Lecture Notes on Probability Theory and Random Processes

Jean Walrand
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720

August 25, 2004

Table of Contents

Abstract

These notes are derived from lectures and office-hour conversations in a junior/senior-level course on probability and random processes in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley.

The notes do not replace a textbook. Rather, they provide a guide through the material. The style is casual, with no attempt at mathematical rigor. The goal to to help the student figure out the meaning of various concepts and to illustrate them with examples.

When choosing a textbook for this course, we always face a dilemma. On the one hand, there are many excellent books on probability theory and random processes. However, we find that these texts are too demanding for the level of the course. On the other hand, books written for the engineering students tend to be fuzzy in their attempt to avoid subtle mathematical concepts. As a result, we always end up having to complement the textbook we select. If we select a math book, we need to help the student understand the meaning of the results and to provide many illustrations. If we select a book for engineers, we need to provide a more complete conceptual picture. These notes grew out of these efforts at filling the gaps.

You will notice that we are not trying to be comprehensive. All the details are available in textbooks. There is no need to repeat the obvious.

The author wants to thank the many inquisitive students he has had in that class and the very good teaching assistants, in particular Teresa Tung, Mubaraq Misra, and Eric Chi, who helped him over the years; they contributed many of the problems.

Happy reading and keep testing hypotheses!

Berkeley, June 2004 - Jean Walrand

Introduction

Engineering systems are designed to operate well in the face of uncertainty of characteristics of components and operating conditions. In some case, uncertainty is introduced in the operations of the system, on purpose.

Understanding how to model uncertainty and how to analyze its effects is – or should be – an essential part of an engineer's education. Randomness is a key element of all systems we design. Communication systems are designed to compensate for noise. Internet routers are built to absorb traffic fluctuations. Building must resist the unpredictable vibrations of an earthquake. The power distribution grid carries an unpredictable load. Integrated circuit manufacturing steps are subject to unpredictable variations. Searching for genes is looking for patterns among unknown strings.

What should you understand about probability? It is a complex subject that has been constructed over decades by pure and applied mathematicians. Thousands of books explore various aspects of the theory. How much do you really need to know and where do you start?

The first key concept is how to model uncertainty (see Chapter 2 - 3). What do we mean by a "random experiment?" Once you understand that concept, the notion of a random variable should become transparent (see Chapters 4 - 5). You may be surprised to learn that a random variable does not vary! Terms may be confusing. Once you appreciate the notion of randomness, you should get some understanding for the idea of expectation (Section 4.5) and how observations modify it (Chapter 6). A special class of random variables (Gaussian)

are particularly useful in many applications (Chapter 7). After you master these key notions, you are ready to look at detection (Chapter 8) and estimation problems (Chapter 9). These are representative examples of how one can process observation to reduce uncertainty. That is, how one learns. Many systems are subject to the cumulative effect of many sources of randomness. We study such effects in Chapter 11 after having provided some background in Chapter 10. The final set of important notions concern random processes: uncertain evolution over time. We look at particularly useful models of such processes in Chapters 12-15. We conclude the notes by discussing a few applications in Chapter 16.

The concepts are difficult, but the math is not (Appendix **??** reviews what you should know). The trick is to know what we are trying to compute. Look at examples and invent new ones to reinforce your understanding of ideas. Don't get discouraged if some ideas seem obscure at first, but do not let the obscurity persist! This stuff is not that hard, it is only new for you.

Chapter 1

Modelling Uncertainty

In this chapter we introduce the concept of a model of an uncertain physical system. We stress the importance of concepts that justify the structure of the theory. We comment on the notion of a hidden variable. We conclude the chapter with a very brief historical look at the key contributors and some notes on references.

1.1 Models and Physical Reality

Probability Theory is a mathematical model of uncertainty. In these notes, we introduce examples of uncertainty and we explain how the theory models them.

It is important to appreciate the difference between uncertainty in the physical world and the models of Probability Theory. That difference is similar to that between laws of theoretical physics and the real world: even though mathematicians view the theory as standing on its own, when engineers use it, they see it as a model of the physical world.

Consider flipping a fair coin repeatedly. Designate by 0 and 1 the two possible outcomes of a coin flip (say 0 for head and 1 for tail). This experiment takes place in the physical world. The outcomes are uncertain. In this chapter, we try to appreciate the probability model of this experiment and to relate it to the physical reality.

1.2 Concepts and Calculations

In our many years of teaching probability models, we have always found that what is most subtle is the interpretation of the models, not the calculations. In particular, this introductory course uses mostly elementary algebra and some simple calculus. However, understanding the meaning of the models, what one is trying to calculate, requires becoming familiar with some new and nontrivial ideas.

Mathematicians frequently state that "definitions do not require interpretation." We beg to disagree. Although as a logical edifice, it is perfectly true that no interpretation is needed; but to develop some intuition about the theory, to be able to anticipate theorems and results, to relate these developments to the physical reality, it is important to have some interpretation of the definitions and of the basic axioms of the theory. We will attempt to develop such interpretations as we go along, using physical examples and pictures.

1.3 Function of Hidden Variable

One idea is that the uncertainty in the world is fully contained in the selection of some hidden variable. (This model does not apply to quantum mechanics, which we do not consider here.) If this variable were known, then nothing would be uncertain anymore. Think of this variable as being picked by nature at the big bang. Many choices were possible, but one particular choice was made and everything derives from it. [In most cases, it is easier to think of nature's choice only as it affects a specific experiment, but we worry about this type of detail later.] In other words, everything that is uncertain is a function of that hidden variable. By function, we mean that if we know the hidden variable, then we know everything else.

Let us denote the hidden variable by ω. Take one uncertain thing, such as the outcome of the fifth coin flip. This outcome is a function of ω. If we designate the outcome of

Figure 1.1: Adrien Marie Legendre

the fifth coin flip by X, then we conclude that X is a function of ω. We can denote that function by $X(\omega)$. Another uncertain thing could be the outcome of the twelfth coin flip. We can denote it by $Y(\omega)$. The key point here is that X and Y are functions of the same ω. Remember, there is only one ω (picked by nature at the big bang).

Summing up, everything that is random is some function X of some hidden variable ω. This is a model. To make this model more precise, we need to explain how ω is selected and what these functions $X(\omega)$ are like. These ideas will keep us busy for a while!

1.4 A Look Back

The theory was developed by a number of inquiring minds. We briefly review some of their contributions. (We condense this historical account from the very nice book by S. M. Stigler [9]. For ease of exposition, we simplify the examples and the notation.)

Adrien Marie LEGENDRE, 1752-1833

Best use of inaccurate measurements: Method of Least Squares.

To start our exploration of "uncertainty" We propose to review very briefly the various attempts at making use of inaccurate measurements.

Say that an amplifier has some gain A that we would like to measure. We observe the

input X and the output Y and we know that $Y = AX$. If we could measure X and Y precisely, then we could determine A by a simple division. However, assume that we cannot measure these quantities precisely. Instead we make two sets of measurements: (X, Y) and (X', Y'). We would like to find A so that $Y = AX$ and $Y' = AX'$. For concreteness, say that $(X, Y) = (2, 5)$ and $(X', Y') = (4, 7)$. No value of A works exactly for both sets of measurements. The problem is that we did not measure the input and the output accurately enough, but that may be unavoidable. What should we do?

One approach is to average the measurements, say by taking the arithmetic means: $((X + X')/2, (Y + Y')/2) = (3, 6)$ and to find the gain A so that $6 = A \times 3$, so that $A = 2$. This approach was commonly used in astronomy before 1750.

A second approach is to solve for A for each pair of measurements: For (X, Y), we find $A = 2.5$ and for (X', Y'), we find $A = 1.75$. We can average these values and decide that A should be close to $(2.5 + 1.75)/2 = 2.125$.

We skip over many variations proposed by Mayer, Euler, and Laplace.

Another approach is to try to find A so as to minimize the sum of the squares of the errors between Y and AX and between Y' and AX'. That is, we look for A that minimizes $(Y - AX)^2 + (Y' - AX')^2$. In our example, we need to find A that minimizes $(5 - 2A)^2 + (7 - 4A)^2 = 74 - 76A + 20A^2$. Setting the derivative with respect to A equal to 0, we find $-76 + 40A = 0$, or $A = 1.9$. This is the solution proposed by Legendre in 1805. He called this approach the *method of least squares*.

The method of least squares is one that produces the "best" prediction of the output based on the input, under rather general conditions. However, to understand this notion, we need to make a short excursion on the characterization of uncertainty.

Jacob BERNOULLI, 1654-1705

Making sense of uncertainty and chance: *Law of Large Numbers*.

Figure 1.2: Jacob Bernoulli

If an urn contains 5 red balls and 7 blue balls, then the odds of picking "at random" a red ball from the urn are 5 out of 12. One can view the likelihood of a complex event as being the ratio of the number of favorable cases divided by the total number of "equally likely" cases. This is a somewhat circular definition, but not completely: from symmetry considerations, one may postulate the existence equally likely events. However, in most situations, one cannot determine – let alone count – the equally likely cases nor the favorable cases. (Consider for instance the odds of having a sunny Memorial Day in Berkeley.) Jacob Bernoulli (one of twelve Bernoullis who contributed to Mathematics, Physics, and Probability) showed the following result. If we pick a ball from an urn with r red balls and b blue balls a large number N of times (always replacing the ball before the next attempt), then the fraction of times that we pick a red ball approaches $r/(r+b)$. More precisely, he showed that the probability that this fraction differs from $r/(r+b)$ by more than any given $\epsilon > 0$ goes to 0 as N increases. We will learn this result as the weak law of large numbers.

Abraham DE MOIVRE, 1667 1754

Bounding the probability of deviation: Normal distribution

De Moivre found a useful approximation of the probability that preoccupied Jacob Bernoulli. When N is large and ϵ small, he derived the normal approximation to the

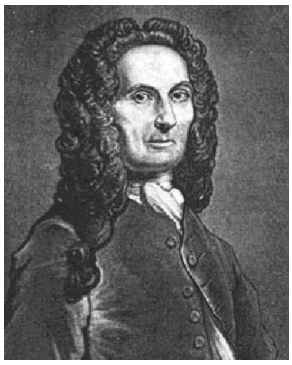


Figure 1.3: Abraham de Moivre

Figure 1.4: Thomas Simpson

probability discussed earlier. This is the first mention of this distribution and an example of the Central Limit Theorem.

Thomas SIMPSON, 1710-1761

A first attempt at posterior probability.

Looking again at Bernoulli's and de Moivre's problem, we see that they assumed $p = r/(r+b)$ known and worried about the probability that the fraction of N balls selected from the urn differs from p by more than a fixed $\epsilon > 0$. Bernoulli showed that this probability goes to zero (he also got some conservative estimates of N needed for that probability to be a given small number). De Moivre improved on these estimates.

Figure 1.5: Thomas Bayes

Simpson (a heavy drinker) worried about the "reverse" question. Assume we do not know p and that we observe the fraction q of a large number N of balls being red. We believe that p should be close to q, but how close can we be confident that it is? Simpson proposed a naïve answer by making arbitrary assumptions on the likelihood of the values of p.

Thomas BAYES, 1701-1761

The importance of the prior distribution: Bayes' rule.

Bayes understood Simpson's error. To appreciate Bayes' argument, assume that $q = 0.6$ and that we have made 100 experiments. What are the odds that $p \in [0.55, 0.65]$? If you are told that $p = 0.5$, then these odds are 0. However, if you are told that the urn was chosen such that $p = 0.5$ or $p = 1$, with equal probabilities, then the odds that $p \in [0.55, 0.65]$ are now close to 1.

Bayes understood how to include systematically the information about the prior distribution in the calculation of the posterior distribution. He discovered what we know today as Bayes' rule, a simple but very useful identity.

Pierre Simon LAPLACE, 1749-1827

Posterior distribution: Analytical methods.

Figure 1.6: Pierre Simon Laplace

Figure 1.7: Carl Friedrich Gauss

Laplace introduced the transform methods to evaluate probabilities. He provided derivations of the central limit theorem and various approximation results for integrals (based on what is known as Laplace's method).

Carl Friedrich GAUSS, 1777 1855

Least Squares Estimation with Gaussian errors.

Gauss developed the systematic theory of least squares estimation when the errors are Gaussian. We explain in the notes the remarkable fact that the best estimate is linear in the observations.

Figure 1.8: Andrei Andreyevich Markov

Andrei Andreyevich MARKOV, 1856 1922

Markov Chains

A sequence of coin flips produces results that are independent. Many physical systems exhibit a more complex behavior that requires a new class of models. Markov introduced a class of such models that enable to capture dependencies over time. His models, called *Markov chains*, are both fairly general and tractable.

Andrei Nikolaevich KOLMOGOROV, 1903-1987

Kolmogorov was one of the most prolific mathematicians of the 20th century. He made fundamental contributions to dynamic systems, ergodic theory, the theory of functions and functional analysis, the theory of probability and mathematical statistics, the analysis of turbulence and hydrodynamics, to mathematical logic, to the theory of complexity, to geometry, and topology.

In probability theory, he formulated probability as part of measure theory and established some essential properties such as the extension theorem and many other fundamental results.

Figure 1.9: Andrei Nikolaevich Kolmogorov

1.5 References

There are many good books on probability theory and random processes. For the level of this course, we recommend Ross [7], Hoel et al. [4], Pitman [5], and Bremaud [2]. The books by Feller [3] are always inspiring. For a deeper look at probability theory, Breiman [1] are a good start. For cute problems, we recommend Sevastyanov et al. [8].

Chapter 2

Probability Space

In this chapter we describe the probability model of "choosing an object at random." Examples will help us come up with a good definition. We explain that the key idea is to associate a likelihood, which we call *probability*, to sets of outcomes, not to individual outcomes. These sets are *events*. The description of the events and of their probability constitute a *probability space* that characterizes completely a random experiment.

2.1 Choosing At Random

First consider picking a card out of a 52-card deck. We could say that the odds of picking any particular card are the same as that of picking any other card, assuming that the deck has been well shuffled. We then decide to assign a "probability" of 1/52 to each card. That probability represents the odds that a given card is picked. One interpretation is that if we repeat the experiment "choosing a card from the deck" a large number N of times (replacing the card previously picked every time and re-shuffling the deck before the next selection), then a given card, say the ace of diamonds, is selected approximated $N/52$ times. Note that this is only an interpretation. There is nothing that tells us that this is indeed the case; moreover, if it is the case, then there is certainly nothing yet in our theory that allows us to expect that result. Indeed, so far, we have simply assigned the number 1/52 to each card

in the deck. Our interpretation comes from what we expect from the physical experiment. This remarkable "statistical regularity" of the physical experiment is a consequence of some deeper properties of the sequences of successive cards picked from a deck. We will come back to these deeper properties when we study independence. You may object that the definition of probability involves implicitly that of "equally likely events." That is correct as far as the interpretation goes. The mathematical definition does not require such a notion.

Second, consider the experiment of throwing a dart on a dartboard. The likelihood of hitting a specific point on the board, measured with pinpoint accuracy, is essentially zero. Accordingly, in contrast with the previous example, we cannot assign numbers to individual outcomes of the experiment. The way to proceed is to assign numbers to sets of possible outcomes. Thus, one can look at a subset of the dartboard and assign some probability that represents the odds that the dart will land in that set. It is not simple to assign the numbers to all the sets in a way that these numbers really correspond to the odds of a given dart player. Even if we forget about trying to model an actual player, it is not that simple to assign numbers to all the subsets of the dartboard. At the very least, to be meaningful, the numbers assigned to the different subsets must obey some basic consistency rules. For instance, if A and B are two subsets of the dartboard such that $A \subset B$, then the number $P(B)$ assigned to B must be at least as large as the number $P(A)$ assigned to A. Also, if A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$. Finally, $P(\Omega) = 1$, if Ω designates the set of all possible outcomes (the dartboard, possibly extended to cover all bases). This is the basic story: probability is defined on sets of possible outcomes and it is additive. [However, it turns out that one more property is required: countable additivity (see below).]

Note that we can lump our two examples into one. Indeed, the first case can be viewed as a particular case of the second where we would define $P(A) = |A|/52$, where A is any subset of the deck of cards and $|A|$ is the number of cards in the deck. This definition is certainly additive and it assigns the probability $1/52$ to any one card.

Some care is required when defining what we mean by a random choice. See Bertrand's paradox in Appendix E for an illustration of a possible confusion. Another example of the possible confusion with statistics is Simpson's paradox in Appendix F.

2.2 Events

The sets of outcomes to which one assigns a probability are called events. It is not necessary (and often not possible, as we may explain later) for every set of outcomes to be an event.

For instance, assume that we are only interested in whether the card that we pick is black or red. In that case, it suffices to define $P(A) = 0.5 = P(A^c)$ where A is the set of all the black cards and A^c is the complement of that set, i.e., the set of all the red cards. Of course, we know that $P(\Omega) = 1$ where Ω is the set of all the cards and $P(\emptyset) = 0$, where $\emptyset$ is the empty set. In this case, there are four events: $\emptyset, \Omega, A, A^c$.

More generally, if A and B are events, then we want $A^c, A \cap B$, and $A \cup B$ to be events also. Indeed, if we want to define the probability that the outcome is in A and the probability that it is in B, it is reasonable to ask that we can also define the probability that the outcome is not in A, that it is in A and B, and that it is in A or in B (or in both). By extension, set operations that are performed on a finite collection of events should always produce an event. For instance, if A, B, C, D are events, then $[(A \setminus B) \cap C] \cup D$ should also be an event. We say that the set of events is closed under finite set operations. [We explain below that we need to extend this property to countable operations.] With these properties, it makes sense to write for disjoint events A and B that $P(A \cup B) = P(A) + P(B)$. Indeed, $A \cup B$ is an event, so that $P(A \cup B)$ is defined.

You will notice that if we want $A \subset \Omega$ (with $A \neq \Omega$ and $A \neq \emptyset$) to be an event, then the smallest collection of events is necessarily $\{\emptyset, \Omega, A, A^c\}$.

If you want to see why, generally for uncountable sample spaces, all sets of outcomes

may not be events, check Appendix C.

2.3 Countable Additivity

This topic is the first serious hurdle that you face when studying probability theory. If you understand this section, you increase considerably your appreciation of the theory. Otherwise, many issues will remain obscure and fuzzy.

We want to be able to say that if the events A_n for $n = 1, 2, \ldots$, are such that $A_n \subset A_{n+1}$ for all n and if $A := \cup_n A_n$, then $P(A_n) \uparrow P(A)$ as $n \to \infty$. Why is this useful? This property, called *σ-additivity* is the key to being able to approximate events. The property specifies that the probability is continuous: if we approximate the events, then we also approximate their probability.

This strategy of "filling the gaps" by taking limits is central in mathematics. You remember that real numbers are defined as limits of rational numbers. Similarly, integrals are defined as limits of sums. The key idea is that different approximations should give the same result. For this to work, we need the continuity property above.

To be able to write the continuity property, we need to assume that $A := \cup_n A_n$ is an event whenever the events A_n for $n = 1, 2, \ldots$, are such that $A_n \subset A_{n+1}$. More generally, we need the set of events to be closed under countable set operations.

For instance, if we define $P([0, x]) = x$ for $x \in [0, 1]$, then we can define $P([0, a)) = a$ because if ϵ is small enough, then $A_n := [0, a - \epsilon/n]$ is such that $A_n \subset A_{n+1}$ and $[0, a) := \cup_n A_n$. We will discuss many more interesting examples.

You may wish to review the meaning of countability (see Appendix **??**).

2.4 Probability Space

Putting together the observations of the sections above, we have defined a probability space as follows.

Definition 2.4.1. Probability Space

A probability space is a triplet $\{\Omega, \mathcal{F}, P\}$ where

- Ω is a nonempty set, called the *sample space*;
- $\mathcal{F}$ is a collection of subsets of Ω closed under countable set operations - such a collection is called a *σ-field*. The elements of $\mathcal{F}$ are called *events*;
- P is a countably additive function from $\mathcal{F}$ into $[0, 1]$ such that $P(\Omega) = 1$, called a *probability measure*.

Examples will clarify this definition. The main point is that one defines the probability of sets of outcomes (the events). The probability should be countably additive (to be continuous). Accordingly (to be able to write down this property), and also quite intuitively, the collection of events should be closed under countable set operations.

2.5 Examples

Throughout the course, we will make use of simple examples of probability space. We review some of those here.

2.5.1 Choosing uniformly in $\{1, 2, \ldots, N\}$

We say that we pick a value ω uniformly in $\{1, 2, \ldots, N\}$ when the N values are equally likely to be selected. In this case, the sample space Ω is $\Omega = \{1, 2, \ldots, N\}$. For any subset $A \subset \Omega$, one defines $P(A) = |A|/N$ where $|A|$ is the number of elements in A. For instance, $P(\{2, 5\}) = 2/N$.

2.5.2 Choosing uniformly in $[0, 1]$

Here, $\Omega = [0, 1]$ and one has, for example, $P([0, 0.3]) = 0.3$ and $P([0.2, 0.7]) = 0.5$. That is, $P(A)$ is the "length" of the set A. Thus, if ω is picked uniformly in $[0, 1]$, then one can write $P([0.2, 0.7]) = 0.5$.

It turns out that one cannot define the length of every subset of $[0, 1]$, as we explain in Appendix C. The collection of sets whose length is defined is the smallest σ-field that contains the intervals. This collection is called the *Borel* σ-field of $[0, 1]$. More generally, the smallest σ-field of $\Re$ that contains the intervals is the Borel σ-field of $\Re$, usually designated by $\mathcal{B}$.

2.5.3 Choosing uniformly in $[0, 1]^2$

Here, $\Omega = [0, 1]^2$ and one has, for example, $P([0.1, 0.4] \times [0.2, 0.8]) = 0.3 \times 0.6 = 0.18$. That is, $P(A)$ is the "area" of the set A. Thus, $P([0.1, 0.4] \times [0.2, 0.8]) = 0.18$. Similarly, in that case, if

$$B = \{\omega = (\omega_1, \omega_2) \in \Omega \mid \omega_1 + \omega_2 \leq 1\} \text{ and } C = \{\omega = (\omega_1, \omega_2) \in \Omega \mid \omega_1^2 + \omega_2^2 \leq 1\},$$

then

$$P(B) = \frac{1}{2} \text{ and } P(C) = \frac{\pi}{4}.$$

As in one dimension, one cannot define the area of every subset of $[0, 1]^2$. The proper σ-field is the smallest that contains the rectangles. It is called the Borel σ-field of $[0, 1]^2$. More generally, the smallest σ-field of $\Re^2$ that contains the rectangles is the Borel σ-field of $\Re^2$ designated by $\mathcal{B}^2$. This idea generalizes to $\Re^n$, with $\mathcal{B}^n$.

2.6 Summary

We have learned that a probability space is $\{\Omega, \mathcal{F}, P\}$ where Ω is a nonempty set, $\mathcal{F}$ is a σ-field of Ω, i.e., a collection of subsets of Ω that is closed under countable set operations,

and $P : \mathcal{F} \to [0, 1]$ is a σ-additive set function such that $P(\Omega) = 1$.

The idea is to specify the likelihood of various outcomes (elements of Ω). If one can specify the probability of individual outcomes (e.g., when Ω is countable), then one can choose $\mathcal{F} = 2^{\Omega}$, so that all sets of outcomes are events. However, this is generally not possible as the example of the uniform distribution on $[0, 1]$ shows. (See Appendix C.)

2.6.1 Stars and Bars Method

In many problems, we use a method for counting the number of ordered groupings of identical objects. This method is called the *stars and bars method*. Suppose we are given identical objects we call *stars*. Any ordered grouping of these stars can be obtained by separating them by *bars*. For example, $||***|*$ separates four stars into four groups of sizes 0, 0, 3, and 1.

Suppose we wish to separate N stars into M ordered groups. We need $M - 1$ bars to form M groups. The number of orderings is the number of ways of placing the N identical stars and $M - 1$ identical bars into $N + M - 1$ spaces, $\binom{N+M-1}{M}$.

Creating compound objects of stars and bars is useful when there are bounds on the sizes of the groups.

2.7 Solved Problems

Example 2.7.1. *Describe the probability space $\{\Omega, \mathcal{F}, P\}$ that corresponds to the random experiment "picking five cards without replacement from a perfectly shuffled 52-card deck."*

1. One can choose Ω to be all the permutations of $A := \{1, 2, \ldots, 52\}$. The interpretation of $\omega \in \Omega$ is then the shuffled deck. Each permutation is equally likely, so that $p_\omega = 1/(52!)$ for $\omega \in \Omega$. When we pick the five cards, these cards are $(\omega_1, \omega_2, \ldots, \omega_5)$, the top 5 cards of the deck.

2. One can also choose Ω to be all the subsets of A with five elements. In this case, each subset is equally likely and, since there are $N := \binom{52}{5}$ such subsets, one defines $p_\omega = 1/N$ for $\omega \in \Omega$.

3. One can choose $\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) \mid \omega_n \in A \text{ and } \omega_m \neq \omega_n, \forall m \neq n, m, n \in \{1, 2, \ldots, 5\}\}$. In this case, the outcome specifies the order in which we pick the cards. Since there are $M := 52!/(47!)$ such ordered lists of five cards without replacement, we define $p_\omega = 1/M$ for $\omega \in \Omega$.

As this example shows, there are multiple ways of describing a random experiment. What matters is that Ω is large enough to specify completely the outcome of the experiment.

Example 2.7.2. *Pick three balls without replacement from an urn with fifteen balls that are identical except that ten are red and five are blue. Specify the probability space.*

One possibility is to specify the color of the three balls in the order they are picked. Then

$$\Omega = \{R, B\}^3, \mathcal{F} = 2^\Omega, P(\{RRR\}) = \frac{10}{15}\frac{9}{14}\frac{8}{13}, \ldots, P(\{BBB\}) = \frac{5}{15}\frac{4}{14}\frac{3}{13}.$$

Example 2.7.3. *You flip a fair coin until you get three consecutive 'heads'. Specify the probability space.*

One possible choice is $\Omega = \{H, T\}^*$, the set of finite sequences of H and T. That is,

$$\{H, T\}^* = \cup_{n=1}^\infty \{H, T\}^n.$$

This set Ω is countable, so we can choose $\mathcal{F} = 2^\Omega$. Here,

$$P(\{\omega\}) = 2^{-n} \text{where } n := \text{length of } \omega.$$

This is another example of a probability space that is bigger than necessary, but easier to specify than the smallest probability space we need.

Example 2.7.4. *Let $\Omega = \{0, 1, 2, \ldots\}$. Let $\mathcal{F}$ be the collection of subsets of Ω that are either finite or whose complement is finite. Is $\mathcal{F}$ a σ-field?*

No, $\mathcal{F}$ is not closed under countable set operations. For instance, $\{2n\} \in \mathcal{F}$ for each $n \geq 0$ because $\{2n\}$ is finite. However,

$$A := \cup_{n=0}^{\infty}\{2n\}$$

is not in $\mathcal{F}$ because both A and A^c are infinite.

Example 2.7.5. *In a class with 24 students, what is the probability that no two students have the same birthday?*

Let $N = 365$ and $n = 24$. The probability is

$$\alpha := \frac{N}{N} \times \frac{N-1}{N} \times \frac{N-2}{N} \times \cdots \times \frac{N-n+1}{N}.$$

To estimate this quantity we proceed as follows. Note that

$$\begin{aligned} ln(\alpha) &= \sum_{k=1}^{n} ln(\frac{N-n+k}{N}) \approx \int_1^n ln(\frac{N-n+x}{N})dx \\ &= N\int_a^1 ln(y)dy = N[yln(y) - y]_a^1 \\ &= -(N-n+1)ln(\frac{N-n+1}{N}) - (n-1). \end{aligned}$$

(In this derivation we defined $a = (N-n+1)/N$.) With $n = 24$ and $N = 365$ we find that $\alpha \approx 0.48$.

Example 2.7.6. *Let A, B, C be three events. Assume that $P(A) = 0.6, P(B) = 0.6, P(C) = 0.7, P(A \cap B) = 0.3, P(A \cap C) = 0.4, P(B \cap C) = 0.4$, and $P(A \cup B \cup C) = 1$. Find $P(A \cap B \cap C)$.*

We know that (draw a picture)

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

Substituting the known values, we find

$$1 = 0.6 + 0.6 + 0.7 - 0.3 - 0.4 - 0.4 + P(A \cap B \cap C),$$

so that

$$P(A \cap B \cap C) = 0.2.$$

Example 2.7.7. *Let $\Omega = \{1, 2, 3, 4\}$ and let $\mathcal{F} = 2^{\Omega}$ be the collection of all the subsets of Ω. Give an example of a collection $\mathcal{A}$ of subsets of Ω and probability measures P_1 and P_2 such that*

(i). $P_1(A) = P_2(A), \forall A \in \mathcal{A}$.

(ii). The σ-field generated by $\mathcal{A}$ is $\mathcal{F}$. (This means that $\mathcal{F}$ is the smallest σ-field of Ω that contains $\mathcal{A}$.)

(iii). P_1 and P_2 are not the same.

Let $\mathcal{A}$= $\{\{1, 2\}, \{2, 4\}\}$.

Assign probabilities $P_1(\{1\}) = \frac{1}{8}, P_1(\{2\}) = \frac{1}{8}, P_1(\{3\}) = \frac{3}{8}, P_1(\{4\}) = \frac{3}{8}$; and $P_2(\{1\}) = \frac{1}{12}, P_2(\{2\}) = \frac{2}{12}, P_2(\{3\}) = \frac{5}{12}, P_2(\{4\}) = \frac{4}{12}$.

Note that P_1 and P_2 are not the same, thus satisfying (iii).

$P_1(\{1, 2\}) = P_1(\{1\}) + P_1(\{2\}) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$

$P_2(\{1, 2\}) = P_2(\{1\}) + P_2(\{2\}) = \frac{1}{12} + \frac{2}{12} = \frac{1}{4}$

Hence $P_1(\{1, 2\}) = P_2(\{1, 2\})$.

$P_1(\{2, 4\}) = P_1(\{2\}) + P_1(\{4\}) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$

$P_2(\{2, 4\}) = P_2(\{2\}) + P_2(\{4\}) = \frac{2}{12} + \frac{4}{12} = \frac{1}{2}$

Hence $P_1(\{2, 4\}) = P_2(\{2, 4\})$.

Thus $P_1(A) = P_2(A) \forall A \in \mathcal{A}$, thus satisfying (i).

To check (ii), we only need to check that $\forall k \in \Omega, \{k\}$ can be formed by set operations on sets in $\mathcal{A} \cup \phi \cup \Omega$. Then any other set in $\mathcal{F}$ can be formed by set operations on $\{k\}$.

$\{1\} = \{1, 2\} \cap \{2, 4\}^C$

$\{2\} = \{1,2\} \cap \{2,4\}$

$\{3\} = \{1,2\}^C \cap \{2,4\}^C$

$\{4\} = \{1,2\}^C \cap \{2,4\}.$

Example 2.7.8. *Choose a number randomly between 1 and 999999 inclusive, all choices being equally likely. What is the probability that the digits sum up to 23? For example, the number 7646 is between 1 and 999999 and its digits sum up to 23 (7+6+4+6=23).*

Numbers between 1 and 999999 inclusive have 6 digits for which each digit has a value in $\{0,1,2,3,4,5,6,7,8,9\}$. We are interested in finding the numbers $x_1+x_2+x_3+x_4+x_5+x_6 = 23$ where x_i represents the ith digit.

First consider all nonnegative x_i where each digit can range from 0 to 23, the number of ways to distribute 23 amongst the x_i's is $\binom{28}{5}$.

But we need to restrict the digits $x_i < 10$. So we need to subtract the number of ways to distribute 23 amongst the x_i's when $x_k \geq 10$ for some k. Specifically, when $x_k \geq 10$ we can express it as $x_k = 10 + y_k$. For all other $j \neq k$ write $y_j = x_j$. The number of ways to arrange 23 amongst x_i when some $x_k \geq 10$ is the same as the number of ways to arrange y_i so that $\sum_{i=1}^{6} y_i = 23 - 10$ is $\binom{18}{5}$. There are 6 possible ways for some $x_k \geq 10$ so there are a total of $6\binom{18}{5}$ ways for some digit to be greater than or equal to 10, as we can see by using the stars and bars method (see 2.6.1).

However, the above counts events multiple times. For instance, $x_1 = x_2 = 10$ is counted both when $x_1 \geq 10$ and when $x_2 \geq 10$. We need to account for these events that are counted multiple times. We can consider when two digits are greater than or equal to 10: $x_j \geq 10$ and $x_k \geq 10$ when $j \neq k$. Let $x_j = 10 + y_j$ and $x_k = 10 + y_k$ and $x_i = y_i \forall i \neq j,k$. Then the number of ways to distribute 23 amongst x_i when there are 2 greater than or equal to 10 is equivalent to the number of ways to distribute y_i when $\sum_{i=1}^{6} y_i = 23 - 10 - 10 = 3$. There are $\binom{8}{5}$ ways to distribute these y_i and there are $\binom{6}{2}$ ways to choose the possible two digits that are greater than or equal to 10.

We are interested in when the sum of x_i's is equal to 23. So we can have at most 2 x_i's greater than or equal to 10. So we are done.

Thus there are $\binom{28}{5} - 6\binom{18}{5} + \binom{6}{2}\binom{8}{5}$ numbers between 1 through 999999 whose digits sum up to 23. The probability that a number randomly chosen has digits that sum up to 23 is $\frac{\binom{28}{5}-6\binom{18}{5}+\binom{6}{2}\binom{8}{5}}{999999}$.

Example 2.7.9. *Let $A_1, A_2, \ldots, A_n, n \geq 2$ be events. Prove that $P(\cup_{i=1}^n A_i) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n)$.*

We prove the result by induction on n.

First consider the base case when $n = 2$. $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$.

Assume the result holds true for n, prove the result for $n + 1$.

$$\begin{aligned}
P(\cup_{i=1}^{n+1} A_i) &= P(\cup_{i=1}^n A_i) + P(A_{n+1}) - P((\cup_{i=1}^n A_i) \cap A_{n+1}) \\
&= P(\cup_{i=1}^n A_i) + P(A_{n+1}) - P(\cup_{i=1}^n (A_i \cap A_{n+1})) \\
&= \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots \\
&\qquad + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n) + P(A_{n+1}) - (\sum_i P(A_i \cap A_{n+1}) \\
&\qquad - \sum_{i<j} P(A_i \cap A_j \cap A_{n+1}) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k \cap A_{n+1}) - \ldots \\
&\qquad + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n \cap A_{n+1})) \\
&= \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots \\
&\qquad + (-1)^{n+2} P(A_1 \cap A_2 \cap \ldots \cap A_{n+1})
\end{aligned}$$

Example 2.7.10. *Let $\{A_n, n \geq 1\}$ be a collection of events in some probability space $\{\Omega, \mathcal{F}, P\}$. Assume that $\sum_{n=1}^{\infty} P(A_n) < \infty$. Show that the probability that infinitely many of those events occur is zero. This result is known as the* Borel-Cantelli Lemma.

To prove this result we must write the event "infinitely many of the events A_n occur"

in terms of the A_n. This event can be written as

$$A = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n.$$

To see this, note that ω is in infinitely many A_n if and only if for all $m \geq 1$ there is some $n \geq m$ such that $\omega \in A_n$.

It follows from this representation of A that $B_m \downarrow A$ where $B_m := \cup_{n=m}^{\infty} A_n$. Now, because of the σ-additivity of $P(\cdot)$, we know that $P(B_m) \downarrow P(A)$. But

$$P(B_m) \leq \sum_{n=m}^{\infty} P(A_n).$$

Also, since $\sum_{n=1}^{\infty} P(A_n) < \infty$, we know that $\sum_{n=m}^{\infty} P(A_n) \downarrow 0$ as $m \to \infty$. Consequently, $P(A) = \lim_m P(B_m) = 0$.

Chapter 3

Conditional Probability and Independence

The theme of this chapter is how to use observations. The key idea is that observations modify our belief about the likelihood of events. The mathematical notion of conditional probability formalizes that idea.

3.1 Conditional Probability

Assume that we know that the outcome is in $B \subset \Omega$. Given that information, what is the probability that the outcome is in $A \subset \Omega$? This probability is written $P[A|B]$ and is read "the conditional probability of A given B," or "the probability of A given B", for short.

For instance, one picks a card at random from a 52-card deck. One knows that the card is black. What is the probability that it is the ace of clubs? The sensible answer is that if one only knows that the card is black, then that card is equally likely to be any one of the 26 black cards. Therefore, the probability that it is the ace of clubs is 1/26. Similarly, given that the card is black, the probability that it is an ace is 2/26, because there are 2 black aces (spades and clubs).

We can formulate that calculation as follows. Let A be the set of aces (4 cards) and B the set of black cards (26 cards). Then, $P[A|B] = P(A \cap B)/P(B) = (2/52)(26/52) = 2/26$.

Indeed, for the outcome to be in A, given that it is in B, that outcome must be in $A \cap B$. Also, given that the outcome is in B, the probability of all the outcomes in B should be renormalized so that they add up to 1. To renormalize these probabilities, we divide them by $P(B)$. This division does not modify the relative likelihood of the various outcomes in B.

More generally, we define the probability of A given B by

$$P[A|B] = \frac{P(A \cap B)}{P(B)}.$$

This definition of conditional probability makes sense if $P(B) > 0$. If $P(B) = 0$, we define $P[A|B] = 0$. This definition is somewhat arbitrary but it makes the formulas valid in all cases.

Note that

$$P(A \cap B) = P[A \mid B]P(B).$$

3.2 Remark

Define $P'(A) = P[A|B]$ for any event A. Then $P'(\cdot)$ is a new probability measure. In particular, the usual formulas apply. For instance, $P'[A \cap C] = P'[A|C]P'(C)$, i.e.,

$$P[A \cap C|B] = P[A|B \cap C]P[C|B],$$

which you can verify by using the definition of $P[\cdot|B]$. After a while, you should be able to write expressions such as the one above by thinking of $P[\cdot \mid B]$ as a new probability.

3.3 Bayes' Rule

Let B_1 and B_2 be disjoint events whose union is Ω. Let also A be another event. We can write

$$P[B_1|A] = \frac{P(B_1 \cap A)}{P(A)} = \frac{P[A|B_1]P(B_1)}{P(A)},$$

and

$$P(A) = P(B_1 \cap A) + P(B_2 \cap A) = P[A|B_1]P(B_1) + P[A|B_2]P(B_2).$$

Hence,

$$P[B_1|A] = \frac{P[A|B_1]P(B_1)}{P[A|B_1]P(B_1) + P[A|B_2]P(B_2)}.$$

This formula extends to a finite number of events B_n that partition Ω. The result is know as Bayes' rule. Think of the B_n as possible "causes" of some effect A. You know the prior probabilities $P(B_n)$ of the causes and also the probability that each cause provokes the effect A. The formula tells you how to calculate the probability that a given cause provoked the observed effect. Applications abound, as we will see in detection theory. For instance, you alarm can sound either if there is a burglar or also if there is no burglar (false alarm). Given that the alarm sounds, what is the probability that it is a false alarm?

3.4 Independence

It may happen that knowing that an event occurs does not change the probability of another event. In that case, we say that the events are independent. Let us look at an example first.

3.4.1 Example 1

We roll two dice and we designate the pair of results by $\omega = (\omega_1, \omega_2)$. Then Ω has 36 elements: $\omega = \{(\omega_1, \omega_2) | \omega_1 = 1, \ldots, 6 \text{ and } \omega_2 = 1, \ldots, 6\}$. Each of these elements has probability 1/36. Let $A = \{\omega \in \Omega | \omega_1 \in \{1, 3, 4\}\}$ and $B = \{\omega \in \Omega | \omega_2 \in \{3, 5\}\}$. Assume that we know that the outcome is in B. What is the probability that it is in A?

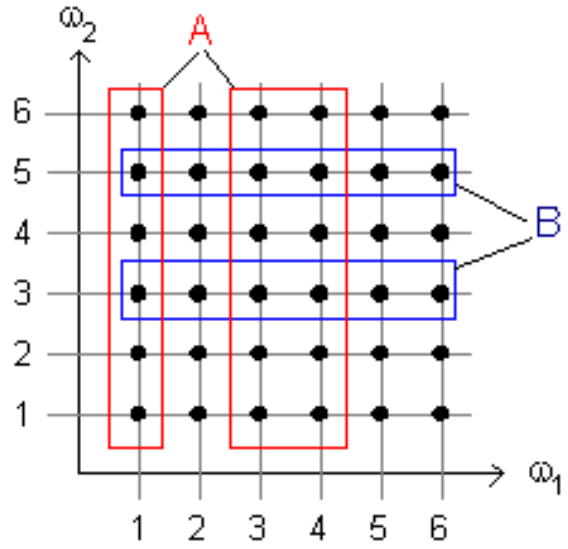


Figure 3.1: Rolling two dice

Using the conditional probability formula, we find $P[A|B] = P(A \cap B)/P(B) = (6/36)/(12/36) = 1/2$. Note also that $\mathrm{P}(A) = 18/36 = 1/2$. Thus, in this example, $P[A|B] = P(A)$.

The interpretation is that if we know the outcome of the second roll, we dont know anything about the outcome of the first roll.

3.4.2 Example 2

We pick two points independently and uniformly in $[0, 1]$. In this case, the outcome $\omega = (\omega_1, \omega_2)$ of the experiment (the pair of points chosen) belongs to the set $\Omega = [0, 1]^2$. That point ω is picked uniformly in $[0, 1]^2$. Let $A = [0.2, 0.5] \times [0, 1]$ and $B = [0, 1] \times [0.2, 0.8]$. The interpretation of A is that the first point is picked in $[0.2, 0.5]$; that of B is that the second point is picked in $[0.2, 0.8]$. Note that $P(A) = 0.3$ and $P(B) = 0.6$. Moreover, since $A \cap B = [0.2, 0.5] \times [0.2, 0.8]$, one finds that $P(A \cap B) = 0.3 \times 0.6 = P(A)P(B)$. Thus, A and B are independent events.

3.4.3 Definition

Motivated by the discussion above, we say that two events A and B are *independent* if

$$P(A \cap B) = P(A)P(B).$$

Note that the independence is a notion that depends on the probability.

Do not confuse "independent" and "disjoint." If two events A and B are disjoint, then they are independent only if at least one of them has probability 0. Indeed, if they are disjoint, $P(A \cap B) = P(\emptyset) = 0$, so that $P(A \cap B) = P(A)P(B)$ only if $P(A) = 0$ or $P(B) = 0$. Intuitively, if A and B are disjoint, then knowing that A occurs implies that B does not, which is some new information about B unless B is impossible in the first place.

3.4.4 General Definition

Generally, we say that a collection of events $\{A_i, i \in I\}$ are mutually independent if for any finite subcollection $\{i, j, \ldots, k\} \subset I$ one has

$$P(A_i \cap A_j \cap \cdots \cap A_k) = P(A_i)P(A_j) \cdots P(A_k).$$

Subtility

The definition seems innocuous, but one has to be a bit careful. For instance, look at the example illustrated in Figure 3.2.

The sample space $\Omega = \{1, 2, 3, 4\}$ has four points that have a probability $1/4$ each. The events A, B, C are defined as $A = \{1, 2\}, B = \{1, 3\}, C = \{2, 3\}$. We can verify that A and B are independent. Indeed, $P(A \cap B) = 1/4 = P(A)P(B)$. Similarly, A and C are independent and so are B and C. However, the events $\{A, B, C\}$ are not mutually independent. Indeed, $P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C) = 1/8$.

The point of the example is the following. Knowing that A has occurred tells us something about outcome ω of the random experiment. This knowledge, by itself, is not sufficient

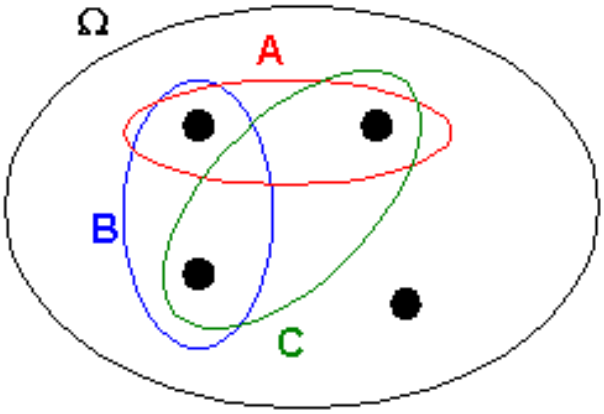


Figure 3.2: Pairwise but not mutual independence

to affect our estimate of the probability that C has occurred. The same is true is we know that B has occurred. However, if we know that both A and B have occurred, then we know that C cannot have occurred. Thus, it is not correct to think that "A does not tell us anything about C, B does not tell us anything about C, therefore A and B do not tell us anything about C." I encourage you to think about this example carefully.

3.5 Summary

We explained the definition of conditional probability $P[A \mid B] = P(A \cap B)/P(B)$ and that of independence of two events and of mutual independence of a collection of events.

Pairwise independence does not imply mutual independence.

3.6 Solved Problems

Example 3.6.1. *Is it true that*

$$P(A \cap B \cap C) = P[A \mid B]P[B \mid C]P(C)?$$

If true, provide a proof; if false, provide a counterexample.

That identity is false. Here is one counterexample. Let $\Omega = \{1, 2, 3, 4\}$ and $p_\omega = 1/4$ for $\omega \in \Omega$. Choose $A = \{1, 2\}, B = \{2, 3\}, C = \{3, 4\}$. Then $P(A \cap B \cap C) = 0, P[A \mid B] = P[B \mid C] = P(C) = 1/2$, so that the identity does not hold.

Example 3.6.2. *There are two coins. The first coin is fair. The second coin is such that* $P(H) = 0.6 = 1 - P(T)$. *You are given one of the two coins, with equal probabilities between the two coins. You flip the coin four times and three of the four outcomes are* H. *What is the probability that your coin is the fair one?*

Let A designate the event "your coin is fair." Let also B designate the event "three of the four outcomes are H."

Bayes' rule implies

$$\begin{aligned} P[A|B] &= \frac{P(A \cap B)}{P(A)} = \frac{P[B|A]P(A)}{P[B|A]P(A) + P[B|A^c]P(A^c)} \\ &= \frac{\binom{4}{3}(1/2)^4}{\binom{4}{3}(1/2)^4 + \binom{4}{3}(0.6)^3(0.4)} = \frac{2^{-4}}{2^{-4} + (0.6)^3 0.4}. \end{aligned}$$

Example 3.6.3. *Choose two numbers uniformly but without replacement in* $\{0, 1, \ldots, 10\}$. *What is the probability that the sum is less than or equal to 10 given that the smallest is less than or equal to 5?*

Draw a picture of

$$\Omega = \{0, 1, \ldots, 10\}^2 \setminus \{(i, i) \mid i = 0, 1, \ldots, 10\}.$$

The outcomes in Ω all have the same probability. Let also

$$A = \{\omega \mid \omega_1 \neq \omega_2 \text{ and } \omega_1 + \omega_2 \leq 10\}, B = \{\omega \mid \omega_1 \neq \omega_2 \text{ and } \min\{\omega_1, \omega_2\} \leq 5\}.$$

The probability we are looking for is

$$\frac{|A \cap B|}{|B|} = \frac{|A|}{|B|}.$$

Your picture shows that $|A| = 10+9+8+\cdots+1 = 55$ and that $|B| = 10\times 5+4\times 5 = 70$. Hence, the answer is $55/70 = 11/14$.

Example 3.6.4. *You flip a fair coin repeatedly. What is the probability that you have to flip it exactly 10 times to see two "heads"?*

There must be exactly one head among the first nine flips and the last flip must be another head. The probability of that event is

$$9(\frac{1}{2})^9 \times (\frac{1}{2}) = \frac{9}{2^{10}}.$$

Example 3.6.5. *a. Let A and B be independent events. Show that A^C and B are independent.*

b. Let A and B be two events. If the occurrence of event B makes A more likely, then does the occurrence of the event A make B more likely? Justify your answer.

c. If event A is independent of itself, show that $P(A)$ is 1 or 0.

d. If $P(A)$ is 1 or 0, show that A is independent of all events B.

a. We have

$$\begin{aligned} P(A^C \cap B) &= P(B \setminus \{A \cap B\}) = P(B) - P(A \cap B) \\ &= P(B) - P(A)P(B) = P(A^C)P(B). \end{aligned}$$

Hence A^C and B are independent.

b. The occurrence of event B makes A more likely can be interpreted as $P[A|B] > P(A)$. Now,

$$P[A|B]) = \frac{P(A \cap B)}{P(B)} = \frac{P[B|A]P(A)}{P(B)} > P(A).$$

Hence $\frac{P[B|A]}{P(B)} > 1$ and $P[B|A] > P(B)$. Thus the occurrence of event A makes B more likely.

c. If A is independent of itself, then $P(A) = P(A \cap A) = P(A)^2$. Hence $P(A) = 0$ or 1.

d. Suppose $P(A) = 0$. Then $P(A)P(B) = 0, \forall B$. Now, $A \cap B \subseteq A$, so $0 \leq P(A \cap B) \leq P(A) = 0$. Hence $P(A \cap B) = 0$.

Suppose $P(A) = 1$. Then $P(A \cap B) = P(B)$, so that $P(A \cap B) = P(A)P(B)$.

Example 3.6.6. *A man has 5 coins in his pocket. Two are double-headed, one is double-tailed, and two are normal. The coins cannot be distinguished unless one looks at them.*

a. The man shuts his eyes, chooses a coin at random, and tosses it. What is the probability that the lower face of the coin is heads?

b. He opens his eyes and sees that the upper face of the coin is a head. What is the probability that the lower face is a head.

c. He shuts his eyes again, picks up the same coin, and tosses it again. What is the probability that the lower face is a head?

d. He opens his eyes and sees that the upper face is a head. What is the probability that the lower face is a head?

Let D denote the event that he picks a double-headed coin, N denote the event that he picks a normal coin, and Z be the event that he picks the double-tailed coin. Let H_{L_i} (and H_{U_i}) denote the event that the lower face (and the upper face) of the coin on the ith toss is a head.

a. One has

$$\begin{aligned} P(H_{L_1}) &= P[H_{L_1}|D]P(D) + P[H_{L_1}|N]P(N) + P[H_{L_1}|Z]P(Z) \\ &= (1)(\frac{2}{5}) + (\frac{1}{2})(\frac{2}{5}) + (0)(\frac{1}{5}) = \frac{3}{5}. \end{aligned}$$

b. We find

$$P[H_{L_1}|H_{U_1}] = \frac{P(H_{L_1} \cap H_{U_1})}{P(H_{U_1})} = \frac{\frac{2}{5}}{\frac{3}{5}} = \frac{2}{3}.$$

c. We write

$$\begin{aligned} P[H_{L_2}|H_{U_1}] &= \frac{P(H_{L_2} \cap H_{U_1})}{P(H_{U_1})} \\ &= \frac{P[H_{L_2} \cap H_{U_1}|D]P(D) + P[H_{L_2} \cap H_{U_1}|N]P(N) + P[H_{L_2} \cap H_{U_1}|Z]P(Z)}{P(H_{U_1})} \\ &= \frac{(1)(\frac{2}{5}) + (\frac{1}{4})(\frac{2}{5}) + (0)(\frac{1}{5})}{\frac{3}{5}} = \frac{5}{6}. \end{aligned}$$

d. Similarly,

$$\begin{aligned} P[H_{L_2}|H_{U_1} \cap H_{U_2}] &= \frac{P(H_{L_2} \cap H_{U_1} \cap H_{U_2})}{P(H_{U_1} \cap H_{U_2})} \\ &= \frac{(1)(\frac{2}{5}) + (0)(\frac{2}{5}) + (0)(\frac{1}{5})}{(1)(\frac{2}{5}) + (\frac{1}{4})(\frac{2}{5}) + (0)(\frac{1}{5})} = \frac{4}{5}. \end{aligned}$$

Chapter 4

Random Variable

In this chapter we define a random variable and we illustrate the definition with examples. We then define the expectation and moments of a random variable. We conclude the chapter with useful inequalities.

A random variable takes real values. The definition is "a random variable is a measurable real-valued function of the outcome of a random experiment."

Mathematically, one is given a probability space and some function $X : \Omega \to \Re := (-\infty, +\infty)$. If the outcome of the random experiment is ω, then the value of the random variable is $X(\omega) \in \Re$.

Physical examples: noise voltage at a given time and place, temperature at a given time and place, height of the next person to enter the room, and so on. The color of a randomly picked apple is not a random variable since its value is not a real number.

4.1 Measurability

An arbitrary real-valued function defined on Ω is not necessarily a random variable.

For instance, let $\Omega = [0, 1]$ and $A = [0, 0.5]$. Assume that the events are $[0, 1]$, $[0, 0.5]$, $(0.5, 1]$, and $\emptyset$. For instance, assume that we have defined $P([0, 0.5]) = 0.73$ and that this is all we know. Consider the function $X(\omega)$ that takes the value 0 when ω is in $[0, 0.3]$ and the

value 1 when ω is in $(0.3, 1]$. This function is not a random variable. We cannot determine $P(X = 0)$ from the information we have. Accordingly, the statistical properties of X are not defined. This is what we mean by measurability. Thus, measurability is not a subtle notion. It is a first order idea: What are the functions whose statistics are defined by the model? These are the measurable functions.

Let $\mathcal{F}$ be a collection of events of Ω. (Recall that $\mathcal{F}$ is closed under countable set operations.)

Definition 4.1.1. A function $X : \Omega \to \Re$ is $\mathcal{F}$-measurable if $X^{-1}((-\infty, a]) \in \mathcal{F}$ for all $a \in \Re$. Thus, we can define $P(X \leq a)$ for all $a \in \Re$.

Equivalently, the function is $\mathcal{F}$-measurable if and only if

$$X^{-1}(B) \in \mathcal{F}, \forall B \in \mathcal{B}$$

where $\mathcal{B}$ is the Borel σ-field of $\Re$, i.e., the smallest σ-field that contains the intervals.

4.2 Distribution

A random variable X is *discrete* if it takes values in a countable set $\{x_n, n \geq 1\} \subset \Re$. We can define $P(X = x_n) = p_n$ with $p_n > 0$ and $\sum_n p_n = 1$. The collection $\{x_n, p_n, n \geq 1\}$ is then called the *Probability Mass Function* (pmf) of the random variable X.

In general, the function $\{P(X \leq x) =: F(x), x \in \Re\}$ – called the *cumulative distribution function* (cdf) of X - completely characterizes the "statistics" of X. For short, we also call the cdf the *distribution* of X.

A function $F(\cdot) : \Re \to \Re$ is the cdf of some random variable if and only if

$$
\begin{aligned}
&lim_{x\to-\infty}F_X(x) = 0;\\
&lim_{x\to\infty}F_X(x) = 1;\\
&\text{If } x < y, \text{ then } F_X(x) \le F_X(y);\\
&F_X(x) \text{ is right-continuous. That is, } y \downarrow x \text{ implies } F(y) \to F(x).
\end{aligned}
\tag{4.2.1}
$$

These properties follow directly from the continuity property of probability: If $A_n \downarrow A$, then $P(A_n) \downarrow P(A)$. For instance, since $(-\infty, x) \downarrow \emptyset$ as $x \downarrow -\infty$, it follows that $F_X(x) \downarrow 0$ as $x \downarrow -\infty$. The fact that a function with these properties is a cdf can be seen from the construction we explain in Section 4.4.

Note also that if $x_n \uparrow x$, then

$$P(X < x) = P(\cup_{n\ge1}\{X \le x_n\}) = \lim_{n\to\infty} P(X \le x_n) = \lim_{n\to\infty} F_X(x_n) = F_X(x-) \tag{4.2.2}$$

where $F_X(x-) := \lim_{y\uparrow x} F_X(y)$. Consequently,

$$P(X = x) = P(\{X \le x\} \setminus \{X < x\}) = P(X \le x) - P(X < x) = F_X(x) - F_X(x-).$$

A random variable X is *continuous* if one can write

$$P(X \in (a, b]) = F_X(x) = \int_a^b f_X(x)dx \tag{4.2.3}$$

for all real numbers $a < b$. In this expression, $f_X(\cdot)$ is a nonnegative function called the *probability density function* (pdf) of X. The pdf $f_X(\cdot)$ is the derivative of the cdf $F_X(\cdot)$.

Obviously, a discrete random variable is not continuous. Also, a random variable may be neither discrete nor continuous.

The function $F_X(\cdot)$ may jump at discrete points x_n by $a_n = F_X(x_n) - F_X(x_n-) = P(X = x_n)$. In that case, we define the pdf as a "formal" derivative $f_X(x)$ as follows:

$$f_X(x) := g(x) + \sum_n a_n\delta(x - x_n) \tag{4.2.4}$$

where $g(\cdot)$ is the derivative of $F_X(\cdot)$ where it is differentiable and $\delta(x-x_n)$ is a *Dirac impulse* at a jump x_n. The formal definition of a Dirac impulse $\delta(x-x_0)$ is that

$$\int_a^b g(x)\delta(x-x_0)dx = g(x_0)1\{a \le x_0 \le b\}$$

whenever $g(\cdot)$ is a function that is continuous at x_0. With this formal definition, you can see that (4.2.3) holds in the general case.

4.3 Examples of Random Variable

We say that the random variable X has a *Bernoulli* distribution with parameter $p \in [0,1]$, and we write $X =_D B(p)$ if

$$P(X=1) = p \text{ and } P(X=0) = 1-p. \tag{4.3.1}$$

The random variable X has a *binomial* distribution with parameters $n \in \{1,2,\ldots\}$ and $p \in [0,1]$, and we write $X =_D B(n,p)$, if

$$P(X=m) = \binom{n}{m} p^m (1-p)^{n-m}, \text{ for } m = 0,1,\ldots,n. \tag{4.3.2}$$

The random variable X has a *geometric* distribution with parameter $p \in (0,1]$, and we write $X =_D G(p)$, if

$$P(X=n) = p(1-p)^{n-1}, \text{ for } n \ge 0. \tag{4.3.3}$$

A random variable X with a geometric distribution has the remarkable property of being *memoryless*. That is,

$$P[X \ge n+m \mid X \ge n] = P(X \ge m), \forall\ m,n \ge 0.$$

Indeed, from (4.3.3) one finds that $P(X \ge n) = (1-p)^n$, so that

$$P[X \ge n+m \mid X \ge n] = \frac{P(X \ge n+m)}{P(X \ge n)} = \frac{(1-p)^{n+m}}{(1-p)^n} = (1-p)^m = P(X \ge m). \tag{4.3.4}$$

The interpretation is that if X is the lifetime of a light bulb (in years, say), then the residual lifetime $X - n$ of that light bulb, if it is still alive after n years, has the same distribution as that of a new light bulb. Thus, if light bulbs had a geometrically distributed lifetime, preventive replacements would be useless.

The random variable X has a *Poisson distribution* with parameter $\lambda > 0$, and we write $X =_D P(\lambda)$, if

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}, \text{ for } n \geq 0. \tag{4.3.5}$$

The random variable X is uniformly distributed in the interval $[a, b]$ where $a < b$, and we write $X =_D U[a, b]$ if

$$f_X(x) = \begin{cases} \frac{1}{b-a}, \text{ if } x \in [a, b] \\ 0, \text{ otherwise.} \end{cases} \tag{4.3.6}$$

The random variable X is exponentially distributed with rate $\lambda > 0$, and we write $X =_D Exd(\lambda)$, if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, \text{ if } x > 0 \\ 0, \text{ otherwise.} \end{cases} \tag{4.3.7}$$

Exponentially distributed random variables are memoryless in the sense that

$$P[X > s + t \mid X > t] = P(X > s), \forall\ s, t > 0. \tag{4.3.8}$$

Indeed, from (4.3.7) one finds that $P(X > t) = e^{-\lambda t}$, so that

$$P[X > s + t \mid X > t] = \frac{P(X > s + t)}{P(X > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s).$$

The interpretation of this property is the same as for the geometric distribution.

We discuss Gaussian random variables later.

4.4 Generating Random Variables

Methods to generate a random variable X with a given distribution from uniform random variables are useful in simulations and provide a good insight into the meaning of the cdf and of the pdf.

The first method is to generate a random variable Z uniform in $[0,1]$ (using the random number generator of your computer) and to define $X(Z) = \min\{a|F(a) \geq Z\}$. Then, for any real number b, $X \leq b$ if $Z \leq F(b)$, which occurs with probability $F(b)$, as desired.

The second method uses the pdf. Assume that X is continuous with pdf $f(x)$ and that $P(a < X < b) = 1$ and $f(x) \leq c$. Pick a point (X, Y) uniformly in $[a, b] \times [0, c]$ (by generating two uniform random variables). If the point falls under the curve $f(\cdot)$, i.e., if $Y \leq f(X)$, then keep the value X; otherwise, repeat. Then, $P(a < X < a + \epsilon) = P[A|B]$ where $A = \{(x, y)|a < x < a + \epsilon, y < f(x)\} = [a, a + \epsilon] \times [0, f(a)]$. Then, $P[A|B] = P(A)/P(B)$ with $P(A) = f(a)\epsilon/[(ba)c]$ and $P(B) = 1/[(ba)c]$. Hence $P[A|B] = f(a)\epsilon$, as desired. (The factor $1/[(ba)c]$ is to normalize our uniform distribution on $[a, b] \times [0, c]$ and the term 1 in the numerator of $P(B)$ comes from the fact that the p.d.f. $f(\cdot)$ integrates to 1.

4.5 Expectation

Imagine that you play a game of chance a large number K of times. Each time you play, you have some probability p_n of winning x_n, for $n \geq 1$. If our interpretation of probability is correct, you expect to win x_n approximately Kp_n times out of K. Accordingly, your total earnings should be approximately equal to $\sum_n Kx_np_n$. Thus, your earnings should average $\sum_n x_np_n$ per instance of the game. That value, the average earnings per experiment, is the *interpretation* of the expected value of the random variable X that represents your earnings.

Formally, we *define* the *expected value* $E(X)$ of a random variable X as follows.

Definition 4.5.1. Expected Value

For a discrete random variable, $E(X) = \sum_n x_np_n$.

For a continuous random variable, $E(X) = \int_{-\infty}^{\infty} xf_X(x)dx$.

For a general random variable with (4.2.4),

$$E(X) = \int_{-\infty}^{\infty} xdF_X(x) := \sum_n x_na_n + \int_{-\infty}^{\infty} xg(x)dx. \tag{4.5.1}$$

There are some potential problems. The sums could yield $\infty - \infty$. In that case, we say that the expectation of the random variable is not defined.

4.6 Function of Random Variable

Let X be a random variable and $h : \Re \to \Re$ be a function. Since X is some function from Ω to $\Re$, so is $h(X)$. Is $h(X)$ a random variable? Well, there is that measurability question. We must check that $h(X)^{-1}((-\infty, a]) \in \mathcal{F}$ for some $a \in \Re$. That property holds if $h(\cdot)$ is *Borel*-measurable, as defined below.

Definition 4.6.1. A function $h : \Re \to \Re$ is *Borel-measurable* if

$$h^{-1}(B) \in \mathcal{B}, \forall B \in \mathcal{B}.$$

All the functions from $\Re$ to $\Re$ we will encounter are Borel-measurable.

Using Definition 4.1.1 we see that if h is Borel-measurable, then, for all $B \in \mathcal{B}$, one has $A = h^{-1}(B) \in \mathcal{B}$, so that $(h(X))^{-1}(B) = X^{-1}(A) \in \mathcal{F}$, which proves that $h(X)$ is a random variable.

In some cases, if X has a pdf, $Y = h(X)$ *may* also have one. For instance, if X has pdf $f_X(.)$ and $Y = aX + b$ with $a > 0$, then

$$P(y < Y < y + dy) = P((y - b)/a < X < (y - b)/a + dy/a) = f_X((y - b)/a)dy/a,$$

so that the pdf of Y, say $f_Y(y)$, is $f_Y(y) = f_X((y-b)/a)/a$. We highlight that useful result below:

$$\text{If } Y = aX + b, \text{ then } f_Y(y) = \frac{1}{|a|} f_X(\frac{y - b}{a}). \tag{4.6.1}$$

By adapting the above argument, you can check that if $h(\cdot)$ is differentiable and one-to-one, then the density of $Y = h(X)$ is equal to

$$f_Y(y) = \frac{1}{|h'(x)|} f_X(x)|_{h(x)=y}. \tag{4.6.2}$$

For instance, if $X =_D U[0,1]$ and $Y = (3+X)^2$, then

$$f_Y(y) = \frac{1}{2(3+x)} 1\{x \in [0,1]\}|_{(3+x)^2=y} = \frac{1}{2\sqrt{y}} 1\{y \in [\sqrt{3}, 2]\}.$$

How do we compute $E(h(X))$? If X is discrete, then so is $h(X)$. One could then look at the values y_n of $Y = h(X)$ and their probabilities, say $q_n = P(Y = y_n)$. Then $E(h(X)) = \sum_n y_n q_n$. There is a clever observation that is very useful. That observation is that

$$E(h(X)) = \sum_n h(x_n) p_n \text{ where } p_n = P(X = x_n).$$

If you think about it, it looks a bit like magic. However, it is not hard to understand what is going on. This expression is useful because you do not have to calculate the probability of the different values of Y.

If X has pdf $f_X(\cdot)$, then

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x) dx.$$

Again, you do not have to calculate the pdf of Y. For instance, with $X =_D U[0,1]$ and $h(X) = (3+X)^2 =: Y$, we find

$$E(h(X)) = \int_0^1 (3+x)^2 dx = [\frac{1}{3}(3+x)^3]_0^1 = \frac{1}{3}[4^3 - 3^3] = \frac{37}{3}.$$

A calculation based on f_Y gives

$$E(h(X)) = E(Y) = \int_{\sqrt{3}}^2 y f_Y(y) dy = \int_{\sqrt{3}}^2 y \frac{1}{2\sqrt{y}} dy.$$

If we do the change of variables $y = (3+x)^2$, so that $dy = 2(3+x)dx = 2\sqrt{y}dx$, then we can write the integral above as

$$E(Y) = \int_0^1 (3+x)^2 dx,$$

which hopefully explains the magic.

In the general case, if $F_X(\cdot)$ is the cdf of X (see (4.2.4), then

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)dF_X(x) := \sum_n h(x_n)a_n + \int_{-\infty}^{\infty} h(x)g(x)dx.$$

Note for instance that the above identities show that

$$E(a_1h_1(X) + a_2h_2(X)) = a_1E(h_1(X)) + a_2E(h_2(X)).$$

Accordingly, if we define $Y = h_1(X)$ and $Z = h_2(X)$, we conclude that

$$E(a_1Y + a_2Z) = a_1E(Y) + a_2E(Z). \tag{4.6.3}$$

This property is called the *linearity of expectation*.

Examples will help you understand this section well.

4.7 Moments of Random Variable

The n-th moment of X is $E(X^n)$.

The *variance* of X is

$$\text{var}(X) := E(X - E(X))^2 = E(X^2) - \{E(X)\}^2.$$

The variance measures the "spread" of the distribution around the mean. A random variable with a zero variance is constant. The larger the variance, the more "uncertain" the random variable is, in the mean square sense. Note that I say "uncertain" and not "variable" since you know by now that a random variable does not vary.

4.8 Inequalities

Inequalities are often useful to estimate some expected values. Here are a few particularly useful ones.

Exponential bound:

$$1 + x \leq \exp\{x\}.$$

Chebychev:

$$P(X \leq a) \leq E(X^2)/a^2. \tag{4.8.1}$$

Markov Inequality: If $f(\cdot)$ is nondecreasing and nonnegative on $[a, \infty)$, then

$$P(X \leq a) \leq \{E(f(X))\}/f(a). \tag{4.8.2}$$

Jensen: If $f(\cdot)$ is convex, then

$$E(f(X)) \geq f(E(X)). \tag{4.8.3}$$

4.9 Summary

A random variable is a function $X : \Omega \to \Re$ such that $X^{-1}((-\infty, x]) \in \mathcal{F}$ for all $x \in \Re$. We then define the cdf $F_X(.)$ of X as

$$F_X(x) = P(X \leq x) := P(X^{-1}((-\infty, x])), x \in \Re.$$

We also defined the pmf and pdf that summarize the distribution of the random variable.

We defined the expected value and explained that

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) dF_X(x).$$

We introduced the moments and the variance and we stated a few useful inequalities.

You should become familiar with the distributions we introduced. We put a table that summarizes those distributions in Appendix G, for ease of reference.

4.10 Solved Problems

Example 4.10.1. *A random variable X has pdf $f_X(\cdot)$ where*

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

a. Find c;

b. Find $P(\frac{1}{2} < X \le \frac{3}{4})$;

c. Find the cdf $F_X(\cdot)$ of X;

d. Calculate $E(X)$ and var(X).

a. Need to have $\int_{-\infty}^{\infty} f_X(x)dx = 1$. Now,

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 cx(1-x)dx = c\int_0^1 (x-x^2)dx = c(\frac{x^2}{2} - \frac{x^3}{3})]_0^1 = c(\frac{1}{2} - \frac{1}{3}) = \frac{c}{6}.$$

Thus $c = 6$.

b. We find

$$\begin{aligned} P(\frac{1}{2} \le X \le \frac{3}{4}) &= \int_{\frac{1}{2}}^{\frac{3}{4}} f_X(x)dx = \int_{\frac{1}{2}}^{\frac{3}{4}} 6x(1-x)dx = 6(\frac{x^2}{2} - \frac{x^3}{3})]_{\frac{1}{2}}^{\frac{3}{4}} \\ &= 6(\frac{9}{32} - \frac{9}{64} - \frac{1}{8} + \frac{1}{24}) = \frac{11}{32} \end{aligned}$$

c. The cdf is $F_X(x) = P(X \le x)$. For $x < 0$, $F_X(x) = 0$. For $0 \le x < 1$, $F_X(x) = \int_0^x 6y(1-y)dy = 3x^2 - 2x^3$. For $x \ge 1$, $F_X(x) = 1$. Hence the cdf is

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0; \\ 3x^2 - 2x^3, & \text{if } 0 \le x < 1; \\ 1, & \text{if } x \ge 1. \end{cases}$$

d. We find

$$E(X) = \int_{-\infty}^{\infty} xf_X(x)dx = \int_0^1 x6x(1-x)dx = [2x^3]_0^1 - [\frac{3}{2}x^4]_0^1 = 2 - \frac{3}{2} = 0.5,$$

which is not surprising since $f_X(\cdot)$ is symmetric around 0.5. Also,

$$E(X^2) = \int_0^1 x^2 6x(1-x)dx = [\frac{3}{2}x^4]_0^1 - [\frac{6}{5}x^5]_0^1 = \frac{3}{2} - \frac{6}{5} = \frac{3}{10}.$$

Hence,

$$\text{var}(X) = E(X^2) - (E(X))^2 = \frac{3}{10} - (\frac{1}{2})^2 = \frac{1}{20}.$$

Example 4.10.2. *Give an example of a probability space and a real-valued function on* Ω *that is* not *a random variable.*

Let $\Omega = \{0,1,2\}, \mathcal{F} = \{\emptyset, \{0\}, \{1,2\}, \Omega\}, P(\{0\}) = 1/2 = P(\{1,2\})$, and $X(\omega) = \omega$ for $\omega \in \Omega$. The function X is not a random variable since

$$X^{-1}((-\infty, 1]) = \{0,1\} \notin \mathcal{F}.$$

The meaning of all this is that the probability space is not rich enough to specify $P(X \leq 1)$.

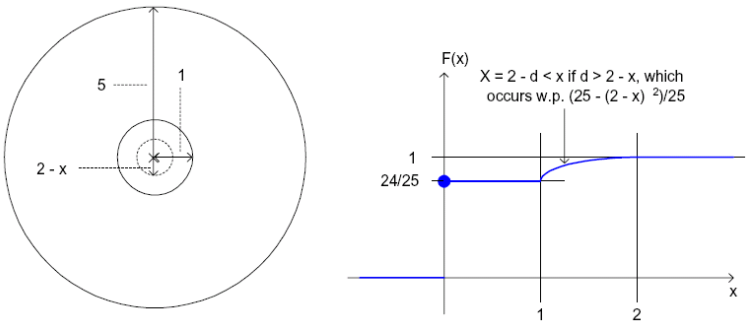
Example 4.10.3. *Define the random variable* X *as follows. You throw a dart uniformly in a circle with radius 5. The random variable* X *is equal to 2 minus the distance between the dart and the center of the circle if this distance is less than or equal to one. Otherwise,* X *is equal to* 0*.*

a. Plot carefully the probability distribution function $F(x) = P(X \leq x)$ *for* $x \in \Re := (-\infty, +\infty)$*.*

b. Give the mathematical expression for the probability density function $f(x)$ *of* X *for* $x \in \Re := (-\infty, +\infty)$*.*

Let Y be the distance between the dart and the center of the circle.

a. When $1 \leq x \leq 2$, $X \leq x$ if $Y \geq 2-x$, which occurs with probability $(25-(2-x)^2)/25$. Also, $X = 0$ if $Y > 1$, which occurs with probability $(25-1)/25 = 24/25$. These observations translate into the plot shown below:



b. Taking the derivative of $F(x)$, one finds

$$f(x) = \frac{24}{25}\delta(x) + \frac{2x-4}{25}1\{1 < x < 2\}.$$

Example 4.10.4. *Express the cdf of the following random variables in terms of* $F_X(\cdot)$.

a. $X^+ := \max\{0, X\}$*;*

b. $-X$.

c. $X^- := \max\{0, -X\}$*;*

d. $|X|$.

a.

$$P(X^+ \le x) = \begin{cases} 0, \text{ if } x < 0; \\ P(X \le x) = F_X(x), \text{ if } x \ge 0 \end{cases}$$

b. We find, using (4.2.2),

$$P(-X \le x) = P(X \ge -x) = 1 - P(X < -x) = 1 - F_X(-x-).$$

c. Note that, if $x \ge 0$, then $P(X^- \le x) = P(-X \le x) = 1 - F_X(-x-)$, as we showed above. Hence,

$$P(X^- \le x) = \begin{cases} 0, \text{ if } x < 0; \\ 1 - F(-x-), \text{ if } x \ge 0. \end{cases}$$

d. First note that, if $x \ge 0$,

$$P(|X| \le x) = P(-x \le X \le x) = P(X \le x) - P(X < -x) = F_X(x) - F_X(-x-).$$

Therefore,

$$P(|X| \leq x) = \begin{cases} 0, \text{ if } x < 0; \\ F(x) - F_X(-x-), \text{ if } x \geq 0. \end{cases}$$

Example 4.10.5. *A dart is flung at a circular dartboard of radius 3. Suppose that the probability that the dart lands in some region A of the dartboard is proportional to the area $|A|$ of A, i.e., is equal to $|A|/9\pi$.*

For each of the three scoring systems defined below:

i. Determine the distribution function of the score X;

ii. Calculate the expected value $E(X)$ of the score.

Here are the scoring systems:

a. $X = 4 - i$ if the distance Z of the dart to the center of the dartboard is in $(i-1, i]$ for $i = 1, 2, 3$.

b. $X = 3 - Z$ where Z is as in part a.

c. Assume now that the player has some probability 0.3 of missing the target altogether. If he does not miss, he hits an area A with a probability proportional to $|A|$. The score X is now 0 if the dart misses the target. Otherwise, it is $3 - Z$, where Z is as before.

a.i. We see that $P(X = 3) = P(Z \leq 1) = \pi/9\pi = 1/9$. Similarly, $P(X = 2) = P(1 < Z \leq 2) = (4\pi - \pi)/9\pi = 1/3$. Finally, $P(X = 1) = P(2 < Z \leq 3) = (9\pi - 4\pi)/9\pi = 5/9$. Hence,

$$F_X(x) = \begin{cases} 0, \text{if } x < 1; \\ 5/9, \text{if } 1 \leq x < 2; \\ 8/9, \text{if } 2 \leq x < 3; \\ 1, \text{if } x \geq 3. \end{cases}$$

a.ii. Accordingly,

$$E(X) = 1 \times \frac{5}{9} + 2 \times \frac{1}{3} + 3 \times \frac{1}{9} = 2.$$

b.i. Let $x > 0$. We see that $X = 3 - Z \leq x$ if $Z \geq 3 - x$, which occurs with probability

$1 - P(Z < 3 - x) = 1 - (3 - x)^2\pi/9\pi$. Hence,

$$F_X(x) = \begin{cases} 0, \text{if } x < 0; \\ 1 - (3-x)^2/9, \text{if } 0 \leq x < 3; \\ 1, \text{if } x \geq 3. \end{cases}$$

b.ii. Accordingly,

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^3 x \frac{2}{9}(3-x)dx = \frac{2}{9}[\frac{3}{2}x^2 - \frac{1}{3}x^3]_0^3 = 1.$$

c.i. Let Y be the score given that the player does not miss the target. Then Y has the pdf that we derived in part b. The score X of the player who misses the target with probability 0.3 is equal to 0 with probability 0.3 and to Y with probability 0.7. Hence,

$$F_X(x) = 0.31\{x \geq 0\} + 0.7F_Y(x).$$

That is,

$$F_X(x) = \begin{cases} 0, \text{if } x < 0; \\ 1 - 0.7(3-x)^2/9, \text{if } 0 \leq x < 3; \\ 1, \text{if } x \geq 3. \end{cases}$$

c.ii. From the definition of X in terms of Y we see that

$$E(X) = 0.3 \times 0 + 0.7E(X) = 0.7.$$

Example 4.10.6. *Suppose you put m balls randomly in n boxes. Each box can hold an arbitrarily large number of balls. What is the expected number of empty boxes?*

Designate by p the probability that the first box is empty. Let X_k be equal to 1 when box k is empty and to zero otherwise, for $k = 1, \ldots, n$. The number of empty boxes is $X = X_1 + \cdots + X_n$. By linearity of expectation, $E(X) = E(X_1) + \cdots + E(X_n) = nE(X_1) = np$. Now,

$$p = (\frac{n-1}{n})^m.$$

Indeed, p is the probability of the intersection of the independent events A_k for $k = 1, \ldots, m$ where event A_k is "ball k is put in a box other than box 1."

Example 4.10.7. *A cereal company is running a promotion for which it is giving a toy in every box of cereal. There are n different toys and each box is equally likely to contain any one of the n toys. What is the expected number of boxes of cereal you have to purchase to collect all n toys.*

Assume that you have just collected the first m toys, for some $m = 0, \dots n-1$. Designate by X_m the random number of boxes you have to purchase until you collect another different toy. Note that $P(X_m = 1) = (n-m)/n$. Also, if $X_m = 0$, then $X_m = 1 + Y_m$ where Y_m designates the additional number of boxes that you purchase until you get another different toy. Observe that X_m and Y_m have the same distribution, so that

$$E(X_m) = \frac{n-m}{n} \times 1 + \frac{m}{n} \times E(1 + Y_m) = 1 + \frac{m}{n} E(Y_m) = 1 + \frac{m}{n} E(X_m).$$

Solving, we find $E(X_m) = n/(n-m)$. Finally, the expected number of boxes we have to purchase is

$$E(X_0 + X_1 + \cdots + X_{n-1}) = \sum_{m=0}^{n-1} E(X_m) = \sum_{m=0}^{n-1} \frac{n}{n-m} = n(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}).$$

Example 4.10.8. *You pick a point P with a uniform distribution in $[0,1]^2$. Let Θ denote the angle made between the x-axis and the line segment that joins $(0,0)$ to the point P. Find the cdf, pdf, and expected value of Θ.*

Since P is chosen uniformly on the square, the probability we are within some region of the square is just proportional to the area of that region.

First, we find the cdf. One has

$$F_\Theta(\theta) = P(\Theta \leq \theta) = \begin{cases} 0, \text{ if } \theta < 0 \\ \frac{1}{2}\tan\theta, \text{ if } 0 \leq \theta \leq \frac{\pi}{4} \\ 1 - \frac{1}{2}\tan(\frac{\pi}{2} - \theta), \text{ if } \frac{\pi}{4} < \theta \leq \frac{\pi}{2} \\ 1, \text{ if } \theta > \frac{\pi}{2} \end{cases}$$

Second, we differentiate the cdf to find the pdf.

$$f_\Theta(\theta) = \frac{d}{d\theta} F_\Theta(\theta) = \begin{cases} \frac{1}{2(\cos\theta)^2}, \text{ if } 0 \le \theta \le \frac{\pi}{4} \\ \frac{1}{2[\cos(\frac{\pi}{2}-\theta)]^2}, \text{ if } \frac{\pi}{4} < \theta \le \frac{\pi}{2} \\ 0, \text{ otherwise.} \end{cases}$$

Finally, we use the pdf to find the expected value.

$$\begin{aligned} E[\Theta] &= \int_0^{\frac{\pi}{4}} \frac{\theta}{2(\cos\theta)^2} d\theta + \int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \frac{\theta}{2[\cos(\frac{\pi}{2}-\theta)]^2} d\theta \\ &= \frac{1}{2}(\ln(\cos\theta) + \theta\tan\theta)]_0^{\frac{\pi}{4}} + \frac{1}{2}(\ln(\cos(\frac{\pi}{2}-\theta)) - \theta\tan(\frac{\pi}{2}-\theta))]_{\frac{\pi}{4}}^{\frac{\pi}{2}} \\ &= \frac{1}{2}[\ln(\frac{\sqrt{2}}{2}) + \frac{\pi}{4} - (\ln(\frac{\sqrt{2}}{2}) - \frac{\pi}{4})] \\ &= \frac{\pi}{4} \end{aligned}$$

Example 4.10.9. *A random variable X has the following cdf:*

$$F_X(x) = \begin{cases} 0, \textit{ for } x < 0 \\ 0.3, \textit{ for } 0 \le x < 2 \\ 0.3 + 0.2x, \textit{for } 2 \le x < 3 \\ 1, \textit{ for } x \ge 3. \end{cases}$$

a. Explain why F_X is a cdf;

b. Find $P(X = 0), P(X = 1), P(X = 2), P(X = 3)$;

c. Find $P(0.5 < X \le 2.3)$;

d. Find $P(0 < X < 3)$ and $P(X < X \le 3)$;

e. Find $f_X(x)$;

f. Find $P[X \le 1.4 \mid X \le 2.2]$;

g. Calculate $E(X)$.

a. Figure 4.1(a) shows the cdf $F_X(x)$. To show that $F_X(x)$ is indeed a cdf we must verify the properties (4.2.1). The figure shows that $F_X(\cdot)$ satisfies these properties.

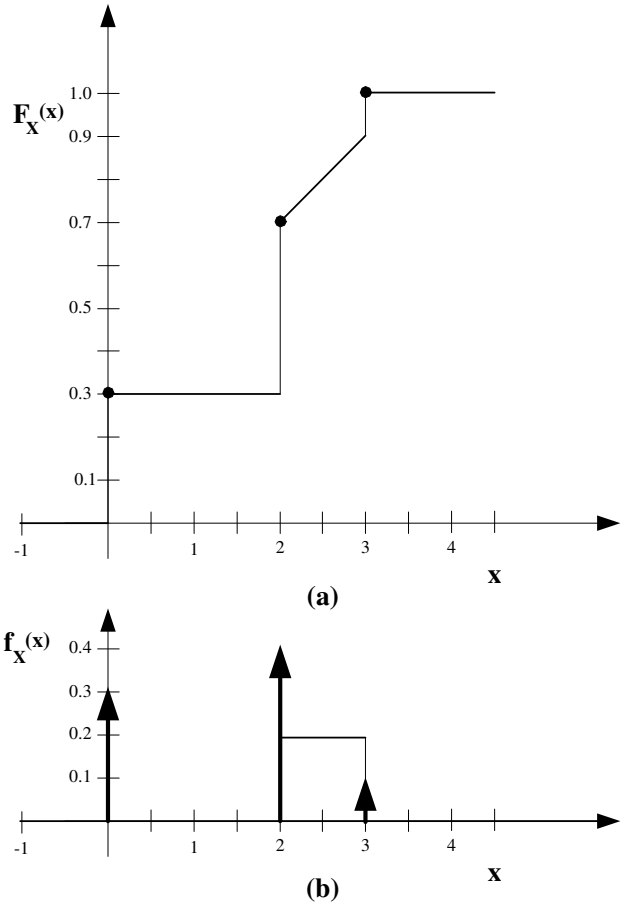


Figure 4.1: (a) cdf of X. (b) pdf of X.

b. We find the following probabilities:

$$\begin{aligned} P(X=0) &= F_X(0) - F_X(0-) = 0.3 \\ P(X=2) &= F_X(2) - F_X(2-) = 0.4 \\ P(X=3) &= F_X(3) - F_X(3-) = 0.1 \end{aligned}$$

c. One has $P(0.5 < X \leq 2.3) = P(X \leq 2.3) - P(X \leq 0.5) = F_X(2.3) - F_X(0.5) = 0.76 - 0.46 = 0.3$.

d. We find $P(0 < X < 3) = P(X < 3) - P(X \leq 0) = F_X(3-) - F_X(0) = 0.9 - 0.3 = 0.6$.

Also, $P(0 < X \leq 3) = P(X \leq 3) - P(X \leq 0) = F_X(3) - F_X(0) = 1 - 0.3 = 0.7$.

e. According to (4.2.4), the pdf $f_X(\cdot)$ of X is as follows:

$$f_X(x) = 0.3\delta(x) + 0.4\delta(x-2) + 0.1\delta(x-3) + 0.2 \times 1\{2 < x < 3\}.$$

See Figure 4.1(b).

f. $P[X \leq 1.4 \mid X \leq 2.2] = \frac{P[X \leq 1.4 \cap X \leq 2.2]}{P[X \leq 2.2]} = \frac{P[X \leq 1.4]}{P[X \leq 2.2]} = \frac{0.3}{0.74} = 0.395$

g. Accordingly to (4.5.1),

$$\begin{aligned} E[X] &= 0 \times 0.3 + 2 \times 0.4 + 3 \times 0.1 + \int_2^3 x \times 0.2 dx \\ &= 0.8 + 0.3 + 0.2 \times \left[\frac{x^2}{2}\right]_2^3 = 1.1 + 0.2 \times 2.5 = 1.6 \end{aligned}$$

Example 4.10.10. *Let X and Y be independent random variables with common distribution function F and density function f.*

a. Show that $V = \max\{X, Y\}$ has distribution function $F_V(v) = F(v)^2$ and density function $f_V(v) = 2f(v)F(v)$.

b. Find the density function of $U = \min\{X, Y\}$.

Let X and Y be independent random variables each having the uniform distribution on $[0, 1]$.

c. Find $E(U)$.

d. Find cov(U, V).

Let X and Y be independent exponential random variables with mean 1.

e. Show that U is an exponential random variable with mean $1/2$.

f. Find $E(V)$ and var(V).

a. We find

$$\begin{aligned} F_V(v) &= P(V \leq v) = P(\max\{X, Y\} \leq v) \\ &= P(\{X \leq v\} \cap \{Y \leq v\}) = P(X \leq v)P(Y \leq v) = F(v)^2. \end{aligned}$$

Differentiate the cdf to get the pdf by using the Chain Rule $f_V(v) = \frac{d}{dv}F_V(v) = 2F(v)f(v)$.

b. First find the cdf of U.

$$\begin{aligned} F_U(u) &= P(U \le u) = P(\min\{X,Y\} \le u) = 1 - P(\min\{X,Y\} > u) \\ &= 1 - P(\{X > u\} \cap \{Y > u\}) = 1 - P(X > u)P(Y > u) \\ &= 1 - (1 - P(X \le u))(1 - P(Y \le u)) = 1 - (1 - F(u))^2. \end{aligned}$$

Differentiate the cdf to get the pdf by using the Chain Rule $f_U(u) = \frac{d}{du}F_U(u) = 2f(u)(1 - F(u))$.

c. X and Y be independent random variables each having the uniform distribution on [0,1]. Hence, $f_U(u) = 2(1-u)$ for $u \in [0,1]$.

$$\begin{aligned} E[U] &= \int_0^1 u f_U(u)du = \int_0^1 2u(1-u)du = 2\int_0^1 (u - u^2)du \\ &= 2(\frac{u^2}{2} - \frac{u^3}{3})]_0^1 = \frac{1}{3}. \end{aligned}$$

d. $E[V] = \int_0^1 v f_V(v)dv = \int_0^1 2v^2 dv = \frac{2}{3}$

$$\begin{aligned} \text{cov}[U,V] &= E[(U - E[U])(V - E[V])] = E[UV] - E[U]E[V] = E[XY] - E[U]E[V] \\ &= \int_0^1 \int_0^1 xy f_{X,Y}(x,y)dxdy - E[U]E[V] = \int_0^1 xdx \int_0^1 ydy - E[U]E[V] \\ &= \frac{1}{2} \times 1 - \frac{1}{3} \times \frac{2}{3} = \frac{1}{36}. \end{aligned}$$

e. X and Y be independent exponential random variables with mean 1, so each has pdf $f(x) = e^{-x}$ and cdf $F(x) = 1 - P(X > x) = 1 - \int_x^\infty e^{-\tilde{x}}d\tilde{x} = 1 - e^{-x}$. Using part b, $f_U(u) = 2f(u)(1-F(u)) = 2e^{-2u}$. Thus U is an exponential random variable with mean $\frac{1}{2}$.

f. From part a, $f_V(v) = 2f(v)F(v) = 2e^{-v}(1 - e^{-v})$.

$$\begin{aligned} E[V] &= \int_0^\infty v f_V(v)dv = \int_0^\infty 2v(e^{-v} - e^{-2v})dv \\ &= 2\int_0^\infty ve^{-v}dv - \int_0^\infty v(2e^{-2v})dv = 2 \times 1 - \frac{1}{2} = \frac{3}{2}. \end{aligned}$$

$$\begin{aligned} E[V^2] &= \int_0^\infty v^2 f_V(v)dv = \int_0^\infty 2v^2(e^{-v} - e^{-2v})dv \\ &= \int_0^\infty 2v^2 e^{-v}dv - \int_0^\infty 2v^2 e^{-2v}dv \\ &= (-2v^2e^{-v}]_0^\infty + \int_0^\infty 4ve^{-v}dv) - (-v^2e^{-2v} + \int_0^\infty 2ve^{-2v}dv) \\ &= (0 - 4ve^{-v}]_0^\infty + \int_0^\infty 4e^{-v}dv) - (0 - ve^{-2v}]_0^\infty + \int_0^\infty e^{-2v}dv) \\ &= (0 - 4e^{-v}]_0^\infty) - (0 - \frac{1}{2}e^{-2v}]_0^\infty) = 4 - \frac{1}{2} = \frac{7}{2}. \end{aligned}$$

$Var[V] = E[V^2] - E[V]^2 = \frac{7}{2} - (\frac{3}{2})^2 = \frac{5}{4}$.

Example 4.10.11. *Choose X in $[0,1]$ as follows. With probability 0.2, $X = 0.3$; with probability 0.3, $X = 0.7$; otherwise, X is uniformly distributed in $[0.2, 0.5] \cup [0.6, 0.8]$. (a). Plot the c.d.f. of X; (b) Find $E(X)$; (c) Find $var(X)$; (d) Calculate $P[X \leq 0.3 \mid X \leq 0.7]$.*

a. Figure 4.2 shows the p.d.f. and the c.d.f. of X. Note that the value of the density (1) is such that f_X integrates to 1.

b. We find (see Figure 4.2)

$$\begin{aligned} E(X) &= 0.2 \times 0.3 + 0.3 \times 0.7 + \int_{0.2}^{0.5} xdx + \int_{0.6}^{0} .8xdx = 0.27 + \frac{1}{2}[x^2]_{0.2}^{0.5} + \frac{1}{2}[x^2]_{0.6}^{0.8} \\ &= 0.27 + \frac{1}{2}[(0.5)^2 - (0.2)^2] + \frac{1}{2}[(0.8)^2 - (0.6)^2] = 0.515. \end{aligned}$$

c. We first calculate $E(X^2)$. We find

$$\begin{aligned} E(X^2) &= 0.2 \times (0.3)^2 + 0.3 \times (0.7)^2 + \int_{0.2}^{0.5} x^2dx + \int_{0.6}^{0} .8x^2dx = 0.165 + \frac{1}{3}[x^3]_{0.2}^{0.5} + \frac{1}{3}[x^3]_{0.6}^{0.8} \\ &= 0.165 + \frac{1}{3}[(0.5)^3 - (0.2)^3] + \frac{1}{3}[(0.8)^3 - (0.6)^3] = 0.3027. \end{aligned}$$

Consequently,

$$\text{var}(X) = E(X^2) - (E(X))^2 = 0.3027 - (0.515)^2 = 0.0374.$$

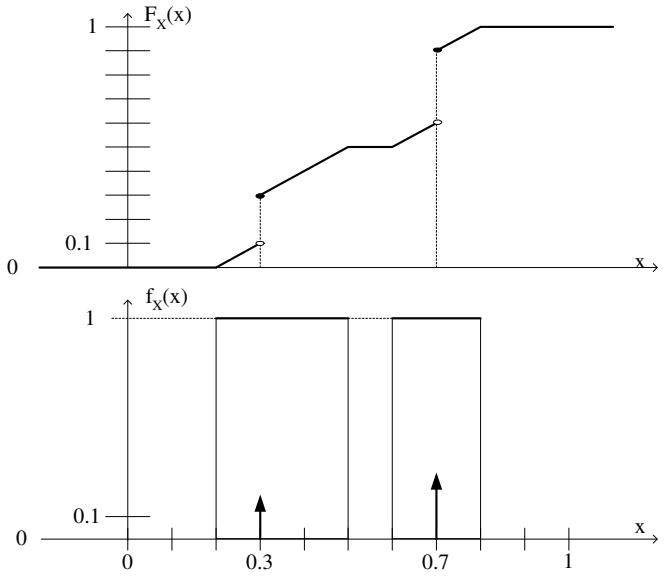


Figure 4.2: The cumulative distribution function of X.

d. We have (see Figure 4.2)

$$P[X \leq 0.3 \mid X \leq 0.7] = \frac{P(X \leq 0.3)}{P(X \leq 0.7)} = \frac{0.3}{0.9} = \frac{1}{3}.$$

Example 4.10.12. *Let X be uniformly distributed in $[0, 10]$. Find the cdf of the following random variables.*

a. $Y := \max\{2, \min\{4, X\}\}$*;*

b. $Z := 2 + X^2$*;*

c. $V := |X - 4|$*;*

d. $W := \sin(2\pi X)$.

a. $F_Y(y) = P(Y \leq y) = P(max\{2, min\{4, X\}\} \leq y)$. Note that $Y \in [2, 4]$, so that $F_Y(2-) = 0$ and $F_Y(4) = 1$.

Let $y \in [2, 4)$. We see that $Y \leq y$ if and only if $X \leq y$, which occurs with probability $y/10$.

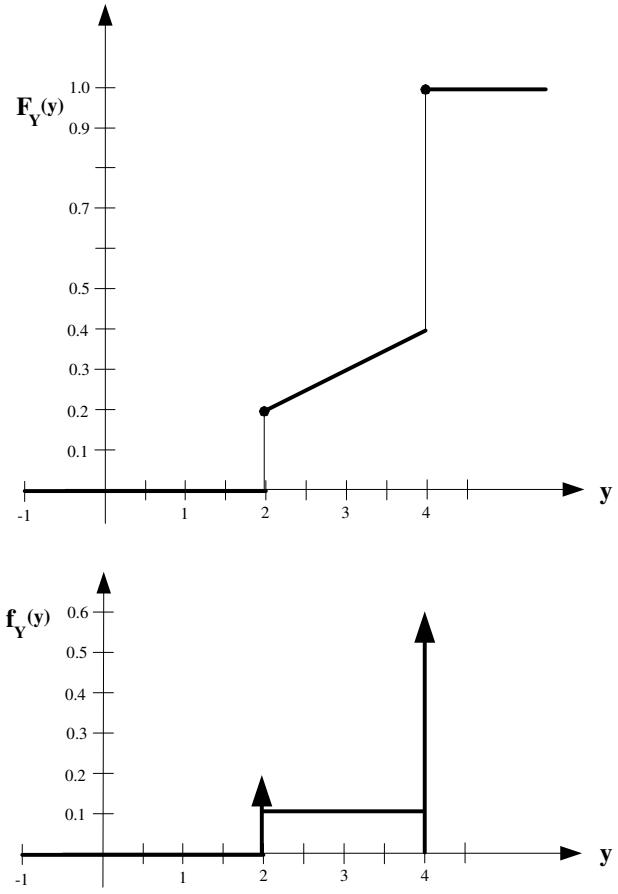


Figure 4.3: cdf and pdf of $\max\{2, min\{4, X\}\}$

Hence,

$$F_Y(y) = \begin{cases} 0, \text{ if } y < 2 \\ y/10, \text{ if } 2 \leq y < 4 \\ 1, \text{ if } y \geq 4. \end{cases}$$

Accordingly,

$$f_Y(y) = 0.2\delta(y-2) + 0.6\delta(y-4) + 0.1 \times 1\{2 < y < 4\}.$$

b. $F_Z(z) = P(Z \leq z) = P(2 + X^2 \leq z) = P(X \leq \sqrt{z-2})$. Consequently,

$$F_Z(z) = \begin{cases} 0, \text{ if } z < 2 \\ \sqrt{z-2}, \text{ if } 2 \leq z < 102 \\ 1, \text{ if } z \geq 102. \end{cases}$$

Also,

$$f_Z(z) = \begin{cases} 0, \text{ if } z \leq 2 \\ \frac{1}{2\sqrt{z-2}}, \text{ if } 2 < z < 102 \\ 0, \text{ if } z \geq 102 \end{cases}$$

c. $F_V(v) = P(V \leq v) = P(\mid X - 4 \mid \leq v) = P(-v + 4 \leq X \leq v + 4)$. Hence,

$$F_V(v) = \begin{cases} 0, \text{ if } v < 0 \\ 0.2v, \text{ if } 0 \leq v < 4 \\ 0.1v + 0.4, \text{ if } 4 \leq v < 6 \\ 1, \text{ if } v \geq 6. \end{cases}$$

Also,

$$f_V(v) = \begin{cases} 0, \text{ if } v \leq 0 \\ 0.2, \text{ if } 0 < v \leq 4 \\ 0.1, \text{ if } 4 < v \leq 6 \\ 0, \text{ if } v > 6 \end{cases}$$

d. $F_W(w) = P(W \leq w) = P(sin(2\pi X) \leq w)$. Note that $W \in [-1, 1]$, so that $F_W(-1-) = 0$ and $F_W(1) = 1$. The interesting case is $w \in (-1, 1)$. A picture shows that

$$F_W(w) = \begin{cases} 0, \text{ if } w < -1 \\ 0.5 + \frac{1}{\pi} sin^{-1}(w), \text{ if } -1 \leq w < 1 \\ 1, \text{ if } w \geq 1. \end{cases}$$

Example 4.10.13. *Assume that a dart flung at a target hits a point ω uniformly distributed in $[0, 1]^2$. The random variables $X(\omega), Y(\omega), Z(\omega)$ are defined as follows. $X(\omega)$ is the maximum distance between ω and any side of the square. $Y(\omega)$ is the minimum distance between ω and any side of the square. $Z(\omega)$ is the distance between ω and a fixed vertex of the square.*

Find the probability density functions f_X, f_Y, f_Z.

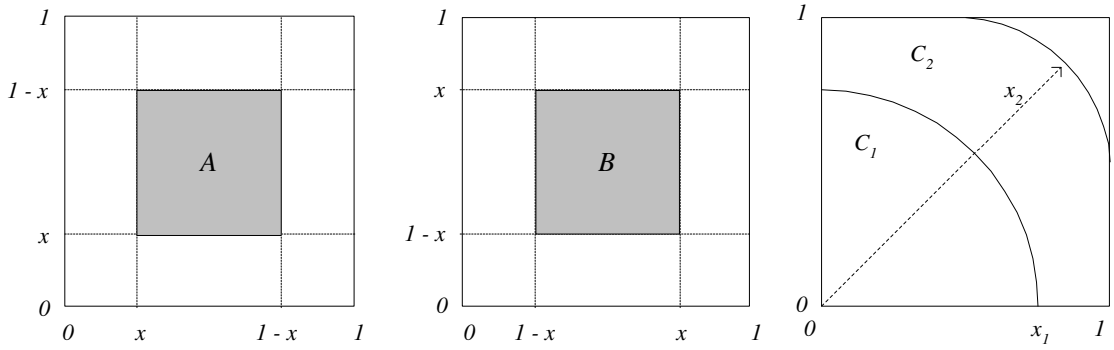


Figure 4.4: Diagram for Example 4.10.13

Figure 4.4 shows the following events:

$$\begin{aligned} A &= \{\omega \mid X \geq x\}; \\ B &= \{\omega \mid Y \leq x\}; \\ C_1 &= \{\omega \mid Z \leq x_1\} \text{ when } x_1 \leq 1; \\ C_2 &= \{\omega \mid Z \leq x_2\} \text{ when } x_2 > 1. \end{aligned}$$

Note the difference in labels on the axes for the events A and B. For C_1 and C_2, the reference vertex is $(0, 0)$.

a. From these figures, we find that

$$F_X(x) = \begin{cases} 0, \text{ if } x < 0 \\ 1 - (1 - 2x)^2, \text{ if } 0 \leq x \leq 0.5 \\ 1, \text{ if } x \geq 0.5. \end{cases}$$

Accordingly,

$$f_X(x) = \begin{cases} 0, \text{ if } x < 0 \\ 4(1 - 2x), \text{ if } 0 \leq x \leq 0.5 \\ 0, \text{ if } x \geq 0.5. \end{cases}$$

b. Similarly,

$$F_Y(x) = \begin{cases} 0, \text{ if } x < 0.5 \\ (2x-1)^2, \text{ if } 0.5 \leq x \leq 1 \\ 1, \text{ if } x \geq 1.5. \end{cases}$$

Accordingly,

$$f_Y(x) = \begin{cases} 0, \text{ if } x < 0 \\ 4(2x-1), \text{ if } 0.5 \leq x \leq 1 \\ 0, \text{ if } x > 1. \end{cases}$$

c. The area of C_1 is $\pi x_1^2/4$. That of C_2 consists of a rectangle $[0, v] \times [0, 1]$ plus the integral over $uin[v, 1]$ of $\sqrt{x_2^2 - u^2}$. One finds

$$f_Z(z) = \begin{cases} \frac{1}{2}\pi z & 0 \leq z < 1 \\ \frac{1}{2}\pi z - 2zcos^{-1}(\frac{1}{z}) & 1 \leq z < \sqrt{2} \\ 0 & z \geq \sqrt{2} \end{cases}$$

Example 4.10.14. *A circle of unit radius is thrown on an infinite sheet of graph paper that is grid-ruled with a square grid with squares of unit side. Assume that the center of the circle is uniformly distributed in the square in which it falls. Find the expected number of vertex points of the grid that fall in the circle.*

There is a very difficult way to solve the problem and a very easy way. The difficult way is as follows. Let X be the number of vertex points that fall in the circle. We find $P(X = k)$ for $k = 1, 2, \ldots$ and we compute the expectation. This is very hard because the sets of possible locations of the center of the circle for these various events are complicated intersections of circles.

The easy way is as follows. We consider the four vertices of the square in which the center of the circle lies. For each of these vertices, there is some probability p that it is in the circle. Accordingly, we can write $X = X_1 + X_2 + X_3 + X_4$ where X_i is 1 if vertex i of the square is in the circle and is 0 otherwise. Now,

$$E(X) = E(X_1) + E(X_2) + E(X_3) + E(X_4) = p + p + p + p = 4p.$$

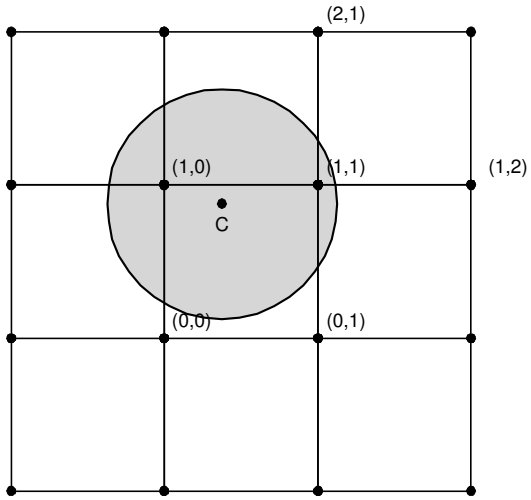


Figure 4.5: Diagram for Example 4.10.14

The key observation here is that the average value of a sum of random variables is the sum of their average values, even when these random variables are not independent.

It remains to calculate p. To do that, note that the set of possible locations of the center of the circle in a given square such that one vertex is in the circle is a quarter-circle with radius 1. Hence, $p = \pi/4$ and we conclude that $E(X) = \pi$.

Example 4.10.15. *Ten numbers are selected from* $\{1, 2, 3, \ldots, 30\}$ *uniformly and without replacement. Find the expected value of the sum of the selected numbers.*

Let $X_1, \ldots, X_{10}$ be the ten numbers you pick in $\{1, 2, \ldots, 30\}$ uniformly and without replacement. Then $E(X_1 + \cdots + X_{10}) = E(X_1) + \cdots + E(X_{10})$. Consider any X_k for some $k \in \{1, \ldots, 10\}$. By symmetry, X_k is uniformly distributed in $\{1, 2, \ldots, 30\}$. Consequently, $E(X_k) = \sum_{n=1}^{30} n \times \frac{1}{30} = 15.5$. Hence $E(X_1 + \cdots + X_{10}) = 10 \times 15.5 = 155$. Once again, the trick is to avoid looking at the joint distribution of the X_i, as in the previous example.

Example 4.10.16. *We select a point* X *according to some probability density function* f_X.

Find the value of a that minimizes the average value of the square distance between the point $(a, 1)$ *and the random point* $(X, 0)$ *in the plane.*

Let the random variable Z be the squared distance between $(X, 0)$ and $(a, 1)$. That is,

$$Z = (X - a)^2 + (0 - 1)^2.$$

We wish to minimize $E[Z]$ i.e minimize $E[(X - a)^2 + (0 - 1)^2]$.

To find the value of a, we solve $\frac{dE[Z]}{da} = 0$. We find

$$\frac{d}{da}(a^2 - 2aE[X] - E[X^2] + a).$$

We find that the value of a for which this expression is equal to zero is $a = E(X)$.

The value of $\frac{d^2}{da^2}(a^2 - 2aE[X] - E[X^2] + a)$ for $a = E(X)$ is equal to 2. Since this is positive, a indeed corresponds to the minimum.

Example 4.10.17. *Construct a pdf of a random variable* X *on* $[0, \infty)$ *so that* $P[X > a+b \mid X > a] > P(X > b)$ *for all* $a, b > 0$.

The idea is that X is a lifetime of an item whose residual lifetime X gets longer as it gets older. An example would be an item whose lifetime is either $Exd(1)$ or $Exd(2)$, each with probability 0.5, say. As the item gets older, it becomes more likely that its lifetime is $Exd(1)$ (i.e., with mean 1) instead of $Exd(2)$ (with mean 1/2). Let's do the math to confirm the intuition.

From the definition,

$$f_X(x) = 0.5 \times e^{-x} + 0.5 \times 2e^{-2x}, x \geq 0.$$

Hence,

$$P(X > a) = \int_a^\infty f_X(x)dx = 0.5e^{-a} + 0.5e^{-2a}, a \geq 0,$$

so that

$$P[X > a + b \mid X > a] = \frac{0.5e^{-(a+b)} + 0.5e^{-2(a+b)}}{0.5e^{-a} + 0.5e^{-2a}}.$$

Simple algebra allows to verify that $P[X > a + b \mid X > a] \geq P(X > b)$.

Example 4.10.18. *Construct a pdf of a random variable X on $[0, \infty)$ so that $P[X > a+b \mid X > a] < P(X > b)$ for all $a, b > 0$.*

Here, we can choose a pdf that decays faster than exponentially. Say that the lifetime has a density

$$f_X(x) = A \exp\{-x^2\}$$

where A is such that $\int_0^\infty f_X(x)dx = 1$. The property we are trying to verify is equivalent to

$$\int_{a+b}^\infty f_X(x)dx < \int_a^\infty f_X(x)dx \int_b^\infty f_X(x)dx,$$

or

$$\int_0^\infty f_X(x)dx \int_{a+b}^\infty f_X(x)dx < \int_a^\infty f_X(x)dx \int_b^\infty f_X(x)dx.$$

We can write this inequality as

$$\int_0^\infty \int_{a+b}^\infty e^{-(x^2+y^2)}dxdy < \int_a^\infty \int_b^\infty e^{-(x^2+y^2)}dxdy.$$

That is,

$$\phi(A) + \phi(B) \leq \phi(B) + \phi(C)$$

where

$$\phi(D) = \int\int_D e^{-(x^2+y^2)}dxdy$$

for a set $D \subset \Re^2$ and $A = [0, b]\times[a+b, \infty)$, $B = [b, \infty)\times[a+b, \infty)$, and $C = [b, \infty)\times[a, a+b]$.

To show $\phi(A) < \phi(C)$, we note that each point $(x, a+b+y)$ in A corresponds to a point $(b + y, a + x)$ in C and

$$e^{-(x^2+(a+b+y)^2)} < e^{-((b+y)^2+(a+x)^2)},$$

by convexity of $g(z) = z^2$.

Example 4.10.19. *Suppose that the number of telephone calls made in a day is a Poisson random variable with mean 1000.*

a. What is the probability that more than 1142 calls are made in a day?

b. Find a bound for this probability using Markov's Inequality.

c. Find a bound for this probability using Chebyshev's Inequality.

Let N denote the number of telephone calls made in a day. N is Poisson with mean 1000 so the pmf is// $P(N = n) = \frac{e^{-1000}1000^n}{n!}$ and $Var[N] = 1000$.

a. $P(N > 1142) = e^{-1000} \sum_{n=1143}^{\infty} \frac{1000^n}{n!}$.

b. $P(N > 1142) = P(N \geq 1143) \leq \frac{E[N]}{1143} = \frac{1000}{1143}$.

c. $P(N > 1142) = P(N \geq 1143) \leq P(|N - E[N]| \geq 143) \leq \frac{Var[N]}{143^2} = \frac{1000}{20449}$.

Chapter 5

Random Variables

A collection of random variables is a collection of functions of the outcome of the same random experiment. We explain how one characterizes the statistics of these random variables.

We have looked at one random variable. The idea somehow was that we made one numerical observation of one random experiment. Here we extend the idea to multiple numerical observations about the same random experiment. Since there is one random experiment, nature chooses a single value of $\omega \in \Omega$. The different observations, say X, Y, Z, are all functions of the *same* ω. That is, one models these observations as $X(\omega), Y(\omega)$, and $Z(\omega)$.

As you may expect, these values are related in some way. Thus, observing $X(\omega)$ provides some information about $Y(\omega)$. In fact, one of the interesting questions is how one can use the information that some observations contain about some other random variables that we do not observe directly.

5.1 Examples

We pick a ball randomly from a bag and we note its weight X and its diameter Y.

We observe the temperature at a few different locations.

We measure the noise voltage at different times.

We track the evolution over time of the value of Cisco shares and we want to forecast future values.

A transmitter sends some signal and the receiver observes the signal it receives and tries to guess which signal the transmitter sent.

5.2 Joint Statistics

Let $\{X(\omega), Y(\omega)\}$ be a pair of random variables.

The *joint distribution* of $\{X(\omega), Y(\omega)\}$ is specified by the *joint cumulative distribution function* (jcdf) $F_{X,Y}$ defined as follows:

$$F_{X,Y}(x, y) := P(X \leq x \text{ and } Y \leq y), x, y \in \Re.$$

In the continuous case,

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) du \; dv$$

for a nonnegative function $f_{X,Y}(x, y)$ that is called the *joint pdf* (jpdf) of the random variables.

These ideas extend to an arbitrary number of random variables.

This joint distribution contains more information than the two individual distributions. For instance, let $\{X(\omega), Y(\omega)\}$ be the coordinates of a point chosen uniformly in $[0, 1]^2$. Define also $Z(\omega) = X(\omega)$. Observe that the individual distributions of the each of the random variables in the pairs $\{X(\omega), Y(\omega)\}$ and $\{X(\omega), Z(\omega)\}$ are the same. The tight dependency of X and Z is revealed by their joint distribution.

The *covariance* of a pair of random variables (X, Y) is defined as

$$\text{cov}(X, Y) = E((X - E(X))(Y - E(Y)) = E(XY) - E(X)E(Y).$$

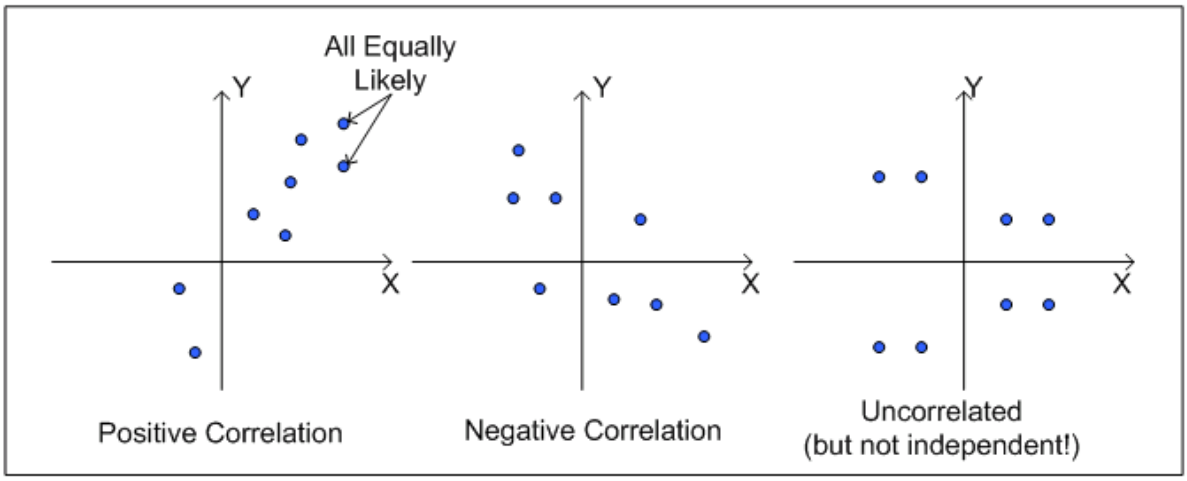


Figure 5.1: Correlation

The random variables are positively (resp. negatively, un-) correlated if $\text{cov}(X,Y) > 0$ (resp. $< 0, = 0$). The covariance is a measure of dependence. The idea is that if $E(XY)$ is larger than $E(X)E(Y)$, then X and Y tend to be large or small together more than if they were independent. In our example above, $E(XZ) = E(X^2) = 1/3 > E(X)E(Z) = 1/4$. Figures 5.1 illustrates the meaning of correlation. Each of the figures shows the possible values of a pair (X,Y) of random variables; all the values are equally likely. In the left-most figure, X and Y tend to be large or small together. These random variables are positively correlated. Indeed, the product XY is larger on average than it would be if a larger value of X did not imply a larger than average value of Y. The other two figures can be understood similarly.

If $h : \Re^2 \to \Re$ is nice (Borel-measurable - once again, all the functions from $\Re^2$ to $\Re$ that we encounter have that property), then $h(X,Y)$ is a random variable. One can show, as in the case of a single random variable, that

$$E(h(X,Y)) = \int \int h(x,y) dF_{X,Y}(x,y).$$

The continuous case is similar to the one-variable case.

It is sometimes convenient to use vector notation. To do that, one defines the expected value of a random vector to be the vector of expected values. Similarly, the expected value of a matrix is the matrix of expected values. Let $\boldsymbol{X}$ be a column vector whose n elements $X_1, \ldots, X_n$ are random variables. That is, $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ where $(\cdot)^T$ indicates the transposition operation. We define $E(\boldsymbol{X}) = (E(X_1), \ldots, E(X_n))^T$. For a random matrix $\boldsymbol{W}$ whose entry (i, j) is the random variable $W_{i,j}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, we define $E(\boldsymbol{W})$ to be the matrix whose entry (i, j) is $E(W_{i,j})$. Recall that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices of compatible dimensions, then $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$. Recall also the definition of the *trace* $tr(\boldsymbol{A})$ of a square matrix $\boldsymbol{A}$ as the sum of its diagonal elements: $tr(\boldsymbol{A}) = \sum_i A_{i,i}$. Note that if $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors with the same dimension, then

$$\boldsymbol{a}^T\boldsymbol{b} = \sum_i a_i b_i = tr(\boldsymbol{a}\boldsymbol{b}^T) = tr(\boldsymbol{b}\boldsymbol{a}^T).$$

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors, we define the following *covariance matrices*:

$$\Sigma_{\boldsymbol{X},\boldsymbol{Y}} := \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) := E((\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T) = E(\boldsymbol{X}\boldsymbol{Y}^T) - E(\boldsymbol{X})E(\boldsymbol{Y}^T)$$

and

$$\Sigma_{\boldsymbol{X}} := E((\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T) = E(\boldsymbol{X}\boldsymbol{X}^T) - E(\boldsymbol{X})E(\boldsymbol{X}^T).$$

Using linearity, one finds that

$$\mathrm{cov}(\boldsymbol{A}\boldsymbol{X}, \boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{A}\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}^T. \tag{5.2.1}$$

Similarly,

$$E(\boldsymbol{X}^T\boldsymbol{Y}) = E(tr(\boldsymbol{X}\boldsymbol{Y}^T)) = trE(\boldsymbol{X}\boldsymbol{Y}^T). \tag{5.2.2}$$

5.3 Independence

We say that two random variables $\{X, Y\}$ are independent if

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B) \tag{5.3.1}$$

for all subsets A and B of the real line (... Borel sets, to be precise).

More generally, a collection of random variables are said to be mutually independent if the probability that any finite subcollection of them belongs to any given subsets is the product of the probabilities.

We have seen examples before: flipping coins, tossing dice, picking (X, Y) uniformly in a square or a rectangle, and so on.

Theorem 5.3.1. *a. The random variables X, Y are independent if and only if the joint cdf $F_{X,Y}(x, y)$ is equal to $F_X(x)F_Y(y)$, for all x, y. A collection of random variables are mutually independent if the jcdf of any finite subcollection is the product of the cdf.*

b. If the random variables X, Y have a joint pdf $f_{X,Y}(x, y)$, they are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, for all x, y. A collection of random variables with a jpdf are mutually independent if the jpdf of any finite subcollection is the product of the pdf.

c. If X, Y are independent, then $f(X)$ and $g(Y)$ are independent.

d. If X and Y are independent, then $E(XY) = E(X)E(Y)$. [The converse is not true!] That is, independent random variables are uncorrelated.

e. More generally, if $X, Y, , W$ are mutually independent, then

$$E(XY \cdots W) = E(X)E(Y) \cdots E(W).$$

f. The variance of the sum of pairwise independent random variables is the sum of their variances.

g. If X and Y are continuous and independent, then

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(u)f_Y(x-u)du. \tag{5.3.2}$$

The expression to the right of the identity is called the convolution *of f_X and f_Y. Hence, the pdf of the sum of two independent random variables is the convolution of their pdf.*

Proof:

We provide sketches of the proof of these important results. The derivation should help you appreciate the results.

a. Assume that X, Y are independent. Then, by (5.3.1),

$$F_{X,Y}(x,y) = P(X \leq x \text{ and } Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y), \text{ for all } x, y \in \Re.$$

Conversely, assume that the identity above holds. It is easy to see that

$$P(X \in (a,b] \text{ and } Y \in (c,d]) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c).$$

Using $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ in this expression, we find after some simple algebra that

$$P(X \in (a,b] \text{ and } Y \in (c,d]) = P(X \in (a,b])P(Y \in (c,d]).$$

Since the probability is countably additive, the expression above implies that

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B)$$

for a collection of sets A and B that is closed under countable operations and that contains the intervals. Consequently, the identity above holds for all $A, B \in \mathcal{B}$ where $\mathcal{B}$ is the Borel σ-field of $\Re$. Hence, X, Y are independent.

The same argument proves the corresponding result for mutual independence of random variables.

b. Assume that X, Y are independent. Then (5.3.1) implies that

$$\begin{aligned} f_{X,Y}(x,y)dxdy &= P(X \in (x, x+dx) \text{ and } Y \in (y, y+dy)) \\ &= P(X \in (x, x+dx))P(Y \in (y, y+dy)) = f_X(x)dx f_Y(y)dy, \end{aligned}$$

so that

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \text{ for all } x, y \in \Re.$$

Conversely, if the identity above holds, then

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dudv \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u)f_Y(v)dudv \\ &= \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv = F_X(x)F_Y(y). \end{aligned}$$

Part (a) then implies that X and Y are independent.

A similar argument proves the result for the mutual independence.

c. Assume X and Y are independent. Note that $g(X) \in A$ if and only if $X \in g^{-1}(A)$, by definition of $g^{-1}(\cdot)$. A similar result holds for $h(Y)$. Hence

$$\begin{aligned} P(g(X) \in A \text{ and } h(Y) \in B) &= P(X \in g^{-1}(A) \text{ and } Y \in h^{-1}(B)) \\ &= P(X \in g^{-1}(A))P(Y \in h^{-1}(B)) = P(g(X) \in A)P(h(Y) \in B), \end{aligned}$$

which shows that $g(X)$ and $h(Y)$ are independent. The derivation of the mutual independence result is similar.

d. Assume that X and Y are independent and that they are continuous. Then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, so that

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_X(y)dy = E(X)E(Y). \end{aligned}$$

The same derivation holds in the discrete case. The hybrid case is similar.

Note that the converse is not true. For instance, assume that (X,Y) is equally likely to take the four values $\{(-1,0),(0,-1),(1,0),(0,1)\}$. We find that $E(XY) = 0 = E(X)E(Y)$. However, these random variables are not independent since $P[Y = 0 \mid X = 1] = 1 \neq P(Y = 0) = 1/2$.

e. We can prove this result by induction by noticing that if $\{X_n, n \geq 1\}$ are mutually independent, then X_{n+1} and $X_1 \times \cdots \times X_n$ are independent. Hence,

$$E(X_1 \times \cdots \times X_n X_{n+1}) = E(X_1 \times \cdots \times X_n)E(X_{n+1}).$$

f. Let $\{X_1, \ldots, X_n\}$ be pairwise independent. By subtracting their means, we can assume that they are zero-mean. Now,

$$\begin{aligned}\text{var}(X_1 + \cdots + X_n) &= E((X_1 + \cdots + X_n)^2) = E(X_1^2 + \cdots + X_n^2 + \sum_{i \neq j} X_i X_j) \\ &= E(X_1^2) + \cdots + E(X_n^2) = \text{var}(X_1) + \cdots + \text{var}(X_n).\end{aligned}$$

In this calculation, we used the fact that $E(X_i X_j) = E(X_i)E(X_j)$ for $i \neq j$ because the random variables are pairwise independent.

g. Note that

$$\begin{aligned}P(X + Y \leq x) &= \int_{-\infty}^{\infty} P(X \leq x - u \text{ and } Y \in (u, u + du)) \\ &= \int_{-\infty}^{\infty} P(X \leq x - u) f_Y(u) du.\end{aligned}$$

Taking the derivative with respect to x, we find

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(x - u) f_Y(u) du.$$

□

5.4 Summary

We explained that multiple random variables are defined on the same probability space. We discussed the joint distribution. We showed how to calculate $E(h(\boldsymbol{X}))$. In particular, we defined the variance, covariance, k-th moment. The vector notation has few secrets for you.

You also know the definition (and meaning) of independence and mutual independence and you know that the mean value of a product of independent random variables is the product of their mean values. You can also prove that functions of independent random variables are independent. We also showed that the variance of the sum of pairwise independent random variables is the sum of their variances.

5.5 Solved Problems

Example 5.5.1. *Let $X_1, \ldots, X_n$ be i.i.d. $B(p)$. Show that $X_1 + \cdots + X_n$ is $B(n, p)$.*

This follows from the definitions by computing the pmf of the sum.

Example 5.5.2. *Let X_1 and X_2 be independent and such that X_i is $Exd(\lambda_i)$ for $i = 1, 2$. Calculate*

$$P[X_1 \leq X_2 | X_1 \wedge X_2 = x].$$

(Note: $a \wedge b = min\{a, b\}$.)

We find

$$P[X_1 \leq X_2 | X_1 \wedge X_2 = x] = A/(A + B)$$

where

$$A = P(X_1 \in (x, x + dx), X_2 \geq x) = \lambda_1 \exp\{-\lambda_1 x\} dx \exp\{-\lambda_2 x\} = \lambda_1 \exp\{-(\lambda_1 + \lambda_2)x\} dx$$

and

$$B = P(X_2 \in (x, x + dx), X_1 \geq x) = \lambda_2 \exp\{-(\lambda_1 + \lambda_2)x\} dx$$

Hence,

$$P[X_1 \leq X_2 | X_1 \wedge X_2 = x] = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Example 5.5.3. *Let $\{X_n, n \geq 1\}$ be i.i.d. $U[0, 1]$. Calculate $var(X_1 + 2X_2 + X_3^2)$.*

We use the following easy but useful properties of the variance:

(i) $\text{var}(X) = E(X^2) - (E(X))^2$;

(ii) If X and Y are independent, then $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$;

(iii) For any X and $a \in \Re$, $\text{var}(aX) = a^2\text{var}(X)$.

Using these properties, we find

$$\text{var}(X_1 + 2X_2 + X_3^2) = \text{var}(X_1) + 4\text{var}(X_2) + \text{var}(X_3^2).$$

Also note that if $X =_D U[0,1]$, then

$$E(X^k) = \int_0^1 x^k dx = 1/(k+1) \text{ and } \text{var}(X) = E(X^2) - (E(X))^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Consequently,

$$\text{var}(X_1 + 2X_2 + X_3^2) = \frac{1}{12} + \frac{4}{12} + E(X_3^8) - (E(X_3^4))^2 = \frac{5}{12} + \frac{1}{9} - (\frac{1}{5})^2 \approx 0.49.$$

Example 5.5.4. *Let X, Y be i.i.d. $U[0,1]$. Compute and plot the pdf of $X+Y$.*

We use (5.3.2):

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(u) f_Y(x-u) du.$$

For a give value of x, the convolution is the integral of the product of $f_Y(u)$ and $f_X(x-u)$. The latter function is obtained by flipping $f_X(u)$ around the vertical axis and dragging it to the right by x. Figure 5.2 shows the result.

Example 5.5.5. *Let $\boldsymbol{X} = (X_1, X_2)^T$ be a vector of two i.i.d. $U[0,1]$ random variables. Let also $\boldsymbol{A}$ be a 2×2 matrix. When do the random variables $\boldsymbol{Y} = \boldsymbol{AX}$ have a j.p.d.f.? What is that j.p.d.f. when it exists? If it does not exist, how do you characterize the distribution of $\boldsymbol{Y}$.*

Assume that the two rows of $\boldsymbol{A}$ are proportional to each other. Then so are Y_1 and Y_2. In that case, the distribution of $\boldsymbol{Y}$ is concentrated on a line segment in $\Re^2$. Consequently,

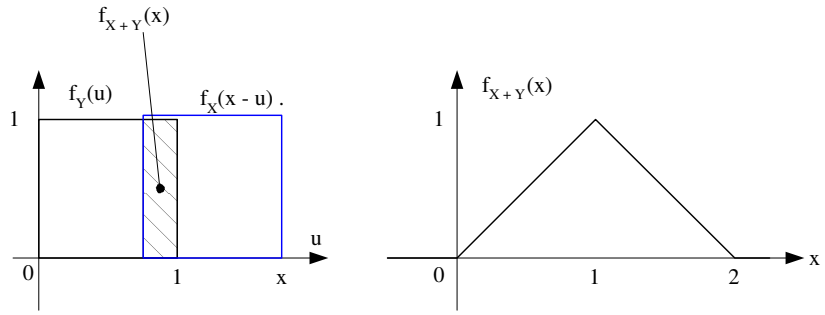


Figure 5.2: Convolution in example 5.5.4

it cannot have a density, for the integral in the plane of any function that is nonzero only on a line is equal to zero, which violates the requirement that the density must integrate to 1. Thus, if $\boldsymbol{A}$ is singular, $\boldsymbol{Y}$ has no density. We can characterize the distribution of $\boldsymbol{Y}$ by writing that $Y_2 = \alpha Y_1$ and Y_1 is a linear combination of i.i.d. $U[0,1]$ random variables. For instance, if $Y_1 = a_1 X_1 + a_2 X_2$, then we can calculate the pdf of Y_1 as the convolution of $U[0, a_1]$ and $U[0, a_2]$ as in the previous example.

If $\boldsymbol{A}$ is nonsingular, then we can write

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|A|} f_{\boldsymbol{X}}(\boldsymbol{A}^{-1}\boldsymbol{y}),$$

using the change of variables formula from calculus.

Example 5.5.6. *Let $X_1, \ldots, X_n$ be mutually independent random variables. Show that $g(X_1, \ldots, X_m)$ and $h(X_{m+1}, \ldots, X_n)$ are independent random variables for any functions $g(\cdot)$ and $h(\cdot)$ and any $m \in \{1, \ldots, n-1\}$.*

Fix A and B as subsets of $\Re$. Note that

$$g(X_1, \ldots, X_m) \in A \text{ if and only if } (X_1, \ldots, X_m) \in g^{-1}(A),$$

and similarly,

$$h(X_{m+1}, \ldots, X_n) \in B \text{ if and only if } (X_{m+1}, \ldots, X_n) \in h^{-1}(B).$$

Hence,

$$\begin{aligned}
&P(g(X_1,\ldots,X_m)\in A \text{ and } h(X_{m+1},\ldots,X_n)\in B)\\
&\quad = P((X_1,\ldots,X_m)\in g^{-1}(A) \text{ and } (X_{m+1},\ldots,X_n)\in h^{-1}(B))\\
&\quad = P((X_1,\ldots,X_m)\in g^{-1}(A))P((X_{m+1},\ldots,X_n)\in h^{-1}(B)))\\
&\quad = P(g(X_1,\ldots,X_m)\in A)P(h(X_{m+1},\ldots,X_n)\in B),
\end{aligned}$$

which proves the independence. (The next-to-last line follows from the mutual independence of the X_i.)

Example 5.5.7. *Let X, Y be two points picked independently and uniformly on the circumference of the unit circle. Define $Z = ||X - Y||^2$. Find $f_Z(.)$.*

By symmetry we can assume that the point X has coordinates $(1, 0)$. The point Y then has coordinates $(\cos(\theta), \sin(\theta))$ where θ is uniformly distributed in $[0, 2\pi]$. Consequently, $X-Y$ has coordinates $(1-\cos(\theta), -\sin(\theta))$ and $Z = (1-\cos(\theta))^2+\sin^2(\theta) = 2(1-\cos(\theta)) =: g(\theta)$.

We now use the basic results on the density of a function of a random variable. To review how this works, note that if $\theta \in (\theta_0, \theta_0 + \epsilon)$, then

$$g(\theta) \in (g(\theta_0), g(\theta_0 + \epsilon)) = (g(\theta_0), g(\theta_0) + g\prime(\theta_0)\epsilon).$$

Accordingly,

$$g(\theta) \in (z, z + \delta)$$

if and only if

$$\theta \in ((\theta_n, \theta_n + \frac{\delta}{g\prime(\theta_n)})$$

for some θ_n such that $g(\theta_n) = z$. It follows that, if $Z = g(\theta)$, then

$$f_Z(z) = \sum_n \frac{1}{|g\prime(\theta_n)|} f_\theta(\theta_n).$$

In this expression, the sum is over all the θ_n such that $g(\theta_n) = z$.

Coming back to our example, $g(\theta) = z$ if $2(1 - \cos(\theta)) = z$. In that case, $|g\prime(\theta)| = 2|\sin(\theta)| = 2\sqrt{1 - (1 - \frac{z}{2})^2}$. Note that there are two values of θ such that $g(\theta) = z$ whenever $z \in (0, 4)$. Accordingly,

$$f_Z(z) = 2 \times \frac{1}{2\sqrt{1 - (1 - \frac{z}{2})^2}} \times \frac{1}{2\pi} = \frac{1}{2\pi\sqrt{z - \frac{z^2}{4}}}, \text{ for } z \in (0, 4).$$

Example 5.5.8. *The two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are selected independently and uniformly in $[-1, 1]^2$. Calculate $E(||\boldsymbol{X} - \boldsymbol{Y}||^2)$.*

Let $\boldsymbol{X} = (X_1, X_2)$ and similarly for $\boldsymbol{Y}$. Then

$$E(||\boldsymbol{X} - \boldsymbol{Y}||^2) = E(|X_1 - Y_1|^2 + |X_2 - Y_2|^2) = 2E(|X_1 - Y_1|^2),$$

by symmetry.

Now,

$$E(|X_1 - Y_1|^2) = E(X_1^2) + E(Y_1)^2 - 2E(X_1Y_1) = 2E(X_1^2)$$

since X_1 and Y_1 are independent and zero-mean.

Also,

$$E(X_1^2) = \int_{-1}^{1} x^2 \frac{1}{2} dx = [\frac{x^3}{6}]_{-1}^{1} = \frac{1}{3}.$$

Finally, putting the pieces together, we get

$$E(||\boldsymbol{X} - \boldsymbol{Y}||^2) = 4E(X_1^2) = \frac{4}{3}.$$

Example 5.5.9. *Let $\{X_n, n \geq 1\}$ be i.i.d. $B(p)$. Assume that $g, h : \Re^n \to \Re$ have the property that if $\boldsymbol{x}, \boldsymbol{y} \in \Re^n$ are such that $x_i \leq y_i$ for $i = 1, \ldots, n$, then $g(\boldsymbol{x}) \leq g(\boldsymbol{y})$ and $h(\boldsymbol{x}) \leq h(\boldsymbol{y})$. Show, by induction on n, that $cov(g(\boldsymbol{X}), h(\boldsymbol{X})) \geq 0$ where $\boldsymbol{X} = (X_1, \ldots, X_n)$.*

The intuition is that $g(\boldsymbol{X})$ and $h(\boldsymbol{X})$ are large together and small together.

For $n = 1$ this is easy. We must show that $\text{cov}(g(X_1)h(X_1)) \geq 0$. By redefining $\tilde{g}(x) = g(x) - g(0)$ and $\tilde{h}(x) = h(x) - h(0)$, we see that it is equivalent to show that $\text{cov}(\tilde{g}(X_1), \tilde{h}(X_1)) \geq 0$. In other words, we can assume without loss of generality that $g(0) = h(0) = 0$. If we do that, we need only show that

$$E(g(X_1)h(X_1)) = pg(1)h(1) \geq E(g(X_1))E(h(X_1)) = pg(1)ph(1)$$

which is seen to be satisfied since $g(1)$ and $h(1)$ are nonnegative and $p \leq 1$.

Assume that the result is true for n. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $V = X_{n+1}$. We must show that

$$E(g(\boldsymbol{X}, V)h(\boldsymbol{X}, V)) \geq E(g(\boldsymbol{X}, V)E(h(\boldsymbol{X}, V)).$$

We know, by the induction hypothesis, that

$$E(g(\boldsymbol{X}, i)h(\boldsymbol{X}, i)) \geq E(g(\boldsymbol{X}, i)E(h(\boldsymbol{X}, i)), \text{ for } i = 0, 1.$$

Assume, without loss of generality, that

$$E(g(\boldsymbol{X}, 0)) = 0.$$

Then we know that

$$E(g(\boldsymbol{X}, 1)) \geq 0 \text{ and } E(g(\boldsymbol{X}, 0)h(\boldsymbol{X}, 0)) \geq 0$$

and

$$E(h(\boldsymbol{X}, V)) \leq E(h(\boldsymbol{X}, 1)),$$

so that

$$\begin{aligned} E(g(\boldsymbol{X}, V))E(h(\boldsymbol{X}, V)) &= pE(g(\boldsymbol{X}, 1))E(h(\boldsymbol{X}, V)) \leq pE(g(\boldsymbol{X}, 1))E(h(\boldsymbol{X}, 1)) \\ &\leq pE(g(\boldsymbol{X}, 1)h(\boldsymbol{X}, 1)) \leq pE(g(\boldsymbol{X}, 1)h(\boldsymbol{X}, 1)) + (1-p)E(g(\boldsymbol{X}, 0)h(\boldsymbol{X}, 0)) \\ &= E(g(\boldsymbol{X}, V)h(\boldsymbol{X}, V)), \end{aligned}$$

which completes the proof.

Example 5.5.10. *Let X be uniformly distributed in $[0, 2\pi]$ and $Y = \sin(X)$. Calculate the p.d.f. f_Y of Y.*

Since $Y = g(X)$, we know that

$$f_Y(y) = \sum \frac{1}{|g'(x_n)|} f_X(x_n)$$

where the sum is over all the x_n such that $g(x_n) = y$.

For each $y \in (-1, 1)$, there are two values of x_n in $[0, 2\pi]$ such that $g(x_n) = \sin(x_n) = y$. For those values, we find that

$$|g'(x_n)| = |\cos(x_n)| = \sqrt{1 - \sin^2(x_n)} = \sqrt{1 - y^2},$$

and

$$f_X(x_n) = \frac{1}{2\pi}.$$

Hence,

$$f_Y(y) = 2\frac{1}{\sqrt{1-y^2}}\frac{1}{2\pi} = \frac{1}{\pi\sqrt{1-y^2}}.$$

Example 5.5.11. *Let $\{X, Y\}$ be independent random variables with X exponentially distributed with mean 1 and Y uniformly distributed in $[0, 1]$. Calculate $E(\max\{X, Y\})$.*

Let $Z = \max\{X, Y\}$. Then

$$\begin{aligned} P(Z \le z) &= P(X \le z, Y \le z) = P(X \le z)P(Y \le z) \\ &= \begin{cases} z(1 - e^{-z}), \text{ for } z \in [0, 1] \\ 1 - e^{-z}, \text{ for } z \ge 1. \end{cases} \end{aligned}$$

Hence,

$$f_Z(z) = \begin{cases} 1 - e^{-z} + ze^{-z}, \text{ for } z \in [0, 1] \\ e^{-z}, \text{ for } z \ge 1. \end{cases}$$

Accordingly,

$$E(Z) = \int_0^\infty z f_Z(z) dz = \int_0^1 z(1 - e^{-z} + ze^{-z}) dz + \int_1^\infty ze^{-z} dz$$

To do the calculation we note that

$$\int_0^1 z dz = [z^2/2]_0^1 = 1/2,$$

$$\begin{aligned} \int_0^1 ze^{-z} dz &= -\int_0^1 z de^{-z} = -[ze^{-z}]_0^1 + \int_0^1 e^{-z} dz \\ &= -e^{-1} - [e^{-z}]_0^1 = 1 - 2e^{-1}. \end{aligned}$$

$$\begin{aligned} \int_0^1 z^2 e^{-z} dz &= -\int_0^1 z^2 de^{-z} = -[z^2 e^{-z}]_0^1 + \int_0^1 2ze^{-z} dz \\ &= -e^{-1} + 2(1 - 2e^{-1}) = 2 - 5e^{-1}. \end{aligned}$$

$$\int_1^\infty ze^{-z} dz = 1 - \int_0^1 ze^{-z} dz = 2e^{-1}.$$

Collecting the pieces, we find that

$$E(Z) = \frac{1}{2} - (1 - 2e^{-1}) + (2 - 5e^{-1}) + 2e^{-1} = 3 - 5e^{-1} \approx 1.16.$$

Example 5.5.12. *Let $\{X_n, n \geq 1\}$ be i.i.d. with $E(X_n) = \mu$ and var$(X_n) = \sigma^2$. Use Chebyshev's inequality to get a bound on*

$$\alpha := P(|\frac{X_1 + \cdots + X_n}{n} - \mu| \geq \epsilon).$$

Chebyshev's inequality (4.8.1) states that

$$\alpha \leq \frac{1}{\epsilon^2} \text{var}(\frac{X_1 + \cdots + X_n}{n}) = \frac{1}{\epsilon^2} \frac{n\text{var}(X_1)}{n^2} = \frac{\sigma^2}{n\epsilon^2}.$$

This calculation shows that the sample mean gets closer and closer to the mean: the variance of the error decreases like $1/n$.

Example 5.5.13. *Let $X =_D P(\lambda)$. You pick X white balls. You color the balls independently, each red with probability p and blue with probability $1-p$. Let Y be the number of red balls and Z the number of blue balls. Show that Y and Z are independent and that $Y =_D P(\lambda p)$ and $Z =_D P(\lambda(1-p))$.*

We find

$$\begin{aligned} P(Y=m, Z=n) &= P(X=m+n)\binom{m+n}{m}p^m(1-p)^n \\ &= \frac{\lambda^{m+n}}{(m+n)!}\binom{m+n}{m}p^m(1-p)^n = \frac{\lambda^{m+n}}{(m+n)!}\times\frac{(m+n)!}{m!n!}p^m(1-p)^n \\ &= [\frac{(\lambda p)^m}{m!}e^{-\lambda p}]\times[\frac{(\lambda(1-p))^n}{n!}e^{-\lambda(1-p)}], \end{aligned}$$

which proves the result.

Chapter 6

Conditional Expectation

Conditional expectation tells us how to use the observation of a random variable $Y(\omega)$ to estimate another random variable $X(\omega)$. This conditional expectation is the best guess about $X(\omega)$ given $Y(\omega)$ if we want to minimize the mean squared error. Of course, the value of this conditional expectation is a function of $Y(\omega)$, so that it is also a random variable. We will learn to calculate the conditional expectation.

6.1 Examples

6.1.1 Example 1

Assume that the pair of random variables (X, Y) is discrete and takes values in $\{x_1, \ldots, x_m\} \times \{y_1, \ldots, y_n\}$. This pair of random variables is defined by specifying $p(i, j) = P(X = x_i, Y = y_j), i = 1, \ldots, m; j = 1, \ldots, n$. From this information, we can derive

$$P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i p(i, j).$$

We can then calculate $P[X = x_i | Y = y_j] = P(X = x_i, Y = y_j)/P(Y = y_j)$ and we define

$$E[X|Y = y_j] = \sum_i x_i P[X = x_i | Y = y_j].$$

We then define $E[X|Y] = E[X|Y = y_j]$ when $Y = y_j$. In other words,

$$E[X|Y] = \sum_j E[X|Y = y_j]1\{Y = y_j\}.$$

Note that $E[X|Y]$ is a random variable.

For instance, your guess about the temperature in San Francisco certainly depends on the temperature you observe in Berkeley. Since the latter is random, so is your guess about the former.

Although this definition is sensible, it is not obvious in what sense this is the best guess about X given $\{Y = y_j\}$. We discuss this below.

6.1.2 Example 2

Consider the case where (X, Y) have a joint density $f(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$. One can then define the conditional density of X given that $Y = y$ as follows. We see that

$$P[x < X \leq x + \epsilon | y < Y \leq y + \delta] = \frac{f(x, y)\epsilon\delta}{f_Y(y)\delta} = \frac{f(x, y)\epsilon}{f_Y(y)} =: f_{X|Y}[x|y]\epsilon.$$

As δ goes down to zero, we see that $f_{X|Y}[x|y]$ is the conditional density of X given $\{Y = y\}$. We then define

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}[x|y]dx. \tag{6.1.1}$$

6.1.3 Example 3

The ideas of Examples 1 and 2 extend to hybrid cases. For instance, consider the situation illustrated in Figure 6.1:

The figure shows the joint distribution of (X, Y). With probability $0.4, (X, Y) = (0.75, 0.25)$. Otherwise (with probability 0.6), the pair (X, Y) is picked uniformly in the

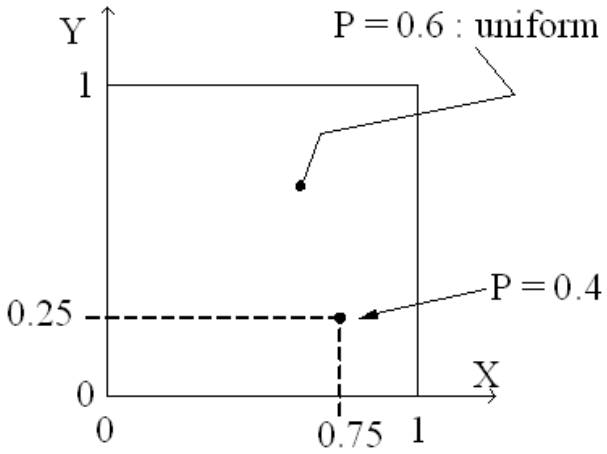


Figure 6.1: Hybrid Distribution: Neither discrete nor continuous

square $[0,1]^2$. You see that $E[X|Y=y] = 0.5$ if $y \neq 0.25$. Also, if $Y = 0.25$, then $X = 0.75$, so that $E[X|Y = 0.25] = 0.75$.

Thus $E[X|Y] = g(Y)$ where $g(0.25) = 0.75$ and $g(y) = 0.5$ for $y \neq 0.25$.

In this case, $E[X|Y]$ is a random variable such that $E[X|Y] = 0.5$ w.p. 0.6 and $E[X|Y] = 0.75$ w.p. 0.4.

(Note that the expected value of $E[X|Y]$ is $0.5 \times 0.6 + 0.75 \times 0.4 = 0.6$ and you can observe that $E(X) = 0.5 \times 0.6 + 0.75 \times 0.4 = 0.6$. That is, $E(E[X|Y]) = E(X)$ and we will learn that this is always the case.)

6.2 MMSE

The examples that we have explored led us to define $E[X|Y]$ as the expected value of X when it has its conditional distribution given the value of Y. In this section, we explain that $E[X|Y]$ can be defined as the function $g(Y)$ of the observed value Y that minimizes $E((X - g(Y)^2)$. That is, $E[X|Y]$ is the best guess about X that is based on Y, where best means that it minimizes the mean squared error.

To verify that fact, choose an arbitrary function $g(Y)$. We show that

$$E((X - g(Y))^2) \geq E((X - E[X|Y])^2).$$

The first step of the derivation is to write

$$\begin{aligned} E((X - g(Y))^2) &= E((X - E[X|Y] + E[X|Y] - g(Y))^2) \\ &= E((X - E[X|Y])^2) + E((E[X|Y] - g(Y))^2) + 2E((X - E[X|Y])(E[X|Y] - g(Y))) \\ &= E((X - E[X|Y])^2) + E((E[X|Y] - g(Y))^2) + 2E((X - E[X|Y])h(Y)) \end{aligned} \tag{6.2.1}$$

where $h(Y) := E[X|Y] - g(Y)$.

The second step is to show that the last term in (6.2.1) is equal to zero. To show that, we calculate

$$\begin{aligned} E(h(Y)E[X|Y]) &= \int h(y)E[X|Y = y]f_Y(y)dy = \int h(y)\{\int x f_{X|Y}[x|y]dx\}\ f_Y(y)dy \\ &= \int\int xh(y)f_{X,Y}(x,y)dxdy = E(h(Y)X). \end{aligned} \tag{6.2.2}$$

The next-to-last identity uses the fact that $f_{X|Y}[x|y]f_Y(y) = f_{X,Y}(x,y)$, by definition of the conditional density.

The final step is to observe that (6.2.1) with the last term equal to zero implies that

$$E((X-g(Y))^2) = E((X-E[X|Y])^2)+E((E[X|Y]-g(Y))^2) \geq E((X-E[X|Y])^2). \tag{6.2.3}$$

This is the story when joint densities exist. The derivation can be adapted to the case when joint densities do not exist.

6.3 Two Pictures

The left-hand part of Figure 6.2 shows that $E[X|Y]$ is the average value of X on sets that correspond to a constant value of Y. The figure also highlights the fact that $E[X|Y]$ is a random variable.

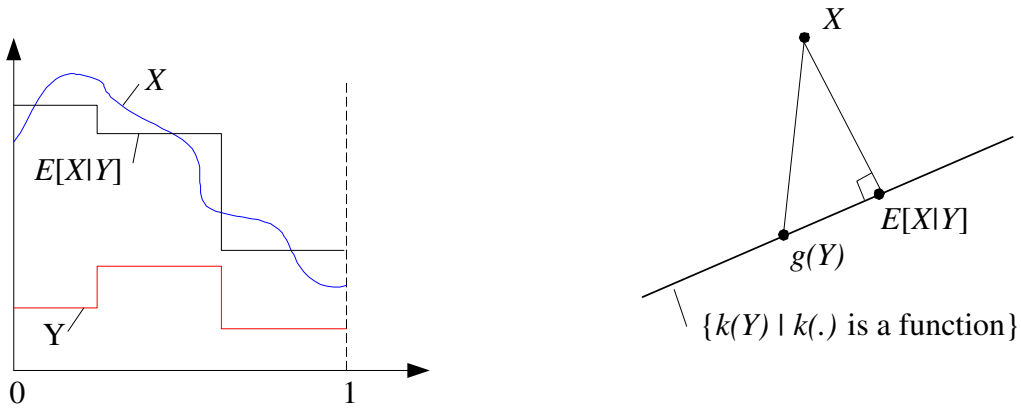


Figure 6.2: Conditional Expectation - Pictures

The right-hand part of Figure 6.2 depicts random variables as points in some vector space. The figure shows that $E[X|Y]$ is the function of Y that is closest to X. The metric in the space is $d(V,W) = (E(V-W)^2)^{1/2}$. That figure illustrates the relations (6.2.3). These relations are a statement of Pythagora's theorem: the square of the length of the hypothenuse $d^2(X, g(Y))$ is the sum of the squares of the sides of the right triangle $d^2(X, E[X|Y]) + d^2(E[X|Y], g(Y))$. This figure shows that $E[X|Y]$ is the *projection* of X onto the hyperplane $\{k(Y) \mid k(\cdot) \text{ is a function }\}$. The figure also shows that for $E[X|Y]$ to be that projection, the vector $X - E[X|Y]$ must be orthogonal to every function $k(Y)$, and in particular to $E[X|Y] - g(Y)$, as (6.2.2) states.

To give you a concrete feel for this vector space, imagine that $\Omega = \{\omega_1, \ldots, \omega_N\}$ and that p_k is the probability that ω is equal to ω_k. In that case, the random variable X corresponds to the vector $(X(\omega_1)(p_1)^{1/2}, \ldots, X(\omega_N)(p_N)^{1/2})$ in $\Re^N$. For a general Ω, the random variable X is a function of ω and it belongs to a function space. This space is a vector space since linear combinations of functions are also functions. If we restrict our attention to random variables X with $E(X^2) < \infty$, that space, with the metric that one defines, turns out to be closed under limits of convergent sequences. (Such a space is called a Hilbert space.) This property is very useful because it implies that as one chooses functions $g_n(Y)$ whose

distances to X approach the minimum distance, these functions converge to a function $g(Y)$. This argument implies the existence of conditional expectation. The uniqueness is intuitively clear: if two random variables $g(Y)$ and $g'(Y)$ achieve the minimum distance to X, they must be equal.

Thus, we can *define* $E[X|Y]$ as the function $g(Y)$ that minimizes $E((X-g(Y))^2)$. This definition does not assume the existence of a conditional density $f_{X|Y}[\cdot|\cdot]$, nor of a joint pmf.

6.4 Properties of Conditional Expectation

One often calculates the conditional expectation by using the properties of that operator. We derive these properties in this section. We highlight the key observation we made in the derivation of the MMSE property of conditional expectation as the following lemma.

Lemma 6.4.1. *A random variable $g(Y)$ is equal to $E[X|Y]$ if and only if*

$$E(g(Y)h(Y)) = E(Xh(Y)) \text{ for any function } h(\cdot). \tag{6.4.1}$$

We proved in (6.2.2) that $E[X|Y]$ satisfies that property. To show that a function that satisfies (6.4.1) is the conditional expectation, one observes that if both $g(Y)$ and $g'(Y)$ satisfy that condition, they must be equal. To see that, note that

$$\begin{aligned} E((X-g(Y))^2) &= E((X-g'(Y)+g'(Y)-g(Y))^2) \\ &= E((X-g'(Y))^2) + E((g'(Y)-g(Y))^2) + 2E((g'(Y)-g(Y))(X-g'(Y))). \end{aligned}$$

Using (6.4.1), we see that the last term is equal to zero. Hence,

$$E((X-g(Y))^2) = E((X-g'(Y))^2) + E((g(Y)-g'(Y))^2).$$

But we assume that $E((X-g(Y))^2) = E((X-g'(Y))^2)$. Consequently, $E((g(Y)-g'(Y))^2) = 0$, which implies that $g(Y)-g'(Y)=0$ (with probability 1).

Using this lemma, we can prove the following properties.

Theorem 6.4.2. *Properties of Conditional Expectation*

a. Linearity:

$$E[a_1X_1 + a_2X_2 \mid Y] = a_1E[X_1 \mid Y] + a_2E[X_2 \mid Y]; \tag{6.4.2}$$

b. Known Factor:

$$E[Xk(Y) \mid Y] = k(Y)E[X \mid Y]; \tag{6.4.3}$$

c. Averaging:

$$E(E[X \mid Y]) = E(X); \tag{6.4.4}$$

d. Independence: If X and Y are independent, then

$$E[X \mid Y] = E(X). \tag{6.4.5}$$

e. Smoothing:

$$E[E[X \mid Y, Z] \mid Y] = E[X \mid Y]. \tag{6.4.6}$$

Proof:

The derivation of these identities is a simple exercise, but going through it should help you appreciate the mechanics.

a. Linearity is fairly clear from our original definition (6.1.1), when the conditional density exists. In the general case, we can use the lemma as follows. We do this derivation step by step as the other properties follow the same pattern.

To show that the function $g(Y) := a_1E[X_1 \mid Y] + a_2E[X_2 \mid Y]$ is in fact $E[X|Y]$ with $X := a_1X_1 + a_2X_2$, we show that it satisfies (6.4.1). That is, we must show that

$$E((a_1E[X_1 \mid Y] + a_2E[X_2 \mid Y])h(Y)) = E((a_1X_1 + a_2X_2)h(Y)) \tag{6.4.7}$$

for all $h(\cdot)$. But we know that, for $i = 1, 2$,

$$E((a_iE[X_i \mid Y])h(Y)) = E(E[X_i \mid Y](a_ih(Y))) = E(X_i(a_ih(Y))) = E(a_iX_ih(Y)). \tag{6.4.8}$$

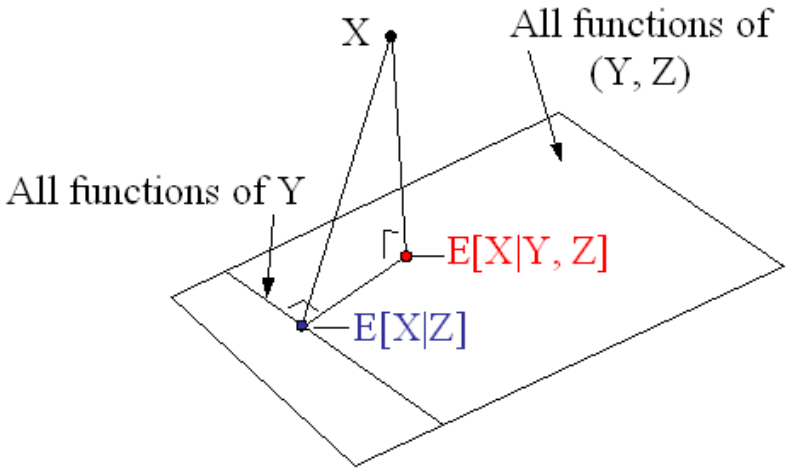


Figure 6.3: Conditional Expectation - Smoothing Property

Indeed, the third identity follows from the property (6.4.1) applied to $E[X_i \mid Y]$ with $h(Y)$ replaced by $a_i h(Y)$. The other identities are obvious.

By adding up the expressions (6.4.8) for $i = 1, 2$, one gets (6.4.7).

b. To show that $g(Y) := k(Y)E[X|Y]$ is equal to $E[Xk(Y)|Y]$ we prove that it satisfies (6.4.1). That is, we must show that

$$E(k(Y)E[X|Y]h(Y)) = E(Xk(Y)h(Y)), \ \forall h(\cdot).$$

This identity follows from the property (6.4.1) of $E[X|Y]$ where one replaces $h(Y)$ by $k(Y)h(Y)$.

c. The averaging property is a particular case of (6.4.1) where $h(Y) = 1$.

d. We use the lemma. Let $g(Y) = E[X|Y]$. We show that for any $h(Y)$ one has $E(g(Y)h(Y)) = E(E[X|Y, Z]h(Y))$. Now, $E(E[X|Y, Z]h(Y)) = E(Xh(Y))$ by (6.4.1) and $E(g(Y)h(Y)) = E(E[X|Y]h(Y)) = E(Xh(Y))$, also by (6.4.1). This completes the proof.

□

Figure 6.3 shows why $E[E[X|Y, Z]|Y] = E[X|Y]$.

6.5 Gambling System

Conditional expectation provides a way to evaluate gambling systems. Let $\{X_n, n \geq 1\}$ be i.i.d. random variables with $P(X_n = -1) = P(X_n = 1) = 0.5$. The random variable X_n represents your gain at the n-th game of roulette, playing black or red and assuming that there is no house advantage (no 0 nor double-zero). Say that you have played n times and observed $(X_1, X_2, \ldots, X_n)$. You then calculate the amount $Y_n = h_n(X_1, X_2, \ldots, X_n)$ that you gamble on the next game. You earn $Y_n X_{n+1}$ on that next game. After a number of such games, you have accumulated

$$Z = Y_0 X_1 + Y_1 X_2 + \cdots + Y_n X_{n+1}.$$

(Here, Y_0 is some arbitrary initial bet.) Assume that the random variables Y_n are bounded (which is not unreasonable since there may be a table limit), then you find that

$$E(Y_n X_{n+1}) = E(E[Y_n X_{n+1} | X_1, X_2, \ldots, X_n]) = E(Y_n E[X_{n+1} | X_1, X_2, \ldots, X_n]) = E(Y_n 0) = 0.$$

Consequently, $E(Z) = 0$. This result shows the "impossibility" of a gambling system and guarantees that the casinos will be doing well.

6.6 Summary

The setup is that (X, Y) are random variables on some common probability space, i.e., with some joint distribution. We observe Y and want to estimate X.

The minimum mean squares estimator of X given Y is defined as the function $g(Y)$ that minimizes $E((X - g(Y))^2)$. We know that the answer is $g(Y) = E[X \mid Y]$. How do we calculate it?

Direct Calculation

The direct calculation uses (6.1.1) or the discrete version. We look at hybrid cases in the examples.

Symmetry

Assume $X_1, \ldots, X_n$ are i.i.d. Then

$$E[X_1 + \cdots + X_m \mid X_1 + \cdots + X_n] = \frac{m}{n} \times (X_1 + \cdots + X_n) \text{ for } 1 \leq m \leq n.$$

Note also that

$$E[X_1 + \cdots + X_m \mid X_1 + \cdots + X_n] = X_1 + \cdots + X_n + (m-n)E(X_1) \text{ for } 1 \leq n \leq m.$$

To derive these identities, we first note that

$$E[X_i \mid X_1 + \cdots + X_n] = Y, \text{ for } i = 1, \ldots, n$$

where Y is some random variable. Second, by (6.4.2), if we add up these identities for $i = 1, \ldots, n$, we find

$$nY = E[X_1 + \cdots + X_n \mid X_1 + \cdots + X_n] = X_1 + \cdots + X_n.$$

Hence,

$$Y = (X_1 + \cdots + X_n)/n,$$

so that

$$E[X_i \mid X_1 + \cdots + X_n] = (X_1 + \cdots + X_n)/n, \text{ for } i = 1, \ldots, n.$$

Using these identities we can now derive the two properties stated above.

Properties

Often one can use the properties of conditional expectation states in Theorem 6.4.2 to calculate $E[X \mid Y]$.

6.7 Solved Problems

Example 6.7.1. *Let (X, Y) be a point picked uniformly in the quarter circle $\{(x, y) \mid x \geq 0, y \geq 0, x^2 + y^2 \leq 1\}$. Find $E[X \mid Y]$.*

Given $Y = y$, X is uniformly distributed in $[0, \sqrt{1 - y^2}]$. Hence

$$E[X \mid Y] = \frac{1}{2}\sqrt{1 - Y^2}.$$

Example 6.7.2. *A customer entering a store is served by clerk i with probability p_i, $i = 1, 2, \ldots, n$. The time taken by clerk i to service a customer is an exponentially distributed random variable with parameter α_i.*

a. Find the pdf of T, the time taken to service a customer.

b. Find $E[T]$.

c. Find $Var[T]$.

Designate by X the clerk who serves the customer.

a. $f_T(t) = \sum_{i=1}^n p_i f_{T|X}[t|i] = \sum_{i=1}^n p_i \alpha_i e^{-\alpha_i t}$

b. $E[T] = E(E[T \mid X]) = E(\frac{1}{\alpha_X}) = \sum_{i=1}^n p_i \frac{1}{\alpha_i}$.

c. We first find $E[T^2] = E(E[T^2 \mid X]) = E(\frac{1}{\alpha_i^2}) = \sum_{i=1}^n p_i \frac{2}{\alpha_i^2}$. Hence, $\text{var}(T) = E(T^2) - (E(T))^2 = \sum_{i=1}^n p_i \frac{2}{\alpha_i^2} - (\sum_{i=1}^n p_i \frac{1}{\alpha_i})^2$.

Example 6.7.3. *The random variables X_i are i.i.d. and such that $E[X_i] = \mu$ and $var(X_i) = \sigma^2$. Let N be a random variable independent of all the X_is taking on nonnegative integer values. Let $S = X_1 + X_2 + \ldots + X_N$.*

a. Find $E(S)$.

b. Find $var(S)$.

a. $E(S)] = E(E[S \mid N]) = E(N\mu) = \mu E(N)$.

b. First we calculate $E(S^2)$. We find

$$\begin{aligned} E(S^2) &= E(E[S^2 \mid N]) = E(E[(X_1 + X_2 + \ldots + X_N)^2 \mid N]) \\ &= E(E[X_1^2 + \cdots + X_N^2 + \sum_{i \neq j} X_i X_j \mid N]) \\ &= E(NE(X_1^2) + N(N-1)E(X_1 X_2)) = E(N(\mu^2 + \sigma^2) + N(N-1)\mu^2) \\ &= E(N)\sigma^2 + E(N^2)\mu^2. \end{aligned}$$

Then,

$$\mathrm{var}(S) = E(S^2) - (E(S))^2 = E(N)\sigma^2 + E(N^2)\mu^2 - \mu^2(E(N))^2 = E(N)\sigma^2 + \mathrm{var}(N)\mu^2.$$

Example 6.7.4. *Let X, Y be independent and uniform in $[0,1]$. Calculate $E[X^2 \mid X+Y]$.*

Given $X + Y = z$, the point (X, Y) is uniformly distributed on the line $\{(x, y) \mid x \geq 0, y \geq 0, x + y = z\}$. Draw a picture to see that if $z > 1$, then X is uniform on $[z-1, 1]$ and if $z < 1$, then X is uniform on $[0, z]$. Thus, if $z > 1$ one has

$$E[X^2 \mid X + Y = z] = \int_{z-1}^{1} x^2 \frac{1}{2-z} dx = \frac{1}{2-z}[\frac{x^3}{3}]_{z-1}^{1} = \frac{1-(z-1)^3}{3(2-z)}.$$

Similarly, if $z < 1$, then

$$E[X^2 \mid X + Y = z] = \int_{0}^{z} x^2 \frac{1}{z} dx = \frac{1}{z}[\frac{x^3}{3}]_{0}^{z} = \frac{z^2}{3}.$$

Example 6.7.5. *Let (X, Y) be the coordinates of a point chosen uniformly in $[0,1]^2$. Calculate $E[X \mid XY]$.*

This is an example where we use the straightforward approach, based on the definition. The problem is interesting because is illustrates that approach in a tractable but nontrivial example. Let $Z = XY$.

$$E[X \mid Z = z] = \int_{0}^{1} x f_{[X|Z]}[x \mid z] dx.$$

Now,

$$f_{[X|Z]}[x \mid z] = \frac{f_{X,Z}(x,z)}{f_Z(z)}.$$

Also,

$$\begin{aligned} f_{X,Z}(x,z)dxdz &= P(X \in (x, x+dx), Z \in (z, z+dz)) \\ &= P(X \in (x, x+dx))P[Z \in (z, z+dz) \mid X = x] = dxP(xY \in (z, z+dz)) \\ &= dxP(Y \in (\frac{z}{x}, \frac{z}{x} + \frac{dz}{x})) = dx\frac{dz}{x}1\{z \le x\}. \end{aligned}$$

Hence,

$$f_{X,Z}(x,z) = \begin{cases} \frac{1}{x}, \text{ if } x \in [0,1] \text{ and } z \in [0,x] \\ 0, \text{ otherwise.} \end{cases}$$

Consequently,

$$f_Z(z) = \int_0^1 f_{X,Z}(x,z)dx = \int_z^1 \frac{1}{x}dx = -ln(z), 0 \le z \le 1.$$

Finally,

$$f_{[X|Z]}[x \mid z] = -\frac{1}{xln(z)}, \text{ for } x \in [0,1] \text{ and } z \in [0,x],$$

and

$$E[X \mid Z = z] = \int_z^1 x(-\frac{1}{xln(z)})dx = \frac{z-1}{ln(z)},$$

so that

$$E[X \mid XY] = \frac{XY-1}{ln(XY)}.$$

Examples of values:

$E[X \mid XY = 1] = 1, E[X \mid XY = 0.1] = 0.39, E[X \mid XY \approx 0] \approx 0.$

Example 6.7.6. *Let X, Y be independent and exponentially distributed with mean 1. Find $E[\cos(X+Y) \mid X]$.*

We have

$$E[\cos(X+Y) \mid X=x] = \int_0^\infty \cos(x+y)e^{-y}dy = Re\{\int_0^\infty e^{i(x+y)-y}dy\}$$
$$= Re\{\frac{e^{ix}}{1-i}\} = \frac{\cos(x)-\sin(x)}{2}.$$

Example 6.7.7. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $U[0,1]$ and $Y = max\{X_1, \ldots, X_n\}$. Calculate $E[X_1 \mid Y]$.*

Intuition suggests, and it is not too hard to justify, that if $Y = y$, then $X_1 = y$ with probability $1/n$, and with probability $(n-1)/n$ the random variable X_1 is uniformly distributed in $[0, y]$. Hence,

$$E[X_1 \mid Y] = \frac{1}{n}Y + \frac{n-1}{n}\frac{Y}{2} = \frac{n+1}{2n}Y.$$

Example 6.7.8. *Let X, Y, Z be independent and uniform in $[0,1]$. Calculate $E[(X+2Y+Z)^2 \mid X]$.*

One has, $E[(X+2Y+Z)^2 \mid X] = E[X^2+4Y^2+Z^2+4XY+4YZ+2XZ \mid X]$. Now,

$$E[X^2+4Y^2+Z^2+4XY+4YZ+2XZ \mid X]$$
$$= X^2 + 4E(Y^2) + E(Z^2) + 4XE(Y) + 4E(Y)E(Z) + 2XE(Z)$$
$$= X^2 + 4/3 + 1/3 + 2X + 1 + X = X^2 + 3X + 8/3.$$

Example 6.7.9. *Let X, Y, Z be three random variables defined on the same probability space. Prove formally that*

$$E(|X - E[X \mid Y]|^2) \geq E(|X - E[X \mid Y, Z]|^2).$$

Let $X_1 = E[X \mid Y]$ and $X_2 = E[X \mid Y, Z]$. Note that

$$E((X - X_2)(X_2 - X_1)) = E(E[(X - X_2)(X_2 - X_1) \mid Y, Z])$$

and

$$E[(X - X_2)(X_2 - X_1) \mid Y, Z] = (X_2 - X_1)E[X - X_2 \mid Y, Z] = X_2 - X_2 = 0.$$

Hence,

$$E((X-X_1)^2) = E((X-X_2+X_2-X_1)^2) = E((X-X_2)^2)+E((X_2-X_1)^2) \geq E((X-X_2)^2).$$

Example 6.7.10. *Pick the point* (X, Y) *uniformly in the triangle* $\{(x, y) \mid 0 \leq x \leq 1$ *and* $0 \leq y \leq x\}$.

a. Calculate $E[X \mid Y]$.

b. Calculate $E[Y \mid X]$.

c. Calculate $E[(X - Y)^2 \mid X]$.

a. Given $\{Y = y\}$, X is $U[y, 1]$, so that $E[X \mid Y = y] = (1 + y)/2$. Hence,

$$E[X \mid Y] = \frac{1+Y}{2}.$$

b. Given $\{X = x\}$, Y is $U[0, x]$, so that $E[Y \mid X = x] = x/2$. Hence,

$$E[Y \mid X] = \frac{X}{2}.$$

c. Since given $\{X = x\}$, Y is $U[0, x]$, we find

$$E[(X - Y)^2 \mid X = x] = \int_0^x (x - y)^2 \frac{1}{x} dy = \frac{1}{x} \int_0^x y^2 dy = \frac{x^2}{3}. \text{ Hence,}$$

$$E[(X - Y)^2 \mid X] = \frac{X^2}{3}.$$

Example 6.7.11. *Assume that the two random variables* X *and* Y *are such that* $E[X \mid Y] = Y$ *and* $E[Y \mid X] = X$. *Show that* $P(X = Y) = 1$.

We show that $E((X - Y)^2) = 0$. This will prove that $X - Y = 0$ with probability one. Note that

$$E((X - Y)^2) = E(X^2) - E(XY) + E(Y^2) - E(XY).$$

Now,

$$E(XY) = E(E[XY \mid X]) = E(XE[Y \mid X]) = E(X^2).$$

Similarly, one finds that $E(XY) = E(Y^2)$. Putting together the pieces, we get $E((X - Y)^2) = 0$.

Example 6.7.12. *Let X, Y be independent random variables uniformly distributed in $[0, 1]$. Calculate $E[X|X < Y]$.*

Drawing a unit square, we see that given $\{X < Y\}$, the pair (X, Y) is uniformly distributed in the triangle left of the diagonal from the upper left corner to the bottom right corner of that square. Accordingly, the p.d.f. $f(x)$ of X is given by $f(x) = 2(1-x)$. Hence,

$$E[X|X < Y] = \int_0^1 x \times 2(1-x)dx = \frac{1}{3}.$$

Chapter 7

Gaussian Random Variables

Gaussian random variables show up frequently. (This is because of the central limit theorem that we discuss later in the class.) Here are a few essential properties that we explain in the chapter.

- The Gaussian distribution is determined by its mean and variance.
- The sum of independent Gaussian random variables is Gaussian.
- Random variables are jointly Gaussian if an arbitrary linear combination is Gaussian.
- Uncorrelated jointly Gaussian random variables are independent.
- If random variables are jointly Gaussian, then the conditional expectation is linear.

7.1 Gaussian

7.1.1 $N(0,1)$: Standard Gaussian Random Variable

Definition 7.1.1. We say that X is a *standard Gaussian* (or *standard Normal*) random variable, and we write $X =_D N(0,1)$, if

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}, \text{ for } x \in \Re.$$

To see that $f_X(\cdot)$ is a proper density we should verify that it integrates to one. We do the calculation next. Let

$$A := \int \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx.$$

Then

$$\begin{aligned} A^2 &= \int \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx \int \frac{1}{\sqrt{2\pi}} \exp\{-y^2/2\} dy \\ &= \int\int \frac{1}{2\pi} \exp\{-(x^2+y^2)/2\} dx dy. \end{aligned}$$

We do a change of variables from Cartesian to polar coordinates by letting $x = r\cos(\theta)$ and $y = r\sin(\theta)$. Then, $dxdy = rdrd\theta$ and $x^2 + y^2 = r^2$. We then rewrite the integral above as follows:

$$\begin{aligned} A^2 &= \int_0^\infty \int_0^{2\pi} \frac{1}{2\pi} r e^{-r^2/2} dr d\theta = \int_0^\infty r e^{-r^2/2} dr \\ &= -\int_0^\infty de^{-r^2/2} = [e^{-r^2/2}]_0^\infty = 1, \end{aligned}$$

which proves that $A = 1$, as we wanted to verify.

We claim that $X =_D N(0,1)$ if and only if

$$E(\exp\{iuX\}) = \exp\{-u^2/2\}.$$

To see this, assume that $X =_D N(0,1)$. Then

$$\phi(u) := E(e^{iuX}) = \int e^{iux} \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx,$$

so that

$$\begin{aligned} \phi'(u) := \frac{d}{du}\phi(u) &= \int ixe^{iux} \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx \\ &= -i \int \frac{1}{\sqrt{2\pi}} e^{iux} de^{-x^2/2} = i \int \frac{1}{\sqrt{2\pi}} e^{-x^2/2} de^{iux} \\ &= -u \int e^{iux} \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx = -u\phi(u). \end{aligned}$$

Thus, $\phi'(u) = -u\phi(u)$. We can rewrite this identity as follows:

$$\frac{d\phi(u)}{\phi(u)} = -udu,$$

or

$$d(\ln(\phi(u))) = d(-u^2/2).$$

Integrating both sides from 0 to u, we find

$$\ln(\phi(u)) - \ln(\phi(0)) = -u^2/2.$$

But $\phi(0) = E(e^{i0X}) = 1$. Hence, $\ln(\phi(u)) = -u^2/2$, which implies $E(e^{iuX}) = e^{-u^2/2}$, as claimed.

We have shown that if $X =_D N(0,1)$, then $E(e^{iuX}) = e^{-u^2/2}$. To see the converse, one observes that

$$E(e^{iuX}) = \int_{-\infty}^{\infty} e^{iux} f_X(x)dx,$$

so that $E(e^{iuX})$ is the Fourier transform of $f_X(\cdot)$. It can be shown that the Fourier transform specifies $f_X(\cdot)$ uniquely. That is, if two random variables X and Y are such that $E(e^{iuX}) = E(e^{iuY})$, then $f_X(x) = f_Y(x)$ for $x \in \Re$. Accordingly, if $E(e^{iuX}) = e^{-u^2/2}$, it must be that $X =_D N(0,1)$.

Since

$$\begin{aligned} E(\exp\{iuX\}) &= E(1 + iuX - \frac{1}{2}u^2X^2 - \frac{1}{3!}iu^3X^3 + \cdots + \frac{1}{n!}(iuX)^n + \cdots) \\ &= \exp\{-u^2/2\} = 1 - \frac{1}{2}u^2 + \frac{1}{8}u^4 - \frac{1}{48}u^6 + \cdots + \frac{1}{m!}(-u^2/2)^m + \cdots, \end{aligned}$$

we see that $E(X^m) = 0$ for m odd and

$$(iu)^{2m}\frac{1}{(2m)!}E(X^{2m}) = \frac{1}{m!}(-u^2/2)^m,$$

so that $E(X^{2m}) = \frac{(2m)!}{2^m\ m!}$.

The cdf of X does not admit a closed form expression. Its values have been tabulated and many software packages provides that function. Table 7.1 shows sample values of the standard Gaussian distribution.

x	1	1.64	1.96	2	2.58	7.13
$P(N(0,1) > x)$	15.9%	5%	2.5%	2.27%	0.5%	5×10^{-13}
$P(\|N(0,1)\| > x)$	31.7%	10%	5%	4.55%	1%	10^{-12}

Table 7.1: Sample values of probabilities for $N(0,1)$.

7.1.2 $N(\mu, \sigma^2)$

Definition 7.1.2. $X = N(\mu, \sigma^2)$ if $X = \mu + \sigma Y$ where $Y = N(0,1)$, so that $E(\exp\{iuX\}) = \exp\{iu\mu - u^2\sigma^2/2\}$.

Using (4.6.1), we see that the pdf of a $N(\mu, \sigma^2)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \text{ for } x \in \Re.$$

7.2 Jointly Gaussian

So far we have considered a single Gaussian random variable. In this section we discuss collections of random variables that have a Gaussian joint distribution. Such random variables occur frequently in applications. A simple example is a collection of independent Gaussian random variables. Another example is a collection of linear combinations of a set of independent Gaussian random variables.

7.2.1 $N(\mathbf{0}, \boldsymbol{I})$

Definition 7.2.1. $\boldsymbol{X} = N(\mathbf{0}, \boldsymbol{I})$ if the components of X are i.i.d. $N(0,1)$.

7.2.2 Jointly Gaussian

Random variables $\boldsymbol{X}$ are said to be jointly Gaussian if $\boldsymbol{u}^T\boldsymbol{X}$ is Gaussian for any vector $\boldsymbol{u}$.

Assume $Y := \boldsymbol{u}^T\boldsymbol{X}$ is Gaussian and let $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $\Sigma := \Sigma_X$. Since $E(Y) = \boldsymbol{u}^T\boldsymbol{\mu}$ and $\text{var}(Y) = \boldsymbol{u}^T\Sigma\boldsymbol{u}$ (see (7.5.7)), we see that

$$E(e^{iY}) = E(e^{i\boldsymbol{u}^T\boldsymbol{X}}) = e^{iE(Y)-\text{var}(Y)/2} = e^{i\boldsymbol{u}^T\boldsymbol{\mu}-\boldsymbol{u}^T\Sigma\boldsymbol{u}/2}. \tag{7.2.1}$$

Now,

$$E(e^{i\boldsymbol{u}^T\boldsymbol{X}}) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{i\boldsymbol{u}^T\boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x})dx_1\ldots dx_n \tag{7.2.2}$$

is the n-dimensional Fourier transform of $f_{\boldsymbol{X}}(\boldsymbol{x})$. As in the one-dimensional case, it turns out that this Fourier transform completely determines the joint density.

Using these preliminary calculations we can derive the following very useful result.

Theorem 7.2.1. *Independence and Correlation*

Jointly Gaussian random variables are independent if and only if they are uncorrelated.

Proof:

We know from Theorem 5.3.1 that the random variables $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ are independent if and only if their joint density is the product of their individual densities. Now, identity (7.2.2) shows that this happens if and only if

$$\begin{aligned} E(e^{i\boldsymbol{u}^T\boldsymbol{X}}) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{i(u_1x_1+\cdots+u_nx_n)} f_{X_1}(x_1)\ldots f_{X_n}(x_n)dx_1\ldots dx_n \\ &= \int_{-\infty}^{\infty} e^{iu_1x_1} f_{X_1}x_1dx_1\cdots\int_{-\infty}^{\infty} e^{iu_nx_n} f_{X_n}x_ndx_n \\ &= E(e^{iu_1X_1})\times\cdots\times E(e^{iu_nX_n}). \end{aligned}$$

That is, random variables are mutually independent if and only if their joint characteristic function $E(e^{i\boldsymbol{u}^T\boldsymbol{X}})$ is the product of their individual characteristic functions $E(e^{iu_mX_m})$. To complete the proof, we use the specific form (7.2.1) of the joint characteristic function of jointly Gaussian random variables and we note that it factorizes if and only if Σ is diagonal, i.e., if and only if the random variables are uncorrelated.

□

Note also that if $\boldsymbol{X} = N(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{AX} = N(\boldsymbol{\mu}, \boldsymbol{AA}^T)$. Assume that $\boldsymbol{A}$ is nonsingular. Then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|A|} f_{\boldsymbol{X}}(\boldsymbol{A}^{-1}\boldsymbol{y}),$$

in view of the change of variables. Hence,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|)^{n/2}} \exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$$

where $\boldsymbol{\Sigma} = \boldsymbol{AA}^T$. We used the fact that the determinant of the product of square matrices is the product of their determinants, so that $|\boldsymbol{\Sigma}| = |\boldsymbol{A}| \times |\boldsymbol{A}^T| = |\boldsymbol{A}|^2$ and $|\boldsymbol{A}| = |\boldsymbol{\Sigma}|^{1/2}$.

7.3 Conditional Expectation J.G.

We explain how to calculate the conditional expectation of jointly Gaussian random variables. The result is remarkably simple: the conditional expectation is linear in the observations!

Theorem 7.3.1. *Let (X, Y) be two jointly Gaussian random variables. Then*

a. One has

$$E[X \mid Y] = E(X) + \frac{cov(X,Y)}{var(Y)}(Y - E(Y)). \tag{7.3.1}$$

Let $(\boldsymbol{X}, \boldsymbol{Y})$ be jointly Gaussian.

b. If $\Sigma_{\boldsymbol{Y}}$ is invertible, then

$$E[\boldsymbol{X}|\boldsymbol{Y}] = E(\boldsymbol{X}) + \Sigma_{\boldsymbol{X},\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y} - E(\boldsymbol{Y})). \tag{7.3.2}$$

c. If $\Sigma_{\boldsymbol{Y}}$ is not invertible, then

$$E[\boldsymbol{X}|\boldsymbol{Y}] = E(\boldsymbol{X}) + \Sigma_{\boldsymbol{X},\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}}^{\dagger}(\boldsymbol{Y} - E(\boldsymbol{Y})) \tag{7.3.3}$$

where $\Sigma_{\boldsymbol{Y}}^{\dagger}$ is such that

$$\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{\dagger}\boldsymbol{\Sigma}_{\boldsymbol{Y}}.$$

Proof:

Then we can look for a vector $\boldsymbol{a}$ and a matrix $\boldsymbol{B}$ of compatible dimensions so that $\boldsymbol{X} - \boldsymbol{a} - \boldsymbol{BY}$ is zero-mean and uncorrelated with $\boldsymbol{Y}$. In that case, these random variables are independent and we can write

$$\begin{aligned} E[\boldsymbol{X}|\boldsymbol{Y}] &= E[\boldsymbol{X} - \boldsymbol{a} - \boldsymbol{BY} + \boldsymbol{a} + \boldsymbol{BY}|\boldsymbol{Y}] \\ &= E[\boldsymbol{X} - \boldsymbol{a} - \boldsymbol{BY}|\boldsymbol{Y}] + E[\boldsymbol{a} + \boldsymbol{BY}|\boldsymbol{Y}], \text{ by (6.4.2)} \\ &= E(\boldsymbol{X} - \boldsymbol{a} - \boldsymbol{BY}) + \boldsymbol{a} + \boldsymbol{BY}, \text{ by (6.4.5)} \\ &= \boldsymbol{a} + \boldsymbol{BY}. \end{aligned}$$

To find the desired a and $\boldsymbol{B}$, we solve $E[\boldsymbol{X} - \boldsymbol{a} - \boldsymbol{BY}] = 0$, or

$$\boldsymbol{a} = E(\boldsymbol{X}) - \boldsymbol{B}E(\boldsymbol{Y})$$

and $E[(\boldsymbol{X} - \boldsymbol{a} - \boldsymbol{BY})\boldsymbol{Y}^T] = 0$, or

$$\Sigma_{\boldsymbol{X},\boldsymbol{Y}} = \boldsymbol{B}\Sigma_{\boldsymbol{Y}}. \tag{7.3.4}$$

If $\Sigma_{\boldsymbol{Y}}$ is invertible, this gives

$$\boldsymbol{B} = \Sigma_{\boldsymbol{X},\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}}^{-1},$$

so that

$$E[\boldsymbol{X}|\boldsymbol{Y}] = E(\boldsymbol{X}) + \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y} - E(\boldsymbol{Y})).$$

If $\boldsymbol{\Sigma}_{\boldsymbol{Y},\boldsymbol{Y}}$ is not invertible is not, we choose a pseudo-inverse $\Sigma_{\boldsymbol{Y}}^{\dagger}$ that solves (7.3.4), i.e., is such that

$$\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{\dagger}\boldsymbol{\Sigma}_{\boldsymbol{Y}}.$$

We then find

$$E[\boldsymbol{X}|\boldsymbol{Y}] = E(\boldsymbol{X}) + \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{\dagger}(\boldsymbol{Y} - E(\boldsymbol{Y})).$$

Examples will help you understand how to use these results.

7.4 Summary

We defined the Gaussian random variables $N(0,1), N(\mu,\sigma^2)$, and $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ both in terms of their density and their characteristic function.

Jointly Gaussian random variables that are uncorrelated are independent.

If X, Y are jointly Gaussian, then $E[X \mid Y] = E(X) + \text{cov}(X,Y)\text{var}(Y)^{-1}(Y - E(Y))$. In the vector case,

$$E[\boldsymbol{X} \mid \boldsymbol{Y}] = E(\boldsymbol{X}) + \Sigma_{\boldsymbol{X},\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y} - E(\boldsymbol{Y}),$$

when $\Sigma_{\boldsymbol{Y}}$ is invertible. We also discussed the non-invertible case.

7.5 Solved Problems

Example 7.5.1. *The noise voltage X in an electric circuit can be modelled as a Gaussian random variable with mean zero and variance equal to 10^{-8}.*

a. What is the probability that it exceeds 10^{-4}? What is the probability that it exceeds 2×10^{-4}? What is the probability that its value is between -2×10^{-4} and 10^{-4}?

b. Given that the noise value is positive, what is the probability that it exceeds 10^{-4}?

c. What is the expected value of $|X|$?

Let $Z = 10^4 X$, then $Z =_D N(0,1)$ and we can reformulate the questions in terms of Z.

a. Using (7.1) we find $P(Z > 1) = 0.159$ and $P(Z > 2) = 0.023$. Indeed, $P(Z > d) = P(|Z| > d)/2$, by symmetry of the density. Moreover,

$$P(-2 < Z < 1) = P(Z < 1) - P(Z \leq -2) = 1 - P(Z > 1) - P(Z > 2) = 1 - 0.159 - 0.023 = 0.818.$$

b. We have

$$P[Z > 1 \mid Z > 0] = \frac{P(Z > 1)}{P(Z > 0)} = 2P(Z > 1) = 0.318.$$

c. Since $Z = 10^4 X$, one has $E(|Z|) = 10^4 E(|X|)$. Now,

$$\begin{aligned} E(|Z|) &= \int_{-\infty}^{\infty} |z| f_Z(z) dz = 2\int_0^{\infty} z f_Z(z) dz = 2\int_0^{\infty} \frac{1}{\sqrt{2\pi}} z \exp\{-\frac{1}{2}z^2\} dz \\ &= -\sqrt{\frac{2}{\pi}} \int_0^{\infty} d[\exp\{-\frac{1}{2}z^2\}] = \sqrt{\frac{2}{\pi}}. \end{aligned}$$

Hence,

$$E(|X|) = 10^{-4}\sqrt{\frac{2}{\pi}}.$$

Example 7.5.2. *Let $\boldsymbol{U} = \{U_n, n \geq 1\}$ be a sequence of independent standard Gaussian random variables. A low-pass filter takes the sequence $\boldsymbol{U}$ and produces the output sequence $X_n = U_n + U_{n+1}$. A high-pass filter produces the output sequence $Y_n = U_n - U_{n+1}$.*

a. Find the joint pdf of X_n and X_{n-1} and find the joint pdf of X_n and X_{n+m} for $m > 1$.

b. Find the joint pdf of Y_n and Y_{n-1} and find the joint pdf of Y_n and Y_{n+m} for $m > 1$.

c. Find the joint pdf of X_n and Y_m.

We start with some preliminary observations. First, since the U_i are independent, they are jointly Gaussian. Second, X_n and Y_n are linear combinations of the U_i and thus are also jointly Gaussian. Third, the jpdf of jointly gaussian random variables $\boldsymbol{Z}$ is

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^n det(C)}} exp[-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{m})C^{-1}(\boldsymbol{z} - \boldsymbol{m})]$$

where n is the dimension of $\boldsymbol{Z}$, $\boldsymbol{m}$ is the vector of expectations of $\boldsymbol{Z}$, and C is the covariance matrix $E[(\boldsymbol{Z} - \boldsymbol{m})(\boldsymbol{Z} - \boldsymbol{m})^T]$. Finally, we need some basic facts from algebra. If $C = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $det(C) = ad - bc$ and $C^{-1} = \frac{1}{det(C)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$. We are now ready to answer the questions.

a. Express in the form $\boldsymbol{X} = A\boldsymbol{U}$.

$$\begin{bmatrix} X_n \\ X_{n-1} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} U_{n-1} \\ U_n \\ U_{n+1} \end{bmatrix}$$

Then $E[\boldsymbol{X}] = AE[\boldsymbol{U}] = \boldsymbol{0}$.

$$C = E[\boldsymbol{X}\boldsymbol{X}^T] = AE[\boldsymbol{U}\boldsymbol{U}^T]A^T = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

Then $det(C) = \frac{1}{4} - \frac{1}{16} = \frac{3}{16}$ and

$$C^{-1} = \frac{16}{3} \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

$f_{X_n X_{n-1}}(x_n, x_{n-1}) = \frac{2}{\pi\sqrt{3}} \exp[-\frac{4}{3}(x_n^2 - x_n x_{n-1} + x_{n-1}^2)]$

For $m > 1$,

$$\begin{bmatrix} X_n \\ X_{n+m} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} U_n \\ U_{n+1} \\ U_{n+m} \\ U_{n+m+1} \end{bmatrix}$$

Then $E[\boldsymbol{X}] = AE[\boldsymbol{U}] = \boldsymbol{0}$.

$$C = E[\boldsymbol{X}\boldsymbol{X}^T] = AE[\boldsymbol{U}\boldsymbol{U}^T]A^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

Then $det(C) = \frac{1}{4}$ and

$$C^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$f_{X_n X_{n+m}}(x_n, x_{n+m}) = \frac{1}{\pi} \exp[-\frac{1}{4}(x_n^2 + x_{n+m}^2)]$

b.

$$\begin{bmatrix} Y_n \\ Y_{n-1} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} U_{n-2} \\ U_{n-1} \\ U_n \end{bmatrix}$$

Then $E[\boldsymbol{Y}] = AE[\boldsymbol{U}] = \boldsymbol{0}$.

$$C = E[\boldsymbol{Y}\boldsymbol{Y}^T] = AE[\boldsymbol{U}\boldsymbol{U}^T]A^T = \begin{bmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

Then $det(C) = \frac{1}{4} - \frac{1}{16} = \frac{3}{16}$ and

$$C^{-1} = \frac{16}{3}\begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

$f_{Y_nY_{n-1}}(y_n, y_{n-1}) = \frac{2}{\pi\sqrt{3}} \exp[-\frac{4}{3}(y_n^2 + y_n y_{n-1} + y_{n-1}^2)]$

For $m > 1$,

$$\begin{bmatrix} Y_n \\ Y_{n+m} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} U_{n-1} \\ U_n \\ U_{n+m-1} \\ U_{n+m} \end{bmatrix}$$

Then $E[\boldsymbol{Y}] = AE[\boldsymbol{U}] = \boldsymbol{0}$.

$$C = E[\boldsymbol{Y}\boldsymbol{Y}^T] = AE[\boldsymbol{U}\boldsymbol{U}^T]A^T = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} & 0 \\ \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

Then $det(C) = \frac{1}{4}$ and

$$C^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$f_{Y_nY_{n+m}}(y_n, y_{n+m}) = \frac{1}{\pi} \exp[-\frac{1}{4}(y_n^2 + y_{n+m}^2)]$

c. We have 3 cases when i. m=n, ii. m=n+1, and iii. otherwise.

i. First consider when m=n.

$$\begin{bmatrix} X_n \\ Y_n \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} U_{n-1} \\ U_n \\ U_{n+1} \end{bmatrix}$$

Then $E[[X_n\ Y_n]^T] = AE[\boldsymbol{U}] = \boldsymbol{0}$.

$$C = AE[\boldsymbol{U}\boldsymbol{U}^T]A^T = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

Then $det(C) = \frac{1}{4} - \frac{1}{14} = \frac{3}{16}$ and

$$C^{-1} = \frac{16}{3}\begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

$f_{X_nY_n}(x_n, y_n) = \frac{2}{\pi\sqrt{3}} \exp[-\frac{4}{3}(x_n^2 - x_n y_n + y_n^2)]$

ii. Consider m=n+1.

$$\begin{bmatrix} X_n \\ Y_{n+1} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} U_n \\ U_{n+1} \end{bmatrix}$$

Then $E[[X_n\ Y_{n+1}]^T] = AE[\boldsymbol{U}] = \mathbf{0}$.

$$C = AE[\boldsymbol{U}\boldsymbol{U}^T]A^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

Then $det(C) = \frac{1}{4}$ and

$$C^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$f_{X_nY_{n+1}}(x_n, y_{n+1}) = \frac{1}{\pi} \exp[-\frac{1}{4}(x_n^2 + y_{n+1}^2)]$

iii. For all other m.

$$\begin{bmatrix} X_n \\ Y_m \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} U_n \\ U_{n+1} \\ U_{m-1} \\ U_m \end{bmatrix}$$

Then $E[[X_n\ Y_m]^T] = AE[\boldsymbol{U}] = \mathbf{0}$.

$$C = AE[\boldsymbol{U}\boldsymbol{U}^T]A^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

Then $det(C) = \frac{1}{4}$ and

$$C^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$f_{X_nY_m}(x_n, y_m) = \frac{1}{\pi} \exp[-\frac{1}{4}(x_n^2 + y_m^2)]$

Example 7.5.3. *Let X, Y, Z, V be i.i.d. $N(0,1)$. Calculate $E[X+2Y|3X+Z, 4Y+2V]$.*

We have

$$E[X+2Y|3X+Z, 4Y+2V] = \boldsymbol{a}\Sigma^{-1} \begin{bmatrix} 3X+Z \\ 4Y+2V \end{bmatrix}$$

where

$$\boldsymbol{a} = [E((X+2Y)(3X+Z)), E((X+2Y)(4Y+2V))] = [3, 8]$$

and

$$\Sigma = \begin{bmatrix} \text{var}(3X+Z) & E((3X+Z)(4Y+2V)) \\ E((3X+Z)(4Y+2V)) & \text{var}(4Y+2V) \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 20 \end{bmatrix}.$$

Hence,

$$E[X+2Y|3X+Z, 4Y+2V] = [3, 8] \begin{bmatrix} 10^{-1} & 0 \\ 0 & 20^{-1} \end{bmatrix} \begin{bmatrix} 3X+Z \\ 4Y+2V \end{bmatrix} = \frac{3}{10}(3X+Z)+\frac{4}{10}(4Y+2V).$$

Example 7.5.4. *Assume that $\{X, Y_n, n \geq 1\}$ are mutually independent random variables with $X = N(0,1)$ and $Y_n = N(0, \sigma^2)$. Let $\hat{X}_n = E[X \mid X+Y_1, \ldots, X+Y_n]$. Find the smallest value of n such that*

$$P(|X - \hat{X}_n| > 0.1) \leq 5\%.$$

We know that $\hat{X}_n = a_n(nX + Y_1 + \cdots + Y_n)$. The value of a_n is such that

$$E((X - \hat{X}_n)(X+Y_j)) = 0, \text{ i.e., } E((X - a_n(nX+Y_j))(X+Y_j)) = 0,$$

which implies that

$$a_n = \frac{1}{n+\sigma^2}.$$

Then

$$\begin{aligned} \text{var}(X - \hat{X}_n) &= \text{var}((1-na_n)X - a_n(Y_1 + \cdots + Y_n)) = (1-na_n)^2 + n(a_n)^2\sigma^2 \\ &= \frac{\sigma^2}{n+\sigma^2}. \end{aligned}$$

Thus we know that $X - \hat{X}_n = N(0, \frac{\sigma^2}{n+\sigma^2})$. Accordingly,

$$P(|X - \hat{X}_n| > 0.1) = P(|N(0, \frac{\sigma^2}{n+\sigma^2})| > 0.1) = P(|N(0,1)| > \frac{0.1}{\alpha_n})$$

where $\alpha_n = \sqrt{\frac{\sigma^2}{n+\sigma^2}}$. For this probability to be at most 5% we need

$$\frac{0.1}{\alpha_n} = 2, \text{ i.e., } \alpha_n = \sqrt{\frac{\sigma^2}{n+\sigma^2}} = \frac{0.1}{2},$$

so that

$$n = 19\sigma^2.$$

The result is intuitively pleasing: If the observations are more noisy (σ^2 large), we need more of them to estimate X.

Example 7.5.5. *Assume that X, Y are i.i.d. $N(0,1)$. Calculate $E[(X+Y)^4 \mid X-Y]$.*

Note that $X+Y$ and $X-Y$ are independent because they are jointly Gaussian and uncorrelated. Hence,

$$E[(X+Y)^4 \mid X-Y] = E((X+Y)^4) = E(X^4+4X^3Y+6X^2Y^2+4XY^3+Y^4) = 3+6+3 = 12.$$

Example 7.5.6. *Let X, Y be independent $N(0,1)$ random variables. Show that $W := X^2 + Y^2 =_D Exd(1/2)$. That is, the sum of the squares of two i.i.d. zero-mean Gaussian random variables is exponentially distributed!*

We calculate the characteristic function of W. We find

$$\begin{aligned} E(e^{iuW}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{iu(x^2+y^2)} \frac{1}{2\pi} e^{-(x^2+y^2)/2} dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{iur^2} \frac{1}{2\pi} e^{-r^2/2} r dr d\theta \\ &= \int_0^{\infty} e^{iur^2} e^{-r^2/2} r dr \\ &= \int_0^{\infty} \frac{1}{2iu-1} d[e^{iur^2 - r^2/2}] = \frac{1}{1-2iu}. \end{aligned}$$

On the other hand, if $W =_D Exd(\lambda)$, then

$$\begin{aligned} E(e^{iuW}) &= \int_0^\infty e^{iux}\lambda e^{-\lambda x}dx \\ &= \frac{\lambda}{\lambda - iu} = \frac{1}{1-\lambda^{-1}iu}. \end{aligned}$$

.

Comparing these expressions shows that $X^2 + Y^2 =_D Exd(1/2)$ as claimed.

Example 7.5.7. *Let $\{X_n, n \geq 0\}$ be Gaussian $N(0,1)$ random variables. Assume that $Y_{n+1} = aY_n + X_n$ for $n \geq 0$ where Y_0 is a Gaussian random variable with mean zero and variance σ^2 independent of the X_n's and $|a| < 1$.*

a. Calculate var(Y_n) for $n \geq 0$. Show that var$(Y_n) \to \gamma^2$ as $n \to \infty$ for some value γ^2.

b. Find the values of σ^2 so that the variance of Y_n does not depend on $n \geq 1$.

a. We see that

$$\text{var}(Y_{n+1}) = \text{var}(aY_n + X_n) = a^2\text{var}(Y_n) + \text{var}(X_n) = a^2\text{var}(Y_n) + 1.$$

Thus, we $\alpha_n := \text{var}(Y_n)$, one has

$$\alpha_{n+1} = a^2\alpha_n + 1 \text{ and } \alpha_0 = \sigma^2.$$

Solving these equations we find

$$\text{var}(Y_n) = \alpha_n = a^{2n}\sigma^2 + \frac{1-a^{2n}}{1-a^2}, \text{ for } n \geq 0.$$

Since $|a| < 1$, it follows that

$$\text{var}(Y_n) \to \gamma^2 := \frac{1}{1-a^2} \text{ as } n \to \infty.$$

b. The obvious answer is $\sigma^2 = \gamma^2$.

Example 7.5.8. *Let the X_n's be as in Example 7.5.7.*

a. Calculate

$$E[X_1 + X_2 + X_3 \mid X_1 + X_2, X_2 + X_3, X_3 + X_4].$$

b. Calculate

$$E[X_1 + X_2 + X_3 \mid X_1 + X_2 + X_3 + X_4 + X_5].$$

a. We know that the solution is of the form $Y = a(X_1 + X_2) + b(X_2 + X_3) + c(X_3 + X_4)$ where the coefficients a, b, c must be such that the estimation error is orthogonal to the conditioning variables. That is,

$$E((X_1 + X_2 + X_3) - Y)(X_1 + X_2)) = E((X_1 + X_2 + X_3) - Y)(X_2 + X_3))$$
$$= E((X_1 + X_2 + X_3) - Y)(X_3 + X_4)) = 0.$$

These equalities read

$$2 - a - (a + b) = 2 - (a + b) - (b + c) = 1 - (b + c) - c = 0,$$

and solving these equalities gives $a = 3/4, b = 1/2$, and $c = 1/4$.

b. Here we use symmetry. For $k = 1, \ldots, 5$, let

$$Y_k = E[X_k \mid X_1 + X_2 + X_3 + X_4 + X_5].$$

Note that $Y_1 = Y_2 = \cdots = Y_5$, by symmetry. Moreover,

$$Y_1+Y_2+Y_3+Y_4+Y_5 = E[X_1+X_2+X_3+X_4+X_5 \mid X_1+X_2+X_3+X_4+X_5] = X_1+X_2+X_3+X_4+X_5.$$

It follows that $Y_k = (X_1 + X_2 + X_3 + X_4 + X_5)/5$ for $k = 1, \ldots, 5$. Hence,

$$E[X_1 + X_2 + X_3 \mid X_1 + X_2 + X_3 + X_4 + X_5] = Y_1 + Y_2 + Y_3 = \frac{3}{5}(X_1 + X_2 + X_3 + X_4 + X_5).$$

Example 7.5.9. *Let the X_n's be as in Example 7.5.7. Find the jpdf of* $(X_1 + 2X_2 + 3X_3, 2X_1 + 3X_2 + X_3, 3X_1 + X_2 + 2X_3)$.

These random variables are jointly Gaussian, zero mean, and with covariance matrix Σ given by

$$\Sigma = \begin{bmatrix} 14 & 11 & 11 \\ 11 & 14 & 11 \\ 11 & 11 & 14 \end{bmatrix}.$$

Indeed, Σ is the matrix of covariances. For instance, its entry (2, 3) is given by

$$E((2X_1 + 3X_2 + X_3)(3X_1 + X_2 + 2X_3)) = 2 \times 3 + 3 \times 1 + 1 \times 2 = 11.$$

We conclude that the jpdf is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\}.$$

We let you calculate $|\Sigma|$ and Σ^{-1}.

Example 7.5.10. *Let X_1, X_2, X_3 be independent $N(0,1)$ random variables. Calculate $E[X_1 + 3X_2|\boldsymbol{Y}]$ where*

$$\boldsymbol{Y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

By now, this should be familiar. The solution is $Y := a(X_1 + 2X_2 + 3X_3) + b(3X_1 + 2X_2 + X_3)$ where a and b are such that

$$0 = E((X_1+3X_2-Y)(X_1+2X_2+3X_3)) = 7-(a+3b)-(4a+4b)-(9a+3b) = 7-14a-10b$$

and

$$0 = E((X_1+3X_2-Y)(3X_1+2X_2+X_3)) = 9-(3a+9b)-(4a+4b)-(3a+b) = 9-10a-14b.$$

Solving these equations gives $a = 1/12$ and $b = 7/12$.

Example 7.5.11. *Find the jpdf of $(2X_1+X_2, X_1+3X_2)$ where X_1 and X_2 are independent $N(0,1)$ random variables.*

These random variables are jointly Gaussian, zero-mean, with covariance Σ given by

$$\Sigma = \begin{bmatrix} 5 & 5 \\ 5 & 10 \end{bmatrix}.$$

Hence,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\{-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\} \\ &= \frac{1}{10\pi} \exp\{-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\} \end{aligned}$$

where

$$\Sigma^{-1} = \frac{1}{25}\begin{bmatrix} 10 & -5 \\ -5 & 5 \end{bmatrix}.$$

Example 7.5.12. *The random variable X is $N(\mu, 1)$. Find an approximate value of μ so that*

$$P(-0.5 \leq X \leq -0.1) \approx P(1 \leq X \leq 2).$$

We write $X = \mu + Y$ where Y is $N(0, 1)$. We must find μ so that

$$g(\mu) := P(-0.5 - \mu \leq Y \leq -0.1 - \mu) - P(1 - \mu \leq Y \leq 2 - \mu) \approx 0.$$

We do a little search using a table of the $N(0, 1)$ distribution or using a calculator. I find that $\mu \approx 0.065$.

Example 7.5.13. *Let X be a $N(0, 1)$ random variable. Calculate the mean and the variance of $\cos(X)$ and $\sin(X)$.*

a. Mean Values. We know that

$$E(e^{iuX}) = e^{-u^2/2} \text{ and } e^{i\theta} = \cos(\theta) + i\sin(\theta).$$

Therefore,

$$E(\cos(uX) + i\sin(uX)) = e^{-u^2/2},$$

so that

$$E(\cos(uX)) = e^{-u^2/2} \text{ and } E(\sin(uX)) = 0.$$

In particular, $E(\cos(X)) = e^{-1/2}$ and $E(\sin(X)) = 0$.

b. Variances. We first calculate $E(\cos^2(X))$. We find

$$E(\cos^2(X)) = E(\frac{1}{2}(1 + \cos(2X))) = \frac{1}{2} + \frac{1}{2}E(\cos(2X)).$$

Using the previous derivation, we find that

$$E(\cos(2X)) = e^{-2^2/2} = e^{-2},$$

so that $E(\cos^2(X)) = (1/2) + (1/2)e^{-2}$. We conclude that

$$\text{var}(\cos(X)) = E(\cos^2(X)) - (E(\cos(uX)))^2 = \frac{1}{2} + \frac{1}{2}e^{-2} - (e^{-1/2})^2 = \frac{1}{2} + \frac{1}{2}e^{-2} - e^{-1}.$$

Similarly, we find

$$E(\sin^2(X)) = E(1 - \cos^2(X)) = \frac{1}{2} - \frac{1}{2}e^{-2} = \text{var}(\sin(X)).$$

Example 7.5.14. *Let X be a $N(0,1)$ random variable. Define*

$$Y = \begin{cases} X, \textit{ if } |X| \leq 1 \\ -X, \textit{ if } |X| > 1. \end{cases}$$

Find the pdf of Y.

By symmetry, X is $N(0,1)$.

Example 7.5.15. *Let $\{X, Y, Z\}$ be independent $N(0,1)$ random variables.*

a. Calculate

$$E[3X + 5Y \mid 2X - Y, X + Z].$$

b. How does the expression change if X, Y, Z are i.i.d. $N(1,1)$?

a. Let $V_1 = 2X - Y, V_2 = X + Z$ and $\boldsymbol{V} = [V_1, V_2]^T$. Then

$$E[3X + 5Y \mid \boldsymbol{V}] = \boldsymbol{a}\Sigma_V^{-1}\boldsymbol{V}$$

where

$$\boldsymbol{a} = E((3X + 5Y)\boldsymbol{V}^T) = [1, 3]$$

and

$$\Sigma_V = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}.$$

Hence,

$$E[3X + 5Y \mid \boldsymbol{V}] = [1, 3] \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}^{-1} \boldsymbol{V} = [1, 3]\frac{1}{6} \begin{bmatrix} 2 & -2 \\ -2 & 5 \end{bmatrix} \boldsymbol{V}$$
$$= \frac{1}{6}[-4, 13]\boldsymbol{V} = -\frac{2}{3}(2X - Y) + \frac{13}{6}(X + Z).$$

b. Now,

$$E[3X + 5Y \mid \boldsymbol{V}] = E(3X + 5Y) + \boldsymbol{a}\Sigma_V^{-1}(\boldsymbol{V} - E(\boldsymbol{V})) = 8 + \frac{1}{6}[-4, 13](\boldsymbol{V} - [1, 2]^T)$$
$$= \frac{26}{6} - \frac{2}{3}(2X - Y) + \frac{13}{6}(X + Z).$$

Example 7.5.16. *Let (X, Y) be jointly Gaussian. Show that $X - E[X \mid Y]$ is Gaussian and calculate its mean and variance.*

We know that

$$E[X \mid Y] = E(X) + \frac{cov(X, Y)}{\text{var}(Y)}(Y - E(Y)).$$

Consequently,

$$X - E[X \mid Y] = X - E(X) - \frac{cov(X, Y)}{\text{var}(Y)}(Y - E(Y))$$

and is certainly Gaussian. This difference is zero-mean. Its variance is

$$\text{var}(X) + [\frac{cov(X, Y)}{\text{var}(Y)}]^2\text{var}(Y) - 2\frac{cov(X, Y)}{\text{var}(Y)}cov(X, Y) = \text{var}(X) - \frac{[cov(X, Y)]^2}{\text{var}(Y)}.$$

Chapter 8

Detection and Hypothesis Testing

The detection problem is roughly as follows. We want to guess which of finitely many possible causes produced an observed effect. For instance, you have a fever (observed effect); do you think you have the flu or a cold or the malaria? As another example, you observe some strange shape on an X-ray; is it a cancer or some infection of the tissues? A receiver gets a particular waveform; did the transmitter send the bit 0 or the bit 1? (Hypothesis testing is similar.) As you can see, these problems are prevalent in applications.

There are two basic formulations: either we know the prior probabilities of the possible causes (Bayesian) or we do not (non-Bayesian). When we do not, we can look for the maximum likelihood detection or we can formulate a hypothesis-testing problem.

8.1 Bayesian

Assume that X takes values in a finite set $\{0, \ldots, M\}$. We know the conditional density or distribution of Y given $\{X = x\}$ and the *prior* distribution of X: $\{P(X = x), x = 0, \ldots, M\}$. We want to choose Z in $\{0, \ldots, N\}$ on the basis of Y to minimize $E(c(X, Z))$ where $c(\cdot, \cdot)$ is a known nonnegative function.

Since $E(c(X, Z)) = E(E[c(X, Z)|Y])$, we should choose $Z = g(Y)$ where

$$g(y) = \arg\min_z E[c(X, z)|Y = y].$$

A particular example is when $M = N$ and $c(m, n) = 1\{m \neq n\}$. In that case,

$$g(y) = MAP[X|Y = y] := \arg\max_x P[X = x|Y = y], \tag{8.1.1}$$

the *maximum a posteriori* estimate of X given $\{Y = y\}$, the most likely value of X given $\{Y = y\}$.

To calculate the MAP, one uses the fact that

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

so that

$$MAP[X|Y = y] = \arg\max_x P(X = x)f_{Y|X}[y|x]. \tag{8.1.2}$$

The common criticism of this formulation is that in many cases the prior distribution of X is not known at all. For instance, consider designing a burglar alarm for your house. What prior probability should you use? You suspect a garbage in, garbage out effect here and you are correct.

8.2 Maximum Likelihood estimation

Instead of choosing the value of X that is most likely given the observation, one can choose the value of X that makes the observation most likely. That is, one can choose $\arg\max_x P[Y = y|X = x]$. This estimator is called the *maximum likelihood* estimator of X given $\{Y = y\}$, or $MLE[X|Y = y]$.

Identity (8.1.2) shows that $MLE[X|Y = y] = MAP[X|Y = y]$ when the prior distribution of X is uniform, i.e., when $P(X = x)$ has the same value for all x. Note also that $MLE[X|Y = y]$ can be calculated without knowing the prior distribution of X. Finally, a deeper property of the MLE is that under weak assumptions it tends to be a good estimator (asymptotically efficient).

8.3 Hypothesis Testing Problem

Consider the problem of designing a fire alarm system. You want to make the alarm as sensitive as possible as long as it does not generate too many false alarms. We formulate and solve that problem in this section.

8.3.1 Simple Hypothesis

We consider the case of a simple hypothesis. We define the problem and state the solution in the form of a theorem. We then examine some examples. We conclude the section by proving the theorem.

There are two possible hypotheses H0: $X = 0$ or H1: $X = 1$. Should one reject H0 on the basis of the observation Y?

One is given the distribution of the observation Y given X. The problem is to choose $Z = g(Y) \in \{0, 1\}$ to minimize the probability of *missed detection* $P[Z = 0|X = 1]$ subject to a bound on the probability of *false alarm*: $P[Z = 1|X = 0] \leq \beta$, for a given $\beta \in (0, 1)$. (Think of $\{X = 1\}$ = "fire" and $\{X = 0\}$ = "no fire".) For convenience, we designate the solution of that problem by $Z = HT[X|Y]$, which means that Z is the solution of the hypothesis testing problem we just described.

Discrete Case

Given X, Y has a known p.m.f. $P[Y = y|X]$. Let $L(y) = P[Y = y|X = 1]/P[Y = y|X = 0]$ (the *likelihood ratio*).

Theorem 8.3.1. *Neyman-Pearson; discrete case*

The solution is randomized:

$$Z = \begin{cases} 1, \text{ if } L(y) > \lambda, \\ 0, \text{ if } L(y) < \lambda \\ 1 \text{ w.p. } \gamma \text{ and } 0 \text{ w.p. } 1-\gamma, \text{ if } L(y) = \lambda. \end{cases}$$

The values of λ *and* γ *are selected so that* $P[Z = 1|X = 0] = \beta$.

If $L(y)$ *is increasing in* y*, then the solution is*

$$Z = \begin{cases} 1, \text{ if } y > y_0, \\ 0, \text{ if } y < y_0 \\ 1 \text{ w.p. } \gamma \text{ and } 0 \text{ w.p. } 1-\gamma, \text{ if } y = y_0 \end{cases}$$

where y_0 *and* γ *are selected so that* $P[Z = 1|X = 0] = \beta$.

It is not too difficult to show that there is a choice of λ and γ for which $[Z = 1|X = 0] = \beta$. We leave you the details.

Continuous Case

Given X, Y has a known p.d.f. $f_{Y|X}[y|x]$. Let $L(y) = f_{Y|X}[y|1]/f_{Y|X}[y|0]$ (the *likelihood ratio*).

Theorem 8.3.2. *Neyman-Pearson; continuous case*

The solution is

$$Z = \begin{cases} 1, \text{ if } L(y) > \lambda \\ 0, \text{ if } L(y) \leq \lambda. \end{cases}$$

The value of λ *is selected so that* $P[Z = 1|X = 0] = \beta$.

If $L(y)$ *is increasing in* y*, the solution becomes*

$$Z = \begin{cases} 1, \text{ if } L(y) > y_0 \\ 0, \text{ if } L(y) \leq y_0. \end{cases}$$

The value of y_0 *is selected so that* $P[Z = 1|X = 0] = \beta$.

8.3.2 Examples

We examine a few examples to illustrate how the theorem is used.

Example 1

If $X = k$, Y is exponentially distributed with mean $\mu(k)$, for $k = 0, 1$ where $0 < \mu(0) < \mu(1)$. Here, $f_{Y|X}[y|x] = \kappa(x) \exp\{-\kappa(x)y\}$ where $\kappa(x) = \mu(x)^{-1}$ for $x = 0, 1$. Thus $Z = 1\{Y > y_0\}$ where y_0 is such that $P[Z = 1|X = 0] = \beta 4$. That is, $\exp\{-\kappa(0)y_0\} = \beta$, or $y_0 = -\ln(\beta)/\kappa(0) = -\mu(0)\ln(\beta)$.

Example 2

If $X = k$, Y is Gaussian with mean $\mu(k)$ and variance 1, for $k = 0, 1$ where $0 < \mu(0) < \mu(1)$. Here, $f_{Y|X}[y|x] = K \exp\{-(x - \mu(k))^2/2\}$ where $K = 1/\sqrt{2\pi}$. Accordingly, $L(y) = B \exp\{x(\mu(1) - \mu(0))\}$ where $B = \exp\{(\mu(0)^2 - \mu(1)^2)/2\}$. Thus $Z = 1\{Y > y_0\}$ where y_0 is such that $P[Z = 1|X = 0] = \beta$. That is, $P(N(\mu(0), 1) > y_0) = \beta$, and one finds the value of y_0 using a table of the c.d.f. of $N(0, 1)$ or using a computer.

Example 3

You flip a coin 100 times and count the number Y of heads. You must decide whether the coin is fair or biased, say with $P(H) = 0.6$. The goal is to minimize the probability of missed detection $P[fair|biased]$ subject to a false alarm probability $P[biased|fair] \leq \beta$.

Solution: You can verify that $L(y)$ in increasing in y. Thus, the best decision is $Z = 1$ if $Y > y_0$, $Z = 0$ (fair) if $Y < y_0$, and $Z = 1$ w.p. γ if $Y = y_0$. You must choose y_0 and γ so that $P[Z = 1|fair] = \beta$. Let us plot $P[Y = y|fair]$:

Figure 8.1 illustrates $P[Z = 1|fair]$. For $\beta = 0.001$, one finds (using a calculator) $y_0 = 66$; for $\beta = 0.01, y_0 = 63$.

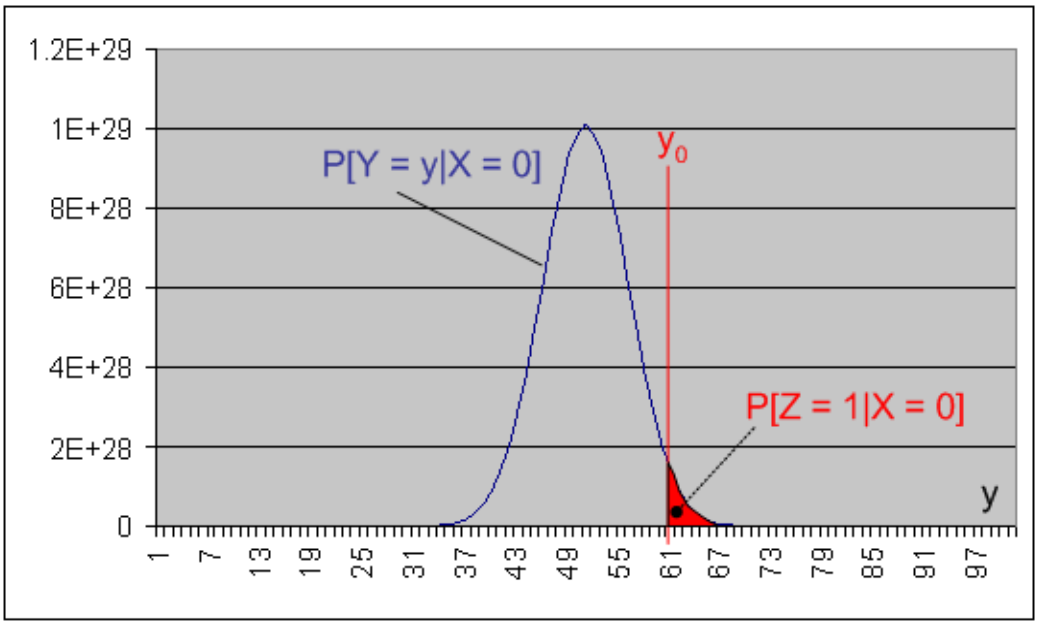


Figure 8.1: Number of heads for unbiased or biased coin

One finds also that $P[Y \geq 58|fair] = 0.066$ and $P[Y \geq 59|fair] = 0.043$; accordingly, if $\beta = 0.05$ one should decide $Z = 1$ w.p. 1 if $Y >= 59$ and $Z = 1$ w.p. 0.3 if $Y = 58$. Indeed, in that case, $P[Z = 1|fair] = P[Y \geq 59|fair] + 0.3P[Y = 58|fair] = 0.043 + 0.3(0.066 - 0.043) = 0.05$.

8.3.3 Proof of the Neyman-Pearson Theorem

Before we give a formal proof, we discuss an analogy that might help you understand the structure of the result. Imagine that you have a finite budget with which to buy food items from a given set. Your objective is to maximize the total number of calories of the items you buy. Intuitively, the best strategy is to rank the items in decreasing order of calories per dollar and to buy the items in that order until you run out of money. When you do that, it might be that you still have some money left after purchasing item $n - 1$ but not quite enough to buy item n. In that case, if you could, you would buy a fraction of item n.

If you cannot, and if we care only about the expected amount of money you spend, then you could buy the next item with some probability between 0 and 1 chosen so that you spend all you money, on average. Now imagine that the items are values of the observation Y when you decide to sound the alarm. Each item y has a cost $P[Y = y|X = 0]$ in terms of false alarm probability and some 'reward' $P[Y = y|X = 1]$ in contributing to the probability of correct detection (the caloric content of the item in our previous example). According to our intuition, we rank the items y in decreasing order of the reward/cost ratio which is precisely the likelihood ratio. Consequently, you sound the alarm when the likelihood ratio exceeds to value λ and you may have to randomize at some item to spend you total budget, on average.

Let Z be as specified by the theorem and let V be another random variable based on Y, possibly with randomization, and such that $P[V = 1|X = 0] \leq \beta$. We want to show that $P[Z = 0|X = 1] \leq P[V = 0|X = 1]$. Note that $(\lambda - L(Y))(Z - V) \leq 0$, so that

$$L(Y)(Z - V) \geq \lambda(Z - V). \tag{8.3.1}$$

For the next step, we need the fact that if W is a function of Y, then $E[WL|X = 0] = E[W|Z = 1]$. We show this fact in the continuous case. The other cases are similar. We find

$$E[WL|X = 0] = \int W(y)L(y)f_{Y|X}[y|0]dy = \int W(y)\frac{f_{Y|X}[y|1]}{f_{Y|X}[y|0]}f_{Y|X}[y|0]dy$$
$$= \int W(y)f_{Y|X}[y|1]dy = E[W|X = 1],$$

as we wanted to show.

Taking $E[\cdot|X = 0]$ of both sides of (8.3.1), we find

$$E[Z - V \mid X = 1] \geq \lambda E[Z - V \mid X = 0],$$

so that

$$P[Z = 1 \mid X = 1] - P[V = 1 \mid X = 1] \geq \lambda(P[Z = 1 \mid X = 0] - P[V = 1 \mid X = 0]) \geq 0$$

where the inequality comes from $P[Z = 1 \mid X = 0] = \beta$ and $P[V = 1 \mid X = 0] \leq \beta$. It follows, since $\lambda \geq 0$, that

$$P[Z = 1 \mid X = 1] \geq P[V = 1 \mid X = 1],$$

which is what we needed to prove. □

8.4 Composite Hypotheses

So far we have learned how to decide between two hypotheses that specify the distribution of the observations. In this section we consider composite hypotheses. Each of the two alternatives corresponds to a set of possible distributions and we want to decide which set is in effect. We explain this problem through examples.

8.4.1 Example 1

Consider once again Examples 1 and 2 in Section 8.3.2. Note that the optimal decision Z does not depend on the value of $\mu(1)$. Consequently, the optimal decision would be the same if the two hypotheses were

H0: $\mu = \mu(0)$

H1: $\mu > \mu(0)$.

The hypothesis H1 is called a composite hypothesis because it does not specify a unique value of the parameter to be tested.

8.4.2 Example 2

Once again, consider Examples 1 and 2 in Section 8.3.2 but with the hypotheses

H0: $\mu \leq \mu(0)$

H1: $\mu > \mu(0)$.

Both hypotheses H0 and H1 are composite. We claim that the optimal decision Z is still the same as in the original simple hypotheses case. To see that, observe that $P[Z = 1|\mu] = P[Y > y_0|\mu] \leq P[Y < y_0|\mu(0)] = \beta$, so that our decision meets the condition that $P[Z = 1|H0] \leq \beta$, and it minimizes $P[Z = 0|H1]$ subject to that requirement.

8.4.3 Example 3

Both examples 8.4.1 and 8.4.2 consider one-sided tests where the values of the parameter μ under H1 are all larger than those permitted under H0. What about a two-sided test with

H0: $\mu = \mu(0)$

H1: $\mu \neq \mu(0)$.

More generally, one might consider a test with

H0: $\mu \in A$

H1: $\mu \in B$

where A and B are two disjoint sets.

In general, optimal tests for such situations do not exist and one resorts to approximations. We saw earlier that the optimal decisions for a simple hypothesis test is based on the value of the likelihood ratio $L(y)$, which is the ratio of the densities of Y under the two hypotheses $X = 1$ and $X = 0$, respectively. One might then try to extend this test by replacing $L(y)$ by the ratio of the two densities under H1 and H0, respectively. How do we define the density under the hypothesis H1 "$\mu \in B$"? One idea is to calculate the MLE of μ given Y and H1, and similarly for H0. This approach works well under some situations. However, the details would carry us a bit to far. Interested students will find expositions of these methods in any good statistics book. Look for the keywords "likelihood ratio test, goodness-of-fit test".

8.5 Summary

The detection story is that X, Y are random variables, $X \in \{0, 1\}$ (we could consider more values), and we want to guess the value of X based on Y. We call this guess $\hat{X}$. There are a few possible formulations. In all cases, we assume that $f_{[Y|X]}[y \mid x]$ is known. It tells us how Y is related to X.

8.5.1 MAP

We know $P(X = x)$ for $x = 0, 1$. We want to minimize $P(X \neq \hat{X})$. The solution is

$$\hat{X} = MAP[X \mid Y = y] := \text{argmax}_x f_{[Y|X]}[y \mid x] P(X = x).$$

8.5.2 MLE

We do not know $P(X = x)$ for $x = 0, 1$. By definition, the MLE of X given Y is

$$\hat{X} = MLE[X \mid Y = y] := \text{argmax}_x f_{[Y|X]}[y \mid x].$$

The MLE is the MAP when the values of X are equally likely prior to any observation.

8.5.3 Hypothesis Test

We do not know $P(X = x)$ for $x = 0, 1$. We are given an acceptable probability β of deciding $\hat{X} = 1$ when $X = 0$ (false alarm) and we want to minimize the probability of deciding $\hat{X} = 0$ when $X = 1$ (missed detection). The solution is

$$\hat{X} = HT[X \mid Y] := \begin{cases} 1, \text{ if } L(Y) > \lambda \\ 0, \text{ if } L(Y) < \lambda \\ 1 \text{ with probability } \gamma, \text{ if } L(Y) = \lambda. \end{cases}$$

Here,

$$L(y) = \frac{f_{[Y|X]}[y \mid 1]}{f_{[Y|X]}[y \mid 0]}$$

is the *likelihood ratio* and we have to choose $\lambda \geq 0$ and $\gamma \in [0,1]$ so that $P[\hat{X} = 1 \mid X = 0] = \beta$.

Easy case 1: Continuous

If $L(Y)$ is a continuous random variable, then we don't have to bother with the case $L(Y) = \lambda$. The solution is then

$$\hat{X} = \begin{cases} 1, \text{ if } L(Y) > \lambda \\ 0, \text{ if } L(Y) < \lambda \end{cases}$$

where we choose λ so that $P[\hat{X} = 1 \mid X = 0] = \beta$.

Easy case 2: Continuous and Monotone

In some cases, $L(Y)$ is a continuous random variable that is strictly increasing in Y. Then the decision is

$$\hat{X} = 1\{Y \geq y_0\}$$

where y_0 is such that

$$P[Y \geq y_0 \mid X = 0] = \alpha.$$

8.6 Solved Problems

Example 8.6.1. *Given X, $Y = N(0.1 + X, 0.1 + X)$, for $X \in \{0,1\}$. (Model inspired from optical communication links.) Assume that $P(X = 0) =: \pi(0) = 0.4$ and $P(X = 1) =: \pi(1) = 0.6$. a. Find $\hat{X} = MAP[X \mid Y]$. b. Calculate $P(\hat{X} \neq X)$.*

a. We find

$$f_{[Y|X]}[y \mid 1] = \frac{1}{\sqrt{2.2\pi}} \exp\{-\frac{(y-1.1)^2}{2.2}\}$$

and

$$f_{[Y|X]}[y \mid 0] = \frac{1}{\sqrt{0.2\pi}} \exp\{-\frac{(y-0.1)^2}{0.2}\}.$$

Hence, $\hat{X} = 1$ if $f_{[Y|X]}[y \mid 1]\pi(1) \geq f_{[Y|X]}[y \mid 0]\pi(0)$. After some algebra, one finds that this condition is equivalent to

$$\hat{X} = 1\{y \notin (-0.53, 0.53)\}.$$

Intuitively, if $X = 0$, then $Y =_D N(0.1, 0.1)$ and is likely to be close to zero. On the other hand, if $X = 1$, then $Y =_D N(1.1, 1.1)$ and is more likely to take large positive value or negative values.

b. The probability of error is computed as follows:

$$\begin{aligned}
P(\hat{X} \neq X) &= P(X = 0)P[\hat{X} = 1 \mid X = 0] + P(X = 1)P[\hat{X} = 1 \mid X = 1] \\
&\quad 0.4P[|Y| > 0.53 \mid X = 0] + 0.6P[|Y| < 0.53 \mid X = 1] \\
&= 0.4P(|N(0.1, 0.1)| > 0.53) + 0.6P(|N(1.1, 1.1)| < 0.53) \\
&= 0.8P(N(0.1, 0.1) < -0.53) + 1.2P(N(1.1, 1.1) < -0.53) \\
&= 0.8P(N(0, 0.1) < -0.63) + 1.2P(N(0, 1.1) < -1.63) \\
&= 0.8P(N(0, 1) < -0.63/\sqrt{0.1}) + 1.2P(N(0, 1) < -1.63/\sqrt{1.1}) \\
&= 0.8P(N(0, 1) < -1.99) + 1.2P(N(0, 1) < -1.554) \approx 0.09.
\end{aligned}$$

We used a calculator to evaluate the last expression.

Example 8.6.2. *Given $X = i$, $Y = Exd(\lambda_i)$, for $i = 0, 1$. Assume $\lambda_0 < \lambda_1$. Find $\hat{X} = MAP[X \mid Y]$ and calculate $P(\hat{X} \neq X)$. We know $\pi(i) = P(X = i)$ for $i = 0, 1$.*

We find, for $i = 0, 1$,

$$f_{[Y|X]}[y \mid i]\pi(i) = \pi(i)\lambda_i e^{-\lambda_i y},$$

so that $\hat{X} = 1$ if

$$\pi(1)\lambda_1 e^{-\lambda_1 y} \geq \pi(0)\lambda_0 e^{-\lambda_0 y}.$$

i.e., if

$$y \leq \frac{1}{\lambda_1 - \lambda_0} \ln\{\frac{\pi(1)\lambda_1}{\pi(0)\lambda_0}\} =: y_0.$$

Consequently,

$$P(\hat{X} \neq X) = P(\hat{X} = 0, X = 1) + P(\hat{X} = 1, X = 0)$$
$$= \pi(1)P[Y > y_0 \mid X = 1] + \pi(0)P[Y < y_0 \mid X = 0] = \pi(1)e^{-\lambda_1 y_0} + \pi(0)(1 - e^{-\lambda_0 y_0}).$$

Example 8.6.3. *When* $X = 1, Y = N(0, 2)$ *and when* $X = 0, Y = N(1, 1)$. *Find* $\hat{X} = HT[X \mid Y]$ *with* $\beta = 10^{-2}$.

We have

$$L(y) = [\frac{1}{\sqrt{2\pi 2}} \exp\{-\frac{y^2}{4}\}]/[\frac{1}{\sqrt{2\pi}} \exp\{-\frac{(y-1)^2}{2}\}].$$

We see that $L(y)$ is a strictly increasing function of $y^2 - 4y$ (not of y!). Thus, $L(y) \geq \lambda$ is equivalent to $y^2 - 4y \geq \tau$ for some τ. Accordingly, we can write the solution as

$$\hat{X} = 1\{y^2 - 4y \geq \tau\}$$

where we choose τ so that $P[\hat{X} = 1 \mid X = 0] = \beta$, i.e., so that

$$P[Y^2 - 4Y \geq \tau | X = 0] = \beta.$$

Note that $Y^2 - 4Y \geq \tau$ if and only if $Y \leq y_0$ or $Y \geq y_1$ where

$$y_0 = 2 - \sqrt{4 + \tau} \text{ and } y_1 = 2 + \sqrt{4 + \tau}.$$

How do we find the value of τ for $\beta = 10^{-2}$? A brute force approach consists in calculating $g(\tau) := P[Y^2 - 4Y \geq \tau | X = 0]$ for different values of τ and to zoom in on what works. I used excel and explored $g(\tau)$. We find $\tau = 40.26$. Thus,

$$\hat{X} = 1\{Y < -4.65 \text{ or } Y > 8.65\}.$$

Example 8.6.4. *If $X = 0$, $Y =_D Exd(\lambda_0)$ and if $X = 1$, $Y =_D Exd(\lambda_1)$ where $\lambda_0 = 1 > \lambda_1 > 0$. Let $\beta = 10^{-2}$. Find $HT[X \mid Y]$.*

We have

$$L(y) = \frac{\lambda_1 e^{-\lambda_1 y}}{\lambda_0 e^{-\lambda_0 y}}.$$

Thus $L(y)$ is strictly increasing in y and is continuous. Thus,

$$\hat{X} = 1\{Y > y_0\}$$

where y_0 is such that

$$P[Y > y_0 \mid X = 0] = e^{-y_0} = \alpha, \text{ i.e., } y_0 = -\ln(\beta).$$

This decision rule does not depend on the value of $\lambda_1 < 1$. Accordingly, $\hat{X}$ solves the problem of deciding between $H_0 : E(Y) = 1$ and $H_1 : E(Y) > 1$ so as to minimize the probability of deciding H_0 when H_1 is in force subject to the probability of deciding H_1 when H_0 is in force being at most β.

Example 8.6.5. *If $X = 0$, $Y = U[-1, 1]$ and if $X = 1$, $Y = U[0, 2]$. Calculate $min P[\hat{X} = 0 \mid X = 1]$ over all $\hat{X}$ based on Y such that $P[\hat{X} = 1 \mid X = 0] \leq \beta$.*

Here,

$$L(y) = \frac{1\{0 \leq y \leq 2\}}{1\{-1 \leq y \leq 1\}}.$$

Thus, $L(y) = 0$ for $y \in [-1, 0)$; $L(y) = 1$ for $y \in [0, 1]$; and $L(y) = \infty$ for $y \in (1, 2]$. The decision is then

$$\hat{X} = HT[X \mid Y] := \begin{cases} 1, \text{ if } Y > 1 \\ 0, \text{ if } Y < 0 \\ 1 \text{ with probability } \gamma, \text{ if } Y \in [0, 1]. \end{cases}$$

We choose γ so that

$$\begin{aligned}\beta &= 0.2 = P[\hat{X} = 1 \mid X = 0] \\ &= \gamma P[Y \in [0,1] \mid X = 0] + P[Y > 1 \mid X = 0] = \gamma\frac{1}{2},\end{aligned}$$

so that $\gamma = 2\beta$. It then follows that

$$\begin{aligned}P[\hat{X} = 0 \mid X = 1] &= P[Y < 0 \mid X = 1] + (1-\gamma)P[Y \in [0,1] \mid X = 1] \\ &= (1-\gamma)\frac{1}{2} = (1-2\beta)\frac{1}{2} = \frac{1}{2} - \beta.\end{aligned}$$

Example 8.6.6. *Pick the point (X, Y) uniformly in the triangle $\{(x, y) \mid 0 \le x \le 1$ and $0 \le y \le x\}$.*

a. Find the function $g : [0,1] \to \{0, 0.5, 1\}$ that minimizes $E((X - g(Y))^2)$.

b. Find the function $g : [0,1] \to \Re$ that minimizes $E(h(X - g(Y)))$ where $h(\cdot)$ is a function whose primitive integral $H(\cdot)$ is anti-symmetric strictly convex over $[0, \infty)$. For instance, $h(u) = u^4$ or $h(u) = |u|$.

a. The key observation is that, as follows from (6.4.4),

$$E((X - g(Y))^2) = E(E[(X - g(Y))^2 \mid Y]).$$

For each y, we should choose the value $g(y) := v \in \{0, 0.5, 1\}$ that minimizes $E[(X - v)^2 \mid Y = y]$. Recall that Given $\{Y = y\}$, X is $U[y, 1]$. Hence,

$$\begin{aligned}E[(X - v)^2 \mid Y = y] &= \frac{1}{1-y}\int_y^1 (x - v)^2 dx = \frac{1}{3(1-y)}[(x - v)^3]_y^1 \\ &= \frac{1}{3(1-y)}[(1 - v)^3 - (y - v)^3].\end{aligned}$$

We expect that the minimizing value $g(y)$ of v is nondecreasing in y. That is, we expect that

$$g(y) = \begin{cases} 0, \text{ if } y \in [0, a) \\ 0.5, \text{ if } y \in [a, b) \\ 1, \text{ if } y \in [b, 1]. \end{cases}$$

The "critical" values $a < b$ are such that the choices are indifferent. That is,

$E[(X-0)^2 \mid Y = a] = E[(X-0.5)^2 \mid Y = a]$ and $E[(X-0.5)^2 \mid Y = b] = E[(X-1)^2 \mid Y = b]$.

Substituting the expression for the conditional expectation, these equations become

$$(1-a)^3 = (0.5)^3 - (a-0.5)^3 \text{ and } (0.5)^3 - (b-0.5)^3 = -(b-1)^3.$$

Solving these equations, we find $a = b = 0.5$. Hence,

$$g(y) = \begin{cases} 0, \text{ if } y < 0.5 \\ 1, \text{ if } y \geq 0.5. \end{cases}$$

b. As in previous part,

$$E(h(X - g(Y))) = E(E[h(X - g(Y)) \mid Y]),$$

so that, for each given y, we should choose $g(y)$ to be the value v that minimizes $E[h(X-v) \mid Y = y]$. Now,

$$E[h(X-v) \mid Y = y] = \frac{1}{1-y}\int_y^1 h(x-v)dx = \frac{1}{1-y}[H(1-v) - H(y-v)].$$

Now, we claim that the minimizing value of v is $v^* = (1+y)/2$. To see that, note that for $v \in (y, 1)$ one has

$$H(1-v)-H(y-v) = H(1-v)+H(v-y) > 2H(\frac{(1-v)+(v-y)}{2}) = H(1-v^*)+H(v^*-y),$$

by anti-symmetry and convexity.

Example 8.6.7. *For $x, y \in \{0,1\}$, let $P[Y = y \mid X = x] = P(x,y)$ where $P(0,0) = 1 - P(0,1) = 0.7$ and $P(1,1) = 1 - P(1,0) = 0.6$. Assume that $P(X = 1) = 1 - P(X = 0) = p \in [0,1]$.*

a. Find the MLE of X given Y.

b. Find the MAP of X given Y.

c. Find the estimate $\hat{X}$ based on Y that minimizes $P[\hat{X} = 0 \mid X = 1]$ subject to $P[\hat{X} = 1 \mid X = 0] \leq \beta$, for $\beta \in (0,1)$.

The solution is a direct application of the definitions plus some algebra.

a. By definition,

$$MLE[X \mid Y = y] = \arg\max_x P[Y = y \mid X = x] = \arg\max_x P(x, y).$$

Hence, $MLE[X \mid Y = 0] = 0$ because $P[Y = 0 \mid X = 0] = 0.7 > P[Y = 0 \mid X = 1] = 0.4$. Similarly, $MLE[X \mid Y = 1] = 1$ because $P[Y = 1 \mid X = 1] = 0.6 > P[Y = 1 \mid X = 0] = 0.3$. Consequently,

$$MLE[X \mid Y] = Y.$$

b. We know that

$$MAP[X \mid Y = y] = \arg\max_x P(X = x)P[Y = y \mid X = x] = \arg\max_x P(X = x)P(x, y).$$

Therefore,

$$MAP[X \mid Y = 0]$$
$$= \arg\max_x \{g(x = 0) = P(X = 0)P(0, 0) = 0.7(1 - p), g(x = 1) = P(X = 1)P(1, 0) = 0.4p\}$$

and

$$MAP[X \mid Y = 1]$$
$$= \arg\max_x \{h(x = 0) = P(X = 0)P(0, 1) = 0.3(1 - p), h(x = 1) = P(X = 1)P(1, 1) = 0.6p\}.$$

Consequently,

$$MAP[X|Y = y] = \begin{cases} 0, \text{ if } y = 0 \text{ and } p < 7/11; \\ 1, \text{ if } y = 0 \text{ and } p \geq 7/11; \\ 0, \text{ if } y = 1 \text{ and } p < 1/3; \\ 1, \text{ if } y = 1 \text{ and } p \geq 1/3. \end{cases}$$

c. We know that

$$\hat{X} = \begin{cases} 1, \text{ if } L(y) = P(1, y)/P(0, y) > \lambda; \\ 1 \text{ w.p. } \gamma, \text{ if } L(y) = P(1, y)/P(0, y) = \lambda; \\ 0, \text{ if } L(y) = P(1, y)/P(0, y) < \lambda. \end{cases}$$

We must find λ and γ so that $P[\hat{X} = 1 \mid X = 0] = \beta$. Note that $L(1) = 2$ and $L(0) = 4/7$. Accordingly,

$$P[\hat{X} = 1 \mid X = 0] = \begin{cases} 0, \text{ if } \lambda > 2; \\ 1 \text{ w.p. } P(0,1)\gamma = 0.3\gamma \text{ if } \lambda = 2; \\ 1 \text{ w.p. } P(0,1) = 0.3 \text{ if } \lambda \in (4/7, 2); \\ 1 \text{ w.p. } P(0,0)\gamma + P(0,1) = 0.7\gamma + 0.3 \text{ if } \lambda = 4/7; 1, \text{ if } \lambda < 4/7. \end{cases}$$

To see this, observe that if $\lambda > 2$, then $L(0) < L(1) < \lambda$, so that $\hat{X} = 0$. Also, if $\lambda = 2$, then $L(0) < \lambda = L(1)$, so that $\hat{X} = 1$ w.p. γ when $Y = 1$. Since $P[Y = 1 \mid X = 0] = P(0,1)$, we see that $P[\hat{X} = 1 \mid X = 0] = P(0,1)\gamma$. The other cases are similar.

It follows that

$$\lambda = 2, \gamma = \beta/0.3, \text{ if } \beta \leq 0.3;$$

$$\lambda = 4/7, \gamma = (\beta - 0.3)/0.7, \text{ if } \beta > 0.3.$$

Example 8.6.8. *Given X, the random variables $\{Y_n, n \geq 1\}$ are exponentially distributed with mean X. Assume that $P(X = 1) = 1 - P(X = 2) = p \in (0,1)$.*

a. Find the MLE of X given $Y^n := \{Y_1, \ldots, Y_n\}$.

b. Find the MAP of X given Y^n.

c. Find the estimate $\hat{X}_n$ based on Y^n that minimizes $P[\hat{X}_n = 1 \mid X = 2]$ subject to $P[\hat{X}_n = 2 \mid X = 1] \leq \beta$, for $\beta \in (0,1\}$.

First we compute the likelihood ratio $L(y^n)$. We find

$$L(y^n) = \frac{f_{Y^n|X=2}(y^n|X=2)}{f_{Y^n|X=1}(y^n|X=1)} = (\frac{1}{2})^n e^{\frac{1}{2}\sum_{i=1}^n y_i}.$$

a. Recall that $MLE[X \mid Y^n = y^n] = 2$ if $L(y^n) > 1$ and is equal to 1 otherwise. Hence,

$$MLE[X \mid Y^n = y^n] = \begin{cases} 1, \text{ if } \frac{1}{n}\sum_{i=1}^n y_i < 2\ln(2); \\ 2, \text{ otherwise.} \end{cases}$$

b. We know that $MAP[X \mid Y^n = y^n] = 2$ if $L(y^n) > P(X = 1)/P(X = 2)$. Hence,

$$MAP[X \mid Y^n = y^n] = \begin{cases} 1, \text{ if } \frac{1}{n}\sum_{i=1}^n y_i < 2\ln(2) + \frac{p}{n(1-p)}; \\ 2, \text{ otherwise.} \end{cases}$$

c. Ignoring the unlikely marginal case when $L(y^n) = \lambda$, we know that $\hat{X} = HT[X \mid Y^n = y^n] = 2$ if $L(y^n) > \lambda$ and is equal to 1 otherwise, for some λ. Equivalently,

$$\hat{X} = \begin{cases} 2, \text{ if } \sum_{i=1}^n Y_i > \rho; \\ 1, \text{ if } \sum_{i=1}^n Y_i \leq \rho. \end{cases}$$

We choose ρ so that $P[\hat{X} = 2 \mid X = 1] = \beta$, i.e., so that

$$P[\sum_{i=1}^n Y_i > \rho \mid X = 1] = \beta.$$

Unfortunately, there is no closed-form solution and one has to resort to a computer to determine the suitable value of ρ. After Chapter 11 we will be able to use a Gaussian approximation. See Example 11.7.7.

Example 8.6.9. *Let X, Y be independent random variables where $P(X = -1) = P(X = 0) = P(X = +1) = 1/3$ and Y is $N(0, \sigma^2)$. Find the function $g : \Re \to \{-1, 0, +1\}$ that minimizes $P(X \neq g(X + Y))$.*

Let $Z = X + Y$. Since the prior distribution of X is uniform, we know that the solution is $\hat{X} = MLE[X \mid Z]$. That is, $g(y) = \arg\max_x f_{Z|X}[z|x]$.

Now,

$$f_{Z|X}[z|x] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(z-x)^2}{2\sigma^2}\}.$$

Hence,

$$g(y) = \arg\min_x |z - x|.$$

Consequently,

$$\hat{X} = \begin{cases} -1, \text{ if } z \leq -0.5; \\ 0, \text{ if } -0.5 < z < 0.5; \\ 1, \text{ if } z \geq 0.5. \end{cases}$$

Example 8.6.10. *Assume that X is uniformly distributed in the set $\{1,2,3,4\}$. When $X=i$, one observes $\boldsymbol{Y}=N(\boldsymbol{v}_i,\boldsymbol{I})=_D \boldsymbol{v}_i+\boldsymbol{Z}$ where $\boldsymbol{v}_i\in\Re^2$ for $i=1,2,3,4$, $\boldsymbol{I}$ is the identity matrix in $\Re^{2\times 2}$, and $\boldsymbol{Z}=_D N(\boldsymbol{0},\boldsymbol{I})$.*

a. Find the function $g:\Re^2\to\{1,2,3,4\}$ that minimizes $P(X\neq g(\boldsymbol{Y}))$.

b. Can you find an estimate of $P(X\neq g(\boldsymbol{Y}))$ where $g(.)$ was found in part (a)?

a. From the statement,

$$f[\boldsymbol{y}|X=i]:=f_{\boldsymbol{Y}|X}[\boldsymbol{y}|i]=\frac{1}{2\pi}\,exp[-\frac{1}{2}||\boldsymbol{y}-\boldsymbol{v_i}||^2].$$

Also, because the prior of X is uniform, we now that $g(\boldsymbol{Y})=MLE[X|\boldsymbol{Y}]$. That is,

$$g(\boldsymbol{y})=\arg\max_i f[\boldsymbol{y}|X=i]=\arg\min_i||\boldsymbol{y}-\boldsymbol{v_i}||.$$

That is, the MAP of X given $\boldsymbol{Y}$ corresponds to the vector $\boldsymbol{v_i}$ that is closest to the received vector $\boldsymbol{Y}$. This is fairly intuitive, given the shape of the Gaussian density.

b. It is difficult to get precise estimates. However, we can say that $P[\hat{X}=i\mid X=i]\geq\alpha_i$ where α_i is the probability that $||\boldsymbol{Z}||<0.5\min\{||\boldsymbol{v}_i-\boldsymbol{v}_j||,j\neq i\}=:0.5d_i$. Indeed, this condition guarantees that $\boldsymbol{Z}$ is closer to $\boldsymbol{v}_i$ than to any $\boldsymbol{v}_j$. Now,

$$P(||\boldsymbol{Z}||\leq d)=P(Z_1^2+Z_2^2\leq\alpha^2)=1-\exp\{-0.5d^2\}$$

where the last inequality follows from the result in Example 7.5.6.

Hence,

$$P[\hat{X}=i\mid X=i]\geq 1-\exp\{-0.5d_i^2\}.$$

Consequently,

$$\begin{aligned}P[\hat{X}=X] &= \sum_{i=1}^4 P[\hat{X}=i\mid X=i]P(X=i)\\ &\geq \frac{1}{4}\sum_{i=1}^4(1-\exp\{-0.5d_i^2\}).\end{aligned}$$

For instance, assume that the vectors $\boldsymbol{v}_i$ are the four corners of the square $[-\sqrt{\rho/2}, \sqrt{\rho/2}]^2$. In that case, $d_i = \sqrt{\rho/2}$ for all i and we find that

$$P[\hat{X} = X] \geq 1 - \exp\{-0.25\rho\}.$$

Note also that $||\boldsymbol{v}_i||^2 = \rho$, so that ρ is the power that the transmitter sends. As ρ increases, so does our lower bound on the probability of decoding the signal correctly.

This type of simple bound is commonly used in the evaluation of communication systems.

Example 8.6.11. *A machine produces steel balls for ball bearings. When the machine operates properly, the radii of the balls are i.i.d. and* $N(100, 4)$. *When the machine is defective, the radii are i.i.d. and* $N(98, 4)$.

a. You measure n balls produced by the machine and you must raise an alarm if you believe that the machine is defective. However, you want to limit the probability of false alarm to 1%. Explain how you propose to do this.

b. Compute the probability of missed detection that you obtain in part (a). This probability depends on the number n of balls, so you cannot get an explicit answer. Select the value of n so that this probability of missed detection is 0.1%.

a. This is a hypothesis test. Let $X = 0$ when the machine operates properly and $X = 1$ otherwise. Designate by $Y_1, \ldots, Y_n$ the radii of the balls. The likelihood ratio is

$$\begin{aligned} L(y_1, \ldots, y_n) &= \frac{\exp\{-\frac{1}{8}\sum_{k=1}^n (y_k - 98)^2\}}{\exp\{-\frac{1}{8}\sum_{k=1}^n (y_k - 100)^2\}} \\ &= \exp\{-\frac{1}{2}\sum_{k=1}^n y_k + 49.5 \times n\}. \end{aligned}$$

Since $L(y_1, \ldots, y_n)$ is decreasing in $\sum_{k=1}^n y_k$, we conclude that

$$\hat{X} = \begin{cases} 1, \text{ if } \sum_{k=1}^n Y_k < \lambda \\ 0, \text{ if } \sum_{k=1}^n Y_k > \lambda \end{cases}$$

where λ is such that $P[\hat{X} = 1|X = 0] = 1\%$. That is,

$$\begin{aligned} 1\% &= P[\sum_{k=1}^{n} Y_k < \lambda | X = 0] = P(N(n \times 100, n \times 4) < \lambda) \\ &= P(N(0, 4n) < \lambda - 100n) = P(N(0,1) < \frac{\lambda - 100n}{2\sqrt{n}}). \end{aligned}$$

Accordingly, we find that

$$\frac{\lambda - 100n}{2\sqrt{n}} = -2.3, \text{ i.e., } \lambda = 100n - 4.6\sqrt{n}.$$

b. We have

$$\begin{aligned} P[\hat{X} = 0|X = 1] &= P[\sum_{k=1}^{n} Y_k > \lambda | X = 1] = P(N(n \times 98, n \times 4) > \lambda) \\ &= P(N(0,1) > \frac{\lambda - 98n}{2\sqrt{n}}) = P(N(0,1) > \frac{2n - 4.6\sqrt{n}}{2\sqrt{n}}) \\ &= P(N(0,1) > \sqrt{n} - 2.3). \end{aligned}$$

For this probability to be equal to 0.1%, we need

$$\sqrt{n} - 2.3 = 3.1, \text{ i.e., } n = (2.3 + 3.1)^2 \approx 29.16.$$

We conclude that we have to measure 30 balls.

Chapter 9

Estimation

The estimation problem is similar to the detection problem except that the unobserved random variable X does not take values in a finite set. That is, one observes $Y \in \Re$ and one must compute an estimate of $X \in \Re$ based on Y that is close to X in some sense.

Once again, one has Bayesian and non-Bayesian formulations. The non-Bayesian case typically uses $MLE[X|Y]$ defined as in the discussion of Detection.

9.1 Properties

An estimator of X given Y is a function $g(Y)$. The estimator $g(Y)$ is *unbiased* if $E(g(Y)) = E(X)$, i.e., if its mean is the same as that of X. Recall that $E[X \mid Y]$ is unbiased.

If we make more and more observations, we look at the estimator $\hat{X}_n$ of X given $(Y_1, \ldots, Y_n) : \hat{X}_n = g_n(Y_1, \ldots, Y_n)$. We say that $\hat{X}_n$ is *asymptotically unbiased* if $\lim E(\hat{X}_n) = E(X)$.

9.2 Linear Least Squares Estimator: LLSE

In this section we study a class of estimators that are linear in the observations. They have the advantage of being easy to calculate.

Definition 9.2.1. LLSE

Let (X, Y) be a pair of random variables on some probability space. The *linear least squares estimator* (LLSE) of X given Y, designated by $L[X \mid Y]$, is the linear function $a + bY$ of Y that minimizes $E((X - a - bY)^2)$.

In the multivariate case, let $\boldsymbol{X}, \boldsymbol{Y}$ be vectors of random variables on some probability space. The LLSE of $\boldsymbol{X}$ given $\boldsymbol{Y}$, designated by $L[\boldsymbol{X} \mid \boldsymbol{Y}]$ is the linear function $\boldsymbol{a} + \boldsymbol{BY}$ that minimizes $E(||\boldsymbol{X} - \boldsymbol{a} - \boldsymbol{BY}||^2)$.

The next theorem summarizes the key result.

Theorem 9.2.1. *Linear Least Squares Estimator*

One has

$$L[X \mid Y] = E(X) + \frac{cov(X, Y)}{var(Y)}(Y - E(Y)). \tag{9.2.1}$$

Also, if $\Sigma_{\boldsymbol{Y}}$ *is invertible,*

$$L[\boldsymbol{X} \mid \boldsymbol{Y}] = E(\boldsymbol{X}) + \Sigma_{\boldsymbol{X},\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y} - E(\boldsymbol{Y})). \tag{9.2.2}$$

Finally, if $\Sigma_{\boldsymbol{Y}}$ *is not invertible,*

$$L[\boldsymbol{X} \mid \boldsymbol{Y}] = E(\boldsymbol{X}) + \Sigma_{\boldsymbol{X},\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}}^{\dagger}(\boldsymbol{Y} - E(\boldsymbol{Y}). \tag{9.2.3}$$

where $\Sigma_{\boldsymbol{Y}}^{\dagger}$ *is such that*

$$\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}}^{\dagger} = \boldsymbol{\Sigma}_{\boldsymbol{Y},\boldsymbol{Y}}.$$

Proof:

We provide the proof in the scalar case. The vector case is very similar and we leave the details to the reader. The key observation is that $Z = LLSE[X|Y]$ if and only if $Z = a+bY$ is such that $E(Z) = E(X)$ and $\text{cov}(X - Z, Y) = 0$.

To see why that is the case, assume that Z has those two properties and let $V = c+dY$ be some other linear estimator. Then,

$$E((X-V)^2) = E((X-Z+Z-V)^2) = E((X-Z)^2)+E((Z-V)^2)+2E((X-Z)(Z-V)).$$

But

$$\begin{aligned} E((X-Z)(Z-V)) &= E((X-Z)(a+bY-c-dY)) \\ &= (a-c)E(X-Z) + (b-d)E((X-Z),Y) \\ &= (a-c)(E(X)-E(Z)) + (b-d)\text{cov}(X-Z,Y) \end{aligned}$$

where the last term is justified by the fact that

$$\text{cov}(X-Z,Y) = E((X-Z)Y) - E(X-Z)E(Y) = E((X-Z)Y),$$

since $E(X) = E(Z)$. Hence, $E((X-Z)(Z-V)) = 0$ as we find that

$$E((X-V)^2) = E((X-Z)^2) + E((Z-V)^2) \geq E((X-Z)^2),$$

so that $Z = LLSE[X \mid Y]$.

Conversely, if $Z = LLSE[X \mid Y] = a+bY$, then $\phi(c,d) := E((X-c-dY)^2)$ is minimized by $c = a$ and $d = b$. Setting the derivative of $\phi(c,d)$ with respect to c to zero and similarly with respect to d, we find

$$0 = \frac{\partial \phi(c,d)}{\partial c}|_{c=a,d=b} = 2E(X-a-bY) = 2E(X-Z)$$

and

$$0 = 2E((X-a-bY)Y) = 2E((X-Z)Y).$$

These two expressions imply that $E(X) = E(Z)$ and $\text{cov}(X-Z,Y) = 0$.

These conditions show that

$$Z = E(X) + \frac{\text{cov}(X,Y)}{\text{var}(Y)}(Y - E(Y)).$$

□

Note that these conditions say that the estimation error $X - Z$ should be *orthogonal* to Y, where orthogonal means zero-mean and uncorrelated with Y. We write $X - Z \perp Y$

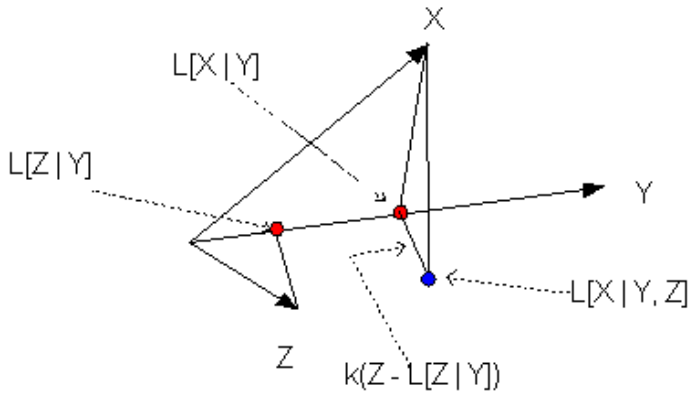


Figure 9.1: Adding observations to improve an estimate

to indicate that $X - Z$ is orthogonal to Y. That is, Z is the *projection* of X on the set of linear functions of Y: $\{V = c + dY \mid c, d \in \Re\}$.

9.3 Recursive LLSE

There are many cases where one keeps on making observations. How do we update the estimate? For instance, how do we calculate $LLSE[X|Y,Z] = bY + cZ$ if one knows $\hat{X} = LLSE[X|Y] = aY$? The answer lies in the observations captured in Figure 9.1 (we assume all the random variables are zero-mean to simplify the notation).

We want $X - bY + cZ \perp Y, Z$ where $\perp$ designates orthogonality. A picture shows that $L[X|Y,Z] = L[X|Y] + k(Z - L[Z|Y])$. We must choose $kX - L[X|Y,Z] \perp Z$.

These ideas lead to the Kalman filter which is a recursive estimator linear in the observations.

9.4 Sufficient Statistics

Assume that the joint pdf of $\boldsymbol{X}$ depends on some parameter Θ. To estimate Θ given $\boldsymbol{X}$ it may be enough to consider functions of $T(\boldsymbol{X})$ instead of all the functions of $\boldsymbol{X}$. This

happens certainly if the density of $\boldsymbol{X}$ given $T(\boldsymbol{X})$ does not depend on Θ. Indeed, in that case, there is no useful information in $\boldsymbol{X}$ about Θ that is not already in $T(\boldsymbol{X})$. In such a situation, we say that $T(\boldsymbol{X})$ is a *sufficient statistic* for (estimating) Θ. Equivalently, $T(\boldsymbol{X})$ is a sufficient statistic for (estimating) Θ if the density $f_{\boldsymbol{X}|\Theta}[\boldsymbol{x}|\theta]$ of $\boldsymbol{X}$ given Θ has the following form:

$$f_{\boldsymbol{X}|\Theta}[\boldsymbol{x}|\theta] = h(\boldsymbol{x})g(T(\boldsymbol{x});\theta).$$

We provide the formal derivation of these ideas in Example 9.6.4.

For instance, if given $\{\Theta = \theta\}$ the random variables $X_1, \ldots, X_n$ are i.i.d. $N(\theta, \sigma^2)$, then one can show that $X_1 + \ldots + X_n$ is a sufficient statistic for Θ. Knowing a sufficient statistic enables us to "compress" observations without loss of relevant information.

Assume that $T(\boldsymbol{X})$ is a sufficient statistic for (estimating) Θ. Then you can verify that $MLE[\Theta|\boldsymbol{X}], MAP[\Theta|\boldsymbol{X}], HT[\Theta|\boldsymbol{X}]$, and $E[\Theta|\boldsymbol{X}]$ are all functions of $T(\boldsymbol{X})$. The corresponding result does not hold for $LLSE[\Theta|\boldsymbol{X}]$.

9.5 Summary

9.5.1 LSSE

We discussed the linear least squares estimator of $\boldsymbol{X}$ given $\boldsymbol{Y}$. The formulas are given in (9.2.1), (9.2.2), and (9.2.3).

The formulas are the same as those for the conditional expectation when the random variables are jointly Gaussian. In the non-Gaussian case, the conditional expectation is not linear and the LLSE is not as close to X as the conditional expectation. That is, in general,

$$E((X - E[X \mid Y])^2) \leq E((X - L[X \mid Y])^2),$$

and the inequality is strict unless the conditional expectation happens to be linear, as in the jointly Gaussian case or other particular cases.

We also explained the notion of sufficient statistic $T(\boldsymbol{X})$ that contains all the information in the observations $\boldsymbol{X}$ that is relevant for estimating some parameter Θ. The necessary and sufficient condition is (9.6.1).

9.6 Solved Problems

Example 9.6.1. *Let X, Y be a pair of random variables. Find the value of a that minimizes the variance of $X - aY$.*

Note that

$$E((X-aY-b)^2) = \text{var}(X-aY-b) + (E(X-aY-b))^2 = \text{var}(X-aY) + (E(X-aY-b))^2.$$

We also know that the values of a and b that minimize $E((X - aY - b)^2)$ are such that $a = \text{cov}(X, Y)/\text{var}(Y)$. Consequently, that value of a minimizes $\text{var}(X - aY)$.

Example 9.6.2. *The random variable X is uniformly distributed in $\{1, 2, \ldots, 100\}$. You are presented a bag with X blue balls and $100 - X$ red balls. You pick 10 balls from the bag and get b blue balls and r red balls. Explain how to calculate the maximum a posteriori estimate of X.*

Your intuition probably suggests that if we get b blue balls out of 10, there should be about $10b$ blue balls out of 100. Thus, we expect the answer to be $x = 10b$. We verify that intuition.

Designate by A the event "we got b blue balls and $r = 10 - b$ red balls." Since the prior p.m.f. of X is uniform,

$$MAP[X \mid A] = MLE[X \mid A] = \text{argmax}_x P[A \mid X = x].$$

Now,

$$P[A \mid X = x] = \frac{\binom{x}{b}\binom{100-x}{r}}{\binom{100}{10}}.$$

Hence,

$$MAP[X \mid A] = \text{argmax}_x \binom{x}{b}\binom{100-x}{r} = \text{argmax}_x \frac{x!r!(100-x-r)!}{b!(x-b)!(100-x)!}.$$

Hence, $MAP[X \mid A] = \text{argmax}_x \alpha(x)$ where

$$\alpha(x) := \frac{x!(90+b-x)!}{(x-b)!(100-x)!}.$$

We now verify that $\alpha(x)$ is minimized by $x = 10b$.

Note that

$$\frac{\alpha(x)}{\alpha(x-1)} = \frac{x(90+b-x)}{(x-b)(100-x)}.$$

Hence,

$$\alpha(x) < \alpha(x-1) \text{ iff } x(90+b-x) < (x-b)(100-x),$$

i.e.,

$$\alpha(x) < \alpha(x-1) \text{ iff } x > 10b.$$

Thus, as x increases from b to $100-r = 90+b$, we see that $\alpha(x)$ increases as long as $x \leq 10b$ and then decreases. It follows that

$$MAP[X \mid A] = 10b.$$

Usually, intuition is much quicker than the algebra... .

Example 9.6.3. *Let X, Y, Z be i.i.d. and with a $B(n,p)$ distribution (that is, Binomial with parameters n and p).*

a. What is var(X)?

b. Calculate $L[X \mid X+2Y, X+3Z]$.

c. Calculate $L[XY \mid X+Y]$.

a. Since X is the sum of n i.i.d. $B(p)$ random variables, we find $\text{var}(X) = np(1-p)$.

b. We use the formula (9.2.2) and find

$$L[X \mid X+2Y, X+3Z] = np + [np(1-p), np(1-p)]\Sigma^{-1}[\begin{pmatrix} X+2Y \\ X+3Z \end{pmatrix} - \begin{pmatrix} 3np \\ 4np \end{pmatrix}]$$

where

$$\Sigma = \begin{pmatrix} 5np(1-p) & np(1-p) \\ np(1-p) & 10np(1-p) \end{pmatrix} = np(1-p)\begin{pmatrix} 5 & 1 \\ 1 & 10 \end{pmatrix}, \text{ so that } \Sigma^{-1} = \frac{1}{np(1-p)}\begin{pmatrix} 10 & -1 \\ -1 & 5 \end{pmatrix}.$$

Putting the pieces together we find

$$L[X \mid X+2Y, X+3Z] = -6np + 9(X+2Y) + 4(X+3Z).$$

c. Similarly,

$$L[XY \mid X+Y] = (np)^2 + \frac{\text{cov}(XY, X+Y)}{\text{var}(X+Y)}(X+Y-2np).$$

Now,

$$\text{cov}(XY, X+Y) = E(X^2Y + XY^2) - E(XY)E(X+Y)$$

and

$$E(X^2Y) = E(X^2)E(Y) = (\text{var}(X) + E(X)^2)E(Y) = (np(1-p) + (np)^2)(np).$$

Hence,

$$\text{cov}(XY, X+Y) = 2(np)^2(1-p) + 2(np)^3 - (np)^2(2np) = 2(np)^2(1-p).$$

Finally

$$L[XY \mid X+Y] = (np)^2 + \frac{2(np)^2(1-p)}{2np(1-p)}(X+Y-2np) = np(X+Y) - (np)^2.$$

Example 9.6.4. *Recall that $T(\boldsymbol{X})$ is a sufficient statistic for estimating Θ given $\boldsymbol{X}$ if*

$$f_{\boldsymbol{X}|\Theta}([\boldsymbol{x} \mid \theta]) = h(\boldsymbol{x})g(T(\boldsymbol{x}); \theta). \tag{9.6.1}$$

a. Show that the identity (9.6.1) holds if and only if the density of $\boldsymbol{X}$ *given* $T(\boldsymbol{X})$ *does not depend on* θ.

b. Show that if given θ *the random variables* $\{X_n, n \geq 1\}$ *are i.i.d.* $N(\theta, \sigma^2)$, *then* $X_1 + \cdots + X_n$ *is a sufficient statistic for estimating* θ *given* $\boldsymbol{X} = \{X_1, \ldots, X_n\}$.

c. Show that if given θ *the random variables* $\{X_n, n \geq 1\}$ *are i.i.d. Poisson with mean* θ, *then* $X_1 + \cdots + X_n$ *is a sufficient statistic for estimating* Θ *given* $\boldsymbol{X} = \{X_1, \ldots, X_n\}$.

d. Give an example where $X_1 + \cdots + X_n$ *is* not *a sufficient statistic for estimating the mean* θ *of i.i.d. random variables* $\{X_1, \ldots, X_n\}$.

a. First assume that (9.6.1) holds. Then

$$P[\boldsymbol{X} \approx \boldsymbol{x} \mid T \approx t, \theta] \approx \frac{P[\boldsymbol{X} \approx \boldsymbol{x}, T \approx t|\theta]}{P[T \approx t|\theta]} \approx \frac{h(\boldsymbol{x})g(t;\theta)d\boldsymbol{x}}{\int_{\{\boldsymbol{x}'|T(\boldsymbol{x}')=t\}} h(x')g(t;\theta)dx'} \approx \frac{h(x)d\boldsymbol{x}}{\int_{\{x'|T(x')=t\}} h(x')dx'}.$$

This shows that the density of $\boldsymbol{X}$ given T does not depend on θ.

For the converse, assume that

$$P[\boldsymbol{X} \approx \boldsymbol{x} \mid T = t, \theta] \approx h(\boldsymbol{x}d\boldsymbol{x}).$$

Then,

$$P[\boldsymbol{X} \approx \boldsymbol{x} \mid \theta] = P[\boldsymbol{X} \approx \boldsymbol{x} \mid T = t, \theta]P[T \approx t \mid \theta] \approx h(\boldsymbol{x})d\boldsymbol{x}P[T \approx t \mid \theta] = h(\boldsymbol{x})g(T(\boldsymbol{x});\theta)d\boldsymbol{x}.$$

(b) This is immediate from (9.6.1) since

$$\begin{aligned} f_{\boldsymbol{X}|\theta}[\boldsymbol{x} \mid \theta] &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\sum_{i=1}^{n}(x_i - \theta)^2/2\sigma^2\} \\ &= [\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\sum_{i=1}^{n} x_i^2\}] \times [\exp\{(-2\sum_{i=1}^{n} x_i + n\theta)^2/2\sigma^2\}] = h(\boldsymbol{x})g(T(\boldsymbol{x});\theta) \end{aligned}$$

where $T(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$,

$$\begin{aligned} h(\boldsymbol{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\sum_{i=1}^{n} x_i^2\}, \text{ and} \\ g(T(\boldsymbol{x});\theta) &= \exp\{(-2T(\boldsymbol{x}) + n\theta)^2/2\sigma^2\}. \end{aligned}$$

(c) This is similar to the previous example. One finds that

$$f_{\boldsymbol{X}|\theta}[\boldsymbol{x} \mid \theta] = \Pi_{i=1}^{n}[\frac{\theta^{x_i}}{x_i!}\exp\{-\theta\}] = \frac{\theta^{\sum_i x_i}}{\Pi_i x_i!}\exp\{-n\theta\} = h(\boldsymbol{x})g(T(\boldsymbol{x});\theta)$$

where $T(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$,

$$\begin{aligned} h(\boldsymbol{x}) &= [\Pi_{i=1}^{n} x_i!]^{-1}, \text{ and} \\ g(T(\boldsymbol{x});\theta) &= \theta^{T(\boldsymbol{x})}\exp\{-n\theta\}. \end{aligned}$$

(d) Assume that, given θ, X_1 and X_2 are i.i.d. $N(\theta,\theta)$. Then, with $\boldsymbol{X} = (X_1, X_2)$ and $T(\boldsymbol{x}) = x_1 + x_2$,

$$f_{\boldsymbol{X}|\theta}[\boldsymbol{x} \mid \theta] = \Pi_{i=1}^{2}[\frac{1}{\sqrt{2\pi\theta}}\exp\{-\frac{(x_i-\theta)^2}{2\theta}\}] = \frac{1}{2\pi\theta}\exp\{-\frac{x_1^2+x_2^2}{2\theta} + T(\boldsymbol{x}) - \theta\}.$$

This function cannot be factorized in the form $h(\boldsymbol{x})g(T(\boldsymbol{x});\theta)$.

Example 9.6.5. *Assume that* $\{X_n, n \geq 1\}$ *are independent and uniformly distributed in* $[0,1]$. *Calculate*

$$L[2X_1 + 3X_2 | X_1^2 + X_3, X_1 + 2X_4].$$

Let $V_1 = 2X_1 + 3X_2, V_2 = X_1^2 + X_3$ and $V_3 = X_1 + 2X_4$

Then $E[V_1] = \frac{5}{2}, E[V_2] = \frac{5}{6}, E[V_3] = \frac{3}{2}$ and $\text{var}(V_1) = \frac{13}{12}, \text{var}(V_2) = \frac{31}{180}, \text{var}(V_3) = \frac{5}{12}$. Also, $\text{cov}(V_1, V_2) = \frac{1}{6}, \text{cov}(V_1, V_3) = \frac{1}{6}$ and $\text{cov}(V_2, V_3) = \frac{8}{3}$. Hence,

$$L[V_1|V_2, V_3] = a + b(V_2 - E[V_2]) + c(V_3 - E[V_3])$$

where

$$a = E[V_1] = \frac{5}{2}$$

and,

$$\begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} Var(V_2) & \text{cov}(V_2, V_3) \\ \text{cov}(V_2, V_3) & Var(V_3) \end{bmatrix}^{-1} \begin{bmatrix} \text{cov}(V_1, V_2) \\ \text{cov}(V_1, V_3) \end{bmatrix} = \begin{bmatrix} \frac{31}{180} & \frac{8}{3} \\ \frac{8}{3} & \frac{5}{12} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{6} \\ \frac{1}{6} \end{bmatrix} = \begin{bmatrix} 0.0533 \\ 0.0591 \end{bmatrix}$$

Hence, $L[V_1|V_2, V_3] = \frac{5}{2} + 0.0533(V_2 - \frac{5}{6}) + 0.0591(V_3 - \frac{3}{2})$.

Example 9.6.6. *Let the point* (X, Y) *be picked randomly in the quarter circle* $\{(x, y) \in \Re_+^2 \mid x^2 + y^2 \leq 1\}$.

a. Find $L[X \mid Y]$.

b. Find $L[X^2 \mid Y]$.

We first calculate the required quantities.

$$f_Y(y) = \int_0^{\sqrt{1-y^2}} \frac{4}{\pi} dx = \frac{4\sqrt{1-y^2}}{\pi}$$

$$\begin{aligned} E(Y^2) &= \int_0^1 y^2 \frac{4\sqrt{1-y^2}}{\pi} dx = \int_0^{\pi/2} \frac{4}{\pi} cos^2(\theta) sin^2(\theta) d\theta \\ &= \int_0^{\pi/2} \frac{1}{\pi} sin^2(2\theta) d\theta = \int_0^{\pi/2} \frac{1}{2\pi}(1 - cos(4\theta)) d\theta \\ &= [\frac{1}{2\pi}(x - \frac{sin(4\theta)}{4})]_0^{\pi/2} = \frac{1}{4}. \end{aligned}$$

Similarly, $E[X^2] = 0.25$.

$$\begin{aligned} E(Y) &= \int_0^1 \frac{4y\sqrt{1-y^2}}{\pi} dy = \int_1^0 \frac{-4}{\pi} z^2 dz \\ &= [\frac{-4z^3}{3\pi}]_1^0 = \frac{4}{3\pi}. \end{aligned}$$

Hence $\text{var}(Y) = \frac{1}{4} - \frac{16}{9\pi^2} = 0.0699$. Similarly, $E[X] = \frac{4}{3\pi}$.

$$\begin{aligned} \text{cov}(XY) &= \int_0^1 \int_0^{\sqrt{1-x^2}} \frac{4}{\pi} xy \, dy dx = \int_0^1 \frac{4}{2\pi} x(1 - x^2) dx \\ &= [\frac{4}{2\pi}(\frac{x^2}{2} - \frac{x^4}{4})]_0^1 = \frac{1}{2\pi} = 0.1592. \end{aligned}$$

$$\begin{aligned}\mathrm{cov}[X^2Y] &= \int_0^1 \int_0^{\sqrt{1-x^2}} \frac{4}{\pi} x^2 y \, dydx = \int_0^1 \frac{4}{2\pi} x^2(1-x^2)dx \\ &= [\frac{4}{2\pi}(\frac{x^3}{3} - \frac{x^5}{5})]_0^1 = \frac{4}{15\pi} = 0.0849.\end{aligned}$$

(a) Using the above quantities, we find that

$$L[X|Y] = \frac{\mathrm{cov}(XY)}{\mathrm{var}(Y)}(Y - E[Y]) + E[X] = \frac{0.1592}{0.0699}(Y - 0.4244) + 0.4244 = 2.2775(Y - 0.4244) + 0.4244.$$

(b) Similarly,

$$L[X^2|Y] = \frac{\mathrm{cov}(X^2Y)}{\mathrm{var}(Y)}(Y - E[Y]) + E[X^2] = \frac{0.0849}{0.0699}(Y - 0.4244) + 0.25 = 1.2146(Y - 0.4244) + 0.25.$$

Example 9.6.7. *Let X, Y be independent random variables uniformly distributed in $[0, 1]$. Calculate $L[Y^2 \mid 2X + Y]$.*

One has

$$\begin{aligned}L[Y^2 \mid 2X + Y] &= E(X^2) + \frac{E(Y^2(2X + Y)) - E(Y^2)E(2X + Y)}{\mathrm{var}(2X + Y)}(2X + Y - E(2X + Y)) \\ &= \frac{1}{3} + \frac{1/3 + 1/4 - (1/3)(3/2)}{4(1/3 - 1/4) + (1/3 - 1/4)}(2X + Y - 3/2).\end{aligned}$$

Example 9.6.8. *Let $\{X_n, n \geq 1\}$ be independent $N(0, 1)$ random variables. Define $Y_{n+1} = aY_n + (1-a)X_{n+1}$ for $n \geq 0$ where Y_0 is a $N(0, \sigma^2)$ random variable independent of $\{X_n, n \geq 0\}$. Calculate*

$$E[Y_{n+m}|Y_0, Y_1, \ldots, Y_n]$$

for $m, n \geq 0$.

Hint: *First argue that observing $\{Y_0, Y_1, \ldots, Y_n\}$ is the same as observing $\{Y_0, X_1, \ldots, X_n\}$. Second, get an expression for Y_{n+m} in terms of $Y_0, X_1, \ldots, X_{n+m}$. Finally, use the independence of the basic random variables.*

One has

$$\begin{aligned} Y_{n+1} &= aY_n + (1-a)X_{n+1}; \\ Y_{n+2} &= aY_{n+1} + (1-a)X_{n+2} = a^2Y_n + (1-a)X_{n+2} + (1-a)^2X_{n+1}; \\ &\cdots \\ Y_{n+m} &= a^mY_n + (1-a)X_{n+m} + (1-a)^2X_{n+m-1} + \cdots + (1-a)^mX_{n+1}. \end{aligned}$$

Hence,

$$E[Y_{n+m} \mid Y_0, Y_1, \ldots, Y_n] = a^mY_n.$$

Example 9.6.9. *Given* $\{\Theta = \theta\}$*, the random variables* $\{X_n, n \geq 1\}$ *are i.i.d.* $U[0,\theta]$*. Assume that* θ *is exponentially distributed with rate* λ*.*

a. Find the MAP $\hat{\theta}_n$ *of* Θ *given* $\{X_1, \ldots, X_n\}$*.*

b. Calculate $E(|\Theta - \hat{\theta}_n|)$*.*

One finds that

$$f_{X|\Theta}[x \mid \theta] f_\Theta(\theta) = \frac{1}{\theta^n} 1\{x_k \leq \theta, k = 1, \ldots, n\} \lambda e^{-\lambda\theta}.$$

Hence,

$$\hat{\theta}_n = max\{X_1, \ldots, X_n\}.$$

Consequently, by symmetry,

$$E[\Theta - \hat{\theta}_n|\theta] = \frac{1}{n+1}\theta.$$

Finally,

$$E(|\Theta - \hat{\theta}_n|) = E(E[\Theta - \hat{\theta}_n|\theta])) = \frac{1}{\lambda(n+1)}.$$

A few words about the symmetry argument. Consider a circle with a circumference length equal to 1. Place $n+1$ point independently and uniformly on that circumference.

By symmetry, the average distance between two points is $1/(n+1)$. Pick any one point and open the circle at that point, calling one end 0 and the other end 1. The other n points are distributed independently and uniformly on $[0,1]$. So, the average distance between 1 and the closest point in $1/(n+1)$. Of course, we could do a direct calculation.

Example 9.6.10. *Let X, Y be independent random variables uniformly distributed in $[0,1]$. a. Calculate $E[X|X^2+Y^2]$. b. Calculate $L[X|X^2+Y^2]$.*

a. Once again, we draw a unit square. Let $R^2 = X^2 + Y^2$. Given $\{R = r\}$, the pair (X, Y) is uniformly distributed on the intersection of the circumference of the circle with radius r centered at the origin and the unit square.

If $r < 1$, then this intersection is the quarter of the circumference and we must calculate $E(r\cos(\theta))$ where θ is uniform in $[0, \pi/2]$. We find

$$E[X|R=r] = E(r\cos(\theta)) = \int_0^{\pi/2} r\cos(x)\frac{2}{\pi}dx = [r\sin(x)\frac{2}{\pi}]_0^{\pi/2} = \frac{2r}{\pi}.$$

If $r > 1$, then θ is uniformly distributed in $[\theta_1, \theta_2]$ where $\cos(\theta_1) = 1/r$ and $\sin(\theta_2) = 1/r$. Hence,

$$\begin{aligned} E[X|R=r] &= E(r\cos(\theta)) = \int_{\theta_1}^{\theta_2} r\cos(x)\frac{1}{\theta_2-\theta_1}dx \\ &= [r\sin(x)\frac{1}{\theta_2-\theta_1}]_{\theta_1}^{\theta_2} = r\frac{\sin(\theta_2)-\sin(\theta_1)}{\theta_2-\theta_1} \\ &= \frac{1-\sqrt{r^2-1}}{\sin^{-1}(1/r)-\cos^{-1}(1/r)}. \end{aligned}$$

b. Let $V = X^2 + Y^2$. Then,

$$L[X|V] = E(X) + \frac{\mathrm{cov}(X,V)}{\mathrm{var(V)}}(V - E(V)).$$

Now, $E(X) = 1/2$. Also,

$$\begin{aligned}\text{cov}(X,V) &= E(XV) - E(X)E(V) = E(X^3 + XY^2) - (1/2)E(X^2 + Y^2) \\ &= \frac{1}{4} + \frac{1}{2} \times \frac{1}{3} - \frac{1}{2} \times (\frac{1}{3} + \frac{1}{3}) \\ &= \frac{1}{12}.\end{aligned}$$

In addition,

$$\text{var(V)} = E(V^2) - (E(V))^2 = E(X^4 + 2X^2Y^2 + Y^4) - (2/3)^2 = \frac{8}{45}$$

and

$$E(V) = \frac{2}{3}.$$

Consequently,

$$L[X|V] = \frac{1}{2} + \frac{1/12}{8/45}(V - \frac{2}{3}) = \frac{3}{16} + \frac{15}{32}(X^2 + Y^2).$$

Example 9.6.11. *Suppose we observe*

$$Y_i = s_i X + W_i, i = 1, \ldots, n$$

where $W_1, \ldots, W_n$ are independent $N(0,1)$ and X takes the values $+1$ and -1 with equal probabilities and is independent of the W_i. The discrete-time signal $s_i (i = 1, \ldots, n)$ is a known deterministic signal. Determine the MAP rule for deciding on X based on $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$.

Since the distribution of X is uniform, $\hat{X} = MAP[X \mid \boldsymbol{Y}] = MLE[X \mid \boldsymbol{Y}]$. That is, $\hat{X}$ is 1 if $L(\boldsymbol{Y}) > 1$ and is -1 otherwise where

$$L(\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|X}[\boldsymbol{y}| + 1]}{f_{\boldsymbol{Y}|X}[\boldsymbol{y}| - 1]}.$$

Now,

$$f_{\boldsymbol{Y}|X}[\boldsymbol{y}|X = x] = \Pi_{i=1}^{N}[\frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(y_i - s_i x)^2\}].$$

Accordingly,

$$\begin{aligned} L(\boldsymbol{y}) &= \frac{\exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - s_i)^2\}}{\exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i + s_i)^2\}} \\ &= \exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - s_i)^2 + \frac{1}{2}\sum_{i=1}^{n}(y_i + s_i)^2\}\} \\ &= \exp\{2\sum_{i=1}^{n} s_i y_i\} = \exp\{2\boldsymbol{y} \centerdot \boldsymbol{s}\} \end{aligned}$$

where

$$\boldsymbol{y} \centerdot \boldsymbol{s} := \sum_{i=1}^{n} s_i y_i.$$

Consequently,

$$\hat{X} = \begin{cases} +1, \text{ if } \boldsymbol{y} \centerdot \boldsymbol{s} > 0; \\ -1, \text{ if } \boldsymbol{y} \centerdot \boldsymbol{s} \leq 0. \end{cases}$$

Notice that $\boldsymbol{y} \centerdot \boldsymbol{s}$ is a sufficient statistic for estimating X given $\boldsymbol{Y}$.

Example 9.6.12. *a. Suppose*

$$Y = gX + W$$

where X and W are independent zero-mean Gaussian random variables with respective variances σ_X^2 and σ_W^2. Find $L[X \mid Y]$ and the resulting mean square error.

b. Suppose now we have two observations:

$$Y_i = g_i X + W_i, i = 1, 2$$

where the W_i are independent and Gaussian with variance σ_W^2 and independent of X. Find $L[X \mid Y_1, Y_2]$.

a. We know that

$$\hat{X} := L[X \mid Y] = E(X) + \frac{\text{cov}(X, Y)}{\text{var}(Y)}(Y - E(Y)) = \frac{\text{cov}(X, Y)}{\text{var}(Y)}Y.$$

Now, because X and Y are zero-mean,

$$\text{cov}(X, Y) = E(XY) = E(X(gX + W)) = gE(X^2) = g\sigma_X^2$$

and

$$\mathrm{var}(Y) = \mathrm{var}(gX + W) = \mathrm{var}(gX) + \mathrm{var}(W) = g^2\mathrm{var}(X) + \mathrm{var}(W) = g^2\sigma_X^2 + \sigma_W^2.$$

Hence,

$$L[X \mid Y] = \frac{g\sigma_X^2}{g^2\sigma_X^2 + \sigma_W^2} Y.$$

The resulting mean square error is

$$\begin{aligned} E((\hat{X} - X)^2) &= E((\frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)}Y - X)^2) \\ &= \frac{\mathrm{cov}^2(X,Y)}{\mathrm{var}^2(Y)}\mathrm{var}(Y) - 2\frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)}\mathrm{cov}(X,Y) + \mathrm{var}(X) \\ &= \mathrm{var}(X) - \frac{\mathrm{cov}^2(X,Y)}{\mathrm{var}(Y)}. \end{aligned}$$

Hence,

$$E((\hat{X} - X)^2) = \sigma_X^2 - \frac{g^2\sigma_X^4}{g^2\sigma_X^2 + \sigma_W^2} = \frac{\sigma_X^2\sigma_W^2}{g^2\sigma_X^2 + \sigma_W^2}.$$

b.Let $\boldsymbol{Y} = (Y_1, Y_2)^T$. Then

$$\hat{X} = L[X \mid \boldsymbol{Y}] = \Sigma_{X\boldsymbol{Y}}\Sigma_{\boldsymbol{Y}}^{-1}\boldsymbol{Y}$$

where

$$\Sigma_{X\boldsymbol{Y}} = E(X\boldsymbol{Y}^T) = E(X(Y_1, Y_2)) = \sigma_X^2(g_1, g_2)$$

and

$$\Sigma_{\boldsymbol{Y}} = E(\boldsymbol{Y}\boldsymbol{Y}^T) = E\begin{bmatrix} Y_1^2 & Y_1Y_2 \\ Y_2Y_1 & Y_2^2 \end{bmatrix} = \begin{bmatrix} g_1^2\sigma_X^2 + \sigma_W^2 & g_1g_2\sigma_X^2 \\ g_1g_2\sigma_X^2 & g_2^2\sigma_X^2 + \sigma_W^2 \end{bmatrix}.$$

Hence,

$$\begin{aligned} \hat{X} &= \sigma_X^2(g_1, g_2)\begin{bmatrix} g_1^2\sigma_X^2 + \sigma_W^2 & g_1g_2\sigma_X^2 \\ g_1g_2\sigma_X^2 & g_2^2\sigma_X^2 + \sigma_W^2 \end{bmatrix}^{-1}\boldsymbol{Y} \\ &= \sigma_X^2(g_1, g_2)\frac{1}{g_1^2\sigma_X^2\sigma_W^2 + g_2^2\sigma_X^2\sigma_W^2 + \sigma_W^4}\begin{bmatrix} g_2^2\sigma_X^2 + \sigma_W^2 & -g_1g_2\sigma_X^2 \\ -g_1g_2\sigma_X^2 & g_1^2\sigma_X^2 + \sigma_W^2 \end{bmatrix}\boldsymbol{Y} \\ &= \frac{\sigma_X^2}{g_1^2\sigma_X^2 + g_2^2\sigma_X^2 + \sigma_W^2}(g_1, g_2)\boldsymbol{Y}. \end{aligned}$$

Thus,

$$\hat{X} = \frac{\sigma_X^2}{g_1^2\sigma_X^2 + g_2^2\sigma_X^2 + \sigma_W^2}(g_1Y_1 + g_2Y_2).$$

Example 9.6.13. *Suppose we observe* $Y_i = 3X_i + W_i$ *where* W_1, W_2 *are independent* $N(0,1)$ *and* $\boldsymbol{X} = (X_1, X_2)$ *is independent of* (W_1, W_2) *and has the following pmf:*

$$P(\boldsymbol{X} = \boldsymbol{x}^1 := (1,0)) = \frac{1}{2}, P(\boldsymbol{X} = \boldsymbol{x}^2 := (0,1)) = \frac{1}{6},$$
$$P(\boldsymbol{X} = \boldsymbol{x}^3 := (-1,0)) = \frac{1}{12}, \text{ and } P(\boldsymbol{X} = \boldsymbol{x}^4 := (0,-1)) = \frac{1}{4}.$$

Find $MAP[(X_1, X_2) \mid (Y_1, Y_2)]$.

We know that

$$MAP[\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}] = \arg\max P(\boldsymbol{X} = \boldsymbol{x}) f_{\boldsymbol{Y}|\boldsymbol{X}}[\boldsymbol{y} \mid \boldsymbol{x}].$$

Now,

$$f_{\boldsymbol{Y}|\boldsymbol{X}}[\boldsymbol{y} \mid \boldsymbol{x}] = \frac{1}{2\pi}\exp\{-\frac{1}{2}[(y_1 - 3x_1)^2 + (y_2 - 3x_2)^2]\}.$$

Hence,

$$MAP[\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}] = \arg\min\{||\boldsymbol{y} - 3\boldsymbol{x}||^2 - 2\ln(P(\boldsymbol{X} = \boldsymbol{x})\} =: \arg\min\, c(\boldsymbol{y}, \boldsymbol{x}).$$

Note that

$$c(\boldsymbol{y}, \boldsymbol{x}^i) < c(\boldsymbol{y}, \boldsymbol{x}^j) \Leftrightarrow \boldsymbol{y} \centerdot (\boldsymbol{x}^i - \boldsymbol{x}^j) > \alpha_i - \alpha_j$$

where, for $\boldsymbol{y}, \boldsymbol{z} \in \Re^2$,

$$\boldsymbol{y} \centerdot \boldsymbol{z} := y_1z_1 + y_2z_2$$

and

$$\alpha_i = \frac{3}{2}||\boldsymbol{x}^i||^2 - \frac{1}{3}\ln(P(\boldsymbol{X} = \boldsymbol{x}^i), \text{ for } i = 1,2,3,4.$$

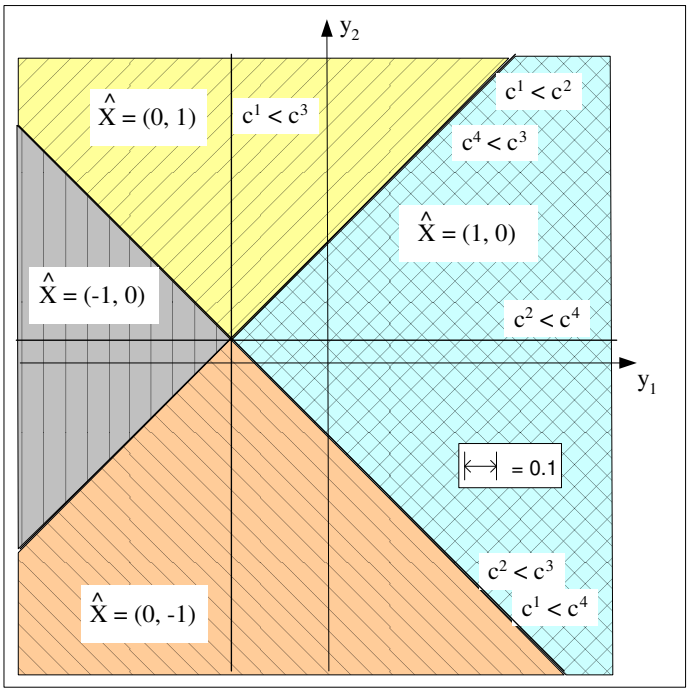


Figure 9.2: The MAP regions for Example 9.6.13

Thus, for every $i, j \in \{1, 2, 3, 4\}$ with $i \neq j$, there is a line that separates the points $\boldsymbol{y}$ where $c(\boldsymbol{y}, \boldsymbol{x}^i) < c(\boldsymbol{y}, \boldsymbol{x}^j)$ from those where $c(\boldsymbol{y}, \boldsymbol{x}^i) > c(\boldsymbol{y}, \boldsymbol{x}^j)$. These lines are the following:

$$c^1 < c^2 \Leftrightarrow y_2 < y_1 + 0.3662$$

$$c^1 < c^3 \Leftrightarrow y_1 > -0.2987$$

$$c^1 < c^4 \Leftrightarrow y_2 > -y_1 - 0.231$$

$$c^2 < c^3 \Leftrightarrow y_2 > -y_1 - 0.231$$

$$c^2 < c^4 \Leftrightarrow y_2 > 0.0676$$

$$c^3 < c^4 \Leftrightarrow y_2 < y_1 + 0.3662$$

We draw these six lines in Figure 9.2.

The figure allows us to identify the regions of $\Re^2$ that correspond to each of the values

of the MAP. We can summarize the results as follows:

$$\hat{X} = \begin{cases} (1,0), \text{ if } -y_1 - 0.231 < y_2 < y_1 + 0.3662; \\ (-1,0), \text{ if } y_1 + 0.3662 < y_2 < -y_1 - 0.231; \\ (0,1), \text{ if } -y_2 - 0.231 < y_1 < y_2 - 0.3662; \\ (0,-1), \text{ if } y_2 - 0.3662 < y_1 < -y_2 - 0.231. \end{cases}$$

Chapter 10

Limits of Random Variables

Random behaviors often become more tractable as some parameter of the system, such as the size or speed, increases. This increased tractability comes from statistical regularity. When many sources of uncertainty combine their effects, their individual fluctuations may compensate one another and the combined result may become more predictable. For instance, if you flip a single coin, the outcome is very unpredictable. However, if you flip a large number of them, the proportion of heads is less variable. To make precise sense of these limiting behaviors one needs to define the limit of random variables.

In this chapter we explain what we mean by $X_n \to X$ as $n \to \infty$. Mathematically, X_n and X are functions. Thus, it takes some care to define the convergence of functions. For the same reason, the meaning of "X_n is close to X" requires some careful definition.

We explain that there are a number of different notions of convergence of random variables, thus a number of ways of defining that X_n approaches X. The differences between these notions may seem subtle at first, however they are significant and correspond to very different types of approximation. We use examples to highlight the differences. We start with convergence in distribution, then explain how to use transform methods to prove that type of convergence. We then discuss almost sure convergence and convergence in probability and in L^2. We conclude with a discussion of the relations between these difference forms of convergence and we comment on the convergence of expectation.

Looking ahead, the strong law of large numbers is an almost sure convergence result; the weak law is a convergence in probability result; the central limit theorem is about convergence in distribution. It is important to appreciate the meaning of these types of convergence in order to understand these important results.

10.1 Convergence in Distribution

Intuitively we can say that the random variables X and Y are similar if their cdf are about the same, i.e., if $P(X \leq x) \approx P(Y \leq x)$ for all $x \in \Re$. For example, we could say that X is almost a standard Gaussian random variable if this approximation holds when $Y = N(0,1)$. It is in this sense that one can show that many random variables that occur in physical systems are almost Gaussian or Poisson.

Correspondingly, we define convergence in distribution as follows.

Definition 10.1.1. Convergence in Distribution

The random variables X_n are said to *converge in distribution* to the random variable X, and we write $X_n \to_D X$, if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \text{ for all } x \in \Re \text{ such that } F_X(x) = F_X(x-). \tag{10.1.1}$$

The restriction $F_X(x) = F_X(x-)$ requires some elaboration. Consider the random variables $X_n = 1 + 1/n$ for $n \geq 1$ and the random variable $X = 1$. Intuitively we want to be able to say that X_n approaches X in distribution. You see that $F_{X_n}(x) \to F_X(x)$ for all $x \neq 1$. However, $F_{X_n}(1) = 0$ for all n and $F_X(1) = 1$. The restriction in (10.1.1) takes care of such discontinuity.

With this definition, you can check that if X_n and X are discrete with $P(X_n = x_{m,n}) = p_{m,n}$ for $m, n \geq 1$ and $P(X = x_m) = p_m$ for $m \geq 1$, then $X_n \to_D X$ if $\lim_{n\to\infty} p_{m,n} = p_m$ and $\lim_{n\to\infty} x_{m,n} = x_m$ for $n \geq 1$. In other words, the definition is conform to our intuition: the possible values get close and so do their probabilities.

If the random variable X is continuous and if $X_n \to_D X$, then

$$\lim_{n\to\infty} P(X_n \in (a,b)) = P(X \in (a,b)), \forall a < b \in \Re.$$

10.2 Transforms

Transform methods are convenient to show convergence in distribution. We give an example here. (See the Central Limit Theorem in Section 11.3 for another example.) For $n \geq 1$ and $p > 0$, let $X(n,p)$ be binomial with the parameters (n,p). That is,

$$P(X(n,p) = m) = \binom{n}{m} p^m (1-p)^{n-m}, m = 0, 1, \ldots, n.$$

We want to show that as $p \downarrow 0$ and $np \to \lambda$, one has $X(n,p) \to_D X$ where X is Poisson with mean λ. We do this by showing that $E(z^{X(n,p)}) \to E(z^X)$ for all complex numbers z. These expected values are the z-transforms of the probability mass functions of the random variables. One then invokes a theorem that says that if the z-transforms converge, then so do the probability mass functions. In these notes, we do the calculation and we accept the theorem. Note that $X(n,p)$ is the sum of n i.i.d. random variables that are 1 with probability p and zero otherwise. If we designate one such generic random variable by $V(p)$, we have

$$E(z^{X(n,p)}) = (E(z^{V(p)}))^n = ((1-p) + pz)^n \to (1 + \lambda(z-1)/n)^n \to \exp\{\lambda(z-1)\},$$

by (**??**). Also,

$$E(z^X) = \sum_n z^n \frac{(\lambda)^n}{n!} \exp\{-\lambda\} = \exp\{\lambda(z-1)\}.$$

10.3 Almost Sure Convergence

A strong form of convergence is when the real numbers $X_n(\omega)$ approach the real number $X(\omega)$ for all ω. In that case we say that the random variables X_n converge to the random variable X almost surely. Formally, we have the following definition.

Definition 10.3.1. Almost Sure Convergence

Let $X_n, n \geq 1$ and X be random variables defined on the same probability space $\{\Omega, \mathcal{F}, P\}$. We say that X_n *converge almost surely* to X, and we write $X_n \to_{\text{a.s.}} X$, if

$$\lim_{n\to\infty} X_n(\omega) = X(\omega), \text{ for almost all } \omega. \tag{10.3.1}$$

The expression "for almost all ω" means for all ω except possibly for a set of ω with probability zero.

Anticipating future results, look at the fraction X_n of heads as you flip a fair coin n times. You expect X_n to approach $1/2$ for every possible realization of this random experiment. However, the outcome where you always gets heads is such that the fraction of heads does not go to $1/2$. In fact, there are many conceivable sequences of coin flips where X_n does not approach $1/2$. However, all these sequences together have probability zero. This example shows that it would be silly to insist that $X_n(\omega) \to X(\omega)$ for all ω. This condition would be satisfied only in trivial models.

Almost sure convergence is a very strong result. It states that the sequence of real numbers $X_n(\omega)$ approaches the real number $X(\omega)$ for every possible outcome of the random experiment. By possible, we mean here "outside of a set of probability zero." Thus, as you perform the random experiment, you find that the numbers $X_n(\omega)$ approach the number $X(\omega)$.

The following observation is very useful in applications. Assume that $X_n \to_{\text{a.s.}} X$ and that $g : \Re \to \Re$ is a continuous function. Note that $X_n(\omega) \to X(\omega)$ implies by continuity

that $g(X_n(\omega)) \to g(X(\omega))$. Accordingly, we find that the random variables $g(X_n)$ converge almost surely to the random variable $g(X)$.

10.3.1 Example

A common technique to prove almost sure convergence is to use the Borel-Cantelli Lemma 2.7.10. We provide an illustration of this technique.

Let $X_n, n \geq 1$ be zero-mean random variables defined on $\{\Omega, \mathcal{F}, P\}$ with $\text{var}(X_n) \leq 1/n^2$. We claim that $X_n \to_{\text{a.s.}} 0$.

To prove that fact, fix $\epsilon > 0$. Note that, by Chebyshev's inequality (4.8.1),

$$P(|X_n| \geq \epsilon) \leq \frac{\text{var}(X_n)}{\epsilon^2} \leq \frac{1}{n^2\epsilon^2}, \text{ for } n \geq 1.$$

Consequently,

$$\sum_{n=1}^{\infty} P(|X_n| \geq \epsilon) \leq \sum_{n=1}^{\infty} \frac{1}{n^2\epsilon^2} < \infty.$$

Using the Borel-Cantelli Lemma 2.7.10, we conclude that

$$P(|X_n| \geq \epsilon \text{ for infinitely many values of } n) = 0.$$

Accordingly, there must be some finite n_0 so that

$$|X_n| < \epsilon \text{ for } n \geq n_0.$$

This almost shows that $X_n \to_{\text{a.s.}} 0$. Now, to be technically precise, we should show that these is some set A of probability 0 so that if $\omega \notin A$, then for all $\epsilon > 0$, there is some n_0 such that $|X_n(\omega)| \leq \epsilon$ whenever $n \geq 0$. What is missing in our derivation is that the set A may depend on ϵ. To fix this problem, let A_n be the set A that corresponds to $\epsilon = 1/n$ and choose $A = \cup_n A_n$. Then $P(A) \leq \sum_n P(A_n) = 0$ and that set has the required property.

10.4 Convergence In Probability

It may be that X_n and X are more and more likely to be close as n increases. In that case, we say that the random variables X_n converge in probability to the random variable X. Formally, we have the following definition.

Definition 10.4.1. Convergence in Probability

Let $X_n, n \geq 1$ and X be random variables defined on the same probability space $\{\Omega, \mathcal{F}, P\}$. We say that the X_n *converge in probability to* X, and we write $X_n \to_P X$, if

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0, \text{ for all } \epsilon > 0. \tag{10.4.1}$$

It takes some reflection to appreciate the difference between convergence in probability and almost sure convergence. Let us try to clarify the difference. If $X_n \to_P X$, there is no guarantee that $X_n(\omega)$ converges to $X(\omega)$.

Here is an example that illustrates the point. Let $\{\Omega, \mathcal{F}, P\}$ be the interval $[0, 1)$ with the uniform distribution. The random variables $\{X_1, X_2, \ldots\}$ are equal to zero everywhere except that they are equal to one on the intervals $[0, 1)$, $[0, 1/2)$, $[1/2, 1)$, $[0, 1/4)$, $[1/4, 2/4)$, $[2/4, 3/4)$, $[3/4, 1)$, $[0, 1/8)$, $[1/8, 2/8)$, $[2/8, 3/8)$, $[3/8, 4/8)$, $[4/8, 5/8)$, ..., respectively. Thus, $X_1 = 1, X_2(\omega) = 1\{\omega \in [0, 1/2)\}, X_9(\omega) = 1\{\omega \in [1/8, 2/8)\}$, and so on. From this definition we see that $P(X_n \neq 0) \to 0$, so that $X_n \to_P 0$. Moreover, for every $\omega < 1$ the sequence $\{X_n(\omega), n \geq 1\}$ contains infinitely many ones. Accordingly, X_n does not converge to 0 for any ω. That is, $P(X_n \to 0) = 0$.

Thus it is possible for the probability of $A_n := \{|X_n - X| > \epsilon\}$ to go to zero but for those sets to keep on "scanning" Ω and, consequently, for $|X_n(\omega) - X(\omega)|$ to be larger than ϵ for infinitely many values of n. In that case, $X_n \to_P X$ but $X_n \nrightarrow_{\text{a.s.}} X$.

Imagine that you simulate the sequence $\{X_n, n \geq 1\}$ and X. If the sequence that you observe from your simulation run is such that $X_n(\omega) \nrightarrow X(\omega)$, then you can conclude that

$X_n \nrightarrow_{\text{a.s.}} X$. However, you cannot conclude that $X_n \nrightarrow_P X$.

10.5 Convergence in L^2

Another way to say that X and Y are close to each other is when $E(|X-Y|^2)$ is small. This corresponds to the meaning of convergence in L^2. Specifically, we have the following definition.

Definition 10.5.1. Convergence in L^2

Let $X_n, n \geq 1$ and X be random variables defined on a common probability space $\{\Omega, \mathcal{F}, P\}$. We say that X_n converges in L^2 to X, and we write $X_n \rightarrow_{L^2} X$ if

$$E|X_n - X|^2 \rightarrow 0 \text{ as } n \rightarrow \infty. \tag{10.5.1}$$

This definition has an intuitive meaning: the error becomes small in the mean squares sense. If you recall our discussion of estimators and the interpretation of MMSE and LLSE in a space with the metric based on the mean squared error, then convergence in L^2 is precisely convergence in that metric. Not surprisingly, this notion of convergence is well matched to the study of properties of estimators.

10.6 Relationships

All these convergence notions make sense. How do they relate to one another? Figure 10.1 provided a summary.

The discussion below of these implications should help you clarify your understanding of the different forms of convergence.

We explained above that convergence in probability does not imply almost sure convergence. We now prove that the opposite implication is true. That is, we prove the following theorem.

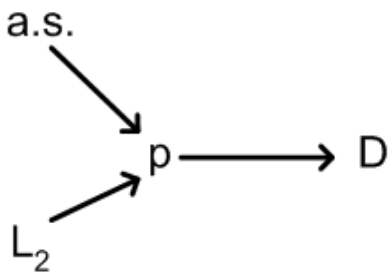


Figure 10.1: Relationships among convergence properties

Theorem 10.6.1. *Let $X_n, n \geq 1$ and X be random variable defined on the same probability space $\{\Omega, \mathcal{F}, P\}$. If $X_n \to_{a.s.} X$, then $X_n \to_P X$.*

Proof:

Assume that $X_n \to_{\text{a.s.}} X$. We show that for all $\epsilon > 0$, $P(A_n) \to 0$ where $A_n := \{\omega \mid |X_n(\omega) - X(\omega)| > \epsilon\}$.

To do this, we define $B_n = \cup_{m=n}^{\infty} A_m$. Note that

$$B_n \downarrow B := \cap_{n=1}^{\infty} B_n = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_n = \{\omega \mid \omega \in A_n \text{ for infinitely many values of } n\}.$$

Thus, $B \subset \{\omega \mid X_n(\omega) \nrightarrow X(\omega)\}$ and we conclude that $P(B) = 0$, so that $P(B_n) \downarrow 0$. Finally, since $A_n \subset B_n$ we find that $P(A_n) \to 0$, as we wanted to show. □

Recapitulating, the idea of the proof is that if $X_n(\omega) \to X(\omega)$, then this ω is not in A_n for all n large enough. Therefore, the only ω's that are in B must be those where X_n does not converge to X, and these ω's have probability zero. The consideration of B_n instead of A_n is needed because the sets A_n may not decrease even though they are contained in the sets B_n whose probability goes down to zero.

We now prove another implication in Figure 10.1.

Theorem 10.6.2. *Let $X_n, n \geq 1$ and X be random variable defined on the same probability space $\{\Omega, \mathcal{F}, P\}$. If $X_n \to_{L^2} X$, then $X_n \to_P X$.*

Proof:

We use Chebychev's inequality (4.8.1):

$$P(|X_n - X| > \epsilon) \leq \frac{E(|X_n - X|^2)}{\epsilon^2}.$$

This inequality shows that if $X_n \to_{L^2} X$, then $P|X_n - X| > \epsilon) \to 0$ for any given $\epsilon > 0$, so that $X_n \to_P X$. □

It is useful to have a simple example in mind that shows that convergence in probability does not imply convergence in L^2. Such an example is as follows. Let $\{\Omega, \mathcal{F}, P\}$ be $[0, 1]$ with the uniform probability. Define $X_n(\omega) = n \times 1\{\omega \leq 1/n\}$. You can see that $X_n \to_{\text{a.s.}} 0$. However, $E(X_n^2) = n \nrightarrow 0$, so that $X_n \nrightarrow_{L^2} 0$.

This example also shows that $X_n \to_{\text{a.s.}} X$ does not necessarily imply that $E(X_n) \to E(X)$. Thus, in general,

$$\lim_{n\to\infty} E(X_n) \neq E(\lim_{n\to\infty} X_n). \tag{10.6.1}$$

We explain in the next section that some simple sufficient condition guarantee the equality in the expression above.

We turn our attention to the last implication in Figure 10.1. We show the following result.

Theorem 10.6.3. *Assume that $X_n \to_P X$. Then $X_n \to_D X$.*

Proof:

The idea of the proof is that if $|X_n - X| \leq \epsilon$, then $P(X \leq x - \epsilon) \leq P(X_n \leq x) \leq P(X \leq x + \epsilon)$. If ϵ is small enough, both side of this inequality are close to $P(X \leq x)$, so that we get $P(X_n \leq x) \approx P(X \leq x)$. Now, if n is large, then $P(|X_n - X| \leq \epsilon) \approx 1$ so that this approximation is off only with a small probability.

We proceed with the formal derivation of this idea.

Fix $\alpha > 0$. First, find $\epsilon > 0$ so that $P(X \in [x-\epsilon, x+\epsilon]) \leq \alpha/2$. This is possible because $F_X(x-) = F_X(x)$. Second, find n_0 so that if $n \geq n_0$, then $P(|X_n - X| \geq \epsilon) \leq \alpha/2$. This is possible because $X_n \to_P X$.

Observe that

$$P(X_n \leq x \text{ and } |X_n - X| \leq \epsilon) \leq P(X \leq x + \epsilon). \tag{10.6.2}$$

Also,

$$P(X \leq x - \epsilon \text{ and } |X_n - X| \leq \epsilon) \leq P(X_n \leq x). \tag{10.6.3}$$

Now, if $P(A) \geq 1 - \delta$, then $P(A \cap B) \geq P(B) - \delta$. Indeed, $P(A \cap B) = 1 - P(A^c \cup B^c) \geq 1 - P(A^c) - P(B^c) = P(B) - P(A^c) \geq P(B) - \delta$. Consequently, (10.6.2) implies that, for $n \geq n_0$,

$$P(X_n \leq x) - \alpha/2 \leq P(X \leq x + \epsilon) \leq F_X(x) + \alpha/2,$$

so that

$$P(X_n \leq x) \leq F_X(x) + \alpha.$$

Similarly, (10.6.3) implies

$$(F_X(x) - \alpha/2) - \alpha/2 \leq P(X \leq x - \epsilon) - \alpha/2 \leq P(X_n \leq x),$$

so that

$$F_X(x) - \alpha \leq P(X_n \leq x).$$

These two inequalities imply that $|F_X(x) - P(X_n \leq x)| \leq \alpha$ for $n \geq n_0$. Hence $X_n \to_D X$.

□

10.7 Convergence of Expectation

Assume $X_n \to_{\text{as}} X$. We gave an example that show that it is generally not the case that $E(X_n) \to E(X)$. (See 10.6.1).) However, two simple sets of sufficient conditions are known.

Theorem 10.7.1. *Lebesgue*

a. Assume $X_n \to_{as} X$ *and* $0 \leq X_n \leq X_{n+1}$ *for all* n. *Then* $E(X_n) \to E(X)$.

b. Assume $X_n \to_{as} X$ *and* $|X_n| \leq Y$ *for all* n *with* $E(Y) < \infty$. *Then* $E(X_n) \to E(X)$.

We refer the reader to probability textbooks for a proof of this result that we use in examples below.

Chapter 11

Law of Large Numbers & Central Limit Theorem

We started the course by saying that, in the long term, about half of the flips of a fair coin yield tail. This is our intuitive understanding of probability. The law of large number explains that our model of uncertain events conforms to that property. The central limit theorem tells us how fast this convergence happens. We discuss these results in this chapter.

11.1 Weak Law of Large Numbers

We first prove an easy convergence result.

Theorem 11.1.1. *Weak Law of Large Numbers*

Let $\{X_n, n \geq 1\}$ *be i.i.d. random variables with mean* μ *and finite variance* σ^2*. Let also* $Y_n = (X_1 + \cdots + X_n)/n$ *be the sample mean of the first* n *random variables Then*

$$Y_n \rightarrow_P \mu \text{ as } n \rightarrow \infty.$$

Proof:

We use Chebychev's inequality (4.8.1). We have

$$P(|\frac{X_1 + \cdots + X_n}{n} - \mu| \geq \epsilon) \leq \frac{1}{\epsilon^2} E((\frac{X_1 + \cdots + X_n}{n} - \mu)^2).$$

Now, $E(\frac{X_1+\cdots+X_n}{n} - \mu) = 0$, so that

$$E((\frac{X_1 + \cdots + X_n}{n} - \mu)^2) = \text{var}(\frac{X_1 + \cdots + X_n}{n}).$$

We know that the variance of a sum of independent random variables is the sum of their variances (see Theorem 5.3.1). Hence,

$$\text{var}(\frac{X_1 + \cdots + X_n}{n}) = \frac{1}{n^2}\text{var}(X_1 + \cdots + X_n) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}.$$

Combining these results, we find that

$$P(|\frac{X_1 + \cdots + X_n}{n} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty.$$

Hence,

$$\frac{X_1 + \cdots + X_n}{n} \to_P 0 \text{ as } n \to \infty.$$

11.2 Strong Law of Large Numbers

The following result is remarkable.

Theorem 11.2.1. *Strong Law of Large Numbers Let $\{X_n, n \geq 1\}$ be i.i.d. random variables with mean μ and finite variance σ^2. Let also $Y_n = (X_1 + \cdots + X_n)/n$ be the sample mean of the first n random variables Then*

$$Y_n \to_{a.s.} \mu \text{ as } n \to \infty.$$

This result is stronger than the weak law:

- Mathematically because almost sure convergence implies convergence in probability;
- In applications because it states that Y_n becomes a good estimate of μ as n increases (always, you cannot be unlucky!).

The proof of this result is a bit too technical for this course.

11.3 Central Limit Theorem

The next result estimates the speed of convergence of the sample mean to the expected value.

Theorem 11.3.1. *CLT Let $\{X_n, n \geq 1\}$ be i.i.d. random variables with mean μ and finite variance σ^2. Then*

$$Z_n := (Y_n \mu)\sqrt{n} \rightarrow_D N(0, \sigma^2) \text{ as } n \rightarrow \infty.$$

This result says (roughly) that the error $Y_n\mu$ is of order $\sigma/\sqrt{n}$ when n is large. Thus, if one makes four times more observations, the error on the mean estimate is reduced by a factor of 2.

Equivalently,

$$(Y_n \mu)\frac{\sqrt{n}}{\sigma} \rightarrow_D N(0,1) \text{ as } n \rightarrow \infty.$$

Proof:

Here is a rough sketch of the proof. We show that $E(e^{iuZ_n}) \rightarrow e^{-u^2/2}$ and we then invoke a theorem that says that if the Fourier transform of f_{Z_n} converge to that of f_V, then $Z_n \rightarrow_D V$.

We find that

$$\begin{aligned} E(e^{iuZ_n}) &= E(\exp\{iu\frac{1}{\sigma\sqrt{n}}((X_1 - \mu) + \cdots + (X_n - \mu))\}) \\ &= [E(\exp\{iu\frac{1}{\sigma\sqrt{n}}(X_1 - \mu)\})]^n \\ &\approx [1 + iu\frac{1}{\sigma\sqrt{n}}E(X_1 - \mu) + \frac{1}{2}(iu\frac{1}{\sigma\sqrt{n}})^2 E((X_1 - \mu)^2)]^n \\ &\approx [1 - \frac{u^2}{2n}]^n \approx e^{-u^2/2}, \end{aligned}$$

by (**??**).

The formal proof must justify the approximations. In particular, one must show that one can ignore all the terms of order higher than two in the Taylor expansion of the exponential.

□

11.4 Approximate Central Limit Theorem

In the CLT, one must know the variance of the random variables to estimate the convergence rate of the sample mean to the mean. In practice it is quite common that one does not know that variance. The following result is then useful.

Theorem 11.4.1. *Approximate CLT Let* $\{X_n, n \geq 1\}$ *be i.i.d. random variables with mean* μ *and such that* $E(X_n^\alpha) < \infty$ *for some* $\alpha > 2$. *Let* $Y_n = (X_1 + \cdots + X_n)/n$, *as before. Then*

$$(Y_n \mu) \frac{\sqrt{n}}{\sigma_n} \rightarrow_D N(0,1) \text{ as } n \rightarrow \infty$$

where

$$\sigma_n^2 = \frac{(X_1 - Y_n)^2 + \cdots + (X_n - Y_n)^2}{n}.$$

That result, which we do not prove here, says that one can replace the variance of the random variables by an estimate of that variance in the CLT.

11.5 Confidence Intervals

Using the approximate CLT, we can construct a confidence interval about our sample mean estimate. Indeed, we can say that when n gets large, (see Table 7.1)

$$P(|(Y_n - \mu)\frac{\sqrt{n}}{\sigma_n}| > 2) \approx 5\%,$$

so that

$$P(\mu \in [Y_n - 2\frac{\sigma_n}{\sqrt{n}}, Y_n + 2\frac{\sigma_n}{\sqrt{n}}]) \approx 95\%.$$

We say that

$$[Y_n - 2\frac{\sigma_n}{\sqrt{n}}, Y_n + 2\frac{\sigma_n}{\sqrt{n}}]$$

is a 95%-confidence interval for the mean.

Similarly,

$$[Y_n - 2.6\frac{\sigma_n}{\sqrt{n}}, Y_n + 2.6\frac{\sigma_n}{\sqrt{n}}]$$

is a 99%-confidence interval for the mean and

$$[\mu_n - \frac{1.6\sigma_n}{\sqrt{n}}, \mu_n + \frac{1.6\sigma_n}{\sqrt{n}}] \text{ is the 90\% confidence interval for } \mu.$$

11.6 Summary

The Strong Law of Large Numbers (SLLN) and the Central Limit Theorem (CLT) are very useful in many applications.

The SLLN says that the sample mean Y_n of n i.i.d. random variables converges to their expected value. The CLT shows that the error multiplied by $\sqrt{n}$ is approximately Gaussian. The CLT enables us to construct confidence intervals

$$[Y_n - \alpha\frac{\sigma_n}{\sqrt{n}}, Y_n - \alpha\frac{\sigma_n}{\sqrt{n}}]$$

where σ_n is the sample estimate of the standard deviation. The probability that the mean is in that interval is approximately $P(|N(0,1)| \leq \alpha)$. Using Table 7.1 one can select the value of α that corresponds to the desired degree of confidence.

11.7 Solved Problems

Example 11.7.1. *Let* $\{X_n, n \geq 1\}$ *be independent and uniformly distributed in* $[0,1]$.

a. Calculate $\lim_{n\to\infty} \frac{\sin(X_1)+\cdots+\sin(X_n)}{n}$.

b. Calculate $\lim_{n\to\infty} \sin(\frac{X_1+\cdots+X_n}{n})$.

c. Calculate $\lim_{n\to\infty} E(\frac{\sin(X_1)+\cdots+\sin(X_n)}{n})$.

d. Calculate $\lim_{n\to\infty} E(\sin(\frac{X_1+\cdots+X_n}{n}))$.

Remember that the Strong Law of Large Numbers says that that if $X_n, n \geq 1$ are i.i.d with mean μ. Then

$$\frac{\sum_{i=1}^{n} X_n}{n} \rightarrow_{\text{as}} \mu.$$

a. We apply the above result to conclude that

$$\frac{\sum_{i=1}^{n} \sin(X_n)}{n} \rightarrow_{\text{as}} E[\sin(X_1)] = 1 - \cos(1) = 0.4597.$$

b. Let $Y_n = \frac{X_1 + \cdots + X_n}{n}$. Then, from the strong law,

$$Y_n \rightarrow_{\text{as}} E(X_1) = 0.5.$$

Hence, since $\sin(\cdot)$ is a continuous function,

$$\lim_{n\to\infty} \sin(Y_n) = \sin(\lim_{n\to\infty} Y_n) = \sin(0.5) = 0.0087.$$

c. Note that

$$|\frac{\sin(X_1) + \cdots + \sin(X_n)}{n}| \leq 1.$$

Consequently, the result of part (a) and Theorem 10.7.1 imply that

$$\lim_{n\to\infty} E(\frac{\sin(X_1) + \cdots + \sin(X_n)}{n}) = 0.4597.$$

d. Using the same argument as in part (c) we find that

$$\lim_{n\to\infty} E(\sin(\frac{X_1 + \cdots + X_n}{n})) = 0.5.$$

Example 11.7.2. *Let $\{X_n, n \geq 1\}$ be i.i.d. $N(0,1)$. What can you say about*

$$\cos(\frac{X_1^4 + X_2^4 + \cdots + X_n^4}{n})$$

as $n \to \infty$?

By the SLLN, we know that

$$\frac{X_1^4 + X_2^4 + \cdots + X_n^4}{n} \rightarrow_{a.s.} E(X_1^4) = 3 \text{ as } n \to \infty.$$

Since cos(.) is a continuous function, it follows that

$$\cos(\frac{X_1^4 + X_2^4 + \cdots + X_n^4}{n}) \rightarrow_{a.s.} \cos(3) \text{ as } n \rightarrow \infty.$$

Example 11.7.3. *Let $\{X_n, n \geq 1\}$ be i.i.d. with mean μ and finite variance σ^2. Let*

$$\mu_n := \frac{X_1 + \cdots + X_n}{n} \text{ and } \sigma_n^2 = \frac{(X_1 - \mu_n)^2 + \cdots + (X_n - \mu_n)^2}{n}. \tag{11.7.1}$$

Show that

$$\sigma_n^2 \rightarrow_{a.s.} \sigma^2 \text{ as } n \rightarrow \infty.$$

By the SLLN, $\mu_n \rightarrow_{a.s.} \mu$. You can then show that $\sigma_n^2 - s_n^2 \rightarrow_{a.s.} 0$ where

$$s_n^2 := \frac{(X_1 - \mu)^2 + \cdots + (X_n - \mu)^2}{n}.$$

But, by SLLN, $s_n^2 \rightarrow_{a.s.} \sigma^2$ as $n \rightarrow \infty$. This completes the proof.

Example 11.7.4. *We want to poll a population to estimate the fraction p of people who will vote for Bush in the next presidential election. We want to to find the smallest number of people we need to poll to estimate p with a margin of error of plus or minus 3% with 95% confidence.*

For simplicity it is assumed that the decisions X_n of the different voters are i.i.d. $B(p)$. Then we know that

$$[\mu_n - \frac{2\sigma_n}{\sqrt{n}}, \mu_n + \frac{2\sigma_n}{\sqrt{n}}] \text{ is the 95\% confidence interval for } \mu.$$

We also know that the variance of the random variables is bounded by 1/4. Indeed, $\text{var}(X_n) = p(1-p) \leq 1/4$. Thus,

$$[\mu_n - \frac{1}{\sqrt{n}}, \mu_n + \frac{1}{\sqrt{n}}] \text{ contains the 95\% confidence interval for } \mu.$$

If we want the error to be 3%, we need

$$\frac{1}{\sqrt{n}} \leq 3\%,$$

i.e.,

$$n \geq (\frac{1}{0.03})^2 = 1112.$$

Example 11.7.5. *If $X = 0$, $Y = B(n, p_0)$ and if $X = 1$, $Y = B(n, p_1)$ where $p_0 = 0.5$ and $p_1 = 0.55$. Let $\alpha = 0.5\%$ and $\hat{X} = HT[X \mid Y]$. Find the value of n needed so that $P[\hat{X} = 0 \mid X = 1] = \alpha$. Use the CLT to estimate the probabilities. (Note that $\hat{X}$ is such that $P[\hat{X} = 1 \mid X = 0] = \alpha$, but we want also $P[\hat{X} = 0 \mid X = 1] = \alpha$.)*

We know that $\hat{X} = 1\{Y \geq y_0\}$ where $P[Y \geq y_0 \mid X = 0] = \alpha$. We can write $Y = Y_1 + \cdots + Y_n$ where the Y_k are i.i.d. $B(1/2)$ if $X = 0$. Now, by CLT,

$$Z_n := \sqrt{n}[\frac{Y_1 + \cdots + Y_n}{n} - 0.5] \approx N(0, 1/4).$$

Then,

$$\begin{aligned} P(Y_1 + \cdots + Y_n > y_0) &= P(Z_n > \sqrt{n}(\frac{y_0}{n} - 0.5)) \\ &\approx P(N(0, \frac{1}{4}) > \sqrt{n}(\frac{y_0}{n} - 0.5))) = P(N(0,1) > 2\sqrt{n}(\frac{y_0}{n} - 0.5)). \end{aligned}$$

We need

$$2\sqrt{n}(\frac{y_0}{n} - 0.5)) \geq 2.6$$

for this probability to be 0.5%. Thus,

$$y_0 = 0.5n + 1.3\sqrt{n}.$$

Then

$$P[\hat{X} = 0 \mid X = 1] = P[Y < y_0 \mid X = 1].$$

Now, when $X = 1$, $Y = W_1 + \cdots + W_n$ where the W_k are $B(0.55)$. Thus,

$$P[Y < y_0 \mid X = 1] = P(W_1 + \cdots + W_n < y_0).$$

By CLT,

$$U_n := \sqrt{n}[\frac{W_1 + \cdots + W_n}{n} - 0.55] \approx N(0, 1/4).$$

(We used the fact that $0.55(1 - 0.55) \approx 1/4$.)

Thus,

$$\begin{aligned} P(W_1 + \cdots + W_n < y_0) &= P(U_n < \sqrt{n}[\frac{y_0}{n} - 0.55]) \\ &\approx P(N(0, \frac{1}{4}) < \sqrt{n}[\frac{y_0}{n} - 0.55]) = P(N(0,1) < 2\sqrt{n}[\frac{y_0}{n} - 0.55]) \\ &= P(N(0,1) < 2\sqrt{n}[\frac{0.5n + 1.3\sqrt{n}}{n} - 0.55]). \end{aligned}$$

For this probability to be equal to $\alpha = 0.5\%$, we need

$$2\sqrt{n}[\frac{0.5n + 1.3\sqrt{n}}{n} - 0.55] = -2.6.$$

This gives $n = (52)^2 = 2704$.

Example 11.7.6. *Let $\{X_n, n \geq 1\}$ be i.i.d. random variables that are uniformly distributed in $[0, 1]$. Express the following limit*

$$\lim_{n \to \infty} P(\sqrt{n}|\frac{X_1 + \cdots + X_n}{n} - \frac{1}{2}| > \epsilon)$$

in terms of $Q(x) := P(N(0,1) > x)$, for $x \in \Re$.

From the Central Limit Theorem,

$$\sqrt{n}|\frac{X_1 + \cdots + X_n}{n} - \frac{1}{2}| \to_D N(0, \sigma^2)$$

where

$$\sigma^2 = \text{var}(X_1) = E(X_1^2) - (E(X_1))^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Hence,

$$P(\sqrt{n}|\frac{X_1+\cdots+X_n}{n}-\frac{1}{2}|>\epsilon)\approx P(|N(0,\frac{1}{12})|>\epsilon).$$

Now,

$$\begin{aligned}P(|N(0,\frac{1}{12})|>\epsilon)&=P(|\frac{1}{\sqrt{12}}N(0,1)|>\epsilon)\\&=P(|N(0,1)|>\epsilon\sqrt{12})=2P(N(0,1)>\epsilon\sqrt{12})=2Q(\epsilon\sqrt{12}).\end{aligned}$$

Consequently,

$$\lim_{n\to\infty}P(\sqrt{n}|\frac{X_1+\cdots+X_n}{n}-\frac{1}{2}|>\epsilon)=2Q(\epsilon\sqrt{12}).$$

Example 11.7.7. *Given X, the random variables $\{Y_n, n\geq 1\}$ are exponentially distributed with mean X. Assume that $P(X=1)=1-P(X=2)=p\in(0,1)$.*

Find the estimate $\hat{X}_n$ based on Y^n that minimizes $P[\hat{X}_n=1\mid X=2]$ subject to $P[\hat{X}_n=2\mid X=1]\leq\beta$, for $\beta=5\%$ and n large.

Hint: *Use the result of Example 8.6.8 and the CLT.*

In Example 8.6.8, we saw that

$$\hat{X}=\begin{cases}2,\ \text{if}\ \sum_{i=1}^n Y_i>\rho;\\ 1,\ \text{if}\ \sum_{i=1}^n Y_i\leq\rho.\end{cases}$$

We choose ρ so that

$$P[\sum_{i=1}^n Y_i>\rho\mid X=1]=\beta.$$

Let us write $\rho=n+\alpha\sqrt{n}$ and find α so that

$$P[\sum_{i=1}^n Y_i>n+\alpha\sqrt{n}\mid X=1]=\beta.$$

Equivalently,

$$P[(\sum_{i=1}^n Y_i-n)/\sqrt{n}>\alpha\mid X=1]=\beta.$$

Now, given $\{X = 1\}$, the Y_i are i.i.d. $Exd(1)$, so that, by CLT,

$$(\sum_{i=1}^{n} Y_i - n)/\sqrt{n} \rightarrow_D N(0, \sigma^2)$$

where $\sigma^2 = \text{var}(Y_i) = 1$. According to (7.1) we need $\alpha = 1.7$. Hence,

$$\hat{X} = \begin{cases} 2, \text{ if } \sum_{i=1}^{n} Y_i > n + 1.7\sqrt{n}; \\ 1, \text{ if } \sum_{i=1}^{n} Y_i \leq n + 1.7\sqrt{n}. \end{cases}$$

Example 11.7.8. *Let $\{X_n, n \geq 1\}$ be a sequence of independent Bernoulli random variables with mean $p \in (0, 1)$. We construct an estimate $\hat{p}_n$ of p from $\{X_1, \ldots, X_n\}$. We know that $p \in (0.4, 0.6)$. Find the smallest value of n so that*

$$P(\frac{|\hat{p}_n - p|}{p} \leq 5\%) \geq 95\%.$$

For Bernoulli Random Variables we have $E[X_n] = p$ and $\text{var}[X_n] = p(1-p)$. Let $\hat{p}_n = (X_1 + \cdots + X_n)/n$. We know that $\hat{p}_n \rightarrow_{\text{as}} p$. Moreover, from the CLT,

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \rightarrow_D N(0, 1).$$

Now,

$$\frac{|\hat{p}_n - p|}{p} \leq 0.05 \Leftrightarrow |\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}}| \leq 0.05\frac{\sqrt{np}}{\sqrt{1-p}}.$$

Hence, for n large,

$$P(\frac{|\hat{p}_n - p|}{p} \leq 0.05) \approx P(|N(0,1)| \leq \frac{0.05 \times \sqrt{np}}{\sqrt{1-p}}).$$

Using (7.1) we find $P(|N(0,1)| \leq 2) \geq 0.95$. Hence, $P(\frac{|\hat{p}_n - p|}{p} \leq 0.05)$ if

$$\frac{0.05 \times \sqrt{np}}{\sqrt{1-p}} \geq 2,$$

i.e.,

$$n \geq 1600\frac{1-p}{p} =: n_0.$$

Since we know that $p \in [0.4, 0.6]$, the above condition implies $1067 \leq n_0 \leq 2400$. Hence the lowest value of n required to ensure a 95% accuracy is 2400.

Example 11.7.9. *Given X, the random variables $\{Y_n, n \geq 1\}$ are independent and Bernoulli(X). We want to decide whether $X = 0.5$ or $X = 0.6$ by observing $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$. Let $\hat{X}$ be the estimator that minimizes $P[\hat{X} = 0.5|X = 0.6]$ subject to $P[\hat{X} = 0.6|X = 0.5] \leq 5\%$. Using the CLT, estimate the smallest value of n so that $P[\hat{X} = 0.5|X = 0.6] \leq 5\%$.*

Lets define the default and the alternative hypothesis as follows:

$H_0 : X = 0.5$

$H_1 : X = 0.6.$

Then the Likelihood ratio $L(\boldsymbol{y})$ can be computed as follows:

$$L(\boldsymbol{y}) = \frac{\prod_{i=1}^n 0.6^{y_i} 0.4^{1-y_i}}{\prod_{i=1}^n 0.5^{y_i} 0.5^{1-y_i}} = \left(\frac{0.6}{0.5}\right)^{\sum_{i=1}^n y_i} \left(\frac{0.4}{0.5}\right)^{\sum_{i=1}^n (1-y_i)} = \left(\frac{0.6}{0.5}\right)^k \left(\frac{0.4}{0.5}\right)^{n-k}$$

where $k = \sum_{i=1}^n y_i$ is the number of heads in n Bernoulli trials. Note that $L(\boldsymbol{y})$ is an increasing function of the number of heads. According to the Neyman-Pearson theorem, we are looking for that value of $k = k_0$ such that

$$0.05 = P[\hat{X} = 0.6|X = 0.5] = P(\sum_{i=1}^n Y_i \geq k_0).$$

We use CLT to estimate k_0. We observe that

$$\sum_{i=1}^n Y_i \geq k_0 \Leftrightarrow \sqrt{\frac{n}{0.25}}(\frac{1}{n}\sum_{i=1}^n Y_i - 0.5) \geq \sqrt{\frac{n}{0.25}}(\frac{1}{n}k_0 - 0.5) \Leftrightarrow N(0,1) \geq \sqrt{\frac{n}{0.25}}(\frac{1}{n}k_0 - 0.5).$$

Using (7.1) we find $P(N(0,1) \geq 1.7) \approx 0.05$. Hence, we need to find k_0 such that

$$\sqrt{\frac{n}{0.25}}(\frac{1}{n}k_0 - 0.5) \approx 1.7. \tag{11.7.2}$$

Next, we consider $P[\hat{X} = 0.5|X = 0.6] \leq 5\%$. Note that

$$\sum_{i=1}^n Y_i \leq k_0 \Leftrightarrow \sqrt{\frac{n}{0.24}}(\frac{1}{n}\sum_{i=1}^n Y_i - 0.6) \leq \sqrt{\frac{n}{0.24}}(\frac{1}{n}k_0 - 0.6) \Leftrightarrow N(0,1) \leq \sqrt{\frac{n}{0.24}}(\frac{1}{n}k_0 - 0.6).$$

Hence, for n large, using (7.1) we find $P[\hat{X} = 0.5|X = 0.6] \leq 5\%$ is equivalent to

$$\sqrt{\frac{n}{0.24}}(\frac{1}{n}k_0 - 0.6) \approx -1.7. \tag{11.7.3}$$

Ignoring the difference between 0.24 and 0.25, we find that the two equations (11.7.2)-(11.7.3) imply

$$\frac{1}{n}k_0 - 0.5 \approx -(\frac{1}{n}k_0 - 0.6),$$

which implies $\frac{1}{n}k_0 \approx 0.55$. Substituting this estimate in (11.7.2), we find

$$\sqrt{\frac{n}{0.25}}(0.55 - 0.5) \approx 1.7,$$

which implies $n \approx 72$, which finally yields $k_0 \approx 0.55n \approx 40$.

Example 11.7.10. *The number of days that a certain type of component functions before failing is a random variable with pdf (one time unit = 1 day)*

$$f_X(x) = 2x, 0 < x < 1.$$

Once the component fails it is immediately replaced by another one of the same type. If we designate by X_i the lifetime of the ith component to be put to use, then $S_n = \sum_{i=1}^{n} X_i$ represents the time of the nth failure.

a. The long-term rate at which failures occurs, call it r, is defined by

$$r = \lim_{n\to\infty} \frac{n}{S_n}.$$

Assuming that the random variables $X_i, i \geq 1$ are independent, determine r.

b. Use the CLT to determine how many components one would need to have to be approximately 95% certain of not running out after 365 days.

a. We first note that $E(X) = \int_0^1 x f_X(x)dx = 2/3$. Using the Strong Law of Large Numbers, we find that

$$r = \lim_{n\to\infty} \frac{n}{S_n} = \lim_{n\to\infty} [\frac{S_n}{n}]^{-1} = [\lim_{n\to\infty} \frac{S_n}{n}]^{-1} = [\frac{2}{3}]^{-1} = 1.5 \text{ failures per day.}$$

(Note: In the third identity we can interchange the function $g(x) = x^{-1}$ and the limit because $g(x)$ is continuous at $x = 2/3$.)

b. Since there are about 1.5 failures per day we will need a few more than $365 \times 1,5 = 548$ components. We use the CLT to estimate how many more we need. You can verify that $\text{var}(X) = 1/18$. Let $Z =_D N(0,1)$. We want to find k so that $P(S_k > 365) \geq 0.95$. Now,

$$P(S_k > 365) = P(\frac{S_k - k(2/3)}{\sqrt{k/18}} > \frac{365 - k(2/3)}{\sqrt{k/18}}) \approx P(Z > \frac{365 - k(2/3)}{\sqrt{k/18}}).$$

Using (7.1), we should choose k so that

$$\frac{365 - k(2/3)}{\sqrt{k/18}} \leq -1.7,$$

or equivalently,

$$k\frac{2}{3} - 365 \geq 1.7\sqrt{k/18}, \text{ i.e., } (k\frac{2}{3} - 365)^2 \geq (1.7)^2 k\frac{1}{18},$$

or

$$\frac{4}{9}k^2 - \frac{4}{3} \times 365k + (365)^2 \geq (1.7)^2 k\frac{1}{18}, \text{ i.e., } k^2\frac{4}{9} - [\frac{4}{3} \times 365 + (1.7)^2\frac{1}{18}]k + (365)^2 \geq 0.$$

That is,

$$0.444k^2 - 486.6k + 133,225 \geq 0.$$

This implies

$$k \geq \frac{486.6 + \sqrt{(486.6)^2 - 4 \times 0.444 \times 133,225}}{2 \times 0.444} \approx 563.$$

Chapter 12

Random Processes Bernoulli - Poisson

We have looked at a finite number of random variables. In many applications, one is interested in the evolution in time of random variables. For instance, one watches on an oscilloscope the noise across two terminals. One may observe packets that arrive at an Internet router, or cosmic rays hitting a detector. If ω designates the outcome of the random experiment (as usual) and t the time, then one is interested in a collection of random variables $X = \{X(t,\omega), t \in \mathcal{T}\}$ where $\mathcal{T}$ designates the set of times. Typically, $\mathcal{T} = \{0,1,2,\ldots\}$ or $\mathcal{T} = \{\ldots,-2,-1,0,1,2,\ldots\}$ or $\mathcal{T} = [0,T]$ for some $T < \infty$ or $\mathcal{T} = [0,\infty)$ or $\mathcal{T} = (-\infty,\infty)$. When $\mathcal{T}$ is countable, one says that X is a *discrete-time random process*. When T is an interval (possibly infinite), one says that X is a *continuous-time random process*.

We explained that a collection of random variables is characterized by their joint cdf. Similarly, a random process is characterized by the joint cdf of any finite collection of the random variables. These joint cdf are called the *finite dimensional distributions* of the random process. For instance, to specify a random process $\{X_t, t \geq 0\}$ one must specify the joint cdf of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ for any value of $n \geq 1$ and any $0 < t_1 < \cdots < t_n$. In most applications, the process is defined by means of a collection of other random

variables that have well-defined joint cdf. In some cases however, one specifies the finite dimensional distributions. An interesting mathematical question is whether there is always a random process that corresponds to a set of finite dimensional distributions. Obviously, to correspond to a random process, the finite dimensional distributions must be consistent. That is, the finite dimensional distribution specifies the joint cdf of a set of random variables must have marginal distributions for subsets of them that agree with the finite dimensional distribution of that subset. For instance, the marginal distribution of X_{t_1} obtained from the joint distribution of (X_{t_1}, X_{t_2}) should be the same as the distribution specified for X_{t_1}. Remarkably, these obviously necessary conditions are also sufficient! This result is known as Kolmogorov's extension theorem.

In this section, we look at two simple random processes: the Bernoulli process and the Poisson process. We consider two other important examples in the next two chapters.

12.1 Bernoulli Process

Definition 12.1.1. Let $X = \{X_n, n \geq 1\}$ be i.i.d. with $P(X_n = 1) = p = 1 - P(X_n = 0)$. The discrete-time random process is called the Bernoulli process. This process models flipping a coin.

12.1.1 Time until next 1

Assume we have watched the first n coin flips. How long do we wait for the next 1? That is, let

$$\tau = \min\{m > 0 | X_{n+m} = 1\}.$$

We want to calculate

$$P[\tau > m | X_1, X_2, \ldots, X_n].$$

Because of independence, we find

$$P[\tau > m|X_1, X_2, \ldots, X_n] = P[X_{n+1} = 0, X_{n+2} = 0, \ldots, X_{n+m} = 0|X_1, X_2, \ldots, X_n] = (1-p)^m.$$

Thus, the random time until the next 1 is $G(p)$.

12.1.2 Time since previous 1

Assume that we have been flipping the coin forever. How has it been since the last 1? Let σ designate that number of steps. We see that, for $m \geq 0$, $\sigma = m$ if $X_n = 0, X_{n-1} = 0, \ldots, X_{n-m-1} = 0, X_{n-m} = 1$ and the probability of that event is $p(1-p)^m$. Thus, $\sigma + 1 = G(p)$. That is, $E(\sigma) = (1/p) - 1$.

12.1.3 Intervals between 1s

There are two ways of looking at the time between two successive 1s.

The first way is to choose some time n and to look at the time since the last 1 and the time until the next 1. We know that these times have mean $(1/p) - 1$ and $1/p$, respectively. In particular, the average time between these two 1s around some time n is $(2/p) - 1$. Note that this time is equal to 1 when $p = 1$, as it should be.

The second way is to start at some time n, wait until the next 1 and then count the time until the next 1. But in this way, we find that the time between two consecutive 1s is geometrically distributed with mean $1/p$.

12.1.4 Saint Petersburg Paradox

The two different answers that we just described are called the Saint Petersburg paradox. The way to explain it is to notice that when we pick some time n and we look at the previous and next 1s, we are more likely to have picked an n that falls in a large gap between two 1s than an n that falls in a small gap. Accordingly, we expect the average duration of the

interval in which we have picked n to be larger than the typical interval between consecutive 1s. In other words, by picking n we face a sampling bias.

12.1.5 Memoryless Property

The geometric distribution is memoryless. That is, if τ is geometrically distributed with mean $1/p$, then if we know that τ is larger than n, then $\tau - n$ is still geometrically distributed with mean $1/p$. Indeed, we find that

$$P[\tau - n > m | \tau > n] = P[\tau > m + n | \tau > n] = \frac{P(\tau > m + n \text{ and } \tau > n)}{P(\tau > n)}$$
$$= \frac{P(\tau > m + n)}{P(\tau > n)} = \frac{(1-p)^{m+n}}{(1-p)^n} = (1-p)^m = P(\tau > m).$$

Intuitively, this result is immediate if we think of τ as the time until the next 1 for a Bernoulli process. Knowing that we have already flipped the coin n times and got 0s does not change how many more times we still have to flip it to get the next 1. (Remember that we assume that we know the probability of getting a 1.)

12.1.6 Running Sum

As before, let $\{X_n, n \geq 1\}$ be i.i.d. $B(p)$. Define $Y_n = Y_0 + 2(X_1 + \ldots + X_n)n$. The interpretation of Y_n is that it represents the accumulated fortune of a gambler who gains 1 with probability p and loses 1 otherwise at each step of a game; the steps are all independent. Two typical evolutions of Y_n are shown in Figure 12.1 when $Y_0 = 0$. The top graph corresponds to $p = 0.54$, the bottom one to p = 0.46. The sequence Y_n is called a *random walk* because its increments $\{Y_n - y_{n-1}, n \geq 1\}$ are i.i.d.

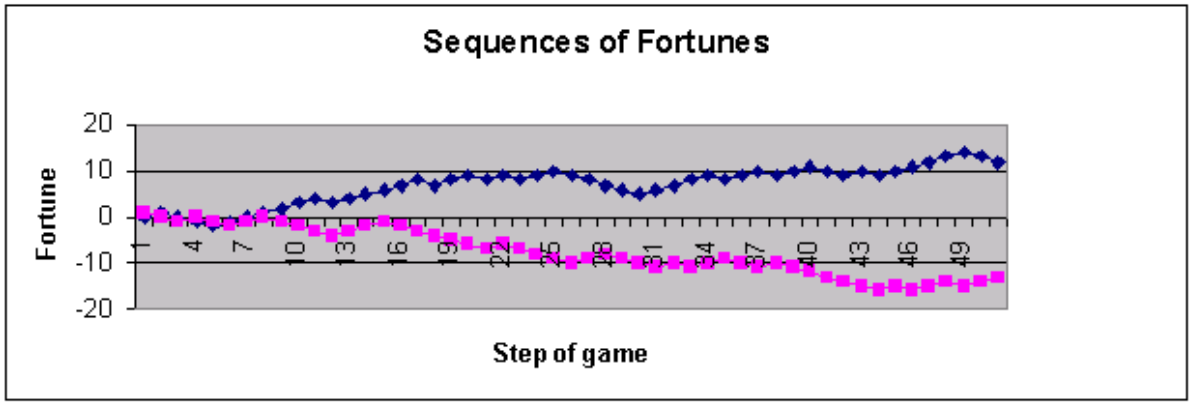


Figure 12.1: Random Walk with $p = 0.54$ (top) and $p = 0.46$ (bottom).

12.1.7 Gamblers Ruin

Assume the gambler plays the above game and starts with an initial fortune $Y_0 = A > 0$. What is the probability that her fortune will reach $B > A$ before it reaches 0 and she is bankrupt? To solve this problem, we define $T_B = \min\{n \geq 0 | Y_n = B\}$. We define T_0 in a similar way. We want to compute $\alpha(A) = P[T_B < T_0 | Y_0 = A]$. A moment (or two) of reflection shows that if $A > 0$, then

$$\begin{aligned}
\alpha(A) &= P[T_B < T_0 \text{ and } X_1 = 1 | Y_0 = A] + P[T_B < T_0 \text{ and } X_1 = -1 | Y_0 = A] \\
&= pP[T_B < T_0 | Y_0 = A, X_1 = 1] + (1-p)P[T_B < T_0 | Y_0 = A, X_1 = -1] \\
&= pP[T_B < T_0 | Y_1 = A+1, X_1 = 1] + (1-p)P[T_B < T_0 | Y_1 = A-1, X_1 = -1] \\
&= pP[T_B < T_0 | Y_1 = A+1] + (1-p)P[T_B < T_0 | Y_1 = A-1] \\
&= p\alpha(A+1) + (1-p)\alpha(A-1).
\end{aligned}$$

Note that these equations are derived by conditioning of the first step of the process. Equations derived in that way are called *first step equations*.

The boundary conditions of these ordinary difference equations are $\alpha(0) = 0$ and $\alpha(B) =$

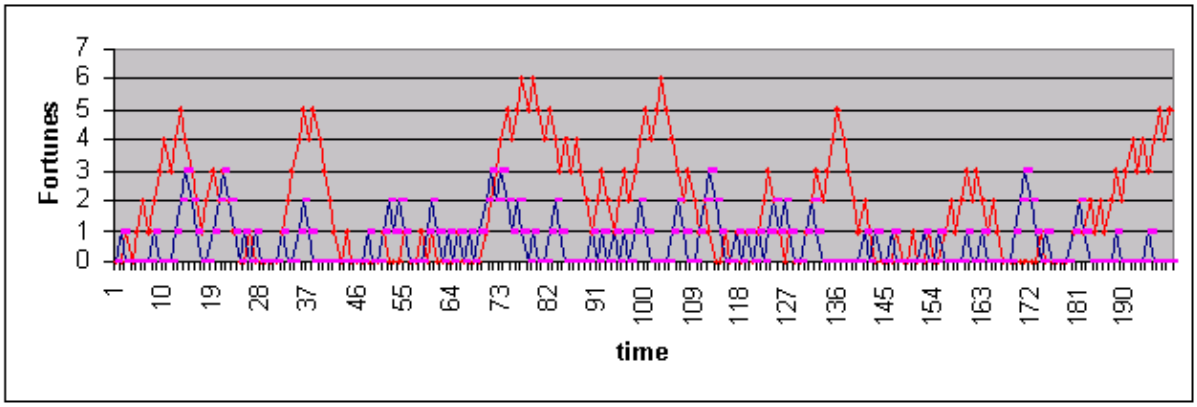


Figure 12.2: Reflected Random Walk with $p = 0.43$ (top) and $p = 0.3$ (bottom).

1. You can verify that the solution is

$$\alpha(A) = \begin{cases} \frac{A}{B}, \text{ if } p = 0.5 \\ \frac{\rho^A - 1}{\rho^B - 1} \text{ with } \rho := \frac{1-p}{p}, \text{ if } p \neq 0.5. \end{cases}$$

For instance, with $p = 0.48, A = 100$, and $B = 1000$, one finds $\alpha(A) = 2 \times 10^{-35}$. Remember this on your next trip to Las Vegas.

12.1.8 Reflected Running Sum

Assume now that you have a rich uncle who gives you \$1.00 every time you go bankrupt, so that you can keep on playing the game forever. To define the game precisely, say that when you hit 0, you can still play the game. If you lose again, you are back at 0, otherwise to have 1, and so on. The resulting process is called a *reflected random walk*. Figure 12.2 shows a typical evolution of your fortune. The top graph corresponds to $p = 0.43$ and the other one to $p = 0.3$. Not surprisingly, with $p = 0.3$ you are always poor.

One interesting question is how much money you have, on average. For instance, looking at the lower graph, we can see that a good fraction of the time your fortune is 0, some other fraction of the time it is 1, and a very small fraction of the time it is larger than 2. How van we calculate these fractions? One way to answer this question is as follows. Assume

that $\pi(k) = P(Y_n = k)$ for $k \geq 1$ and all $n \geq 0$. That is, we assume that the distribution of Y_n does not depend on n. Such a distribution π is said to be *invariant*. Then, for $k > 0$, one has

$$P(Y_{n+1} = k) = P(Y_{n+1} = k, Y_n = k-1) + P(Y_{n+1} = k, Y_n = k+1),$$

so that

$$\begin{aligned}\pi(k) &= P[Y_{n+1} = k | Y_n = k-1]P(Y_n = k-1) + P[Y_{n+1} = k | Y_n = k+1]P(Y_n = k+1) \\ &= p\pi(k-1) + (1-p)\pi(k+1), k = 1, 2, \ldots.\end{aligned}$$

For $k = 0$, one has

$$P(Y_{n+1} = 0) = P(Y_{n+1} = 0, Y_n = 0) + P(Y_{n+1} = 0, Y_n = 1),$$

so that

$$\begin{aligned}\pi(0) &= P[Y_{n+1} = 0 | Y_n = 0]P(Y_n = 0) + P[Y_{n+1} = 0 | Y_n = 1]P(Y_n = 1) \\ &= p\pi(0) + (1-p)\pi(1).\end{aligned}$$

The above identities are called the *balance equations*. You can verify that the solution of these equations is

$$\pi(k) = A\rho^k, k \geq 0 \text{ where } \rho = \frac{p}{1-p}.$$

Since $\sum_k \pi(k) = 1$, we see that a solution is possible if $\rho < 1$, i.e.. if $p < 0.5$. The solution corresponds to $A = 1 - \rho$, so that

$$\pi(k) = (1-\rho)\rho^k, k \geq 0 \text{ where } \rho = \frac{p}{1-p} \text{ when } p < 0.5.$$

When $p \geq 0.5$, there is not solution where $P(Y_n = k)$ does not depend on n. To understand what happens in that case, lets look at the case $p = 0.6$. The evolution of the fortune in that case is shown in Figure 12.3.

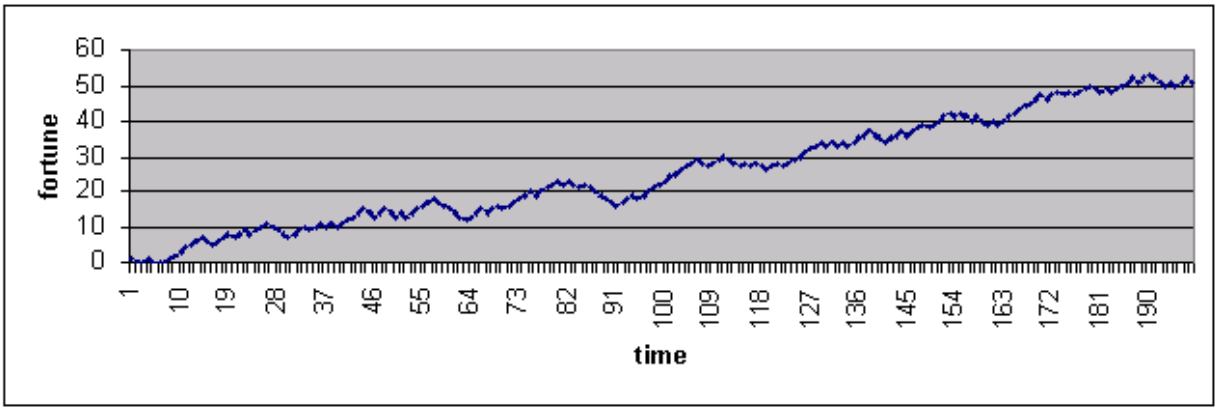


Figure 12.3: Random Walk with $p = 0.6$.

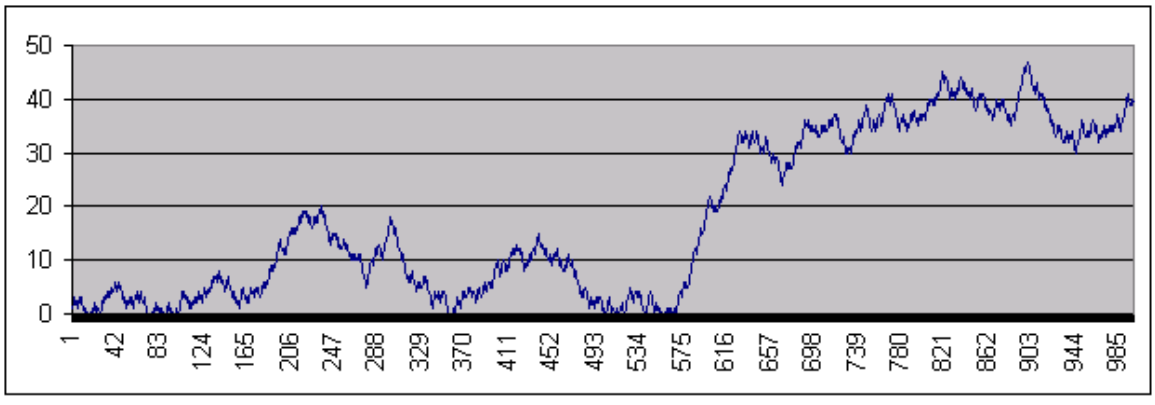


Figure 12.4: Random Walk with $p = 0.5$.

The graph shows that, as time evolves, you get richer and richer. The case $p = 0.5$ is more subtle. Figure 12.4 shows a simulation result.

In this case, the fortune fluctuates and makes very long excursions. One can show that the fortune comes back to 0 infinitely often but that it takes a very long time to come back. So long in fact that the fraction of time that the fortune is zero is negligible. In fact, the fraction of the time that the fortune is any given value is zero!

How do we show all this? Let $p = 0.5$. We know that $P[T_B < T_0|Y_0 = A] = A/B$. Thus, $P[T_0 < T_B|Y_0 = A] = (B - A)/B$. As B increases to infinity, we see that T_B also increases to infinity, to that $P[T_0 < T_B|Y_0 = A]$ increases to $P[T_0 < \infty|Y_0 = A]$ (by continuity of probabilities). Thus, $P[T_0 < \infty|Y_0 = A] = 1$ (because $(B - A)/B$ tends to 1 as B increases to infinity). This shows that the fortune comes back to 0 infinitely often. We can calculate how longs it takes to come back to 0. To do this, define $\beta(A) = E[\min\{T_0, T_B\}|Y_0 = A]$. Arguing as we have for the gamblers ruin problem, you can justify that

$$\beta(A) = 1 + 0.5\beta(A - 1) + 0.5\beta(A + 1) \text{ for } A > 0.$$

Also, $\beta(0) = 0 = \beta(B)$. You can verify that the solution is $\beta(A) = A(B - A)$. Now, as B increases to infinity, T_B also increases to infinity. Since T_0 is finite, it follows that $\min\{T_0, T_B\}$ increases to T_0. It follows from Theorem 10.7.1 that $E[T_0|Y_0 = A] = \infty$ for all $A > 0$.

12.1.9 Scaling: SLLN

Going back to the case $p > 0.5$, we see that the fortune appears to be increasing at an approximately constant rate. In fact, using the SLLN we can see that

$$\frac{Y_{n+m} - Y_n}{m} = \frac{X_{n+1} + \cdots + X_{n+m}}{m} \approx E(X_n) = p + (-1)(1 - p) = 2p - 1 \text{ for } m \gg 1.$$

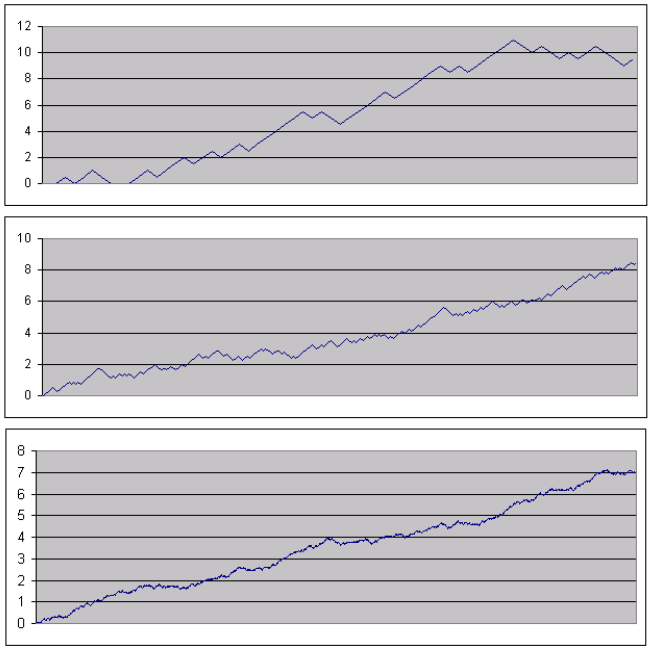


Figure 12.5: Scaled versions or Random Walk.

We can use that observation to say that Y_n grows at rate $2p-1$. We can make this more precise by defining the scaled process $Z_k(t) := Y_{[kt]}/k$ where $[kt]$ is the smallest integer larger than or equal to kt. We can then say that $Z_k(t) \to (2p-1)t$ as $k \to \infty$. The convergence is almost sure for any given t. However, we would like to say that this is true for the trajectories of the process. To see what we mean by this convergence, lets look as a few scaled versions shown in Figure 12.5.

These three scaled versions show that the fluctuations gets smaller as we scale and that the trajectory becomes closer to a straight line, in some uniform sense.

12.1.10 Scaling: Brownian

Of course, if we look very closely at the last graph above, we can still see some fluctuations. One way to see these well is to blow up the y-axis. We show a portion of this graph in Figure 12.6.

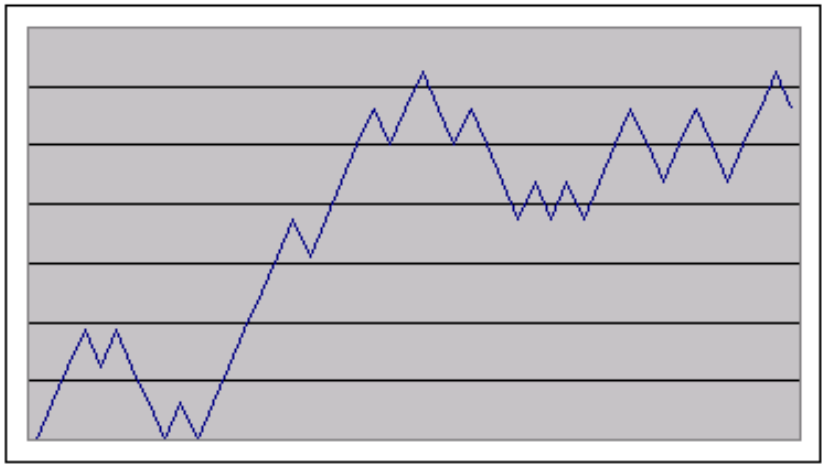


Figure 12.6: Blowing up a scaled random walk.

The fluctuations are still there, obviously. One way to analyze these fluctuations is to use the central limit theorem. Using the CLT, we find that

$$\frac{Y_{n+m} - Y_n - m(2p-1)}{\sqrt{m}} = \frac{X_{n+1} + \cdots + X_{n+m} - m(2p-1)}{\sqrt{m}} \approx N(0, \sigma^2)$$

where $\sigma^2 = p(1-p)$. This shows that, properly scaled, the increments of the fortune look Gaussian. The case $p = 0.5$ is particularly interesting, because then we do not have to worry about the mean. In that case,

$$\frac{Y_{n+m} - Y_n}{\sqrt{m}} \approx N(0, 1/4).$$

We can then scale the process differently and look at $Z_k(t) = Y_{[kt]}/\sqrt{k}$. We find that as k becomes very large, the increments of $Z_k(t)$ become independent and Gaussian. In fact, $Z_k(t+u) - Z_k(t)$ is $N(0, u/4)$. If we multiply the process by 2, we end up with a process $W(t)$ with independent increments such that $W(t+u) - W(t) = N(0, u)$. Such a process is called a standard Brownian motion process or Wiener process.

12.2 Poisson Process

Definition 12.2.1. Let $X = \{X(t), t \geq 0\}$ be defined as follows. For $t \geq 0$, $X(t)$ is the number of jumps in $[0, t]$. The jump times are $\{T_n, n \geq 1\}$ where $\{T_1, T_2 - T_1, T_3 - T_2, T_4 - T_3, \ldots\}$ are i.i.d. and exponentially distributed with mean $1/\lambda$ where $\lambda > 0$. Thus, the times between jumps are exponentially distributed with mean $1/\lambda$ and are all independent. The process X is called a Poisson process with rate λ.

12.2.1 Memoryless Property

Recall that the exponential distribution is memoryless. This implies that, for any $t > 0$, given $\{X(s), 0 \leq s \leq t\}$, the process $\{X(s) - X(t), s \geq t\}$ is again Poisson with rate λ.

12.2.2 Number of jumps in $[0, t]$

The number of jumps in $[0, t]$, $X(t)$, is a Poisson random variable with mean λt. That is, $P(X(t) = n) = (\lambda t)^n \exp\{-\lambda t\}/n!$ for $n \geq 0$. In view of the memoryless property, the increments of the Poisson process are independent and Poisson distributed.

Indeed, the jump times $\{T_1, T_2, \ldots\}$ are separated by i.i.d. $Exp(\lambda)$ random variables. That is, for any selection of $0 < t_1 < \ldots < t_n < t$,

$$\begin{aligned} &P(T_1 \in (t_1, t_1 + \epsilon), \ldots, T_n \in (t_n, t_n + \epsilon), T_{n+1} > t) \\ &\quad = \lambda\epsilon \exp\{-\lambda t_1\} \lambda\epsilon \exp\{-\lambda(t_2 - t_1)\} \cdots \lambda\epsilon \exp\{-\lambda(t_n - t_{n-1})\} \exp\{-\lambda(t - t_n)\} \\ &\quad = (\lambda\epsilon)^n \exp\{-\lambda t\}. \end{aligned}$$

Hence, the probability that there are n jumps in $[0, t]$ is the integral of the above density on the set $S = \{t_1, \ldots, t_n | 0 < t_1 < \ldots < t_n < t\}$, i.e., $|S|\lambda^n \exp\{-\lambda t\}$ where $|S|$ designates the volume of the set S. This set is the fraction $1/n!$ of the cube $[0, t]^n$. Hence $|S| = t^n/n!$ and

$$P(n \text{ jumps in } [0,t]) = [(\lambda t)^n/n!] \exp\{-\lambda t\}.$$

12.2.3 Scaling: SLLN

We know, from the strong law of large numbers, that $X(nt)/n \to \lambda t$ almost surely as $n \to \infty$. As in the case of the Bernoulli process, the process $\{X(nt)/n, t \geq 0\}$ converges to $\{\lambda t, t \geq 0\}$ in a "trajectory" sense that we have not defined.

12.2.4 Scaling: Bernoulli → Poisson

Imagine a Bernoulli process $\{X_n, n \geq 1\}$ with parameter p. We look at $Y_n = X_1 + \cdots + X_n$. The claim is that if $p \to 0$ and $k \to \infty$ in a way that $kp \to \lambda$, then the process $\{W_k(t) = Y_{kt}, t \geq 0\}$ approaches a Poisson process with rate λ. This property follows from the corresponding approximation of a Poisson random variable by a binomial random variable. (See Section 10.2.)

12.2.5 Sampling

Imagine customers that arrive as a Poisson process with rate λ. With probability p, a customer is male, independently of the other customers and of the arrival times. The claim is that the processes $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$ that count the arrivals of male and female customers, respectively, are independent Poisson processes with rates λp and $\lambda(1-p)$. Using Example 5.5.13, we know that $X(t)$ and $Y(t)$ are independent random variables with means λpt and $\lambda(1-p)t$, respectively. To conclude the proof, one needs to show that the increments of $X(t)$ and $Y(t)$ are independent. We leave you the details.

12.2.6 Saint Petersburg Paradox

The Saint Petersburg paradox holds for Poisson processes. In fact, the Poisson process seen from an arrival time is the Poisson process plus one point at that time.

12.2.7 Stationarity

Let $X = \{X(t), t \in \Re\}$ be a random process. We say that it is *stationary* if $\{X(t+u), t \in \Re\}$ has the same distribution for all $u \in \Re$. In other words, the statistics do not change over time. We cannot use this process to measure time. This would be the case of the weather if there were no long-term trend such as global warming or even seasonal effects. As an exercise, you can show that our reflected fortune process with $p < 0.5$ and started with $P(Y_0 = k) = \pi(k)$ is stationary. Also, the Poisson process is stationary.

12.2.8 Time reversibility

Let $X = \{X(t), t \in \Re\}$ be a random process. We say that it is *time-reversible* if $\{X(t), t \in \Re\}$ and $\{X(u-t), t \in \Re\}$ have the same distribution for all $u \in \Re$. In other words, the statistics do not change when we watch the movie in reverse.

You should verify that a time-reversible process is necessarily stationary.

Also, you should show that our reflected fortune process with $p < 0.5$ and started with $P(Y_0 = k) = \pi(k)$ for $k \geq 0$ is time-reversible. Also, the Poisson process is time-reversible.

12.2.9 Ergodicity

Roughly, a stochastic process is *ergodic* if statistics that do not depend on the initial phase of the process are constant. That is, such statistics do not depend on the realization of the process. For instance, if you simulate an ergodic process, you need only one simulation run; it is representative of all possible runs.

Let $\{X(t), t \in \Re\}$ be a random process. We compute some statistic $Z(\omega, u) = \phi(X(\omega, t+u), t \in \Re)$. That is, we perform calculations on the process starting at time u. We are interested in calculations such that $Z(\omega, u) = Z(\omega, 0)$ for all u. (We give examples shortly; don't worry.) Let us call such calculations "invariant" random variables of the process X. The process X is ergodic if all its invariant random variables are constant.

As an example, let $Z(\omega, u) = \lim_{T\to\infty}(1/T)\int_0^T X(u+t)dt$. This random variable is invariant. If the process is ergodic, then Z is the same for all ω: it is constant.

You should show that our reflected fortune process with $p \leq 0.5$ is ergodic. The trick is that this process must eventually go back to 0. One can then couple two versions of the process that start off independently and merge the first time they meet. Since they remained glued forever, the long-term statistics are the same. (See Example 14.8.14 for an illustration of a coupling argument.)

12.2.10 Markov

A random process $\{X(t), t \in \Re\}$ is Markov if, given $X(t)$, the past $\{X(s), s < t\}$ and the future $\{X(s), s > t\}$ are independent. Markov chains are examples of Markov process.

A process with independent increments is Markov. For instance, the Poisson process and the Brownian motion process are Markov.

The reflected fortune process is Markov.

For a simple example of a process that is not Markov, let Y_n be the reflected fortune process and define $W_n = 1\{Y_n < 2\}$.

It is not Markov because if you see that $W_{n-1} = 0$ and $W_n = 1$, then you know that $Y_n = 1$. Consequently, we can write

$$P[W_2 = 0 | W_1 = 1, W_0 = 0] = p.$$

However,

$$P[W_2 = 0 | W_1 = 1, W_0 = 1] < p.$$

Note that this example shows that a function of a Markov process may not be a Markov process.

12.2.11 Solved Problems

Example 12.2.1. *Let $\{N_t, t \geq 0\}$ be a Poisson process with rate λ.*

a. What is the p.m.f. (probability mass function) of N_1?

b. What is the p.m.f. of $N_2 - N_1$?

c. Calculate $L[N_2 \mid N_1, N_3]$.

d. Calculate the Maximum Likelihood Estimator of λ given $\{N_1, N_2, N_3\}$.

a. We know that $N_1 =_d P(\lambda)$, so that $P(N_1 = n) = \frac{\lambda^n}{n!}e^{-\lambda}$.

b. Same.

c. Let $U = N_1, V = N_2 - N_1, W = N_3 - N_2$. We want to calculate

$$L[U + V \mid U, U + V + W]$$

where the random variables U, V, W are i.i.d. $P(\lambda)$. We could use the straightforward approach, with the general formula. However, a symmetry argument turns out to be easier.

Note that, by symmetry

$$X := L[U + V \mid U, U + V + W] = L[U + W \mid U, U + V + W].$$

Consequently,

$$2X = L[(U + V) + (U + W) \mid U, U + V + W] = U + (U + V + W).$$

Hence,

$$X = L[N_2 \mid N_1, N_3] = \frac{1}{2}(U + (U + V + W)) = \frac{1}{2}(N_1 + N_3).$$

d. We are given the three i.i.d. $P(\lambda)$ random variables U, V, W defined earlier. Then

$$P[U = u, V = v, W = w \mid \lambda] = \frac{\lambda^{u+v+w}}{u!v!w!}e^{-3\lambda}.$$

Consequently, the MLE is the value of λ that maximizes

$$\lambda^n e^{-3\lambda},$$

with $n = u + v + w$. Taking the derivative with respect to λ and setting it to 0, we get

$$\lambda = \frac{n}{3}.$$

Hence,

$$MLE[\lambda \mid N_1, N_2, N_3] = \frac{N_3}{3}.$$

Example 12.2.2. *Let $\boldsymbol{N} = \{N_t, t \geq 0\}$ be a counting process. That is, $N_{t+s} - N_s \in \{0, 1, 2, \ldots\}$ for all $s, t \geq 0$. Assume that the process has independent and stationary increments. That is, assume that $N_{t+s} - N_s$ is independent of $\{N_u, u \leq s\}$ for all $s, t \geq 0$ and has a p.m.f. that does not depend on s. Assume further that $P(N_t > \mathbf{1}) = o(t)$. Show that $\boldsymbol{N}$ is a Poisson process. (Hint: Show that the times between jumps are exponentially distributed and independent.)*

Let

$$a(s) = P(N_{t+s} = N_t).$$

Then

$$a(s+u) = P(N_{t+s+u} = N_t) = P(N_{t+s+u} = N_{t+s})P(N_{t+s} = N_t) = a(s)a(u), s, u > 0.$$

Taking the derivative with respect to s, we find

$$a'(s+u) = a'(s)a(u).$$

At $s = 0$, this gives

$$a'(u) = a'(0)a(u).$$

The solution is $a(u) = a(0)\exp\{a'(0)u\}$, which shows that the distribution of the first jump time is exponentially distributed. The independence of the increments allows to conclude that the times between successive jumps are i.i.d. and exponentially distributed.

Example 12.2.3. *Construct a counting process* $\boldsymbol{N}$ *such that for all* $t > 0$ *the random variable* N_t *is Poisson with mean* λt *but the process is not a Poisson process.*

Assume that A_t is a Poisson process with rate λ. Let $F(t;x) = P(N_t \leq x)$ for $x, t \geq 0$. Let also $G(t;u)$ be the inverse function of $F(t;.)$. That is, $G(t;u) = min\{x \mid F(t;x) \geq u\}$ for $u \in [0,1]$. Then we know that if we pick ω uniformly distributed in $[0,1]$, the random variable $G(t;\omega)$ has p.d.f. $F(t;x)$, i.e., is Poisson with mean λt. Consider the process $\boldsymbol{N} = \{N_t(\omega) := G(t;\omega), t \geq 0\}$. It is such that N_t is Poisson with mean λt for all t. However, if you know the first jump time of N_t, then you know ω and all the other jump times. hence $\boldsymbol{N}$ is not Poisson.

Example 12.2.4. *Let* $\boldsymbol{N}$ *be a Poisson process with rate* λ *and define* $X_t = X_0(-1)^{N_t}$ *for* $t \geq 0$ *where* X_0 *is a* $\{-1,+1\}$*-valued random variable independent of* $\boldsymbol{N}$*.*

a. Does the process $\boldsymbol{X}$ *have independent increments?*

b. Calculate $P(X_t = 1)$ *if* $P(X_0 = 1) = p$*.*

c. Assume that $p = 0.5$*, so that, by symmetry,* $P(X_t = 1) = 0.5$ *for all* $t \geq 0$*. Calculate* $E(X_{t+s}X_s)$ *for* $s, t \geq 0$*.*

a. Yes, the process has independent increments. This can be seen by considering the following conditional probabilities:

$$\begin{aligned} P(X_{t+s} = 1 | X_t = 1, (X(u), 0 \leq u \leq s)) &= P(N(t, s+t] \text{ is even} | X_t = 1, (X(u), 0 \leq u \leq s)) \\ &= P(N(t, s+t] \text{ is even}) \end{aligned}$$

Similarly,

$$\begin{aligned} P(X_{t+s} = -1 | X_t = 1, (X(u), 0 \leq u \leq s)) &= P(N(t, s+t] \text{ is odd} | X_t = 1, (X(u), 0 \leq u \leq s)) \\ &= P(N(t, s+t] \text{ is odd}) \end{aligned}$$

Similarly, we can show independent increments for the case when $X(t) = -1$

b. First lets calculate $P(N(0,t] \; is \; even)$

$$\begin{aligned} P(N(0,t] \; is \; even) &= \sum_{i=0}^{\infty} \frac{(\lambda t)^{2i} e^{-\lambda t}}{2i!} \\ &= \frac{e^{-\lambda t}}{2}(\sum_{i=0}^{\infty} \frac{(\lambda t)^i}{i!} + \sum_{i=0}^{\infty} \frac{(-\lambda t)^i}{i!}) \\ &= e^{-\lambda t}(\frac{e^{-\lambda t} + e^{\lambda t}}{2}) \\ &= \frac{1 + e^{-2\lambda t}}{2} \end{aligned}$$

Similarly we get $P(N(0,t] \; is \; odd) = \frac{1-e^{-2\lambda t}}{2}$

$$\begin{aligned} P(X_t = 1) &= P(X_t = 1 | X_0 = 1) P(X_0 = 1) + P(X_t = 1 | X_0 = 0) P(X_0 = 0) \\ &= P(N(0,t] \; is \; even) p + P(N(0,t] \; is \; odd)(1-p) \\ &= \frac{1 + e^{-2\lambda t}}{2} p + \frac{1 - e^{-2\lambda t}}{2}(1-p) \end{aligned}$$

c. Assume that $p = 0.5$, so that, by symmetry, $P(X_t = 1) = 0.5$ for all $t \geq 0$. Calculate $E(X_{t+s} X_s)$ for $s, t \geq 0$.

$$\begin{aligned} E(X_{t+s} X_s) &= 1 \; P(X_{t+s} = X_s) + -1 \; P(X_{t+s} \neq X_s) \\ &= \frac{1}{2}(1 + e^{-2\lambda t}) - \frac{1}{2}(1 - e^{-2\lambda t}) \\ &= e^{-2\lambda t} \end{aligned}$$

Example 12.2.5.

Problem 5. Let $\boldsymbol{N}$ be a Poisson process with rate λ. At each jump time T_n of $\boldsymbol{N}$, a random number X_n of customers arrive at a cashier waiting line. The random variables $\{X_n, n \geq 1\}$ are i.i.d. with mean μ and variance σ^2. Let A_t be the number of customers who arrived by time t, for $t \geq 0$.

a. Calculate $E(A_t)$ and var(A_t).

b. What can you say about A_t/t as $t \to \infty$?

c. What can you say about

$$\frac{A_t - \mu\lambda t}{\sqrt{t}}$$

as $t \to \infty$?

a. We have, $E(A_t|N_t = n) = E(\sum_{i=1}^{n} X_i) = n\mu$ and

$$\begin{aligned} E(A_t) &= \sum_{i=1}^{\infty} E(A_t|N_t = n)P(N_t = n) \\ &= \mu \sum_{i=1}^{\infty} nP(N_t = n) \\ &= \mu\lambda t \end{aligned}$$

$$\begin{aligned} E(A_t^2|N_t = n) &= E(\sum_{i=1}^{n} X_i)^2 \\ &= \mu \sum_{i=1}^{n} X_i^2 + 2 \sum_{i=1, 0<j<i}^{n} X_i X_j \\ &= n(\mu^2 + \sigma^2) + n(n-1)\mu^2 \end{aligned}$$

$$\begin{aligned}
E(A_t^2) &= \sum_{i=1}^{\infty} E(A_t^2|N_t = n)P(N_t = n) \\
&= \sum_{i=1}^{\infty} (n(\mu^2 + \sigma^2) + n(n-1)\mu^2)P(N_t = n) \\
&= (\mu^2 + \sigma^2) \sum_{i=1}^{\infty} n\ P(N_t = n) + \mu^2 \sum_{i=1}^{\infty} n(n-1)\ P(N_t = n) \\
&= (\mu^2 + \sigma^2)\lambda t + \mu^2(\lambda t + (\lambda t)^2 - \lambda t) \\
&= (\mu^2 + \sigma^2)\lambda t + \mu^2(\lambda t)^2
\end{aligned}$$

Hence,

$$\begin{aligned}
Var(A_t) &= E(A_t^2) - E(A_t)^2 \\
&= (\mu^2 + \sigma^2)\lambda t + \mu^2(\lambda t)^2 - \mu^2(\lambda t)^2 \\
&= (\mu^2 + \sigma^2)\lambda t
\end{aligned}$$

b. We can write A_t as:

$$A_t = \frac{\sum_{i=1}^{N_t} X_i}{N_t}\ \frac{N_t}{t}$$

We see that as $t \to \infty, N(t) \to \infty$

Then, $\frac{N_t}{t}$ approaches the long term rate of the Poisson process (λ). Similarly, by the Strong law of Large numbers, $\frac{\sum_{i=1}^{N_t} X_i}{N_t} \overset{a.s}{\to} E[X_1] = \mu$

Hence,

$$lim_{t\to\infty} \frac{A_t}{t} = \mu\ \lambda$$

c. Define the RV Y_1 as the number of customers added in the first 1 sec. Similarly, define RV Y_i as the number of customers added in the i^{th} second. Then $(Y_n, n \geq 1)$ are iid RVs with mean $\mu\lambda$ and variance $(\mu^2 + \sigma^2)\lambda$.

Let $t_c = \lceil t \rceil$. Define $\Delta_c = t_c - t$ and let Z_c denote the number of customers arriving in time interval $(t, t_c]$. Then $E(Z_c) = \mu\lambda\Delta_c$ and $var(Z_c) = (\mu^2 + \sigma^2)\lambda\Delta_c$. Hence,

$$\begin{aligned} lim_{t\to\infty}\frac{A_t-\mu\lambda t}{\sqrt{t}} &= lim_{t_c\to\infty}\frac{\sum_{i=1}^{t_c}Y_i-\mu\lambda t_c-Z_c+\mu\lambda\Delta}{\sqrt{t_c-\Delta_c}} \\ &\geq lim_{t_c\to\infty}\frac{\sum_{i=1}^{t_c}Y_i-\mu\lambda t_c}{\sqrt{t_c}}-\frac{Z_c-\mu\lambda\Delta}{\sqrt{t_c}} \end{aligned}$$

The last term in the expression above goes to zero as $t \to \infty$. And using CLT we can show that $lim_{t_c\to\infty}\frac{\sum_{i=1}^{t_c}Y_i-\mu\lambda t_c}{\sqrt{t_c}} \xrightarrow{d} N(0,\lambda(\mu^2+\sigma^2))$

Similarly, we can show that $lim_{t\to\infty}\frac{A_t-\mu\lambda t}{\sqrt{t}}$ is bounded above by another RV which also has the distribution $N(0,\lambda(\mu^2+\sigma^2))$.

Hence, $lim_{t\to\infty}\frac{A_t-\mu\lambda t}{\sqrt{t}} \to N(0,\lambda(\mu^2+\sigma^2))$.

Chapter 13

Filtering Noise

When your FM radio is tuned between stations it produces a noise that covers a wide range of frequencies. By changing the settings of the bass or treble control, you can make that noise sound harsher or softer. The change in the noise is called filtering. More generally, filtering is a transformation of a random process. Typically, one filters a random process to extract desired information.

In this chapter, we define the power that a random process has at different frequencies and we describe how one can design filters that modify that power distribution across frequencies. We also explain how one can use a filter to compute the LLSE of the current value of some random process given a set of observations made so far.

We discuss the results in discrete time. That is, we consider sequences of values instead of functions of a continuous time. Digital filters operate on discrete sets of values. For instance, a piece of music is sampled before being processes, so that continuous time is made discrete. Moreover, for simplicity, we consider that time takes only a finite number of values. Physically, we can take a set of time that is large enough to represent an arbitrarily long time. As an example, we represent a song by a finite number of sample values.

13.1 Linear Time-Invariant Systems

The terminology is that a *system* transforms its input into an output. The input and output are sequences of values describing, for instance, a voltage as a function of time. Accordingly, a system maps functions to functions.

The functions that we consider are sequences such as $\{x(n), n = 0, 1, \ldots, N-1\}$. Such sequences represent signals that, for instance, encode voice or audio. For mathematical convenience, we extend these sequences periodically over all the integers in $\mathbb{Z} := \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. That is, given $\{x(n), n = 0, 1, \ldots, N-1\}$ we define $\{x(n), n \in \mathbb{Z}\}$ so that $x(n) = x(n-N)$ for all $n \in \mathbb{Z}$. We will consider systems whose input are the extension of the original finite sequence. For instance, imagine that we want to study how a particular song is processed. We consider that the song is played over and over and we look at the result of the processing. Intuitively, this artifice does not change the nature of the result, except possibly at the very beginning or end of the song. However, mathematically, this construction greatly simplifies the analysis

13.1.1 Definition

A system is linear if the output that corresponds to a sum of inputs is the sum of their respective outputs. A system is time-invariant if delaying the input only delays the output by the same amount.

Definition 13.1.1. Linear Time-Invariant (LTI) System

A *linear time-invariant system* (LTI) is the transformation of an input $\{x(n), n = 0, 1, \ldots, N-1\}$ into the output $\boldsymbol{y} = \{y(n), n \in \mathbb{Z}\}$ defined as follows:

$$y(n) = \sum_{n=0}^{N-1} h(m)x(n-m), n \in \mathbb{Z} \tag{13.1.1}$$

where the sequence $\{x(n), n \in \mathbb{Z}\}$ is the periodic extension of $\{x(n), n = 0, 1, \ldots, N-1\}$.

The function $\{h(n), n = 0, 1, \ldots, N-1\}$ is called the *impulse response* of the system. Thus, the output is the convolution of the (extended) input with the impulse response.

Roughly, the impulse response is the output that corresponds to the input $\{1, 0, \ldots, 0\}$. We say roughly, because the extension of this input is $x(n) = \sum_{k=-\infty}^{\infty} 1\{n = kN\}$, so that the corresponding output is

$$y(n) = \sum_{k=-\infty}^{\infty} h(n - kN).$$

For instance, imagine that $h(n) = 0$ for $n \notin \{0, 1, \ldots, N-1\}$. Then $y(n) = h(n)$ for $n \in \{0, 1, \ldots, N-1\}$. In that case, the impulse response is the output that corresponds to the impulse input $\{1, 0, \ldots, 0\}$. If $h(\cdot)$ is nonzero over a long period of time that exceeds the period N of the periodic repetition of the input, then the outputs of these different repetitions superpose each other and the result is complicated. In applications, one chooses N large enough to exceed the duration of the impulse response. Assuming that $h(n) = 0$ for $n < 0$ means that the system does not produce an output before the input is applied. Such a system is said to be causal. Obviously, all physical systems are causal.

Consider a system that delays the input by k time units. Such a system has impulse response $h(n) = 1\{n - k\}$. Similarly, a system which averages k successive input values and produces $y(n) = (x(n) + x(n-1) + \cdots + x(n-k+1))/k$ for some $k < N$ has impulse response

$$h(n) = \frac{1}{k} 1\{0 \le n \le k-1\}, n = 0, \ldots, N-1.$$

As another examples, consider a system described by the following identities:

$$y(n) + a_1 y(n-1) + \cdots + a_k y(n-k) = b_0 x(n) + b_1 x(n-1) + \cdots + b_m x(n-m), m \in \mathbb{Z}. \quad (13.1.2)$$

We assume $m < N - 1$. This system is LTI because the identities are linear and their coefficients do not depend on n. To find the impulse response, assume that $x(n) = 1\{n = 0\}$ and designate by $h(n)$ the corresponding output $y(n)$ when $y(n) = 0$ for $n < 0$. For instance,

by considering (13.1.2) successively for $n = 0, 1, 2, \ldots$:

$$\begin{aligned} h(0) &= b_0; \\ h(1) + a_1 h(0) &= b_1, \text{ so that } h(1) = b_1 - a_1 b_0; \\ h(2) + a_1 h(1) + a_2 h(0) &= b_2, \text{ so that } h(2) = b_2 - a_1(b_1 - a_1 b_0) - a_2 b_0; \end{aligned}$$

and so on.

We want to analyze the effect of such LTI systems on random processes. Before doing this, we introduce some tools to study the systems in the frequency domain.

13.1.2 Frequency Domain

LTI systems are easier to describe in the "frequency domain" than in the "time domain." To understand these ideas, we need to review (or introduce) the notion of frequencies.

As a preliminary, assume that $x(n) = e^{i2\pi un} = \beta^{-nu}$ for $n = 0, 1 \ldots, N-1$ and for a fixed $u \in \{0, 1, \ldots, N-1\}$. Then (13.1.1) implies that

$$\begin{aligned} y(n) &= \sum_{n=0}^{N-1} h(m)x(n-m) = \sum_{n=0}^{N-1} h(m)\beta^{-(n-m)u} \\ &= \beta^{-nu} \sum_{n=0}^{N-1} h(m)\beta^{mu} = x(n)H(u) = x(n)|H(u)|e^{i\theta(u)} \end{aligned}$$

where $H(u) = |H(u)|e^{i\theta(u)} = \sum_{m=0}^{N-1} h(m)\beta^{mu}$. Thus, if the input is a complex sine wave, then so is the output. Using complex-valued signals is a mathematical artefact that simplifies the analysis. In a physical system, the quantities are real. For instance, $h(n)$ is real. If we take the imaginary part of the expression above, we find that

$$\mathrm{Im}\{y(n)\} = \mathrm{Im}\{\sum_{n=0}^{N-1} h(m)x(n-m)\} = \mathrm{Im}\{x(n)|H(u)|e^{i\theta(u)}\}.$$

This gives

$$\begin{aligned} \sum_{n=0}^{N-1} h(m)\,\mathrm{Im}\{x(n-m)\} &= \sum_{n=0}^{N-1} h(m)\sin(2\pi u(n-m)/N) \\ &= \mathrm{Im}\{e^{i2\pi nu/N}|H(u)|e^{i\theta(u)}\} = |H(u)|\sin(2\pi nu/N + \theta(u)). \end{aligned}$$

This expression shows that if the input is the sine wave $\sin(2\pi un/T)$, then so is the output, except that its amplitude is multiplied by some gain $|H(u)|$ and that it is delayed. Note that the analysis is easier with complex functions. One can then take the imaginary part to recover the output to a sine wave.

The example above motivates the following definition.

Definition 13.1.2. Discrete Fourier Transform (DFT)

The *Discrete Fourier Transform* (DFT) of $\{x(n), n = 0, 1, \ldots, N-1\}$ is the sequence of complex number $\{X(u), u = 0, 1, \ldots, N-1\}$ where

$$X(u) = \sum_{n=0}^{N-1} x(n)\beta^{nu}, u = 0, 1, \ldots, N-1 \tag{13.1.3}$$

with $\beta := e^{-i2\pi/N}$.

It turns out that one can recover $x(\cdot)$ from $X(\cdot)$ by computing the *Inverse Discrete Fourier Transform*:

$$x(n) = \frac{1}{N}\sum_{u=0}^{N-1} X(u)\beta^{-nu}, n = 0, 1, \ldots, N-1. \tag{13.1.4}$$

Indeed,

$$\frac{1}{N}\sum_{u=0}^{N-1} X(u)\beta^{-nu} = \frac{1}{N}\sum_{u=0}^{N-1}[\sum_{m=0}^{N-1} x(m)\beta^{mu}]\beta^{-nu} = \sum_{m=0}^{N-1} x(m)[\frac{1}{N}\sum_{u=0}^{N-1}\beta^{u(m-n)}].$$

But,

$$\frac{1}{N}\sum_{u=0}^{N-1}\beta^{u(m-n)} = 1, \text{ for } m = n$$

and, for $m \neq n$,

$$\frac{1}{N}\sum_{u=0}^{N-1}\beta^{u(m-n)} = \frac{1}{N}\frac{\beta^{N(m-n)}-1}{\beta^{m-n}-1} = 0,$$

because $\beta^N = e^{-2i\pi} = 1$. Hence,

$$\frac{1}{N}\sum_{u=0}^{N-1} X(u)\beta^{-nu} = \sum_{m=0}^{N-1} x(m)1\{m = n\} = x(n),$$

as claimed.

The following result shows that LTI systems are easy to analyze in the frequency domain.

Theorem 13.1.1. *Let x and y be the input and output of an LTI system, as in (13.1.1). Then*

$$Y(u) = H(u)X(u), u = 0, 1, \ldots, N-1. \tag{13.1.5}$$

The theorem shows that the convolution (13.1.1) in the time domain becomes a multiplication in the frequency domain.

Proof:

We find

$$\begin{aligned} Y(u) &= \sum_{n=0}^{N-1} y(n)\beta^{nu} = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} h(m)x(n-m)\beta^{nu} \\ &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} h(m)x(n-m)\beta^{(n-m)u}\beta^{mu} = \sum_{m=0}^{N-1} h(m)\beta^{mu}[\sum_{n=0}^{N-1} x(n-m)\beta^{(n-m)u}] \\ &= \sum_{m=0}^{N-1} h(m)\beta^{mu}X(u) = H(u)X(u). \end{aligned}$$

The next-to-last identity follows from the observation that $\sum_{n=0}^{N-1} x(n-m)\beta^{(n-m)u}$ does not depend on m because both $x(n)$ and β^{nu} are periodic with period N. □

As an example, consider the LTI system (13.1.2). Assume that the input is $x(n) = e^{i2\pi un} = \beta^{-nu}$ for $n \in \mathbb{Z}$. We know that $y(n) = H(u)\beta^{-nu}, n \in \mathbb{Z}$ for some $H(u)$ that we calculate next. Using (13.1.2) we find

$$H(u)\beta^{-nu} + a_1 H(u)\beta^{-(n-1)u} + \cdots + a_k H(u)\beta^{-(n-k)u} = b_0\beta^{-nu} + b_1\beta^{-(n-1)u} + \cdots + b_m\beta^{-(n-m)u}.$$

Hence, after dividing both sides by β^{-nu},

$$H(u)[1 + a_1\beta^u + \cdots + a_k\beta^{ku}) = b_0 + b_1\beta^u + \cdots + \beta^{mu}.$$

Consequently,

$$H(u) = \frac{b_0 + b_1\beta^u + \cdots + \beta^{mu}}{1 + a_1\beta^u + \cdots + a_k\beta^{ku}}. \tag{13.1.6}$$

As an example, consider the moving average filter with

$$y(n) = \frac{1}{m}(x(n) + x(n-1) + \cdots + x(n-m+1)).$$

Here, $a_1 = \cdots = a_m = 0$ and $b_0 = \cdots = b_{m-1} = 1/m$. Consequently,

$$H(u) = \frac{1}{m}(1 + \beta^u + \cdots + \beta^{(m-1)u}) = \frac{1}{m}\frac{1-\beta^{mu}}{1-\beta^u} = \frac{1}{m}\frac{1-e^{-i2\pi mu/N}}{1-e^{-i2\pi u/N}}. \tag{13.1.7}$$

13.2 Wide Sense Stationary Processes

We are interested in analyzing systems when their inputs are random processes. One could argue that most real systems are better modelled that way than by assuming that the input is purely deterministic. Moreover, we are concerned by the power of the noise, as we see that noise as a disturbance that clouds our view of a signal of interest.

The power of a random process has to do with the average value of its squared magnitude. This should not be surprising. The power of an electrical signal is the product of its voltage by its current. For a given load, the current is proportional to the voltage, so that that the power is proportional to the square of the voltage. If the signal fluctuates, the power is proportional to the long term average value of the square of the voltage. If the law of large numbers applies (say, if the process is ergodic), then this long term average value is the expected value. This discussion motivates the *definition* of power as the expected value of the square.

We introduce a few key ideas on a simple model. Assume the the input of the system at time n is the random variable $X(n)$ that the output is $Y(n) = X(n) + X(n-1)$. The power of the input is $E(X(n)^2)$. For this quantity to be well-defined, we assume that it does not depend on n. The average power of the output is $E(Y(n;\omega)) = E((X(n) + X(n-1))^2) = E(X(n)^2) + 2E(X(n)X(n-1)) + E(X(n-1)^2)$. For this quantity to be well-defined, we assume that $E(X(n)X(n-1))$ does not depend on n. More generally, this

example shows that the definitions become much simpler if we consider processes such that $E(X(n)X(n-m))$ does not depend on n for $m \geq 0$. These considerations motivate the following definition. For convenience, we consider complex-valued processes.

Definition 13.2.1. Wide Sense Stationary

The random process $\{X(n), n \in \mathbb{Z}\}$ is *wide sense stationary* (wss) if

$$E(X(n)) = \mu, n \in \mathbb{Z} \tag{13.2.1}$$

and

$$E(X(n)X^*(n-m)) = R_X(m), n, m \in \mathbb{Z} \tag{13.2.2}$$

As we did in the deterministic case, we will extend periodically a random sequence $\{X(n), n = 0, \ldots, N-1\}$ into a sequence defined for $n \in \mathbb{Z}$. For the extended sequence to be wss, the original sequence must satisfy a related condition that $E(X(n)X^*(n-m)) = R_X(m)$ for $n, m \in \{0, \ldots, N-1\}$ with the convention that $n-m$ is replaced by $n-m+N$ if $n-m < 0$. When this condition holds, we say that the sequence $\{X(0), \ldots, X(N-1)\}$ is wss. Note also in that case that $R_X(n)$ is periodic with period N. Although this condition may seem somewhat contrived, we see below that it is satisfied for many processes.

One first example is when $\{X(n), n = 0, \ldots, N-1\}$ are i.i.d. with mean μ and variance σ^2. In that case, one finds $R_X(m) = (\mu^2 + \sigma^2)1\{m = 0\}$ for $m = 0, 1, \ldots, N-1$. The sequence $R_X(n)$ for $n \in \mathbb{Z}$ is the periodic extension of $\{R_X(0), \ldots, R_X(N-1)\}$.

The following result explains how one can generate many examples of wss process.

Theorem 13.2.1. *Assume that the input of a LTI system with impulse response* $\{h(n), n = 0, 1, \ldots, N-1\}$ *is the wss process* $\{X(n), n = 0, \ldots, N-1\}$. *Then the output* $\{Y(n), n \in \mathbb{Z}\}$ *is wss.*

Proof::

First, we notice that

$$E(Y(n)) = E(\sum_{m=0}^{N-1} h(m)X(n-m)) = \sum_{m=0}^{N-1} h(m)E(X(n-m)) = \mu \sum_{m=0}^{N-1} h(m), n \in \mathbb{Z}.$$

Second, we calculate

$$\begin{aligned} E(Y(n)Y^*(n-m)) &= E(\sum_{k=0}^{N-1} h(k)X(n-k) \sum_{k'=0}^{N-1} h(k')X^*(n-m-k') \\ &= \sum_{k=0}^{N-1}\sum_{k'=0}^{N-1} h(k)h(k')R_X(m+k'-k), \end{aligned}$$

which shows that the result does not depend on n. We designate that result as $R_Y(m)$. Note also that $R_Y(m)$ is periodic in m with period N, since R_X has that property.

We take note of the result of the calculation above, for further reference:

$$R_Y(m) = \sum_{k=0}^{N-1}\sum_{k'=0}^{N-1} h(k)h(k')R_X(m+k'-k). \tag{13.2.3}$$

□

13.3 Power Spectrum

The result of the calculation in (13.2.3) is cumbersome and hard to interpret. We introduce the notion of power spectrum which will give us a simpler interpretation of the result.

Definition 13.3.1. Let X be a wss process. We define the *power spectrum* of the process X as $\{S_X(u), u = 0, \ldots, N-1\}$, the DFT of R_X. That is,

$$S_X(u) = \sum_{n=0}^{N-1} R_X(n)\beta^{nu}, u = 0, 1, \ldots, N-1 \tag{13.3.1}$$

with $\beta := e^{-i2\pi/N}$.

We look at a few examples to clarify the meaning of the power spectrum.

First consider the random process

$$X(n) = \beta^{nu_0} e^{i\Theta}, n = 0, 1, \ldots, N-1$$

where $\Theta =_D U[0, 2\pi]$.

This process is a sine wave with frequency u_0 and a random phase. (The unit of frequency is $1/N$.) Without the random phase, the process would certainly not be wss since its mean would depend on n. However, as defined, this process is wss. Indeed,

$$E(X(n)) = \beta^{nu_0} E(e^{i\Theta}) = 0, \text{ by symmetry,}$$

and

$$E(X(n)X^*(n-m)) = \beta^{-mu_0} =: R_X(m).$$

We find that

$$S_X(u) = N1\{u = u_0\}, u = 0, 1, \ldots, N-1. \tag{13.3.2}$$

Indeed, by the identity (13.1.4) for the inverse DFT, we see that

$$R_X(m) = \frac{1}{N} \sum_{u=0}^{N-1} S_X(u)\beta^{-mu}, n = 0, 1, \ldots, N-1 \tag{13.3.3}$$

which agrees with the expression (13.3.2) for S_X. This expression (13.3.2) says that the power spectrum of the process is concentrated on frequency u_0, which is consistent with the definition of $X(n)$ as a sine wave at frequency u_0.

Second, we look at the process with i.i.d. random variables $\{X(n), n = 0, \ldots, N-1\}$. We saw earlier that $R_X(u) = (\mu^2 + \sigma^2)1\{n = 0\}$ for $n = 0, 1, \ldots, N-1$. Hence,

$$S_X(u) = \sum_{n=0}^{N-1} R_X(n)\beta^{nu} = \sum_{n=0}^{N-1} (\mu^2 + \sigma^2)1\{n = 0\}\beta^{nu} = \mu^2 + \sigma^2, u = 0, 1, \ldots, N-1.$$

This process has a constant power spectrum. A process with that property is said to be a *white noise*. In a sense, it is the opposite of a pure sine wave with a random phase.

13.4 LTI Systems and Spectrum

The following theorem is the central result about filtering.

Theorem 13.4.1. *Let X be a wss input of a LTI system with impulse response $h(\cdot)$ and output Y. Then*

$$S_Y(u) = |H(u)|^2 S_X(u), u = 0, 1, \ldots, N-1. \tag{13.4.1}$$

Proof:

We use (13.2.3) to find

$$\begin{aligned} S_Y(u) &= \sum_{m=0}^{N-1} R_Y(m)\beta^{mu} \\ &= \sum_{m=0}^{N-1}[\sum_{k=0}^{N-1}\sum_{k'=0}^{N-1} h(k)h(k')R_X(m+k'-k)]\beta^{mu} \\ &= \sum_{m=0}^{N-1}\sum_{k=0}^{N-1}\sum_{k'=0}^{N-1} h(k)h(k')R_X(m+k'-k)\beta^{ku}\beta^{-k'u}\beta^{(m+k'-k)u} \\ &= \sum_{k=0}^{N-1} h(k)\beta^{ku}\{\sum_{k'=0}^{N-1} h(k')\beta^{-k'u}[\sum_{m=0}^{N-1} R_X(m+k'-k)\beta^{(m+k'-k)u}]\}. \end{aligned}$$

Now,

$$\sum_{m=0}^{N-1} R_X(m+k'-k)\beta^{(m+k'-k)u} = S_X(u)$$

because both $R_X(m+k'-k)$ and $\beta^{(m+k'-k)u}$ are periodic in $N-1$, so that we can set $k'-k=0$ in the calculation without changing the result. Hence,

$$S_Y(u) = \sum_{k=0}^{N-1} h(k)\beta^{ku}\{\sum_{k'=0}^{N-1} h(k')\beta^{-k'u}S_X(u)\} = H(u)H^*(u)S_X(u) = |H(U)|^2 S_X(u),$$

as claimed. □

The theorem provides a way to interpret the meaning of the power spectrum. Imagine a wss random process X. We want to understand the meaning of $S_X(u)$. Assume that we build an ideal LTI system with transfer function $H(u) = \sqrt{N}1\{u = u_0\}$. This filter lets

only the frequency u_0 go through. The process X goes through the LTI system. The power of the output, $E(Y_n^2)$ is calculated as follows: (we use (13.3.3))

$$E(|Y_n|^2) = R_Y(0) = \frac{1}{N}\sum_{u=0}^{N-1} S_Y(u) = \frac{1}{N}\sum_{u=0}^{N-1} |H(u)|^2 S_X(u) = S_X(u_0).$$

Thus, $S_X(u_0)$ measures the power of the random process X at frequency u_0.

13.5 Solved Problems

Example 13.5.1. *Averaging a process over time should make it smoother and reduce its rapid changes. Accordingly, we expect that such processing should cut down the high frequencies. Verify that fact.*

We use (13.1.7) to find that

$$|H(u)|^2 = |\frac{1}{m}\frac{1-e^{-i2\pi mu/N}}{1-e^{-i2\pi u/N}}|^2 = \frac{1}{m^2}\frac{\sin^2(2\pi mu/N) + (1-\cos(2\pi mu/N))^2}{\sin^2(2\pi u/N) + (1-\cos(2\pi u/N))^2}$$
$$= \frac{1}{m^2}\frac{1-\cos(2\pi mu/N)}{1-\cos(2\pi u/N)}.$$

Figure 13.1 plots the value of $|H(u)|^2$. The figure shows the "low-pass" filtering effect: the low frequencies go through but the high frequencies are greatly attenuated. This plot is the power spectrum of the output when the input is a white noise such as an uncorrelated sequence. That output is a "colored" noise with most of the power in the low frequencies.

Example 13.5.2. *By calculating the differences between successive values of a process one should highlight its high-frequency components. Verify that fact.*

We consider the LTI system

$$y(n) = x(n) - x(n-1).$$

Using (13.1.7) we find that its transfer function $H(u)$ is given by

$$H(u) = 1 - \beta^u = 1 - e^{i2\pi u/N}.$$

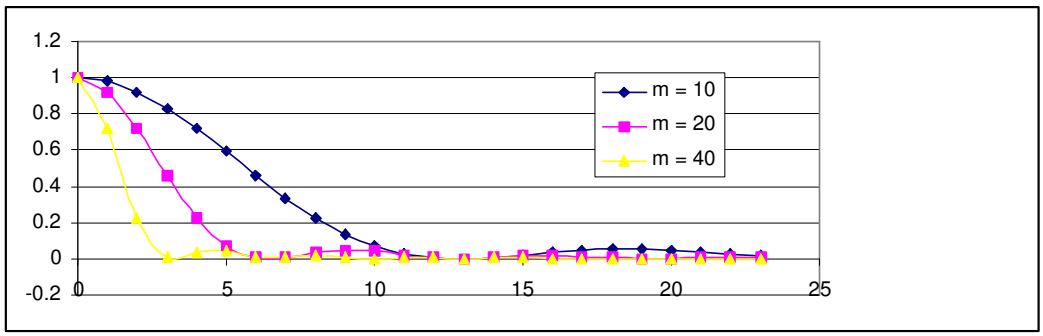


Figure 13.1: $|H(u)|^2$ for Moving Average.

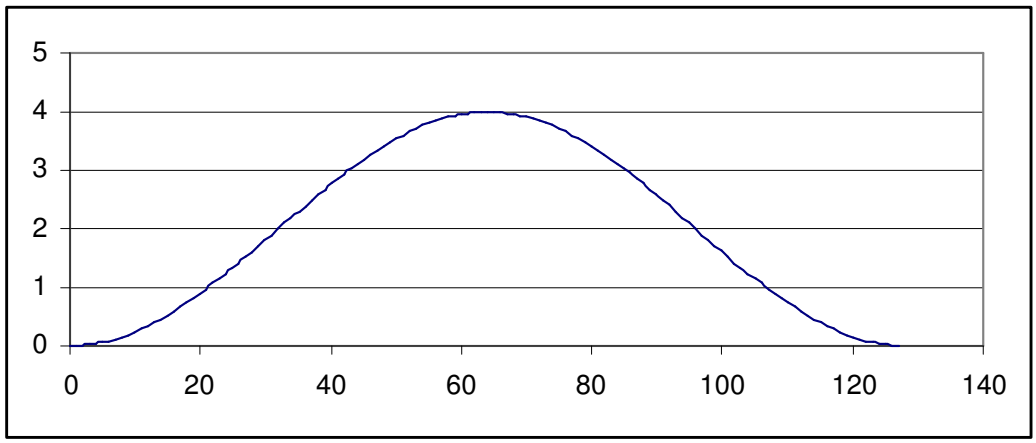


Figure 13.2: $|H(u)|^2$ for Difference Filter.

Consequently,

$$|H(u)|^2 = (1 - \cos(2\pi u/N))^2 + \sin^2(2\pi u/N) = 2 - 2\cos(2\pi u/N) = 4\sin^2(\pi u/N).$$

Figure 13.2 plots that expression and shows that the filter boosts the high frequencies, as expected. You will note that this system acts as a high-pass filter for frequencies up to 64 (in this example, $N = 128$). In practice, one should choose N large enough so that the filter covers most of the range of frequencies of the input process. Thus, if the power spectrum of the input process is limited to K, one can choose $N = 2K$ and this system will act as a high-pass filter.

Chapter 14

Markov Chains - Discrete Time

A Markov chain models the random motion in time of an object in a countable set. The key feature of that motion is that the object has no memory of its past and does not carry a watch. That is, the future motion depends only on the current location. Consequently, the law of motion is specified by the one-step transition probabilities from any given location. Markov chains are an important class of models because they are fairly general and good numerical techniques exist for computing statistics about Markov chains.

14.1 Definition

A discrete-time Markov chain is a process $X = \{X_n, n \geq 0\}$ that takes values in a countable set S and is such that

$$P[X_{n+1} = j | X_0, \ldots, X_{n-1}, X_n = i] = P(i, j) \text{ for all } i, j \in S \text{ and } n \geq 0.$$

The matrix $P = [P(i, j), i, j \in S]$ is the *transition probability matrix* of the Markov chain. The matrix P is any nonnegative matrix whose rows sum to 1. Such a matrix is called a *stochastic matrix*.

The finite dimensional distributions of X are specified by the initial distribution $\pi_0 = \{\pi_0(i) = P(X_0 = i), i \in S\}$ and by P. Indeed,

$$
\begin{aligned}
&P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) \\
&\quad = P(X_0 = i_0)P[X_1 = i_1|X_0 = i_0]P[X_2 = i_2|X_0 = i_0, X_1 = i_1] \\
&\qquad \times \cdots \times P[X_n = i_n|X_0 = i_0, \ldots, X_{n-1} = i_{n-1}] \\
&\quad = P(X_0 = i_0)P[X_1 = i_1|X_0 = i_0]P[X_2 = i_2|X_1 = i_1] \cdots P[X_n = i_n|X_{n-1} = i_{n-1}] \\
&\quad = \pi_0(i_0)P(i_0, i_1)P(i_1, i_2) \cdots P(i_{n-1}, i_n).
\end{aligned}
$$

Note in particular (by summing over $i_1, i_2, \ldots, i_{n-1}$) that

$$P[X_n = j|X_0 = i] = P^n(i, j) \text{ and } P(X_n = j) = \pi_0 P^n(j)$$

where P^n is the n-th power of the matrix P and π_0 is the row vector with entries $\pi_0(j)$. (The power of a stochastic matrix is defined as that of a finite matrix.)

A *state transition diagram* can represent the transition probability matrix. Such a diagram shows the states and the probabilities are represented by numbers on arrows between states. By convention, no arrow between two states means that the corresponding transition probability is 0. (See examples below.)

14.2 Examples

We first look at the following state transition diagram.

Diagram 14.1 represents a Markov chain with $S = \{0, 1\}$. Its transition probability matrix is shown next to the diagram.

The following state transition diagram 14.2 corresponds to a Markov chain with $S = \{0, 1, 2, \ldots\}$.

This is the state transition diagram of the sequence of fortunes with the rich uncle.

Also, for future reference, we introduce a few other examples. Markov chain 14.3 has

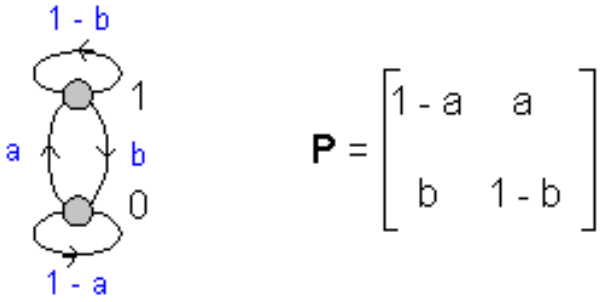


Figure 14.1: Markov chain with two states.

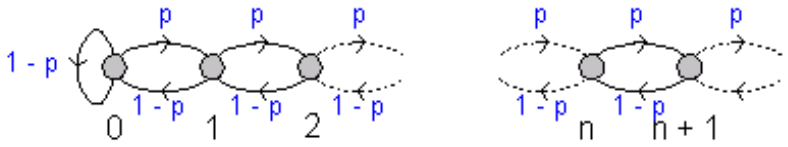


Figure 14.2: State transitions diagram of reflected random walk.

two sets of states that do not communicate. (Recall that the absence of an arrow means that the corresponding transition probability is 0.)

Markov chain 14.4 has one state, state 4, that cannot be exited from. Such a state is said to be *absorbing*. Note that $P(4,4) = 1$.

Markov chains 14.5 and 14.6 have characteristics that we will discuss later.

Consider the following "non-Markov" chain. The possible values are $0, 1, 2$. The initial

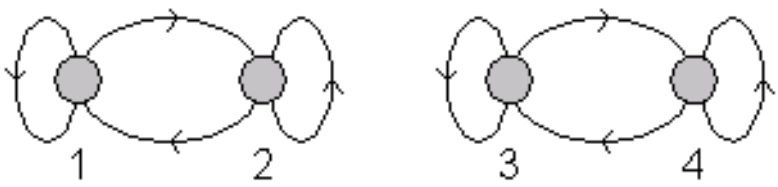


Figure 14.3: Markov chain with two disconnected sets of states.

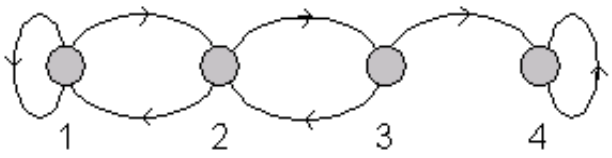


Figure 14.4: Markov chain with absorbing state.

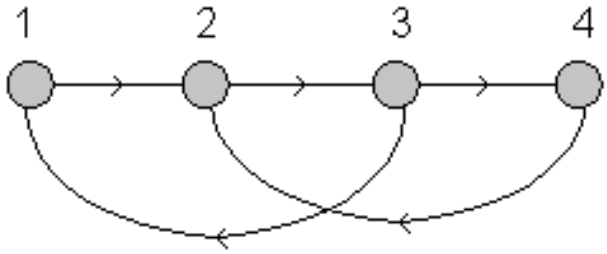


Figure 14.5: Periodic Markov chain.

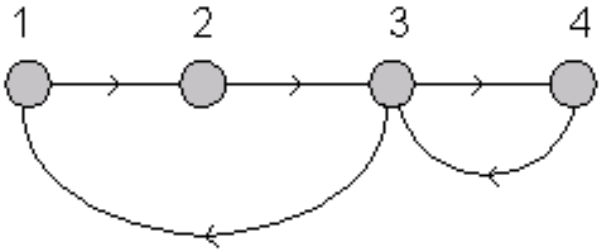


Figure 14.6: Aperiodic Markov chain.

value is picked randomly. Also, with probability 1/2, the sequence increases (modulo 3), otherwise, it decreases. Thus, with probability $\pi(0)/2$, the sequence is $\{0, 1, 2, 0, 1, 2, \ldots\}$ and with probability $\pi(0)/2$ it is $\{0, 2, 1, 0, 2, 1, 0, \ldots\}$. Similarly for the other two possible starting values. This is not Markov since by looking at the previous two values you can predict exactly the next one, which you cannot do if you only see the current value. Note that you can "complete the state" by considering the pair of two successive values. This pair is a Markov chain. More generally, if a sequence has a finite memory of duration k, the vector of k successive values is Markov.

Here is another example. Let $X_n = (X_0 + n) mod 3$ where X_0 is uniformly distributed in $\{0, 1, 2\}$. That is, if $X_0 = 0$, then $(X_n, n \geq 0) = (0, 1, 2, 0, 1, 2, \ldots)$ whereas if $X_0 = 1$, then $(X_n, n \geq 0) = (1, 2, 0, 1, 2, \ldots)$, and similarly if $X_0 = 2$. Let $g(0) = g(1) = 5$ and $g(2) = 6$. Then $\{Y_n = g(X_n), n \geq 0\}$ is not a Markov chain. Indeed,

$$P[Y_2 = 6 \mid Y_1 = 5, Y_0 = 5] = 1 \neq P[Y_2 = 6 \mid Y_1 = 5] = \frac{1}{2}.$$

14.3 Classification

The properties of Markov chains are determined largely (completely for finite Markov chains) by the "topology" of their state transition diagram. We need some terminology.

A Markov chain (or its probability transition matrix) is said to be *irreducible* if it can reach every state from every other state (not necessarily in one step). For instance, the Markov chains in Figures 14.1, 14.2, 14.5, and 14.6 are irreducible but those in Figures 14.3 and 14.4 are not.

Define $d(i) = \text{g.c.d.}\{n > 0 |$ it is possible to go from i to i in n steps $\}$. That is,

$$d(i) = \text{g.c.d.}\{n > 0 | P^n(i, i) > 0\}.$$

Here, g.c.d. means the greatest common divisor of the integers in the set. For instance, $\text{g.c.d.}\{6, 9, 15\} = 3$ and $\text{g.c.d.}\{12, 15, 25\} = 1$. For instance, for the Markov chain in Figure

14.1, $d(1) = \text{g.c.d.}\{1, 2, 3, \ldots\} = 1$. For Figure 14.2, $d(2) = \text{g.c.d.}\{2, 4, 5, 6, \ldots\} = 1$. For Figure 14.5, $d(1) = \text{g.c.d.}\{3, 6, 9, \ldots\} = 3$. For Figure 14.6, $d(1) = \text{g.c.d.}\{3, 5, 6, \ldots\} = 1$.

If the Markov chain is irreducible, then it can be shown that $d(i)$ has the same value for all $i \in S$. If this common value d is equal to 1, then the Markov chain is said to be *aperiodic*. Otherwise, the Markov chain is said to be *periodic* with period d. Accordingly, the Markov chains [1], [2], [6] are aperiodic and Markov chain [5] is periodic with period 3.

Define $T_i = \min\{n \geq 0 | X_n = i\}$. If the Markov chain is irreducible, then one can show that $P[T_i < \infty | X_0 = i]$ has the same value for all $i \in S$. Moreover, that value is either 1 or 0. If it is 1, the Markov chain is said to be *recurrent*. If it is 0, then the Markov chain is said to be *transient*.

Moreover, if the irreducible Markov chain is recurrent, then $E[T_i | X_0 = i]$ is either finite for all $i \in S$ or infinite for all $i \in S$. If $E[T_i | X_0 = i]$ is finite, then the Markov chain is said to be *positive recurrent*. Every finite irreducible Markov chain is positive recurrent. If $E[T_i | X_0 = i]$ is infinite, then the Markov chain is *null recurrent*. Also, one can show that

$$\lim_{N \to \infty} \frac{1}{n}(1\{X_1 = j\} + 1\{X_2 = j\} + \cdots + 1\{X_N = j\}) = 0$$

for all j if the Markov chain is null recurrent and

$$\lim_{N \to \infty} \frac{1}{n}(1\{X_1 = j\} + 1\{X_2 = j\} + \cdots + 1\{X_N = j\}) =: \pi(j) > 0$$

for all j if the Markov chain is positive recurrent.

Finally, if the Markov chain is irreducible, aperiodic, and positive recurrent, then

$$P[X_n = j | X_0 = i] \to \pi(j)$$

for all $i, j \in S$, as $n \to \infty$. The Markov chain is said to be *asymptotically stationary*.

The Markov chain in Figure 14.2 is transient when $p > 0.5$, null recurrent when $p = 0.5$, and positive recurrent when $p < 0.5$.

14.4 Invariant Distribution

If $P(X_n = i) = \pi(i)$ for $i \in S$ (i.e., does not depend on n), the distribution π is said to be *invariant*. Since

$$\begin{aligned} P(X_{n+1} = i) &= \sum_j P(X_n = j, X_{n+1} = i) \\ &= \sum_j P[X_{n+1} = i | X_n = j] P(X_n = j) = \sum_j P(X_n = j) P(j, i), \end{aligned}$$

if π is invariant, then

$$\pi(i) = \sum_j \pi(j) P(j, i), \text{ for } i \in S.$$

These identities are called the *balance equations*. Thus, a distribution is invariant if and only if it satisfies the balance equations.

An irreducible Markov chain has at most one invariant distribution. It has one if and only if it is positive recurrent. In that case, the Markov chain is ergodic and asymptotically stationary in that the distribution of X_n converges to the stationary distribution as we saw in the previous section.

Examples of calculations of stationary distribution abound.

The following theorem summarizes the discussion of the previous two sections.

Theorem 14.4.1. *Consider an irreducible Markov chain. It is either transient, null recurrent, or positive recurrent. Only the last case is possible for a finite-state Markov chain.*

If the Markov chain is transient or null recurrent, then it has no invariant distribution; the fraction of time it is in state j converges to zero for all j, and the probability that it is in state j converges to 0 for all j.

If the Markov chain is positive recurrent, it has a unique invariant distribution π. The fraction of time that the Markov chain is in state j converges to $\pi(j)$ for all j. If the Markov chain is aperiodic, then the probability that it is in state j converges to $\pi(j)$ for all j.

14.5 First Passage Time

We can extend the example of the gambler's ruin to a general Markov chain. For instance, let X be a Markov chain on S with transition probability matrix P. Let also $A \subset S$ be a given subset and $T = \min\{n \geq 0 | X_n \in A\}$ and define $\beta(i) = E[T|X_0 = i]$. Then one finds that

$$\beta(i) = 1 + \sum_j P(i,j)\beta(j), \text{ for } i \notin A.$$

Of course, $\beta(i) = 0$ for $i \in A$. In finite cases, these equations suffice to determine $\beta(i)$. In infinite cases, one may have to introduce a boundary as we did in the case of the reflected fortune process. In many cases, no simple solution can be found. These are the first step equations for the first passage time.

14.6 Time Reversal

Assume that X is a stationary irreducible Markov chain with invariant distribution π and transition probability matrix P on S. What does the time-reversed process $X' = \{X'(n) := X_{N-n}, n \geq 0\}$ look like? It turns out that X' is also a stationary Markov chain with the same invariant distribution (obviously) and with transition probability matrix P' given by

$$P'(i,j) = \frac{\pi(i)P(i,j)}{\pi(j)}, i, j \in S.$$

In some cases, $P = P'$. In those cases, X and X' have the same finite dimensional distributions (are indistinguishable statistically) and X is said to be *time-reversible*. Thus, X is time-reversible if and only if it is stationary and

$$\pi(i)P(i,j) = \pi(j)P(j,i), i, j \in S.$$

These equations are called the *detailed balance equations*. You can use these equations to verify that the stationary reflected fortune process is time-reversible.

14.7 Summary

Recall that a sequence of random variables $\boldsymbol{X} = \{X_n, n \geq 0\}$ taking values in a countable set $\boldsymbol{S}$ is a Markov chain if

$$P[X_{n+1} = j \mid X_n = i, X_m, m \leq n-1] = P(i,j), \forall i, j \in \boldsymbol{S}, n \geq 0.$$

The key point of this definition is that, given the present value of X_n, the future $\{X_m, m \geq n+1\}$ and the past $\{X_m, m \leq n-1\}$ are independent. That is, the evolution of $\boldsymbol{X}$ starts afresh from X_n. In other words, the *state* X_n contains all the information that is useful for predicting the future evolution.

The *First Step Equations* are difference equations about some statistics of a Markov chain $\{X_n, n \geq 0\}$ that are derived by considering the different possible values of the first step, i.e., for X_1.

Finally, a stationary Markov chain reversed in time is again a Markov chain, generally with a different transition probability matrix, unless the detailed balanced equations hold, in which case the Markov chain is time-reversible.

14.8 Solved Problems

Example 14.8.1. *We flip a coin repeatedly until we get three successive heads. What is the average number of coin flips?*

We can model the problem as a Discrete Time Markov chain where the states denote the number of successive heads obtained so far. Figure 14.7 shows the transition diagram of this Markov chain.

If we are in state 0, we jump back to state 0 if we receive a Tail else we jump to state 1. The probabilities of each of these actions is 1/2. Similarly if we are in state 2, we jump back to state 0 if we receive a Tail else we jump state 3, and similarly for the other transitions.

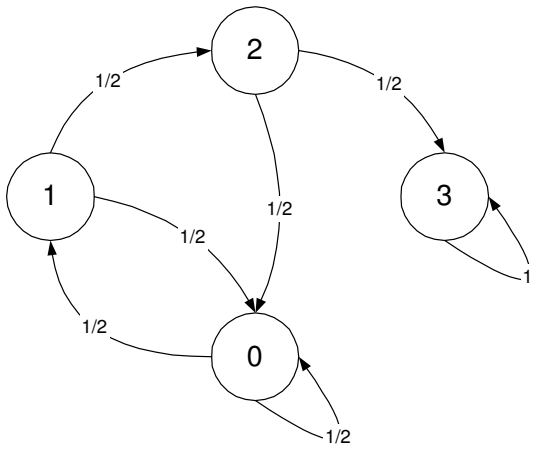


Figure 14.7: (a) Markov chain for example 14.8.1

We are interested in finding the mean number of coin flips until we get 3 successive heads. This is the mean number of steps taken to reach state 3 from state 0. Let $N = min\{n \geq 0, X_n = 3\}$, i.e., N is a random variable which specifies the first time we hit state 3. Let $\beta(i) = E[N \mid X_0 = i]$ for $i = 0, 1, 2, 3$.

The first step equations are as follows:

$$\begin{aligned}
\beta(0) &= \frac{1}{2}\beta(0) + \frac{1}{2}\beta(1) + 1; \\
\beta(1) &= \frac{1}{2}\beta(2) + \frac{1}{2}\beta(0) + 1; \\
\beta(2) &= \frac{1}{2}\beta(3) + \frac{1}{2}\beta(0) + 1; \\
\beta(3) &= 0.
\end{aligned}$$

Solving these three equations we get $\beta(0) = 14, \beta(1) = 12, \beta(2) = 8$.

Hence the expected numbed of tosses to until we get 3 successive heads is 14.

Example 14.8.2. *Let $\{X_n, n \geq 0\}$ be a Markov chain on $\boldsymbol{S}$ with transition probability matrix P and initial distribution π. Specify the probability space.*

The simplest choice is the *canonical* probability space defined as follows. $\Omega = \boldsymbol{S}^\infty$; $\mathcal{F}$ is

the smallest σ-field that contains all the events of the form

$$\{\omega \mid \omega_0 = i_0, \ldots \omega_n = i_n\};$$

P is the σ-additive set function on $\mathcal{F}$ such that

$$P(\{\omega \mid \omega_0 = i_0, \ldots \omega_n = i_n\}) = \pi(i_0)P(i_0, i_1) \times \cdots \times P(i_{n-1}, i_n).$$

Example 14.8.3. *We flip a biased coin forever. Let $X_1 = 0$ and, for $n \geq 2$, let $X_n = 1$ if the outcomes of the n-th and $(n-1)$-st coin flips are identical and $X_n = 0$ otherwise. Is $\boldsymbol{X} = \{X_n, n \geq 1\}$ a Markov chain?*

Designate by Y_n the outcome of the n-th coin flip. Let $P(H) = p = 1 - q$. If $\boldsymbol{X}$ is a Markov chain, then

$$P[X_4 = 1 \mid X_3 = 1, X_2 = 1] = P[X_4 = 1 \mid X_3 = 1, X_2 = 0].$$

The left-hand side is

$$\begin{aligned} &P[(Y_1, Y_2, Y_3, Y_4) \in \{HHHH, TTTT\} \mid (Y_1, Y_2, Y_3) \in \{HHH, TTT\}] \\ &= \frac{P((Y_1, Y_2, Y_3, Y_4) \in \{HHHH, TTTT\})}{P((Y_1, Y_2, Y_3) \in \{HHH, TTT\})} = \frac{p^4 + q^4}{p^3 + q^3}. \end{aligned}$$

Similarly, the right-hand side is

$$\begin{aligned} &P[(Y_1, Y_2, Y_3, Y_4) \in \{THHH, HTTT\} \mid (Y_1, Y_2, Y_3) \in \{THH, HTT\}] \\ &= \frac{P((Y_1, Y_2, Y_3, Y_4) \in \{THHH, HTTT\})}{P((Y_1, Y_2, Y_3) \in \{THH, HTT\})} = \frac{qp^3 + pq^3}{qp^2 + pq^2} = p^2 + q^2. \end{aligned}$$

Algebra shows that the expressions are equal if and only if $p = 0.5$. Thus, if $\boldsymbol{X}$ is a Markov chain, $p = 0.5$. Conversely, if $p = 0.5$, then we see that the random variables $\{X_n, n \geq 2\}$ are i.i.d. $B(0.5)$ and $\boldsymbol{X}$ is therefore a Markov chain.

Example 14.8.4. *A clumsy man tries to go up a ladder. At each step, he manages to go up one rung with probability p, otherwise he falls back to the ground. What is the average time he takes to go up to the n-th rung.*

Let $\beta(m)$ be the average time to reach the n-th rung, starting from the m-th one, for $m \in \{0, 1, 2, \ldots, n\}$. The FSE are

$$\begin{aligned} \beta(m) &= 1 + p\beta(m+1) + (1-p)\beta(0), \text{ for } m \in \{0, 1, \ldots, n-1\} \\ \beta(n) &= 0 \end{aligned}$$

The first equation is of the form $\beta(m+1) = a\beta(m) + b$ with $a = 1/p$ and $b = -1/p - (1-p)\beta(0)/p$. The solution is

$$\beta(m) = a^m\beta(0) + \frac{1-a^m}{1-a}b, m = 0, 1, \ldots, n.$$

Since $\beta(n) = 0$, we find that

$$a^n\beta(0) + \frac{1-a^n}{1-a}b = 0.$$

Substituting the values of a and b, we find

$$\beta(0) = \frac{1-p^n}{p^n - p^{n+1}}.$$

For instance, with $p = 0.8$ and $n = 10$, one finds $\beta(0) = 41.5$.

Example 14.8.5. *You toss a fair coin repeatedly with results $Y_0, Y_1, Y_2, \ldots$ that are 0 or 1 independently with probability 1/2 each. For $n \geq 1$ let $X_n = Y_n + Y_{n-1}$. Is X_n a Markov chain?*

No because

$$P[X_{n+1} = 0 \mid X_n = 1, X_{n-1} = 2] = \frac{1}{2} \neq P[X_{n+1} = 0 \mid X_n = 1] = \frac{1}{4}.$$

Example 14.8.6. *Consider a small deck of three cards 1, 2, 3. At each step, you take the middle card and you place it first with probability 1/2 or last with probability 1/2. What is the average time until the cards are in the reversed order 3, 2, 1?*

The possible states are the six permutations $\{123, 132, 312, 321, 231, 213\}$. The state transition diagram consists of these six states placed around a circle (in the order indicated), with a probability 1/2 of transition of one step clockwise or counterclockwise. Relabelling the states 1, 2, ..., 6 for simplicity, with 1 = 321 and 4 = 123, we write the FSE for the average time $\beta(i)$ from state i to state 1 as follows:

$$\begin{aligned} \beta(i) &= 1 + 0.5\beta(i+1) + 0.5\beta(i-1), i \neq 1 \\ \beta(1) &= 0. \end{aligned}$$

In these equations, the conventions are that 6 + 1 = 1 and 1 - 1 = 6. Solving the equations gives $\beta(1) = 0, \beta(2) = \beta(6) = 5, \beta(3) = \beta(5) = 8, \beta(4) = 9$. Accordingly, the answer to our problem is that it takes an average of 9 steps to reverse the order of the cards.

Example 14.8.7. *For the same Markov chain as in the previous example, what it the probability $F(n)$ that it takes at most n steps to reverse the order of the cards?*

Let $F(n; i)$ be the probability that it takes at most n steps to reach state 1 from state i, for $i \in \{1, 2, \ldots, 6\}$. The FSE for $F(n; i)$ are

$$\begin{aligned} F(n;i) &= 0.5F(n-1;i+1) + 0.5F(n-1;i-1), i \neq 1, n \geq 1 \\ F(n;1) &= 1, n \geq 0 \\ F(0;i) &= 1\{i = 1\}. \end{aligned}$$

Again we adopt the conventions that 6 + 1 = 1 and 1 - 1 = 6. We can solve the equations numerically and plot the values of $F(n) = F(4; n)$. The graph is shown in Figure 14.8.

Example 14.8.8. *Is the previous Markov chain periodic?*

Yes, it takes 2, 4, 6, ... steps to go from state i to itself. Thus, the Markov chain is periodic with period 2. Recall that this implies that the probability of being in state i does not converge to the invariant distribution $(1/6, 1/6, \ldots, 1/6)$. The graph in Figure 14.9

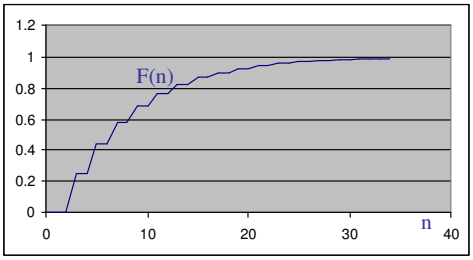


Figure 14.8: Graph of $F(n)$ in example 14.8.7

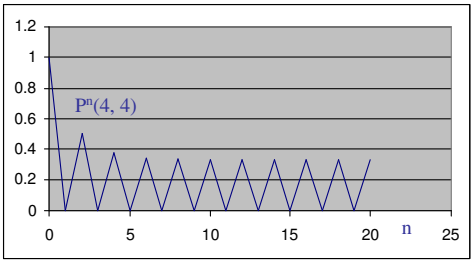


Figure 14.9: Graph of $P^n(4,4)$ in example 14.8.8

shows the probability of being in state 4 at time n given that $X_0 = 4$. This is derived by calculating $P(4,4)^n$. Since $P^{n+1}(4,j) = \sum_i P^n(4,i)P(i,j) = 0.5P^n(4,j-1) + 0.5P^n(4,j+1)$, one can compute recursively by iterating a vector with 6 elements instead of a matrix with 36.

Example 14.8.9. *We flip a fair coin repeatedly until we get either the pattern HHH or HTH. What is the average number of coin flips?*

Let X_n be the last two outcomes. After two flips, we start with X_0 that is equally likely to be any of the four pairs in $\{H,T\}^2$. Look at the transition diagram of Figure 14.10.

The FSE for the average time to hit one of the two states HTH or HHH from the

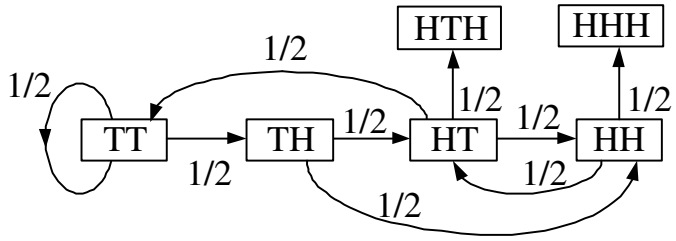


Figure 14.10: Transition diagram of X_n in example 14.8.9

other states are as follows:

$$\begin{aligned}
\beta(TT) &= 1 + 0.5\beta(TT) + 0.5\beta(TH) \\
\beta(TH) &= 1 + 0.5\beta(HH) + 0.5\beta(HT) \\
\beta(HT) &= 1 + 0.5\beta(TT) \\
\beta(HH) &= 1 + 0.5\beta(HT)
\end{aligned}$$

Solving these equations we find

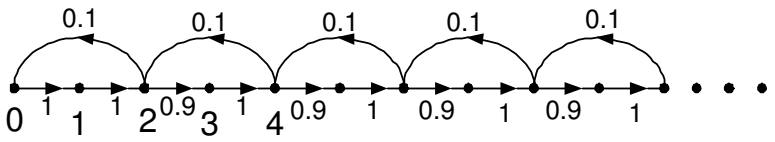
$$(\beta(TT), \beta(HT), \beta(HH), \beta(TH)) = \frac{1}{5}(34, 22, 16, 24)$$

and the answer to our problem is then

$$2 + \frac{1}{4}(\beta(TT) + \beta(TH) + \beta(HT) + \beta(HH)) = \frac{34}{5} = 6.8.$$

Example 14.8.10. *Give an example of a discrete time irreducible Markov chain with period 3 that is transient.*

The figure below shows the state diagram of an irreducible Markov chain with period 3. Indeed, the Markov chain can go from 0 to 0 in $3, 6, 9, \ldots$ steps. The probability of motion to the right is much larger than to the left. Accordingly, one can expect that the Markov chain goes to infinity.



Example 14.8.11. *For $n = 1, 2, \ldots$, during year n, barring any catastrophe, your company makes a profit equal to X_n, where the random variables X_n are i.i.d. and uniformly distributed in $\{0, 1, 2\}$. Unfortunately, during year n, there is also a probability of a catastrophe that sets back your company's total profit to 0. Such a catastrophe occurs with probability 5% independently each year. Explain precisely how to calculate the average time until your company's total profit reaches 100. (The company does not invest its money; the total profit is the sum of the profits since the last catastrophe.) Do not perform the calculations but provide the equations to be solved and explain a complete algorithm that I could use to perform the calculations.*

Let $\beta(i)$ be the average time until the profits reach 100 starting from i.

Then

$$\begin{aligned} \beta(100) &= \beta(101) = 0 \\ \beta(i) &= 1 + 0.05\beta(0) + \frac{0.95}{3}\beta(i) + \frac{0.95}{3}\beta(i+1) + \frac{0.95}{3}\beta(i+2), \text{ for } i = 0, 1, \ldots, 99. \end{aligned}$$

Fix $\beta(0)$ and $\beta(1)$ arbitrarily. use the second equation to find $\beta(2), \beta(3), \ldots, \beta(100), \beta(101)$, in that order. Use the first two equations to determine the two unknowns $\beta(0), \beta(1)$.

Example 14.8.12. *Consider the discrete time Markov chain on $\{0, 1, 2, 3, 4\}$ with transition probabilities such that*

$$P(i,j) = \begin{cases} 0.5, & \text{if } j = (i+1) mod 5 \\ 0.5, & \text{if } j = (i+2) mod 5 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, $P(0,1) = P(0,2) = P(1,2) = P(1,3) = \ldots = P(3,4) = P(3,0) = P(4,0) = P(4,1) = 0.5$. If you picture the states $\{0, 1, \ldots, 4\}$ as the vertices of a pentagon whose labels increase clockwise, then the Markov chain makes one or two steps clockwise with probability 0.5 each at every time instant.

a. Is this Markov chain periodic and, if it is, what is the period?

b. What is an invariant distribution of the Markov chain?

c. Is that invariant distribution unique?

d. Write the first step equations to calculate the average time for the Markov chain to go from state 0 back to state 0. Can you guess the answer from the result of part (b)?

a. The Markov chain is aperiodic. For instance, it can go from state 0 to itself in 4 or 5 steps.

b. By symmetry it has to be $(1/5, 1/5, 1/5, 1/5, 1/5)$.

c. A finite irreducible Markov chain has always one and only one invariant distribution.

d. Let $\beta(i)$ be the average time to reach 0 starting from i. Then

$$\begin{aligned} \beta(0) &= 0, \\ \beta(i) &= 1 + 0.5\beta(i+1) + 0.5\beta(i+2), \text{ for } i \neq 0. \end{aligned}$$

In the last equation, the addition is modulo 5.

Example 14.8.13. *Let $\{X_n, n \geq 0\}$ be i.i.d. Bernoulli with mean p. Define $Y_0 = X_0$. For $n \geq 1$, let $Y_n = \max\{X_n, X_{n-1}\}$. Is $\{Y_n, n \geq 0\}$ a Markov chain? Prove or disprove.*

The sequence $\{Y_n, n \geq 0\}$ is not a Markov chain unless $p = 1$ or $p = 0$. To see this, note that

$$P[Y_3 = 1 \mid Y_2 = 1, Y_1 = 0] = P[X_0 = 0, X_1 = 0, X_2 = 1 \mid X_0 = 0, X_1 = 0, X_2 = 1] = 1.$$

However, if $p \in (0, 1)$, one finds that

$$\begin{aligned} P[Y_3 = 1 \mid Y_2 = 1, Y_1 = 1] &= P[\{0101, 101, 111\} \mid \{010, 101, 111\} \\ &= \frac{p^2q^2 + p^2q + p^3}{pq^2 + p^2q + p^3} < 1. \end{aligned}$$

In this derivation, by 0101 we mean $X_0 = 0, X_1 = 1, X_2 = 0, X_3 = 1$; by 101 we mean $X_0 = 1, X_1 = 0, X_2 = 1$, and similarly for the other terms.

The average time from state 0 to itself can be guessed from the invariant distribution as follows. Imagine that once the Markov chain reaches state 0 it takes on average β steps for it to return to 0. Then, the Markov chain spends one step out every $1 + \beta$ steps in state 0, on average. Hence, the probability of being in state 0 should be equal to $1/(1 + \beta)$. Thus, $1/(1 + \beta) = 1/5$, so that $\beta = 4$.

Example 14.8.14. *Consider a modified random walk defined as follows. If $Y_n = k$, then $Y_{n+1} = max\{0, min\{k + X_{n+1}, N\}\}$ where the $\{X_n, n \geq 1\}$ are i.i.d. with $P(X_n = +1) = p = 1 - P(X_n = -1)$. The random variable Y_0 is independent of $\{X_n, n \geq 1\}$. Assume that $0 < p < 1$.*

a. Calculate the stationary distribution of $\{Y_n, n \geq 0\}$.

b. Use a probabilistic (coupling) argument to show that $\{Y_n, n \geq 0\}$ is asymptotically stationary.

c. Write the first step equations for the average time until $Y_n = 0$ given $Y_0 = k$.

a. We write and solve the balance equations:

$$\begin{aligned} \pi(0) &= (1-p)\pi(0) + (1-p)\pi(1) \Rightarrow \pi(1) = \rho\pi(0) \text{ where } \rho := \frac{p}{1-p} \\ \pi(1) &= p\pi(0) + (1-p)\pi(2) = (1-p)\pi(1) + (1-p)\pi(2) \Rightarrow \pi(2) = \rho^2\pi(0) \end{aligned}$$

Continuing this way shows that

$$\pi(n) = \rho^n\pi(0), n = 0, 1, \ldots, N.$$

Since the probabilities add up to 1, find $\pi(0)$ and we conclude that

$$\pi(n) = \frac{\rho^n(1-\rho)}{1-\rho^{N+1}}, n = 0, 1, \ldots, N.$$

b. Let $\{Z_n, n \geq 0\}$ be a stationary version of the Markov chain. That is, its initial value Z_0 is selected with the invariant distribution and if $Z_n = k$, then $Z_{n+1} = max\{0, min\{k + X_{n+1}, N\}\}$. Define the sequence $\{Y_n, n \geq 0\}$ so that Y_0 is arbitrary and if $Y_n = k$, then $Y_{n+1} = max\{0, min\{k + X_{n+1}, N\}\}$. The random variables X_n are the same for the two sequences. Note that Z_n and Y_n will be equal after some finite random time τ. For instance, we can choose τ to be the first time that N successive X_n's are equal to -1. Indeed, at that time, both Y and Z must be zero and they remain equal thereafter. Now,

$$|P(Y_n = k) - P(Z_n = k)| \leq P(Y_n \neq Z_n) \leq P(n < \tau) \to 0 \text{ as } n \to \infty.$$

Since $P(Z_n = k) = \pi(k)$ for all n, this shows that $P(Z_n = k) \to \pi(k)$.

c. Let

$$\beta(k) = E[T_0 \mid Y_0 = k].$$

Then

$$\begin{aligned}
\beta(k) &= 1 + p\beta(k+1) + (1-p)\beta(k-1), \text{ for } 0 < k < N; \\
\beta(0) &= 1 + (1-p)\beta(0) + p\beta(1); \\
\beta(N) &= 1 + (1-p)\beta(N-1) + p\beta(N); \\
\beta(0) &= 0.
\end{aligned}$$

Chapter 15

Markov Chains - Continuous Time

We limit our discussion to a simple case: the regular Markov chains. Such a Markov chain visits states in a countable set. When it reaches a state, it stays there for an exponentially distributed random time (called the state holding time) with a mean that depends only on the state. The Markov chain then jumps out of that state to another state with transition probabilities that depend only on the current state. Given the current state, the state holding time, the next state, and the evolution of the Markov chain prior to hitting the current state are independent. The Markov chain is regular if jumps do not accumulate, i.e., if it makes only finitely many jumps in finite time. We explain this construction and we give some examples. We then state some results about the stationary distribution.

15.1 Definition

A random process $X = \{X(t), t \in \Re\}$ is a continuous-time Markov chain on the countable set S with *generator* (or *rate matrix*) $Q = [q(i,j), i,j \in S]$ if

$$P[X(t+\epsilon) = j | X(t) = i, X(s), s \leq t] = 1\{i = j\} + q(i,j)\epsilon + o(\epsilon), i,j \in S. \qquad (15.1.1)$$

Here, Q is a matrix with nonnegative off-diagonal elements, finite nonpositive diagonal elements, and row sums equal to zero. Such a matrix is called a *rate matrix* or *generator*.

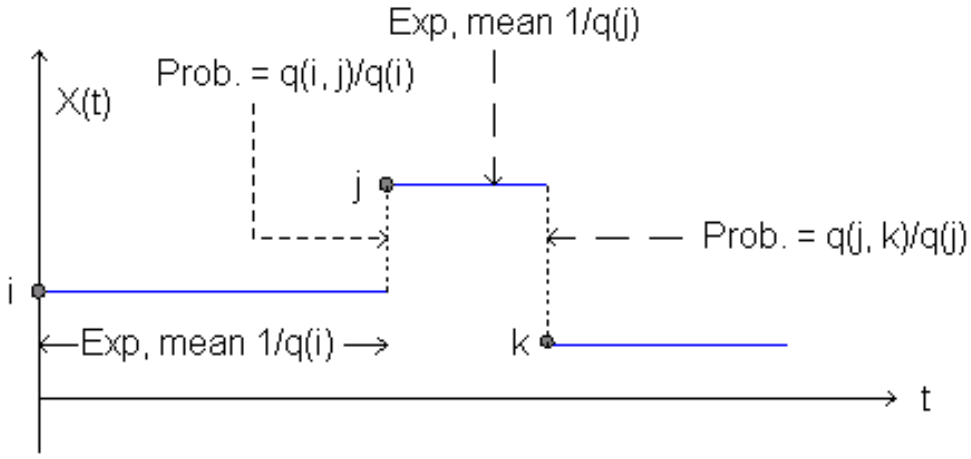


Figure 15.1: Construction of continuous-time Markov chain.

The formula says (if you can read between the symbols) that given $X(t)$, the future and the past are independent.

One represents the rate matrix by a state transition diagram that shows the states; an arrow from i to $j \neq i$ marked with $q(i,j)$ shows the transition rate between these states. No arrow is drawn when the transition rate is zero.

15.2 Construction (regular case)

Let Q be a rate matrix. Define the process $X = \{X(t), t \geq 0\}$ as follow. Start by choosing $X(0)$ according to some distribution in S. When X reaches a state i (and when it starts in that state), it stays there for some exponentially distributed time with mean $1/q(i)$ where $q(i) = -q(i,i)$. When it leaves state i, the process jumps to state $j \neq i$ with probability $q(i,j)/q(i)$. The evolution of X then continues as before. Figure 15.1 illustrates this construction.

This construction defines a process on the positive real line if the jumps do not accumulate, which we assume here. A simple argument to shows that this construction corresponds

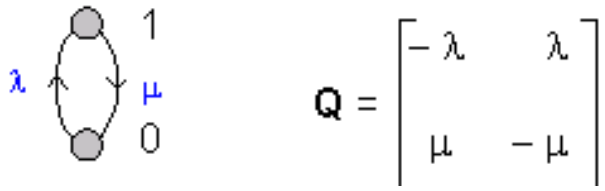


Figure 15.2: Markov chain with two states.

Figure 15.3: State transition diagram of Poisson process.

to the definition. Essentially, the memoryless property of the exponential distribution and the definition of the jump probabilities yield (15.1.1).

15.3 Examples

The example of Figure 15.2 corresponds to a Markov chain with two states. The example of Figure 15.3 corresponds to a Poisson process with rate λ. The example of Figure 15.4 corresponds to a reflected difference between two Poisson processes. (See Applications.)

Figure 15.4: Reflected difference of two Poisson processes.

15.4 Invariant Distribution

The classification results are similar to the discrete time case. An irreducible Markov chain (can reach every state from every other state) is either null recurrent, positive recurrent, or transient (defined as in discrete time). The positive recurrent Markov chains have a unique invariant distribution, the others do not have any. Also, a distribution π is invariant if and only if it solves the following *balance equations*:

$$\pi Q = 0. \tag{15.4.1}$$

For instance, the Markov chain in Figure 15.4 is transient if $\rho := \lambda/\mu > 1$; it is null recurrent if $\rho = 1$; it is positive recurrent if $\rho < 1$ and then its invariant distribution is

$$\pi(n) = (1-\rho)\rho^n, n = 0, 1, 2, \ldots.$$

15.5 Time-Reversibility

A stationary irreducible Markov chain is time-reversible if and only if π satisfies the detailed balance equations:

$$\pi(i)q(i,j) = \pi(j)q(j,i) \text{ for all } i, j \in S. \tag{15.5.1}$$

For instance, in Figure 15.4 is time-reversible.

15.6 Summary

The random process $\boldsymbol{X} = \{X_t, t \geq 0\}$ taking values in the countable set $\boldsymbol{S}$ is a Markov chain with rate matrix $\boldsymbol{Q}$ if it satisfies (15.1.1).

The definition specifies the Markov property that given X_t the past and the future are independent. Recall that the Markov chain stays in state i for an exponentially distributed

time with rate $q(i)$, then jumps to state j with probability $q(i,j)/q(i)$ for $j \neq i$, and the evolution continues in that way.

We discussed the classification of Markov chains. In particular, we explained the notions of irreducibility, positive and null recurrence, and transience.

A distribution π is invariant if and only if it satisfies the balance equations (15.4.1). A stationary Markov chain is time-reversible if the detailed balance equations (15.5.1) are satisfied.

15.7 Solved Problems

Example 15.7.1. *Consider n light bulbs that have independent lifetimes exponentially distributed with mean 1. What is the average time until the last bulb dies?*

Let X_t be the number of bulbs still alive at time $t \geq 0$. Because of the memoryless property of the exponential distribution, $\{X_t, t \geq 0\}$ is a Markov chain. Also, the rate matrix is seen to be such that

$$q(m) = q(m, m-1) = m, m \in \{1, 2, \ldots, n\}.$$

The average time in state m is $1/m$ and the Markov chain goes from state n to $n-1$ to $n-2$, and so on until it reaches 0. The average time to hit 0 is then

$$\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{3} + \frac{1}{2} + 1.$$

To fix ideas, one finds the average time to be about 3.6 when $n = 20$.

Example 15.7.2. *In the previous example, assume that the janitor replaces a burned out bulb after an exponentially distributed time with mean 0.1. What is the average time until all the bulbs are out?*

The rate matrix now corresponds to the state diagram shown in Figure 15.5. Defining

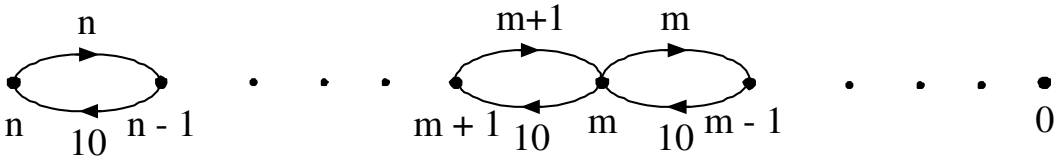


Figure 15.5: State transitions diagram in example 15.7.2

$\beta(m)$ as the average time from state m to state 0, for $m \in \{0, 1, \ldots, n\}$, we can write the FSE as

$$\begin{aligned}
\beta(m) &= \frac{1}{m+10} + \frac{m}{m+10}\beta(m-1) + \frac{10}{m+10}\beta(m+1), \text{ for } m \in \{1, 2, \ldots, n-1\} \\
\beta(n) &= \frac{1}{n} + \beta(n-1) \\
\beta(0) &= 0.
\end{aligned}$$

If we knew $\beta(n-1)$, we could solve recursively for all values of $\beta(m)$. We could then check that $\beta(0) = 0$. Choosing $n = 20$ and adjusting $\beta(19)$ so that $\beta(0) = 0$, we find numerically that $\beta(20) \approx 2,488$.

Example 15.7.3. *Let $\boldsymbol{A} = \{A_t, t \geq 0\}$ and $\boldsymbol{B} = \{B_t, t \geq 0\}$ be two independent Poisson processes with rates λ and μ, respectively. Let also X_0 be a random variable independent of $\boldsymbol{A}$ and $\boldsymbol{B}$. Show that $\{X_t = X_0 + A_t - B_t, t \geq 0\}$ is a Markov chain. What is its rate matrix? Show that it is irreducible (unless $\lambda = \mu = 0$). For what values of λ and μ is the Markov chain positive recurrent, null recurrent, transient? Explain.*

To prove that X_t is a CTMC, we need to show (15.1.1). We find

$$\begin{aligned}
& lim_{h\to\infty} P[A_{t+h} - A_t = 1, B_{t+h} - B_t = 0 | X_u, 0 \leq u \leq t] \\
= \; & lim_{h\to\infty} P[A_{t+h} - A_t = 1 | A_u, B(u), X_0, 0 \leq u \leq t] P[B_{t+h} - B_t = 0 | A_u, B(u), X_0, 0 \leq u \leq t] \\
= \; & lim_{h\to\infty} P[A_{t+h} - A_t = 1 | A_u, 0 \leq u \leq t] P[B_{t+h} - B_t = 0 | B(u), 0 \leq u \leq t] \\
= \; & (\lambda h + o(h))(1 - \mu h + o(h)) \\
= \; & \lambda h + o(h).
\end{aligned}$$

Similarly, we can show that

$$lim_{h\to\infty} P[X_{t+h} - X_t = -1 | X_u, 0 \leq u \leq t] \quad = \quad \mu h + o(h)$$

To study the recurrence or transience of the Markov chain, consider the *jump chain* with the following transition probabilities:

$$P_{ij} = \begin{cases} \frac{\lambda}{\lambda+\mu} & j = i+1 \\ \frac{\mu}{\lambda+\mu} & j = i-1 \\ 0 & otherwise. \end{cases}$$

We know that this DTMC is transient if $\frac{\lambda}{\lambda+\mu} \neq \frac{\mu}{\lambda+\mu}$ and is null recurrent if $\frac{\lambda}{\lambda+\mu} = \frac{\mu}{\lambda+\mu}$. Hence the original CTMC is null recurrent if $\lambda = \mu$ and transient if $\lambda \neq \mu$.

Example 15.7.4. *Let Q be a rate matrix on a finite state space $\boldsymbol{S}$. For each pair of states $(i,j) \in \boldsymbol{S}^2$ with $i \neq j$, let $\boldsymbol{N}(i,j) = \{N_t(i,j), t \geq 0\}$ be a Poisson process with rate $q(i,j)$. Assume that these Poisson processes are all mutually independent. Construct the process $\boldsymbol{X} = \{X_t, t \geq 0\}$ as follows. Pick X_0 randomly in $\boldsymbol{S}$, independently of the Poisson processes. If $X_t = i$, let s be the first jump time after time t of one of the Poisson processes $\boldsymbol{N}(i,j)$ for $j \neq i$. If s is a jump time of $\boldsymbol{N}(i,j)$, then let $X_u = i$ for $u \in [t,s)$ and let $X_s = j$. Continue the construction using the same procedure. Show that $\boldsymbol{X}$ is a Markov chain with rate matrix Q.*

First note the following. If we have two random variables $X =_D Exd(\lambda)$ and $Y =_D Exd(\mu)$, then $P(X < Y) = \lambda/(\lambda+\mu)$. Also, $\min\{A,B\} =_D Exd(\lambda+\mu)$.

Since Q is a rate matrix, $q(i,i) = -\sum_{j=0,j\neq i}^{S} q(i,j)$. Imagine that we are in state i at time t. $S-1$ exponential clocks, each with rate $q(i,j), 1 \leq j \leq S, j \neq i$ are running simultaneously. If clock k expires first we jump to state k. Probability of the k^{th} clock expiring first is given by $\frac{q(i,k)}{\sum_{j=0,j\neq i}^{S} q(i,j)}$. The distribution of the time spent in state i is exponential with rate $\sum_{j=0,j\neq i}^{S} q(i,j)$.

Next, we must look at the infinitesimal rate of jumping from state i to state j.

$$lim_{h\to\infty}P(X_{t+h}=j|X_t=i,X(u),0\le u\le t) \quad = \quad q(i,j)h+o(h)$$

Also,

$$lim_{h\to\infty}P(X_{t+h}=i|X_t=i,X(u),0\le u\le t) \quad = \quad 1-\sum_{j=0,j\neq i}^{S} q(i,j)+o(h)$$

Hence the rate matrix obtained for X_t is also Q.

Example 15.7.5. *Let $\boldsymbol{X}$ be a Markov chain with rate matrix Q on a finite state space $\boldsymbol{S}$. Let λ be such that $\lambda \ge -q(i,i)$ for all $i \in \boldsymbol{S}$. Define the process $\boldsymbol{Y}$ as follows. Choose $Y_0 = X_0$. Let $\{N_t, t \ge 0\}$ be a Poisson process with rate λ and let T be its first jump time. If $Y_0 = i$, then let $Y_u = i$ for $u \in [0,T)$. Let also $Y_T = j$ with probability $q(i,j)/\lambda$ for $j \neq i$ and $Y_T = i$ with probability $1 + q(i,i)/\lambda$. Continue the construction in the same way. Show that $\boldsymbol{Y}$ is a Markov chain with rate Q.*

Consider the infinitesimal rate of jumping from state i to state j. In the original chain we had:

$$lim_{h\to\infty}P(X_{t+h}=j|X_t=i,X(u),0\le u\le t) \quad = \quad q(i,j)h+o(h)$$

In the new chain Y, when in state i, we start an exponentially distributed clock of rate λ instead of rate $-q(i,i)$ as in the original chain. When the clock expires we jump to state j with probability $\frac{q(i,j)}{\lambda}$. (In chain X this probability was $\frac{q(i,j)}{-q(i,i)}$).

So in chain Y, the probability of jumping from state i to state j in an infinitesimal time h is the probability that the exponential clock expires and we actually jump.

$$\begin{aligned} lim_{h\to\infty}P(Y_{t+h}=j|Y_t=i,X(u),0\le u\le t) &= \frac{q(i,j)}{\lambda}*(\lambda h+o(h)) \\ &= q(i,j)h+o(h) \end{aligned}$$

In Markov chain X, the probability of staying in state i in an infinitesimal time period h was given by.

$$lim_{h\to\infty}P(X_{t+h} = i|X_t = i, X(u), 0 \leq u \leq t) \quad = \quad 1 + q(i,i)h + o(h)$$

In the new chain Y, the probability of staying in the current state is equal to the probability that exponential clock does not expire in the infinitesimal time interval h or that clock expires but we do not jump.

$$\begin{aligned} lim_{h\to\infty}P(Y_{t+h} = i|Y_t = i, Y(u), 0 \leq u \leq t) &= 1 - \lambda h + o(h) + (\lambda h + o(h))(1 + \frac{q(i,i)}{\lambda}) \\ &= 1 + q(i,i)h + o(h) \end{aligned}$$

Hence the new chain Y exhibits the same rate matrix as chain X.

Example 15.7.6. *Let* $\boldsymbol{X}$ *be a Markov chain on* $\{1,2,3,4\}$ *with the rate matrix* Q *given by*

$$Q = \begin{bmatrix} -2 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & -2 \\ 1 & 0 & 1 & -2 \end{bmatrix}.$$

a. Calculate $E[T_2 \mid X_0 = 1]$ *where* $T_2 = \min\{t \geq 0 \mid X_t = 2\}$.

b. How would you calculate $E[e^{iuT_2} \mid X_0 = 1]$ *for* $u \in \Re$*?*

Define $\beta(i) = E[T_2|X_0 = i]$. The general form of the first step equations for the CTMC is:

$$\beta(i) = \frac{1}{-q(i,i)} + \sum_{j=1}^{N} \frac{q(i,j)}{-q(i,i)}\beta(j)$$

where N is the size of the state space. For our example,

$$\begin{aligned}
\beta(1) &= \frac{1}{2} + \frac{1}{2}\beta(2) + \frac{1}{2}\beta(4) \\
\beta(2) &= 0 \\
\beta(3) &= \frac{1}{2} + \frac{1}{2}\beta(1) + \frac{1}{2}\beta(2) \\
\beta(i4) &= \frac{1}{2} + \frac{1}{2}\beta(1) + \frac{1}{2}\beta(3).
\end{aligned}$$

Solving these equations we get: $\beta(1) = 1.4, \beta(3) = 1.2, \beta(4) = 1.8$.

b. We can find $E[e^{iuT_2}|X_0 = 1]$ by writing the first step equations for the characteristic function. For instance, note that the time $T(1)$ to hit 2 starting from 1 is equal to an exponential time with rate 2, say τ, plus, with probability 1/2, the time $T(4)$ to hit 2 from state 4. That is, we can write

$$\begin{aligned}
v_1(u) &:= E(e^{iuT(1)}) = \frac{1}{2}E(e^{iu\tau}) + \frac{1}{2}E(e^{iu(\tau+T(4))}) \\
&= \frac{1}{2}\frac{2}{2-iu} + \frac{1}{2}\frac{2}{2-iu}v_4(u) \\
&= \frac{1}{2-iu} + \frac{1}{2-iu}v_4(u).
\end{aligned}$$

In this derivation, we used the fact that $E(e^{iu\tau}) = 2/(2-iu)$ and we defined $v_4(u) = E(e^{iuT(4)})$.

Similarly, we find

$$\begin{aligned}
v_3(u) &:= E(e^{iuT(3)}) = \frac{1}{2-iu} + \frac{1}{2-iu}v_1(u), \\
v_4(u) &= \frac{1}{2-iu}v_1(u) + \frac{1}{2-iu}v_3(u).
\end{aligned}$$

Solving these equations for $v_1(u)$ we find

$$E[e^{iuT_2} \mid X_0 = 1] = \frac{(2-iu)^3 + (2-iu)}{(2-iu)^4 - (2-iu)^2 - 1}.$$

Chapter 16

Applications

Of course, one week of lectures on applications of this theory is vastly inadequate. Many courses are devoted to such applications, including courses on communication systems, digital communication theory, stochastic control, performance evaluation of communication networks, information theory and coding, and many others. In this brief chapter we take a look at a few representative applications.

16.1 Optical Communication Link

Consider the following model of an optical communication link: [Laser] → [Fiber] → [Photodetector] → [Receiver].

To send a bit "1" we turn the laser on for T seconds; to send a bit "0" we turn it off for T seconds. We agree that we start by turning the laser on for T seconds before the first bit and that we send always groups of N bits. [More sophisticated systems exist that can send a variable number of bits.] When the laser is on, it produces light that reaches the photodetector with some intensity λ_1. This light is "seen" by the photodetector that converts it into electricity. When the laser is off, no light reaches the photodetector. Unfortunately, the electronic circuitry adds some thermal noise. As a result, the receiver sees light with an intensity $\lambda_0 + \lambda_1$ when the laser is on and with an intensity λ_0 when

the laser is off. (That is, λ_0 is the equivalent intensity of light that corresponds to the noise current.) By "light with intensity λ" we mean a Poisson process of photons that has intensity λ. Indeed, a laser does not produce photons are precise times. Rather, it produces a Poisson stream of photons. The brighter the laser, the larger the intensity of the Poisson process.

The problem of the receiver is to decide whether it has received a bit 0 or a bit 1 during a given time interval of T seconds. For simplicity, we assume that the boundary between time intervals is known. (What would you do to find these boundaries?)

Let Y be the number of photons that the receiver sees during the time interval. Let $X = 0$ if the bit is 0 and $X = 1$ if the bit is 1. Then $f_{Y|X}[y|0]$ is Poisson with mean $\lambda_0 T$ and $f_{Y|X}[y|1]$ is Poisson with mean $(\lambda_0 + \lambda_1)T$. The problem can then be formulated as an MLE, MAP, or HT problem.

Assume that the bits 0 and 1 are equally likely. To minimize the probability of detection error we should detect using the MLE (which is the same as the MAP in this case). That is, we decide 1 if $P[Y = y|X = 1] > P[Y = y|X = 0]$. To simplify the math, we use a Gaussian approximation of the Poisson random variable. Specifically, we approximate a Poisson random variable with mean μ with a $N(\mu, \mu)$. With this approximation,

$$f_{Y|X}[y|0] = \frac{1}{\sqrt{2\pi\lambda_0 T}} \exp\{-(y - \lambda_0 T)^2/(2\lambda_0 T)\}$$

and

$$f_{Y|X}[y|1] = \frac{1}{\sqrt{2\pi(\lambda_0 + \lambda_1)T}} \exp\{-(y - (\lambda_0 + \lambda_1)T)^2/(2(\lambda_0 + \lambda_1)T)\}$$

Figure 16.1 shows these densities when $\lambda_0 T = 10$ and $(\lambda_0 + \lambda_1)T = 25$.

Using the graphs, one sees that the receiver decides that it got a bit 1 whenever $Y > 16.3$. (In fact, I used the actual values of the densities to identify this threshold.) You also see that $P[\text{Decide } "1''|X = 0] = P(N(10, 10) > 16.3) = 0.023$ and $P[\text{Decide } "0''|X = 1] = P(N(25, 25) < 16.3) = 0.041$. To find the numerical values, you use a calculator or a

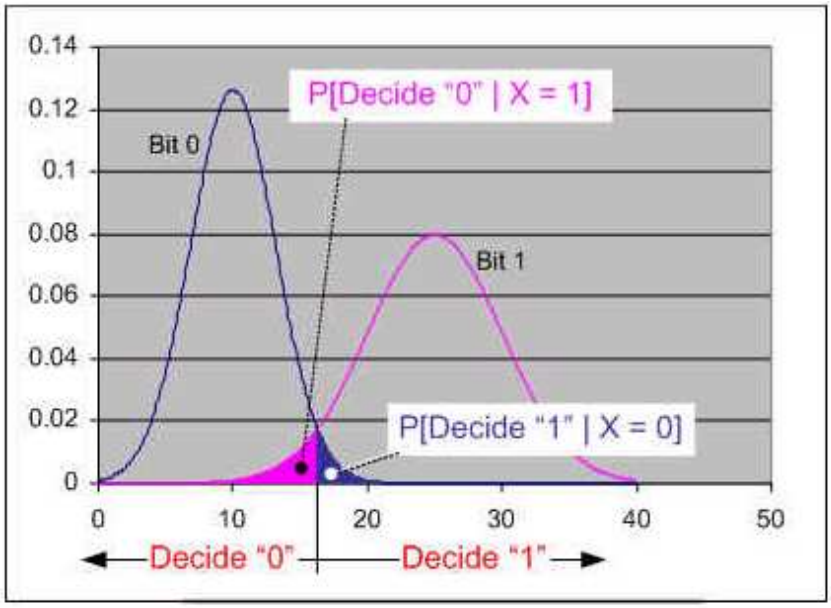


Figure 16.1: Approximate densities of number of photons under bits 0 and 1.

table of the c.d.f. of $N(0,1)$ after you write that $P(N(10,10) > 16.3) = P(N(0,1) > (16.3-10)/(10^{0.5})$ and similarly for the other value.

One interesting question is to figure out how to design the link so that the probability of errors are acceptably small. A typical target for an optical link is a probability of error of the order of 10^{-12}, which is orders of magnitude smaller than what we have achieved so far. To reduce the probability of error, one must reduce the amount of "overlap" of the two densities shown in the figure. If the values of λ_0 and λ_1 are given (λ_0 depends on the noise and λ_1 depends on the received light power from the transmitter laser), one solution is to increase T, i.e., to transmit the bits more slowly by spending more time for each bit. Note that the graphs will separate if one multiplies the means and variances by some constant larger than 1.

16.2 Digital Wireless Communication Link

Consider the following model of an wireless communication link: [Transmitter and antenna] → [Free Space] → [Antenna and receiver]

For simplicity, assume a discrete time model of the system. To transmit bit 0 the transmitter sends a signal $\boldsymbol{a} := \{a_n, n = 1, 2, \ldots, N\}$. To transmit bit 1 the transmitter sends a signal $\boldsymbol{b} := \{b_n, n = 1, 2, \ldots, N\}$. The actual values are selected based on the efficiency of the antenna at transmitting such signals and on some other reasons. In any case, it seems quite intuitive that the two signals should be quite different if the receiver must be able to distinguish a 0 from a 1. We try to understand how the receiver makes its choice. As always, the difficulty is that the receiver gets the transmitted signal corrupted by noise. A good model of the noise is that the receiver gets

$$Y_n = a_n + Z_n, n = 1, 2, \ldots, N$$

when the transmitter sends a bit 0 and

$$Y_n = b_n + Z_n, n = 1, 2, \ldots, N$$

when the transmitter sends a bit 1, where $\{Z_n, n = 1, 2, \ldots, N\}$ are i.i.d., $N(0, \sigma^2)$.

The MLE decides 0 if $||\boldsymbol{Y} - \boldsymbol{a}||^2 := \sum_n (Y_n - a_n)^2 < \sum_n (Y_n - b_n)^2 =: ||\boldsymbol{Y} - \boldsymbol{b}||^2$, and decides 1 otherwise. That is, the MLE decides 0 if the received signal $\boldsymbol{Y}$ is closer to the signal $\boldsymbol{a}$ than to the signal $\boldsymbol{b}$. Nothing counterintuitive; of course the key is to measure "closeness" correctly. One can then study the probability of making an error and one finds that it depends on the energy of the signals $\boldsymbol{a}$ and $\boldsymbol{b}$; again this is not too surprising. The problem can be extended to more than 2 signals so that we can send more than one bit in N steps. The problem of designing the best signals (that minimize the probability of error) with a given energy is then rather tricky.

16.3 M/M/1 Queue

Consider jobs that arrive at a queue according to a Poisson process with rate λ. The jobs are served by a single server and they require i.i.d. $Exd(\mu)$ service times. This queue is called the M/M/1 queue. The notation designates (inter-arrivals/service/number of servers) and M means memoryless (i.e., Poisson arrivals and exponential service times).

Because of the memoryless property of the exponential distribution, the number $X(t)$ of jobs in the queue at time t is a Markov chain with the state transition diagram shown in Figure 15.4. The balance equations of that MC are as follows:

$$\begin{aligned} \lambda\pi(0) &= \mu\pi(1) \\ (\lambda+\mu)\pi(n) &= \lambda\pi(n-1)+\mu\pi(n+1), n = 1, 2, \ldots. \end{aligned}$$

You can check that the solution is $\pi(n) = \rho^n(1-\rho)$ for $n = 0, 1, \ldots$, with $\rho = \lambda/\mu$, provided that $\lambda < \mu$. (Otherwise, there is no invariant distribution.)

Consider a job that arrives at the queue and the queue is in steady-state (i.e., $X(\cdot)$ has its invariant distribution). What is the probability that it finds n other jobs already in the queue? Say that the job arrives during $(t, t+\epsilon)$, then we want

$$P[X(t) = n | X(t+\epsilon) = X(t)+1] = P(X(t) = n) = \pi(n)$$

because $X(t+\epsilon)-X(t)$ and $X(t)$ are independent. (The memoryless property of the Poisson process.) This result is know as PASTA: Poisson arrivals see time averages.

How long will the job spend in the queue? To do the calculation, first recall that a random variable W is $Exd(\alpha)$ if and only if $E(\exp\{-uW\}) = \alpha/(u+\alpha)$. Next, observe that with probability $\pi(n)$, the job has to wait for $n+1$ exponential service times $Z_1, \ldots, Z_{n+1}$

before it leaves. Consequently, if we designate its time in the queue by V,

$$\begin{aligned} E(\exp\{-uV\}) &= \sum_n \pi(n) \exp\{-u(Z_1 + \cdots + Z_{n+1})\} \\ &= \sum_n \rho^n (1-\rho)(\lambda/(u+\lambda))^{n+1} = \cdots = \frac{\mu - \lambda}{u + \mu - \lambda}. \end{aligned}$$

This calculation shows that the job spends an exponential time in the queue with mean $1/(\mu - \lambda)$.

16.4 Speech Recognition

We explain a simplified model of speech recognition and the algorithm that computes the MAP.

A speaker pronounces a string of syllables $(X_1, \ldots, X_n)$ that is modelled as a Markov chain:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \pi(x_1) P(x_1, x_2) \cdots P(x_{n-1}, x_n)$$

for $x_1, \ldots, x_n \in S$. Here, $\pi(\cdot)$ and $P(.,.)$ are supposed to model the language. The listener hears sounds $(Y_1, \ldots, Y_n)$ such that

$$P[Y_1 = y - 1, \ldots, Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n] = Q(x_1, y_1) \cdots Q(x_n, y_n).$$

This model is called a *hidden Markov chain* model: The string that the speaker pronounces is a Markov chain that is hidden from the listener who cannot read her mind but instead only hears the sounds. We want to calculate $MAP[X|Y]$. Note that

$$\begin{aligned} P[X = x | Y = y] &= \frac{P(X = x) P[Y = y | X = x]}{P(X = x)} \\ &= \pi(x_1) P(x_1, x_2) \cdots P(x_{n-1}, x_n) Q(x_1, y_1) \cdots Q(x_n, y_n) / P(X = x). \end{aligned}$$

Consequently,

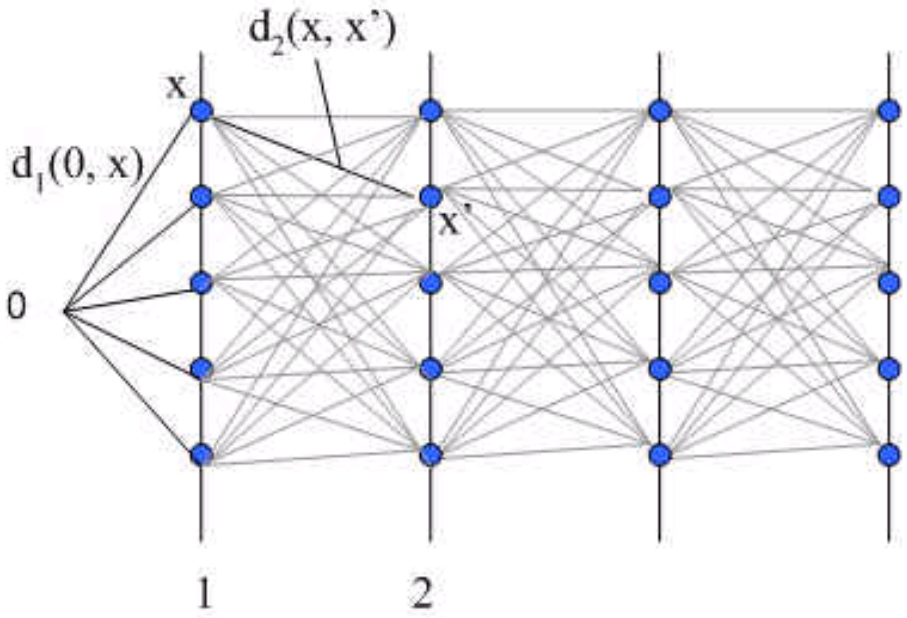


Figure 16.2: Shortest path in speech recognition

$$MAP[X|Y = y] = \arg\max \pi(x_1)P(x_1, x_2) \cdots P(x_{n-1}, x_n)Q(x_1, y_1) \cdots Q(x_n, y_n)$$

where the maximization is over $x \in S_n$. To maximize this product, we minimize the negative of its logarithm, which we write as $d_1(0, x_1) + d_2(x_1, x_2) + \cdots + d_n(x_{n-1}, x_n)$ where

$$d_1(0, x) = -\log(\pi(x)Q(x, y_1))$$

and

$$d_k(x, x') = -\log(P(x, x')Q(x', y_k))$$

for $x, x' \in S$ and $k = 1, 2, \ldots, n$. (Recall that y is given.)

Minimizing $d_1(0, x_1) + d_2(x_1, x_2) + \cdots + d_n(x_{n-1}, x_n)$ is equivalent to finding the shortest path in the graph shown in Figure 16.2.

A shortest path algorithm is due Bellman-Ford. Let $A(x,k)$ be the length of the shortest path from 0 to x at step k. Then we can calculate recursively

$$A(x,k+1) = \min\{A(x',k) + d_{k+1}(x',x)\}$$

where the minimum is over all x' in S. This algorithm, applied to calculate the MAP is called Viterbi's algorithm.

16.5 A Simple Game

Consider the following "matching pennies" game. Alice and Bob both have a penny and they select which face they show. If they both show the same face, Alice wins \$1.00 from Bob. If they show different faces, Bob wins \$1.00 from Alice. Intuitively it is quite clear that the best way to play the game is for both players to choose randomly and with equal probabilities which face to show. This strategy constitutes an "equilibrium" in the sense that no player has an incentive to deviate unilaterally from it. Indeed, if Bob plays randomly (50/50), then the average reward of Alice is 0 no matter how she plays, so she might as well play randomly (50/50). It is also not hard to see that this is the only equilibrium. Such an equilibrium is called a *Nash Equilibrium*

This example is a particular case of a general type of games that can be described as follows. If Alice chooses the action $a \in A$ and Bob the action $b \in B$, then Alice gets the payoff $A(a,b)$ and Bob the payoff $B(a,b)$. If Alice and Bob choose a and b randomly and independently in A and B, then they get the corresponding expected rewards. Nash proved the remarkable result that if A and B are finite, then there must be at least one random choice that is an equilibrium, thus generalizing the matching pennies result.

16.6 Decisions

One is given a perfectly shuffled 52-card deck. The cards will be turned over one at a time. You must try to guess when an ace is about to be turned over. If you guess correctly, you win \$1.00, otherwise you lose. One strategy might be to wait until a number of cards are turned over; if you are lucky, the fraction of aces left in the deck will get larger than 4/52 and this will increase your odds of guessing right. Unfortunately, things might go the other way and you might see a number of aces being turned over quickly, thus reducing your odds of winning. What should you do?

A simple argument shows that it does not really matter how you play. To see this, designate by $V(n, m)$ your maximum expected reward given that there remain m aces and a total of n cards in the deck. By maximum, we mean the expected reward you can get by playing the game in the best possible way. The key to the analysis is to observe that you have two choices as the game starts with n cards and m aces: either you gamble on the next card or you don't. If you do, that next card is an ace with probability m/n and you win \$1.00 with that probability. If you don't, then, after the next card is turned over, with probability m/n you find yourself with a deck of $n-1$ cards with $m-1$ aces; with probability $1-m/n$, you face a deck of $n-1$ cards and m aces. Accordingly, if you play the game optimally after the next card, your expected reward is $(m/n)V(n-1, m-1) + (1-m/n)V(n-1, m)$. Hence, the maximum reward $V(n, m)$ should be either m/n (if you gamble on the next card) or $(m/n)V(n-1, m-1) + (1-m/n)V(n-1, m)$ (if you don't). We conclude that

$$V(n, m) = \max\{m/n, (m/n)V(n-1, m-1) + (1-m/n)V(n-1, m)\}. \tag{16.6.1}$$

You can verify that the solution of the above equations is $V(n, m) = m/n$ whenever $n > 0$. Thus, both alternatives (gamble on the next card or don't) yield the same expected reward.

The equations (16.6.1) are called the *Dynamic Programming Equations* (DPE). They

express the maximum expected reward by comparing the consequences of the different decisions that can be made at a stage of the game. The trick to write down that equation is to identify correctly the "state" of the game (here, the pair (n, m)). By solving the DPE and choosing at each stage the decision that corresponds to the maximum, one derives the optimal strategy. These ideas are due to Richard Bellman. (I borrowed this simple example from the great little book by Sheldon Ross [6].)

Appendix A

Mathematics Review

A.1 Numbers

A.1.1 Real, Complex, etc

You are familiar with whole, rational, real, and complex numbers. You know how to perform operations on complex numbers and how to convert them to and from the $r \times e^{i\theta}$ notation. Recall that $|a + ib| = \sqrt{a^2 + b^2}$.

For instance, you can show that

$$\frac{3+i}{1-i} = 1 + 2i$$

and that

$$2 + i = \sqrt{5}e^{i\pi/6}.$$

A.1.2 Min, Max, Inf, Sup

Let A be a set of real numbers. An *upper bound* of A is a finite number b such that $b \geq a$ for all a in A. If there is an upper bound of A that is in A, it is called the *maximal element* of A and is designated by $\max\{A\}$. If A has an upper bound, it has a *lowest upper bound*

that is designated by $\sup\{A\}$. One defines a lower bound, the minimal element $\min\{A\}$, and the greatest lower bound $\inf\{A\}$ similarly.

For instance, let $A = (2, 5]$. Then 6 is an upper bound of A, 1 is a lower bound, $5 = \max\{A\} = \sup\{A\}$, $2 = \inf\{A\}$ and A has no minimal element.

For any real number x one defines $x^+ = \max\{x, 0\}$ and $x^- = (-x)^+$. Note that $|x| = x^+ + x^-$. We also use the notation $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. For instance, $(-5)^+ = 0, (-5)^- = 5, 3 \vee 6 = 6$, and $3 \wedge 6 = 3$.

A.2 Summations

You recall the notations

$$\sum_{n=0}^{N} x_n \text{ and } \Pi_{n=0}^{N} x_n$$

and you can calculate the corresponding expressions for some specific examples of sequences $\{x_n, n \geq 1\}$. For instance, you remember and you can prove that if $a \neq 1$, then

$$\sum_{n=0}^{N} a^n = \frac{1 - a^{N+1}}{1 - a}.$$

You also remember that

$$\sum_{n=0}^{\infty} x_n := \lim_{N \to \infty} \sum_{n=0}^{N} x_n$$

when the limit exists. For instance, you remember and you can prove that if $|a| < 1$, then

$$\sum_{n=0}^{\infty} a^n = \frac{1}{1 - a}.$$

By taking the derivative of the above expression with respect to a, you find that, when $|a| < 1$,

$$\sum_{n=0}^{\infty} n a^{n-1} = \frac{1}{(1 - a)^2}.$$

By taking the derivative one more time, we get

$$\sum_{n=0}^{\infty} n(n-1) a^{n-2} = \frac{2}{(1 - a)^3}.$$

It is sometimes helpful to exchange the order of summation. For instance,

$$\sum_{n=0}^{N}\sum_{m=0}^{n} x_{m,n} = \sum_{m=0}^{N}\sum_{n=m}^{N} x_{m,n}.$$

A.3 Combinatorics

A.3.1 Permutations

There are $n!$ ways to order n distinct elements, where

$$n! = 1 \times 2 \times 3 \times \cdots \times n.$$

By convention, $0! = 1$.

For instance, there are 120 ways of seating 5 people at a table with 5 chairs.

A.3.2 Combinations

There are $\binom{N}{n}$ distinct groups of n objects selected without replacement from a set of N distinct objects, where

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

For instance, there are about 2.6×10^6 distinct sets of five cards picked from a 52-card deck.

You remember that

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

A.3.3 Variations

You should be able to apply these ideas and their variations. For instance, you can count the number of strings of five letters that have exactly one E.

A.4 Calculus

You remember the meaning of

$$\int_a^b f(x)dx.$$

In particular, you know how to calculate some simple integrals. You recall that, for $n = 0, 1, 2, \ldots$,

$$\int_0^1 x^n dx = 1/(n+1).$$

Also,

$$\int_1^A \frac{1}{x} dx = \ln A.$$

You know the integration by parts formula and you can calculate

$$\int_0^y x^n e^x dx.$$

A useful fact is that, for any complex number a,

$$(1 + \frac{a}{n})^n \to e^a \text{ as } n \to \infty.$$

You also remember that

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots$$

and

$$\ln(1+x) \approx x \text{ whenever } |x| \ll 1.$$

A.5 Sets

A set is a well-defined collection of elements. That is, for every element one can determine whether it is in the set or not. Recall the notation $x \in A$ meaning that x is an element of the set A. We also say that x belongs to A.

It is usual to characterize a set by a proposition that its elements satisfy. For instance one can define

$$A = \{x \mid 0 < x < 1 \text{ and } x \text{ is a rational number}\}.$$

You recall the definition

$$A \cap B = \{x \mid x \in A \text{ and } a \in B\}.$$

Similarly, you know how to define $A \cup B, A \setminus B$, and $A \Delta B$. You also know the meaning of $A \subset B$ and of the complement A^c of A.

You can show that

$$(A \cup B)^c = A^c \cap B^c \text{ and } (A \cap B)^c = A^c \cup B^c.$$

You are not confused by the notation and you would never write $[1, 2] \in [0, 3]$ because you know that $[1, 2] \subset [0, 3]$. Similarly, you would never write $1 \subset [0, 3]$ but you would write $1 \in [0, 3]$ or $\{1\} \subset [0, 3]$. Along the same lines, you know that $0 \in [0, 3]$ but you would never write $0 \in (0, 3]$.

You could meditate on the meaning of

$$S = \{x \mid x \text{ is not a member of } x\}.$$

A.6 Countability

A set A is *countable* if it is finite or if one can enumerate its elements as $A = \{a_1, a_2, a_3, \ldots\}$. A subset of a countable set is countable. If the sets A_n are countable for $n \geq 1$, then so is their union

$$A = \cup_{n=1}^{\infty} A_n := \{a \mid a \in A_n \text{ for some } n \geq 1\}.$$

The *cartesian product* $A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}$ of countable sets is countable. The set of rational numbers is countable.

The set $[0, 1]$ is not countable. To see that,k imagine that one can enumerate its elements and write them as decimal expansions, for instance as $\{0.23145..., 0.42156.., 0.13798..., \ldots\}$. This list does not contain the number 0.135... selected in a way that the first digit in the expansion differs from the first digit of the first element in the list, its second digit differs from that of the second element, and so on. This *diagonal argument* shows that there is no possible list that contains all the elements of $[0, 1]$.

A.7 Basic Logic

A.7.1 Proof by Contradiction

Let p and q be two propositions. We say "if p then q" if the proposition q is true whenever p is. For instance, if p means "it rains" and q means "the roof gets wet," then we can postulate "if p then q."

You know that if the statement "if p then q" is true, then so is the statement "if not q then not p." However, the statement "if not p then not q" may not be true.

Therefore, if we know that the statement "if p then q" is true, a method for proving "not p" is to prove "not q".

As an example, let us prove by contradiction that the statement "$\sqrt{2}$ is irrational" is true. Let p designate the statement "$\sqrt{2}$ is rational." We know that "if p then q" where q is the statement "$\sqrt{2} = a/b$ where a and b are integers". We will prove that "not q" is true. To do this, assume that q is true, i.e., that $\sqrt{2} = a/b$. We can simplify that fraction until a and b are not both multiples of 2. Taking the square, we get $2 = a^2/b^2$. This implies that $a^2 = 2b^2$ is even, which implies that a is even and that b is not (since a and b are not both multiples of 2). But then $a = 2c$ and $a^2 = 4c^2 = 2b^2$, which shows that $b^2 = 2c^2$ is even, so

that b must also be even, which contradicts our assumption.

A.7.2 Proof by Induction

Assume that for $n \geq 1$, $p(n)$ designates a proposition. The induction method for proving that $p(n)$ is true for all finite $n \geq 1$ consists in showing first that $p(1)$ is true and second that if $p(n)$ is true, then so is $p(n+1)$. The second step is called the *induction step.*

As an example, we show that if $a \neq 1$ then $a + a^2 + \cdots + a^N = (a - a^{N+1})/(1-a)$. The identity is certainly true for $n = 1$. Assume that it is true for some N. Then

$$\begin{aligned} a &+ a^2 + \cdots + a^N + a^{N+1} \\ &= (a - a^{N+1})/(1-a) + a^{N+1} \\ &= (a - a^{N+1})/(1-a) + (a^{N+1} - a^{N+2})/(1-a) \\ &= (a - a^{N+2})/(1-a), \end{aligned}$$

which proves the identity for $N + 1$.

If $p(n)$ is true for all finite n, this does not imply that $p(\infty)$ is true, even if $p(\infty)$ is well-defined. For instance, the set $\{1, 2, \ldots, n\}$ is finite for all finite n, but $\{1, 2, \ldots\}$ is infinite.

A.8 Sample Problems

Problem A.8.1. *Express* $(1+3i)/(2+i)$ *in the form* $a + bi$ *and in the form* $r \times e^{i\theta}$.

Problem A.8.2. *Prove by induction that*

$$\sum_{k=1}^{n} k^3 = (\sum_{k=1}^{n} k)^2.$$

Note: We want a proof by induction, not a direct proof. You may use the fact that

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

Problem A.8.3. *Give an example of a function $f(x)$ defined on $[0,1]$ such that*

$$\sup_{0\leq x\leq 1} f(x) = 1 \text{ and } \inf_{0\leq x\leq 1} f(x) = 0$$

and the function $f(x)$ does not have a maximum on $[0,1]$.

Problem A.8.4. *Calculate $\int_0^1 \frac{x+1}{x+2}dx$.*

Problem A.8.5. *Which of the following is/are true?*

1. $0 \in (0,1)$
2. $0 \subset (-1,3)$
3. $(0,1) \cup (1,2) = (0,2)$
4. *The set of integers is uncountable.*

Problem A.8.6. *Calculate $\int_0^\infty x^2 e^{-x} dx$.*

Problem A.8.7. *Let $A = (1,5), B = [0,3)$, and $C = (2,4)$. What is $A \setminus (B \Delta C)$?*

Problem A.8.8. *Calculate*

$$\sum_{m=0}^{N} \sum_{n=m}^{N} \frac{1}{n+1}.$$

Problem A.8.9. *Let $A = [3, 4.7)$. What are*

$$\min\{A\}, \max\{A\}, \inf\{A\}, \text{ and } \sup\{A\}?$$

Problem A.8.10. *Let A be a set of numbers and define $B = \{-a | a \in A\}$. Show that* $\inf\{A\} = -\sup\{B\}$.

Problem A.8.11. *Calculate, for* $|a| < 1$,

$$\sum_{n=0}^{\infty} na^n \text{ and } \sum_{n=0}^{\infty} n^2 a^n.$$

Problem A.8.12. *How many distinct sets of five cards with three red cards can one draw from a deck of 52 cards?*

Problem A.8.13. *Let* A *be a set of real numbers with an upper bound* b. *Show that* $\sup\{A\}$ *exists.*

Problem A.8.14. *Derive the expression for* $\sum_{n=0}^{N} a^n$.

Problem A.8.15. *Let* $\{x_n, n \geq 1\}$ *be real numbers such that* $x_n \leq x_{n+1}$ *and* $x_n \leq a < \infty$ *for* $n \geq 1$. *Prove that* $x_n \to x$ *as* $n \to \infty$ *for some* $x < \infty$.

Problem A.8.16. *Show that the set of finite sentences in English is countable.*

Appendix B

Functions

A *function* $f(\cdot)$ is a mapping from a set D into another set V. To each point x of D, the function attaches a single point $f(x)$ of V.

The function $f : D \to V$ is said to be *one-to-one* if $x \neq y$ implies $f(x) \neq f(y)$. The function is *onto* if $\{f(x)|x \in D\} = V$. The function is a *bijection* if it is onto and one-to-one.

For any function $f : D \to V$ one can define the inverse of a set $A \subset V$ by

$$f^{-1}(A) = \{x \in D|f(x) \in A\}.$$

Figures B.1 and B.2 illustrate these notions.

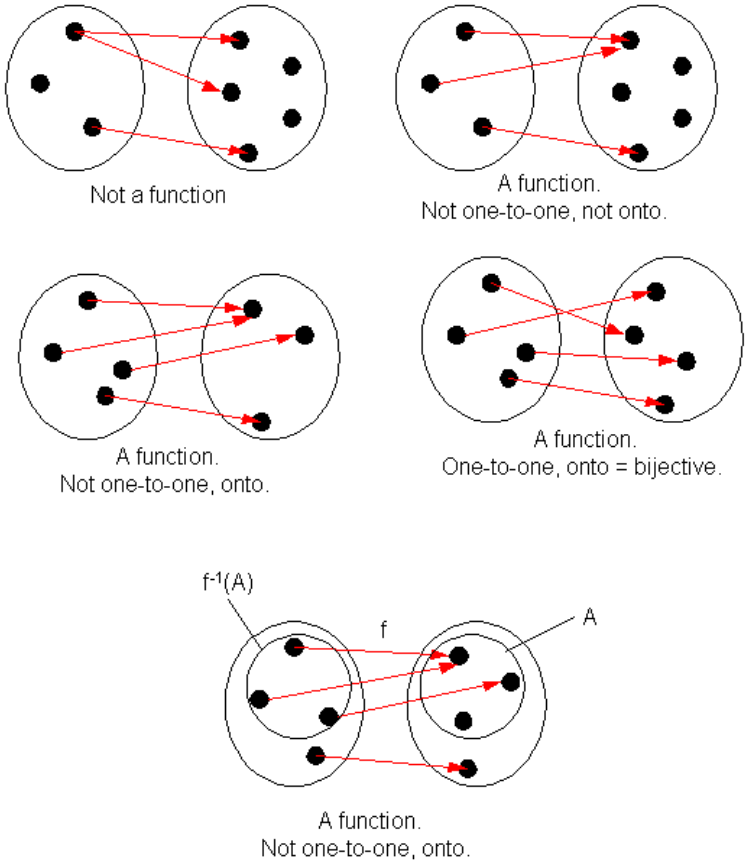


Figure B.1: Examples of mappings

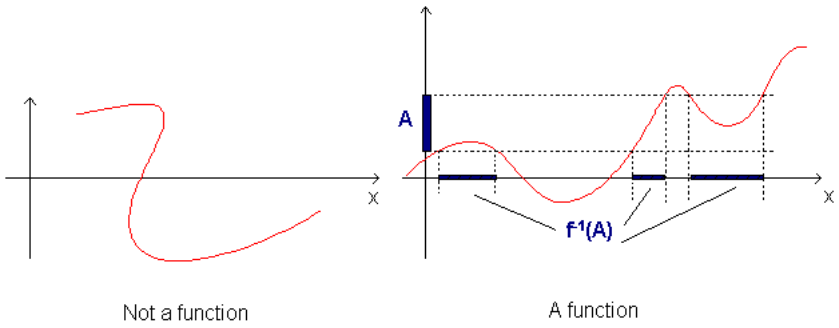


Figure B.2: Graphs

Appendix C

Nonmeasurable Set

C.1 Overview

We defined events as being some sets of outcomes. The collection of events is closed under countable set operations. When the sample space is countable, we can define the probability of every set of outcomes. However, in general this is not possible. For instance, one cannot define the length of every subset of the real line. We explain that fact in this note. These ideas a bit subtle. We explain them only because some students always ask for a justification.

C.2 Outline

We construct a set S of real numbers between 0 and 1 with the following properties. Define $S_x = \{y + x | y \in S\}$. That is, S_x is the set S shifted by x. Then there is a countable collection C of numbers such that the union A of S_x for x in C is such that $[1/3, 2/3]$ is a subset of A and A is a subset of $[0, 1]$. Morever, S_x and S_y are disjoint whenever x and y are distinct in C.

Assume that the "length" of S is L. The length of S_x is also L for all x. (Indeed, the length is first defined for intervals and is shift-invariant.) The length of A must be the sum

of the length of the sets S_x for x in C since it is a countable union of these disjoint sets. If $L > 0$, then the length of A is infinite, which is not possible since A is contained in $[0, 1]$. If $L = 0$, then the length of A must be 0, which is not possible since A contains $[1/3, 2/3]$. Thus the length of S cannot be defined. It remains to construct S. We do that next.

C.3 Constructing S

We start by defining x and y in $[0, 1/3]$ to be equivalent if they differ by a rational number. For instance, $x = (20.5)/8$ and $x + 0.12$ are equivalent. We can then look at all the equivalence classes of $[0, 1/3]$, i.e., all the sets of equivalent numbers. Two different equivalence classes must be disjoint. We then form a set S by picking one element from each equivalence class. [Some philosophers will object to this selection, arguing that it is not reasonable. They refuse the axiom of choice that postulates that such as set is well defined.] Note that all the numbers in S are in $[0, 1/3]$ and that any two numbers in S cannot be equivalent since they were picked in different equivalence classes. That is, any two numbers x, y in S differ by an irrational number. Moreover, any number u in $[0, 1/3]$ is equivalent to some s in S, since S contains a representative of all the equivalence classes of points in $[0, 1/3]$.

Next, let C be all the rational numbers in $[0, 2/3]$. For x in C, S_x is a subset of $[0, 1]$. Also, for any two distinct x, y in C the sets S_x and S_y must be disjoint. Otherwise, they contain a common element z such that $z = u + x = v + y$ for some distinct u and v in S; but this implies that $x - y = v - u$ is rational, which is not possible. It remains only to show that the union of the sets S_x for x in C contains $[1/3, 2/3]$. Pick any number w in $[1/3, 2/3]$. Note that $w - 1/3$ is in $[0, 1/3]$ and must be equivalent to some s in S. That implies that $x = w - s$ is rational and it is in $[0, 2/3]$, and therefore in C. Thus, w is in S_x for some x in C.

Appendix D

Key Results

We cover a number of important results in this course. In the table below we list these results. We indicate the main reference and their applications. In the table, Ex. refers to an Example, S. to a Section, C. to a Chapter, and T. to a Theorem.

Result	Main Discussion	Applications
Bayes' Rule	S. 3.3	Ex. 3.6.2
Borel-Cantelli	T. 2.7.10	S. 10.3.1
Chebyshev's $\leq$	(4.8.1)	Ex. 4.10.19, 5.5.12, **??**
CLT	S. 11.3 , T. 11.4.1	Ex. 11.7.5, 11.7.6, 11.7.8, 11.7.9, 11.7.10, 12.2.5; S. 11.5, 12.1.10
Continuity of $P(\cdot)$	S. 2.3	(4.2.2); T. 2.7.10
Convergence	S. 10.3, S. 10.4, S. 10.5	Ex. 11.7.1; S. 10.6
Coupling	Ex. 14.8.14	S. 12.2.9
$E[X\|Y]$	C. 6, (6.4.1), T. 6.4.2	Ex. 8.6.6, 9.6.8, 9.6.10; S. 6.2, 6.7
FSE	S. 12.1.7	Ex. 14.8.4, 14.8.6, 14.8.7; S. 12.1.8, 14.5, 14.8 , 14.8.11, 14.8.12 , 14.8.14, 15.7.1, 15.7.2, 15.7.6
Gaussian	C. 7, (7.1)	T. 7.3.1; S. 7.5
$HT[X\|Y]$	C. 8, T. 8.3.1, T. 8.3.2	Ex. 8.6.3-8.6.5, 8.6.7, 8.6.8, 8.6.11, 11.7.7, 11.7.8, 11.7.9; S. 8.3.2
Independence	S. 3.4.4, S. 5.3, T. 5.3.1	Ex. 3.6.5, 4.10.10; S. 5.5
Lebesgue C.T.	T. 10.7.1	Ex. 11.7.1
Linearity $E(.)$	(4.6.3)	(5.2.1), (5.2.2)
LLSE	S. 9.2	Ex. 9.6.1, 9.6.3, 9.6.5, 9.6.6, 9.6.7, 9.6.10, 9.6.12, 12.2.1
Markov Chain	C. 14, C. 15	S. 14.8
MAP, MLE	(8.1.2), S. 8.2	Ex. 8.6.1, 8.6.2, 8.6.7, 8.6.8, 8.6.10, 9.6.2, 9.6.9, 9.6.11, 12.2.1
Memoryless	S. 12.1.5, (4.3.4), (4.3.8)	Ex. 15.7.4, 15.7.5
$\{\Omega, \mathcal{F}, P\}$	S. 2.4	S. 2.5, S. 2.7, Ex. 4.10.2
SLLN	S. 11.2	Ex. 11.7.2, 11.7.3, 11.7.10.a; S. 12.1.9, 12.2.3
Sufficient Statistics	S. 9.4	Ex. 9.6.4, 9.6.11
Symmetry		Ex. 7.5.8, 7.5.8, 9.6.9, 12.2.1, 12.2.4, 7.5.1
Transforms	S. 10.2	S. 7.1.1, 7.2.2

Appendix E

Bertrand's Paradox

The point of this note is that one has to be careful about the meaning of "choosing at random."

Consider the following question: What is the probability that a chord selected at random in a circle is larger than the side of an inscribed equilateral triangle? There are three plausible answers to this question: 1/2, 1/3, and 1/4. Of course, the answer depends on how we choose the chord.

Answer 1: 1/3

The first choice is shown in the left-most part of Figure E.1. To choose the chord, we fix a point A on the circle; it will be one of the ends of the chord. We then choose another point X at random on the circumference of the circle. If X happens to be between B and C (where ABC is equilateral), then AX is longer than the sides of ABC. Thus, the requested probability is 1/3.

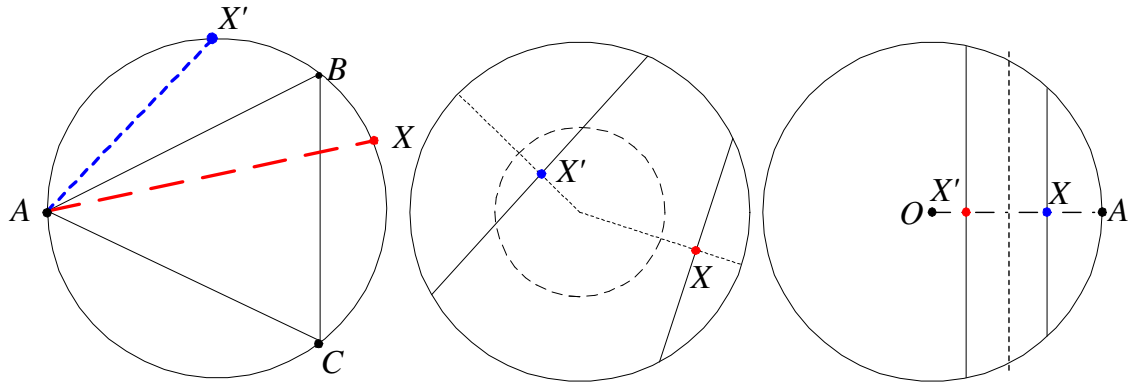


Figure E.1: Three ways of choosing a chord.

Answer 2: 1/4

The second choice is illustrated in the middle part of Figure E.1. We choose the chord by choosing its midpoint (e.g., X) at random inside the circle. The chord is longer than the side of the inscribed equilateral triangle if and only if X falls inside the circle with half the radius of the original circle, which happens with probability 1/4.

Answer 2: 1/2

The third choice is illustrated in the right-most part of Figure E.1. We choose the chord by choosing its midpoint (e.g., X) at random on a given radius OA of the circle. The chord is longer than the side of the inscribed triangle if an only if the point is closer to the center than half a radius, which happens with probability 1/2.

Appendix F

Simpson's Paradox

The point of this note is that proportions do not add up and that one has to be careful with statistics.

Consider a university where 80% of the male applicants are accepted but only 51% of the female applicants are accepted. You will be tempted to conclude that the university discriminates against female applicants. However, a closer look at this university shows that it has only two colleges with the admission records shown in the table.

Note that each college admits a larger fraction of female applicants than of male applicants, so that the university cannot be accused of discrimination against the female students. It happens that more female students apply to a more difficult college.

College	F. Appl.	F. Adm.	% F. Adm.	M. Appl.	M. Adm.	% M. Adm.
A	980	490	50%	200	80	40%
B	20	20	100%	800	720	90%
Total	1000	510	51%	1000	800	80%

Appendix G

Familiar Distributions

We collect here the few distributions that we encounter repeatedly in the text.

G.1 Table

Distribution	Shorthand	Definition	Mean	Variance
Bernoulli	$B(p)$	1 w.p. p; 0 w.p. $1-p$	p	$p(1-p)$
Binomial	$B(n,p)$	m w.p. $\binom{n}{m}p^m(1-p)^{n-m}, m=0,\ldots,n$	np	$np(1-p)$
Geometric	$G(p)$	m w.p. $p(1-p)^{m-1}, m \geq 1$	$1/p$	$p^{-2}-p^{-1}$
Poisson	$P(\lambda)$	m w.p. $\lambda^m e^{-\lambda}/m!$	λ	λ
Uniform	$U[0,1]$	$f_X(x) = 1\{0 \leq x \leq 1\}$	$1/2$	$1/12$
Exponential	$Exd(\lambda)$	$f_X(x) = \lambda e^{-\lambda x} 1\{x \geq 0\}$	λ^{-1}	λ^{-2}
Std. Gaussian	$N(0,1)$	$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \Re$	0	1

G.2 Examples

Here are typical random experiments that give rise to these distributions. We also comment on the properties of those distributions.

- Bernoulli: Flip a coin; $X = 1$ if outcome is H, $X = 0$ if it is T.

- Binomial: Number of Hs when flipping n coins.
- Geometric: Number of flips until the first H. Memoryless. Holding time of state of discrete-time Markov chain.
- Poisson: Number of photons that hit a given area in a given time interval. Limit of $B(n,p)$ as $np = \lambda$ and $n \to \infty$. The sum of independent $P(\lambda_i)$ is $P(\sum_i \lambda_i)$. Random sampling (coloring) of a $P(\lambda)$-number of objects yields independent Poisson random variables.
- Uniform: A point picked "uniformly" in $[0,1]$. Returned by function "random(.)". Useful to generate random variables.
- Exponential: Time until next photon hits. Memoryless. Holding time of state of continuous-time Markov chain. Limit of $G(p)/n$ as $np = \lambda$ and $n \to \infty$. The minimum of independent $Exd(\lambda_i)$ is $Exd(\sum_i \lambda_i)$.
- Gaussian: Thermal noise. Sum of many small independent random variables (CLT). By definition, $\mu + \sigma N(0,1) =_D N(\mu, \sigma^2)$. The sum of independent $N(\mu_i, \sigma_i^2)$ is $N(\sum_i \mu_i, \sum_i \sigma_i^2)$.

Index

Bibliography

[1] L. Breiman, *Probability*, Addison-Wesley, Reading, Mass, 1968.

[2] P. Bremaud, *An introduction to probabilistic modeling*, Springer Verlag, 1988.

[3] W. Feller, *Introduction to probability theory and its applications*, Wiley, New York.

[4] Port S.C. Hoel, P.G. and C. J. Stone, *An introduction to probability theory*, Houghton Mifflin, 1971.

[5] J. Pitman, *Probability*, Springer-Verlag, 1997.

[6] S. Ross, *Introduction to stochastic dynamic programming*, Academic Press, New York, NY, 1984.

[7] ———, *Introduction to probability models, seventh edition*, Harcourt, Academic Press, Burlington, MA, 2000.

[8] Chisyakov V.P. Sevastyanov, B. A. and A. M. Zubkov, *Problems in the theory of probability*, MIR Publishers, Moscow, 1985.

[9] S. M. Stigler, *The history of statistics – the measurement of uncertainty before 1900*, Belknap, Harvard, 1999.