SIMULATION AND THE MONTE CARLO METHOD

SIMULATION AND THE MONTE CARLO METHOD

Reuven Y. Rubinstein
Technion

Dirk P. Kroese
University of Queensland

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

ISBN 978-0-470-17793-8

To my friends and colleagues Søren Asmussen and Peter Glynn

— *RYR*

In memory of my parents Albert and Anna Kroese

— *DPK*

CONTENTS

CHAPTER 1

PRELIMINARIES

1.1 RANDOM EXPERIMENTS

The basic notion in probability theory is that of a *random experiment*: an experiment whose outcome cannot be determined in advance. The most fundamental example is the experiment where a fair coin is tossed a number of times. For simplicity suppose that the coin is tossed three times. The *sample space*, denoted Ω, is the set of all possible outcomes of the experiment. In this case Ω has eight possible outcomes:

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

where, for example, HTH means that the first toss is heads, the second tails, and the third heads.

Subsets of the sample space are called *events*. For example, the event A that the third toss is heads is

$$A = \{HHH, HTH, THH, TTH\} .$$

We say that event A *occurs* if the outcome of the experiment is one of the elements in A. Since events are sets, we can apply the usual set operations to them. For example, the event $A \cup B$, called the *union* of A and B, is the event that A or B or both occur, and the event $A \cap B$, called the *intersection* of A and B, is the event that A and B both occur. Similar notation holds for unions and intersections of more than two events. The event A^c, called the *complement* of A, is the event that A does not occur. Two events A and B that have no outcomes in common, that is, their intersection is empty, are called *disjoint* events. The main step is to specify the probability of each event.

Simulation and the Monte Carlo Method, Second Edition. By R.Y. Rubinstein and D. P. Kroese

Definition 1.1.1 (Probability) A *probability* $\mathbb{P}$ is a rule that assigns a number $0 \leqslant \mathbb{P}(A) \leqslant 1$ to each event A, such that $\mathbb{P}(\Omega) = 1$, and such that for any sequence $A_1, A_2, \ldots$ of disjoint events

$$\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i) \; . \tag{1.1}$$

Equation (1.1) is referred to as the *sum rule* of probability. It states that if an event can happen in a number of different ways, but not simultaneously, the probability of that event is simply the sum of the probabilities of the comprising events.

For the fair coin toss experiment the probability of any event is easily given. Namely, because the coin is fair, each of the eight possible outcomes is equally likely, so that $\mathbb{P}(\{HHH\}) = \cdots = \mathbb{P}(\{TTT\}) = 1/8$. Since any event A is the union of the "elementary" events $\{HHH\}, \ldots, \{TTT\}$, the sum rule implies that

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \; , \tag{1.2}$$

where $|A|$ denotes the number of outcomes in A and $|\Omega| = 8$. More generally, if a random experiment has finitely many and equally likely outcomes, the probability is always of the form (1.2). In that case the calculation of probabilities reduces to counting.

1.2 CONDITIONAL PROBABILITY AND INDEPENDENCE

How do probabilities change when we know that some event $B \subset \Omega$ has occurred? Given that the outcome lies in B, the event A will occur if and only if $A \cap B$ occurs, and the relative chance of A occurring is therefore $\mathbb{P}(A \cap B)/\mathbb{P}(B)$. This leads to the definition of the *conditional probability* of A given B:

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \; . \tag{1.3}$$

For example, suppose we toss a fair coin three times. Let B be the event that the total number of heads is two. The conditional probability of the event A that the first toss is heads, given that B occurs, is $(2/8)/(3/8) = 2/3$.

Rewriting (1.3) and interchanging the role of A and B gives the relation $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \,|\, A)$. This can be generalized easily to the *product rule* of probability, which states that for any sequence of events $A_1, A_2, \ldots, A_n$,

$$\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}) \; , \tag{1.4}$$

using the abbreviation $A_1 A_2 \cdots A_k \;\equiv\; A_1 \cap A_2 \cap \cdots \cap A_k$.

Suppose $B_1, B_2, \ldots, B_n$ is a *partition* of Ω. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is Ω. Then, by the sum rule, $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \cap B_i)$ and hence, by the definition of conditional probability, we have the *law of total probability*:

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i) \; . \tag{1.5}$$

Combining this with the definition of conditional probability gives *Bayes' rule*:

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\,\mathbb{P}(B_j)}{\sum_{i=1}^n \mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i)} \; . \tag{1.6}$$

Independence is of crucial importance in probability and statistics. Loosely speaking, it models the lack of information between events. Two events A and B are said to be *independent* if the knowledge that B has occurred does not change the probability that A occurs. That is, A, B independent $\Leftrightarrow \mathbb{P}(A \,|\, B) = \mathbb{P}(A)$. Since $\mathbb{P}(A \,|\, B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$, an alternative definition of independence is

$$A, B \text{ independent} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \ .$$

This definition covers the case where $B = \emptyset$ (empty set). We can extend this definition to arbitrarily many events.

Definition 1.2.1 (Independence) The events $A_1, A_2, \ldots,$ are said to be *independent* if for any k and any choice of distinct indices $i_1, \ldots, i_k$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\,\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) \ .$$

Remark 1.2.1 In most cases, independence of events is a model assumption. That is, we assume that there exists a $\mathbb{P}$ such that certain events are independent.

■ **EXAMPLE 1.1**

We toss a biased coin n times. Let p be the probability of heads (for a fair coin $p = 1/2$). Let A_i denote the event that the i-th toss yields heads, $i = 1, \ldots, n$. Then $\mathbb{P}$ should be such that the events $A_1, \ldots, A_n$ are independent, and $\mathbb{P}(A_i) = p$ for all i. These two rules completely specify $\mathbb{P}$. For example, the probability that the first k throws are heads and the last $n - k$ are tails is

$$\begin{aligned} \mathbb{P}(A_1 \cdots A_k A_{k+1}^c \cdots A_n^c) &= \mathbb{P}(A_1) \cdots \mathbb{P}(A_k)\,\mathbb{P}(A_{k+1}^c) \cdots \mathbb{P}(A_n^c) \\ &= p^k (1-p)^{n-k}. \end{aligned}$$

1.3 RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

Specifying a model for a random experiment via a complete description of Ω and $\mathbb{P}$ may not always be convenient or necessary. In practice we are only interested in various observations (that is, numerical measurements) in the experiment. We incorporate these into our modeling process via the introduction of *random variables*, usually denoted by capital letters from the last part of the alphabet, e.g., $X, X_1, X_2, \ldots, Y, Z$.

■ **EXAMPLE 1.2**

We toss a biased coin n times, with p the probability of heads. Suppose we are interested only in the number of heads, say X. Note that X can take any of the values in $\{0, 1, \ldots, n\}$. The *probability distribution* of X is given by the *binomial formula*

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n \ . \tag{1.7}$$

Namely, by Example 1.1, each elementary event $\{HTH \cdots T\}$ with exactly k heads and $n - k$ tails has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ such events.

The probability distribution of a general random variable X — identifying such probabilities as $\mathbb{P}(X = x), \mathbb{P}(a \leqslant X \leqslant b)$, and so on — is completely specified by the *cumulative distribution function* (cdf), defined by

$$F(x) = \mathbb{P}(X \leqslant x), \;\; x \in \mathbb{R} \; .$$

A random variable X is said to have a *discrete* distribution if, for some finite or countable set of values $x_1, x_2, \ldots$, $\mathbb{P}(X = x_i) > 0$, $i = 1, 2, \ldots$ and $\sum_i \mathbb{P}(X = x_i) = 1$. The function $f(x) = \mathbb{P}(X = x)$ is called the *probability mass function* (pmf) of X — but see Remark 1.3.1.

■ EXAMPLE 1.3

Toss two fair dice and let M be the largest face value showing. The pmf of M is given by

m	1	2	3	4	5	6	$\sum$
$f(m)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

For example, to get $M = 3$, either $(1, 3), (2, 3), (3, 3), (3, 2)$, or $(3, 1)$ has to be thrown, each of which happens with probability $1/36$.

A random variable X is said to have a *continuous* distribution if there exists a positive function f with total integral 1, such that for all a, b

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f(u)\, du \; . \tag{1.8}$$

The function f is called the *probability density function* (pdf) of X. Note that in the continuous case the cdf is given by

$$F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)\, du \; ,$$

and f is the derivative of F. We can interpret $f(x)$ as the probability “density” at $X = x$ in the sense that

$$\mathbb{P}(x \leqslant X \leqslant x + h) = \int_x^{x+h} f(u)\, du \approx h\, f(x) \; .$$

Remark 1.3.1 (Probability Density) Note that we have deliberately used the *same* symbol, f, for both pmf and pdf. This is because the pmf and pdf play very similar roles and can, in more advanced probability theory, both be viewed as particular instances of the general notion of *probability density*. To stress this viewpoint, we will call f in *both* the discrete and continuous case the pdf or (probability) density (function).

1.4 SOME IMPORTANT DISTRIBUTIONS

Tables 1.1 and 1.2 list a number of important continuous and discrete distributions. We will use the notation $X \sim f$, $X \sim F$, or $X \sim \mathsf{Dist}$ to signify that X has a pdf f, a cdf F or a distribution Dist. We sometimes write f_X instead of f to stress that the pdf refers to the random variable X. Note that in Table 1.1, Γ is the gamma function:

$$\Gamma(\alpha) = \int_0^\infty \mathrm{e}^{-x} x^{\alpha-1}\, dx\ , \quad \alpha > 0\ .$$

Table 1.1 Commonly used continuous distributions.

Name	Notation	$f(x)$	$x \in$	Params.
Uniform	$\mathsf{U}[\alpha,\beta]$	$\dfrac{1}{\beta-\alpha}$	$[\alpha,\beta]$	$\alpha < \beta$
Normal	$\mathsf{N}(\mu,\sigma^2)$	$\dfrac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{R}$	$\sigma > 0,\ \mu \in \mathbb{R}$
Gamma	$\mathsf{Gamma}(\alpha,\lambda)$	$\dfrac{\lambda^\alpha x^{\alpha-1}\mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}$	$\mathbb{R}_+$	$\alpha,\lambda > 0$
Exponential	$\mathsf{Exp}(\lambda)$	$\lambda\,\mathrm{e}^{-\lambda x}$	$\mathbb{R}_+$	$\lambda > 0$
Beta	$\mathsf{Beta}(\alpha,\beta)$	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}$	$[0,1]$	$\alpha,\beta > 0$
Weibull	$\mathsf{Weib}(\alpha,\lambda)$	$\alpha\lambda\,(\lambda x)^{\alpha-1}\,\mathrm{e}^{-(\lambda x)^\alpha}$	$\mathbb{R}_+$	$\alpha,\lambda > 0$
Pareto	$\mathsf{Pareto}(\alpha,\lambda)$	$\alpha\lambda\,(1+\lambda x)^{-(\alpha+1)}$	$\mathbb{R}_+$	$\alpha,\lambda > 0$

Table 1.2 Commonly used discrete distributions.

Name	Notation	$f(x)$	$x \in$	Params.
Bernoulli	$\mathsf{Ber}(p)$	$p^x(1-p)^{1-x}$	$\{0,1\}$	$0 \leqslant p \leqslant 1$
Binomial	$\mathsf{Bin}(n,p)$	$\dbinom{n}{x}\,p^x(1-p)^{n-x}$	$\{0,1,\ldots,n\}$	$0 \leqslant p \leqslant 1,$ $n \in \mathbb{N}$
Discrete uniform	$\mathsf{DU}\{1,\ldots,n\}$	$\dfrac{1}{n}$	$\{1,\ldots,n\}$	$n \in \{1,2,\ldots\}$
Geometric	$\mathsf{G}(p)$	$p(1-p)^{x-1}$	$\{1,2,\ldots\}$	$0 \leqslant p \leqslant 1$
Poisson	$\mathsf{Poi}(\lambda)$	$\mathrm{e}^{-\lambda}\,\dfrac{\lambda^x}{x!}$	$\mathbb{N}$	$\lambda > 0$

1.5 EXPECTATION

It is often useful to consider various numerical characteristics of a random variable. One such quantity is the expectation, which measures the mean value of the distribution.

Definition 1.5.1 (Expectation) Let X be a random variable with pdf f. The *expectation* (or expected value or mean) of X, denoted by $\mathbb{E}[X]$ (or sometimes μ), is defined by

$$\mathbb{E}[X] = \begin{cases} \sum_x x\, f(x) & \text{discrete case,} \\ \int_{-\infty}^{\infty} x\, f(x)\, dx & \text{continuous case.} \end{cases}$$

If X is a random variable, then a function of X, such as X^2 or $\sin(X)$, is again a random variable. Moreover, the expected value of a function of X is simply a weighted average of the possible values that this function can take. That is, for any real function h

$$\mathbb{E}[h(X)] = \begin{cases} \sum_x h(x)\, f(x) & \text{discrete case,} \\ \int_{-\infty}^{\infty} h(x)\, f(x)\, dx & \text{continuous case.} \end{cases}$$

Another useful quantity is the variance, which measures the spread or dispersion of the distribution.

Definition 1.5.2 (Variance) The *variance* of a random variable X, denoted by Var(X) (or sometimes σ^2), is defined by

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

The square root of the variance is called the *standard deviation*. Table 1.3 lists the expectations and variances for some well-known distributions.

Table 1.3 Expectations and variances for some well-known distributions.

Dist.	$\mathbb{E}[X]$	Var(X)	Dist.	$\mathbb{E}[X]$	Var(X)
Bin(n,p)	np	$np(1-p)$	Gamma(α,λ)	$\dfrac{\alpha}{\lambda}$	$\dfrac{\alpha}{\lambda^2}$
G(p)	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$	N(μ,σ^2)	μ	σ^2
Poi(λ)	λ	λ	Beta(α,β)	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$
U(α,β)	$\dfrac{\alpha+\beta}{2}$	$\dfrac{(\beta-\alpha)^2}{12}$	Weib(α,λ)	$\frac{\Gamma(1/\alpha)}{\alpha\,\lambda}$	$\frac{2\Gamma(2/\alpha)}{\alpha} - \left(\frac{\Gamma(1/\alpha)}{\alpha\,\lambda}\right)^2$
Exp(λ)	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$			

The mean and the variance do not give, in general, enough information to completely specify the distribution of a random variable. However, they may provide useful bounds.

We discuss two such bounds. Suppose X can only take nonnegative values and has pdf f. For any $x > 0$, we can write

$$\begin{aligned} \mathbb{E}[X] &= \int_0^x t f(t)\,dt + \int_x^\infty t f(t)\,dt \geqslant \int_x^\infty t f(t)\,dt \\ &\geqslant \int_x^\infty x f(t)\,dt = x\,\mathbb{P}(X \geqslant x)\ , \end{aligned}$$

from which follows the *Markov inequality:* if $X \geqslant 0$, then for all $x > 0$,

$$\mathbb{P}(X \geqslant x) \leqslant \frac{\mathbb{E}[X]}{x}\ . \tag{1.9}$$

If we also know the variance of a random variable, we can give a tighter bound. Namely, for any random variable X with mean μ and variance σ^2, we have

$$\mathbb{P}(|X - \mu| \geqslant x) \leqslant \frac{\sigma^2}{x^2}\ . \tag{1.10}$$

This is called the *Chebyshev inequality*. The proof is as follows: Let $D^2 = (X - \mu)^2$; then, by the Markov inequality (1.9) and the definition of the variance,

$$\mathbb{P}(D^2 \geqslant x^2) \leqslant \frac{\sigma^2}{x^2}\ .$$

Also, note that the event $\{D^2 \geqslant x^2\}$ is equivalent to the event $\{|X - \mu| \geqslant x\}$, so that (1.10) follows.

1.6 JOINT DISTRIBUTIONS

Often a random experiment is described by more than one random variable. The theory for multiple random variables is similar to that for a single random variable.

Let $X_1, \dots, X_n$ be random variables describing some random experiment. We can accumulate these into a *random vector* $\mathbf{X} = (X_1, \dots, X_n)$. More generally, a collection $\{X_t, t \in \mathscr{T}\}$ of random variables is called a *stochastic process*. The set $\mathscr{T}$ is called the *parameter set* or *index set* of the process. It may be discrete (such as $\mathbb{N}$ or $\{1, \dots, 10\}$) or continuous (for example, $\mathbb{R}_+ = [0, \infty)$ or $[1, 10]$). The set of possible values for the stochastic process is called the *state space*.

The joint distribution of $X_1, \dots, X_n$ is specified by the *joint cdf*

$$F(x_1, \dots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \dots, X_n \leqslant x_n)\ .$$

The *joint pdf* f is given, in the discrete case, by $f(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$, and in the continuous case f is such that

$$\mathbb{P}(\mathbf{X} \in \mathscr{B}) = \int_{\mathscr{B}} f(x_1, \dots, x_n)\ dx_1 \dots dx_n$$

for any (measurable) region $\mathscr{B}$ in $\mathbb{R}^n$. The marginal pdfs can be recovered from the joint pdf by integration or summation. For example, in the case of a continuous random vector (X, Y) with joint pdf f, the pdf f_X of X is found as

$$f_X(x) = \int f(x, y)\,dy\ .$$

Suppose X and Y are both discrete or both continuous, with joint pdf f, and suppose $f_X(x) > 0$. Then the *conditional pdf* of Y given $X = x$ is given by

$$f_{Y|X}(y \,|\, x) = \frac{f(x,y)}{f_X(x)} \quad \text{for all } y \; .$$

The corresponding *conditional expectation* is (in the continuous case)

$$\mathbb{E}[Y \mid X = x] = \int y \, f_{Y|X}(y \,|\, x) \, dy \; .$$

Note that $\mathbb{E}[Y \mid X = x]$ is a function of x, say $h(x)$. The corresponding random variable $h(X)$ is written as $\mathbb{E}[Y \mid X]$. It can be shown (see, for example, [4]) that its expectation is simply the expectation of Y, that is,

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y] \; . \tag{1.11}$$

When the conditional distribution of Y given X is identical to that of Y, X and Y are said to be independent. More precisely:

Definition 1.6.1 (Independent Random Variables) The random variables $X_1, \ldots, X_n$ are called *independent* if for all events $\{X_i \in A_i\}$ with $A_i \subset \mathbb{R}$, $i = 1, \ldots, n$

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n) \; .$$

A direct consequence of the above definition for independence is that random variables $X_1, \ldots, X_n$ with joint pdf f (discrete or continuous) are independent if and only if

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \tag{1.12}$$

for all $x_1, \ldots, x_n$, where $\{f_{X_i}\}$ are the marginal pdfs.

■ EXAMPLE 1.4 Bernoulli Sequence

Consider the experiment where we flip a biased coin n times, with probability p of heads. We can model this experiment in the following way. For $i = 1, \ldots, n$ let X_i be the result of the i-th toss: $\{X_i = 1\}$ means heads (or success), $\{X_i = 0\}$ means tails (or failure). Also, let

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0), \quad i = 1, 2, \ldots, n \; .$$

Finally, assume that $X_1, \ldots, X_n$ are independent. The sequence $\{X_i, i = 1, 2, \ldots\}$ is called a *Bernoulli sequence* or *Bernoulli process* with success probability p. Let $X = X_1 + \cdots + X_n$ be the total number of successes in n trials (tosses of the coin). Denote by $\mathscr{B}$ the set of all binary vectors $\mathbf{x} = (x_1, \ldots, x_n)$ such that $\sum_{i=1}^n x_i = k$. Note that $\mathscr{B}$ has $\binom{n}{k}$ elements. We now have

$$\begin{aligned}
\mathbb{P}(X = k) &= \sum_{\mathbf{x} \in \mathscr{B}} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \\
&= \sum_{\mathbf{x} \in \mathscr{B}} \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) = \sum_{\mathbf{x} \in \mathscr{B}} p^k (1-p)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k} \; .
\end{aligned}$$

In other words, $X \sim \mathsf{Bin}(n, p)$. Compare this with Example 1.2.

Remark 1.6.1 An *infinite* sequence $X_1, X_2, \ldots$ of random variables is called independent if for any finite choice of parameters $i_1, i_2, \ldots, i_n$ (none of them the same) the random variables $X_{i_1}, \ldots, X_{i_n}$ are independent. Many probabilistic models involve random variables $X_1, X_2, \ldots$ that are *independent and identically distributed*, abbreviated as *iid*. We will use this abbreviation throughout this book.

Similar to the one-dimensional case, the expected value of any real-valued function h of $X_1, \ldots, X_n$ is a weighted average of all values that this function can take. Specifically, in the continuous case,

$$\mathbb{E}[h(X_1, \ldots, X_n)] = \int \cdots \int h(x_1, \ldots, x_n)\, f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \; .$$

As a direct consequence of the definitions of expectation and independence, we have

$$\mathbb{E}[a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n] = a + b_1 \mu_1 + \cdots + b_n \mu_n \tag{1.13}$$

for any sequence of random variables X_1, X_2, ..., X_n with expectations $\mu_1, \mu_2, \ldots, \mu_n$, where a, b_1, b_2, ..., b_n are constants. Similarly, for *independent* random variables one has

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mu_1\, \mu_2 \cdots \mu_n \; .$$

The *covariance* of two random variables X and Y with expectations $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$, respectively, is defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \; .$$

This is a measure for the amount of linear dependency between the variables. A scaled version of the covariance is given by the *correlation coefficient*,

$$\varrho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\, \sigma_Y},$$

where $\sigma_X^2 = \text{Var}(X)$ and $\sigma_Y^2 = \text{Var}(Y)$. It can be shown that the correlation coefficient always lies between -1 and 1; see Problem 1.13.

For easy reference, Table 1.4 lists some important properties of the variance and covariance. The proofs follow directly from the definitions of covariance and variance and the properties of the expectation.

Table 1.4 Properties of variance and covariance.

1	$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
2	$\text{Var}(aX + b) = a^2 \text{Var}(X)$
3	$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\, \mathbb{E}[Y]$
4	$\text{Cov}(X, Y) = \text{Cov}(Y, X)$
5	$\text{Cov}(aX + bY, Z) = a\, \text{Cov}(X, Z) + b\, \text{Cov}(Y, Z)$
6	$\text{Cov}(X, X) = \text{Var}(X)$
7	$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\, \text{Cov}(X, Y)$
8	X and Y indep. $\Longrightarrow \text{Cov}(X, Y) = 0$

As a consequence of properties 2 and 7, for any sequence of *independent* random variables $X_1, \ldots, X_n$ with variances $\sigma_1^2, \ldots, \sigma_n^2$

$$\mathrm{Var}(a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n) = b_1^2\,\sigma_1^2 + \cdots + b_n^2\,\sigma_n^2 \tag{1.14}$$

for any choice of constants a and $b_1, \ldots, b_n$.

For random vectors, such as $\mathbf{X} = (X_1, \ldots, X_n)^T$, it is convenient to write the expectations and covariances in vector notation.

Definition 1.6.2 (Expectation Vector and Covariance Matrix) For any random vector $\mathbf{X}$ we define the *expectation vector* as the vector of expectations

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])^T \ .$$

The *covariance matrix* Σ is defined as the matrix whose (i,j)-th element is

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \ .$$

If we define the expectation of a vector (matrix) to be the vector (matrix) of expectations, then we can write

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$$

and

$$\Sigma = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \ .$$

Note that $\boldsymbol{\mu}$ and Σ take on the same role as μ and σ^2 in the one-dimensional case.

Remark 1.6.2 Note that any covariance matrix Σ is *symmetric*. In fact (see Problem 1.16), it is *positive semidefinite*, that is, for any (column) vector $\mathbf{u}$,

$$\mathbf{u}^T\,\Sigma\,\mathbf{u} \geqslant 0 \ .$$

1.7 FUNCTIONS OF RANDOM VARIABLES

Suppose $X_1, \ldots, X_n$ are measurements of a random experiment. Often we are only interested in certain *functions* of the measurements rather than the individual measurements themselves. We give a number of examples.

■ **EXAMPLE 1.5**

Let X be a continuous random variable with pdf f_X and let $Z = aX + b$, where $a \neq 0$. We wish to determine the pdf f_Z of Z. Suppose that $a > 0$. We have for any z

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}\big(X \leqslant (z-b)/a\big) = F_X\big((z-b)/a\big) \ .$$

Differentiating this with respect to z gives $f_Z(z) = f_X\big((z-b)/a\big)/a$. For $a < 0$ we similarly obtain $f_Z(z) = f_X\big((z-b)/a\big)/(-a)$. Thus, in general,

$$f_Z(z) = \frac{1}{|a|}\, f_X\left(\frac{z-b}{a}\right) \ . \tag{1.15}$$

■ **EXAMPLE 1.6**

Generalizing the previous example, suppose that $Z = g(X)$ for some monotonically increasing function g. To find the pdf of Z from that of X we first write

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}\left(X \leqslant g^{-1}(z)\right) = F_X\left(g^{-1}(z)\right) \ ,$$

where g^{-1} is the inverse of g. Differentiating with respect to z now gives

$$f_Z(z) = f_X(g^{-1}(z))\,\frac{d}{dz}g^{-1}(z) = \frac{f_X(g^{-1}(z))}{g'(g^{-1}(z))} \ . \tag{1.16}$$

For monotonically decreasing functions, $\frac{d}{dz}g^{-1}(z)$ in the first equation needs to be replaced with its negative value.

■ **EXAMPLE 1.7 Order Statistics**

Let $X_1, \ldots, X_n$ be an iid sequence of random variables with common pdf f and cdf F. In many applications one is interested in the distribution of the *order statistics* $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, where $X_{(1)}$ is the smallest of the $\{X_i, i = 1, \ldots, n\}$, $X_{(2)}$ is the second smallest, and so on. The cdf of $X_{(n)}$ follows from

$$\mathbb{P}(X_{(n)} \leqslant x) = \mathbb{P}(X_1 \leqslant x, \ldots, X_n \leqslant x) = \prod_{i=1}^{n} \mathbb{P}(X_i \leqslant x) = (F(x))^n \ .$$

Similarly,

$$\mathbb{P}(X_{(1)} > x) = \mathbb{P}(X_1 > x, \ldots, X_n > x) = \prod_{i=1}^{n} \mathbb{P}(X_i > x) = (1 - F(x))^n \ .$$

Moreover, because all orderings of $X_1, \ldots, X_n$ are equally likely, it follows that the joint pdf of the ordered sample is, on the wedge $\{(x_1, \ldots, x_n) : x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n\}$, simply $n!$ times the joint density of the unordered sample and zero elsewhere.

1.7.1 Linear Transformations

Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ be a column vector in $\mathbb{R}^n$ and A an $m \times n$ matrix. The mapping $\mathbf{x} \mapsto \mathbf{z}$, with $\mathbf{z} = A\mathbf{x}$, is called a *linear transformation*. Now consider a *random* vector $\mathbf{X} = (X_1, \ldots, X_n)^T$, and let

$$\mathbf{Z} = A\mathbf{X} \ .$$

Then $\mathbf{Z}$ is a random vector in $\mathbb{R}^m$. In principle, if we know the joint distribution of $\mathbf{X}$, then we can derive the joint distribution of $\mathbf{Z}$. Let us first see how the expectation vector and covariance matrix are transformed.

Theorem 1.7.1 *If $\mathbf{X}$ has an expectation vector $\boldsymbol{\mu}_\mathbf{X}$ and covariance matrix $\Sigma_\mathbf{X}$, then the expectation vector and covariance matrix of $\mathbf{Z} = A\mathbf{X}$ are given by*

$$\boldsymbol{\mu}_\mathbf{Z} = A\boldsymbol{\mu}_\mathbf{X} \tag{1.17}$$

and

$$\Sigma_\mathbf{Z} = A\,\Sigma_\mathbf{X}\,A^T \ . \tag{1.18}$$

Proof: We have $\boldsymbol{\mu}_{\mathbf{Z}} = \mathbb{E}[\mathbf{Z}] = \mathbb{E}[A\mathbf{X}] = A\,\mathbb{E}[\mathbf{X}] = A\boldsymbol{\mu}_{\mathbf{X}}$ and

$$\begin{aligned}\Sigma_{\mathbf{Z}} &= \mathbb{E}[(\mathbf{Z}-\boldsymbol{\mu}_{\mathbf{Z}})(\mathbf{Z}-\boldsymbol{\mu}_{\mathbf{Z}})^T] = \mathbb{E}[A(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(A(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}))^T] \\ &= A\,\mathbb{E}[(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})^T]A^T \\ &= A\,\Sigma_{\mathbf{X}}\,A^T\ .\end{aligned}$$

□

Suppose that A is an invertible $n \times n$ matrix. If $\mathbf{X}$ has a joint density $f_{\mathbf{X}}$, what is the joint density $f_{\mathbf{Z}}$ of $\mathbf{Z}$? Consider Figure 1.1. For any fixed $\mathbf{x}$, let $\mathbf{z} = A\mathbf{x}$. Hence, $\mathbf{x} = A^{-1}\mathbf{z}$. Consider the n-dimensional cube $C = [z_1, z_1+h] \times \cdots \times [z_n, z_n+h]$. Let D be the image of C under A^{-1}, that is, the parallelepiped of all points $\mathbf{x}$ such that $A\mathbf{x} \in C$. Then,

$$\mathbb{P}(\mathbf{Z} \in C) \approx h^n\, f_{\mathbf{Z}}(\mathbf{z})\ .$$

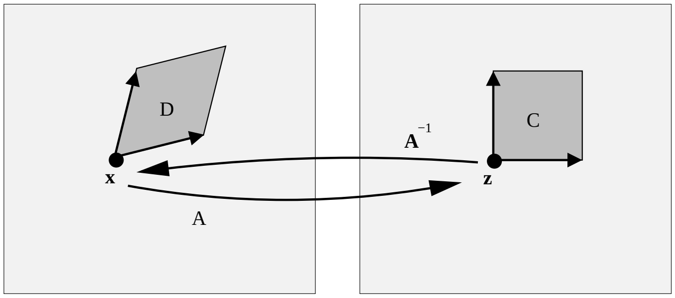


Figure 1.1 Linear transformation.

Now recall from linear algebra (see, for example, [6]) that any matrix B linearly transforms an n-dimensional rectangle with volume V into an n-dimensional parallelepiped with volume $V\,|B|$, where $|B| = |\det(B)|$. Thus,

$$\mathbb{P}(\mathbf{Z} \in C) = \mathbb{P}(\mathbf{X} \in D) \approx h^n |A^{-1}|\, f_{\mathbf{X}}(\mathbf{x}) = h^n |A|^{-1}\, f_{\mathbf{X}}(\mathbf{x})\ .$$

Letting h go to 0, we obtain

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{f_{\mathbf{X}}(A^{-1}\mathbf{z})}{|A|},\ \ \mathbf{z} \in \mathbb{R}^n. \tag{1.19}$$

1.7.2 General Transformations

We can apply reasoning similar to that above to deal with general transformations $\mathbf{x} \mapsto \boldsymbol{g}(\mathbf{x})$, written out:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{pmatrix}.$$

For a fixed $\mathbf{x}$, let $\mathbf{z} = \boldsymbol{g}(\mathbf{x})$. Suppose $\boldsymbol{g}$ is invertible; hence, $\mathbf{x} = \boldsymbol{g}^{-1}(\mathbf{z})$. Any infinitesimal n-dimensional rectangle at $\mathbf{x}$ with volume V is transformed into an n-dimensional parallelepiped at $\mathbf{z}$ with volume $V\,|J_{\mathbf{x}}(\boldsymbol{g})|$, where $J_{\mathbf{x}}(\boldsymbol{g})$ is the *matrix of Jacobi* at $\mathbf{x}$ of the transformation $\boldsymbol{g}$, that is,

$$J_{\mathbf{x}}(\boldsymbol{g}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix} .$$

Now consider a random column vector $\mathbf{Z} = \boldsymbol{g}(\mathbf{X})$. Let C be a small cube around $\mathbf{z}$ with volume h^n. Let D be the image of C under $\boldsymbol{g}^{-1}$. Then, as in the linear case,

$$\mathbb{P}(\mathbf{Z} \in C) \approx h^n\, f_{\mathbf{Z}}(\mathbf{z}) \approx h^n |J_{\mathbf{z}}(\boldsymbol{g}^{-1})|\, f_{\mathbf{X}}(\mathbf{x}) \ .$$

Hence, we have the transformation rule

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\boldsymbol{g}^{-1}(\mathbf{z}))\, |J_{\mathbf{z}}(\boldsymbol{g}^{-1})|, \ \ \mathbf{z} \in \mathbb{R}^n. \tag{1.20}$$

(Note: $|J_{\mathbf{z}}(\boldsymbol{g}^{-1})| = 1/|J_{\mathbf{x}}(\boldsymbol{g})|$.)

Remark 1.7.1 In most coordinate transformations it is $\boldsymbol{g}^{-1}$ that is given — that is, an expression for $\mathbf{x}$ as a function of $\mathbf{z}$ rather than $\boldsymbol{g}$.

1.8 TRANSFORMS

Many calculations and manipulations involving probability distributions are facilitated by the use of transforms. Two typical examples are the *probability generating function* of a positive integer-valued random variable N, defined by

$$G(z) = \mathbb{E}[z^N] = \sum_{k=0}^{\infty} z^k\, \mathbb{P}(N = k) \ , \quad |z| \leqslant 1 \ ,$$

and the *Laplace transform* of a positive random variable X defined, for $s \geqslant 0$, by

$$L(s) = \mathbb{E}[\mathrm{e}^{-sX}] = \begin{cases} \sum_x \mathrm{e}^{-sx}\, f(x) & \text{discrete case,} \\ \int_0^\infty \mathrm{e}^{-sx}\, f(x)\, dx & \text{continuous case.} \end{cases}$$

All transforms share an important *uniqueness property*: two distributions are the same if and only if their respective transforms are the same.

■ **EXAMPLE 1.8**

Let $M \sim \mathsf{Poi}(\mu)$; then its probability generating function is given by

$$G(z) = \sum_{k=0}^{\infty} z^k\, \mathrm{e}^{-\mu}\, \frac{\mu^k}{k!} = \mathrm{e}^{-\mu} \sum_{k=0}^{\infty} \frac{(z\mu)^k}{k!} = \mathrm{e}^{-\mu}\mathrm{e}^{z\mu} = \mathrm{e}^{-\mu(1-z)} \ . \tag{1.21}$$

Now let $N \sim \mathsf{Poi}(\nu)$ independently of M. Then the probability generating function of $M + N$ is given by

$$\mathbb{E}[z^{M+N}] = \mathbb{E}[z^M]\, \mathbb{E}[z^N] = \mathrm{e}^{-\mu(1-z)}\mathrm{e}^{-\nu(1-z)} = \mathrm{e}^{-(\mu+\nu)(1-z)} \ .$$

Thus, by the uniqueness property, $M + N \sim \mathsf{Poi}(\mu + \nu)$.

EXAMPLE 1.9

The Laplace transform of $X \sim \mathsf{Gamma}(\alpha, \lambda)$ is given by

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{-sX}] &= \int_0^\infty \frac{\mathrm{e}^{-\lambda x}\,\lambda^\alpha\,x^{\alpha-1}}{\Gamma(\alpha)}\,\mathrm{e}^{-sx}\,dx \\ &= \left(\frac{\lambda}{\lambda+s}\right)^\alpha \int_0^\infty \frac{\mathrm{e}^{-(\lambda+s)x}\,(\lambda+s)^\alpha\,x^{\alpha-1}}{\Gamma(\alpha)}\,dx \\ &= \left(\frac{\lambda}{\lambda+s}\right)^\alpha .\end{aligned}$$

As a special case, the Laplace transform of the $\mathsf{Exp}(\lambda)$ distribution is given by $\lambda/(\lambda+s)$. Now let $X_1, \ldots, X_n$ be iid $\mathsf{Exp}(\lambda)$ random variables. The Laplace transform of $S_n = X_1 + \cdots + X_n$ is

$$\mathbb{E}[\mathrm{e}^{-sS_n}] = \mathbb{E}[\mathrm{e}^{-sX_1}\cdots\mathrm{e}^{-sX_n}] = \mathbb{E}[\mathrm{e}^{-sX_1}]\,\cdots\,\mathbb{E}[\mathrm{e}^{-sX_n}] = \left(\frac{\lambda}{\lambda+s}\right)^n ,$$

which shows that $S_n \sim \mathsf{Gamma}(n, \lambda)$.

1.9 JOINTLY NORMAL RANDOM VARIABLES

It is helpful to view normally distributed random variables as simple transformations of *standard normal* — that is, $\mathsf{N}(0,1)$-distributed — random variables. In particular, let $X \sim \mathsf{N}(0,1)$. Then, X has density f_X given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{x^2}{2}}.$$

Now consider the transformation $Z = \mu + \sigma X$. Then, by (1.15), Z has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}}.$$

In other words, $Z \sim \mathsf{N}(\mu, \sigma^2)$. We can also state this as follows: if $Z \sim \mathsf{N}(\mu, \sigma^2)$, then $(Z-\mu)/\sigma \sim \mathsf{N}(0,1)$. This procedure is called *standardization*.

We now generalize this to n dimensions. Let $X_1, \ldots, X_n$ be independent and standard normal random variables. The joint pdf of $\mathbf{X} = (X_1, \ldots, X_n)^T$ is given by

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}\mathrm{e}^{-\frac{1}{2}\,\mathbf{x}^T\mathbf{x}}, \quad \mathbf{x} \in \mathbb{R}^n. \tag{1.22}$$

Consider the *affine* transformation (that is, a linear transformation plus a constant vector)

$$\mathbf{Z} = \boldsymbol{\mu} + B\,\mathbf{X} \tag{1.23}$$

for some $m \times n$ matrix B. Note that, by Theorem 1.7.1, $\mathbf{Z}$ has expectation vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma = BB^T$. Any random vector of the form (1.23) is said to have a *jointly normal* or *multivariate normal* distribution. We write $\mathbf{Z} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$. Suppose B is an invertible $n \times n$ matrix. Then, by (1.19), the density of $\mathbf{Y} = \mathbf{Z} - \boldsymbol{\mu}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|B|\sqrt{(2\pi)^n}}\,\mathrm{e}^{-\frac{1}{2}\,(B^{-1}\mathbf{y})^T B^{-1}\mathbf{y}} = \frac{1}{|B|\sqrt{(2\pi)^n}}\,\mathrm{e}^{-\frac{1}{2}\,\mathbf{y}^T(B^{-1})^T B^{-1}\mathbf{y}} .$$

We have $|B| = \sqrt{|\Sigma|}$ and $(B^{-1})^T B^{-1} = (B^T)^{-1} B^{-1} = (BB^T)^{-1} = \Sigma^{-1}$, so that

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\, \mathrm{e}^{-\frac{1}{2}\,\mathbf{y}^T\Sigma^{-1}\mathbf{y}}\ .$$

Because $\mathbf{Z}$ is obtained from $\mathbf{Y}$ by simply adding a constant vector $\boldsymbol{\mu}$, we have $f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{Y}}(\mathbf{z} - \boldsymbol{\mu})$ and therefore

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\, \mathrm{e}^{-\frac{1}{2}\,(\mathbf{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})}, \quad \mathbf{z} \in \mathbb{R}^n\ . \tag{1.24}$$

Note that this formula is very similar to the one-dimensional case.

Conversely, given a covariance matrix $\Sigma = (\sigma_{ij})$, there exists a unique lower triangular matrix

$$B = \begin{pmatrix} b_{11} & 0 & \cdots & 0 \\ b_{21} & b_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix} \tag{1.25}$$

such that $\Sigma = BB^T$. This matrix can be obtained efficiently via the *Cholesky square root method*, see Section A.1 of the Appendix.

1.10 LIMIT THEOREMS

We briefly discuss two of the main results in probability: the law of large numbers and the central limit theorem. Both are associated with sums of independent random variables.

Let $X_1, X_2, \ldots$ be iid random variables with expectation μ and variance σ^2. For each n let $S_n = X_1 + \cdots + X_n$. Since $X_1, X_2, \ldots$ are iid, we have $\mathbb{E}[S_n] = n\,\mathbb{E}[X_1] = n\mu$ and $\mathrm{Var}(S_n) = n\,\mathrm{Var}(X_1) = n\sigma^2$.

The law of large numbers states that S_n/n is close to μ for large n. Here is the more precise statement.

Theorem 1.10.1 (Strong Law of Large Numbers) *If $X_1, \ldots, X_n$ are iid with expectation μ, then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1\ .$$

The central limit theorem describes the limiting distribution of S_n (or S_n/n), and it applies to both continuous and discrete random variables. Loosely, it states that the random sum S_n has a distribution that is approximately normal, when n is large. The more precise statement is given next.

Theorem 1.10.2 (Central Limit Theorem) *If $X_1, \ldots, X_n$ are iid with expectation μ and variance $\sigma^2 < \infty$, then for all $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) = \Phi(x)\ ,$$

where Φ is the cdf of the standard normal distribution.

In other words, S_n has a distribution that is approximately normal, with expectation $n\mu$ and variance $n\sigma^2$. To see the central limit theorem in action, consider Figure 1.2. The left part shows the pdfs of $S_1, \ldots, S_4$ for the case where the $\{X_i\}$ have a $\mathsf{U}[0,1]$ distribution. The right part shows the same for the $\mathsf{Exp}(1)$ distribution. We clearly see convergence to a bell-shaped curve, characteristic of the normal distribution.

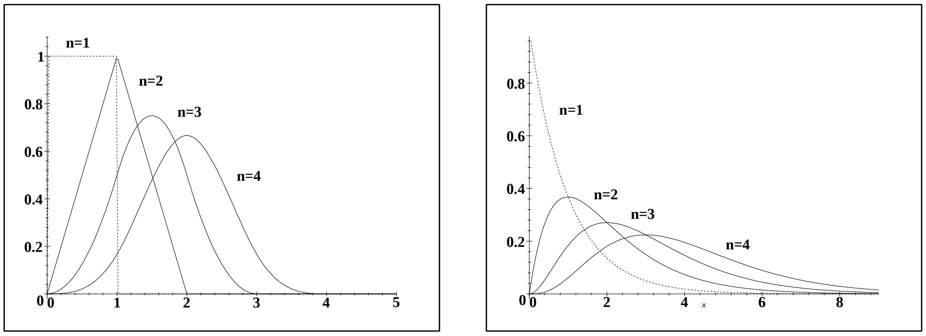


Figure 1.2 Illustration of the central limit theorem for (left) the uniform distribution and (right) the exponential distribution.

A direct consequence of the central limit theorem and the fact that a $\mathsf{Bin}(n,p)$ random variable X can be viewed as the sum of n iid $\mathsf{Ber}(p)$ random variables, $X = X_1 + \cdots + X_n$, is that for large n

$$\mathbb{P}(X \leqslant k) \approx \mathbb{P}(Y \leqslant k) \,, \tag{1.26}$$

with $Y \sim \mathsf{N}(np, np(1-p))$. As a rule of thumb, this *normal approximation to the binomial distribution* is accurate if both np and $n(1-p)$ are larger than 5.

There is also a central limit theorem for random vectors. The multidimensional version is as follows: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be iid random vectors with expectation vector $\boldsymbol{\mu}$ and covariance matrix Σ. Then for large n the random vector $\mathbf{X}_1 + \cdots + \mathbf{X}_n$ has approximately a multivariate normal distribution with expectation vector $n\boldsymbol{\mu}$ and covariance matrix $n\Sigma$.

1.11 POISSON PROCESSES

The Poisson process is used to model certain kinds of arrivals or patterns. Imagine, for example, a telescope that can detect individual photons from a faraway galaxy. The photons arrive at random times $T_1, T_2, \ldots$. Let N_t denote the number of arrivals in the time interval $[0,t]$, that is, $N_t = \sup\{k : T_k \leqslant t\}$. Note that the number of arrivals in an interval $I = (a,b]$ is given by $N_b - N_a$. We will also denote it by $N(a,b]$. A sample path of the arrival counting process $\{N_t, t \geqslant 0\}$ is given in Figure 1.3.

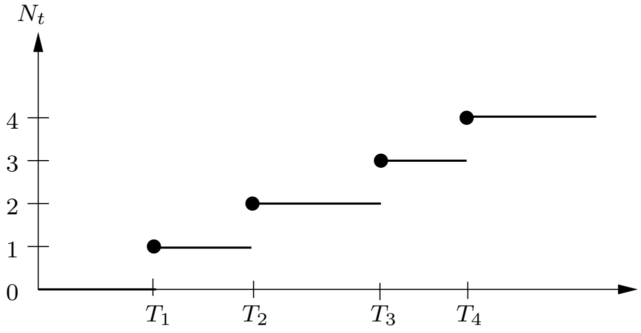


Figure 1.3 A sample path of the arrival counting process $\{N_t, t \geqslant 0\}$.

For this particular arrival process, one would assume that the number of arrivals in an interval (a, b) is independent of the number of arrivals in interval (c, d) when the two intervals do not intersect. Such considerations lead to the following definition:

Definition 1.11.1 (Poisson Process) An arrival counting process $N = \{N_t\}$ is called a *Poisson process* with *rate* $\lambda > 0$ if

(a) The numbers of points in nonoverlapping intervals are independent.

(b) The number of points in interval I has a Poisson distribution with mean $\lambda \times \text{length}(I)$.

Combining (a) and (b) we see that the number of arrivals in any small interval $(t, t+h]$ is independent of the arrival process up to time t and has a $\mathsf{Poi}(\lambda h)$ distribution. In particular, the conditional probability that exactly one arrival occurs during the time interval $(t, t+h]$ is $\mathbb{P}(N(t, t+h] = 1 \,|\, N_t) = \mathrm{e}^{-\lambda h}\, \lambda\, h \approx \lambda\, h$. Similarly, the probability of no arrivals is approximately $1 - \lambda h$ for small h. In other words, λ is the *rate* at which arrivals occur. Notice also that since $N_t \sim \mathsf{Poi}(\lambda t)$, the expected number of arrivals in $[0, t]$ is λt, that is, $\mathbb{E}[N_t] = \lambda t$. In Definition 1.11.1 N is seen as a random counting measure, where $N(I)$ counts the random number of arrivals in set I.

An important relationship between N_t and T_n is

$$\{N_t \geqslant n\} = \{T_n \leqslant t\}\ . \tag{1.27}$$

In other words, the number of arrivals in $[0, t]$ is at least n if and only if the n-th arrival occurs at or before time t. As a consequence, we have

$$\begin{aligned} \mathbb{P}(T_n \leqslant t) &= \mathbb{P}(N_t \geqslant n) = 1 - \sum_{k=0}^{n-1} \mathbb{P}(N_t = k) \\ &= 1 - \sum_{k=0}^{n-1} \frac{\mathrm{e}^{-\lambda\, t} (\lambda\, t)^k}{k!}\ , \end{aligned}$$

which corresponds exactly to the cdf of the $\mathsf{Gamma}(n, \lambda)$ distribution; see Problem 1.17. Thus,

$$T_n \sim \mathsf{Gamma}(n, \lambda). \tag{1.28}$$

Hence, each T_n has the same distribution as the sum of n independent $\mathsf{Exp}(\lambda)$-distributed random variables. This corresponds with the second important characterization of a Poisson process:

> *An arrival counting process $\{N_t\}$ is a Poisson process with rate λ if and only if the interarrival times $A_1 = T_1, A_2 = T_2 - T_1, \ldots$ are independent and $\mathsf{Exp}(\lambda)$-distributed random variables.*

Poisson and Bernoulli processes are akin, and much can be learned about Poisson processes via the following *Bernoulli approximation*. Let $N = \{N_t\}$ be a Poisson process with parameter λ. We divide the time axis into small time intervals $[0, h), [h, 2h), \ldots$ and count how many arrivals occur in each interval. Note that the number of arrivals in any small time interval of length h is, with high probability, either 1 (with probability $\lambda\, h\, \mathrm{e}^{-\lambda h} \approx \lambda h$) or 0 (with probability $\mathrm{e}^{-\lambda\, h} \approx 1 - \lambda h$). Next, define $X = \{X_n\}$ to be a Bernoulli process with success parameter $p = \lambda\, h$. Put $Y_0 = 0$ and let $Y_n = X_1 + \cdots + X_n$ be the total number of successes in n trials. $Y = \{Y_n\}$ is called the *Bernoulli approximation* to N. We can view N as a limiting case of Y as we decrease h.

As an example of the usefulness of this interpretation, we now demonstrate that the Poisson property (b) in Definition 1.11.1 follows basically from the *independence* assumption (a). For small h, N_t should have approximately the same distribution as Y_n, where n is the integer part of t/h (we write $n = \lfloor t/h \rfloor$). Hence,

$$\begin{aligned}
\mathbb{P}(N_t = k) &\approx \mathbb{P}(Y_n = k) \\
&= \binom{n}{k} (\lambda\, h)^k (1 - (\lambda\, h))^{n-k} \\
&\approx \binom{n}{k} (\lambda t/n)^k (1 - (\lambda t/n))^{n-k} \\
&\approx \mathrm{e}^{\lambda\, t}\, \frac{(\lambda\, t)^k}{k!} \; . \qquad (1.29)
\end{aligned}$$

Equation (1.29) follows from the Poisson approximation to the binomial distribution; see Problem 1.22.

Another application of the Bernoulli approximation is the following. For the Bernoulli process, given that the total number of successes is k, the positions of the k successes are uniformly distributed over points $1, \ldots, n$. The corresponding property for the Poisson process N is that given $N_t = n$, the arrival times $T_1, \ldots, T_n$ are distributed according to the order statistics $X_{(1)}, \ldots, X_{(n)}$, where $X_1, \ldots, X_n$ are iid $\mathsf{U}[0, t]$.

1.12 MARKOV PROCESSES

Markov processes are stochastic processes whose futures are conditionally independent of their pasts given their present values. More formally, a stochastic process $\{X_t, t \in \mathscr{T}\}$, with $\mathscr{T} \subseteq \mathbb{R}$, is called a *Markov process* if, for every $s > 0$ and t,

$$(X_{t+s} \,|\, X_u, u \leqslant t) \quad \sim \quad (X_{t+s} \,|\, X_t) \; . \qquad (1.30)$$

In other words, the conditional distribution of the future variable X_{t+s}, given the entire past of the process $\{X_u, u \leqslant t\}$, is the same as the conditional distribution of X_{t+s} given only the present X_t. That is, in order to predict future states, we only need to know the present one. Property (1.30) is called the *Markov property*.

Depending on the index set $\mathscr{T}$ and state space $\mathscr{E}$ (the set of all values the $\{X_t\}$ can take), Markov processes come in many different forms. A Markov process with a discrete index set is called a *Markov chain*. A Markov process with a discrete state space and a continuous index set (such as $\mathbb{R}$ or $\mathbb{R}_+$) is called a *Markov jump process*.

1.12.1 Markov Chains

Consider a Markov chain $X = \{X_t, t \in \mathbb{N}\}$ with a discrete (that is, countable) state space $\mathscr{E}$. In this case the Markov property (1.30) is:

$$\mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_0 = x_0, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_t = x_t) \tag{1.31}$$

for all $x_0, \ldots, x_{t+1}, \in \mathscr{E}$ and $t \in \mathbb{N}$. We restrict ourselves to Markov chains for which the conditional probability

$$\mathbb{P}(X_{t+1} = j \,|\, X_t = i),\ i, j \in \mathscr{E} \tag{1.32}$$

is independent of the time t. Such chains are called *time-homogeneous*. The probabilities in (1.32) are called the *(one-step) transition probabilities* of X. The distribution of X_0 is called the *initial distribution* of the Markov chain. The one-step transition probabilities and the initial distribution completely specify the distribution of X. Namely, we have by the product rule (1.4) and the Markov property (1.30)

$$\begin{aligned}
&\mathbb{P}(X_0 = x_0, \ldots, X_t = x_t) \\
&\quad = \mathbb{P}(X_0 = x_0)\,\mathbb{P}(X_1 = x_1 \,|\, X_0 = x_0) \cdots \mathbb{P}(X_t = x_t \,|\, X_0 = x_0, \ldots X_{t-1} = x_{t-1}) \\
&\quad = \mathbb{P}(X_0 = x_0)\,\mathbb{P}(X_1 = x_1 \,|\, X_0 = x_0) \cdots \mathbb{P}(X_t = x_t \,|\, X_{t-1} = x_{t-1})\ .
\end{aligned}$$

Since $\mathscr{E}$ is countable, we can arrange the one-step transition probabilities in an array. This array is called the (one-step) *transition matrix* of X. We usually denote it by P. For example, when $\mathscr{E} = \{0, 1, 2, \ldots\}$ the transition matrix P has the form

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}\ .$$

Note that the elements in every row are positive and sum up to unity.

Another convenient way to describe a Markov chain X is through its *transition graph*. States are indicated by the nodes of the graph, and a strictly positive (> 0) transition probability p_{ij} from state i to j is indicated by an arrow from i to j with weight p_{ij}.

■ **EXAMPLE 1.10 Random Walk on the Integers**

Let p be a number between 0 and 1. The Markov chain X with state space $\mathbb{Z}$ and transition matrix P defined by

$$P(i, i+1) = p, \quad P(i, i-1) = q = 1 - p, \quad \text{for all } i \in \mathbb{Z}$$

is called a *random walk on the integers*. Let X start at 0; thus, $\mathbb{P}(X_0 = 0) = 1$. The corresponding transition graph is given in Figure 1.4. Starting at 0, the chain takes subsequent steps to the right with probability p and to the left with probability q.

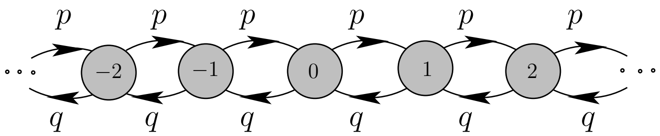


Figure 1.4 Transition graph for a random walk on $\mathbb{Z}$.

We shall show next how to calculate the probability that, starting from state i at some (discrete) time t, we are in j at (discrete) time $t+s$, that is, the probability $\mathbb{P}(X_{t+s} = j \mid X_t = i)$. For clarity, let us assume that $\mathscr{E} = \{1, 2, \ldots, m\}$ for some fixed m, so that P is an $m \times m$ matrix. For $t = 0, 1, 2, \ldots$, define the row vector

$$\boldsymbol{\pi}^{(t)} = (\mathbb{P}(X_t = 1), \ldots, \mathbb{P}(X_t = m)).$$

We call $\boldsymbol{\pi}^{(t)}$ the *distribution vector*, or simply the *distribution*, of X at time t and $\boldsymbol{\pi}^{(0)}$ the *initial distribution* of X. The following result shows that the t-step probabilities can be found simply by matrix multiplication.

Theorem 1.12.1 *The distribution of X at time t is given by*

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^{(0)} P^t \tag{1.33}$$

for all $t = 0, 1, \ldots$. (Here P^0 denotes the identity matrix.)

Proof: The proof is by induction. Equality (1.33) holds for $t = 0$ by definition. Suppose it is true for some $t = 0, 1, \ldots$. We have

$$\mathbb{P}(X_{t+1} = k) = \sum_{i=1}^{m} \mathbb{P}(X_{t+1} = k \mid X_t = i)\, \mathbb{P}(X_t = i)\ .$$

But (1.33) is assumed to be true for t, so $\mathbb{P}(X_t = i)$ is the i-th element of $\boldsymbol{\pi}^{(0)} P^t$. Moreover, $\mathbb{P}(X_{t+1} = k \mid X_t = i)$ is the (i, k)-th element of P. Therefore, for every k

$$\sum_{i=1}^{m} \mathbb{P}(X_{t+1} = k \mid X_t = i)\, \mathbb{P}(X_t = i) = \sum_{i=1}^{m} P(i, k)(\boldsymbol{\pi}^{(0)} P^t)(i)\ ,$$

which is just the k-th element of $\boldsymbol{\pi}^{(0)} P^{t+1}$. This completes the induction step, and thus the theorem is proved. □

By taking $\boldsymbol{\pi}^{(0)}$ as the i-th unit vector, $\mathbf{e}_i$, the t-step transition probabilities can be found as $\mathbb{P}(X_t = j \mid X_0 = i) = (\mathbf{e}_i\, P^t)(j) = P^t(i, j)$, which is the (i, j)-th element of matrix P^t. Thus, to find the t-step transition probabilities, we just have to compute the t-th power of P.

1.12.2 Classification of States

Let X be a Markov chain with discrete state space $\mathscr{E}$ and transition matrix P. We can characterize the relations between states in the following way: If states i and j are such that $P^t(i, j) > 0$ for some $t \geqslant 0$, we say that i *leads to* j and write $i \to j$. We say that i and j

communicate if $i \to j$ and $j \to i$, and write $i \leftrightarrow j$. Using the relation "$\leftrightarrow$" we can divide $\mathscr{E}$ into *equivalence classes* such that all the states in an equivalence class communicate with each other but not with any state outside that class. If there is only one equivalent class ($= \mathscr{E}$), the Markov chain is said to be *irreducible*. If a set of states $\mathscr{A}$ is such that $\sum_{j \in \mathscr{A}} P(i,j) = 1$ for all $i \in \mathscr{A}$, then $\mathscr{A}$ is called a *closed* set. A state i is called an *absorbing* state if $\{i\}$ is closed. For example, in the transition graph depicted in Figure 1.5, the equivalence classes are $\{1,2\}$, $\{3\}$, and $\{4,5\}$. Class $\{1,2\}$ is the only closed set: the Markov chain cannot escape from it. If state 1 were missing, state 2 would be absorbing. In Example 1.10 the Markov chain is irreducible since all states communicate.

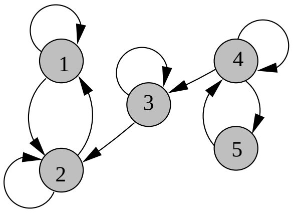


Figure 1.5 A transition graph with three equivalence classes.

Another classification of states is obtained by observing the system from a local point of view. In particular, let T denote the time the chain first visits state j, or first returns to j if it started there, and let N_j denote the total number of visits to j from time 0 on. We write $\mathbb{P}_j(A)$ for $\mathbb{P}(A \mid X_0 = j)$ for any event A. We denote the corresponding expectation operator by $\mathbb{E}_j$. State j is called a *recurrent* state if $\mathbb{P}_j(T < \infty) = 1$; otherwise, j is called *transient*. A recurrent state is called *positive recurrent* if $\mathbb{E}_j[T] < \infty$; otherwise, it is called *null recurrent*. Finally, a state is said to be *periodic, with period* δ, if $\delta \geqslant 2$ is the largest integer for which $\mathbb{P}_j(T = n\delta$ for some $n \geqslant 1) = 1$; otherwise, it is called *aperiodic*. For example, in Figure 1.5 states 1 and 2 are recurrent, and the other states are transient. All these states are aperiodic. The states of the random walk of Example 1.10 are periodic with period 2.

It can be shown that recurrence and transience are class properties. In particular, if $i \leftrightarrow j$, then i recurrent (transient) $\Leftrightarrow$ j recurrent (transient). Thus, in an irreducible Markov chain, one state being recurrent implies that all other states are also recurrent. And if one state is transient, then so are all the others.

1.12.3 Limiting Behavior

The limiting or "steady-state" behavior of Markov chains as $t \to \infty$ is of considerable interest and importance, and is often simpler to describe and analyze than the "transient" behavior of the chain for fixed t. It can be shown (see, for example, [4]) that in an irreducible, aperiodic Markov chain with transition matrix P the t-step probabilities converge to a constant that does not depend on the initial state. More specifically,

$$\lim_{t \to \infty} P^t(i,j) = \pi_j \tag{1.34}$$

for some number $0 \leqslant \pi_j \leqslant 1$. Moreover, $\pi_j > 0$ if j is positive recurrent and $\pi_j = 0$ otherwise. The intuitive reason behind this result is that the process "forgets" where it was initially if it goes on long enough. This is true for both finite and countably infinite Markov chains. The numbers $\{\pi_j, j \in \mathscr{E}\}$ form the *limiting distribution* of the Markov chain, provided that $\pi_j \geqslant 0$ and $\sum_j \pi_j = 1$. Note that these conditions are not always satisfied: they are clearly not satisfied if the Markov chain is transient, and they may not be satisfied if the Markov chain is recurrent (namely when the states are null-recurrent). The following theorem gives a method for obtaining limiting distributions. Here we assume for simplicity that $\mathscr{E} = \{0, 1, 2, \ldots\}$. The limiting distribution is identified with the row vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$.

Theorem 1.12.2 *For an irreducible, aperiodic Markov chain with transition matrix P, if the limiting distribution $\boldsymbol{\pi}$ exists, then it is uniquely determined by the solution of*

$$\boldsymbol{\pi} = \boldsymbol{\pi} P \ , \tag{1.35}$$

with $\pi_j \geqslant 0$ and $\sum_j \pi_j = 1$. Conversely, if there exists a positive row vector $\boldsymbol{\pi}$ satisfying (1.35) *and summing up to 1, then $\boldsymbol{\pi}$ is the limiting distribution of the Markov chain. Moreover, in that case $\pi_j > 0$ for all j, and all states are positive recurrent.*

Proof: (Sketch). For the case where $\mathscr{E}$ is finite, the result is simply a consequence of (1.33). Namely, with $\boldsymbol{\pi}^{(0)}$ being the i-th unit vector, we have

$$P^{t+1}(i,j) = \left(\boldsymbol{\pi}^{(0)}\, P^t\, P\right)(j) = \sum_{k \in \mathscr{E}} P^t(i,k) P(k,j) \ .$$

Letting $t \to \infty$, we obtain (1.35) from (1.34), provided that we can change the order of the limit and the summation. To show uniqueness, suppose that another vector $\mathbf{y}$, with $y_j \geqslant 0$ and $\sum_j y_j = 1$, satisfies $\mathbf{y} = \mathbf{y}P$. Then it is easy to show by induction that $\mathbf{y} = \mathbf{y}P^t$, for every t. Hence, letting $t \to \infty$, we obtain for every j

$$y_j = \sum_i y_i\, \pi_j = \pi_j \ ,$$

since the $\{y_j\}$ sum up to unity. We omit the proof of the converse statement. □

■ EXAMPLE 1.11 Random Walk on the Positive Integers

This is a slightly different random walk than the one in Example 1.10. Let X be a random walk on $\mathscr{E} = \{0, 1, 2, \ldots\}$ with transition matrix

$$P = \begin{pmatrix} q & p & 0 & \ldots & & \\ q & 0 & p & 0 & \ldots & \\ 0 & q & 0 & p & 0 & \ldots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where $0 < p < 1$ and $q = 1 - p$. X_t could represent, for example, the number of customers who are waiting in a queue at time t.

All states can be reached from each other, so the chain is irreducible and every state is either recurrent or transient. The equation $\boldsymbol{\pi} = \boldsymbol{\pi} P$ becomes

$$\begin{aligned} \pi_0 &= q\,\pi_0 + q\,\pi_1 \ , \\ \pi_1 &= p\,\pi_0 + q\,\pi_2 \ , \\ \pi_2 &= p\,\pi_1 + q\,\pi_3 \ , \\ \pi_3 &= p\,\pi_2 + q\,\pi_4 \ , \end{aligned}$$

and so on. We can solve this set of equation sequentially. If we let $r = p/q$, then we can express the $\pi_1, \pi_2, \ldots$ in terms of π_0 and r as

$$\pi_j = r^j \, \pi_0, \;\; j = 0, 1, 2, \ldots \; .$$

If $p < q$, then $r < 1$ and $\sum_{j=0}^{\infty} \pi_j = \pi_0/(1-r)$, and by choosing $\pi_0 = 1 - r$, we can make the sum $\sum \pi_j = 1$. Hence, for $r < 1$ we have found the limiting distribution $\boldsymbol{\pi} = (1-r)(1, r, r^2, r^3, \ldots)$ for this Markov chain, and all the states are therefore positive recurrent. On the other hand, when $p \geqslant q$, $\sum \pi_j$ is either 0 or infinite, and hence all states are either null-recurrent or transient. (It can be shown that only the case $p = q$ leads to null-recurrent states.)

Let X be a Markov chain with limiting distribution $\boldsymbol{\pi}$. Suppose $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$. Then, combining (1.33) and (1.35), we have $\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}$. Thus, if the initial distribution of the Markov chain is equal to the limiting distribution, then the distribution of X_t is the same for all t (and is given by this limiting distribution). In fact, it is not difficult to show that for any k the distribution of $X_k, X_{k+1}, X_{k+2} \ldots$ is the same as that of $X_0, X_1, \ldots$. In other words, when $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$, the Markov chain is a stationary stochastic process. More formally, a stochastic process $\{X_t, t \in \mathbb{N}\}$ is called *stationary* if, for any positive $\tau, t_1, \ldots, t_n$, the vector $(X_{t_1}, \ldots, X_{t_n})$ has the same distribution as $(X_{t_1+\tau}, \ldots, X_{t_n+\tau})$. Similar definitions hold when the index set is $\mathbb{Z}$, $\mathbb{R}_+$ or $\mathbb{R}$. For this reason, any distribution $\boldsymbol{\pi}$ for which (1.35) holds is called a *stationary distribution*.

Noting that $\sum_j p_{ij} = 1$, we can rewrite (1.35) as the system of equations

$$\sum_j \pi_i \, p_{ij} = \sum_j \pi_j \, p_{ji} \quad \text{for all } i \in \mathscr{E} \; . \tag{1.36}$$

These are called the *global balance equations*. We can interpret (1.35) as the statement that the "probability flux" out of i is balanced by the probability flux into i. An important generalization, which follows directly from (1.36), states that the same balancing of probability fluxes holds for an arbitrary set $\mathscr{A}$. That is, for every set $\mathscr{A}$ of states we have

$$\sum_{i \in \mathscr{A}} \sum_{j \notin \mathscr{A}} \pi_i \, p_{ij} = \sum_{i \in \mathscr{A}} \sum_{j \notin \mathscr{A}} \pi_j \, p_{ji} \; . \tag{1.37}$$

1.12.4 Reversibility

Reversibility is an important notion in the theory of Markov and more general processes. A stationary stochastic process $\{X_t\}$ with index set $\mathbb{Z}$ or $\mathbb{R}$ is said to be *reversible* if, for any positive integer n and for all $t_1, \ldots, t_n$, the vector $(X_{t_1}, \ldots, X_{t_n})$ has the same distribution as $(X_{-t_1}, \ldots, X_{-t_n})$. One way to visualize this is to imagine that we have taken a video of the stochastic process, which we may run in forward and reverse time. If we cannot determine whether the video is running forward or backward, the process is reversible. The main result for reversible Markov chains is that a stationary Markov process is reversible if and only if there exists a collection of positive numbers $\{\pi_i, \; i \in \mathscr{E}\}$ summing to unity that satisfy the *detailed (or local) balance equations*

$$\pi_i \, p_{ij} = \pi_j \, p_{ji} \; , \quad i, j \in \mathscr{E}. \tag{1.38}$$

Whenever such a collection $\{\pi_j\}$ exists, it is the stationary distribution of the process.

A good way to think of the detailed balance equations is that they balance the probability flux from state i to state j with that from state j to state i. Contrast this with the equilibrium equations (1.36), which balance the probability flux out of state i with that into state i.

Kolmogorov's criterion is a simple criterion for reversibility based on the transition probabilities. It states that a stationary Markov process is reversible if and only if its transition rates satisfy

$$p(i_1, i_2)\, p(i_2, i_3) \ldots p(i_{n-1}, i_n)\, p(i_n, i_1) = p(i_1, i_n)\, p(i_n, i_{n-1}) \ldots p(i_2, i_1) \qquad (1.39)$$

for all finite loops of states $i_1, \ldots, i_n, i_1$. (For clarity, we have used the notation $p(i,j)$ rather than p_{ij} for the transition probabilities.) The idea is quite intuitive: if the process in forward time is more likely to traverse a certain closed loop in one direction than in the opposite direction, then in backward time it will exhibit the opposite behavior, and hence we have a criterion for detecting the direction of time. If such "looping" behavior does not occur, the process must be reversible.

1.12.5 Markov Jump Processes

A *Markov jump process* $X = \{X_t, t \geqslant 0\}$ can be viewed as a continuous-time generalization of a Markov chain and also of a Poisson process. The Markov property (1.30) now reads

$$\mathbb{P}(X_{t+s} = x_{t+s} \mid X_u = x_u, u \leqslant t) = \mathbb{P}(X_{t+s} = x_{t+s} \mid X_t = x_t)\;. \qquad (1.40)$$

As in the Markov chain case, one usually assumes that the process is *time-homogeneous*, that is, $\mathbb{P}(X_{t+s} = j \mid X_t = i)$ does not depend on t. Denote this probability by $P_s(i,j)$. An important quantity is the *transition rate* q_{ij} from state i to j, defined for $i \neq j$ as

$$q_{ij} = \lim_{t \downarrow 0} \frac{P_t(i,j)}{t}\;.$$

The sum of the rates out of state i is denoted by q_i. A typical sample path of X is shown in Figure 1.6. The process jumps at times $T_1, T_2, \ldots$ to states $Y_1, Y_2, \ldots$, staying some length of time in each state.

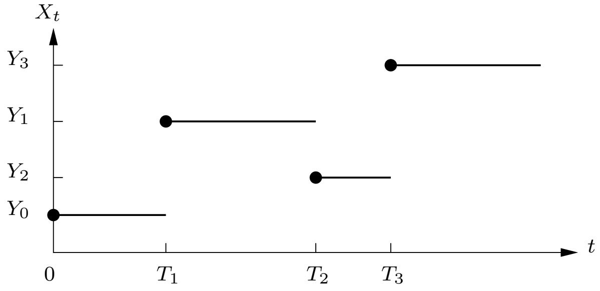


Figure 1.6 A sample path of a Markov jump process $\{X_t, t \geqslant 0\}$.

More precisely, a Markov jump process X behaves (under suitable regularity conditions; see [4]) as follows:

1. Given its past, the probability that X jumps from its current state i to state j is $K_{ij} = q_{ij}/q_i$.
2. The amount of time that X spends in state j has an exponential distribution with mean $1/q_j$, independent of its past history.

The first statement implies that the process $\{Y_n\}$ is in fact a Markov chain, with transition matrix $K = (K_{ij})$.

A convenient way to describe a Markov jump process is through its *transition rate graph*. This is similar to a transition graph for Markov chains. The states are represented by the nodes of the graph, and a transition rate from state i to j is indicated by an arrow from i to j with weight q_{ij}.

■ EXAMPLE 1.12 Birth and Death Process

A *birth and death process* is a Markov jump process with a transition rate graph of the form given in Figure 1.7. Imagine that X_t represents the total number of individuals in a population at time t. Jumps to the right correspond to births, and jumps to the left to deaths. The *birth rates* $\{b_i\}$ and the *death rates* $\{d_i\}$ may differ from state to state. Many applications of Markov chains involve processes of this kind. Note that the process jumps from one state to the next according to a Markov chain with transition probabilities $K_{0,1} = 1$, $K_{i,i+1} = b_i/(b_i + d_i)$, and $K_{i,i-1} = d_i/(b_i + d_i)$, $i = 1, 2, \ldots$. Moreover, it spends an $\mathsf{Exp}(b_0)$ amount of time in state 0 and $\mathsf{Exp}(b_i+d_i)$ in the other states.

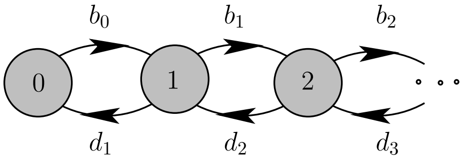


Figure 1.7 The transition rate graph of a birth and death process.

Limiting Behavior We now formulate the continuous-time analogues of (1.34) and Theorem 1.12.2. Irreducibility and recurrence for Markov jump processes are defined in the same way as for Markov chains. For simplicity, we assume that $\mathscr{E} = \{1, 2, \ldots\}$. If X is a recurrent and irreducible Markov jump process, then irrespective of i,

$$\lim_{t\to\infty} \mathbb{P}(X_t = j \mid X_0 = i) = \pi_j \tag{1.41}$$

for some number $\pi_j \geqslant 0$. Moreover, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ is the solution to

$$\sum_{j\neq i} \pi_i \, q_{ij} = \sum_{j\neq i} \pi_j \, q_{ji}, \quad \text{for all } i = 1, \ldots, m \tag{1.42}$$

with $\sum_j \pi_j = 1$, if such a solution exists, in which case all states are positive recurrent. If such a solution does not exist, all π_j are 0.

As in the Markov chain case, $\{\pi_j\}$ is called the *limiting distribution* of X and is usually identified with the row vector $\boldsymbol{\pi}$. Any solution $\boldsymbol{\pi}$ of (1.42) with $\sum_j \pi_j = 1$ is called a *stationary distribution*, since taking it as the initial distribution of the Markov jump process renders the process stationary.

The equations (1.42) are again called the *global balance equations* and are readily generalized to (1.37), replacing the transition probabilities with transition rates. More importantly, if the process is reversible, then, as with Markov chains, the stationary distribution can be found from the *local balance equations*:

$$\pi_i \, q_{ij} = \pi_j \, q_{ji} \; , \quad i, j \in \mathscr{E} \; . \tag{1.43}$$

Reversibility can be easily verified by checking that looping does not occur, that is, via Kolmogorov's criterion (1.39), replacing the probabilities p with rates q.

■ **EXAMPLE 1.13** $M/M/1$ **Queue**

Consider a service facility where customers arrive at certain random times and are served by a single server. Arriving customers who find the server busy wait in the queue. Customers are served in the order in which they arrive. The interarrival times are exponential random variables with rates λ, and the service times of customers are iid exponential random variables with rates μ. Finally, the service times are independent of the interarrival times. Let X_t be the number of customers in the system at time t. By the memoryless property of the exponential distribution (see Problem 1.7), it is not difficult to see that $X = \{X_t, t \geqslant 0\}$ is a Markov jump process, and in fact a birth and death process with birth rates $b_i = \lambda,\ i = 0, 1, 2, \ldots$ and death rates $d_i = \mu,\ i = 1, 2, \ldots$.

Solving the global balance equations (or, more easily, the local balance equations, since X is reversible), we see that X has a limiting distribution given by

$$\lim_{t \to \infty} \mathbb{P}(X_t = n) = (1 - \varrho)\, \varrho^n, \quad n = 0, 1, 2, \ldots, \tag{1.44}$$

provided that $\varrho = \lambda/\mu < 1$. This means that the expected service time needs to be less than the expected interarrival time for a limiting distribution to exist. In that case, the limiting distribution is also the stationary distribution. In particular, if X_0 is distributed according to (1.44), then X_t has the same distribution for all $t > 0$.

1.13 EFFICIENCY OF ESTIMATORS

In this book we will frequently use

$$\widehat{\ell} = \frac{1}{N} \sum_{i=1}^{N} Z_i \; , \tag{1.45}$$

which presents an *unbiased* estimator of the unknown quantity $\ell = \mathbb{E}[\widehat{\ell}] = \mathbb{E}[Z]$, where $Z_1, \ldots, Z_N$ are independent replications of some random variable Z.

By the central limit theorem, $\widehat{\ell}$ has approximately a $\mathsf{N}(\ell, N^{-1}\mathrm{Var}(Z))$ distribution for large N. We shall estimate $\mathrm{Var}(Z)$ via the *sample variance*

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Z_i - \widehat{\ell})^2 \; .$$

By the law of large numbers, S^2 converges with probability 1 to $\text{Var}(Z)$. Consequently, for $\text{Var}(Z) < \infty$ and large N, the approximate $(1 - \alpha)$ confidence interval for ℓ is given by

$$\left(\widehat{\ell} - z_{1-\alpha/2}\,\frac{S}{\sqrt{N}},\ \widehat{\ell} + z_{1-\alpha/2}\,\frac{S}{\sqrt{N}}\right) ,$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. For example, for $\alpha = 0.05$ we have $z_{1-\alpha/2} = z_{0.975} = 1.96$. The quantity

$$\frac{S/\sqrt{N}}{\widehat{\ell}}$$

is often used in the simulation literature as an accuracy measure for the estimator $\widehat{\ell}$. For large N it converges to the *relative error* of $\widehat{\ell}$, defined as

$$\kappa = \frac{\sqrt{\text{Var}(\widehat{\ell})}}{\mathbb{E}[\widehat{\ell}]} = \frac{\sqrt{\text{Var}(Z)/N}}{\ell} . \tag{1.46}$$

The square of the relative error

$$\kappa^2 = \frac{\text{Var}(\widehat{\ell})}{\ell^2} \tag{1.47}$$

is called the *squared coefficient of variation*.

■ **EXAMPLE 1.14 Estimation of Rare-Event Probabilities**

Consider estimation of the tail probability $\ell = \mathbb{P}(X \geqslant \gamma)$ of some random variable X for a *large* number γ. If ℓ is very small, then the event $\{X \geqslant \gamma\}$ is called a *rare event* and the probability $\mathbb{P}(X \geqslant \gamma)$ is called a *rare-event probability*.

We may attempt to estimate ℓ via (1.45) as

$$\widehat{\ell} = \frac{1}{N}\sum_{i=1}^{N} I_{\{X_i \geqslant \gamma\}} , \tag{1.48}$$

which involves drawing a random sample $X_1, \ldots, X_N$ from the pdf of X and defining the indicators $Z_i = I_{\{X_i \geqslant \gamma\}}$, $i = 1, \ldots, N$. The estimator ℓ thus defined is called the *crude Monte Carlo* (CMC) estimator. For small ℓ the relative error of the CMC estimator is given by

$$\kappa = \frac{\sqrt{\text{Var}(\widehat{\ell})}}{\mathbb{E}[\widehat{\ell}]} = \sqrt{\frac{1-\ell}{N\,\ell}} \approx \sqrt{\frac{1}{N\,\ell}} . \tag{1.49}$$

As a numerical example, suppose that $\ell = 10^{-6}$. In order to estimate ℓ accurately with relative error (say) $\kappa = 0.01$, we need to choose a sample size

$$N \approx \frac{1}{\kappa^2 \ell} = 10^{10} .$$

This shows that estimating small probabilities via CMC estimators is computationally meaningless.

1.13.1 Complexity

The theoretical framework in which one typically examines rare-event probability estimation is based on *complexity theory*, as introduced in [1, 12]. In particular, the estimators are classified either as *polynomial-time* or as *exponential-time*. It is shown in [1, 15] that for an arbitrary estimator, $\widehat{\ell}$ of ℓ, to be polynomial-time as a function of some γ, it suffices that its squared coefficient of variation, κ^2, or its relative error, κ, is bounded in γ by some polynomial function, $p(\gamma)$. For such polynomial-time estimators, the required sample size to achieve a fixed relative error does not grow too fast as the event becomes rarer.

Consider the estimator (1.48) and assume that ℓ becomes very small as $\gamma \to \infty$. Note that

$$\mathbb{E}[Z^2] \geqslant (\mathbb{E}[Z])^2 = \ell^2 \; .$$

Hence, the best one can hope for with such an estimator is that its second moment of Z^2 decreases proportionally to ℓ^2 as $\gamma \to \infty$. We say that the rare-event estimator (1.48) has *bounded relative error* if for all γ

$$\mathbb{E}[Z^2] \leqslant c\,\ell^2 \tag{1.50}$$

for some fixed $c \geqslant 1$. Because bounded relative error is not always easy to achieve, the following weaker criterion is often used. We say that the estimator (1.48) is *logarithmically efficient* (sometimes called *asymptotically optimal*) if

$$\lim_{\gamma\to\infty} \frac{\ln \mathbb{E}[Z^2]}{\ln \ell^2} = 1 \; . \tag{1.51}$$

■ **EXAMPLE 1.15 The CMC Estimator Is Not Logarithmically Efficient**

Consider the CMC estimator (1.48). We have

$$\mathbb{E}[Z^2] = \mathbb{E}[Z] = \ell \; ,$$

so that

$$\lim_{\gamma\to\infty} \frac{\ln \mathbb{E}[Z^2]}{\ln \ell^2(\gamma)} = \frac{\ln \ell}{\ln \ell^2} = \frac{1}{2} \; .$$

Hence, the CMC estimator is not logarithmically efficient, and therefore alternative estimators must be found to estimate small ℓ.

1.14 INFORMATION

In this section we discuss briefly various measures of information in a random experiment. Suppose we describe the measurements in a random experiment via a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with pdf f. Then all the information about the experiment (all of our probabilistic knowledge) is obviously contained in the pdf f. However, in most cases we wish to characterize our information about the experiments with just a few key numbers, such as the *expectation* and the *covariance matrix* of $\mathbf{X}$, which provide information about the mean measurements and the variability of the measurements, respectively. Another informational measure comes from coding and communications theory, where the *Shannon entropy* characterizes the average number of bits needed to transmit a message $\mathbf{X}$ over a (binary) communication channel. Yet another approach to information can be found in statistics. Specifically, in the theory of point estimation, the pdf f depends on a parameter vector $\boldsymbol{\theta}$. The question is how well $\boldsymbol{\theta}$ can be estimated via an outcome of $\mathbf{X}$ — in other words, how much information about $\boldsymbol{\theta}$ is contained in the "data" $\mathbf{X}$. Various measures for this type of information are associated with the *maximum likelihood*, the *score*, and the *(Fisher) information matrix*. Finally, the amount of information in a random experiment can often be quantified via a *distance* concept, such as the *Kullback–Leibler* "distance" (divergence), also called the *cross-entropy*.

1.14.1 Shannon Entropy

One of the most celebrated measures of uncertainty in information theory is the *Shannon entropy*, or simply *entropy*. A good reference is [5], where the entropy of a discrete random variable X with density f is defined as

$$\mathbb{E}\left[\log_2 \frac{1}{f(X)}\right] = -\mathbb{E}\left[\log_2 f(X)\right] = -\sum_{\mathscr{X}} f(x) \log_2 f(x) \; .$$

Here X is interpreted as a random character from an alphabet $\mathscr{X}$, such that $X = x$ with probability $f(x)$. We will use the convention $0 \ln 0 = 0$.

It can be shown that the most efficient way to transmit characters sampled from f over a binary channel is to encode them such that the number of bits required to transmit x is equal to $\log_2(1/f(x))$. It follows that $-\sum_{\mathscr{X}} f(x) \log_2 f(x)$ is the expected bit length required to send a random character $X \sim f$; see [5].

A more general approach, which includes continuous random variables, is to define the entropy of a random variable X with density f by

$$\mathcal{H}(X) = -\mathbb{E}[\ln f(X)] = \begin{cases} -\sum f(x) \ln f(x) & \text{discrete case,} \\ -\int f(x) \ln f(x)\, dx & \text{continuous case.} \end{cases} \tag{1.52}$$

Definition (1.52) can easily be extended to random vectors $\mathbf{X}$ as (in the continuous case)

$$\mathcal{H}(\mathbf{X}) = -\mathbb{E}[\ln f(\mathbf{X})] = -\int f(\mathbf{x}) \ln f(\mathbf{x})\, d\mathbf{x} \; . \tag{1.53}$$

Often $\mathcal{H}(\mathbf{X})$ is called the *joint* entropy of the random variables $X_1, \ldots, X_n$ and is also written as $\mathcal{H}(X_1, \ldots, X_n)$. In the continuous case, $\mathcal{H}(\mathbf{X})$ is frequently referred to as the *differential entropy* to distinguish it from the discrete case.

■ **EXAMPLE 1.16**

Let X have a $\mathsf{Ber}(p)$ distribution for some $0 \leqslant p \leqslant 1$. The density f of X is given by $f(1) = \mathbb{P}(X = 1) = p$ and $f(0) = \mathbb{P}(X = 0) = 1 - p$, so that the entropy of X is

$$\mathcal{H}(X) = -p \, \ln p - (1 - p) \, \ln(1 - p) \; .$$

The graph of the entropy as a function of p is depicted in Figure 1.8. Note that the entropy is maximal for $p = 1/2$, which gives the "uniform" density on $\{0, 1\}$.

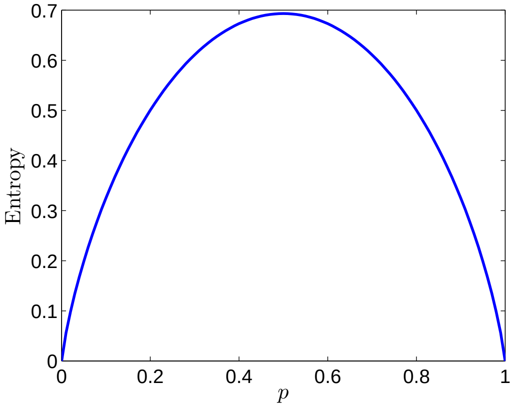


Figure 1.8 The entropy for the $\mathsf{Ber}(p)$ distribution as a function of p.

Next, consider a sequence $X_1, \ldots, X_n$ of iid $\mathsf{Ber}(p)$ random variables. Let $\mathbf{X} = (X_1, \ldots, X_n)$. The density of $\mathbf{X}$, say g, is simply the product of the densities of the X_i, so that

$$\mathcal{H}(\mathbf{X}) = -\mathbb{E}\left[\ln g(\mathbf{X})\right] = -\mathbb{E}\left[\ln \prod_{i=1}^{n} f(X_i)\right] = \sum_{i=1}^{n} -\mathbb{E}\left[\ln f(X_i)\right] = n\,\mathcal{H}(X)\ .$$

The properties of $\mathcal{H}(\mathbf{X})$ in the continuous case are somewhat different from those in the discrete one. In particular:

1. The differential entropy can be negative, whereas the discrete entropy is always positive.

2. The discrete entropy is insensitive to invertible transformations, whereas the differential entropy is not. Specifically, if $\mathbf{X}$ is discrete, $\mathbf{Y} = g(\mathbf{X})$, and g is an invertible mapping, then $\mathcal{H}(\mathbf{X}) = \mathcal{H}(\mathbf{Y})$, because $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))$. However, in the continuous case, we have an additional factor due to the Jacobian of the transformation.

It is not difficult to see that of any density f, the one that gives the maximum entropy is the uniform density on $\mathscr{X}$. That is,

$$\mathcal{H}(\mathbf{X}) \text{ is maximal } \Leftrightarrow f(\mathbf{x}) = \frac{1}{|\mathscr{X}|} \text{ (constant)}\ . \tag{1.54}$$

For two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with joint pdf f, we define the *conditional entropy* of $\mathbf{Y}$ given $\mathbf{X}$ as

$$\mathcal{H}(\mathbf{Y}\,|\,\mathbf{X}) = -\mathbb{E}\left[\ln \frac{f(\mathbf{X},\mathbf{Y})}{f_{\mathbf{X}}(\mathbf{X})}\right] = \mathcal{H}(\mathbf{X},\mathbf{Y}) - \mathcal{H}(\mathbf{X})\ , \tag{1.55}$$

where $f_{\mathbf{X}}$ is the pdf of $\mathbf{X}$ and $\frac{f(\mathbf{x},\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}$ is the conditional density of $\mathbf{Y}$ (at $\mathbf{y}$), given $\mathbf{X} = \mathbf{x}$. It follows that

$$\mathcal{H}(\mathbf{X},\mathbf{Y}) = \mathcal{H}(\mathbf{X}) + \mathcal{H}(\mathbf{Y}\,|\,\mathbf{X}) = \mathcal{H}(\mathbf{Y}) + \mathcal{H}(\mathbf{X}\,|\,\mathbf{Y})\ . \tag{1.56}$$

It is reasonable to impose that any sensible additive measure describing the average amount of uncertainty should satisfy at least (1.56) and (1.54). It follows that the uniform density carries the least amount of information, and the entropy (average amount of uncertainty) of $(\mathbf{X}, \mathbf{Y})$ is equal to the sum of the entropy of $\mathbf{X}$ and the amount of entropy in $\mathbf{Y}$ after the information in $\mathbf{X}$ has been accounted for. It is argued in [11] that any concept of entropy that includes the general properties (1.54) and (1.56) must lead to the definition (1.53).

The *mutual information* of $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathcal{M}(\mathbf{X}, \mathbf{Y}) = \mathcal{H}(\mathbf{X}) + \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{X}, \mathbf{Y}) \ , \tag{1.57}$$

which, as the name suggests, can be interpreted as the amount of information shared by $\mathbf{X}$ and $\mathbf{Y}$. An alternative expression, which follows from (1.56) and (1.57), is

$$\mathcal{M}(\mathbf{X}, \mathbf{Y}) = \mathcal{H}(\mathbf{X}) - \mathcal{H}(\mathbf{X} \,|\, \mathbf{Y}) = \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{Y} \,|\, \mathbf{X}) \ , \tag{1.58}$$

which can be interpreted as the reduction of the uncertainty of one random variable due to the knowledge of the other. It is not difficult to show that the mutual information is always positive. It is also related to the cross-entropy concept, which follows.

1.14.2 Kullback–Leibler Cross-Entropy

Let g and h be two densities on $\mathscr{X}$. The Kullback–Leibler cross-entropy between g and h (compare with (1.53)) is defined (in the continuous case) as

$$\begin{aligned} \mathcal{D}(g, h) &= \mathbb{E}_g \left[\ln \frac{g(\mathbf{X})}{h(\mathbf{X})} \right] \\ &= \int g(\mathbf{x}) \ln g(\mathbf{x}) \, d\mathbf{x} - \int g(\mathbf{x}) \ln h(\mathbf{x}) \, d\mathbf{x} \ . \end{aligned} \tag{1.59}$$

$\mathcal{D}(g, h)$ is also called the *Kullback–Leibler divergence*, the *cross-entropy*, and the *relative entropy*. If not stated otherwise, we shall call $\mathcal{D}(g, h)$ the *cross-entropy* (CE) between g and h. Notice that $\mathcal{D}(g, h)$ is not a distance between g and h in the formal sense, since in general $\mathcal{D}(g, h) \neq \mathcal{D}(h, g)$. Nonetheless, it is often useful to think of $\mathcal{D}(g, h)$ as a distance because

$$\mathcal{D}(g, h) \geqslant 0$$

and $\mathcal{D}(g, h) = 0$ if and only if $g(x) = h(x)$. This follows from Jensen's inequality (if ϕ is a convex function, such as $-\ln$, then $\mathbb{E}[\phi(X)] \geqslant \phi(\mathbb{E}[X])$). Namely,

$$\mathcal{D}(g, h) = \mathbb{E}_g \left[-\ln \frac{h(\mathbf{X})}{g(\mathbf{X})} \right] \geqslant -\ln \left\{ \mathbb{E}_g \left[\frac{h(\mathbf{X})}{g(\mathbf{X})} \right] \right\} = -\ln 1 = 0 \ .$$

It can be readily seen that the mutual information $\mathcal{M}(\mathbf{X}, \mathbf{Y})$ of vectors $\mathbf{X}$ and $\mathbf{Y}$ defined in (1.57) is related to the CE in the following way:

$$\mathcal{M}(\mathbf{X}, \mathbf{Y}) = \mathcal{D}(f, f_{\mathbf{X}} f_{\mathbf{Y}}) = \mathbb{E}_f \left[\ln \frac{f(\mathbf{X}, \mathbf{Y})}{f_{\mathbf{X}}(\mathbf{X}) \, f_{\mathbf{Y}}(\mathbf{Y})} \right] \ ,$$

where f is the (joint) pdf of $(\mathbf{X}, \mathbf{Y})$ and $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$ are the (marginal) pdfs of $\mathbf{X}$ and $\mathbf{Y}$, respectively. In other words, the mutual information can be viewed as the CE that measures the distance between the joint pdf f of $\mathbf{X}$ and $\mathbf{Y}$ and the product of their marginal pdfs $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$, that is, under the assumption that the vectors $\mathbf{X}$ and $\mathbf{Y}$ are *independent*.

Remark 1.14.1 (Other Distance Measures) Instead of Kullback–Leibler distance, one can use several other distance or divergence measures between pdfs. An important class of such "distances" is formed by Csiszár's ϕ*-divergence* [10],

$$d(g,\, h) = \int_{\mathscr{X}} p(\mathbf{x})\, \phi\!\left(\frac{g(\mathbf{x})}{h(\mathbf{x})}\right) d\mathbf{x}\ , \tag{1.60}$$

where ϕ is any function such that $\phi(1) = 0$ and $\phi''(x) > 0,\ x > 0$ (in particular, ϕ is convex). Below is a list of important divergence measures that can be found as special cases of the ϕ-divergence.

- *Burg CE distance*:

$$d(g,\, h) = \int h(\mathbf{x}) \ln \frac{h(\mathbf{x})}{g(\mathbf{x})}\, d\mathbf{x}$$

- *Kullback–Leibler CE distance*:

$$d(g,\, h) = \int g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{h(\mathbf{x})}\, d\mathbf{x}$$

- *Hellinger distance*:

$$d(g,\, h) = 2 \int \left(\sqrt{g(\mathbf{x})} - \sqrt{h(\mathbf{x})}\right)^2 d\mathbf{x}$$

- *Pearson χ^2 discrepancy measure*:

$$d(g,\, h) = \frac{1}{2} \int \frac{\big[g(\mathbf{x}) - h(\mathbf{x})\big]^2}{h(\mathbf{x})}\, d\mathbf{x}$$

- *Neymann χ^2 goodness of fit measure*:

$$d(g,\, h) = \frac{1}{2} \int \frac{\big[g(\mathbf{x}) - h(\mathbf{x})\big]^2}{g(\mathbf{x})}\, d\mathbf{x}$$

1.14.3 The Maximum Likelihood Estimator and the Score Function

We introduce here the notion of the *score function* (SF) via the classical *maximum likelihood estimator*. Consider a random vector $\mathbf{X} = (X_1, \ldots, X_n)$, which is distributed according to a fixed pdf $f(\cdot; \boldsymbol{\theta})$ with unknown parameter (vector) $\boldsymbol{\theta} \in \Theta$. Assume that we wish to estimate $\boldsymbol{\theta}$ on the basis of a given outcome $\mathbf{x}$ (the data) of $\mathbf{X}$. For a given $\mathbf{x}$, the function $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta})$ is called the *likelihood function*. Note that $\mathcal{L}$ is a function of $\boldsymbol{\theta}$ for a fixed parameter $\mathbf{x}$, whereas for the pdf f it is the other way around. The maximum likelihood *estimate* $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$ of $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}; \mathbf{x})\ . \tag{1.61}$$

Because the function $\ln$ is monotone increasing, we also have

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) \ . \tag{1.62}$$

The random variable $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ with $\mathbf{X} \sim f(\cdot; \boldsymbol{\theta})$ is the corresponding maximum likelihood *estimator*, which is again written as $\widehat{\boldsymbol{\theta}}$. Note that often the data $X_1, \ldots, X_n$ form a random sample from some pdf $f_1(\cdot; \boldsymbol{\theta})$, in which case $f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{N} f_1(x_i; \boldsymbol{\theta})$ and

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \ln f_1(X_i; \boldsymbol{\theta}) \ . \tag{1.63}$$

If $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ is a continuously differentiable concave function with respect to $\boldsymbol{\theta}$ and the maximum is attained in the interior of Θ, then we can find the maximum likelihood estimator of $\boldsymbol{\theta}$ by solving

$$\nabla_{\boldsymbol{\theta}} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \mathbf{0} \ .$$

The function $\mathcal{S}(\cdot; \mathbf{x})$ defined by

$$\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}) = \nabla_{\boldsymbol{\theta}} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{x}; \boldsymbol{\theta})} \tag{1.64}$$

is called the *score function*. For the exponential family (A.9) it is easy to see that

$$\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}) = \frac{\nabla c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})} + \mathbf{t}(\mathbf{x}) \ . \tag{1.65}$$

The *random vector* $\mathcal{S}(\boldsymbol{\theta}) = \mathcal{S}(\boldsymbol{\theta}; \mathbf{X})$ with $\mathbf{X} \sim f(\cdot; \boldsymbol{\theta})$ is called the *(efficient) score*. The expected score is always equal to the zero vector, that is

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{S}(\boldsymbol{\theta})] = \int \nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta}) \, \mu(d\mathbf{x}) = \nabla_{\boldsymbol{\theta}} \int f(\mathbf{x}; \theta) \, \mu(d\mathbf{x}) = \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0} \ ,$$

where the interchange of differentiation and integration is justified via the bounded convergence theorem.

1.14.4 Fisher Information

The covariance matrix $\mathcal{I}(\boldsymbol{\theta})$ of the score $\mathcal{S}(\boldsymbol{\theta})$ is called the *Fisher information matrix*. Since the expected score is always $\mathbf{0}$, we have

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\mathcal{S}(\boldsymbol{\theta}) \mathcal{S}(\boldsymbol{\theta})^T \right] \ . \tag{1.66}$$

In the one-dimensional case we thus have

$$\mathcal{I}(\theta) = \mathbb{E}_{\theta} \left[\left(\frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2 \right] \ .$$

Because

$$\frac{\partial^2}{\partial \theta^2} \ln f(x; \theta) = \frac{\frac{\partial^2}{\partial \theta^2} f(x; \theta)}{f(x; \theta)} - \left(\frac{\frac{\partial}{\partial \theta} f(x; \theta)}{f(x; \theta)} \right)^2 ,$$

we see that (under straightforward regularity conditions) the Fisher information is also given by

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2} \right] \; .$$

In the multidimensional case we have similarly

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}} \left[\nabla \mathcal{S}(\boldsymbol{\theta}) \right] = -\mathbb{E}_{\boldsymbol{\theta}} \left[\nabla^2 \ln f(\mathbf{X};\boldsymbol{\theta}) \right] \; , \tag{1.67}$$

where $\nabla^2 \ln f(\mathbf{X};\boldsymbol{\theta})$ denotes the *Hessian* of $\ln f(\mathbf{X};\boldsymbol{\theta})$, that is, the (random) matrix

$$\left(\frac{\partial^2 \ln f(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \; .$$

The importance of the Fisher information in statistics is corroborated by the famous *Cramér-Rao inequality*, which (in a simplified form) states that the variance of any unbiased estimator Z of $g(\boldsymbol{\theta})$ is bounded from below via

$$\mathrm{Var}(Z) \geqslant (\nabla g(\boldsymbol{\theta}))^T \, \mathcal{I}^{-1}(\boldsymbol{\theta}) \, \nabla g(\boldsymbol{\theta}) \; . \tag{1.68}$$

For more details see [13].

1.15 CONVEX OPTIMIZATION AND DUALITY

Let $f(x),\ x \in \mathbb{R}$, be a real-valued function with continuous derivatives — also called a C^1 function. The standard approach to minimizing $f(x)$ is to solve the equation

$$f'(x) = 0 \; . \tag{1.69}$$

The solutions to (1.69) are called *stationary points*. If, in addition, the function has continuous second derivatives (a so-called C^2 function), the condition

$$f''(x^*) > 0 \tag{1.70}$$

ensures that a stationary point x^* is a *local minimizer*, that is, $f(x^*) < f(x)$ for all x in a small enough neighborhood of x^*.

For a C^1 function on $\mathbb{R}^n$, (1.69) generalizes to

$$\nabla f(\mathbf{x}) \equiv \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix} = \mathbf{0} \; , \tag{1.71}$$

where $\nabla f(\mathbf{x})$ is the *gradient* of f at $\mathbf{x}$. Similarly, a stationary point $\mathbf{x}^*$ is a local minimizer of f if the *Hessian matrix* (or simply *Hessian*) at $\mathbf{x}^*$,

$$\nabla^2 f(\mathbf{x}^*) \equiv \begin{pmatrix} \frac{\partial^2 f(\mathbf{x}^*)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\mathbf{x}^*)}{\partial x_1 \partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 f(\mathbf{x}^*)}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f(\mathbf{x}^*)}{\partial x_n^2} \end{pmatrix} , \tag{1.72}$$

is *positive definite*, that is, $\mathbf{x}^T \left[\nabla^2 f(\mathbf{x}^*) \right] \mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

The situation can be further generalized by introducing *constraints*. A general constrained optimization problems can be written as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{1.73}$$

$$\text{subject to:} \quad h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, m \tag{1.74}$$

$$g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \ldots, k. \tag{1.75}$$

Here $f,\ g_i$ and h_i are given functions, $f(\mathbf{x})$ is called the *objective function*, and $h_i(\mathbf{x}) = 0$ and $g_i(\mathbf{x}) \leqslant 0$ represent the *equality* and *inequality* constraints, respectively.

The region of the domain where the objective function is defined and where all the constraints are satisfied is called the *feasible region*. A *global solution* to the optimization problem is a point $\mathbf{x}^* \in \mathbb{R}^n$ such that there exists no other point $\mathbf{x} \in \mathbb{R}^n$ for which $f(\mathbf{x}) < f(\mathbf{x}^*)$. Alternative names are *global minimizer* and *global minimum*, although the latter could be confused with the minimum value of the function. Similarly, for a *local* solution/minimizer, the condition $f(\mathbf{x}) < f(\mathbf{x}^*)$ only needs to hold in some neighborhood of $\mathbf{x}^*$.

Within this formulation fall many of the traditional optimization problems. An optimization problem in which the objective function and the equality and inequality constraints are linear functions, is called a *linear program*. An optimization problem in which the objective function is quadratic, while the constraints are linear functions is called a *quadratic program*. Convexity plays an important role in many practical optimization problems.

Definition 1.15.1 (Convex Set) A set $\mathscr{X} \in \mathbb{R}^n$ is called *convex* if, for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$ and $\theta \in (0, 1)$, the point $(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \in \mathscr{X}$.

Definition 1.15.2 (Convex Function) A function $f(\mathbf{x})$ on a convex set $\mathscr{X}$ is called *convex* if, for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$ and $\theta \in (0, 1)$,

$$f\big(\theta\mathbf{x} + (1-\theta)\mathbf{y}\big) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})\ . \tag{1.76}$$

If a strict inequality in (1.76) holds, the function is said to be *strictly convex*. If a function f is (strictly) convex, then $-f$ is said to be (strictly) *concave*. Assuming $\mathscr{X}$ is an open set, convexity for $f \in C^1$ is equivalent to

$$f(\mathbf{y}) \geqslant f(\mathbf{x}) + (\mathbf{y}-\mathbf{x})^T \nabla f(\mathbf{x}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathscr{X}\ .$$

Moreover, for $f \in C^2$, convexity is equivalent to the Hessian matrix being positive semidefinite for all $\mathbf{x} \in \mathscr{X}$:

$$\mathbf{y}^T \left[\nabla^2 f(\mathbf{x})\right] \mathbf{y} \geqslant 0, \quad \text{for all } \mathbf{y} \in \mathbb{R}^n\ .$$

The problem (1.73) is said to be a *convex programming problem* if

1. the objective function f is convex,
2. the inequality constraint functions $\{g_i(\mathbf{x})\}$ are convex, and
3. the equality constraint functions $\{h_i(\mathbf{x})\}$ are *affine*, that is, of the form $\boldsymbol{a}_i^T\mathbf{x} - b_i$.

Note that the last requirement follows from the fact that an equality constraint $h_i(\mathbf{x}) = 0$ can be viewed as a combination of the inequality constraints $h_i(\mathbf{x}) \leqslant 0$ and $-h_i(\mathbf{x}) \leqslant 0$, so that both h_i and $-h_i$ need to be convex. Both the linear and quadratic programs (with positive definite matrix C) are convex.

1.15.1 Lagrangian Method

The main components of the Lagrangian method are the Lagrange multipliers and the Lagrange function. The method was developed by Lagrange in 1797 for the optimization problem (1.73) with equality constraints (1.74). In 1951 Kuhn and Tucker extended Lagrange's method to inequality constraints.

Definition 1.15.3 (Lagrange Function) Given an optimization problem (1.73) containing only equality constraints $h_i(\mathbf{x}) = 0,\ i = 1, \ldots, m$, the *Lagrange function*, or *Lagrangian*, is defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_i \beta_i \, h_i(\mathbf{x}) \ ,$$

where the coefficients $\{\beta_i\}$ are called the *Lagrange multipliers*.

A necessary condition for a point $\mathbf{x}^*$ to be a local minimizer of $f(\mathbf{x})$ subject to the equality constraints $h_i(\mathbf{x}) = 0,\ i = 1, \ldots, m$, is

$$\begin{aligned} \nabla_{\mathbf{x}} \, \mathcal{L}(\mathbf{x}^*, \boldsymbol{\beta}^*) &= \mathbf{0} \ , \\ \nabla_{\boldsymbol{\beta}} \, \mathcal{L}(\mathbf{x}^*, \boldsymbol{\beta}^*) &= \mathbf{0} \end{aligned}$$

for some value $\boldsymbol{\beta}^*$. The above conditions are also sufficient if $\mathcal{L}(\mathbf{x}, \boldsymbol{\beta}^*)$ is a convex function of $\mathbf{x}$.

■ **EXAMPLE 1.17 Maximum Entropy Distribution**

Let $p = \{p_i, i = 1, \ldots, n\}$ be a probability distribution. Consider the following program, which maximizes the (Shannon) entropy:

$$\begin{aligned} \max_{\mathbf{p}} \quad & -\sum_{i=1}^{n} p_i \ln p_i \\ \text{subject to:} \quad & \sum_{i=1}^{n} p_i = 1 \ . \end{aligned}$$

The Lagrangian is

$$\mathcal{L}(\mathbf{p}, \beta) = \sum_{i=1}^{n} p_i \ln p_i + \beta \left(\sum_{i=1}^{n} p_i - 1 \right)$$

over the domain $\{(\mathbf{p}, \beta) : p_i \geq 0, i = 1, \ldots, n,\ \beta \in \mathbb{R}\}$. The optimal solution $\mathbf{p}^*$ of the problem is the uniform distribution, that is, $\mathbf{p}^* = (1/n, \ldots, 1/n)$; see Problem 1.35.

Definition 1.15.4 (Generalized Lagrange Function) Given the original optimization problem (1.73), containing both the equality and inequality constraints, the *generalized Lagrange function*, or simply *Lagrangian*, is defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{i=1}^{k} \alpha_i \, g_i(\mathbf{x}) + \sum_{i=1}^{m} \beta_i \, h_i(\mathbf{x}) \ .$$

A necessary condition for a point $\mathbf{x}^*$ to be a local minimizer of $f(\mathbf{x})$ in the optimization problem (1.73) is the existence of an $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ such that

$$\begin{aligned}
\nabla_{\mathbf{x}}\, \mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) &= \mathbf{0}, \\
\nabla_{\boldsymbol{\beta}}\, \mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) &= \mathbf{0}\ , \\
g_i(\mathbf{x}^*) &\leqslant 0, \quad i = 1, \ldots, k\ , \\
\alpha_i^* &\geqslant 0, \quad i = 1, \ldots, k\ , \\
\alpha_i^*\, g_i(\mathbf{x}^*) &= 0, \quad i = 1, \ldots, k\ .
\end{aligned}$$

These equations are usually referred as the *Karush-Kuhn-Tucker (KKT) conditions*. For *convex* programs we have the following important results:

1. Every local solution $\mathbf{x}^*$ to a convex programming problem is a global solution and the set of global solutions is convex. If, in addition, the objective function is strictly convex, then any global solution is unique.

2. For a strictly convex programming problem with C^1 objective and constraint functions, the KKT conditions are necessary and sufficient for a unique global solution.

1.15.2 Duality

The aim of duality is to provide an alternative formulation of an optimization problem that is often more computationally efficient or has some theoretical significance (see [8], page 219). The original problem (1.73) is referred to as the *primal* problem, whereas the reformulated problem, based on Lagrange multipliers, is referred to as the *dual* problem. Duality theory is most relevant to convex optimization problems. It is well known that if the primal optimization problem is (strictly) convex, then the dual problem is (strictly) concave and has a (unique) solution from which the optimal (unique) primal solution can be deduced.

Definition 1.15.5 (Lagrange Dual Program) The *Lagrange dual program* of the primal program (1.73), is

$$\begin{aligned}
\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \quad & \mathcal{L}^*(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\
\text{subject to:} \quad & \boldsymbol{\alpha} \geqslant 0\ ,
\end{aligned}$$

where $\mathcal{L}^*$ is the *Lagrange dual function*:

$$\mathcal{L}^*(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\mathbf{x} \in \mathscr{X}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})\ . \tag{1.77}$$

It is not difficult to see that if f^* is the minimal value of the primal problem, then $\mathcal{L}^*(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant f^*$ for any $\boldsymbol{\alpha} \geqslant 0$ and any $\boldsymbol{\beta}$. This property is called *weak duality*. The Lagrangian dual program thus determines the best lower bound on f^*. If d^* is the optimal value for the dual problem then $d^* < f^*$. The difference $f^* - d^*$ is called the *duality gap*.

The duality gap is extremely useful for providing lower bounds for the solutions of primal problems that may be impossible to solve directly. It is important to note that for linearly constrained problems, if the primal is infeasible (does not have a solution satisfying the constraints), then the dual is either infeasible or unbounded. Conversely, if the dual is infeasible, then the primal has no solution. Of crucial importance is the *strong duality* theorem, which states that for convex programs (1.73) with linear constrained functions h_i and g_i the duality gap is zero, and any $\mathbf{x}^*$ and $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ satisfying the KKT conditions are (global) solutions to the primal and dual programs, respectively. In particular, this holds for linear and convex quadratic programs (note that not all quadratic programs are convex).

For a convex primal program with C^1 objective and constraint functions, the Lagrangian dual function (1.77) can be obtained by simply setting the gradient (with respect to $\mathbf{x}$) of the Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ to zero. One can further simplify the dual program by substituting into the Lagrangian the relations between the variables thus obtained.

■ EXAMPLE 1.18 Linear Programming Problem

Consider the following linear programming problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{c}^T \mathbf{x} \\ \text{subject to:} \quad & A\mathbf{x} \geqslant \mathbf{b} \; . \end{aligned}$$

The Lagrangian is $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{c}^T\mathbf{x} - \boldsymbol{\alpha}^T(A\mathbf{x} - \mathbf{b})$. The Lagrange dual function is the infimum of $\mathcal{L}$ over all $\mathbf{x}$; thus,

$$\mathcal{L}^*(\boldsymbol{\alpha}) = \begin{cases} \mathbf{b}^T\boldsymbol{\alpha} & \text{if} \quad A^T\boldsymbol{\alpha} = \mathbf{c} \; , \\ -\infty & \text{otherwise,} \end{cases}$$

so that the Lagrange dual program becomes

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \mathbf{b}^T \boldsymbol{\alpha} \\ \text{subject to:} \quad & A^T\boldsymbol{\alpha} = \mathbf{c} \\ & \boldsymbol{\alpha} \geqslant \mathbf{0} \; . \end{aligned}$$

It is interesting to note that for the linear programming problem the dual of the dual problem always gives back the primal problem.

■ EXAMPLE 1.19 Quadratic Programming Problem

Consider the following quadratic programming problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2}\,\mathbf{x}^T C \mathbf{x} \\ \text{subject to:} \quad & C\mathbf{x} \geqslant \mathbf{b} \; , \end{aligned}$$

where the $n \times n$ matrix C is assumed to be positive definite (for a general quadratic programming problem the matrix C can always be assumed to be symmetric, but it is not necessarily positive definite). The Lagrangian is $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{x}^T C\mathbf{x} - \boldsymbol{\alpha}^T(C\mathbf{x} - \mathbf{b})$. We can minimize this by taking its gradient with respect to $\mathbf{x}$ and setting it to zero. This gives $C\mathbf{x} - C\boldsymbol{\alpha} = C(\mathbf{x} - \boldsymbol{\alpha}) = \mathbf{0}$. The positive definiteness of C implies that $\mathbf{x} = \boldsymbol{\alpha}$. The maximization of the Lagrangian is now reduced to maximizing $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\alpha}) = \frac{1}{2}\,\boldsymbol{\alpha}^T C\boldsymbol{\alpha} - \boldsymbol{\alpha}^T(C\boldsymbol{\alpha} - \mathbf{b}) = -\frac{1}{2}\,\boldsymbol{\alpha}^T C\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\mathbf{b}$ subject to $\boldsymbol{\alpha} \geqslant \mathbf{0}$. Hence, we can write the dual problem as

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2}\,\boldsymbol{\alpha}^T C \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \mathbf{b} \\ \text{subject to:} \quad & \boldsymbol{\alpha} \geqslant \mathbf{0} \; . \end{aligned}$$

Notice that the dual problem involves only simple nonnegativity constraints.

Now suppose that we are given the Cholesky factorization $C = BB^T$. It turns out (see Problem 1.36) that the Lagrange dual of the above dual problem can be written as

$$\begin{aligned} \min_{\boldsymbol{\mu}} \quad & \frac{1}{2}\,\boldsymbol{\mu}^T \boldsymbol{\mu} \\ \text{subject to:} \quad & B\boldsymbol{\mu} \geqslant \mathbf{b}\,, \end{aligned} \tag{1.78}$$

with $\boldsymbol{\mu} = B^T \boldsymbol{\alpha}$. This is a so-called *least distance* problem, which, provided we know the Cholesky factorization of C, is easier to solve than the original quadratic programming problem.

A final example of duality is provided by the widely used *minimum cross-entropy method* [10].

■ EXAMPLE 1.20 Minimum Cross-Entropy (MinxEnt) Method

Let $\mathbf{X}$ be a discrete random variable (or vector) taking values $\mathbf{x}_1, \ldots, \mathbf{x}_r$, and let $\mathbf{q} = (q_1, \ldots, q_r)^T$ and $\mathbf{p} = (p_1, \ldots, p_r)^T$ be two strictly positive distribution (column) vectors for $\mathbf{X}$. Consider the following primal program of minimizing the cross-entropy of $\mathbf{p}$ and $\mathbf{q}$, that is, $\sum_{i=1}^n p_i \ln(p_i/q_i)$, for a fixed $\mathbf{q}$, subject to linear equality constraints:

$$\min_{\mathbf{p}} \quad \sum_{k=1}^r p_k \ln \frac{p_k}{q_k} \tag{1.79}$$

$$\text{subject to:} \quad \mathbb{E}_{\mathbf{p}}[S_i(\mathbf{X})] = \sum_{k=1}^r S_i(\mathbf{x}_k)\, p_k = \gamma_i, \quad i = 1, \ldots, m \tag{1.80}$$

$$\sum_{k=1}^r p_k = 1\,, \tag{1.81}$$

where $S_1, \ldots, S_m$ are arbitrary functions.

Here the objective function is convex, since it is a linear combination of functions of the form $p \ln(p/c)$, which are convex on $\mathbb{R}_+$, for any $c > 0$. In addition, the equality constraint functions are affine (of the form $\mathbf{a}^T\mathbf{p} - \gamma$). Therefore, this problem is convex. To derive the optimal solution $\mathbf{p}^*$ of the above primal program, it is typically easier to solve the associated *dual* program [10]. Below we present the corresponding procedure.

1. The Lagrangian of the primal problem is given by

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta) = \sum_{k=1}^r p_k \ln \frac{p_k}{q_k} - \sum_{i=1}^m \lambda_i \Big(\sum_{k=1}^r S_i(\mathbf{x}_k)\, p_k - \gamma_i \Big) + \beta \Big(\sum_{k=1}^r p_k - 1 \Big)\,, \tag{1.82}$$

   where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^T$ is the Lagrange multiplier vector corresponding to (1.80) and β is the Lagrange multiplier corresponding to (1.81). Note that we can use either a plus or a minus sign in the second sum of (1.82). We choose the latter, because later on we generalize the above problem to inequality ($\geqslant$) constraints in (1.80), giving rise to a minus sign in the Lagrangian.

2. Solve (for fixed $\boldsymbol{\lambda}$ and β)
$$\min_{\mathbf{p}} \mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta) \tag{1.83}$$
by solving
$$\nabla_{\mathbf{p}} \mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta) = \mathbf{0} \ ,$$
which gives the set of equations
$$\nabla_{p_k} \mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta) = \ln \frac{p_k}{q_k} + 1 - \sum_{i=1}^{m} \lambda_i \, S_i(\mathbf{x}_k) + \beta = 0, \quad k = 1, \ldots, r \ .$$
Denote the optimal solution and the optimal function value obtained from the program (1.83) as $\mathbf{p}(\boldsymbol{\lambda}, \beta)$ and $\mathcal{L}^*(\boldsymbol{\lambda}, \beta)$, respectively. The latter is the Lagrange dual function. We have
$$p_k(\boldsymbol{\lambda}, \beta) = q_k \exp\left(-\beta - 1 + \sum_{i=1}^{m} \lambda_i \, S_i(\mathbf{x}_k)\right), \quad k = 1, \ldots, r \ . \tag{1.84}$$
Since the sum of the $\{p_k\}$ must be 1, we obtain
$$\mathrm{e}^{\beta} = \sum_{k=1}^{r} q_k \exp\left(-1 + \sum_{i=1}^{m} \lambda_i \, S_i(\mathbf{x}_k)\right) \ . \tag{1.85}$$
Substituting $\mathbf{p}(\boldsymbol{\lambda}, \beta)$ back into the Lagrangian gives
$$\mathcal{L}^*(\boldsymbol{\lambda}, \beta) = -1 + \sum_{i=1}^{m} \lambda_i \, \gamma_i - \beta \ . \tag{1.86}$$

3. Solve the *dual* program
$$\max_{\boldsymbol{\lambda}, \beta} \mathcal{L}^*(\boldsymbol{\lambda}, \beta) \ . \tag{1.87}$$
Since β and $\boldsymbol{\lambda}$ are related via (1.85), solving (1.87) can be done by substituting the corresponding $\beta(\boldsymbol{\lambda})$ into (1.86) and optimizing the resulting function:
$$D(\boldsymbol{\lambda}) = -1 + \sum_{i=1}^{m} \lambda_i \, \gamma_i - \ln \left\{ \sum_{k=1}^{m} q_k \, \exp\{-1 + \textstyle\sum_{i=1}^{m} \lambda_i \, S_i(\mathbf{x}_k)\} \right\}. \tag{1.88}$$
Since $D(\boldsymbol{\lambda})$ is continuously differentiable and concave with respect to $\boldsymbol{\lambda}$, we can derive the optimal solution, $\boldsymbol{\lambda}^*$, by solving
$$\nabla_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda}) = \mathbf{0} \ , \tag{1.89}$$
which can be written componentwise in the following explicit form:
$$\begin{aligned} \nabla_{\lambda_j} D(\boldsymbol{\lambda}) &= \gamma_i - \frac{\sum_{k=1}^{r} S_i(\mathbf{x}_k) \, q_k \exp\left\{-1 + \sum_{j=1}^{m} \lambda_j \, S_j(\mathbf{x}_k)\right\}}{\sum_{k=1}^{r} q_k \exp\left\{-1 + \sum_{j=1}^{m} \lambda_j \, S_j(\mathbf{x}_k)\right\}} \\ &= \gamma_i - \frac{\mathbb{E}_{\mathbf{q}}\left[S_i(\mathbf{X}) \exp\left\{-1 + \sum_{j=1}^{m} \lambda_j \, S_j(\mathbf{X})\right\}\right]}{\mathbb{E}_{\mathbf{q}}\left[\exp\left\{-1 + \sum_{j=1}^{m} \lambda_j \, S_j(\mathbf{X})\right\}\right]} = 0 \end{aligned} \tag{1.90}$$

for $j = 1, \ldots, m$. The optimal vector $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ can be found by solving (1.90) numerically. Note that if the primal program has a nonempty interior optimal solution, then the dual program has an optimal solution $\boldsymbol{\lambda}^*$.

4. Finally, substitute $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ and $\beta = \beta(\boldsymbol{\lambda}^*)$ back into (1.84) to obtain the solution to the original MinxEnt program.

It is important to note that we do not need to explicitly impose the conditions $p_i \geqslant 0,\ i = 1, \ldots, n$, because the quantities $\{p_i\}$ in (1.84) are automatically strictly positive. This is a crucial property of the CE distance; see also [2]. It is instructive (see Problem1.37) to verify how adding the nonnegativity constraints affects the above procedure.

When inequality constraints $\mathbb{E}_{\mathbf{p}}[S_i(\mathbf{X})] \geqslant \gamma_i$ are used in (1.80) instead of equality constraints, the solution procedure remains almost the same. The only difference is that the Lagrange multiplier vector $\boldsymbol{\lambda}$ must now be nonnegative. It follows that the dual program becomes

$$\begin{aligned} \max_{\boldsymbol{\lambda}} &\quad D(\boldsymbol{\lambda}) \\ \text{subject to:} &\quad \boldsymbol{\lambda} \geqslant \mathbf{0}\ , \end{aligned}$$

with $D(\boldsymbol{\lambda})$ given in (1.88).

A further generalization is to replace the above discrete optimization problem with a *functional* optimization problem. This topic will be discussed in Chapter 9. In particular, Section 9.5 deals with the MinxEnt method, which involves a functional MinxEnt problem.

PROBLEMS

Probability Theory

1.1 Prove the following results, using the properties of the probability measure in Definition 1.1.1 (here A and B are events):

a) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

b) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

1.2 Prove the product rule (1.4) for the case of three events.

1.3 We draw three balls consecutively from a bowl containing exactly five white and five black balls, without putting them back. What is the probability that all drawn balls will be black?

1.4 Consider the random experiment where we toss a biased coin until heads comes up. Suppose the probability of heads on any one toss is p. Let X be the number of tosses required. Show that $X \sim \mathsf{G}(p)$.

1.5 In a room with many people, we ask each person his/her birthday, for example May 5. Let N be the number of people queried until we get a "duplicate" birthday.

a) Calculate $\mathbb{P}(N > n)$, $n = 0, 1, 2, \ldots$.

b) For which n do we have $\mathbb{P}(N \leqslant n) \geqslant 1/2$?

c) Use a computer to calculate $\mathbb{E}[N]$.

1.6 Let X and Y be independent standard normal random variables, and let U and V be random variables that are derived from X and Y via the linear transformation

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} \sin\alpha & -\cos\alpha \\ \cos\alpha & \sin\alpha \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} .$$

a) Derive the joint pdf of U and V.
b) Show that U and V are independent and standard normally distributed.

1.7 Let $X \sim \mathsf{Exp}(\lambda)$. Show that the *memoryless property* holds: for all $s, t \geqslant 0$,

$$\mathbb{P}(X > t + s \,|\, X > t) = \mathbb{P}(X > s) .$$

1.8 Let X_1, X_2, X_3 be independent Bernoulli random variables with success probabilities $1/2, 1/3$, and $1/4$, respectively. Give their conditional joint pdf, given that $X_1 + X_2 + X_3 = 2$.

1.9 Verify the expectations and variances in Table 1.3.

1.10 Let X and Y have joint density f given by

$$f(x, y) = c\,x\,y, \quad 0 \leqslant y \leqslant x, \quad 0 \leqslant x \leqslant 1 .$$

a) Determine the normalization constant c.
b) Determine $\mathbb{P}(X + 2\,Y \leqslant 1)$.

1.11 Let $X \sim \mathsf{Exp}(\lambda)$ and $Y \sim \mathsf{Exp}(\mu)$ be independent. Show that

a) $\min(X, Y) \sim \mathsf{Exp}(\lambda + \mu)$,
b) $\mathbb{P}(X < Y \mid \min(X, Y)) = \dfrac{\lambda}{\lambda + \mu}$.

1.12 Verify the properties of variance and covariance in Table 1.4.

1.13 Show that the correlation coefficient always lies between -1 and 1. [Hint, use the fact that the variance of $aX + Y$ is always non-negative, for any a.]

1.14 Consider Examples 1.1 and 1.2. Define X as the function that assigns the number $x_1 + \cdots + x_n$ to each outcome $\omega = (x_1, \ldots, x_n)$. The event that there are exactly k heads in n throws can be written as

$$\{\omega \in \Omega : X(\omega) = k\} .$$

If we abbreviate this to $\{X = k\}$, and further abbreviate $\mathbb{P}(\{X = k\})$ to $\mathbb{P}(X = k)$, then we obtain exactly (1.7). Verify that one can always view random variables in this way, that is, as real-valued functions on Ω, and that probabilities such as $\mathbb{P}(X \leqslant x)$ should be interpreted as $\mathbb{P}(\{\omega \in \Omega : X(\omega) \leqslant x\})$.

1.15 Show that

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j) .$$

1.16 Let Σ be the covariance matrix of a random column vector $\mathbf{X}$. Write $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the expectation vector of $\mathbf{X}$. Hence, $\Sigma = \mathbb{E}[\mathbf{Y}\mathbf{Y}^T]$. Show that Σ is positive semidefinite. That is, for any vector $\mathbf{u}$, we have $\mathbf{u}^T \Sigma \mathbf{u} \geqslant 0$.

1.17 Suppose $Y \sim \mathsf{Gamma}(n, \lambda)$. Show that for all $x \geqslant 0$

$$\mathbb{P}(Y \leqslant x) = 1 - \sum_{k=0}^{n-1} \frac{\mathrm{e}^{-\lambda\, x}(\lambda\, x)^k}{k!} \; . \tag{1.91}$$

1.18 Consider the random experiment where we draw uniformly and independently n numbers, $X_1, \ldots, X_n$, from the interval [0,1].

- **a)** Let M be the smallest of the n numbers. Express M in terms of $X_1, \ldots, X_n$.
- **b)** Determine the pdf of M.

1.19 Let $Y = \mathrm{e}^X$, where $X \sim \mathsf{N}(0, 1)$.

- **a)** Determine the pdf of Y.
- **b)** Determine the expected value of Y.

1.20 We select a point (X, Y) from the triangle $(0, 0) - (1, 0) - (1, 1)$ in such a way that X has a uniform distribution on $(0, 1)$ and the conditional distribution of Y given $X = x$ is uniform on $(0, x)$.

- **a)** Determine the joint pdf of X and Y.
- **b)** Determine the pdf of Y.
- **c)** Determine the conditional pdf of X given $Y = y$ for all $y \in (0, 1)$.
- **d)** Calculate $\mathbb{E}[X \,|\, Y = y]$ for all $y \in (0, 1)$.
- **e)** Determine the expectations of X and Y.

Poisson Processes

1.21 Let $\{N_t, t \geqslant 0\}$ be a Poisson process with rate $\lambda = 2$. Find

- **a)** $\mathbb{P}(N_2 = 1, N_3 = 4, N_5 = 5)$,
- **b)** $\mathbb{P}(N_4 = 3 \,|\, N_2 = 1, N_3 = 2)$,
- **c)** $\mathbb{E}[N_4 \,|\, N_2 = 2]$,
- **d)** $\mathbb{P}(N[2, 7] = 4, N[3, 8] = 6)$,
- **e)** $\mathbb{E}[N[4, 6] \,|\, N[1, 5] = 3]$.

1.22 Show that for any fixed $k \in \mathbb{N}$, $t > 0$ and $\lambda > 0$,

$$\lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} = \frac{(\lambda t)^k}{k!}\, \mathrm{e}^{-\lambda t} \; .$$

(Hint: write out the binomial coefficient and use the fact that $\lim_{n\to\infty}\left(1 - \frac{\lambda t}{n}\right)^n = \mathrm{e}^{-\lambda t}$.)

1.23 Consider the Bernoulli approximation in Section 1.11. Let $U_1, U_2, \ldots$ denote the times of success for the Bernoulli process X.

- **a)** Verify that the "intersuccess" times $U_1, U_2 - U_1, \ldots$ are independent and have a geometric distribution with parameter $p = \lambda h$.
- **b)** For small h and $n = \lfloor t/h \rfloor$, show that the relationship $\mathbb{P}(A_1 > t) \approx \mathbb{P}(U_1 > n)$ leads in the limit, as $n \to \infty$, to

$$\mathbb{P}(A_1 > t) = \mathrm{e}^{-\lambda t}.$$

1.24 If $\{N_t, t \geqslant 0\}$ is a Poisson process with rate λ, show that for $0 \leqslant u \leqslant t$ and $j = 0, 1, 2, \ldots, n$,

$$\mathbb{P}(N_u = j \,|\, N_t = n) = \binom{n}{j} \left(\frac{u}{t}\right)^j \left(1 - \frac{u}{t}\right)^{n-j},$$

that is, the conditional distribution of N_u given $N_t = n$ is binomial with parameters n and u/t.

Markov Processes

1.25 Determine the (discrete) pdf of each X_n, $n = 0, 1, 2, \ldots$ for the random walk in Example 1.10. Also, calculate $\mathbb{E}[X_n]$ and the variance of X_n for each n.

1.26 Let $\{X_n, n \in \mathbb{N}\}$ be a Markov chain with state space $\{0, 1, 2\}$, transition matrix

$$P = \begin{pmatrix} 0.3 & 0.1 & 0.6 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.7 & 0.2 \end{pmatrix},$$

and initial distribution $\pi = (0.2, 0.5, 0.3)$. Determine

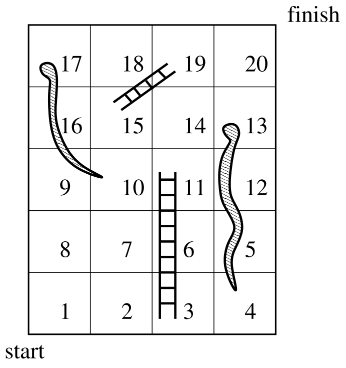
a) $\mathbb{P}(X_1 = 2)$,
b) $\mathbb{P}(X_2 = 2)$,
c) $\mathbb{P}(X_3 = 2 \,|\, X_0 = 0)$,
d) $\mathbb{P}(X_0 = 1 \,|\, X_1 = 2)$,
e) $\mathbb{P}(X_1 = 1, X_3 = 1)$.

1.27 Consider two dogs harboring a total number of m fleas. Spot initially has b fleas and Lassie has the remaining $m - b$. The fleas have agreed on the following immigration policy: at every time $n = 1, 2 \ldots$ a flea is selected at random from the total population and that flea will jump from one dog to the other. Describe the flea population on Spot as a Markov chain and find its stationary distribution.

1.28 Classify the states of the Markov chain with the following transition matrix:

$$P = \begin{pmatrix} 0.0 & 0.3 & 0.6 & 0.0 & 0.1 \\ 0.0 & 0.3 & 0.0 & 0.7 & 0.0 \\ 0.3 & 0.1 & 0.6 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 & 0.9 & 0.0 \\ 0.1 & 0.1 & 0.2 & 0.0 & 0.6 \end{pmatrix}.$$

1.29 Consider the following snakes-and-ladders game. Let N be the number of tosses required to reach the finish using a fair die. Calculate the expectation of N using a computer.



1.30 Ms. Ella Brum walks back and forth between her home and her office every day. She owns three umbrellas, which are distributed over two umbrella stands (one at home and one at work). When it is not raining, Ms. Brum walks without an umbrella. When it is raining, she takes one umbrella from the stand at the place of her departure, provided there is one available. Suppose the probability that it is raining at the time of any departure is p. Let X_n denote the number of umbrellas available at the place where Ella arrives after walk number n; $n = 1, 2, \ldots$, including the one that she possibly brings with her. Calculate the limiting probability that it rains and no umbrella is available.

1.31 A mouse is let loose in the maze of Figure 1.9. From each compartment the mouse chooses one of the adjacent compartments with equal probability, independent of the past. The mouse spends an exponentially distributed amount of time in each compartment. The mean time spent in each of the compartments 1, 3, and 4 is two seconds; the mean time spent in compartments 2, 5, and 6 is four seconds. Let $\{X_t, t \geqslant 0\}$ be the Markov jump process that describes the position of the mouse for times $t \geqslant 0$. Assume that the mouse starts in compartment 1 at time $t = 0$.

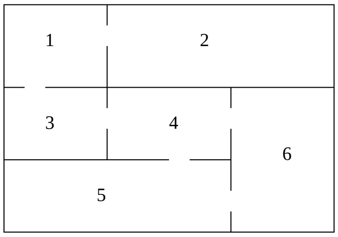


Figure 1.9 A maze.

What are the probabilities that the mouse will be found in each of the compartments $1, 2, \ldots, 6$ at some time t far away in the future?

1.32 In an $M/M/\infty$-queueing system, customers arrive according to a Poisson process with rate a. Every customer who enters is immediately served by one of an infinite number of servers; hence, there is no queue. The service times are exponentially distributed, with mean $1/b$. All service and interarrival times are independent. Let X_t be the number of customers in the system at time t. Show that the limiting distribution of X_t, as $t \to \infty$, is Poisson with parameter a/b.

Optimization

1.33 Let $\mathbf{a}$ and let $\mathbf{x}$ be n-dimensional column vectors. Show that $\nabla_{\mathbf{x}}\, \mathbf{a}^T\mathbf{x} = \mathbf{a}$.

1.34 Let A be a symmetric $n \times n$ matrix and $\mathbf{x}$ be an n-dimensional column vector. Show that $\nabla_{\mathbf{x}}\, \frac{1}{2}\, \mathbf{x}^T A\mathbf{x} = A\mathbf{x}$. What is the gradient if A is not symmetric?

1.35 Show that the optimal distribution $\mathbf{p}^*$ in Example 1.17 is given by the uniform distribution.

1.36 Derive the program (1.78).

1.37 Consider the MinxEnt program

$$\min_{\mathbf{p}} \quad \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

$$\text{subject to:} \quad \mathbf{p} \geqslant \mathbf{0}, \quad A\mathbf{p} = \mathbf{b}, \quad \sum_{i=1}^{n} p_i = 1 \ ,$$

where $\mathbf{p}$ and $\mathbf{q}$ are probability distribution vectors and A is an $m \times n$ matrix.

a) Show that the Lagrangian for this problem is of the form

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \beta, \boldsymbol{\mu}) = \mathbf{p}^T \boldsymbol{\xi}(\mathbf{p}) - \boldsymbol{\lambda}^T (A\mathbf{p} - \mathbf{b}) - \boldsymbol{\mu}^T \mathbf{p} + \beta(\mathbf{1}^T \mathbf{p} - 1) \ .$$

b) Show that $p_i = q_i \exp(-\beta - 1 + \mu_i + \sum_{j=1}^{m} \lambda_j \, a_{ji})$, for $i = 1, \ldots, n$.

c) Explain why, as a result of the KKT conditions, the optimal $\boldsymbol{\mu}^*$ must be equal to the zero vector.

d) Show that the solution to this MinxEnt program is exactly the same as for the program where the nonnegativity constraints are omitted.

Further Reading

An easy introduction to probability theory with many examples is [14], and a more detailed textbook is [9]. A classical reference is [7]. An accurate and accessible treatment of various stochastic processes is given in [4]. For convex optimization we refer to [3] and [8].

REFERENCES

1. S. Asmussen and R. Y. Rubinstein. Complexity properties of steady-state rare-events simulation in queueing models. In J. H. Dshalalow, editor, *Advances in Queueing: Theory, Methods and Open Problems*, pages 429–462, New York, 1995. CRC Press.
2. Z. I. Botev, D. P. Kroese, and T. Taimre. Generalized cross-entropy methods for rare-event simulation and optimization. *Simulation: Transactions of the Society for Modeling and Simulation International*, 2007. In press.
3. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.
4. E. Çinlar. *Introduction to Stochastic Processes*. Prentice Hall, Englewood Cliffs, NJ, 1975.
5. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
6. C. W. Curtis. *Linear Algebra: An Introductory Approach*. Springer-Verlag, New York, 1984.
7. W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, New York, 2nd edition, 1970.
8. R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, New York, 1987.
9. G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, 3rd edition, 2001.
10. J. N. Kapur and H. K. Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, New York, 1992.

11. A. I. Khinchin. *Information Theory*. Dover Publications, New York, 1957.

12. V. Kriman and R. Y. Rurbinstein. Polynomial time algorithms for estimation of rare events in queueing models. In J. Dshalalow, editor, *Frontiers in Queueing: Models and Applications in Science and Engineering*, pages 421–448, New York, 1995. CRC Press.

13. E. L. Lehmann. *Testing Statistical Hypotheses*. Springer-Verlag, New York, 1997.

14. S. M. Ross. *A First Course in Probability*. Prentice Hall, Englewood Cliffs, NJ, 7th edition, 2005.

15. R. Y. Rubinstein and B. Melamed. *Modern Simulation and Modeling*. John Wiley & Sons, New York, 1998.

APPENDIX

A.1 CHOLESKY SQUARE ROOT METHOD

Let Σ be a covariance matrix. We wish to find a matrix B such that $\Sigma = BB^T$. The *Cholesky square root method* computes a lower triangular matrix B via a set of recursive equations as follows: From (1.23) we have

$$Z_1 = b_{11}X_1 + \mu_1 \; . \tag{A.1}$$

Therefore, $\text{Var}(Z_1) = \sigma_{11} = b_{11}^2$ and $b_{11} = \sigma_{11}^{1/2}$. Proceeding with the second component of (1.23), we obtain

$$Z_2 = b_{21}X_1 + b_{22}X_2 + \mu_2 \tag{A.2}$$

and thus

$$\sigma_{22} = \text{Var}(Z_2) = \text{Var}(b_{21}X_1 + b_{22}X_2) = b_{21}^2 + b_{22}^2 \, . \tag{A.3}$$

Further, from (A.1) and (A.2),

$$\sigma_{12} = \mathbb{E}[(Z_1 - \mu_1)(Z_2 - \mu_2)] = \mathbb{E}[b_{11}X_1(b_{21}X_1 + b_{22}X_2)] = b_{11}b_{21} \; . \tag{A.4}$$

Hence, from (A.3) and (A.4) and the symmetry of Σ,

$$\begin{aligned} b_{21} &= \frac{\sigma_{12}}{b_{11}} = \frac{\sigma_{12}}{\sigma_{11}^{1/2}} && \text{(A.5)} \\ b_{22} &= \left(\sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}}\right)^{1/2} . && \text{(A.6)} \end{aligned}$$

Simulation and the Monte Carlo Method, Second Edition. By R.Y. Rubinstein and D. P. Kroese

Generally, the b_{ij} can be found from the recursive formula

$$b_{ij} = \frac{\sigma_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}}{\left(\sigma_{jj} - \sum_{k=1}^{j-1} b_{jk}^2\right)^{1/2}} \;, \tag{A.7}$$

where, by convention,

$$\sum_{k=1}^{0} b_{ik} b_{jk} = 0, \quad 1 \leqslant j \leqslant i \leqslant n \;.$$

A.2 EXACT SAMPLING FROM A CONDITIONAL BERNOULLI DISTRIBUTION

Suppose the vector $\mathbf{X} = (X_1, \ldots, X_n)$ has independent components, with $X_i \sim \mathsf{Ber}(p_i)$, $i = 1, \ldots, n$. It is not difficult to see (see Problem A.1) that the conditional distribution of $\mathbf{X}$ given $\sum_i X_i = k$ is given by

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \,\middle|\, \sum_{i=1}^{n} X_i = k\right) = \frac{\prod_{i=1}^{n} w_i^{x_i}}{c} \;, \tag{A.8}$$

where c is a normalization constant and $w_i = p_i/(1-p_i)$, $i = 1, \ldots, n$. Generating random variables from this distribution can be done via the so-called *drafting* procedure, described, for example, in [2]. The Matlab code below provides a procedure for calculating the normalization constant c and drawing from the conditional joint pdf above.

■ **EXAMPLE A.1**

Suppose $\mathbf{p} = (1/2, 1/3, 1/4, 1/5)$ and $k = 2$. Then $\mathbf{w} = (w_1, \ldots, w_4) = (1, 1/2, 1/3, 1/4)$. The first element of `Rgens(k,w)`, with `k` $= 2$ and `w` $= \mathbf{w}$ is $35/24 \approx 1.45833$. This is the normalization constant c. Thus, for example,

$$\mathbb{P}\left(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = x_1 \,\middle|\, \sum_{i=1}^{4} X_i = 2\right) = \frac{w_2\, w_4}{35/24} = \frac{3}{35} \approx 0.08571 \;.$$

To generate random vectors according to this conditional Bernoulli distribution call `condbern(p,k)`, where `k` is the number of unities (here 2) and `p` is the probability vector $\mathbf{p}$. This function returns the positions of the unities, such as (1, 2) or (2,4).

```
function sample = condbern(k,p)
% k = no of units in each sample, P = probability vector
W=zeros(1,length(p));
sample=zeros(1,k);
ind1=find(p==1);
sample(1:length(ind1))=ind1;
k=k-length(ind1);
ind=find(p<1 & p>0);
W(ind)=p(ind)./(1-p(ind));
for i=1:k
```

```
    Pr=zeros(1,length(ind));
    Rvals=Rgens(k-i+1,W(ind));
    for j=1:length(ind)
        Pr(j)=W(ind(j))*Rvals(j+1)/((k-i+1)*Rvals(1));
    end
    Pr=cumsum(Pr);
    entry=ind(min(find(Pr>rand)));
    ind=ind(find(ind~=entry));
    sample(length(ind1)+i)=entry;
end
sample=sort(sample);
return


function Rvals = Rgens(k,W)
N=length(W);
T=zeros(k,N+1);
R=zeros(k+1,N+1);
for i=1:k
    for j=1:N, T(i,1)=T(i,1)+W(j)^i; end
    for j=1:N, T(i,j+1)=T(i,1)-W(j)^i; end
end
R(1,:)=ones(1,N+1);
for j=1:k
    for l=1:N+1
        for i=1:j
            R(j+1,l)=R(j+1,l)+(-1)^(i+1)*T(i,l)*R(j-i+1,l);
        end
    end
    R(j+1,:)=R(j+1,:)/j;
end
Rvals=[R(k+1,1),R(k,2:N+1)];
return
```

A.3 EXPONENTIAL FAMILIES

Exponential families play an important role in statistics; see, for example, [1]. Let $\mathbf{X}$ be a random variable or vector (in this section, vectors will always be interpreted as *column* vectors) with pdf $f(\mathbf{x};\boldsymbol{\theta})$ (with respect to some measure), where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$ is an m-dimensional parameter vector. $\mathbf{X}$ is said to belong to an m-parameter *exponential family* if there exist real-valued functions $t_i(\mathbf{x})$ and $h(\mathbf{x}) > 0$ and a (normalizing) function $c(\boldsymbol{\theta}) > 0$ such that

$$f(\mathbf{x};\boldsymbol{\theta}) = c(\boldsymbol{\theta})\,\mathrm{e}^{\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{x})}\,h(\mathbf{x}) \;, \tag{A.9}$$

where $\mathbf{t}(\mathbf{x}) = (t_1(\mathbf{x}), \ldots, t_m(\mathbf{x}))^T$ and $\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{x})$ is the inner product $\sum_{i=1}^m \theta_i t_i(\mathbf{x})$. The representation of an exponential family is in general not unique.

Remark A.3.1 (Natural Exponential Family) The standard definition of an exponential family involves a family of densities $\{g(\mathbf{x};\mathbf{v})\}$ of the form

$$g(\mathbf{x};\mathbf{v}) = d(\mathbf{v})\,\mathrm{e}^{\boldsymbol{\theta}(\mathbf{v})\cdot\mathbf{t}(\mathbf{x})}\,h(\mathbf{x}) \;, \tag{A.10}$$

where $\boldsymbol{\theta}(\mathbf{v}) = (\theta_1(\mathbf{v}),\ldots,\theta_m(\mathbf{v}))^T$, and the $\{\theta_i\}$ are real-valued functions of the parameter $\mathbf{v}$. By *reparameterization* — by using the θ_i as parameters — we can represent (A.10) in so-called *canonical form* (A.9). In effect, $\boldsymbol{\theta}$ is the natural parameter of the exponential family. For this reason, a family of the form (A.9) is called a *natural exponential family*.

Table A.1 displays the functions $c(\boldsymbol{\theta}),\ t_k(x)$, and $h(x)$ for several commonly used distributions (a dash means that the corresponding value is not used).

Table A.1 The functions $c(\boldsymbol{\theta}),\ t_k(x)$ and $h(x)$ for commonly used distributions.

Distr.	$t_1(x),\ t_2(x)$	$c(\boldsymbol{\theta})$	$\theta_1,\ \theta_2$	$h(x)$
$\mathsf{Gamma}(\alpha,\lambda)$	$x,\ \ln x$	$\dfrac{(-\theta_1)^{\theta_2+1}}{\Gamma(\theta_2+1)}$	$-\lambda,\ \alpha-1$	1
$\mathsf{N}(\mu,\sigma^2)$	$x,\ x^2$	$\dfrac{\mathrm{e}^{\theta_1^2/(4\theta_2)}}{\sqrt{-\pi/\theta_2}}$	$\dfrac{\mu}{\sigma^2},\ -\dfrac{1}{2\sigma^2}$	1
$\mathsf{Weib}(\alpha,\lambda)$	$x^\alpha,\ \ln x$	$-\theta_1(\theta_2+1)$	$-\lambda^\alpha,\ \alpha-1$	1
$\mathsf{Bin}(n,p)$	$x,\ \ -$	$(1+\mathrm{e}^{\theta_1})^{-n}$	$\ln\left(\dfrac{p}{1-p}\right),\ \ -$	$\dbinom{n}{x}$
$\mathsf{Poi}(\lambda)$	$x,\ \ -$	$\mathrm{e}^{-\mathrm{e}^{\theta_1}}$	$\ln\lambda,\ \ -$	$\dfrac{1}{x!}$
$\mathsf{G}(p)$	$x-1,\ \ -$	$1-\mathrm{e}^{\theta_1}$	$\ln(1-p),\ \ -$	1

As an important instance of a natural exponential family, consider the univariate, single-parameter ($m=1$) case with $t(x)=x$. Thus, we have a family of densities $\{f(x;\theta),\theta\in\Theta\subset\mathbb{R}\}$ given by

$$f(x;\theta) = c(\theta)\,\mathrm{e}^{\theta x}\,h(x) \;. \tag{A.11}$$

If $h(x)$ is a pdf, then $c^{-1}(\theta)$ is the corresponding *moment generating function*:

$$c^{-1}(\theta) = \int \mathrm{e}^{\theta x}\,h(x)\,dx \;.$$

It is sometimes convenient to introduce instead the logarithm of the moment generating function:

$$\zeta(\theta) = \ln \int \mathrm{e}^{\theta x} h(x)\,dx \;,$$

which is called the *cumulant function*. We can now write (A.11) in the following convenient form:

$$f(x;\theta) = \mathrm{e}^{\theta x - \zeta(\theta)}\,h(x) \;. \tag{A.12}$$

■ **EXAMPLE A.2**

If we take h as the density of the $\mathsf{N}(0, \sigma^2)$-distribution, $\theta = \lambda/\sigma^2$ and $\zeta(\theta) = \sigma^2\,\theta^2/2$, then the family $\{f(\cdot;\theta), \theta \in \mathbb{R}\}$ is the family of $\mathsf{N}(\lambda, \sigma^2)$ densities, where σ^2 is fixed and $\lambda \in \mathbb{R}$.

Similarly, if we take h as the density of the $\mathsf{Gamma}(a, 1)$-distribution, and let $\theta = 1 - \lambda$ and $\zeta(\theta) = -a\ln(1-\theta) = -a\ln\lambda$, we obtain the class of $\mathsf{Gamma}(a, \lambda)$ distributions, with a fixed and $\lambda > 0$. Note that in this case $\Theta = (-\infty, 1)$.

Starting from any pdf f_0, we can easily generate a natural exponential family of the form (A.12) in the following way: Let Θ be the largest interval for which the cumulant function ζ of f_0 exists. This includes $\theta = 0$, since f_0 is a pdf. Now define

$$f(x;\theta) = \mathrm{e}^{\theta x - \zeta(\theta)}\, f_0(x)\ . \tag{A.13}$$

Then $\{f(\cdot;\theta), \theta \in \Theta\}$ is a natural exponential family. We say that the family is obtained from f_0 by an *exponential twist* or *exponential change of measure* (ECM) with a *twisting* or *tilting* parameter θ.

Remark A.3.2 (Reparameterization) It may be useful to reparameterize a natural exponential family of the form (A.12) into the form (A.10). Let $X \sim f(\cdot;\theta)$. It is not difficult to see that

$$\mathbb{E}_\theta[X] = \zeta'(\theta) \quad \text{and} \quad \mathrm{Var}_\theta(X) = \zeta''(\theta)\ . \tag{A.14}$$

$\zeta'(\theta)$ is increasing in θ, since its derivative, $\zeta''(\theta) = \mathrm{Var}_\theta(X)$, is always greater than 0. Thus, we can reparameterize the family using the mean $v = \mathbb{E}_\theta[X]$. In particular, to the above natural exponential family there corresponds a family $\{g(\cdot;v)\}$ such that for each pair (θ, v) satisfying $\zeta'(\theta) = v$ we have $g(x;v) = f(x;\theta)$.

■ **EXAMPLE A.3**

Consider the second case in Example A.2. Note that we constructed in fact a natural exponential family $\{f(\cdot;\theta), \theta \in (-\infty, 1)\}$ by exponentially twisting the $\mathsf{Gamma}(\alpha, 1)$ distribution, with density $f_0(x) = x^{\alpha-1}\mathrm{e}^{-x}/\Gamma(a)$. We have $\zeta'(\theta) = \alpha/(1-\theta) = v$. This leads to the reparameterized density

$$g(x;v) \;=\; \exp\left(\theta x + \alpha\ln(1-\theta)\right) f_0(x) \;=\; \frac{\exp\left(-\frac{\alpha}{v}\,x\right)\left(\frac{\alpha}{v}\right)^{\alpha}\, x^{\alpha-1}}{\Gamma(\alpha)}\ ,$$

corresponding to the $\mathsf{Gamma}(\alpha, \alpha v^{-1})$ distribution, $v > 0$.

CE Updating Formulas for Exponential Families

We now obtain an *analytic* formula for a general one-parameter exponential family. Let $X \sim f(x;u)$ for some nominal reference parameter u. For simplicity, assume that $\mathbb{E}_u[H(X)] > 0$ and that X is nonconstant. Let $f(x;u)$ be a member of a one-parameter exponential family $\{f(x;v)\}$. Suppose the parameterization $\eta = \psi(v)$ puts the family in canonical form. That is,

$$f(x;v) = g(x;\eta) = \mathrm{e}^{\eta x - \zeta(\eta)}h(x)\ .$$

Moreover, let us assume that v corresponds to the expectation of X. This can always be established by reparameterization; see Remark A.3.2. Note that, in particular, $v = \zeta'(\eta)$. Let $\theta = \psi(u)$ correspond to the nominal reference parameter. Since $\max_v \mathbb{E}_u[H(X) \ln f(X; v)] = \max_\eta \mathbb{E}_\theta[H(X) \ln g(X; \eta)]$, we may obtain the optimal solution v^* to (5.61) by finding, as in (5.62), the solution η^* to

$$\mathbb{E}_\theta \left[H(X) \frac{d}{d\eta} \ln g(X; \eta) \right] = 0$$

and putting $v^* = \psi^{-1}(\eta^*)$. Since $(\ln g(X; \eta))' = x - \zeta'(\eta)$, and $\zeta'(\eta) = v$, we see that v^* is given by the solution of $\mathbb{E}_u[H(X)\,(-v + X)] = 0$. Hence v^* is given by

$$v^* = \frac{\mathbb{E}_u[H(X)\, X]}{\mathbb{E}_u\,[H(X)]} = \frac{\mathbb{E}_w[H(X)\, W(X; u, w)\, X]}{\mathbb{E}_w[H(X)\, W(X; u, w)]} \tag{A.15}$$

for any reference parameter w. It is not difficult to check that v^* is indeed a unique global maximum of $D(v) = \mathbb{E}_u[H(X) \ln f(X; v)]$. The corresponding estimator $\widehat{v}$ of v^* in (A.15) is

$$\widehat{v} = \frac{\sum_{i=1}^N H(X_i)\, W(X_i; u, w)\, X_i}{\sum_{i=1}^N\ H(X_i) W(X_i; u, w)}\ , \tag{A.16}$$

where $X_1, \ldots, X_N$ is a random sample from the density $f(\cdot; w)$.

A similar explicit formula can be found for the case where $\mathbf{X} = (X_1, \ldots, X_n)$ is a vector of *independent* random variables such that each component X_j belongs to a one-parameter exponential family parameterized by the mean; that is, the density of each X_j is given by

$$f_j(x; u_j) = \mathrm{e}^{x\theta(u_j) - \zeta(\theta(u_j))}\, h_j(x),$$

where $\mathbf{u} = (u_1, \ldots, u_n)$ is the nominal reference parameter. It is easy to see that problem (5.64) under the independence assumption becomes “separable,” that is, it reduces to n subproblems of the form above. Thus, we find that the optimal reference parameter vector $\mathbf{v}^* = (v_1^*, \ldots, v_n^*)$ is given as

$$v_j^* = \frac{\mathbb{E}_{\mathbf{u}}[H(\mathbf{X})\, X_j]}{\mathbb{E}_{\mathbf{u}}[H(\mathbf{X})]} = \frac{\mathbb{E}_{\mathbf{w}}[H(\mathbf{X})\, W(\mathbf{X}; \mathbf{u}, \mathbf{w})\, X_j]}{\mathbb{E}_{\mathbf{w}}[(\mathbf{X})\, W(\mathbf{X}; \mathbf{u}, \mathbf{w})]}\ . \tag{A.17}$$

Moreover, we can estimate the j-th component of $\mathbf{v}^*$ as

$$\widehat{v}_j = \frac{\sum_{i=1}^N H(\mathbf{X}_i)\, W(\mathbf{X}_i; \mathbf{u}, \mathbf{w}) X_{ij}}{\sum_{i=1}^N H(\mathbf{X}_i)\, W(\mathbf{X}_i; \mathbf{u}, \mathbf{w})}\ , \tag{A.18}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ is a random sample from the density $f(\cdot; \mathbf{w})$ and X_{ij} is the j-th component of $\mathbf{X}_i$.

A.4 SENSITIVITY ANALYSIS

The crucial issue in choosing a good importance sampling density $f(\mathbf{x}; \mathbf{v})$ to estimate $\nabla^k \ell(\mathbf{u})$ via (7.16) is to ensure that the corresponding estimators have low variance. We consider this issue for the cases $k = 0$ and $k = 1$. For $k = 0$ this means minimizing the variance of $\widehat{\ell}(\mathbf{u}; \mathbf{v})$ with respect to $\mathbf{v}$, which is equivalent to solving the minimization program

$$\min_{\mathbf{v}} \mathcal{L}^0(\mathbf{v};\mathbf{u}) = \min_{\mathbf{v}} \mathbb{E}_{\mathbf{v}}[H^2(\mathbf{X})\, W^2(\mathbf{X};\mathbf{u},\mathbf{v})] \; . \tag{A.19}$$

For the case $k = 1$, note that $\nabla\widehat{\ell}(\mathbf{u};\mathbf{v})$ is a vector rather than a scalar. To obtain a good reference vector $\mathbf{v}$, we now minimize the *trace of the associated covariance matrix*, which is equivalent to minimizing

$$\min_{\mathbf{v}} \mathcal{L}^1(\mathbf{v};\mathbf{u}) = \min_{\mathbf{v}} \mathbb{E}_{\mathbf{v}}\left[H^2(\mathbf{X})\, W^2(\mathbf{X};\mathbf{u},\mathbf{v})\ \mathrm{tr}\left(\mathcal{S}(\mathbf{u};\mathbf{x})\mathcal{S}(\mathbf{u};\mathbf{x})^T\right)\right], \tag{A.20}$$

where tr denotes the trace. For exponential families both optimization programs are *convex*, as demonstrated in the next proposition. To conform with our earlier notation for exponential families in Section A.3, we use $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ instead of $\mathbf{u}$ and $\mathbf{v}$, respectively.

A.4.1 Convexity Results

Proposition A.4.1 *Let $\mathbf{X}$ be a random vector from an m-parameter exponential family of the form (A.9). Then $\mathcal{L}^k(\boldsymbol{\eta};\boldsymbol{\theta}),\ k = 0, 1,$ defined in (A.19) and (A.20), are convex functions of $\boldsymbol{\eta}$.*

Proof: Consider first the case $k = 0$. One has (see (7.23))

$$\mathcal{L}^0(\boldsymbol{\eta};\boldsymbol{\theta}) = c(\boldsymbol{\theta})^2 \int \frac{H^2(\mathbf{x})}{c(\boldsymbol{\eta})}\, \mathrm{e}^{(2\boldsymbol{\theta}-\boldsymbol{\eta})\cdot\mathbf{t}(\mathbf{x})} h(\mathbf{x})\mathrm{d}\mathbf{x} \; , \tag{A.21}$$

where

$$c(\boldsymbol{\eta})^{-1} = \int \mathrm{e}^{\boldsymbol{\eta}\cdot\mathbf{t}(\mathbf{z})} h(\mathbf{z})\,\mathrm{d}\mathbf{z} \; .$$

Substituting the above into (A.21) yields

$$\mathcal{L}^0(\boldsymbol{\eta};\boldsymbol{\theta}) = c(\boldsymbol{\theta})^2 \int\int H^2(\mathbf{x})\, \mathrm{e}^{2\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{x}) + \boldsymbol{\eta}\cdot(\mathbf{t}(\mathbf{z}) - \mathbf{t}(\mathbf{x}))} h(\mathbf{x})\, h(\mathbf{z})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z} \; . \tag{A.22}$$

Now, for any linear function, $a(\boldsymbol{\eta})$ of $\boldsymbol{\eta}$, the function $\mathrm{e}^{a(\boldsymbol{\eta})}$ is convex. Since $H^2(\mathbf{x})$ is nonnegative, it follows that for any fixed $\boldsymbol{\theta}$, $\mathbf{x}$, and $\mathbf{z}$, the function under the integral sign in (A.22) is convex in $\boldsymbol{\eta}$. This implies the convexity of $\mathcal{L}^0(\boldsymbol{\eta};\boldsymbol{\theta})$.

The case $k = 1$ follows in exactly the same way, noting that the trace $\mathrm{tr}\left(\mathcal{S}(\boldsymbol{\theta};\mathbf{x})\mathcal{S}(\boldsymbol{\theta};\mathbf{x})^T\right)$ is a nonnegative function for $\mathbf{x}$ for any $\boldsymbol{\theta}$. □

Remark A.4.1 Proposition A.4.1 can be extended to the case where

$$\ell(\mathbf{u}) = \varphi(\ell_1(\mathbf{u}), \ldots, \ell_k(\mathbf{u}))$$

and

$$\ell_i(\mathbf{u}) = \mathbb{E}_{\mathbf{u}}[H_i(\mathbf{X})] = \mathbb{E}_{\mathbf{v}}[H_i(\mathbf{X})W(\mathbf{X};\mathbf{u};\mathbf{v})] = \mathbb{E}_{\mathbf{v}}[H_i W], \quad i = 1, \ldots, k \; .$$

Here the $\{H_i(\mathbf{X})\}$ are sample functions associated with the same random vector $\mathbf{X}$ and $\varphi(\cdot)$ is a real-valued differentiable function. We prove its validity for the case $k = 2$. In this case, the estimators of $\ell(\mathbf{u})$ can be written as

$$\widehat{\ell}(\mathbf{u};\mathbf{v}) = \varphi(\widehat{\ell}_1(\mathbf{u};\mathbf{v}), \widehat{\ell}_2(\mathbf{u};\mathbf{v})) \; ,$$

where $\widehat{\ell}_1(\mathbf{u};\mathbf{v})$ and $\widehat{\ell}_2(\mathbf{u};\mathbf{v})$ are the usual importance sampling estimators of $\ell_1(\mathbf{u})$ and $\ell_2(\mathbf{u})$, respectively. By virtue of the delta method (see Problem 7.11), $N^{1/2}(\widehat{\ell}(\mathbf{u};\mathbf{v})-\ell(\mathbf{u}))$ is asymptotically normal, with mean 0 and variance

$$\begin{aligned}\sigma^2(\mathbf{v};\mathbf{u}) &= a^2\,\mathrm{Var}_{\mathbf{v}}(H_1W)+b^2\,\mathrm{Var}_{\mathbf{v}}(H_2W)+2\,a\,b\,\mathrm{Cov}_{\mathbf{v}}\,(H_1W,H_2W)\\ &= \mathbb{E}_{\mathbf{v}}\left[(aH_1+bH_2)^2\,W^2\right]+R(\mathbf{u})\;. \qquad \text{(A.23)}\end{aligned}$$

Here $R(\mathbf{u})$ consists of the remaining terms that are independent of $\mathbf{v}$, $a=\partial\varphi(x_1,x_2)/\partial x_1$ and $b=\partial\varphi(x_1,x_2)/\partial x_2$ at $(x_1,x_2)=(\ell_1(\mathbf{u}),\ell_2(\mathbf{u}))$. For example, for $\varphi(x_1,x_2)=x_1/x_2$, one gets $a=1/\ell_2(\mathbf{u})$ and $b=-\ell_1(\mathbf{u})/\ell_2(\mathbf{u})^2$.

The convexity of $\sigma^2(\mathbf{v};\mathbf{u})$ in $\mathbf{v}$ now follows similarly to the proof of Proposition A.4.1.

A.4.2 Monotonicity Results

Consider optimizing the functions $\mathcal{L}^k(\mathbf{v};\mathbf{u})$, $k=0,1$ in (A.19) and (A.20) with respect to $\mathbf{v}$. Let $\mathbf{v}^*(k)$ be the optimal solutions for $k=0,1$. The following proposition states that the optimal reference parameter always leads to a "fatter" tail for $f(\mathbf{x};\mathbf{v}^*)$ than that of the original pdf $f(\mathbf{x};\mathbf{u})$. This important phenomenon is the driving force for all of our beautiful results in this book, as well as for preventing the degeneracy of our importance sampling estimates. For simplicity, the result is given for the gamma distribution only. Similar results can be established with respect to some other parameters of the exponential family and for the CE approach.

Proposition A.4.2 *Let $X\sim\mathsf{Gamma}(\alpha,u)$. Suppose that $H^2(x)$ is a monotonically increasing function on the interval $[0,\infty)$. Then*

$$v^*(k)<u,\ k=0,1\;. \qquad \text{(A.24)}$$

Proof: The proof will be given for $k=0$ only. The proof for $k=1$, using the trace operator, is similar. To simplify the notation, we write $\mathcal{L}(v)$ for $\mathcal{L}^0(v;u)$.

Since $\mathcal{L}(v)$ is convex, it suffices to prove that its derivative with respect to v is positive at $v=u$. To this end, represent $\mathcal{L}(v)$ as

$$\mathcal{L}(v)=c\int_0^\infty v^{-\alpha}H^2(x)\,x^{\alpha-1}\mathrm{e}^{-(2u-v)x}\,\mathrm{d}x\;,$$

where the constant $c=u^{2\alpha}\Gamma(\alpha)^{-1}$ is independent of v. Differentiating $\mathcal{L}(v)$ above with respect to v at $v=u$, one has

$$\mathcal{L}'(v)|_{v=u}=\mathcal{L}'(u)=c\int_0^\infty(x-\alpha u^{-1})\,u^{-\alpha}H^2(x)\,x^{\alpha-1}\,\mathrm{e}^{-ux}\,\mathrm{d}x\;.$$

Integrating by parts yields

$$\begin{aligned}\mathcal{L}'(u) &= \lim_{R\to\infty}-c\,u^{-\alpha-1}R^\alpha\mathrm{e}^{-uR}\,H^2(R)+c\,u^{-\alpha-1}\int_0^\infty x^\alpha\,\mathrm{e}^{-ux}\,\mathrm{d}H^2(x)\\ &= c\,u^{-\alpha-1}\int_0^\infty x^\alpha\,\mathrm{e}^{-ux}\,\mathrm{d}H^2(x)\;, \qquad \text{(A.25)}\end{aligned}$$

provided $H^2(R)R^\alpha\exp(-uR)$ tends to 0 as $R\to\infty$. Finally, since $H^2(x)$ is monotonically increasing in x, we conclude that the integral (A.25) is positive, and consequently, $\mathcal{L}'(u)>0$. This fact, and the convexity of $\mathcal{L}(v)$, imply that $v^*(0)<u$. □

Proposition A.4.2 can be extended to the multidimensional gamma distribution, as well as to some other exponential family distributions. For details see [5].

A.5 A SIMPLE CE ALGORITHM FOR OPTIMIZING THE PEAKS FUNCTION

The following Matlab code provides a simple implementation of a CE algorithm to solve the `peaks` function; see Example 8.12 on page 268.

```
n = 2;                              % dimension
mu = [-3,-3]; sigma = 3*ones(1,n); N = 100; eps = 1E-5; rho = 0.1;

while max(sigma) > eps
   X = randn(N,n)*diag(sigma)+ mu(ones(N,1),:);
   SX= S(X);                        %Compute the performance
   sortSX = sortrows([X, SX],n+1);
   Elite = sortSX((1-rho)*N:N,1:n); % elite samples
   mu = mean(Elite,1);              % take sample mean row-wise
   sigma = std(Elite,1);            % take sample st.dev. row-wise
   [S(mu),mu,max(sigma)]            % output the result
end

function out = S(X)
out =  3*(1-X(:,1)).^2.*exp(-(X(:,1).^2) - (X(:,2)+1).^2) ...
   - 10*(X(:,1)/5 - X(:,1).^3 - X(:,2).^5).*exp(-X(:,1).^2-X(:,2).^2) ...
   - 1/3*exp(-(X(:,1)+1).^2 - X(:,2).^2);
end
```

A.6 DISCRETE-TIME KALMAN FILTER

Consider the hidden Markov model

$$\begin{aligned} X_t &= A\,X_{t-1} + \varepsilon_{1t} \\ Y_t &= B\,X_t + \varepsilon_{2t}, \qquad t = 1, 2, \ldots, \end{aligned} \tag{A.26}$$

where A and B are matrices (B does not have to be a square matrix). We adopt the notation of Section 5.7.1. The initial state X_0 is assumed to be $\mathsf{N}(\mu_0, \Sigma_0)$ distributed. The objective is to obtain the filtering pdf $f(x_t \,|\, \mathbf{y}_{1:t})$ and the *predictive* pdf $f(x_t \,|\, \mathbf{y}_{1:t-1})$. Observe that the joint pdf of $\mathbf{X}_{1:t}$ and $\mathbf{Y}_{1:t}$ must be Gaussian, since these random vectors are linear transformations of independent standard Gaussian random variables. It follows that $f(x_t \,|\, \mathbf{y}_{1:t}) \sim \mathsf{N}(\mu_t, \Sigma_t)$ for some mean vector μ_t and covariance matrix Σ_t. Similarly, $f(x_t \,|\, \mathbf{y}_{1:t-1}) \sim \mathsf{N}(\widetilde{\mu}_t, \widetilde{\Sigma}_t)$ for some mean vector $\widetilde{\mu}_t$ and covariance matrix $\widetilde{\Sigma}_t$. We wish to compute μ_t, $\widetilde{\mu}_t$, Σ_t and $\widetilde{\Sigma}_t$ recursively. The argument goes as follows: by assumption, $(X_{t-1} \,|\, \mathbf{y}_{1:t-1}) \sim \mathsf{N}(\mu_{t-1}, \Sigma_{t-1})$. Combining this with the fact that $X_t = A\,X_{t-1} + \varepsilon_{1t}$ yields

$$(X_t \,|\, \mathbf{y}_{1:t-1}) \sim \mathsf{N}(A\,\mu_{t-1},\; A\Sigma_{t-1}A^T + C_1)\;.$$

In other words,

$$\begin{aligned} \widetilde{\mu}_t &= A\,\mu_{t-1}, \\ \widetilde{\Sigma}_t &= A\Sigma_{t-1}A^T + C_1\;. \end{aligned} \tag{A.27}$$

Next, we determine the joint pdf of X_t and Y_t given $\mathbf{Y}_{1:t-1} = \mathbf{y}_{1:t-1}$. Decomposing $\widetilde{\Sigma}_t$ and C_2 as $\widetilde{\Sigma}_t = RR^T$ and $C_2 = QQ^T$, respectively (e.g., via the Cholesky square root method), we can write (see (1.23))

$$\left(\begin{matrix} X_t \\ Y_t \end{matrix} \,\middle|\, \mathbf{y}_{1:t-1} \right) = \begin{pmatrix} \widetilde{\mu}_t \\ B\widetilde{\mu}_t \end{pmatrix} + \begin{pmatrix} R & 0 \\ BR & Q \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix} ,$$

where, conditional on $Y_{t-1} = \mathbf{y}_{1:t-1}$, U and V are independent standard normal random vectors. The corresponding covariance matrix is

$$\begin{pmatrix} R & 0 \\ BR & Q \end{pmatrix} \begin{pmatrix} R^T & R^T B^T \\ 0 & Q^T \end{pmatrix} = \begin{pmatrix} RR^T & RR^T B^T \\ BRR^T & BRR^T B^T + QQ^T \end{pmatrix} ,$$

so that we have

$$\left(\begin{matrix} X_t \\ Y_t \end{matrix} \,\middle|\, \mathbf{y}_{1:t-1} \right) \sim \mathsf{N}\left(\begin{pmatrix} \widetilde{\mu}_t \\ B\widetilde{\mu}_t \end{pmatrix}, \; \begin{pmatrix} \widetilde{\Sigma}_t & \widetilde{\Sigma}_t B^T \\ B\widetilde{\Sigma}_t & B\widetilde{\Sigma}_t B^T + C_2 \end{pmatrix} \right) \tag{A.28}$$

(note that $\widetilde{\Sigma}_t$ is symmetric).

The result (A.28) enables us to find the conditional pdf $f(x_t \,|\, \mathbf{y}_t)$ with the aid of the following general result (see Problem A.2 below for a proof): If

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \; \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \right) ,$$

then

$$(X \,|\, Y = y) \sim \mathsf{N}\left(m_1 + S_{12} S_{22}^{-1} (y - m_2), \;\; S_{11} - S_{12} S_{22}^{-1} S_{12}^T \right) \; . \tag{A.29}$$

Because $f(x_t \,|\, \mathbf{y}_{1:t}) = f(x_t \,|\, \mathbf{y}_{1:t-1}, y_t)$, an immediate consequence of (A.28) and (A.29) is

$$\begin{aligned} \mu_t &= \widetilde{\mu}_t + \widetilde{\Sigma}_t B^T (B\widetilde{\Sigma}_t B^T + C_2)^{-1} (y_t - B\widetilde{\mu}_t) \; , \\ \Sigma_t &= \widetilde{\Sigma}_t - \widetilde{\Sigma}_t B^T (B\widetilde{\Sigma}_t B^T + C_2)^{-1} B\widetilde{\Sigma}_t \; . \end{aligned} \tag{A.30}$$

Updating formulas (A.27) and (A.30) form the (discrete-time) *Kalman filter*. Starting with some known μ_0 and Σ_0, one determines $\widetilde{\mu}_1$ and $\widetilde{\Sigma}_1$, then μ_1 and Σ_1, and so on. Notice that $\tilde{\Sigma}_t$ and Σ_t do not depend on the observations $y_1, y_2, \ldots$ and can therefore be determined *off-line*. The Kalman filter discussed above can be extended in many ways, for example by including control variables and time-varying parameter matrices. The nonlinear filtering case is often dealt with by linearizing the state and observation equations via a Taylor expansion. This leads to an approximative method called the *extended Kalman filter*.

A.7 BERNOULLI DISRUPTION PROBLEM

As an example of a finite-state hidden Markov model, we consider the following *Bernoulli disruption problem*. In Example 6.8 a similar type of “changepoint” problem is discussed in relation to the Gibbs sampler. However, the crucial difference is that in the present case the detection of the changepoint can be done *sequentially*.

Let $Y_1, Y_2, \ldots$ be Bernoulli random variables and let T be a geometrically distributed random variable with parameter r. Conditional upon T the $\{Y_i\}$ are mutually independent, and $Y_1, Y_2, \ldots, Y_{T-1}$ all have a success probability a, whereas $Y_T, Y_{T+1}, \ldots$ all have a success probability b. Thus, T is the change or disruption point. Suppose that T cannot be observed, but only $\{Y_t\}$. We wish to decide if the disruption has occurred based on the outcome $\mathbf{y}_{1:t} = (y_1, \ldots, y_t)$ of $\mathbf{Y}_{1:t} = (Y_1, \ldots, Y_t)$. An example of the observations is depicted in Figure A.1, where the dark lines indicate the times of successes ($Y_i = 1$).

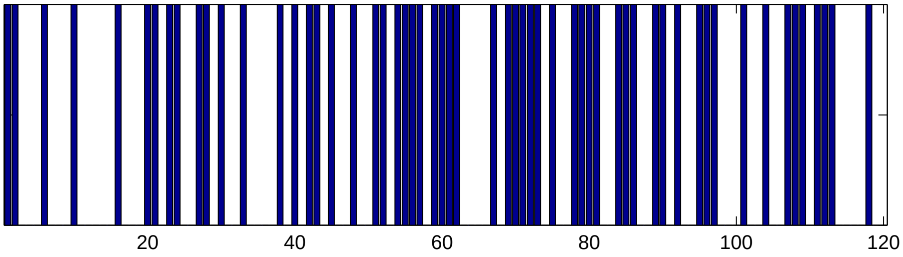


Figure A.1 The observations for the disruption problem.

The situation can be described via the HMM illustrated in Figure A.2. Namely, let $\{X_t, t = 0, 1, 2, \ldots\}$ be a Markov chain with state space $\{0, 1\}$, transition matrix

$$P = \begin{pmatrix} 1-r & r \\ 0 & 1 \end{pmatrix} ,$$

and initial state $X_0 = 0$. Then the objective is to find $\mathbb{P}(T \leqslant t \,|\, \mathbf{Y}_{1:t} = \mathbf{y}_{1:t}) = \mathbb{P}(X_t = 1 \,|\, \mathbf{Y}_t = \mathbf{y}_{1:t})$.

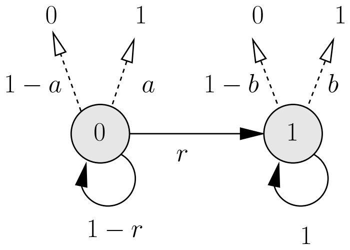


Figure A.2 The HMM diagram for the disruption problem.

This can be done efficiently by introducing

$$\alpha_t(j) = \mathbb{P}(X_t = j,\ \mathbf{Y}_{1:t} = \mathbf{y}_{1:t}) \ .$$

By conditioning on X_{t-1} we have

$$\begin{aligned} \alpha_t(j) &= \sum_i \mathbb{P}(X_t = j, X_{t-1} = i, \mathbf{Y}_{1:t} = \mathbf{y}_{1:t}) \\ &= \sum_i \mathbb{P}(X_t = j, Y_t = y_t \,|\, X_{t-1} = i, \mathbf{Y}_{1:t-1} = \mathbf{y}_{1:t-1})\, \alpha_{t-1}(i) \\ &= \sum_i \mathbb{P}(X_t = j, Y_t = y_t \,|\, X_{t-1} = i)\, \alpha_{t-1}(i). \\ &= \sum_i \mathbb{P}(Y_t = y_t \,|\, X_t = j)\, \mathbb{P}(X_t = j \,|\, X_{t-1} = i)\, \alpha_{t-1}(i) \ . \end{aligned}$$

In particular, we find the recurrence relation

$$\alpha_t(0) = a_{0\,y_t}\,(1-r)\,\alpha_{t-1}(0) \quad \text{and} \quad \alpha_t(1) = a_{1\,y_t}\{r\,\alpha_{t-1}(0) + \alpha_{t-1}(1)\}\ ,$$

with $a_{ij} = \mathbb{P}(Y = j \,|\, X = i)$, $i, j \in \{0, 1\}$ (thus, $a_{00} = 1 - a$, $a_{01} = a$, $a_{10} = 1 - b$, $a_{11} = b$), and initial values

$$\alpha_1(0) = a^{y_1}(1-a)^{1-y_1}(1-r) \quad \text{and} \quad \alpha_1(1) = b^{y_1}(1-b)^{1-y_1}r\ .$$

In Figure A.3 a plot is given of the probability $\mathbb{P}(X_t = 1 \,|\, \mathbf{Y}_{1:t} = \mathbf{y}_{1:t}) = \alpha_t(1)/(\alpha_t(1) + \alpha_t(2))$, as a function of t, for a test case with $a = 0.4$, $b = 0.6$, and $r = 0.01$. In this particular case $T = 49$. We see a dramatic change in the graph after the disruption takes effect.

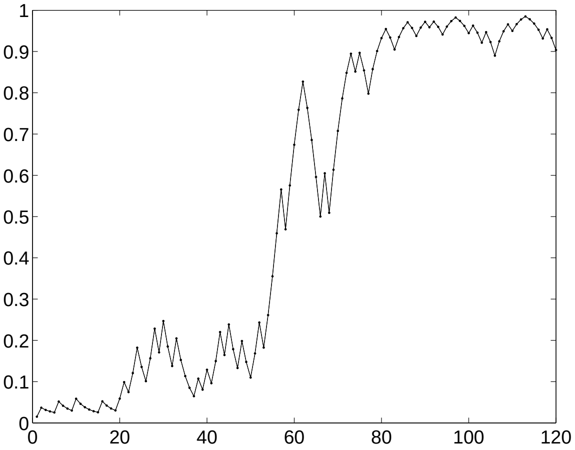


Figure A.3 The probability $\mathbb{P}(X_t = 1 \,|\, \mathbf{Y}_{1:t} = \mathbf{y}_{1:t})$ as a function of t.

A.8 COMPLEXITY OF STOCHASTIC PROGRAMMING PROBLEMS

Consider the following optimization problem:

$$\ell^* = \min_{\mathbf{u} \in \mathscr{U}} \ell(\mathbf{u}) = \min_{\mathbf{u} \in \mathscr{U}} \mathbb{E}_f[H(\mathbf{X}; \mathbf{u})]\ , \tag{A.31}$$

where it is assumed that $\mathbf{X}$ is a random vector with known pdf f having support $\mathscr{X} \subset \mathbb{R}^n$, and $H(\mathbf{X}; \mathbf{u})$ is the sample function depending on $\mathbf{X}$ and the decision vector $\mathbf{u} \in \mathbb{R}^m$.

As an example, consider a two-stage stochastic programming problem with recourse, which is an optimization problem that is divided into two stages. At the first stage, one has to make a decision on the basis of some available information. At the second stage, after a realization of the uncertain data becomes known, an optimal second-stage decision is made. Such a stochastic programming problem can be written in the form (A.31), with $H(\mathbf{X}; \mathbf{u})$ being the optimal value of the second-stage problem.

We now discuss the issue of how difficult it is to solve a stochastic program of type (A.31). We should expect that this problem is at least as difficult as minimizing $\ell(\mathbf{u})$, $\mathbf{u} \in \mathscr{U}$ in the case where $\ell(\mathbf{u})$ is given *explicitly*, say by a closed-form analytic expression or, more generally, by an "oracle" capable of computing the values and the derivatives of

$\ell(\mathbf{u})$ at every given point. As far as problems of minimization of $\ell(\mathbf{u})$, $\mathbf{u} \in \mathscr{U}$, with an explicitly given objective are concerned, the solvable case is known: this is the convex programming case. That is, $\mathscr{U}$ is a closed convex set and $\ell : \mathscr{U} \to \mathbb{R}$ is a convex function. It is known that generic convex programming problems satisfying mild computability and boundedness assumptions can be solved in polynomial time. In contrast to this, typical nonconvex problems turn out to be NP-hard.

We should also stress that a claim that "such and such problem is difficult" relates to a *generic* problem and does *not* imply that the problem has no solvable particular cases. When speaking about conditions under which the stochastic program (A.31) is efficiently solvable, it makes sense to assume that $\mathscr{U}$ is a closed convex set and $\ell(\cdot)$ is convex on $\mathscr{U}$. We gain from a technical viewpoint (and do not lose much from a practical viewpoint) by assuming $\mathscr{U}$ to be bounded. These assumptions, plus mild technical conditions, would be sufficient to make (A.31) easy (manageable) if $\ell(\mathbf{u})$ were given explicitly. However, in stochastic programming, it makes no sense to assume that we can compute efficiently the expectation in (A.31), thus arriving at an explicit representation of $\ell(\mathbf{u})$. If this were the case, there would be no necessity to treat (A.31) as a stochastic program.

We argue now that stochastic programming problems of the form (A.31) can be solved reasonably efficiently by using Monte Carlo sampling techniques, provided that the probability distribution of the random data is not "too bad" and certain general conditions are met. In this respect, we should explain what we mean by "solving" stochastic programming problems. Let us consider, for example, two-stage linear stochastic programming problems with recourse. Such problems can be written in the form (A.31) with

$$\mathscr{U} = \{\mathbf{u} : A\mathbf{u} = \mathbf{b},\ \mathbf{u} \geqslant \mathbf{0}\} \ \text{ and } \ H(\mathbf{X};\mathbf{u}) = \langle \boldsymbol{c}, \mathbf{u}\rangle + Q(\mathbf{X};\mathbf{u})\ ,$$

where $\langle \mathbf{c}, \mathbf{u}\rangle$ is the cost of the first-stage decision and $Q(\mathbf{X};\mathbf{u})$ is the optimal value of the second-stage problem:

$$\min_{\mathbf{y}\geqslant \mathbf{0}} \langle \mathbf{q}, \mathbf{y}\rangle \ \text{ subject to } \ \mathbf{Tu} + \mathbf{Wy} \geqslant \mathbf{h}\ . \tag{A.32}$$

Here, $\langle \cdot, \cdot \rangle$ denotes the inner product. $\mathbf{X}$ is a vector whose elements are composed from elements of vectors $\mathbf{q}$ and $\mathbf{h}$ and matrices $\mathbf{T}$ and $\mathbf{W}$, which are assumed to be random.

If we assume that the random data vector $\mathbf{X} = (\mathbf{q}, \mathbf{W}, \mathbf{T}, \mathbf{h})$ takes K different values (called *scenarios*) $\{\mathbf{X}_k, k = 1, \ldots, K\}$, with respective probabilities $\{p_k, k = 1, \ldots, K\}$, then the obtained two-stage problem can be written as one large linear programming problem:

$$\begin{array}{cl} \displaystyle\min_{\mathbf{u},\mathbf{y}_1,\ldots,\mathbf{y}_K} & \langle \mathbf{c}, \mathbf{u}\rangle + \sum_{k=1}^{K} p_k \, \langle \mathbf{q}_k, \mathbf{y}_k\rangle \\ \text{s.t} & \mathbf{Au} = \mathbf{b},\ \mathbf{T}_k\mathbf{u} + \mathbf{W}_k\mathbf{y}_k \geqslant \mathbf{h}_k,\ k = 1, \ldots, K, \\ & \mathbf{u} \geqslant \mathbf{0},\ \mathbf{y}_k \geqslant \mathbf{0},\ k = 1, \ldots, K. \end{array} \tag{A.33}$$

If the number of scenarios K is not too large, then the above linear programming problem (A.33) can be solved accurately in a reasonable period of time. However, even a crude discretization of the probability distribution of $\mathbf{X}$ typically results in an exponential growth of the number of scenarios with the increase of the dimension of $\mathbf{X}$. Suppose, for example, that the components of the random vector $\mathbf{X}$ are mutually independently distributed, each having a small number r of possible realizations. Then the size of the corresponding input data grows linearly in n (and r), while the number of scenarios $K = r^n$ grows exponentially.

We would like to stress that from a practical point of view, it does not make sense to try to solve a stochastic programming problem with high precision. A numerical error resulting from an inaccurate estimation of the involved probability distributions, modeling errors, and so on, can be far bigger than the optimization error. We argue now that two-stage stochastic problems can be solved efficiently with reasonable accuracy, provided that the following conditions are met:

(a) The feasible set $\mathscr{U}$ is fixed (deterministic).

(b) For all $\mathbf{u} \in \mathscr{U}$ and $\mathbf{X} \in \mathscr{X}$, the objective function $H(\mathbf{X};\mathbf{u})$ is real-valued.

(c) The considered stochastic programming problem can be solved efficiently (by a deterministic algorithm) if the number of scenarios is not too large.

When applied to two-stage stochastic programming, the above conditions (a) and (b) mean that the recourse is relatively complete and the second-stage problem is bounded from below. Note that it is said that the recourse is *relatively complete*, if for every $\mathbf{u} \in \mathscr{U}$ and every possible realization of random data, the second-stage problem is feasible. The above condition (c) certainly holds in the case of two-stage *linear* stochastic programming with recourse.

In order to proceed, let us consider the following Monte Carlo sampling approach. Suppose that we can generate an iid random sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ from $f(\mathbf{x})$, and we can estimate the expected value function $\ell(\mathbf{u})$ by the sample average

$$\widehat{\ell}(\mathbf{u}) = \frac{1}{N} \sum_{j=1}^{N} H(\mathbf{X}_j;\mathbf{u}) \; . \tag{A.34}$$

Note that $\widehat{\ell}$ depends on the sample size N and on the generated sample, and in that sense is random. Consequently, we approximate the true problem (A.31) by the following approximated one:

$$\min_{\mathbf{u} \in \mathscr{U}} \widehat{\ell}(\mathbf{u}) \; . \tag{A.35}$$

We refer to (A.35) as the *stochastic counterpart* or *sample average approximation* problem. The optimal value $\widehat{\ell^*}$ and the set $\widehat{\mathscr{U}^*}$ of optimal solutions of the stochastic counterpart problem (A.35) provide estimates of their true counterparts, ℓ^* and $\mathscr{U}^*$, of problem (A.31). It should be noted that once the sample is generated, $\widehat{\ell}(\mathbf{u})$ becomes a deterministic function and problem (A.35) becomes a stochastic programming problem with N scenarios $\mathbf{X}_1, \dots, \mathbf{X}_N$ taken with equal probabilities $1/N$. It also should be mentioned that the stochastic counterpart method is *not* an algorithm. One still has to solve the obtained problem (A.35) by employing an appropriate (deterministic) algorithm.

By the law of large numbers (see Theorem 1.10.1) $\widehat{\ell}(\mathbf{u})$ converges (point-wise in $\mathscr{U}$) with probability 1 to $\ell(\mathbf{u})$ as N tends to infinity. Therefore, it is reasonable to expect for $\widehat{\ell^*}$ and $\widehat{\mathscr{U}^*}$ to converge to their counterparts of the true problem (A.31) with probability 1 as N tends to infinity. And indeed, such convergence can be proved under mild regularity conditions. However, for a fixed $\mathbf{u} \in \mathscr{U}$, convergence of $\widehat{\ell}(\mathbf{u})$ to $\ell(\mathbf{u})$ is notoriously slow. By the central limit theorem (see Theorem 1.10.2) it is of order $\mathcal{O}(N^{-1/2})$. The rate of convergence can be improved, sometimes significantly, by variance reduction methods. However, using Monte Carlo techniques, one cannot evaluate the expected value $\ell(\mathbf{u})$ very accurately.

The following analysis is based on the exponential bounds of the *large deviations* theory. Denote by $\mathscr{U}^{\varepsilon}$ and $\widehat{\mathscr{U}^{\varepsilon}}$ the sets of ε-optimal solutions of the true and stochastic counterpart problems, respectively, that is, $\bar{\mathbf{u}} \in \mathscr{U}^{\varepsilon}$ iff $\bar{\mathbf{u}} \in \mathscr{U}$ and $\ell(\bar{\mathbf{u}}) \leqslant \inf_{\mathbf{u} \in \mathscr{U}} \ell(\mathbf{u}) + \varepsilon$. Note that for $\varepsilon = 0$ the set $\mathscr{U}^0$ coincides with the set of the optimal solutions of the true problem.

Choose accuracy constants $\varepsilon > 0$ and $0 \leqslant \delta < \varepsilon$ and the confidence (significance) level $\alpha \in (0,1)$. Suppose for the moment that the set $\mathscr{U}$ is finite, although its cardinality $|\mathscr{U}|$ can be very large. Then, using Cramér's large deviations theorem, it can be shown [4] that there exists a constant $\eta(\varepsilon,\delta)$ such that

$$N \geqslant \frac{1}{\eta(\varepsilon,\delta)} \ln\left(\frac{|\mathscr{U}|}{\alpha}\right) \tag{A.36}$$

guarantees that the probability of the event $\{\widehat{\mathscr{U}}^\delta \subset \mathscr{U}^\varepsilon\}$ is at least $1-\alpha$. That is, for any N bigger than the right-hand side of (A.36), we are guaranteed that any δ-optimal solution of the corresponding stochastic counterpart problem provides an ε-optimal solution of the true problem with probability at least $1-\alpha$. In other words, solving the stochastic counterpart problem with accuracy δ guarantees solving the true problem with accuracy ε with probability at least $1-\alpha$.

The number $\eta(\varepsilon,\delta)$ in the estimate (A.36) is defined as follows. Consider a mapping $\pi : \mathscr{U} \setminus \mathscr{U}^\varepsilon \to \mathscr{U}$ such that $\ell(\pi(\mathbf{u})) \leqslant \ell(\mathbf{u}) - \varepsilon$ for all $\mathbf{u} \in \mathscr{U} \setminus \mathscr{U}^\varepsilon$. Such mappings do exist, although not uniquely. For example, any mapping $\pi : \mathscr{U} \setminus \mathscr{U}^\varepsilon \to \mathscr{U}^0$ satisfies this condition. The choice of such a mapping gives a certain flexibility to the corresponding estimate of the sample size. For $\mathbf{u} \in \mathscr{U}$, consider the random variable

$$Y_\mathbf{u} = H(\mathbf{X};\pi(\mathbf{u})) - H(\mathbf{X};\mathbf{u}) \ ,$$

its moment generating function $M_\mathbf{u}(t) = \mathbb{E}\left[\mathrm{e}^{tY_\mathbf{u}}\right]$, and the large deviations *rate function*

$$I_\mathbf{u}(z) = \sup_{t\in\mathbb{R}} \left\{tz - \ln M_\mathbf{u}(t)\right\} \ .$$

Note that $I_\mathbf{u}(\cdot)$ is the conjugate of the function $\ln M_\mathbf{u}(\cdot)$ in the sense of convex analysis. Note also that, by construction of mapping $\pi(\mathbf{u})$, the inequality

$$\mu_\mathbf{u} = \mathbb{E}\left[Y_\mathbf{u}\right] = \ell(\pi(\mathbf{u})) - \ell(\mathbf{u}) \leqslant -\varepsilon \tag{A.37}$$

holds for all $\mathbf{u} \in \mathscr{U} \setminus \mathscr{U}^\varepsilon$. Finally, we define

$$\eta(\varepsilon,\delta) = \min_{\mathbf{u}\in\mathscr{U}\setminus\mathscr{U}^\varepsilon} I_\mathbf{u}(-\delta) \ . \tag{A.38}$$

Because of (A.37) and since $\delta < \varepsilon$, the number $I_\mathbf{u}(-\delta)$ is positive, provided that the probability distribution of $Y_\mathbf{u}$ is not too bad. Specifically, if we assume that the moment generating function $M_\mathbf{u}(t)$, of $Y_\mathbf{u}$, is finite-valued for all t in a neighborhood of 0, then the random variable $Y_\mathbf{u}$ has finite moments and $I_\mathbf{u}(\mu_\mathbf{u}) = I'(\mu_\mathbf{u}) = 0$, and $I''(\mu_\mathbf{u}) = 1/\sigma_\mathbf{u}^2$ where $\sigma_\mathbf{u}^2 = \mathrm{Var}\left[Y_\mathbf{u}\right]$. Consequently, $I_\mathbf{u}(-\delta)$ can be approximated by using the second-order Taylor expansion, as follows:

$$I_\mathbf{u}(-\delta) \approx \frac{(-\delta-\mu_\mathbf{u})^2}{2\sigma_\mathbf{u}^2} \geqslant \frac{(\varepsilon-\delta)^2}{2\sigma_\mathbf{u}^2} \ .$$

This suggests that one can expect the constant $\eta(\varepsilon,\delta)$ to be of order $(\varepsilon-\delta)^2$. And indeed, this can be ensured by various conditions. Consider the following ones.

(A1) *There exists a constant $\sigma > 0$ such that for any $\mathbf{u} \in \mathscr{U} \setminus \mathscr{U}^\varepsilon$, the moment generating function $M_\mathbf{u}^*(t)$ of the random variable $Y_\mathbf{u} - \mathbb{E}\left[Y_\mathbf{u}\right]$ satisfies*

$$M_\mathbf{u}^*(t) \leqslant \exp\left(\sigma^2 t^2/2\right), \quad \forall\, t \in \mathbb{R} \ . \tag{A.39}$$

Note that the random variable $Y_{\mathbf{u}} - \mathbb{E}\left[Y_{\mathbf{u}}\right]$ has zero mean. Moreover, if it has a normal distribution, with variance $\sigma_{\mathbf{u}}^2$, then its moment generating function is equal to the right-hand side of (A.39). Condition (A.39) means that the tail probabilities $\mathbb{P}\big(|H(\mathbf{X};\pi(\mathbf{u})) - H(\mathbf{X};\mathbf{u})| > t\big)$ are bounded from above by $\mathcal{O}(1)\exp\left(-t^2/(2\sigma_{\mathbf{u}}^2)\right)$. Note that by $\mathcal{O}(1)$ we denote generic absolute constants. This condition certainly holds if the distribution of the considered random variable has a bounded support. Condition (A.39) implies that $M_{\mathbf{u}}(t) \leqslant \exp(\mu_{\mathbf{u}} t + \sigma^2 t^2/2)$. It follows that

$$I_{\mathbf{u}}(z) \geqslant \sup_{t\in\mathbb{R}} \left\{tz - \mu_{\mathbf{u}} t - \sigma^2 t^2/2\right\} = \frac{(z-\mu_{\mathbf{u}})^2}{2\sigma^2}\,, \tag{A.40}$$

and hence, for any $\varepsilon > 0$ and $\delta \in [0,\varepsilon)$,

$$\eta(\varepsilon,\delta) \geqslant \frac{(-\delta-\mu_{\mathbf{u}})^2}{2\sigma^2} \geqslant \frac{(\varepsilon-\delta)^2}{2\sigma^2}\,. \tag{A.41}$$

It follows that, under assumption (A1), the estimate (A.36) can be written as

$$N \geqslant \frac{2\sigma^2}{(\varepsilon-\delta)^2} \ln\left(\frac{|\mathscr{U}|}{\alpha}\right)\,. \tag{A.42}$$

Remark A.8.1 Condition (A.39) can be replaced by a more general one,

$$M_{\mathbf{u}}^*(t) \leqslant \exp(\psi(t)), \ \ \forall\, t \in \mathbb{R}\,, \tag{A.43}$$

where $\psi(t)$ is a convex even function with $\psi(0) = 0$. Then $\ln M_{\mathbf{u}}(t) \leqslant \mu_{\mathbf{u}} t + \psi(t)$ and hence $I_{\mathbf{u}}(z) \geqslant \psi^*(z - \mu_{\mathbf{u}})$, where ψ^* is the conjugate of the function ψ. It follows then that

$$\eta(\varepsilon,\delta) \geqslant \psi^*(-\delta-\mu_{\mathbf{u}}) \geqslant \psi^*(\varepsilon-\delta)\,. \tag{A.44}$$

For example, instead of assuming that the bound (A.39) holds for all $t \in \mathbb{R}$, we can assume that it holds for all t in a finite interval $[-a,a]$, where $a > 0$ is a given constant. That is, we can take $\psi(t) = \sigma^2 t/2$ if $|t| \leqslant a$ and $\psi(t) = +\infty$ otherwise. In that case, $\psi^*(z) = z^2/(2\sigma^2)$ for $|z| \leqslant a\sigma^2$ and $\psi^*(z) = a|z| - a^2\sigma^2/2$ for $|z| > a\sigma^2$.

A key feature of the estimate (A.42) is that the required sample size N depends *logarithmically* both on the size of the feasible set $\mathscr{U}$ and on the significance level α. The constant σ, postulated in assumption (A1), measures, in some sense, the variability of the considered problem. For, say, $\delta = \varepsilon/2$, the right-hand side of the estimate (A.42) is proportional to $(\sigma/\varepsilon)^2$. For Monte Carlo methods, such dependence on σ and ε seems to be unavoidable. In order to see this, consider a simple case when the feasible set $\mathscr{U}$ consists of just two elements: $\mathscr{U} = \{u_1, u_2\}$, with $\ell(u_2) - \ell(u_1) > \varepsilon > 0$. By solving the corresponding stochastic counterpart problem, we can ensure that u_1 is the ε-optimal solution if $\widehat{\ell}(u_2) - \widehat{\ell}(u_1) > 0$. If the random variable $H(X;u_2) - H(X;u_1)$ has a normal distribution with mean $\mu = \ell(u_2) - \ell(u_1)$ and variance σ^2, then $\widehat{\ell}(u_2) - \widehat{\ell}(u_1) \sim \mathsf{N}(\mu, \sigma^2/N)$ and the probability of the event $\{\widehat{\ell}(u_2) - \widehat{\ell}(u_1) > 0\}$ (that is, of the correct decision) is $\Phi(\mu\sqrt{N}/\sigma)$, where Φ is the cdf of $\mathsf{N}(0,1)$. We have that $\Phi(\varepsilon\sqrt{N}/\sigma) < \Phi(\mu\sqrt{N}/\sigma)$, and in order to make the probability of the incorrect decision less than α, we have to take the sample size $N > z_{1-\alpha}^2\,\sigma^2/\varepsilon^2$, where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution. Even if $H(X;u_2) - H(X;u_1)$ is not normally distributed, the sample size of order σ^2/ε^2 could be justified asymptotically, say by applying the central limit theorem.

Let us also consider the following simplified variant of the estimate (A.42). Suppose that:

(A2) *There is a positive constant C such that the random variable $\mathbf{Y}_{\mathbf{u}}$ is bounded in absolute value by a constant C for all $\mathbf{u} \in \mathscr{U} \setminus \mathscr{U}^{\varepsilon}$.*

Under assumption (A2) we have that for any $\varepsilon > 0$ and $\delta \in [0, \varepsilon]$:

$$I_{\mathbf{u}}(-\delta) \geqslant \mathcal{O}(1) \frac{(\varepsilon - \delta)^2}{C^2}, \quad \text{for all } \mathbf{u} \in \mathscr{U} \setminus \mathscr{U}^{\varepsilon} , \tag{A.45}$$

and hence $\eta(\varepsilon, \delta) \geqslant \mathcal{O}(1)(\varepsilon - \delta)^2 / C^2$. Consequently, the bound (A.36) for the sample size that is required to solve the true problem with accuracy $\varepsilon > 0$ and probability at least $1 - \alpha$, by solving the stochastic counterpart problem with accuracy $\delta = \varepsilon/2$, takes the form

$$N \geqslant \mathcal{O}(1) \left(\frac{C}{\varepsilon} \right)^2 \ln \left(\frac{|\mathscr{U}|}{\alpha} \right) . \tag{A.46}$$

Now let $\mathscr{U}$ be a bounded, not necessarily a finite, subset of $\mathbb{R}^m$ of diameter

$$D = \sup\nolimits_{\mathbf{u}', \mathbf{u} \in \mathscr{U}} \|\mathbf{u}' - \mathbf{u}\| .$$

Then for $\tau > 0$, we can construct a set $\mathscr{U}_\tau \subset \mathscr{U}$ such that for any $\mathbf{u} \in \mathscr{U}$ there is $\mathbf{u}' \in \mathscr{U}_\tau$ satisfying $\|\mathbf{u} - \mathbf{u}'\| \leqslant \tau$, and $|\mathscr{U}_\tau| = (\mathcal{O}(1) D / \tau)^m$.

Suppose next that the following condition holds:

(A3) *There exists a constant $\sigma > 0$ such that for any $\mathbf{u}', \mathbf{u} \in \mathscr{U}$ the moment generating function $M_{\mathbf{u}', \mathbf{u}}(t)$, of random variable $H(\mathbf{X}; \mathbf{u}') - H(\mathbf{X}; \mathbf{u}) - \mathbb{E}[H(\mathbf{X}; \mathbf{u}') - H(\mathbf{X}; \mathbf{u})]$, satisfies*

$$M_{\mathbf{u}', \mathbf{u}}(t) \leqslant \exp\left(\sigma^2 t^2 / 2\right) , \quad \forall t \in \mathbb{R} . \tag{A.47}$$

The above assumption (A3) is slightly stronger than assumption (A1), that is, assumption (A3) follows from (A1) by taking $\mathbf{u}' = \pi(\mathbf{u})$. Then by (A.42), for $\varepsilon' > \delta$, we can estimate the corresponding sample size required to solve the reduced optimization problem, obtained by replacing $\mathscr{U}$ with $\mathscr{U}_\tau$, as

$$N \geqslant \frac{2\sigma^2}{(\varepsilon' - \delta)^2} \left[n \left(\ln D - \ln \tau \right) + \ln \left(\mathcal{O}(1) / \alpha \right) \right] . \tag{A.48}$$

Suppose further that there exists a function $\kappa : \mathscr{X} \to \mathbb{R}_+$ and $\varrho > 0$ such that

$$|H(\mathbf{X}; \mathbf{u}') - H(\mathbf{X}; \mathbf{u})| \leqslant \kappa(\mathbf{X}) \, \|\mathbf{u}' - \mathbf{u}\|^{\varrho} \tag{A.49}$$

holds for all $\mathbf{u}', \mathbf{u} \in \mathscr{U}$ and all $\mathbf{X} \in \mathscr{X}$. It follows by (A.49) that

$$|\widehat{\ell}(\mathbf{u}') - \widehat{\ell}(\mathbf{u})| \leqslant N^{-1} \sum_{j=1}^{N} |H(\mathbf{X}_j; \mathbf{u}') - H(\mathbf{X}_j; \mathbf{u})| \leqslant \widehat{\kappa} \, \|\mathbf{u}' - \mathbf{u}\|^{\varrho} , \tag{A.50}$$

where $\widehat{\kappa} = N^{-1} \sum_{j=1}^{N} \kappa(\mathbf{X}_j)$.

Let us further assume the following:

(A4) *The moment generating function $M_\kappa(t) = \mathbb{E}\left[e^{t\kappa(\mathbf{X})}\right]$ of $\kappa(\mathbf{X})$ is finite-valued for all t in a neighborhood of 0.*

It follows then that the expectation $L = \mathbb{E}[\kappa(\mathbf{X})]$ is finite, and moreover, by Cramér's large deviations theorem that for any $L' > L$ there exists a positive constant $\beta = \beta(L')$ such that

$$\mathbb{P}\left(\widehat{\kappa} > L'\right) \leqslant \mathrm{e}^{-N\beta} \; . \tag{A.51}$$

Let $\widehat{\mathbf{u}}$ be a δ-optimal solution of the stochastic counterpart problem and let $\tilde{\mathbf{u}} \in \mathscr{U}_\tau$ be a point such that $\|\widehat{\mathbf{u}} - \tilde{\mathbf{u}}\| \leqslant \tau$. Let us take $N \geqslant \beta^{-1}\ln(2/\alpha)$, so that by (A.51) we have

$$\mathbb{P}\left(\widehat{\kappa} > L'\right) \leqslant \alpha/2 \; . \tag{A.52}$$

Then with probability at least $1-\alpha/2$, the point $\tilde{\mathbf{u}}$ is a $(\delta + L'\tau^\varrho)$-optimal solution of the reduced stochastic counterpart problem. Setting

$$\tau = [(\varepsilon - \delta)/(2L')]^{1/\varrho} \; ,$$

we find that with probability at least $1-\alpha/2$, the point $\tilde{\mathbf{u}}$ is an ε'-optimal solution of the reduced stochastic counterpart problem with $\varepsilon' = (\varepsilon+\delta)/2$. Moreover, by taking a sample size satisfying (A.48), we find that $\tilde{\mathbf{u}}$ is an ε'-optimal solution of the reduced expected-value problem with probability at least $1-\alpha/2$. It follows that $\widehat{\mathbf{u}}$ is an ε''-optimal solution of the stochastic counterpart problem (A.31) with probability at least $1-\alpha$ and $\varepsilon'' = \varepsilon' + L\tau^\varrho \leqslant \varepsilon$. We obtain the following estimate

$$N \geqslant \frac{4\sigma^2}{(\varepsilon-\delta)^2}\left[n\left(\ln D + \varrho^{-1}\ln\frac{2L'}{\varepsilon-\delta}\right) + \ln\left(\frac{\mathcal{O}(1)}{\alpha}\right)\right] \vee \left[\beta^{-1}\ln\left(2/\alpha\right)\right] \tag{A.53}$$

for the sample size, where $\vee$ denotes the maximum.

The above result is quite general and does not involve the convexity assumption. The estimate (A.53) of the sample size contains various constants and is too conservative for practical applications. However, it can be used as an estimate of the complexity of two-stage stochastic programming problems. In typical applications (e.g., in the convex case) the constant $\varrho = 1$, in which case condition (A.49) means that $H(\mathbf{X};\cdot)$ is Lipschitz continuous on $\mathscr{U}$ with constant $\kappa(\mathbf{X})$. Note that there are also some applications where ϱ could be less than 1. We obtain the following basic result.

Theorem A.8.1 *Suppose that assumptions* (A3) *and* (A4) *hold and $\mathscr{U}$ has a finite diameter D. Then for $\varepsilon > 0$, $0 \leqslant \delta < \varepsilon$ and sample size N satisfying* (A.53), *we are guaranteed that any δ-optimal solution of the stochastic counterpart problem is an ε-optimal solution of the true problem with probability at least $1-\alpha$.*

In particular, if we assume that $\varrho = 1$ and $\kappa(\mathbf{X}) = L$ for all $\mathbf{X} \in \mathscr{X}$, that is, $H(\mathbf{X};\cdot)$ is Lipschitz continuous on $\mathscr{U}$ with constant L independent of $\mathbf{X} \in \mathscr{X}$, then we can take $\sigma = \mathcal{O}(1)DL$ and remove the term $\beta^{-1}\ln(2/\alpha)$ on the right-hand side of (A.53). Further, by taking $\delta = \varepsilon/2$ we find in that case the following estimate of the sample size (compare with estimate (A.46)):

$$N \geqslant \mathcal{O}(1)\left(\frac{DL}{\varepsilon}\right)^2\left[n\ln\left(\frac{DL}{\varepsilon}\right) + \ln\left(\frac{\mathcal{O}(1)}{\alpha}\right)\right] \; . \tag{A.54}$$

We can write the following simplified version of Theorem A.8.1.

Theorem A.8.2 *Suppose that $\mathscr{U}$ has a finite diameter D and condition* (A.49) *holds with $\varrho = 1$ and $\kappa(\mathbf{X}) = L$ for all $\mathbf{X} \in \mathscr{X}$. Then with sample size N satisfying* (A.54)*, we are guaranteed that every $(\varepsilon/2)$-optimal solution of the stochastic counterpart problem is an ε-optimal solution of the true problem with probability at least $1 - \alpha$.*

The above estimates of the required sample size suggest complexity of order σ^2/ε^2 with respect to the desirable accuracy. This is in sharp contrast to deterministic (convex) optimization, where complexity usually is bounded in terms of $\ln(\varepsilon^{-1})$. In view of the above discussion, it should not be surprising that (even linear) two-stage stochastic programs usually cannot be solved with high accuracy. On the other hand, the estimates (A.53) and (A.54) depend *linearly* on the dimension n of the first-stage decision vector. They also depend linearly on $\ln(\alpha^{-1})$. This means that by increasing confidence, say, from 99% to 99.99%, we need to increase the sample size by a factor of $\ln 100 \approx 4.6$ at most. This also suggests that by using Monte Carlo sampling techniques, one can solve a two-stage stochastic program with reasonable accuracy, say with relative accuracy of 1% or 2%, in a reasonable time, provided that (a) its variability is not too large, (b) it has relatively complete recourse, and (c) the corresponding stochastic counterpart problem can be solved efficiently. And indeed, this was verified in numerical experiments with two-stage problems having a linear second-stage recourse. Of course, the estimate (A.53) of the sample size is far too conservative for the actual calculations. For practical applications, there are techniques that allow us to estimate the error of the feasible solution $\bar{\mathbf{u}}$ for a given sample size N; see, for example, [6].

The above estimates of the sample size are quite general. For convex problems, these bounds can be tightened in some cases. That is, suppose that the problem is convex, that is, the set $\mathscr{U}$ is convex and functions $H(\mathbf{X};\cdot)$ are convex for all $\mathbf{X} \in \mathscr{X}$. Suppose further that $\kappa(\mathbf{X}) \equiv L$, the set $\mathscr{U}^0$, of optimal solutions of the true problem, is nonempty and bounded and for some $r \geqslant 1$, $c > 0$ and $a > 0$, the following growth condition holds:

$$\ell(\mathbf{u}) \geqslant \ell^* + c\,[\mathrm{dist}(\mathbf{u}, \mathscr{U}^0)]^r, \quad \forall\, \mathbf{u} \in \mathscr{U}^a\ , \tag{A.55}$$

where $a > 0$ and $\mathscr{U}^a = \{\mathbf{u} \in \mathscr{U} : \ell(\mathbf{u}) \leqslant \ell^* + a\}$ is the set of a-optimal solutions of the true problem. Then for any $\varepsilon \in (0, a)$ and $\delta \in [0, \varepsilon/2)$ we have the following estimate of the required sample size:

$$N \geqslant \left(\frac{\mathcal{O}(1)L}{c^{1/r}\varepsilon^{(r-1)/r}}\right)^2 \left[n \ln\left(\frac{\mathcal{O}(1)LD_a^*}{\varepsilon}\right) + \ln\left(\frac{1}{\alpha}\right)\right]\ , \tag{A.56}$$

where D_a^* is the diameter of $\mathscr{U}^a$. Note that if $\mathscr{U}^0 = \{\mathbf{u}^*\}$ is a singleton, then it follows from (A.55) that $D_a^* \leqslant 2(a/c)^{1/r}$.

In particular, if $r = 1$ and $\mathscr{U}^0 = \{\mathbf{u}^*\}$ is a singleton, that is, the solution $\mathbf{u}^*$ is *sharp*, then D_a^* can be bounded by $4c^{-1}\varepsilon$ and hence we obtain the following estimate:

$$N \geqslant \mathcal{O}(1)c^{-2}L^2 \left[n \ln\left(\mathcal{O}(1)c^{-1}L\right) + \ln\left(\alpha^{-1}\right)\right]\ , \tag{A.57}$$

which does not depend on ε. That is, in that case, convergence to the exact optimal solution $\mathbf{u}^*$ happens with probability 1 in finite time.

For $r = 2$, condition (A.55) is called the *second-order* or *quadratic* growth condition. Under the quadratic growth condition, the first term on the right-hand side of (A.56) becomes of order $c^{-1}L^2\varepsilon^{-1}$.

PROBLEMS

A.1 Prove (A.8).

A.2 Let X and Y be Gaussian random vectors, with joint distribution given by

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathsf{N}\left(\underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}}_{\mu}, \underbrace{\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}}_{\Sigma} \right) .$$

a) Defining $S = \Sigma_{12}\Sigma_{22}^{-1}$, show that

$$\begin{pmatrix} I & -S \\ 0 & I \end{pmatrix} \Sigma \begin{pmatrix} I & 0 \\ -S^T & I \end{pmatrix} = \begin{pmatrix} \Sigma_{11} - S\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} .$$

b) Using the above result, show that for any vectors u and v

$$(u^T \ v^T)\Sigma^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = (u^T - v^T S^T)\widetilde{\Sigma}^{-1}(u - Sv) + v^T \Sigma_{22}^{-1} v ,$$

where $\widetilde{\Sigma} = (\Sigma_{11} - S\Sigma_{21})$.

c) The joint pdf of X and Y is given by

$$f(x, y) = c_1 \ \exp\left[-\frac{1}{2}(x^T - \mu_1^T \quad y^T - \mu_2^T)\, \Sigma^{-1} \begin{pmatrix} x - \mu_1 \\ y - \mu_2 \end{pmatrix} \right]$$

for some constant c_1. Using b), show that the conditional pdf $f(x \,|\, y)$ is of the form

$$f(x \,|\, y) = c_2(y) \ \exp\left[-\frac{1}{2}(x^T - \widetilde{\mu}^T)\, \widetilde{\Sigma}^{-1}\, (x - \widetilde{\mu}) \right],$$

with $\tilde{\mu} = \mu_1 + S(y - \mu_2)$, and where $c_2(y)$ is some function of y (need not be specified). This proves that

$$(X \,|\, Y = y) \sim \mathsf{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y - \mu_2), \ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \right) .$$

Further Reading

More details on exponential families and their role in statistics may be found in [1]. An accessible account of hidden Markov models is [3].

The estimate (A.42) of the sample size, for finite feasible set $\mathscr{U}$, was obtained in [4]. For a general discussion of such estimates and extensions to the general case, see [6]. For a discussion of the complexity of *multistage* stochastic programming problems, see, for example, [8]. Finite time convergence in cases of sharp optimal solutions is discussed in [7].

REFERENCES

1. G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2001.
2. S. X. Chen and J. S. Liu. Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7:875–892, 1997.
3. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2001.
4. A. J. Kleywegt, A. Shapiro, and T. Homem de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12:479–502, 2001.
5. R. Y. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the Score Function Method*. John Wiley & Sons, New York, 1993.
6. A. Shapiro. Monte Carlo sampling methods. In A. Ruszczyński and A. Shapiro, editors, *Handbook in Operations Research and Management Science*, volume 10. Elsevier, Amsterdam, 2003.
7. A. Shapiro and T. Homem de Mello. On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. *SIAM Journal on Optimization*, 11(1):70–86, 2001.
8. A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In V. Jeyakumar and A.M. Rubinov, editors, *Continuous Optimization: Current Trends and Application*. Springer-Verlag, New York, 2005.