# 1
# Review of probability concepts

In this chapter, we give a brief summary of the main concepts and results on probability and statistics which will be needed in the rest of the book. Readers who are familiar with the theory of probability might not need to read this chapter in detail, but we urge them to check that effectively this is the case.

## 1.1
## Random Variables

In most occasions, we cannot predict with absolute certainty the outcome of an experiment (otherwise it might not be necessary to perform the experiment). We understand here the word ''experiment'' in a broad sense. We can count the number of electrons emitted by a $\beta$-radioactive substance in a given time interval, determine the time at which a projectile hits its target or a bus reaches the station, measure an electron's spin, toss a coin and look at the appearing side, or have a look through the window to observe whether it rains or not. We will denote by $E$ the set of possible results of the experiment. For the $\beta$-radioactive substance, $E = \{0, 1, 2, \ldots\}$ is the set of natural numbers $\mathbb{N}$; the hitting times of the projectile or the arrival times of the bus (in some units) both belong to the set of real numbers $E = \mathbb{R}$; the possible outcomes of a measure of an electron's spin are $E = \{-\hbar/2, \hbar/2\}$; when tossing a dice, the possible results are $E = \{\textbf{heads}, \textbf{tails}\}$; and, finally, for the rain observation the set of results is $E = \{\textbf{yes}, \textbf{no}\}$. In all these cases, we have no way (or no effective way) of knowing *a priori* which one of the possible outcomes will be observed. Hence, we abandon the deterministic point of view and adopt a ''probabilistic'' description in which subsets of results (which are called ''events'') are assigned a number measuring their likelihood of appearance. The ''theory of probability'' is the branch of mathematics that allows us to perform such an assignation in a logically consistent way and compatible with our intuition of how this likelihood of events should behave.

It is useful for the theory to consider that the set of results contains only numbers. In this way, we can use the rules of calculus (add, multiply, differentiate, integrate, etc.). If the results themselves are numbers (case of counting the number of electrons, determine the time of impact of the projectile, etc.), this requires no

special consideration. In other cases (to observe whether it rains or not), we need to label each result with a number. This assignation is arbitrary, but usually it responds to some logics of the problem under consideration. For instance, when tossing a coin, it might be that we win €1 every time heads show up and we lose €1 when tails appear. The "natural" identification is +1 for heads and −1 for tails. This assignation of a number to the result of an experiment is called a "random variable." Random variables are, hence, an application of the set of results to the set of real numbers. This application maps each result of the experiment $\xi \in E$ into one, and only one, number. The application need not be one-to-one, but it can be many-to-one. For instance, if the experiment is to extract cards from a shuffled deck, we could assign +2 to all hearts cards, +1 to spades, and 0 to diamonds and clubs. It is customary to denote random variables by using a "hat" on top of its name, say $\hat{\mathbf{x}}$, or $\hat{\mathbf{y}}$, or whatever name we choose for it. If we choose the name $\hat{\mathbf{x}}$, the number associated with the result $\xi$ is $\hat{\mathbf{x}}(\xi) \in \mathbb{R}$. This distinction between the result of the experiment and the real number associated with it is important from the mathematical point of view, but in many cases of interest they both coincide because the result of the experiment is already a real number, $\xi = x$, and it is natural to define $\hat{\mathbf{x}}(x) = x$ in a somewhat redundant notation.

In summary, a random variable $\hat{\mathbf{x}}$ is a real number that is obtained as a result of an experiment.

The next step in the theory is to assign numbers called "probabilities" to the possible results of the experiment or, equivalently, to the different values of the random variable. The assignation should match our *a priori* expectations (if any) about the likelihood (expected frequency of appearance) of the different outcomes. For instance, when tossing a coin, it is natural (but not necessarily useful or convenient) to assign a probability equal to $1/2$ to the appearance of heads, such that $P(\mathbf{heads}) = 1/2$ or, equivalently, to the random variable $\hat{\mathbf{x}}(\mathbf{heads})$ taking the value +1 (as assigned arbitrarily before), $P(\hat{\mathbf{x}} = +1) = 1/2$. The assignation of probabilities to events might follow some physical law (as in the case of the radioactive substance, the Boltzmann law for the distribution of energies, or the quantum mechanical postulates), might come after some lengthy calculation (the probability of rain tomorrow), or might follow other arguments such as symmetry (the probability of heads is equal to $1/2$), Jayne's principle (based on the extremization of the information function), and so on; however, whatever its origin, we consider the assignation to be known. A typical consequence of the theory is the calculation of probabilities for more or less complicated events: What is the probability of obtaining five heads in a row if we toss a coin 10 times? What is the probability that the next emission of an electron by the radioactive substance occurs in the next 10 ms? and so on.

In practice, the assignation of probabilities to values of the random variable is performed differently if the random variable is continuous (i.e., it can take continuous values in a given interval $\hat{\mathbf{x}} \in (\alpha, \beta)$ where $\alpha$ can also be $-\infty$ or $\beta$ can be $+\infty$) or discrete (can take only a finite or infinite numerable set of values $\hat{\mathbf{x}} \in \{x_1, x_2, x_3, \dots\}$). For example, the random variable counting the number of times a coin must be tossed before heads appear can take an infinite numerable

set of values $\{1, 2, 3, \ldots\}$. The time at which the daily bus reaches the station can take continuous values in a finite interval $(0, 24)$ h.

For a discrete random variable taking values $\hat{\mathbf{x}} \in \{x_1, x_2, x_3, \ldots\}$, we assign to each value $x_i$ its probability $p_i = P(\hat{\mathbf{x}} = x_i)$ such that the following two conditions, namely nonnegativity and normalization, are satisfied:

$$p_i \geq 0, \quad \forall i, \tag{1.1}$$

$$\sum_{\forall i} p_i = 1. \tag{1.2}$$

One can check that the assigned probabilities $p_i$ are consistent with the actual results of the experiment. For instance, quantum mechanics might predict that, in a given experiment with an electron's spin, the random variable $\hat{\mathbf{x}}$ takes the value $x_1 = +\hbar/2$ with probability $p_1 = 1/3$ and the value $x_2 = -\hbar/2$ with probability $p_2 = 2/3$. To check whether this prediction is correct, one repeats the experiment $M$ (a large number) times and computes the frequency $f_i = n_i/M$, $n_i$ being the number of times that result $x_i$ appears, and checks whether $f_1$ is close to $1/3$ and $f_2$ to $2/3$. If they are not, then the predictions of the theory or the implementation of the experiment are wrong.[1]

For a continuous random variable $\hat{\mathbf{x}}$, we assign, instead, a probability to the random variable taking a value in a finite interval $[a, b]$ as

$$P(\hat{\mathbf{x}} \in [a, b]) = \int_a^b f_{\hat{\mathbf{x}}}(x)\, dx. \tag{1.3}$$

Here, $f_{\hat{\mathbf{x}}}(x)$ is known as the *probability density function* of the random variable $\hat{\mathbf{x}}$, or pdf for short. It is one of the most important concepts in the theory of random variables. To be able to consider $f_{\hat{\mathbf{x}}}(x)$ as a *bona fide* pdf, it must satisfy the nonnegativity and normalization conditions:

$$f_{\hat{\mathbf{x}}}(x) \geq 0, \tag{1.4}$$

$$\int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x)\, dx = 1. \tag{1.5}$$

The interpretation of the pdf is that, in the limit $dx \to 0$, $f_{\hat{\mathbf{x}}}(x)\, dx$ gives the probability that the random variable $\hat{\mathbf{x}}$ takes values between $x$ and $x + dx$, that is

$$P(x < \hat{\mathbf{x}} \leq x + dx) = f_{\hat{\mathbf{x}}}(x)\, dx. \tag{1.6}$$

In this way, the probability that the random variable $\hat{\mathbf{x}}$ takes a value within an arbitrary region $\Omega \subset \mathbb{R}$ of the real numbers is given by the integral of the pdf over that region:

$$P(\hat{\mathbf{x}} \in \Omega) = \int_{\Omega} f_{\hat{\mathbf{x}}}(x)\, dx. \tag{1.7}$$

Note that $f_{\hat{\mathbf{x}}}(x)$ has units of the inverse of the units of $x$, and it is not limited to taking values smaller than or equal to 1. A pdf governing the probability

---

1) An important issue in probability theory is to be able to conclude whether the observed frequencies $f_i$ are indeed compatible, *within unavoidable statistical errors*, with the postulated probabilities $p_i$.

of the next emission of an electron by a $\beta$-radioactive substance has units of inverse of time, or $T^{-1}$. A pdf can be computed from the experimental data. We first generate $M$ data of the random variable $\hat{\mathbf{x}}$ by repeating the experiment $M$ times and recording the outcomes $\{x_1, x_2, \ldots, x_M\}$. We choose an interval $\Delta x$ and count the number of times $n(x, x + \Delta x)$ in which the random variable has taken values in the interval $(x, x + \Delta x)$. According to the interpretation of $f_{\hat{\mathbf{x}}}(x)$, it is $f_{\hat{\mathbf{x}}}(x) \approx n(x, x + \Delta x)/(M\Delta x)$, from which $f_{\hat{\mathbf{x}}}(x)$ can be estimated. A good estimate for $f_{\hat{\mathbf{x}}}(x)$ requires $M$ to be large and $\Delta x$ to be small. Again, if the estimated $f_{\hat{\mathbf{x}}}(x)$ is not equal to the theoretical prediction, then something is wrong with the theory or with the experiment.

Further calculations can be simplified if one introduces the cumulative distribution function or cdf, $F_{\hat{\mathbf{x}}}(x)$, as

$$F_{\hat{\mathbf{x}}}(x) = \int_{-\infty}^{x} f_{\hat{\mathbf{x}}}(x')\,dx'. \tag{1.8}$$

From this definition, it follows that the cdf $F_{\hat{\mathbf{x}}}(x)$ is the probability that the random variable $\hat{\mathbf{x}}$ takes values less or equal than $x$:

$$P(\hat{\mathbf{x}} \leq x) = F_{\hat{\mathbf{x}}}(x) \tag{1.9}$$

and that

$$P(x_1 < \hat{\mathbf{x}} \leq x_2) = F_{\hat{\mathbf{x}}}(x_2) - F_{\hat{\mathbf{x}}}(x_1) \tag{1.10}$$

which is a relation that will be useful later. The following general properties arise from the definition, the nonnegativity (1.4), and the normalization condition (1.5) of the pdf $f_{\hat{\mathbf{x}}}(x)$:

$$F_{\hat{\mathbf{x}}}(x) \geq 0, \tag{1.11}$$

$$\lim_{x \to -\infty} F_{\hat{\mathbf{x}}}(x) = 0, \tag{1.12}$$

$$\lim_{x \to +\infty} F_{\hat{\mathbf{x}}}(x) = 1, \tag{1.13}$$

$$x_2 > x_1 \Rightarrow F_{\hat{\mathbf{x}}}(x_2) \geq F_{\hat{\mathbf{x}}}(x_1). \tag{1.14}$$

The last property tells us that $F_{\hat{\mathbf{x}}}(x)$ is a nondecreasing function of its argument.

If $f_{\hat{\mathbf{x}}}(x)$ is piecewise continuous, then the probability of the random variable $\hat{\mathbf{x}}$ taking a particular value $x$ is equal to zero, as it must be understood as the limit $P(\hat{\mathbf{x}} = x) = \lim_{\Delta x \to 0} \int_{x}^{x+\Delta x} f_{\hat{\mathbf{x}}}(x)\,dx = 0$. It is possible to treat discrete variables in the language of pdfs if we use the "Dirac-delta function" $\delta(x)$. This mathematical object is not a proper function, but it can be understood[2] as the limit of a succession of functions $\delta_n(x)$ such that $\delta_n(x)$ decays to zero outside a region of width $1/n$ around $x = 0$ and has a height at $x = 0$ or order $n$ such that the integral $\int_{-\infty}^{\infty} dx\,\delta_n(x) = 1$. There are many examples of such functions: for instance, $\delta_n(x) = n/\sqrt{2\pi}e^{-n^2x^2/2}$, or $\delta_n(x) = \begin{cases} 0, & x \notin (-1/n, 1/n) \\ n(1 - n|x|), & x \in (-1/n, 1/n) \end{cases}$, see Figure 1.1. The detailed shape is

2) Another way, more rigorous from the mathematical point of view, to introduce the Dirac-delta is by the use of distribution theory, but this is beyond the scope of this book.
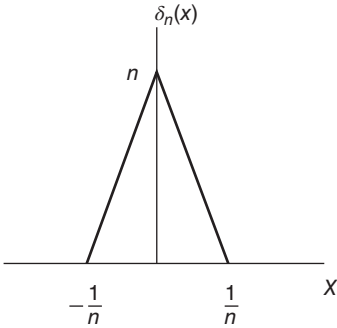
**Figure 1.1** Function $\delta_n(x)$. It has the property that $\int_{-\infty}^{\infty} dx \delta_n(x) = 1$ and, when $n \to \infty$, it tends to the delta function $\delta(x)$.

not important. What matters is that, in the limit $n \to \infty$, these functions tend to yield a nonzero value only at $x = 0$ while keeping their integral over all $\mathbb{R}$ constant. As, for an arbitrary function $f(x)$, we have

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} dx \, \delta_n(x) f(x) = f(0) \tag{1.15}$$

and so we can (in a nonrigorous way) exchange the limit and the integral and understand the Dirac-delta function as satisfying

$$\delta(x) = 0, \quad \text{if } x \neq 0, \tag{1.16}$$

$$\int_{-\infty}^{\infty} dx f(x)\delta(x) = f(0). \tag{1.17}$$

When the random variable takes a discrete (maybe infinite numerable) set of values $\hat{x} \in \{x_1, x_2, x_3, \dots\}$ such that the value $x_i$ has probability $p_i$, then the pdf can be considered as a sum of Dirac-delta functions:

$$f_{\hat{x}}(x) = \sum_{\forall i} p_i \delta(x - x_i) \tag{1.18}$$

because now $P(\hat{x} = x_i) = \lim_{\Delta x \to 0} \int_{x_i - \Delta x}^{x_i + \Delta x} f_{\hat{x}}(x) \, dx = p_i$. The corresponding cumulative function is a sum of Heaviside step functions:

$$F_{\hat{x}}(x) = \sum_{\forall i} p_i \theta(x - x_i) \tag{1.19}$$

with the usual definition

$$\theta(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases} \tag{1.20}$$

## 1.2
## Average Values, Moments

As a random variable $\hat{\mathbf{x}}$ assigns a real number $\hat{\mathbf{x}}(\xi)$ to the result of the experiment $\xi$, it is possible to use a given real function $G(x)$ to define a new random variable $\hat{\mathbf{G}}$ as $\hat{\mathbf{G}}(\xi) = G(\hat{\mathbf{x}}(\xi))$. One defines the average or expected value $E[\hat{\mathbf{G}}]$ of this random variable as

$$E[\hat{\mathbf{G}}] = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x) G(x) \, dx. \tag{1.21}$$

The alternative notations $\langle \hat{\mathbf{G}} \rangle$ or simply $E[G]$ and $\langle G \rangle$ are very common and will also be used in this book. In particular, for a discrete random variable with pdf given by (1.18), the average value is

$$E[\hat{\mathbf{G}}] = \sum_{\forall i} p_i G(x_i). \tag{1.22}$$

Some important expected values are as follows:

- Mean or average value of the random variable: $\mu[\hat{\mathbf{x}}] = E[\hat{\mathbf{x}}]$;
- Moments of order $n$: $E[\hat{\mathbf{x}}^n]$;
- Central moments of $n$: $E[(\hat{\mathbf{x}} - \mu[\hat{\mathbf{x}}])^n]$;
- Variance: $\sigma^2[\hat{\mathbf{x}}] = E[(\hat{\mathbf{x}} - \mu[\hat{\mathbf{x}}])^2] = E[\hat{\mathbf{x}}^2] - E[\hat{\mathbf{x}}]^2$. The value $\sigma[\hat{\mathbf{x}}]$ is the standard deviation of the random variable $\hat{\mathbf{x}}$.

If two random variables $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$ are related by a known function $\hat{\mathbf{y}} = y(\hat{\mathbf{x}})$, then their respective pdfs are also related.

$$f_{\hat{\mathbf{y}}}(y) = \sum_{\mu} \frac{f_{\hat{\mathbf{x}}}(x_{\mu})}{\left| \dfrac{dy}{dx} \right|_{x = x_{\mu}}} \tag{1.23}$$

where $x_{\mu}$ are the solutions of the equation $y = y(x)$. For instance, if the change is $\hat{\mathbf{y}} = \hat{\mathbf{x}}^2$, then the equation $y = x^2$ has no solutions for $y < 0$ and two solutions $x_1 = +\sqrt{y}$, $x_2 = -\sqrt{y}$ for $y \geq 0$, and the pdf for $\hat{\mathbf{y}}$ is

$$f_{\hat{\mathbf{y}}}(y) = \begin{cases} 0, & y < 0, \\ \dfrac{f_{\hat{\mathbf{x}}}(\sqrt{y}) + f_{\hat{\mathbf{x}}}(-\sqrt{y})}{2\sqrt{y}}, & y \geq 0. \end{cases} \tag{1.24}$$

## 1.3
## Some Important Probability Distributions with a Given Name

### 1.3.1
### Bernoulli Distribution

The so-called Bernoulli distribution describes a binary experiment in which only two exclusive options are possible: $A$ or $\overline{A}$ (''heads or tails'', ''either it rains or not''), with respective probabilities $p$ and $1 - p$, being $p \in [0, 1]$. We define the discrete
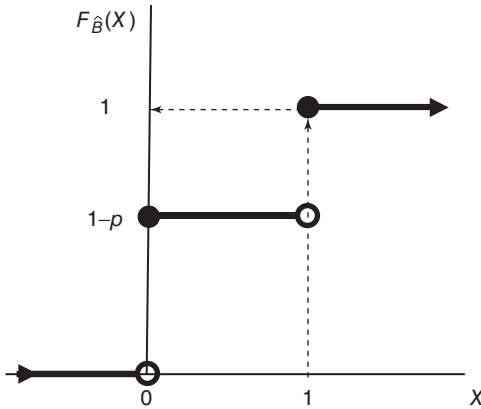
**Figure 1.2** Cumulative distribution function (cdf) $F_{\hat{B}}(x)$ of the Bernoulli random variable $\hat{B}(p)$.

Bernoulli random variable $\hat{B}$ as taking the value 1 (respectively 0) if the experiment yields $A$ (respectively $\overline{A}$). The probabilities are

$$P(\hat{B} = 1) = p, \tag{1.25}$$
$$P(\hat{B} = 0) = 1 - p. \tag{1.26}$$

The mean value and variance can be computed as

$$E[\hat{B}] = p, \tag{1.27}$$
$$\sigma^2[\hat{B}] = p(1 - p). \tag{1.28}$$

Eventually, and when needed, we will use the notation $\hat{B}(p)$ to denote a random variable that follows a Bernoulli distribution with parameter $p$. According to the general expression, the cdf of this random variable is $F_{\hat{B}}(x) = (1 - p)\theta(x) + p\theta(x - 1)$, or

$$F_{\hat{B}}(x) = \begin{cases} 0, & x < 0, \\ 1 - p, & 0 \le x < 1, \\ 1, & x \ge 1 \end{cases} \tag{1.29}$$

which is plotted in Figure 1.2.

## 1.3.2
## Binomial Distribution

We now repeat $M$ times the binary experiment of the previous case and count how many times $A$ appears (independently of the order of appearance). This defines a random variable which we call $\hat{N}_B$. It is a discrete variable that can take any integer value between 0 and $M$ with probabilities

$$p(\hat{N}_B = n) = \binom{M}{n} p^n (1 - p)^{M-n}. \tag{1.30}$$

$\hat{\mathbf{N}}_B$ is said to follow a binomial distribution. The mean value and the variance are given by

$$E[\hat{\mathbf{N}}_B] = Mp, \tag{1.31}$$

$$\sigma^2[\hat{\mathbf{N}}_B] = Mp(1 - p). \tag{1.32}$$

We will denote by $\hat{\mathbf{N}}_B(p, M)$ a random variable that follows a binomial distribution with probability $p$ and number of repetitions $M$.

### 1.3.3
### Geometric Distribution

We consider, again, repetitions of the binary experiment, but now the random variable $\hat{\mathbf{N}}_G$ is defined as the number of times we must repeat the experiment before the result $A$ appears (not including the case in which $A$ does appear). This is a discrete random variable that can take any integer value $0, 1, 2, 3, \ldots$. The probability that it takes a value equal to $n$ is

$$p(\hat{\mathbf{N}}_G = n) = (1 - p)^n p, \quad n = 0, 1, 2, \ldots. \tag{1.33}$$

The mean value and variance are

$$E[\hat{\mathbf{N}}_G] = \frac{1 - p}{p}, \tag{1.34}$$

$$\sigma^2[\hat{\mathbf{N}}_G] = \frac{1 - p}{p^2}. \tag{1.35}$$

### 1.3.4
### Uniform Distribution

This is our first example of a continuous random variable. We want to describe an experiment in which all possible results are real numbers within the interval $(a, b)$ occurring with the same probability, while no result can appear outside this interval. We will use the notation $\hat{\mathbf{U}}(a, b)$ to indicate a uniform random variable in the interval $(a, b)$. The pdf is then constant within the interval $(a, b)$ and 0 outside it. Applying the normalization condition, it is precisely

$$f_{\hat{\mathbf{x}}}(x) = \begin{cases} \dfrac{1}{b - a}, & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases} \tag{1.36}$$

The cumulative function is

$$F_{\hat{\mathbf{x}}}(x) = \begin{cases} 0, & x < a, \\ \dfrac{x - a}{b - a}, & a \leq x < b, \\ 1, & x \geq b. \end{cases} \tag{1.37}$$
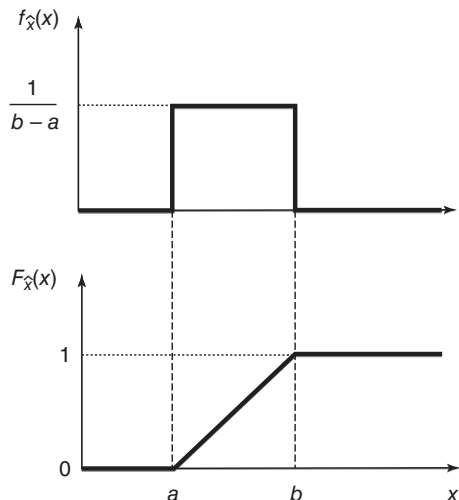
These two functions are plotted in Figure 1.3.

**Figure 1.3**  Top: Probability density function (pdf) $f_{\hat{\mathbf{x}}}(x)$ of the $\hat{\mathbf{U}}(a, b)$ distribution uniformly distributed in the interval $(a, b)$. Bottom: The corresponding cumulative distribution function (cdf) $F_{\hat{\mathbf{x}}}(x)$.

The mean value and variance are

$$E[\hat{\mathbf{x}}] = \frac{a + b}{2}, \tag{1.38}$$

$$\sigma^2[\hat{\mathbf{x}}] = \frac{(b - a)^2}{12}. \tag{1.39}$$

The uniform distribution $\hat{\mathbf{U}}(0, 1)$ appears in a very important result. Let us consider an arbitrary random variable $\hat{\mathbf{x}}$ (discrete or continuous) whose cdf is $F_{\hat{\mathbf{x}}}(x)$, and let us define the new random variable $\hat{\mathbf{u}} = F_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$. We will prove that $\hat{\mathbf{u}}$ is a $\hat{\mathbf{U}}(0, 1)$ variable.

The proof is as follows: Let us compute the cdf of $\hat{\mathbf{u}}$ starting from its definition

$$F_{\hat{\mathbf{u}}}(u) = P(\hat{\mathbf{u}} \leq u) = P(F_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) \leq u). \tag{1.40}$$

As $F_{\hat{\mathbf{x}}}(x) \in [0, 1]$, the condition $F_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) \leq u$ requires necessarily $u \geq 0$, so $F_{\hat{\mathbf{u}}}(u) = 0$ if $u < 0$. If, on the other hand, $u > 1$, then the condition $F_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) \leq u$ is always satisfied and its probability is 1, or $F_{\hat{\mathbf{u}}}(u) = 1$ if $u \geq 1$. Finally, for $u \in (0, 1)$, the condition $F_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) \leq u$ is equivalent to $\hat{\mathbf{x}} \leq F_{\hat{\mathbf{x}}}^{-1}(u)$, as the function $F_{\hat{\mathbf{x}}}(x)$ is a nondecreasing function. This gives

$$F_{\hat{\mathbf{u}}}(u) = P(\hat{\mathbf{x}} \leq F_{\hat{\mathbf{x}}}^{-1}(u)) = F_{\hat{\mathbf{x}}}(F_{\hat{\mathbf{x}}}^{-1}(u)) = u. \tag{1.41}$$

Summing up,

$$F_{\hat{\mathbf{u}}}(u) = \begin{cases} 0, & u < 0, \\ u, & 0 \leq u \leq 1, \\ 1, & u > 1, \end{cases} \tag{1.42}$$

which is nothing but the cdf of a uniform random variable $\hat{\mathbf{U}}(0, 1)$.

## 1.3.5
**Poisson Distribution**

Let us consider the binomial distribution in the limit of infinitely many repetitions $M$. If we take the double limit $M \to \infty$, $p \to 0$ but keeping $Mp \to \lambda$, a finite value, the binomial distribution $\hat{\mathbf{N}}_B(p)$ tends to the so-called Poisson distribution $\hat{\mathbf{P}}(\lambda)$. With the help of the Stirling approximation $m! \approx m^m e^{-m} \sqrt{2\pi m}$, which is valid in the limit $m \to \infty$, we can prove, starting from (1.30), the following expression for the probabilities of the Poisson distribution:

$$P(\hat{\mathbf{P}} = n) = e^{-\lambda} \frac{\lambda^n}{n!}, \qquad n = 0, 1, \dots, \infty. \tag{1.43}$$

The Poisson distribution is one of the most important distributions in nature, probably second only to the Gaussian distribution (to be discussed later). The Poisson distribution has both mean and variance equal to the parameter $\lambda$:

$$E[\hat{\mathbf{P}}] = \sigma^2[\hat{\mathbf{P}}] = \lambda \tag{1.44}$$

which is a typical property that characterizes the Poisson distribution.

We can think of the Poisson distribution simply as a convenient limit that simplifies the calculations in many occasions. For instance, the probability that a person was born on a particular day, say 1 January, is $p = 1/365$, approximately.[3] Imagine that we have now a large group of $M = 500$ people. What is the probability that exactly three people were born on 1 January? The correct answer is given by the binomial distribution by considering the events $A$="being born on 1 January" with probability $p = \frac{1}{365}$ and $\overline{A}$="not being born on 1 January" with probability $1 - p = \frac{364}{365}$:

$$P(\hat{\mathbf{N}}_B = 3) = \binom{500}{3} \left(\frac{1}{365}\right)^3 \left(\frac{364}{365}\right)^{497} = 0.108919\dots \tag{1.45}$$

As $p$ is small and $M$ large, we might find it justified to use the Poisson approximation, $\lambda = pM \approx 500/365 = 1.37$, to obtain

$$P(\hat{\mathbf{P}} = 3) = e^{-1.37} \frac{1.37^3}{3!} = 0.108900\dots \tag{1.46}$$

which is good enough. Let us compute now, using this limit, the probability that at least two persons were born on 11 May.

$$P(\hat{\mathbf{P}} \geq 2) = 1 - P(\hat{\mathbf{P}} \leq 1) = 1 - P(\hat{\mathbf{P}} = 0) - P(\hat{\mathbf{P}} = 1)$$
$$= 1 - e^{\lambda} - \lambda e^{\lambda} = 0.3977\dots \tag{1.47}$$

which is to be compared with the exact result $1 - P(\hat{\mathbf{N}}_B = 0) - P(\hat{\mathbf{N}}_B = 1) = 1 - \binom{500}{0} \left(\frac{1}{365}\right)^0 \left(\frac{364}{365}\right)^{500} - \binom{500}{1} \left(\frac{1}{365}\right)^1 \left(\frac{364}{365}\right)^{499} = 0.397895\dots$, which is again a reasonable approximation.

There are occasions in which the Poisson limit occurs exactly. Imagine we distribute $M$ dots randomly with a distribution $\hat{\mathbf{U}}[0, T]$, uniform in the interval

---

3) Neglecting leap years and assuming that all birth days are equally probable.

[0, *T*] (we will think immediately of this as events occurring randomly in time with a uniform rate, hence the notation). We call $\omega = M/T$ the "rate" (or "frequency") at which points are distributed. We now ask the question: what is the probability that exactly $k$ of the $M$ dots lie in the interval $[t_1, t_1 + t] \in [0, T]$? The event $A$="one given dot lies in the interval $[t_1, t_1 + t]$" has probability $p = \frac{t}{T}$, whereas the event $\overline{A}$="the given dot does not lie in the interval $[t_1, t_1 + t]$" has probability $q = 1 - p$. The required probability is given by the binomial distribution, $\hat{\mathbf{B}}(p, M)$, defined by (1.30). We now make the limit $M \to \infty$, $T \to \infty$ but $\omega = M/T$ finite. This limit corresponds to the distribution in which the events occur uniformly in time with a rate (frequency) $\omega$. As mentioned earlier, it can be proven, using Stirling's approximation, that, in this limit, the binomial distribution $\hat{\mathbf{B}}(p, M)$ tends to a Poisson distribution $\hat{\mathbf{P}}(\lambda)$, of parameter $\lambda = pM = \omega t$, finite. Let us give an example. Consider $N$ atoms of a $\beta$-radioactive substance. Each atom emits one electron independently of the others. The probability that the given atom will disintegrate is constant with time, but it is not known which atoms will disintegrate in a given time interval. All we observe is the emission of electrons with a given rate. It is true that, as time advances, the number of atoms that can disintegrate decreases, although for some radioactive elements the decay rate is extremely slow (on the order of billions of years for the radioactive element $^{40}_{19}$K, for example). We can hence assume a constant rate $\omega$ which can be estimated simply by counting the number of electrons $M$ emitted in a time interval $T$ as $\omega = M/T$. Under these circumstances, the number $k$ of electrons emitted in the time interval $[t_1, t_1 + t]$ follows a Poisson distribution of the parameter $\lambda = pM = \frac{t}{T}M = \omega t$, or

$$P(k; t) = \mathrm{e}^{-\omega t} \frac{(\omega t)^k}{k!}. \tag{1.48}$$

### 1.3.6
### Exponential Distribution

A continuous random variable $\hat{\mathbf{x}}$ follows an exponential distribution if its pdf is

$$f_{\hat{\mathbf{x}}}(x) = \begin{cases} 0, & x < 0, \\ a\mathrm{e}^{-ax}, & x \geq 0. \end{cases} \tag{1.49}$$

The mean value and variance are

$$E[\hat{\mathbf{x}}] = \frac{1}{a}, \tag{1.50}$$

$$\sigma^2[\hat{\mathbf{x}}] = \frac{1}{a^2} \tag{1.51}$$

with $a > 0$ being a parameter.

An interesting example is related to the Poisson distribution. Consider the emission of electrons by a radioactive substance which we know is governed by the Poisson distribution for those time intervals such that the emission rate can be considered constant. Let us set our clock at $t = 0$ and then measure the time $t$ of the first observed emission of an electron. This time is a random variable $\hat{\mathbf{t}}$ (a number

associated with the result of an experiment) and has a pdf which we call $f_{\hat{\mathbf{t}}}^{1\text{st}}(t)$. By definition, $f_{\hat{\mathbf{t}}}^{1\text{st}}(t)dt$ is the probability that the first electron is emitted during the interval $(t, t + dt)$, and, accordingly, the probability that the first electron is emitted after time $t$ is $\int_t^\infty f_{\hat{\mathbf{t}}}^{1\text{st}}(t')dt'$. This is equal to the probability that no electrons have been emitted during $(0, t)$ or $P(0; t) = \mathrm{e}^{-\omega t}$, that is

$$\int_t^\infty f_{\hat{\mathbf{t}}}^{1\text{st}}(t')dt' = \mathrm{e}^{-\omega t}, \ t \geq 0. \tag{1.52}$$

Taking the time derivate on both sides of this equation, we obtain $f_{\hat{\mathbf{t}}}^{1\text{st}}(t) = \omega \mathrm{e}^{-\omega t}$, which is valid for $t \geq 0$, the exponential distribution. Alternatively, if $\hat{\mathbf{t}}$ follows this exponential distribution, then the number of events occurring in a time interval $(0, 1)$ follows a Poisson $\hat{\mathbf{P}}(\lambda)$ distribution with $\lambda = \omega \times 1 = \omega$.

### 1.3.7
### Gaussian Distribution

A continuous random variable $\hat{\mathbf{x}}$ follows a Gaussian distribution if its pdf is

$$f_{\hat{\mathbf{x}}}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]. \tag{1.53}$$

The average and variance are given by

$$E[\hat{\mathbf{x}}] = \mu, \tag{1.54}$$
$$\sigma^2[\hat{\mathbf{x}}] = \sigma^2. \tag{1.55}$$

We will use the notation that $\hat{\mathbf{x}}$ is a $\hat{\mathbf{G}}(\mu, \sigma)$ random variable. The cdf is

$$F_{\hat{\mathbf{x}}}(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right) \tag{1.56}$$

with $\mathrm{erf}(z)$ being the error function

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-\gamma^2} \, d\gamma. \tag{1.57}$$

$f_{\hat{\mathbf{x}}}(x)$ and $F_{\hat{\mathbf{x}}}(x)$ are plotted in Figure (1.4).

Gaussian random variables are very important in practice because they appear in a large number of problems, either as an exact distribution in some limit or, simply, as providing a sufficient approximation to the real distribution. After all, it is not unusual that many distributions have a maximum value and this can in many cases be approximated by a Gaussian distribution (the so-called bell-shaped curve). One of the reasons for the widespread appearance of Gaussian distributions is the central-limit theorem, which states that the sum of a large number of independent random variables, whatever their distribution, will approach a Gaussian distribution.
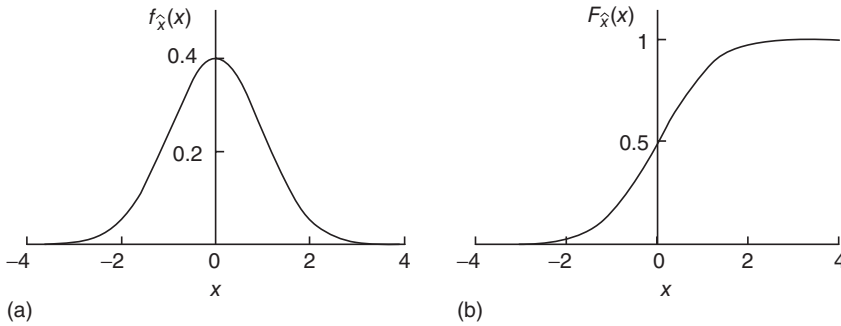
**Figure 1.4** pdf and cdf of the Gaussian distribution of mean 0 and variance 1.

As a first example, one we can prove that the binomial distribution $\hat{\mathbf{N}}_B(p, M)$ tends to the Gaussian distribution $\hat{\mathbf{G}}(Mp, \sqrt{Mp(1-p)})$. More precisely

$$P(\hat{\mathbf{N}}_B = n) = \binom{M}{n} p^n (1-p)^{M-n}$$
$$\approx \frac{\exp\left[-(n-Mp)^2/2Mp(1-p)\right]}{\sqrt{2\pi Mp(1-p)}}. \tag{1.58}$$

The theorem of de Moivre–Laplace, loosely speaking, establishes the equality of both sides of the previous equation in the limit $M \to \infty$ if $|n - Mp| / \sqrt{Mp(1-p)}$ remains finite. In practice, the above approximation is sufficiently good for $M \geq 100$ if $p = 0.5$ or $M \geq 1000$ if $p = 0.1$ (Figure 1.5). As can be seen in these figures, the Gaussian approximation to the binomial distribution is best around the maximum of the distribution and worsens in the tails. An equivalent way of stating the equivalence of both distributions is to define the random variable $\hat{\mathbf{x}} = \frac{\hat{\mathbf{N}}_B - Mp}{\sqrt{Mp(1-p)}}$. As $\langle \hat{\mathbf{x}} \rangle = 0$ and $\sigma[\hat{\mathbf{x}}] = 1$, it follows that $\hat{\mathbf{x}}$ obeys a Gaussian distribution $\hat{\mathbf{G}}(0, 1)$ in the limit $M \to \infty$.

The Gaussian distribution can also be obtained as the limit of the Poisson distribution for large parameter $\lambda \to \infty$. This yields a Gaussian distribution of the same mean and variance, or $\hat{\mathbf{G}}(\lambda, \sqrt{\lambda})$. Again, although the exact result refers to the limit $\lambda \to \infty$, in practice the approximation can be considered sufficient for $\lambda \geq 100$, especially around the maximum of the distribution (Figure 1.5).

### 1.3.8
### Gamma Distribution

A random variable $\hat{\mathbf{x}}$ is said to follow a gamma distribution $\hat{\mathbf{\Gamma}}(\alpha, \theta)$ with shape $\alpha$ and scale parameter $\theta$ if the pdf is

$$f_{\hat{\mathbf{x}}}(x) = \begin{cases} 0, & x < 0, \\ \dfrac{x^{\alpha-1} e^{-x/\theta}}{\theta^\alpha \Gamma(\alpha)}, & x \geq 0 \end{cases} \tag{1.59}$$
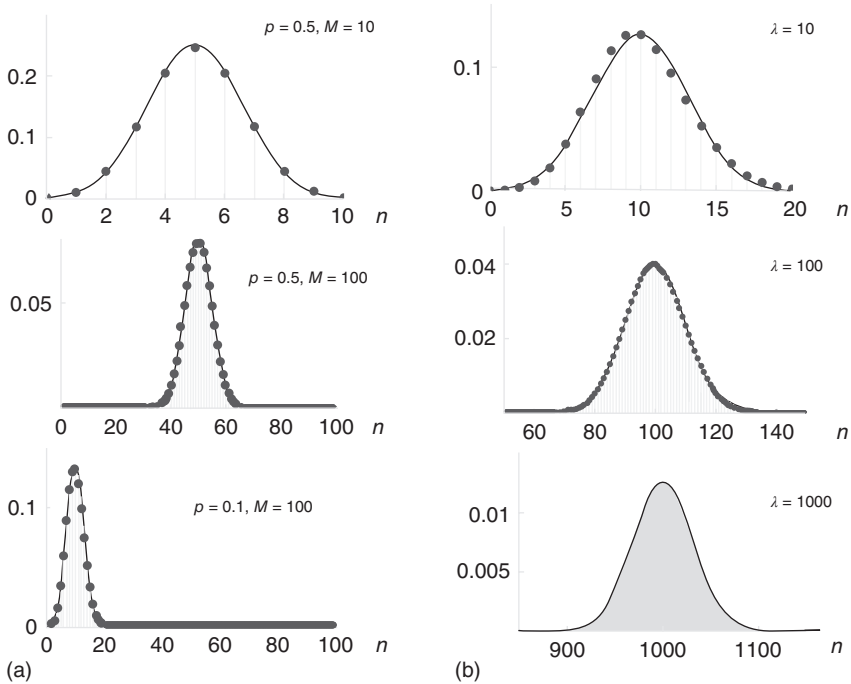
**Figure 1.5** (a) Binomial distribution (dots) and its Gaussian approximation (solid line). (b) Poisson distribution (dots) and its binomial approximation (solid line).

where $\alpha > 0$, $\theta > 0$ are real numbers. Here, $\Gamma(\alpha)$ is the Gamma function which has the interesting property that $\Gamma(x + 1) = x\Gamma(x)$ and that equals the factorial when the argument is an integer number $\Gamma(n + 1) = n!$. The mean value and variance of this distribution are given by

$$E[\hat{\mathbf{x}}] = \alpha\theta, \tag{1.60}$$
$$\sigma^2[\hat{\mathbf{x}}] = \alpha\theta^2. \tag{1.61}$$

The shape depends on the value of $\alpha$. For $0 < \alpha < 1$, it diverges at $x = 0$, while for $\alpha \geq 1$ it has a single maximum located at $x = \alpha - 1$. In Figure 1.6, we plot the gamma distribution for different values of the parameter $\alpha$.

Note that, for $\theta = 1$, the Gamma and the Poisson distributions share the property that the mean value is equal to the variance. In fact, one can prove that the Poisson distribution $\hat{\mathbf{P}}(\lambda)$ can be approximated for large values of $\lambda$ by the $\hat{\mathbf{\Gamma}}(\lambda, \theta = 1)$. Evidence for this is given in Figure 1.7.

### 1.3.9
### Chi and Chi-Square Distributions

If $\alpha = k/2$, with $k \in \mathbb{N}$ and $\theta = 2$, the Gamma distribution $\hat{\mathbf{\Gamma}}(k/2, 2)$ is also called the $\hat{\chi}^2(k)$ or chi-square distribution with $k$ degrees of freedom. This is the distribution
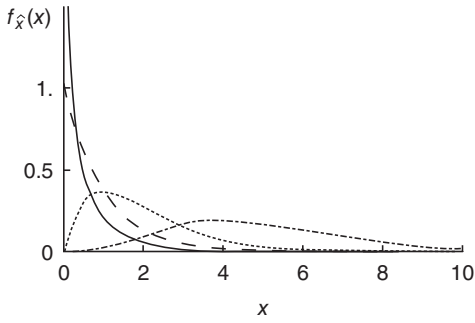
**Figure 1.6** pdf for the Gamma distribution $\hat{\Gamma}(\alpha, \theta = 1)$ for $\alpha = 0.5$ (solid), 1 (dashed), 2 (dotted), 5 (dot-dashed).
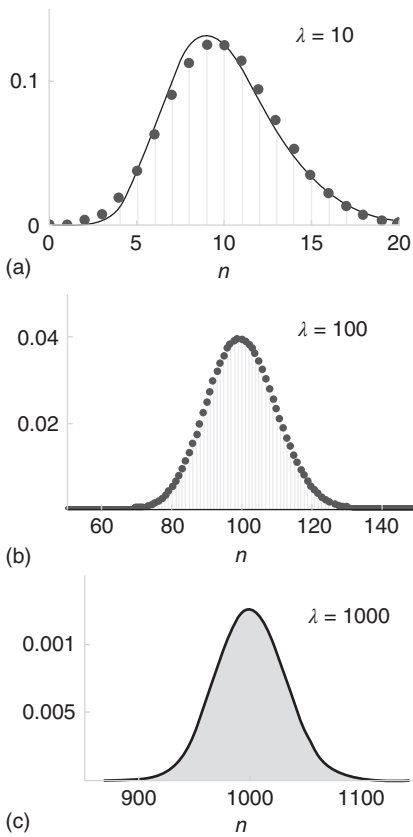


**Figure 1.7** Comparison between the Poisson $\hat{P}(\lambda)$ (dots) and the $\hat{\Gamma}(\lambda, \theta = 1)$ (solid line) distributions for different values of the parameter $\lambda$.

followed by the sum of the squares of $k$ independent $\hat{\mathbf{G}}(0,1)$ Gaussian variables of mean 0 and variance 1. Its square root, or chi distribution $\hat{\chi}(k)$, follows the pdf

$$f_{\hat{\chi}}(\chi) = \frac{2^{1-\frac{k}{2}} \chi^{k-1} e^{-\chi^2/2}}{\Gamma\left(k/2\right)} \tag{1.62}$$

with mean value and variance

$$\langle \hat{\chi} \rangle = \sqrt{2} \frac{\Gamma((k+1)/2)}{\Gamma(k/2)}, \tag{1.63}$$

$$\sigma^2[\hat{\chi}] = k - \langle \hat{\chi} \rangle^2. \tag{1.64}$$

## 1.4
## Successions of Random Variables

It is, of course, possible (and useful) to assign more that one random variable to the result of an experiment. For example, we can measure in a $\beta$-radioactive sample the time $t$ and the speed $v$ at which an electron is emitted; we can measure the time of arrival of the bus and the number of people in the waiting queue; we can observe whether it rains or not and measure the air temperature and pressure, and so on. In general, given an experiment, let us consider $N$ random variables assigned to it: $(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N)$. The joint pdf of all these random variables is a function of $N$ real variables $f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N)$, which allows us to compute the probability that the vector of results $(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N)$ belongs to a region $\Omega$ of $\mathbb{R}^N$ as

$$P((\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N) \in \Omega) = \int_\Omega dx_1 \ldots dx_N f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N). \tag{1.65}$$

In other words, $f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N)\, dx_1 \ldots dx_N$ is the probability that in a measurement of the $N$ random variables the value of $\hat{\mathbf{x}}_1$ lies in $(x_1, x_1 + dx_1)$, the value of $\hat{\mathbf{x}}_2$ in $(x_2, x_2 + dx_2)$, and so on. The cdf is defined as

$$F_{\hat{\mathbf{x}}_1,\hat{\mathbf{x}}_2,\ldots,\hat{\mathbf{x}}_N}(x_1, x_2, \ldots, x_N) = \tag{1.66}$$
$$\int_{-\infty}^{x_1} dx_1' \int_{-\infty}^{x_2} dx_2' \ldots \int_{-\infty}^{x_N} dx_N' \, f_{\hat{\mathbf{x}}_1,\hat{\mathbf{x}}_2,\ldots,\hat{\mathbf{x}}_N}(x_1', x_2', \ldots, x_N').$$

We do have an intuitive idea of when some random variables can be considered independent of each other. A precise statement is that the $N$ random variables $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$ are defined to be statistically independent if the joint pdf factorizes as the product of pdfs for each variable, that is

$$f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N) = f_{\hat{\mathbf{x}}_1}(x_1) f_{\hat{\mathbf{x}}_2}(x_2) \cdots f_{\hat{\mathbf{x}}_N}(x_N). \tag{1.67}$$

The mean value of a function of $N$ variables $G(x_1, \ldots, x_N)$ is computed as

$$\langle G(x_1, \ldots, x_N) \rangle = \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_N \, G(x_1, \ldots, x_N) f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_N). \tag{1.68}$$

In particular, if $G(x_1, \ldots, x_N) = \lambda_1 G_1(x_1) + \cdots + \lambda_N G_N(x_N)$, then

$$\langle G(x_1, \ldots, x_N) \rangle = \lambda_1 \langle G_1(x_1) \rangle + \cdots + \lambda_N \langle G_N(x_N) \rangle \tag{1.69}$$

and if the random variables $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$ are independent of each other, then

$$\sigma^2[G(x_1, \ldots, x_N)] = \lambda_1^2 \sigma^2[G_1(x_1)] + \cdots + \lambda_N^2 \sigma^2[G_N(x_N)]. \tag{1.70}$$

The covariance between two of the random variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{x}}_j$ is defined as

$$C[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j] \equiv C_{ij} \equiv \langle (\hat{\mathbf{x}}_i - \langle \hat{\mathbf{x}}_i \rangle)(\hat{\mathbf{x}}_j - \langle x_j \rangle) \rangle. \tag{1.71}$$

Note that a trivial consequence of that definition is that the matrix whose entries are the covariances is symmetrical, $C_{ij} = C_{ji}$. If the variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{x}}_j$ are statistically independent, then it is easy to verify that

$$C_{ij} = \sigma^2[\hat{\mathbf{x}}_i] \delta_{ij} \tag{1.72}$$

although the inverse statement (if $C_{ij} = \sigma^2[\hat{\mathbf{x}}_i]\delta_{ij}$ then variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{x}}_j$ are statistically independent) need not be true.

In general, the variance of the sum of two functions $G_1(x)$, $G_2(x)$,

$$\sigma^2[G_1 + G_2] = \langle (G_1 + G_2)^2 \rangle - \langle G_1 + G_2 \rangle^2 \tag{1.73}$$

can be written as

$$\sigma^2[G_1 + G_2] = \sigma^2[G_1] + \sigma^2[G_2] + 2C[G_1, G_2] \tag{1.74}$$

with $C[G_1, G_2]$ being the covariance of $G_1$ and $G_2$.

The correlation coefficient $\rho[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j]$ of the random variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{x}}_j$ is defined as a suitable normalization of the covariance:

$$\rho[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j] = \frac{C[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j]}{\sigma[\hat{\mathbf{x}}_i]\sigma[x_j]}. \tag{1.75}$$

From the definition, it follows that

$$|\rho[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j]| \le 1. \tag{1.76}$$

Even if there are $N$ random variables $(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N)$ defined in an experiment, we can still "forget" about some of them and consider the pdfs of only a subset of variables, for instance, $f_{\hat{\mathbf{x}}_1}(x_1)$ or $f_{\hat{\mathbf{x}}_2\hat{\mathbf{x}}_4}(x_2, x_4)$. These are called, in this context, "marginal" probabilities and can be obtained by integrating out the variables that are not of interest. For example

$$f_{\hat{\mathbf{x}}_1}(x_1) = \int_{-\infty}^{\infty} dx_2 f_{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2}(x_1, x_2), \tag{1.77}$$

or

$$f_{\hat{\mathbf{x}}_2\hat{\mathbf{x}}_4}(x_2, x_4) = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_3 \, f_{\hat{\mathbf{x}}_1\hat{\mathbf{x}}_2\hat{\mathbf{x}}_3\hat{\mathbf{x}}_4}(x_1, x_2, x_3, x_4). \tag{1.78}$$

It is possible to relate the joint pdfs of two sets of related random variables: $\hat{\mathbf{y}}_1 = \gamma_1(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_n), \dots, \hat{\mathbf{y}}_n = \gamma_n(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_n)$. The result generalizing (1.23) is

$$f_{\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_n}(\gamma_1, \dots, \gamma_n) = \sum_{\mu} \frac{f_{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_N}(x_{1_\mu}, \dots, x_{n_\mu})}{\left| J\left(\frac{\gamma_1, \dots, \gamma_n}{x_1, \dots, x_n}\right)\right|_{x_i = x_{i_\mu}}} \tag{1.79}$$

where the sum runs again over all solutions of $\gamma_1 = \gamma_1(x_1, x_2, \dots, x_n), \dots, \gamma_n = \gamma_n(x_1, x_2, \dots, x_n)$ and $J$ is the Jacobian matrix of coefficients $J_{ij} = \frac{\partial \gamma_i}{\partial x_j}$.

## 1.5
## Jointly Gaussian Random Variables

There are not many examples (besides those derived from statistically independent variables) of specific forms for a joint pdf $f_{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_N}(x_1, \dots, x_N)$. A particularly useful case is that of jointly Gaussian random variables for which the pdf is the exponential of a quadratic form, namely

$$f_{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_N}(x_1, \dots, x_N) = \sqrt{\frac{|\mathbf{A}|}{(2\pi)^N}} \, \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - \mu_i) A_{ij}(x_j - \mu_j)\right]. \tag{1.80}$$

Here, $\mathbf{A} = \{A_{ij}\}_{i=1,\dots,N; j=1,\dots,N}$ is a symmetric matrix, and $\mu_1, \mu_2, \dots, \mu_N$ are real constants. The joint pdf has, hence, $\frac{N(N+1)}{2} + N = \frac{N(N+3)}{2}$ constants. As the pdf must be normalizable, it must go to zero as any of the variables $(x_1, \dots, x_N)$ tends to $\pm\infty$. This implies that the quadratic form in the exponential must be positive definite. A simple way of checking this out is to make sure that all eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_N$ of matrix $\mathbf{A}$ are strictly positive (remember that they will be real as the matrix is symmetric). The determinant of $\mathbf{A}$ is nothing but the product of the eigenvalues $|\mathbf{A}| = \prod_{i=1}^{N} \lambda_i$. The shape of the Gaussian distribution is such that it has a maximum at the point $(\mu_1, \dots, \mu_N)$ and it decays to zero in the typical bell-shaped curve for large values of the coordinates, see Figure 1.8.

Some average values and correlations are given by

$$\langle \hat{\mathbf{x}}_i \rangle = \mu_i, \tag{1.81}$$

$$C_{ij} = (A^{-1})_{ij}. \tag{1.82}$$

Hence, the correlation matrix $\mathbf{C} = \{C_{ij}\}_{i=1,\dots,N; j=1,\dots,N}$ is the inverse of the matrix $\mathbf{A}$, $\mathbf{C} = \mathbf{A}^{-1}$. It is common, instead of writing out in full the expression (1.80), to characterize a jointly Gaussian distribution by giving the mean values $\mu_i$ and the correlations $C_{ij}$. In fact, most of the integrals needed in the calculations can be obtained directly from these numbers by the application of Wick's theorem. The theorem gives an expression for the calculation of averages of an even product of terms as

$$\langle (x_{i_1} - \mu_{i_1})(x_{i_2} - \mu_{i_2}) \dots (x_{i_{2n}} - \mu_{i_{2n}}) \rangle = \sum_{\text{all possible pairings}} C_{j_1 k_1} C_{j_2 k_2} \dots C_{j_n k_n}. \tag{1.83}$$
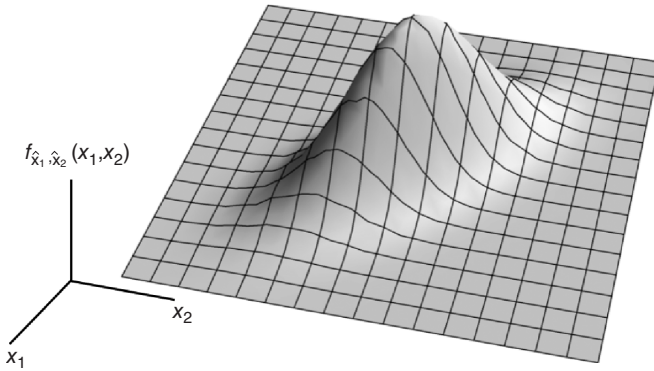
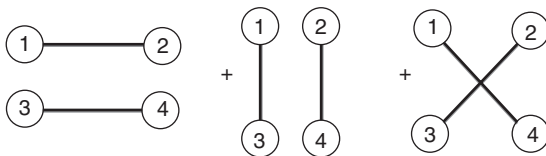**Figure 1.8** Joint Gaussian pdf in $n = 2$ variables.



**Figure 1.9** Graphs for the calculation of $\langle x_1 x_2 x_3 x_4 \rangle$ using Wick's theorem.

But, if the number of terms in the left-hand side is odd, then the average value would be 0.

Wick's theorem is better understood by specific examples. Let us take a set of four random variables $(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{x}}_3, \hat{\mathbf{x}}_4)$ which follow a jointly Gaussian distribution with average values $\langle x_1 \rangle = \langle x_2 \rangle = \langle x_3 \rangle = \langle x_4 \rangle = 0$ and a (symmetric) correlation matrix $C_{ij}$. To compute, for example, $\langle x_1 x_2 x_3 x_4 \rangle$, we need to take all possible pairings of the four numbers $1, 2, 3, 4$. They are $(1, 2)(3, 4)$, $(1, 3)(2, 4)$, and $(1, 4)(2, 3)$. This gives, according to Wick's theorem

$$\langle x_1 x_2 x_3 x_4 \rangle = C_{12} C_{34} + C_{13} C_{24} + C_{14} C_{23}. \tag{1.84}$$

It is useful to use a diagrammatic representation of this theorem. Each factor $(x_i - \mu_i)^k$ is represented by a dot with $k$ "branches" coming out of it. All possible pairings can be realized by joining all the branches in all possible ways. Figure 1.9 exemplifies this case.

Let us consider another example for the same set of four random variables (with zero means, $\mu_i = 0$). We want to compute $\langle x_1^2 x_2 x_3 \rangle$. As this is equal to $\langle x_1 x_1 x_2 x_3 \rangle$, Wick's theorem tells us about all pairings of the variables $x_1$, $x_1$, $x_2$, $x_3$. The diagrammatic representation is shown in Figure 1.10: Or, in equation form,

$$\langle x_1^2 x_2 x_3 \rangle = C_{11} C_{23} + 2 C_{12} C_{13}. \tag{1.85}$$

The factor 2 of the second term on the right-hand side is the symmetry factor of the diagram. It corresponds to the two ways in which the branches coming out of dot number 1 can be joined to the branches of dot number 2 and dot number 3.
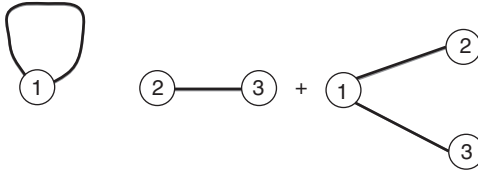
**Figure 1.10** Graphs for the calculation of $\langle x_1^2 x_2 x_3 \rangle$ using Wick's theorem.

Two very important results apply to a set of $N$ jointly Gaussian random variables:

1) The marginal pdf of a subset of $m$ random variables is also a jointly Gaussian;
2) Defining new random variables $\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_n$ as linear combinations $\hat{\mathbf{y}}_i = \sum_{j=1}^{N} B_{ij} \hat{\mathbf{x}}_j$, then $\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_n$ are also jointly Gaussian variables. The new correlation matrix is $\mathbf{BCB}^{-1}$.

## 1.6
### Interpretation of the Variance: Statistical Errors

Let us consider a random variable $\hat{\mathbf{x}}$ assigned to an experiment. In general, every time we execute an experiment and obtain a result $\xi$, we do not know *a priori* which numerical value, $\hat{\mathbf{x}}(\xi)$, will the random variable take (unless there exists an event with probability 1). That is why it is called a *random variable*. Imagine we do know the average value $\mu = E[\hat{\mathbf{x}}]$ and the variance $\sigma^2 = E[\hat{\mathbf{x}}^2] - E[\hat{\mathbf{x}}]^2$. Maybe this knowledge comes from some theory that provides us with the values of $\mu$ and $\sigma$. What can we say about a single outcome $\hat{\mathbf{x}}(\xi)$ of the random variable? Not much in general. But we can say something about the probability of $\hat{\mathbf{x}}(\xi)$ taking values far away from $\mu$, the mean value. Intuitively, we expect that it is unlikely to obtain values very far away from $\mu$. But how unlikely? Chebyshev's theorem quantifies this probability.

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \geq k\,\sigma) \leq \frac{1}{k^2} \tag{1.86}$$

for an arbitrary $k > 1$. In words, the probability that a single measurement $\hat{\mathbf{x}}(\xi)$ of a random variable differs from the mean value $\mu$ by an amount larger than $k$ times the standard deviation $\sigma$ is smaller than $k^{-2}$. The result can be written with the equivalent expression

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \leq k\,\sigma) \geq 1 - \frac{1}{k^2}. \tag{1.87}$$

For instance, if $k = 3$, it is less than $1/3^2 \approx 11\%$ probable that the result of a single experiment lies outside the interval $(\mu - 3\sigma, \mu + 3\sigma)$. In other words, we cannot predict the result of a single experiment but we can affirm with an 11% confidence (about 89 out of every 100 times we make the experiment) that it will lie in the interval $(\mu - 3\sigma, \mu + 3\sigma)$. Of course, if $\sigma$ is a large number, this prediction might be useless, but the reverse is also true, that is, if $\sigma$ is small, then we can

be pretty sure of the result. Imagine that the experiment is to measure the radius of one polystyrene bead taken at random from a large number we bought from a manufacturer who tells us that the average radius of the set is $\mu = 3.5$ mm and the standard deviation is $\sigma = 1.0\,\mu$m. How confident can we be that the radius of that particular bead lies in the interval $(3.49, 3.51)$ mm? To apply Chebyshev's inequality to this data, we need to take $(3.49, 3.51) = (\mu - k\sigma, \mu + k\sigma)$ or $0.01$mm $= k \times 1\mu$ m or $k = 10$. This means that, on average, 1 out of $k^2 = 100$ beads will not have a radius within these limits (or, from the positive side, 99 out of 100 beads will have a radius within these limits). This interpretation of Chebyshev's theorem allows us to identify (in the precise manner defined before) the standard deviation of a distribution with the error (e.g., the uncertainty) in a *single* measurement of a random variable.

Once we have understood this, we should understand the expression

$$\hat{\mathbf{x}}(\xi) = \mu \pm \sigma \tag{1.88}$$

with $\mu = E[\hat{\mathbf{x}}]$ and $\sigma^2 = E[\hat{\mathbf{x}}^2] - E[\hat{\mathbf{x}}]^2$ as a short-hand notation of the exact statement of Chebyshev's theorem (1.86). It does not mean that experimental values $\hat{\mathbf{x}}(\xi)$ that differ from $\mu$ by an amount greater than $\sigma$ cannot appear, or are forbidden; it simply means that they are unlikely. How unlikely? Exactly by $1/k^2$, with $k = \frac{|\hat{\mathbf{x}}(\xi) - \mu|}{\sigma}$.

Chebyshev's theorem is very general. It applies to any random variable whatever its pdf. In most cases, however, the Gaussian distribution is a sufficient approximation to the true (maybe unknown) distribution. In the case of a Gaussian distribution, we have the more precise result

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \le k\,\sigma) = \mathrm{erf}\left(\frac{k}{\sqrt{2}}\right) > 1 - \frac{1}{k^2} \tag{1.89}$$

which takes the following values:

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \le \sigma) = 0.68269\ldots, \tag{1.90}$$

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \le 2\sigma) = 0.95450\ldots, \tag{1.91}$$

$$P(|\hat{\mathbf{x}}(\xi) - \mu| \le 3\sigma) = 0.99736\ldots \tag{1.92}$$

which means that we can be certain with a 68% probability that the result of the measurement will lie in the interval $(\mu - \sigma, \mu + \sigma)$; with a 95% probability in $(\mu - 2\sigma, \mu + 2\sigma)$; and with 99.7% probability in $(\mu - 3\sigma, \mu + 3\sigma)$. Note that, if we take $\sigma$ as the error of the measurement, in 32% (nearly 1/3 ) of the cases the observed value $\hat{\mathbf{x}}(\xi)$ will lie outside the error interval.

In most cases, one does not know the distribution function of the experiment, neither the mean $\mu$ nor the standard deviation $\sigma$. Chebyshev's theorem can be read in the inverse sense

$$\mu = \hat{\mathbf{x}}(\xi) \pm \sigma. \tag{1.93}$$

Given the result of a single measurement $\hat{\mathbf{x}}(\xi)$, this allows us to predict the value of $\mu$ within a certain interval of confidence which depends on the generally unknown standard deviation $\sigma$. However, it is clear that we cannot use this single measurement to obtain information about $\sigma$ (which is ultimately related to the

dispersion in a set of measurements). To obtain estimates for both $\mu$ and $\sigma$, we have to repeat the experiment $M$ times, each one independent of the others. We call, then, $\Xi = (\xi_1, \xi_2, \ldots, \xi_M)$ the result of the experiment which consists of $M$ independent repetitions and use some properties of the sum of random variables.

## 1.7
## Sums of Random Variables

Let $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2$, $\ldots, \hat{\mathbf{x}}_M$ be independent random variables, all of them described by the same pdf $f_{\hat{\mathbf{x}}}(x)$ with mean $\mu$ and variance $\sigma^2$. The natural idea is to consider them as independent repetitions of the same experiment. Associated with the result $\Xi = (\xi_1, \xi_2, \ldots, \xi_M)$, we define the random variables *sample mean* $\hat{\boldsymbol{\mu}}_M$ and *sample variance* $\hat{\boldsymbol{\sigma}}^2_M$:

$$\hat{\boldsymbol{\mu}}_M(\Xi) = \frac{1}{M} \sum_{i=1}^{M} \hat{\mathbf{x}}_i(\xi_i), \tag{1.94}$$

$$\begin{aligned} \hat{\boldsymbol{\sigma}}^2_M(\Xi) &= \frac{1}{M-1} \sum_{i=1}^{M} \left( \hat{\mathbf{x}}_i(\xi_i) - \hat{\boldsymbol{\mu}}_M \right)^2 \\ &= \frac{M}{M-1} \left( \frac{1}{M} \sum_{i=1}^{M} \hat{\mathbf{x}}_i(\xi_i)^2 - \left( \frac{1}{M} \sum_{i=1}^{M} \hat{\mathbf{x}}_i(\xi_i) \right)^2 \right). \end{aligned} \tag{1.95}$$

(Note that the notation stresses the fact that both random variables depend on the number of repetitions $M$.) It is easy to obtain the average of these two sample random variables[4] as

$$E[\hat{\boldsymbol{\mu}}_M] = E[\hat{\mathbf{x}}_i] = \mu, \tag{1.96}$$
$$E[\hat{\boldsymbol{\sigma}}^2_M] = \sigma^2. \tag{1.97}$$

Furthermore, the variance of the sample mean is given by

$$\sigma^2[\hat{\boldsymbol{\mu}}_M] = \frac{1}{M} \sigma^2. \tag{1.98}$$

If we now repeat the experiment $M$ times and obtain a value for $\hat{\boldsymbol{\mu}}_M(\Xi)$, we can use Chebyshev's theorem in its inverse short-hand notation (1.93) applied to the random variable $\hat{\boldsymbol{\mu}}_M$ to write $\mu = \hat{\boldsymbol{\mu}}_M(\Xi) \pm \sigma[\hat{\boldsymbol{\mu}}_M]$, or using (1.98)

$$\mu = \hat{\boldsymbol{\mu}}_M(\Xi) \pm \frac{\sigma}{\sqrt{M}}. \tag{1.99}$$

Still, we do not know the true value of $\sigma$ on the right-hand side of this equation. It seems intuitive, though, given (1.97), that we can replace it by the sample variance $\sigma \approx \hat{\boldsymbol{\sigma}}_M[\Xi]$, leading to the final result

$$\mu = \hat{\boldsymbol{\mu}}_M(\Xi) \pm \frac{\hat{\boldsymbol{\sigma}}_M[\Xi]}{\sqrt{M}} \tag{1.100}$$

---

4) The presence of the factor $M - 1$ in (1.95) avoids its presence in (1.97).

which yields an estimate of the average value $\mu$ together with its error. As discussed before, this error has to be interpreted in the statistical sense. There is some good news here. As the sum of $M$ independent random variables tends to a Gaussian distribution as $M$ increases, we can take the Gaussian confidence limits and conclude that in 68% of the cases the true value of $\mu$ will lie in the interval $\left( \hat{\boldsymbol{\mu}}_M(\Xi) - \frac{\hat{\boldsymbol{\sigma}}_M[\Xi]}{\sqrt{M}}, \hat{\boldsymbol{\mu}}_M(\Xi) + \frac{\hat{\boldsymbol{\sigma}}_M[\Xi]}{\sqrt{M}} \right)$, and so on.

If we worry about the replacement $\sigma \approx \hat{\boldsymbol{\sigma}}_M[\Xi]$ in the previous formulas, we can estimate the error of this replacement (again in a statistical sense) by applying Chebyshev's theorem to the random variable $\hat{\boldsymbol{\sigma}}_M$. In the limit of large $M$, we assume that $\sqrt{\sum_{i=1}^{M} \left( \frac{\hat{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_M}{\sigma} \right)^2}$, and hence $\sqrt{M-1}\,\hat{\boldsymbol{\sigma}}_M/\sigma$ can be approximated by a $\hat{\chi}(M)$ distribution with $M$ degrees of freedom. Using (1.63) and (1.64) and the result of Exercise 5 in the limit of large $M$, we conclude that $\hat{\boldsymbol{\sigma}}_M/\sigma$ follows a $\hat{\mathbf{G}}(1, 1/\sqrt{2M})$ Gaussian distribution or that $\hat{\boldsymbol{\sigma}}_M$ follows a $\hat{\mathbf{G}}(\sigma, \sigma/\sqrt{2M})$ distribution. Using again Chebyshev's theorem in its short-hand notation, we can write

$$\sigma = \hat{\boldsymbol{\sigma}}_M(\Xi) \pm \frac{\sigma}{\sqrt{2M}} \tag{1.101}$$

thereby justifying the replacement $\sigma \approx \hat{\boldsymbol{\sigma}}_M[\Xi]$, which is valid in the limit of large $M$. This result also leads to the standard deviation error estimation, as

$$\sigma = \hat{\boldsymbol{\sigma}}_M(\Xi) \pm \frac{\hat{\boldsymbol{\sigma}}_M(\Xi)}{\sqrt{2M}}. \tag{1.102}$$

## 1.8
## Conditional Probabilities

As the final ingredient in this brief summary of probability theory, we review now the concept of conditional probability. For the sake of simplicity, we will consider the case of two random variables $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, but similar ideas can be easily generalized in the case of more random variables.

The joint probability density $f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)$ is defined such that the probability that a measurement of the random variable $\hat{\mathbf{x}}$ *and* the random variable $\hat{\mathbf{y}}$ gives for each one of them a value in the interval $(x, x + dx)$ and $(y, y + dy)$, respectively, is

$$P(x < \hat{\mathbf{x}} \le x + dx, y < \hat{\mathbf{y}} \le y + dy) = f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dx\, dy. \tag{1.103}$$

The cdf

$$F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, p)\, dq\, dp \tag{1.104}$$

is such that

$$P(x_1 < \hat{\mathbf{x}} \le x_2, y_1 < \hat{\mathbf{y}} \le y_2) = \tag{1.105}$$
$$F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_2, y_2) - F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_1, y_2) - F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_2, y_1) + F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_1, y_1).$$

Some results follow straightforwardly from the definition:

$$\frac{\partial F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\partial x} = \int_{-\infty}^{y} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, p)\, dp,$$

(1.106)

$$\frac{\partial F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\partial y} = \int_{-\infty}^{x} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, y)\, dq,$$

(1.107)

$$\frac{\partial^2 F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\partial x \partial y} = f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y).$$

(1.108)

The marginal probabilities are

$$f_{\hat{\mathbf{x}}}(x) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dy,$$

(1.109)

$$f_{\hat{\mathbf{y}}}(y) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dx.$$

(1.110)

Let us recall the definition of conditional probability. For any two events $A$ and $B$ such that $P(B) \neq 0$, the conditional probability of $A$ given $B$ is defined as

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

(1.111)

This suggests the definition of the conditional distribution function

$$F_{\hat{\mathbf{y}}}(y|B) = P(\hat{\mathbf{y}} \leq y|B) = \frac{P(\hat{\mathbf{y}} \leq y, B)}{P(B)}$$

(1.112)

and the conditional density function

$$f_{\hat{\mathbf{y}}}(y|B) = \frac{\partial F_{\hat{\mathbf{y}}}(y|B)}{\partial y}.$$

(1.113)

In the particular case of the event $B = \{\hat{\mathbf{x}} \leq x\}$, we have

$$F_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} \leq x) = \frac{P(\hat{\mathbf{x}} \leq x, \hat{\mathbf{y}} \leq y)}{P(\hat{\mathbf{x}} \leq x)} = \frac{F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{F_{\hat{\mathbf{x}}}(x)}$$

(1.114)

and the pdf can be written as

$$f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} \leq x) = \frac{\partial F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)/\partial y}{F_{\hat{\mathbf{x}}}(x)} = \frac{\int_{-\infty}^{x} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, y)\, dq}{\int_{-\infty}^{\infty} \int_{-\infty}^{x} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, y)\, dq\, dy}.$$

(1.115)

If we now take $B = \{x_1 < \hat{\mathbf{x}} \leq x_2\}$, we get

$$F_{\hat{\mathbf{y}}}(y|x_1 < \hat{\mathbf{x}} \leq x_2) = \frac{P(x_1 < \hat{\mathbf{x}} \leq x_2, \hat{\mathbf{y}} \leq y)}{P(x_1 < \hat{\mathbf{x}} \leq x_2)}$$
$$= \frac{F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_2, y) - F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x_1, y)}{F_{\hat{\mathbf{x}}}(x_2) - F_{\hat{\mathbf{x}}}(x_1)}$$

(1.116)

and the pdf

$$f_{\hat{\mathbf{y}}}(y|x_1 < \hat{\mathbf{x}} \leq x_2) = \frac{\int_{x_1}^{x_2} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dx}{\int_{x_1}^{x_2} f_{\hat{\mathbf{x}}}(x)\, dx}.$$

(1.117)

Let us consider, finally, the set $B = \{\hat{\mathbf{x}} = x\}$ as the limit $x_1 \rightarrow x_2$ of the previous case. Consequently, we define

$$F_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x) = \lim_{\Delta x \rightarrow 0} F_{\hat{\mathbf{y}}}(y|x < \hat{\mathbf{x}} \leq x + \Delta x). \tag{1.118}$$

From (1.116), we obtain

$$\begin{aligned} F_{\hat{\mathbf{y}}}(y \mid \hat{\mathbf{x}} = x) &= \lim_{\Delta x \rightarrow 0} \frac{F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x + \Delta x, y) - F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{F_{\hat{\mathbf{x}}}(x + \Delta x) - F_{\hat{\mathbf{x}}}(x)} \\ &= \frac{\partial F_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)/\partial x}{d F_{\hat{\mathbf{x}}}(x)/dx} \end{aligned} \tag{1.119}$$

which can be expressed as

$$F_{\hat{\mathbf{y}}}(y \mid \hat{\mathbf{x}} = x) = \frac{\int_{-\infty}^{y} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, p)\, dp}{f_{\hat{\mathbf{x}}}(x)}. \tag{1.120}$$

By taking the derivative with respect to $x$, we obtain the conditional pdf as

$$f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x) = \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{f_{\hat{\mathbf{x}}}(x)} = \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dy}. \tag{1.121}$$

Exchanging the role of $x$ and $y$, we obtain

$$F_{\hat{\mathbf{x}}}(x \mid \hat{\mathbf{y}} = y) = \frac{\int_{-\infty}^{x} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(q, y)\, dq}{f_{\hat{\mathbf{y}}}(y)} \tag{1.122}$$

and

$$f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y) = \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{f_{\hat{\mathbf{y}}}(y)} = \frac{f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)}{\int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dx}. \tag{1.123}$$

For the sake of simplicity, and if no confusion can arise, we will shorten the notation of the four last defined functions to

$$F_{\hat{\mathbf{y}}}(y|x), \; f_{\hat{\mathbf{y}}}(y|x), \; F_{\hat{\mathbf{x}}}(x|y), \; f_{\hat{\mathbf{x}}}(x|y). \tag{1.124}$$

Recall that, by Bayes theorem, if $A$ and $B$ are events and $B_1, B_2, \ldots$ is a partition of $B$, that is, $B = \cup_i B_i$ and $B_i \cap B_j = \emptyset$, then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}. \tag{1.125}$$

We now rephrase an equivalent of Bayes theorem in terms of pdfs. It follows from (1.121) and (1.123) that

$$f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y) = f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x)f_{\hat{\mathbf{x}}}(x), \tag{1.126}$$

$$f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y) = f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y)f_{\hat{\mathbf{y}}}(y), \tag{1.127}$$

from which we obtain

$$f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x) = \frac{f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y)f_{\hat{\mathbf{y}}}(y)}{f_{\hat{\mathbf{x}}}(x)}. \tag{1.128}$$

We now use (1.109) and (1.127) to derive

$$f_{\hat{\mathbf{x}}}(x) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y) f_{\hat{\mathbf{y}}}(y)\, dy \tag{1.129}$$

which when replaced in the denominator of (1.128) yields

$$f_{\hat{\mathbf{y}}}(y|\hat{\mathbf{x}} = x) = \frac{f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y) f_{\hat{\mathbf{y}}}(y)}{\int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y) f_{\hat{\mathbf{y}}}(y)\, dy} \tag{1.130}$$

which is a version of Bayes theorem in terms of pdfs.

In the application of these formulas in the following chapters, we will consider the case in which one of the random variables, say $\hat{\mathbf{y}}$, takes only discrete values. This means that its pdf takes the form

$$f_{\hat{\mathbf{y}}}(y) = \sum_i \mathrm{Prob}(\hat{\mathbf{y}} = y_i)\delta(y - y_i), \tag{1.131}$$

and

$$f_{\hat{\mathbf{y}}}(y|x) = \sum_i \mathrm{Prob}(\hat{\mathbf{y}} = y_i|x)\delta(y - y_i). \tag{1.132}$$

Replacing both expressions in

$$f_{\hat{\mathbf{y}}}(y) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}\hat{\mathbf{y}}}(x, y)\, dx = \int_{-\infty}^{\infty} f_{\hat{\mathbf{y}}}(y|x) f_{\hat{\mathbf{x}}}(x)\, dx \tag{1.133}$$

we derive

$$\mathrm{Prob}(\hat{\mathbf{y}} = y_i) = \int_{-\infty}^{\infty} \mathrm{Prob}(\hat{\mathbf{y}} = y_i|x) f_{\hat{\mathbf{x}}}(x)\, dx. \tag{1.134}$$

Also, replacing in (1.129), we obtain

$$f_{\hat{\mathbf{x}}}(x) = \sum_i f_{\hat{\mathbf{x}}}(x|\hat{\mathbf{y}} = y_i)\mathrm{Prob}(\hat{\mathbf{y}} = y_i). \tag{1.135}$$

Formulas that will be of interest in later chapters.

## 1.9
## Markov Chains

As stated before, it is not difficult to generalize these concepts of joint probabilities to more than two random variables. For example, the pdf of $n$ random variables $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n$ can be written in terms of conditional probabilities as

$$f_{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N}(x_1, \ldots, x_n) = f_{\hat{\mathbf{x}}_1}(x_1) f_{\hat{\mathbf{x}}_2}(x_2|x_1) f_{\hat{\mathbf{x}}_3}(x_3|x_1, x_2) \ldots f_{\hat{\mathbf{x}}_n}(x_n|x_1, \ldots, x_{n-1}). \tag{1.136}$$

This complicated expression adopts a much simpler form for a particular kind of random variables. A succession of random variables $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n$, is called a *Markov chain* if for any value of $m = 1, \ldots, n$ it fulfills

$$f_{\hat{\mathbf{x}}_m}(x_m|x_1, \ldots, x_{m-1}) = f_{\hat{\mathbf{x}}_m}(x_m|x_{m-1}). \tag{1.137}$$

That is, the pdf of $\hat{\mathbf{x}}_m$ conditioned to $\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{m-1}$ is equal to the pdf of $\hat{\mathbf{x}}_m$ conditioned only to $\hat{\mathbf{x}}_{m-1}$. From this property, (1.136) simplifies to

$$f_{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_N}(x_1, \dots, x_n) = f_{\hat{\mathbf{x}}_n}(x_n | x_{n-1}) f_{\hat{\mathbf{x}}_{n-1}}(x_{n-1} | x_{n-2}) \dots f_{\hat{\mathbf{x}}_2}(x_2 | x_1) f_{\hat{\mathbf{x}}_1}(x_1). \quad (1.138)$$

Therefore, the joint pdf of $\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n$ is determined only by the knowledge of $f_{\hat{\mathbf{x}}_1}(x_1)$ and the conditional pdfs $f_{\hat{\mathbf{x}}_m}(x|y)$ (also known in this context as the transition pdf). We recall that $f_{\hat{\mathbf{x}}_m}(x|y)$ is actually a short way to write $f_{\hat{\mathbf{x}}_m}(x|\hat{\mathbf{x}}_{m-1} = y)$ with the meaning that $f_{\hat{\mathbf{x}}_m}(x|y)\,dx$ is the probability that the random variable $\hat{\mathbf{x}}_m$ adopts a value in the interval $(x, x + dx)$ provided that the random variable has taken the value $\hat{\mathbf{x}}_{m-1} = y$.

A Markov chain is called *homogeneous* if the transition probabilities $f_{\hat{\mathbf{x}}_m}(x|y)$ are independent of $m$. Thus, for a homogeneous Markov chain, we write the transition probabilities simply as $f(x|y)$.

It is easy to establish a relationship between $f_{\hat{\mathbf{x}}_{m+1}}(x)$ and $f_{\hat{\mathbf{x}}_m}(y)$ using the definition of conditional probability:

$$f_{\hat{\mathbf{x}}_{m+1}}(x) = \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}_{m+1}, \hat{\mathbf{x}}_m}(x, y)\,dy$$
$$= \int_{-\infty}^{\infty} f_{\hat{\mathbf{x}}_{m+1}}(x|y) f_{\hat{\mathbf{x}}_m}(y)\,dy, \quad m \geq 1 \quad (1.139)$$

which for a homogeneous chain reduces to

$$f_{\hat{\mathbf{x}}_{m+1}}(x) = \int_{-\infty}^{\infty} f(x|y) f_{\hat{\mathbf{x}}_m}(y)\,dy, \quad m \geq 1. \quad (1.140)$$

We can use this relation to *construct* the Markov chain. Starting from a given $f_{\hat{\mathbf{x}}_1}(x)$ initial pdf and a transition pdf $f(x|y)$, we can obtain the succession of random variables $\hat{\mathbf{x}}_m$, $m = 1, 2, \dots$ with the respective pdfs $f_{\hat{\mathbf{x}}_m}(x)$.

If the resulting pdfs $f_{\hat{\mathbf{x}}_m}(x)$ are all identical, $f_{\hat{\mathbf{x}}_m}(x) = f_{\hat{\mathbf{x}}}^{\text{st}}(x)$, $m = 1, 2, \dots$, we say that the Markov chain is stationary. For a stationary Markov chain, (1.140) becomes

$$f_{\hat{\mathbf{x}}}^{\text{st}}(x) = \int_{-\infty}^{\infty} f(x|y) f_{\hat{\mathbf{x}}}^{\text{st}}(y)\,dy. \quad (1.141)$$

It is not easy, in general, to solve the above integral equation to find the stationary pdf of a Markov chain with a given transition pdf $f(x|y)$. However, using

$$\int_{-\infty}^{\infty} f(y|x)\,dy = 1 \quad (1.142)$$

one can write (1.141) as

$$\int_{-\infty}^{\infty} f(y|x) f_{\hat{\mathbf{x}}}^{\text{st}}(x)\,dy = \int_{-\infty}^{\infty} f(x|y) f_{\hat{\mathbf{x}}}^{\text{st}}(y)\,dy \quad (1.143)$$

or, equivalently, as

$$\int_{-\infty}^{\infty} \left[ f(y|x) f_{\hat{\mathbf{x}}}^{\text{st}}(x) - f(x|y) f_{\hat{\mathbf{x}}}^{\text{st}}(y) \right]\,dy = 0. \quad (1.144)$$

A way to satisfy this equation is by requiring the *detailed balance* condition

$$f(y|x) f_{\hat{\mathbf{x}}}^{\text{st}}(x) = f(x|y) f_{\hat{\mathbf{x}}}^{\text{st}}(y). \quad (1.145)$$

This is a simpler functional equation for $f_{\hat{\mathbf{x}}}^{st}(x)$ than the integral (1.141). Any solution $f_{\hat{\mathbf{x}}}^{st}(x)$ of the detailed balance condition[5] will satisfy (1.141), but the reverse is not always true.

Certainly, if a pdf $f_{\hat{\mathbf{x}}}^{st}(x)$ satisfies (1.141), then it is a stationary solution of the recurrence relation (1.140) such that $f_{\hat{\mathbf{x}}_m}(x) = f_{\hat{\mathbf{x}}}^{st}(x)$, $\forall m$, provided that $f_{\hat{\mathbf{x}}_1}(x) = f_{\hat{\mathbf{x}}}^{st}(x)$.

What happens when $f_{\hat{\mathbf{x}}_1}(x) \neq f_{\hat{\mathbf{x}}}^{st}(x)$? Will the recurrence (1.140) converge toward the stationary solution $f_{\hat{\mathbf{x}}}^{st}(x)$? A partial, but important, answer can be formulated as follows: If for every point $x$ such that $f_{\hat{\mathbf{x}}}^{st}(x) > 0$ and for every initial condition $f_{\hat{\mathbf{x}}_1}(x)$, there exists a number $m$ of iterations such that $f_{\hat{\mathbf{x}}_m}(x) > 0$ (*irreducibility* condition) and the recurrence relation (1.140) does not get trapped in cyclic loops, then $f_{\hat{\mathbf{x}}}^{st}(x)$ is the unique stationary solution and, furthermore, $\lim_{m\to\infty} f_{\hat{\mathbf{x}}_m}(x) = f_{\hat{\mathbf{x}}}^{st}(x)$. These conditions (irreducibility and noncyclic behavior) are summarized by saying that the Markov chain is *ergodic*. The irreducibility condition has a simple intuitive interpretation. It states that, independent of the initial condition, the recurrence relation (1.140) does not have ''forbidden'' zones, meaning that it is able to provide eventually a pdf with a nonzero probability to any point $x$ such that $f_{\hat{\mathbf{x}}}^{st}(x) > 0$.

Finally, we can consider that the variable $m$ of the Markov chain represents, in some suitable units, a ''time.'' In this sense, (1.140) introduces a dynamics in the space of pdfs. We will often make use of this dynamical interpretation of a Markov chain in the rest of the book. This dynamical interpretation, which can also be framed in the theory of Markov processes, can be made clear by introducing a ''time'' variable $t = m\delta t$, where $\delta t$ is just the time unit. Then the function $f_{\hat{\mathbf{x}}_m}(x)$ becomes a two-variable function $f(x, t)$ and the evolution equation is given by

$$f_{\hat{\mathbf{x}}_{m+1}}(x) - f_{\hat{\mathbf{x}}_m}(x) = \int f(x|y) f_{\hat{\mathbf{x}}_m}(y)\, dy - f_{\hat{\mathbf{x}}_m}(x) \tag{1.146}$$

or using (1.142)

$$f_{\hat{\mathbf{x}}_{m+1}}(x) - f_{\hat{\mathbf{x}}_m}(x) = \int \left[ f(x|y) f_{\hat{\mathbf{x}}_m}(y) - f(y|x) f_{\hat{\mathbf{x}}_m}(x) \right] dy. \tag{1.147}$$

With the interpretation $f_{\hat{\mathbf{x}}_m}(x) \to f(x, t)$, we can write it as

$$f_{\hat{\mathbf{x}}}(x, t + \delta t) - f_{\hat{\mathbf{x}}}(x, t) = \int \left[ f(x|y) f_{\hat{\mathbf{x}}}(y, t) - f(y|x) f_{\hat{\mathbf{x}}}(x, t) \right] dy. \tag{1.148}$$

And, obviously, one is tempted to interpret the left-hand side as a partial derivative $\delta t \dfrac{\partial f(x, t)}{\partial t}$. We will explore this possibility in Chapter 8.

### Further Reading and References

The material of this chapter is rather standard, and there are many books that cover it in more detail than our brief summary. The book by Papoulis and Pillai [1] is

---

5) Note that the detailed balance condition might have no solution, as we require that $f_{\hat{\mathbf{x}}}^{st}(x)$ is a pdf, that is, nonnegative and normalizable.

a good general introduction. The book by Grimmett and Stirzaker [2] is also of a more advanced level and is particularly interesting for the topic of Markov chains.

**Exercises**

1) A rod of length $a$ is randomly dropped on a floor which has parallel lines drawn on it at a distance $\ell = 1$. Define what you think is meant by "randomly dropped" and, consequently, compute the probability that the rod intersects any of the lines (a problem due to Buffon).

2) A typist makes on average five mistakes in every 100 words. Find the probability that in a text of 1000 words the typist has made (a) exactly 10 mistakes and (b) at least 10 mistakes.

3) Use the Gaussian approximation to the Poisson distribution to find the probability that in a group of 10 000 people, at least 10 people were born on 1 January.

4) A study on the influence of the contraceptive pill on cervical cancer, published in the Lancet, **380**, 1740 (24 November 2007), analyzed a group of 52 082 women, 16 573 of which had taken the pill, and 35 509 have used other contraceptive methods. The study shows that the incidence of cervical cancer in the group of women that took the pill is 4.5 cases per 1000 women while in the group that did not take the pill it is 3.8 per 1000. Calculate the overall incidence rate in the group of 52 082 women. With this result, calculate the probability that a group of 16 573 women chosen randomly out of the 52 082 has a rate of 4.5 per 1000 or larger. Can we conclude that the pill increases the risk of cervical cancer, as was reported in some newspapers?

5) Prove that in the large $k$ limit, the mean value and variance of the $\hat{\chi}$ distribution, as given by (1.63) and (1.64) tend to $\langle \hat{\chi} \rangle = \sqrt{k}$ and $\sigma^2[\hat{\chi}] = 1/2$.

6) Prove that the correlation coefficient satisfies $|\rho[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j]| = 1$ if and only if there is a linear relationship $\hat{\mathbf{x}}_i = a\hat{\mathbf{x}}_j + b$ between the two random variables.

7) Compute the following integral:

$$L[J_1, J_2, \ldots, J_N] =$$
$$\sqrt{\frac{|A|}{(2\pi)^N}} \int_{\mathbb{R}^N} dx_1 \cdots dx_N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i A_{ij} x_j + \sum_{i=1}^{N} J_i x_i\right]$$

and use the result to prove Wick's theorem. Show, first, the equalities

$$\langle x_i \rangle = \left.\frac{\partial L}{\partial J_i}\right|_{\bar{J}=0}, \qquad \left\langle \prod_{k=1}^{n} x_{i_k} \right\rangle = \left.\frac{\partial^n L}{\prod_{k=1}^{n} \partial J_{i_k}}\right|_{\bar{J}=0}.$$

8) Prove that the sum of all symmetry factors in a Wick's diagram with $n$ branches is $(n-1)!!$.

9) Use Wick's theorem to compute $\langle x_1^3 x_2^3 x_3 x_4 \rangle$, where the $x_i$'s form a set a Gaussian variables of zero mean.

10) Let $x$ be a Gaussian random variable of zero mean and variance $\sigma^2$, and let $g(x)$ be an arbitrary function. Prove Novikov's theorem:

$$\langle xg(x)\rangle = \sigma^2 \langle g'(x)\rangle.$$

If $x = (x_1, \ldots, x_N)$ is a set of jointly Gaussian random variables, generalize the theorem to

$$\langle x_i g(x)\rangle = \sum_{k=1}^{N} \langle x_i x_k\rangle \left\langle \frac{\partial g(x)}{\partial x_k} \right\rangle$$

11) Prove that, for a Gaussian variable $x$ of mean $\mu$ and variance $\sigma^2$, it is $\langle e^{-x}\rangle = e^{-\mu + \sigma^2/2}$. For a general random variable, prove Jensen's inequality $\langle e^{-x}\rangle \geq e^{-\langle x\rangle}$.

12) Consider a homogeneous Markov chain with the following transition pdf:

$$f(x|y) = \frac{1}{\sigma\sqrt{2\pi(1-\lambda^2)}} e^{-\frac{(x-\lambda y)^2}{2\sigma^2(1-\lambda^2)}}$$

with $|\lambda| < 1$. Prove that it is irreducible (hint: simply consider $f_{\hat{x}_2}(x)$). Find its stationary pdf $f_{\hat{x}}^{\text{st}}(x)$ by searching for solutions of the detailed balance equation. Take as initial condition

$$f_{\hat{x}_1}(x) = \frac{1}{a\sqrt{2\pi}} e^{-\frac{x^2}{2a^2}},$$

compute $f_{\hat{x}_n}(x)$, and prove that, effectively, $\lim_{n\to\infty} f_{\hat{x}_n}(x) = f_{\hat{x}}^{\text{st}}(x)$.

13) Consider a homogeneous Markov chain with the following transition pdf:

$$f(x|y) = \begin{cases} e^{-(x-y)}, & x \geq y, \\ 0, & x < y. \end{cases}$$

Prove that it is not irreducible. Furthermore, show that there are no solutions $f_{\hat{x}}^{\text{st}}(x)$ to the detailed balance condition (1.145) or the recurrence relation (1.140).

14) Consider a homogeneous Markov chain with the following transition pdf:

$$f(x|y) = \frac{b}{\pi} \frac{1}{1 + b^2(x - \lambda y)^2}.$$

Show that it is irreducible. Prove that, if $|\lambda| < 1$, the stationary solution is

$$f_{\hat{x}}^{\text{st}}(x) = \frac{a}{\pi} \frac{1}{1 + a^2 x^2}$$

with $a = b(1 - \lambda)$; prove also that there are no solutions $f_{\hat{x}}^{\text{st}}(x)$ to the detailed balance condition (1.145) if $\lambda \neq 0$.

15) Consider a homogeneous Markov chain with the following transition pdf:

$$f(x|y) = \begin{cases} \frac{2x}{y}, & x \in (0, y), \\ \frac{2(1-x)}{1-y}, & x \in (y, 1) \end{cases}$$

if $y \in (0, 1)$ and $f(x|y) = 0$ otherwise. Show that is irreducible, and find its stationary distribution.