so that matrix multiplication may not be commutative in general.
- **Matrix-Vector Multiplication:** This is an important special case of matrix multiplication where one of the matrices is a vector. If $\mathbf{A} \in \Re^{m \times n}$ and $\underline{x}$ is an n-dimensional vector, we treat $\underline{x}$ as an $n \times 1$ matrix, and define the matrix-vector product as

$$\mathbf{A}\underline{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n a_{1i}x_i \\ \sum_{i=1}^n a_{2i}x_i \\ \vdots \\ \sum_{i=1}^n a_{mi}x_i \end{bmatrix}$$

It is sometimes useful to view this as summing the n columns of $\mathbf{A}$ with weights given by x_i:

$$\mathbf{A}\underline{x} = \sum_{i=1}^n x_i\underline{a}_i \quad \text{where } \underline{a}_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{mi} \end{bmatrix}$$

Similarly, for an m-dimensional vector $\underline{y}$, we can view $\underline{y}^T$ as a $1 \times m$ dimensional matrix, and define

$$\underline{y}^T\mathbf{A} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m y_i a_{i1} & \sum_{i=1}^m y_i a_{i2} & \cdots & \sum_{i=1}^m y_i a_{in} \end{bmatrix}$$

It is sometimes useful to view this as summing the m rows of $\mathbf{A}$ with weights given by y_i:

$$\underline{y}^T\mathbf{A} = \sum_{i=1}^m y_i\underline{b}_i^T \quad \text{where } \underline{b}_i^T = [a_{i1}\ a_{i2}\ \cdots\ a_{in}]$$

- Outer product of vectors: Let $\underline{x} \in R^n$ and $\underline{y} \in R^m$. Consider $\underline{x}$ as an $n \times 1$ matrix, and $\underline{y}^T$ as a $1 \times n$ matrix. Then the *dyadic* or *outer product* of the vectors $\underline{x}$ and $\underline{y}$ is the $n \times m$ matrix that results from the matrix product of $\underline{x}$ and $\underline{y}^T$, denoted as

$$\underline{x}\underline{y}^T = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_m \\ x_2y_1 & x_2y_2 & \cdots & x_2y_m \\ \vdots & \vdots & & \vdots \\ x_ny_1 & x_ny_2 & \cdots & x_ny_m \end{bmatrix}$$

- Orthogonal matrices: A square $n \times n$ matrix $\mathbf{A}$ is **orthogonal** if $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}_n$. If we think of $\mathbf{A}$ as consisting of a set of columns, i.e. $\mathbf{A} = \begin{bmatrix} \underline{a}_1 & \underline{a}_2 & \cdots & \underline{a}_n \end{bmatrix}$, then in general

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} \underline{x}_1^T\underline{x}_1 & \underline{x}_1^T\underline{x}_2 & \cdots & \underline{x}_1^T\underline{x}_n \\ \underline{x}_2^T\underline{x}_1 & \underline{x}_2^T\underline{x}_2 & \cdots & \underline{x}_2^T\underline{x}_n \\ \vdots & \vdots & & \vdots \\ \underline{x}_n^T\underline{x}_1 & \underline{x}_n^T\underline{x}_2 & \cdots & \underline{x}_n^T\underline{x}_n \end{bmatrix}$$

Consequently, we see that $\mathbf{A}$ is orthogonal if and only if its columns are *orthogonal* and normalized, i.e. $\underline{x}_k \perp \underline{x}_j$, $k \neq j$, and $||\underline{x}_k|| = I$.

- **Rank:** The **rank** of an $m \times n$ matrix $\mathbf{A}$ is equal to the largest number of linearly independent columns in $\mathbf{A}$ and is written as rank($\mathbf{A}$). It is also equal to the largest number of linearly independent rows in $\mathbf{A}$.
  - Rank is preserved under transpose rank($\mathbf{A}^T$) = rank($\mathbf{A}$).
  - rank($\mathbf{A}$) $\leq \min(m, n)$.
  - If rank($\mathbf{A}$) = $\min(m, n)$, then we say the matrix has **full rank.**
- **Trace:** The *trace* of a square matrix $\mathbf{A} \in \Re^{n \times n}$ is the sum of its diagonal elements: $\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$ For any square matrix $\mathbf{A}$ that tr($\mathbf{A}$) = tr($\mathbf{A}^T$).

  For two square matrices $\mathbf{A}, \mathbf{B}$ with the same dimensions, tr($\mathbf{A} + \mathbf{B}$) = tr($\mathbf{A}$) + tr($\mathbf{B}$). For two matrices $\mathbf{A}, \mathbf{B}$ with dimensions so that $\mathbf{AB}$ and $\mathbf{BA}$ are defined, tr($\mathbf{AB}$) = tr($\mathbf{BA}$). In particular, note that

$$||\underline{x}||^2 = \underline{x}^T \underline{x} = \text{tr}\left(\underline{x}^T \underline{x}\right) = \text{tr}\left(\underline{x}\underline{x}^T\right)$$

A.2.2 Matrix Inverses and Determinants

A square $n \times n$ matrix $\mathbf{A}$ is *invertible* or *nonsingular* if the only solution of the equation $\mathbf{A}\underline{x} = 0$ is $\underline{x} = \underline{0}$. That is, the only vector producing zero output is the zero vector. Otherwise, it is called *non-invertible* or *singular*. If $\mathbf{A}$ is invertible, then there exists another $n \times n$ matrix $\mathbf{A}^{-1}$ called the *inverse* of $\mathbf{A}$, so that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$.

The property of invertibility is related to the solution of sets of equations. To see this, consider the set of equations $\mathbf{A}\underline{x} = \underline{y}$ where $\mathbf{A}$ is $n \times n$. This equation has a unique solution $\underline{x}$ for any $\underline{y}$ if and only if $\mathbf{A}$ is invertible (in which case the solution is $\mathbf{A}^{-1}\underline{y}$). Conversely, if $\mathbf{A}$ is singular, then there exists a non-zero vector $\underline{x}'$ such that $A\underline{x}' = 0$. In this case, we can add any multiple of $\underline{x}_N$ to a solution of the linear equation and produce another solution. Thus if $\mathbf{A}$ is singular, the system of equations will not have a unique solution.

The **determinant** det($\mathbf{A}$) of a square $n \times n$ matrix $\mathbf{A}$ can be defined in two equivalent ways:

1. *Recursive:* For any scalar a, define $\det(a) = a$. If $\mathbf{A}$ is $n \times n$, then we can compute det($\mathbf{A}$)| by "expanding by minors" using any row or column, as follows: Let $\mathbf{M}_{k\ell}$ denote the matrix obtained by deleting the kth row and ℓth column from $\mathbf{A}$; this is called the "minor" of $\mathbf{A}$ at (k, ℓ). Note that this is an $n - 1 \times n - 1$ matrix. Then, for any $1 \leq \ell \leq n$

$$\det(\mathbf{A}) = \sum_{k=1}^{n} (-1)^{k+\ell} a_{k\ell} \det(\mathbf{M}_{k\ell})$$

   . Equivalently, for any $1 \leq k \leq n$,

$$\det(\mathbf{A}) = \sum_{\ell=1}^{n} (-1)^{k+\ell} a_{k\ell} \det(\mathbf{M}_{k\ell}) \ .$$

   For example, if $n = 2$, then $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$.

2. *Alternating Sum:* Let $\mathcal{S}$ denote the set of all $n!$ possible permutations of the set $\{1, 2, \ldots, n\}$. The sign of a permutation $\sigma \in \mathcal{S}$, denoted by sign(σ), is equal to $+1$ if the minimum number of pairwise interchanges needed to arrive at σ from $\{1, 2, \ldots, n\}$ is even and equal to -1 if it is odd. Then,

$$\det(\mathbf{A}) = \sum_{\sigma \in \mathcal{S}} \text{sign}(\sigma) \prod_{k=1}^{n} a_{k\sigma(k)} \ .$$

For most small matrices, the recursive way is an efficient way of computing the determinant. Consider the following example:

Example A.1
Define the matrix $\mathbf{A}$ as:

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 & 3 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 5 & 1 & 1 & 9 \end{bmatrix}.$$

Then,

$$\begin{aligned}
\det(\mathbf{A}) &= 2(-1)^{1+1}\det\left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 9 \end{bmatrix}\right) + 0(-1)^{1+2}\det\left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 5 & 1 & 9 \end{bmatrix}\right) \\
&\quad + 0(-1)^{1+3}\det\left(\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 5 & 1 & 9 \end{bmatrix}\right) 3(-1)^{1+4}\det\left(\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 5 & 1 & 1 \end{bmatrix}\right) \\
&= 2\det\left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 9 \end{bmatrix}\right) - 3\det\left(\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 5 & 1 & 1 \end{bmatrix}\right) \\
\det\left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 9 \end{bmatrix}\right) &= 1(-1)^{1+1}\det\left(\begin{bmatrix} 1 & 0 \\ 1 & 9 \end{bmatrix}\right) + 0(-1)^{1+2}\det\left(\begin{bmatrix} 1 & 0 \\ 1 & 9 \end{bmatrix}\right) + 0(-1)^{1+3}\det\left(\begin{bmatrix} 1 & 1 \\ 5 & 1 \end{bmatrix}\right) \\
&= 9 \\
\det\left(\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 5 & 1 & 1 \end{bmatrix}\right) &= 1(-1)^{1+1}\det\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right) + 1(-1)^{1+2}\det\left(\begin{bmatrix} 1 & 1 \\ 5 & 1 \end{bmatrix}\right) \\
&= 1(0) - 1(-4) = 4 \\
\det(\mathbf{A}) &= 2 \cdot 9 - 3 \cdot 4 = 6
\end{aligned}$$

Some basic properties of determinants are:

- For square $n \times n$ matrices $\mathbf{A}, \mathbf{B}$, $\det(\mathbf{AB}) = \det(\mathbf{BA}) = \det(\mathbf{A})\det(\mathbf{B})$.
- For a scalar α, $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$.
- $\det(\mathbf{A}) = \det(\mathbf{A}^T)$.
- If $\mathbf{A}$ is invertible, $\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})}$.

The invertibility of a square matrix $n \times n$ matrix $\mathbf{A}$ is equivalent to each of the following statements:

1. $\mathbf{A}$ is invertible.
2. $\det(A) \neq 0$.
3. $\text{rank}(\mathbf{A}) = n$.
4. All of the columns of $\mathbf{A}$ are linearly independent.
5. All of the rows of $\mathbf{A}$ are linearly independent.
6. The equation $\mathbf{A}\underline{x} = \underline{y}$ has a unique solution $\underline{x}$ for each choice of $\underline{y}$.

The determinant is a useful tool for computing the inverse of an invertible matrix $\mathbf{A}$. The inverse of $\mathbf{A}$ can be expressed as

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\mathbf{C}^T$$

where the matrix $\mathbf{C}$ is computed using the determinants of the minors of $\mathbf{A}$. Recall that the minor $\mathbf{M}_{k\ell}$ is the matrix obtained by deleting the kth row and ℓth column from $\mathbf{A}$. The $k\ell$-th element of $\mathbf{C}$ is obtained as $[\mathbf{C}]_{k\ell} = (-1)^{k+\ell}\det(\mathbf{M}_{k\ell})$. The matrix $\mathbf{C}$ is known as the matrix of cofactors.

For example, let $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. The minors are $\mathbf{M}_{11} = a_{22}, \mathbf{M}_{12} = a_{21}, \mathbf{M}_{21} = a_{12}, \mathbf{M}_{22} = a_1 1$. Since they are scalars, their determinant is equal to their value, so the matrix of cofactors is $\mathbf{C} = \begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix}$. As computed before, we know $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$. Then,

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}}\mathbf{C}^T = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

We can verify that this matrix is indeed the inverse of $\mathbf{A}$ by computing the matrix product.

Some useful properties of inverses are

- $\left(\mathbf{A}^T\right)^{-1} = \left(A^{-1}\right)^T$.
- If $\mathbf{A}, \mathbf{B}$ are invertible square $n \times n$ matrices, $(AB)^{-1} = B^{-1}A^{-1}$.
- If $\mathbf{A}$ is diagonal, such that $\mathbf{A} = \mathsf{diag}\,(\mu_1, \ldots, \mu_n)$, then $\mathbf{A}^{-1} = \mathsf{diag}\left(\frac{1}{\mu_1}, \cdots, \frac{1}{\mu_n}\right)$.
- The matrix $\mathbf{A}$ is orthogonal if and only if $\mathbf{A}^{-1} = \mathbf{A}^T$.

A.2.3 Eigenvalues and Eigenvectors

Let $\mathbf{A}$ be an $n \times n$ real matrix. A scalar λ is called an *eigenvalue* of A with associated nonzero *eigenvector* $\underline{x}$ if

$$A\underline{x} = \lambda\underline{x}.$$

The above equation can be rewritten as $(\lambda\mathbf{I}_n - \mathbf{A})\underline{x} = \underline{0}$. Thus λ is an eigenvalue of $\mathbf{A}$ if and only if the above equation has a solution $\underline{x} \neq \underline{0}$. This will be the case if and only if $\lambda\mathbf{I}_n - \mathbf{A}$ is singular.

Recall that a square matrix is singular if and only if its determinant is zero. Define the characteristic polynomial of $\mathbf{A}$ as the following polynomial in the variable s:

$$p_{\mathbf{A}(s)} = \det(\lambda\mathbf{I}_n - \mathbf{A}) =$$

This will be an $n-th$ degree polynomial, as diagonal elements of the matrix $\lambda\mathbf{I}_n - \mathbf{A}$ consist of terms $s - a_{ii}$. The eigenvalues of the matrix $\mathbf{A}$ must then be solution of the characteristic equation $p_{\mathbf{A}}(s) = 0$.

The characteristic polynomial can always be factored in terms its roots, as $p_{\mathbf{A}}(s) = (s-\lambda_1)(s-\lambda_2)\cdots(s-\lambda_n)$, where the roots may be complex numbers, and they may be repeated. In terms of unique roots, we can factor it as

$$p_{\mathbf{A}}(s) = (s-\lambda_1)^{n_1}(s-\lambda_2)^{n_2}\cdots(s-\lambda_k)^{n_k} \ ,$$

where the scalars $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the k unique eigenvalues of $\mathbf{A}$ where $1 \leq k \leq n$ and n_i is the **algebraic multiplicity** of eigenvalue λ_i. The algebraic multiplicities sum up to the dimension, $\sum_{i=1}^{k} n_i = n$.

Note that the eigenvalues may be complex numbers; however, if a complex λ_i is an eigenvalue, then its complex conjugate λ_i^* is also an eigenvalue, because the coefficients of the polynomial $p_{\mathbf{A}}(s)$ are real numbers.

For each eigenvalue λ_i, there can be m_i linearly independent vectors that satisfy $\left(\lambda_i\mathbf{I}_n - \mathbf{A}\right)\underline{x} = 0$. We refer to m_i as the **geometric multiplicity** of eigenvalue λ_i and note that it is bounded by the algebraic multiplicity, $1 \leq m_i \leq n_i$. Note that, if the eigenvalue λ_i is complex, its associated eigenvectors will be complex also.

If λ_i is an eigenvalue of A, then we can determine an associated eigenvector by solving the set of linear equations $\mathbf{A}\underline{x} = \lambda_i \underline{x}$. Note that if $\underline{x}$ is an eigenvector, so is $\alpha\underline{x}$ for any scalar α. Consequently, we can always adjust the length of the eigenvectors arbitrarily. Note that each distinct λ_i has a linearly independent $\underline{x}_i$ corresponding to it. If λ_i has multiplicity $k > 1$, i.e. if λ_i is a k-th order root of $p_{\mathbf{A}}(\lambda)$, then there may be anywhere from 1 to k linearly independent eigenvectors associated with λ_i.

For the special case where $\mathbf{A}$ is symmetric, we can show that all the eigenvalues λ_i are real-valued, as will the eigenvectors. Let u^{ast} denote the complex conjugate of u. Suppose that λ is a (possibly complex) eigenvalue of the real symmetric matrix $\mathbf{A}$. Thus there is a nonzero vector $\underline{v}$, also with complex entries, such that $\mathbf{A}\underline{v} = \lambda\underline{v}$. Let u^{ast} denote the complex conjugate of u. By taking the complex conjugate of both sides, and noting that $\mathbf{A}^* = \mathbf{A}$ since $\mathbf{A}$ has real entries, we get

$$\mathbf{A}\underline{v} = \lambda\underline{v} \Rightarrow \mathbf{A}\underline{v}^* = \lambda^*\underline{v}^*.$$

Then,

$$\begin{aligned}(\underline{v}^*)^T\mathbf{A}\underline{v}\lambda^*(\underline{v}^*)^T\underline{v} &= \lambda^*||\underline{v}||^2 \\ &= (\underline{v}^*)^T\lambda\underline{v} = \lambda||\underline{v}||^2\end{aligned}$$

which implies $\lambda^* = \lambda$, and thus λ is real. Similarly, if $\underline{v}$ is an eigenvector for real eigenvalue λ, then

$$\mathbf{A}\underline{v} = \lambda\underline{v} \Rightarrow \mathbf{A}^*\underline{v}^* = \mathbf{A}\underline{v}^* = \lambda^*\underline{v}^* = \lambda\underline{v}^*$$

which implies that $\underline{v}^*$ is also an eigenvector for eigenvalue λ. If the multiplicity of λ is 1, this implies that $\underline{v} = \underline{v}^*$ and hence it is real. If the multiplicity is greater than 1, one can show with a lengthier argument that we can pick real eigenvectors corresponding to the eigenvalue λ.

Another property of real symmetric matrices is that there exist a set of real eigenvectors that are orthogonal. This is easy to show for two eigenvectors $\underline{v}_1$, $\underline{v}_2$ corresponding to distinct eigenvalues λ_1, λ_2, because

$$\begin{aligned}\underline{v}_1^T\mathbf{A}\underline{v}_2 &= (\mathbf{A}\underline{v}_1)^T\underline{v}_2 = \lambda_1\underline{v}_1^T\underline{v}_2 \\ &= \underline{v}_1^T(\mathbf{A}\underline{v}_2) = \lambda_2\underline{v}_1^T\underline{v}_2\end{aligned}$$

which shows that $\underline{v}_1^T\underline{v}_2 = 0$, because $\lambda_1 \neq \lambda_2$. Again, a more sophisticated analysis extends this result to the fact that we can pick orthogonal eigenvectors for eigenvalues that are repeated. Since multiples of eigenvalues are also eigenvalues, this means we can pick an orthonormal set of eigenvectors for any real symmetric $\mathbf{A}$.

For a real, symmetric matrix $\mathbf{A}$, let λ_i denote its eigenvalues and $\underline{v}_i$ denote a set of corresponding orthonormal eigenvectors. We can represent the matrix $\mathbf{A}$ by its spectral eigendecompositon as:

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \underline{v}_i \underline{v}_i^T \ .$$

This is also expressed in terms of a matrix decomposition as

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$$

where the matrix $\mathbf{V}$ is an orthogonal $n \times n$ matrix with columns corresponding to the eigenvectors $\underline{v}_i$, and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$.

Some additional facts about eigenvalues of matrices $\mathbf{A}$. Using the definition of the characteristic polynomial, we see the following: $p_{\mathbf{A}}(s) = (s-\lambda_1)(s-\lambda_2)\cdots(s-\lambda_n)$, where $\lambda_i, i = 1, \ldots, n$ are the eigenvalues of $\mathbf{A}$, possibly repeated. By definition, $p_{\mathbf{A}}(s) = \det(s\mathbf{I}_n - \mathbf{A})$, so $p_{\mathbf{A}}(0) = \det(-\mathbf{A}) = (-1)^n\det(\mathbf{A})$. Thus,

$$p_{\mathbf{A}}(0) = (-1)^n\det(\mathbf{A}) = (-1)^n \prod_{i=1}^{n} \lambda_i$$

which shows that the determinant of $\mathbf{A}$ is the product of its eigenvalues. With a little more work, we can show that the trace of $\mathbf{A}$ is the negative of the coefficient of s^{n-1} in $p_{\mathbf{A}}(s)$, and is given by:

$$\mathrm{tr}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

which is the sum of the eigenvalues!

A.3 Similarity Transformations and Change of Bases

A matrix $\mathbf{A}$ with dimensions $m \times n$ specifies a linear transformation from vectors in $\Re^n$ to vectors in $\Re^m$. The specific elements of the matrix correspond to how the basis vectors in $\Re^n$ are mapped into combinations of the basis vectors in $\Re^m$. Specifically, column i of $\mathbf{A}$ represents the coefficients in the basis for $\Re^m$ that represent the vector that is the image of the $i-th$ basis vector in $\Re^n$. If we change the choice of basis in $\Re^m$ or $\Re^n$, we will obtain different values for the coefficient of the matrix $\mathbf{A}$.

Let's consider only matrices $\mathbf{A}$ that are $n \times n$. Let $\mathbf{C}$ be an invertible matrix of the same size. We can then define a *similarity transformation* of $\mathbf{A}$ as $\mathbf{B} = \mathbf{C}\mathbf{A}\mathbf{C}^{-1}$. We say that "$\mathbf{B}$ is similar to $\mathbf{A}$". A similarity transformation corresponds essentially to a change of coordinates. Specifically, suppose

$$\underline{y} = \mathbf{A}\underline{x}$$

and consider a change of coordinates

$$\underline{u} = \mathbf{C}\underline{x}, \qquad \underline{v} = \mathbf{C}\underline{y}$$

(so that each component of $\underline{u}$, for example, is a weighted sum of components of $\underline{x}$ and vice versa, since $\underline{x} = \mathbf{C}^{-1}\underline{u}$). Then $\underline{v} = \mathbf{B}\underline{u}$.

Note that the determinant is not changed by similarity transformations, because

$$\det(\mathbf{B}) = \det(\mathbf{C}\mathbf{A}\mathbf{C}^{-1}) = \det(\mathbf{C})\det(\mathbf{A})\det(\mathbf{C}^{-1}) = \det(\mathbf{A})$$

Furthermore, the characteristic polynomial does not change:

$$\begin{aligned} p_{\mathbf{B}}(s) &= \det(s\mathbf{I}_n - \mathbf{B}) = \det(s\mathbf{I}_n - \mathbf{C}\mathbf{A}\mathbf{C}^{-1}) \\ &= \det(s\mathbf{C}\mathbf{I}_n\mathbf{C}^{-1} - \mathbf{C}\mathbf{A}\mathbf{C}^{-1}) \\ &= \det(\mathbf{C}(s\mathbf{I}_n - \mathbf{A})\mathbf{C}^{-1}) = \det(s\mathbf{I}_n - \mathbf{A}) \end{aligned}$$

Thus, the eigenvalues of $\mathbf{B}$ and $\mathbf{A}$ are the same.

Suppose that the $n \times n$ matrix A has a full set of linearly independent eigenvectors $\underline{x}_1, \ldots, \underline{x}_n$, so that $\mathbf{A}\underline{x}_i = \lambda_i \underline{x}_i,\ i = 1, \ldots, n$. The existence of such a complete set of eigenvectors is guaranteed, for example, if the λ_i are all distinct or if $\mathbf{A}$ is symmetric.

We can rewrite these as one equation

$$\mathbf{A}\begin{bmatrix}\underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_n\end{bmatrix} = \begin{bmatrix}\underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_n\end{bmatrix}\begin{bmatrix}\lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & \lambda_n\end{bmatrix} \tag{A.2}$$

Let $\mathbf{C}^{-1} = \begin{bmatrix}c|c|c|c\underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_n\end{bmatrix}$ which is invertible, since the columns $\underline{x}_1, \ldots, \underline{x}_n$ are linearly independent. Then (A.2)implies that $\mathbf{C}A\mathbf{C}^{-1} = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Note that if A is symmetric we can choose the $\underline{x}_i$ to be orthonormal so that $\mathbf{C}^{-1} = \mathbf{C}^T$.

A.4 Positive-semidefinite and Positive-definite Matrices

A symmetric square matrix $\mathbf{A}$ is *positive semidefinite*, written $\mathbf{A} \geq 0$, if and only if $ullx^T\mathbf{A}\underline{x} \geq 0$ for all vectors $\underline{x}$. This matrix $\mathbf{A}$ is *positive definite*, written $\mathbf{A} > 0$, if $\underline{x}^T\mathbf{A}\underline{x} > 0$ for all $\underline{x}$. It is not difficult to see that a positive semidefinite matrix is positive definite if and only if it is invertible.

Some basic facts about positive semidefinite matrices are the following:

- If $\mathbf{A} \geq 0$ and $\mathbf{B} \geq 0$, then $\mathbf{A} + \mathbf{B} \geq 0$, since $\underline{x}^T(\mathbf{A} + \mathbf{B})\underline{x} = \underline{x}^T\mathbf{A}\underline{x} + \underline{x}^T\mathbf{B}\underline{x}$
- If either $\mathbf{A}$ or $\mathbf{B}$ in (i) is positive definite, then so is $\mathbf{A} + \mathbf{B}$.
- If $A > 0$, then $A^{-1} > 0$ since
$$\underline{x}^T A^{-1}\underline{x} = (A^{-1}\underline{x})^T A(A^{-1}\underline{x}) > 0$$
- If $\mathbf{A} \geq 0$ then $\mathbf{C}^T\mathbf{A}\mathbf{C} \geq 0$ for *any* (not necessarily square) matrix $\mathbf{C}$ for which $\mathbf{C}^T\mathbf{A}\mathbf{C}$ is defined.
- If $\mathbf{A} > 0$ and $\mathbf{C}$ is invertible, $\mathbf{C}^T\mathbf{A}\mathbf{C} > 0$.
- $\mathbf{A} \geq 0$ if and only if all of its eigenvalues are real and non-negative. This is because, for any eigenvector $\underline{x}$ with eigenvalue λ, we have
$$\underline{x}^T\mathbf{A}\underline{x} = \lambda\underline{x}^T\underline{x} = \lambda||\underline{x}^2 \geq 0.$$
- $\mathbf{A} > 0$ if and only if all of its eigenvalues are strictly positive.

One test for positive definiteness is *Sylvester's Test*. Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix}$$

Then $\mathbf{A}$ is positive semidefinite (positive definite) if and only if

$$\begin{aligned} \det(a_{11}) &\geq 0 \\ \det\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}\right) &\geq 0 \\ \det\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}\right) &\geq 0 \\ &\vdots \text{etc.} \end{aligned}$$

Let $\mathbf{A} \geq 0$, and let $\mathbf{C}$ be the orthogonal matrix of eigenvectors so that $\mathbf{C}^T\mathbf{A}\mathbf{C} = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. The, we can construct a simple square root of the matrix $\mathbf{A}$ so that $\mathbf{A} = (\mathbf{A}^{1/2})^T\mathbf{A}^{1/2}$. Specifically, we take

$$\mathbf{A}^{1/2} = \mathsf{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right)\mathbf{C}^T.$$

Note that the square root is invertible if and only if $\mathbf{A} > 0$ so all the eigenvalues are positive. Also, note that the above square root is far from unique: We can multiply it by an orthogonal matrix $\mathbf{B}$ and get another square root.

A.5 Subspaces

A subset $S \subseteq \Re^n$ is a subspace if S is closed under vector addition and scalar multiplication. That is, if $\underline{x} \in S$, and α is a scalar, then $\alpha\underline{x} \in S$. Furthermore, if $\underline{x}, \underline{y} \in S, then \underline{x} + \underline{y} \in S$.

Examples of subspaces of $\Re^2$ are[1]

$$S_1 = \left\{ \begin{bmatrix} a \\ 0 \end{bmatrix} \mid a \in \Re \right\}$$

[1]Here $\Re$ denotes the set of real numbers.

$$S_2 = \left\{ \begin{bmatrix} a \\ 2a \end{bmatrix} \mid a \in \Re \right\}$$

The *dimension* of a subspace equals the maximum number of vectors in S that can form a linearly independent set.

Let K be any subset of $\Re^n$. The *orthogonal complement* of K is defined as follows:

$$K^{\perp} = \left\{ \underline{x} \in \Re^n \mid \underline{x} \perp \underline{y} \; \forall \underline{y} \in K \right\}$$

$K^{\perp}$ is a subspace whether or not K is, since if $\underline{x}_1, \underline{x}_2 \in K^{\perp}$, $\underline{y} \in K$, we have $(\underline{x}_1 + \underline{x}_2)^T \underline{y} = 0$ and $(\alpha \underline{x}_1)^T \underline{y} = 0$.

Let $\underline{d}$ be a single nonzero vector in $\Re^n$ and consider $\{d\}^{\perp}$. This is a subspace of dimension $n-1$. If $n-2$, then the set of all vectors $\underline{x}$ such that $\underline{d}^T \underline{x} = 0$ is a line through the origin perpendicular to d. In 3-dimensions this set is a plane through the origin, again perpendicular to d. Note that the subspace $\{\underline{d}\}^{\perp}$ splits $\Re^n$ into two *half-spaces*, one corresponding to those $\underline{x}$ for which $\underline{d}^T \underline{x} > 0$, the other to those $\underline{x}$ for which $\underline{d}^T \underline{x} < 0$.

Appendix B

Examples of Subsets that are not Events

We are going to construct subsets of an uncountable sample space Ω that cannot be considered events, because which we will not be able to define a probability measure that is consistent with the axioms of probability. Consider the sample space $\Omega = [0,1]$: For a number $s \in [0,1]$, define the set $A_s = \{x \in [0,1] : |x - s|$is a rational number. $\}$ Since the rational numbers are countable, the set $A_s \subset \Omega$ has a countable number of elements. Since s is a real number in $[0,1]$, there are an uncountable number of possible sets A_s. Furthermore, we can find an uncountable collection of $s \in [0,1]$, denoted by C, such that, if $s, t \in C$, then $A_s \cap A_t = \emptyset$, and $\cup_{s \in C} A_s = [0,1]$. In this manner, we have constructed an uncountable partition of the unit interval $[0,1]$, where each set in the partition has a countable number of elements, and any two sets in the partition are disjoint.

Define the set B by selecting one element from each $A_s, s \in C$. Since there are an uncountable number of $s \in C$, the set A_s has an uncountable number of elements. For any rational number $r_i \in [0,1]$, define the translation of B by r_i as $B_i = \{y : y = (x + r_i) \bmod 1$ for some $x \in B\}$. In this definition, we use the modular operation $x \bmod 1 = x - \lfloor x \rfloor$.

Note the following:

- There are a countable number of B_i, because there are a countable number of rational numbers in [0,1].
- Each B_i has an uncountable number of elements, which are translations of the elements of B.
- $B_i \cap B_j = \emptyset$ if $i \neq j$, because B contains one and only one element from each $A_s \in C$. Note that, if the conclusion were not true, then there is are $x, y \in B, x \neq y$ such that $x + r_i = y + r_j$, which would imply that $x, y \in A_s$ for some s, contradicting the construction of B.
- $\cup_{i=1}^{\infty} B_i = \cup_s A_s = [0,1] = \Omega$. This last property follows because the sets A_s consist of the rational translations of s, so that every element of A_s is in some B_i.

Thus, the sets B_i form a countable partition of the interval [0,1].

Denote the probability measure $\mathbb{P}$ such that the probability of an interval (a,b) is its length: $\mathbb{P}[(a,b)] = b - a$ for $0 \leq a \leq b \leq 1$. If we were to make B an event, what would be $\mathbb{P}[B]$? By construction, since each B_i is a simple translation of B, then $\mathbb{P}[B] = \mathbb{P}[B_i]$ for all i. If we were to assign $\mathbb{P}[B] = 0$, then the countable additivity property would require

$$\mathbb{P}[\cup_{i=1}^{\infty} B_i] = \mathbb{P}[\Omega] = \sum_{i=1}^{\infty} \mathbb{P}[B_i] = 0,$$

which would violate the normality property that requires $\mathbb{P}[\Omega] = 1$. If we assign $\mathbb{P}[B] = a > 0$, then

$$\mathbb{P}[\cup_{i=1}^{\infty} B_i] = \mathbb{P}[\Omega] = \sum_{i=1}^{\infty} \mathbb{P}[B_i] = \infty,$$

violating the normalization property of probability measures. Thus, we have a subset of $\Omega = [0,1]$ for which we cannot assign a probability which is compatible with the axioms of probability theory and the definition of the uniform probability measure. Therefore, B cannot be an event in this probability space, although $B \subset [0,1]$.

Appendix C

Standard Normal Cumulative Distribution Function

This appendix contains the table of the standard Normal CDF $\Phi(x)$ described in 3, subsection 3.4.3. To compute the complementary CDF, recall that $Q(x) = 1 - \Phi(x)$. Also, $Q(x) = \Phi(-x)$.

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

Figure C.1: Part 1 of Standard Normal Cumulative Distribution Function: Negative x.

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Figure C.2: Part 2 of Standard Normal Cumulative Distribution Function: positive x.