CHAPTER

# PROBABILITY AND STOCHASTIC PROCESSES

2

## CHAPTER OUTLINE

**Machine Learning. http://dx.doi.org/10.1016/B978-0-12-801522-3.00002-1**

## 2.1 INTRODUCTION

The goal of this chapter is to provide the basic definitions and properties related to probability theory and stochastic processes. It is assumed that the reader has attended a basic course on probability and statistics prior to reading this book. So, the aim is to help the reader refresh her/his memory and to establish a common language and a commonly understood notation.

Besides probability and random variables, random processes will be briefly reviewed and some basic theorems will be stated. Finally, at the end of the chapter, basic definitions and properties related to information theory will be summarized.

The reader who is familiar with all these notions can bypass this chapter.

## 2.2 PROBABILITY AND RANDOM VARIABLES

A random variable, x, is a variable whose variations are due to chance/randomness. A random variable can be considered as a function, which assigns a value to the outcome of an experiment. For example, in a coin tossing experiment, the corresponding random variable, x, can assume the values $x_1 = 0$ if the result of the experiment is “heads” and $x_2 = 1$ if the result is “tails.”

We will denote a random variable with a lower case roman, such as x, and the values it takes once an experiment has been performed, with mathmode italics, such as $x$.

A random variable is described in terms of a set of *probabilities* if its values are of a discrete nature, or in terms of a *probability density function* (pdf) if its values lie anywhere within an interval of the real axis (non-countably infinite set). For a more formal treatment and discussion, see [4, 6].

### 2.2.1 PROBABILITY

Although the words “probability” and “probable” are quite common in our everyday vocabulary, the mathematical definition of probability is not a straightforward one, and there are a number of different definitions that have been proposed over the years. Needless to say, whatever definition is adopted, the end result is that the properties and rules, which are derived, remain the same. Two of the most commonly used definitions are:

#### *Relative frequency definition*

The probability, $P(A)$, of an event, $A$, is the limit

$$P(A) = \lim_{n \longrightarrow \infty} \frac{n_A}{n}, \tag{2.1}$$

where $n$ is the number of total trials and $n_A$ the number of times event $A$ occurred. The problem with this definition is that in practice in any physical experiment, the numbers $n_A$ and $n$ can be large, yet they are always finite. Thus, the limit can only be used as a *hypothesis* and not as something that can be attained experimentally. In practice, often, we use

$$P(A) \approx \frac{n_A}{n} \tag{2.2}$$

for large values of $n$. However, this has to be used with caution, especially when the probability of an event is very small.

#### *Axiomatic definition*

This definition of probability is traced back to 1933 to the work of Andrey Kolmogorov, who found a close connection between probability theory and the mathematical theory of sets and functions of a real variable, in the context of measure theory, as noted in [5].

The probability, $P(A)$, of an event is a nonnegative number assigned to this event, or

$$P(A) \geq 0. \tag{2.3}$$

The probability of an event, $C$, which is certain to occur, equals to one

$$P(C) = 1. \tag{2.4}$$

If two events, $A$ and $B$, are mutually exclusive (they cannot occur simultaneously), then the probability of occurrence of either $A$ *or* $B$ (denoted as $A \cup B$) is given by

$$P(A \cup B) = P(A) + P(B). \tag{2.5}$$

It turns out that these three defining properties, which can be considered as the respective *axioms*, suffice to develop the rest of the theory. For example, it can be shown that the probability of an impossible event is equal to zero, as noted in [6].

The previous two approaches for defining probability are not the only ones. Another interpretation, which is in line with the way we are going to use the notion of probability in a number of places in this book in the context of *Bayesian learning*, has been given by Cox [2]. There, probability was seen as a measure of *uncertainty* concerning an event. Take, for example, the uncertainty whether the Minoan civilization was destroyed as a consequence of the earthquake that happened close to the island of Santorini. This is obviously not an event whose probability can be tested with repeated trials. However, putting together historical as well as scientific evidence, we can quantify our expression of uncertainty concerning such a conjecture. Also, we can modify the degree of our uncertainty once more historical evidence comes to light due to new archeological findings. Assigning numerical values to represent degrees of belief, Cox developed a set of axioms encoding common sense properties of such beliefs, and he came to a set of rules equivalent to the ones we are going to review soon; see also [4].

The origins of probability theory are traced back to the middle 17th century in the works of Pierre Fermat (1601-1665), Blaise Pascal (1623-1662), and Christian Huygens (1629-1695). The concepts of probability and the mean value of a random variable can be found there. The motivation for developing the theory is not related to any purpose for "serving society"; the purpose was to serve the needs of gambling and games of chance!

## 2.2.2 DISCRETE RANDOM VARIABLES

A discrete random variable, x, can take any value from a finite or *countably* infinite set $\mathcal{X}$. The probability of the event, "$\mathrm{x} = x \in \mathcal{X}$," is denoted as

$$P(\mathrm{x} = x) \quad \text{or simply } P(x). \tag{2.6}$$

The function $P(\cdot)$ is known as the *probability mass function* (pmf). Being a probability, it has to satisfy the first axiom, so $P(x) \geq 0$. Assuming that no two values in $\mathcal{X}$ can occur simultaneously and that after any experiment a single value will always occur, the second and third axioms combined give

$$\sum_{x \in \mathcal{X}} P(x) = 1. \tag{2.7}$$

The set $\mathcal{X}$ is also known as the *sample* or *state space*.

### Joint and conditional probabilities

The *joint probability* of two events, $A, B$, is the probability that both events occur simultaneously, and it is denoted as $P(A, B)$. Let us now consider two random variables, x, y, with sample spaces $\mathcal{X} = \{x_1, \ldots, x_{n_x}\}$ and $\mathcal{Y} = \{y_1, \ldots, y_{n_y}\}$, respectively. Let us adopt the relative frequency definition and assume that we carry out $n$ experiments and that each one of the values in $\mathcal{X}$ occurred $n_1^x, \ldots, n_{n_x}^x$ times and each one of the values in $\mathcal{Y}$ occurred $n_1^y, \ldots, n_{n_y}^y$ times, respectively. Then,

$$P(x_i) \approx \frac{n_i^x}{n}, \; i = 1, 2, \ldots, n_x, \quad \text{and} \quad P(y_j) \approx \frac{n_j^y}{n}, \; j = 1, 2, \ldots, n_y.$$

Let us denote by $n_{ij}$ the number of times the values $x_i$ and $y_j$ occurred simultaneously. Then, $P(x_i, y_j) \approx \frac{n_{ij}}{n}$. Simple reasoning dictates that the total number, $n_i^x$, that value $x_i$ occurred, is equal to

$$n_i^x = \sum_{j=1}^{n_y} n_{ij}. \tag{2.8}$$

Dividing both sides in the above by $n$, the following *sum rule* readily results.

$$\boxed{P(x) = \sum_{y \in \mathcal{Y}} P(x, y) : \quad \text{Sum Rule.}} \tag{2.9}$$

The *conditional probability* of an event, $A$, *given* another event, $B$, is denoted as $P(A|B)$ and it is defined as

$$\boxed{P(A|B) := \frac{P(A,B)}{P(B)} : \quad \text{Conditional Probability},} \tag{2.10}$$

provided $P(B) \neq 0$. It can be shown that this is indeed a probability, in the sense that it respects all three axioms [6]. We can better grasp its physical meaning if the relative frequency definition is adopted. Let $n_{AB}$ be the number of times that both events occurred simultaneously, and $n_B$ the times event $B$ occurred, out of $n$ experiments. Then, we have

$$P(A|B) = \frac{n_{AB}}{n} \frac{n}{n_B} = \frac{n_{AB}}{n_B}. \tag{2.11}$$

In other words, the conditional probability of an event, $A$, given another one, $B$, is the relative frequency that $A$ occurred, not with respect to the total number of experiments performed, but relative to the times event $B$ occurred.

Viewed differently and adopting similar notation in terms of random variables, in conformity with Eq. (2.9), the definition of the conditional probability is also known as the *product rule* of probability, written as

$$\boxed{P(x,y) = P(x|y)P(y) : \quad \text{Product Rule.}} \tag{2.12}$$

To differentiate from the joint and conditional probabilities, probabilities, $P(x)$ and $P(y)$ are known as *marginal probabilities*.

*Statistical Independence*: Two random variables are said to be statistically independent *if and only if* their joint probability is written as the product of the respective marginals,

$$P(x,y) = P(x)P(y). \tag{2.13}$$

### Bayes theorem

Bayes theorem is a direct consequence of the product rule and of the symmetry property of the joint probability, $P(x,y) = P(y,x)$, and it is stated as

$$\boxed{P(y|x) = \frac{P(x|y)P(y)}{P(x)} : \quad \text{Bayes Theorem},} \tag{2.14}$$

where the marginal, $P(x)$, can be written as

$$P(x) = \sum_{y \in \mathcal{Y}} P(x,y) = \sum_{y \in \mathcal{Y}} P(x|y)P(y),$$

and it can be considered as the normalizing constant of the numerator on the right-hand side in Eq. (2.14), which guarantees that summing up $P(y|x)$ with respect to all possible values of $y \in \mathcal{Y}$ results in one.

Bayes theorem plays a central role in machine learning, and it will be the basis for developing Bayesian techniques for estimating the values of unknown parameters.

### 2.2.3 CONTINUOUS RANDOM VARIABLES

So far, we have focused on discrete random variables. Our interest now turns to the extension of the notion of probability to random variables, which take values on the real axis, $\mathbb{R}$.

The starting point is to compute the probability of a random variable, $\mathrm{x}$, to lie in an interval, $x_1 < \mathrm{x} \leq x_2$. Note that the two events, $\mathrm{x} \leq x_1$ and $x_1 < \mathrm{x} \leq x_2$, are mutually exclusive. Thus, we can write that

$$P(\mathrm{x} \leq x_1) + P(x_1 < \mathrm{x} \leq x_2) = P(\mathrm{x} \leq x_2). \tag{2.15}$$

Define the *cumulative distribution function* (cdf) of $\mathrm{x}$, as

$$\boxed{F_{\mathrm{x}}(x) := P(\mathrm{x} \leq x): \quad \text{Cumulative Distribution Function.}} \tag{2.16}$$

Then, Eq. (2.15) can be written as

$$P(x_1 < \mathrm{x} \leq x_2) = F_{\mathrm{x}}(x_2) - F_{\mathrm{x}}(x_1). \tag{2.17}$$

Note that $F_{\mathrm{x}}$ is a monotonically increasing function. Furthermore, if it is continuous, the random variable $\mathrm{x}$ is said to be of a *continuous* type. Assuming that it is also differentiable, we can define the *pdf* (pdf) of $\mathrm{x}$ as

$$\boxed{p_{\mathrm{x}}(x) := \frac{dF_{\mathrm{x}}(x)}{dx}: \quad \text{Probability Density Function,}} \tag{2.18}$$

which then leads to

$$P(x_1 < \mathrm{x} \leq x_2) = \int_{x_1}^{x_2} p_{\mathrm{x}}(x)dx. \tag{2.19}$$

Also,

$$F_{\mathrm{x}}(x) = \int_{-\infty}^{x} p_{\mathrm{x}}(z)dz. \tag{2.20}$$

Using familiar logic from calculus arguments, the pdf can be interpreted as

$$\Delta P(x < \mathrm{x} \leq x + \Delta x) \approx p_{\mathrm{x}}(x)\Delta x, \tag{2.21}$$

which justifies its name as a "density" function, being the probability ($\Delta P$) of $\mathrm{x}$ lying in a small interval $\Delta x$, divided by the length of this interval. Note that as $\Delta x \longrightarrow 0$ this probability tends to zero. Thus, the probability of a continuous random variable taking any single value is zero. Moreover, since $P(-\infty < \mathrm{x} < +\infty) = 1$, we have

$$\int_{-\infty}^{+\infty} p_{\mathrm{x}}(x)dx = 1. \tag{2.22}$$

Usually, in order to simplify notation, the subscript $\mathrm{x}$ is dropped and we write $p(x)$, unless it is necessary for avoiding possible confusion. Note, also, that we have adopted the lower case "$p$" to denote a pdf and the capital "$P$" to denote a probability.

All previously stated rules for the probability are readily carried out for the case of pdfs, in the following way

$$p(x|y) = \frac{p(x,y)}{p(y)}, \quad p(x) = \int_{-\infty}^{+\infty} p(x,y)\,\mathrm{d}y. \tag{2.23}$$

### 2.2.4 MEAN AND VARIANCE

Two of the most common and useful quantities associated with any random variable are the respective mean value and variance. The mean value (or sometimes called expected value) is denoted as

$$\boxed{\mathbb{E}[\mathrm{x}] := \int_{-\infty}^{+\infty} xp(x)\,\mathrm{d}x: \quad \text{Mean Value,}} \tag{2.24}$$

where for discrete random variables the integration is replaced by summation $\left(\mathbb{E}[\mathrm{x}] = \sum_{x \in \mathcal{X}} xP(x)\right)$.

The variance is denoted as $\sigma_x^2$ and it is defined as

$$\boxed{\sigma_x^2 := \int_{-\infty}^{+\infty} (x - \mathbb{E}[\mathrm{x}])^2 p(x)\,\mathrm{d}x: \quad \text{Variance,}} \tag{2.25}$$

where integration is replaced by summation for discrete variables. The variance is a measure of the spread of the values of the random variable around its mean value.

The definition of the mean value is generalized for any function, $f(\mathrm{x})$, i.e.,

$$\mathbb{E}[f(\mathrm{x})] := \int_{-\infty}^{+\infty} f(x)p(x)dx. \tag{2.26}$$

It is readily shown that the mean value with respect to two random variables, y, x, can be written as the product

$$\mathbb{E}_{\mathrm{x,y}}[f(\mathrm{x},\mathrm{y})] = \mathbb{E}_{\mathrm{x}}\left[\mathbb{E}_{\mathrm{y}|\mathrm{x}}[f(\mathrm{x},\mathrm{y})]\right]. \tag{2.27}$$

This is a direct consequence of the definition of the mean value and the product rule of probability.

Given two random variables x, y, their *covariance* is defined as

$$\mathrm{cov}(\mathrm{x},\mathrm{y}) := \mathbb{E}[(\mathrm{x} - \mathbb{E}[\mathrm{x}])(\mathrm{y} - \mathbb{E}[\mathrm{y}])], \tag{2.28}$$

and their *correlation* as

$$r_{xy} := \mathbb{E}[\mathrm{xy}] = \mathrm{cov}(\mathrm{x},\mathrm{y}) + \mathbb{E}[\mathrm{x}]\,\mathbb{E}[\mathrm{y}]. \tag{2.29}$$

A *random vector* is a collection of random variables, $\mathbf{x} = [\mathrm{x}_1, \ldots, \mathrm{x}_l]^T$, and $p(\mathbf{x})$ is the joint pdf (probability for discrete variables),

$$p(\mathbf{x}) = p(\mathrm{x}_1, \ldots, \mathrm{x}_l). \tag{2.30}$$

The *covariance matrix* of a random vector, $\mathbf{x}$, is defined as

$$\boxed{\mathrm{Cov}(\mathbf{x}) := \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right]: \quad \text{Covariance Matrix,}} \tag{2.31}$$

or

$$\mathrm{Cov}(\mathbf{x}) = \begin{bmatrix} \mathrm{cov}(\mathrm{x}_1,\mathrm{x}_1) & \ldots & \mathrm{cov}(\mathrm{x}_1,\mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathrm{x}_l,\mathrm{x}_1) & \ldots & \mathrm{cov}(\mathrm{x}_l,\mathrm{x}_l) \end{bmatrix}. \tag{2.32}$$

Similarly, the *correlation matrix* of a random vector, $\mathbf{x}$, is defined as

$$\boxed{R_x := \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right] : \quad \text{Correlation Matrix},} \tag{2.33}$$

or

$$\begin{aligned} R_x &= \begin{bmatrix} \mathbb{E}[\mathrm{x}_1, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_1, \mathrm{x}_l] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\mathrm{x}_l, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_l, \mathrm{x}_l] \end{bmatrix} \\ &= \mathrm{Cov}(\mathbf{x}) + \mathbb{E}[\mathbf{x}]\, \mathbb{E}[\mathbf{x}^T]. \end{aligned} \tag{2.34}$$

Both the covariance and correlation matrices have a very rich structure, which will be exploited in various parts of this book to lead to computational savings whenever they are present in calculations. For the time being, observe that both are symmetric and positive semidefinite. The symmetry, $\Sigma = \Sigma^T$, is readily deduced from the definition. An $l \times l$ symmetric matrix, $A$, is called *positive semidefinite* if

$$\boldsymbol{y}^T A \boldsymbol{y} \geq 0, \quad \forall \boldsymbol{y} \in \mathbb{R}^l. \tag{2.35}$$

If the inequality is a strict one, the matrix is said to be *positive definite*. For the covariance matrix, we have

$$\boldsymbol{y}^T\, \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])\, (\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \boldsymbol{y} = \mathbb{E}\left[\left(\boldsymbol{y}^T\, (\mathbf{x} - \mathbb{E}[\mathbf{x}])\right)^2\right] \geq 0,$$

and the claim has been proved.

### Complex random variables

A complex random variable, $\mathrm{z} \in \mathbb{C}$, is a sum

$$\mathrm{z} = \mathrm{x} + j\mathrm{y}, \tag{2.36}$$

where $\mathrm{x}, \mathrm{y}$ are real random variables and $j := \sqrt{-1}$. Note that for complex random variables, the pdf *cannot* be defined since inequalities of the form, $\mathrm{x} + j\mathrm{y} \leq x + jy$, have no meaning. When we write $p(z)$, we mean the joint pdf of the real and imaginary parts, expressed as

$$p(z) := p(x, y). \tag{2.37}$$

For complex random variables, the notions of mean and covariance are defined as

$$\mathbb{E}[\mathrm{z}] := \mathbb{E}[\mathrm{x}] + j\, \mathbb{E}[\mathrm{y}], \tag{2.38}$$

and

$$\mathrm{cov}(\mathrm{z}_1, \mathrm{z}_2) := \mathbb{E}\left[(\mathrm{z}_1 - \mathbb{E}[\mathrm{z}_1])\, (\mathrm{z}_2 - \mathbb{E}[\mathrm{z}_2])^*\right], \tag{2.39}$$

where "*" denotes complex conjugation. The latter definition leads to the variance of a complex variable,

$$\sigma_z^2 = \mathbb{E}\left[|\mathrm{z} - \mathbb{E}[\mathrm{z}]|^2\right] = \mathbb{E}\left[|\mathrm{z}|^2\right] - |\mathbb{E}\,[\mathrm{z}]|^2\,. \tag{2.40}$$

Similarly, for complex random vectors, $\mathbf{z} = \mathbf{x} + j\mathbf{y} \in \mathbb{C}^l$, we have

$$p(z) := p(x_1, \dots, x_l, y_1, \dots, y_l), \tag{2.41}$$

where $x_i, y_i, i = 1, 2, \ldots, l$, are the components of the involved real vectors, respectively. The covariance and correlation matrices are similarly defined as

$$\mathrm{Cov}(\mathbf{z}) := \mathbb{E}\left[(\mathbf{z} - \mathbb{E}[\mathbf{z}])\,(\mathbf{z} - \mathbb{E}[\mathbf{z}])^H\right], \tag{2.42}$$

where "$H$" denotes the Hermitian (transposition and conjugation) operation.

For the rest of the chapter, we are going to deal mainly with real random variables. Whenever needed, differences with the case of complex variables will be stated.

### 2.2.5 TRANSFORMATION OF RANDOM VARIABLES

Let $\mathbf{x}$ and $\mathbf{y}$ be two random vectors, which are related via the vector transform,

$$\mathbf{y} = \boldsymbol{f}(\mathbf{x}), \tag{2.43}$$

where $\boldsymbol{f} : \mathbb{R}^l \longmapsto \mathbb{R}^l$ is an *invertible* transform. That is, given $\boldsymbol{y}$, then $\boldsymbol{x} = \boldsymbol{f}^{-1}(\boldsymbol{y})$ can be uniquely obtained. We are given the joint pdf, $p_{\mathrm{x}}(x)$, of $\mathbf{x}$ and the task is to obtain the joint pdf, $p_{\mathrm{y}}(y)$, of $\mathbf{y}$.

The Jacobian matrix of the transformation is defined as

$$J(\mathbf{y}; \mathbf{x}) := \frac{\partial(y_1, y_2, \ldots, y_l)}{\partial(x_1, x_2, \ldots, x_l)} := \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_l} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_l}{\partial x_1} & \cdots & \frac{\partial y_l}{\partial x_l} \end{bmatrix}. \tag{2.44}$$

Then, it can be shown (e.g., [6]) that

$$\boxed{p_{\mathbf{y}}(\boldsymbol{y}) = \left.\frac{p_{\mathbf{x}}(\boldsymbol{x})}{|\det(J(\mathbf{y}; \mathbf{x}))|}\right|_{\boldsymbol{x}=\boldsymbol{f}^{-1}(\boldsymbol{y})},} \tag{2.45}$$

where $|\det(\cdot)|$ denotes the absolute value of the determinant of a matrix. For real random variables, as in $\mathrm{y} = f(\mathrm{x})$, Eq. (2.45) simplifies to

$$p_{\mathrm{y}}(y) = \left.\frac{p_{\mathrm{x}}(x)}{|\frac{dy}{dx}|}\right|_{x=f^{-1}(y)}. \tag{2.46}$$

The latter can be graphically understood from Figure 2.1. The following two events have equal probabilities,

$$P(x < \mathrm{x} \leq x + \Delta x) = P(y + \Delta y < \mathrm{y} \leq y), \quad \Delta x > 0,\ \Delta y < 0.$$

Hence, by the definition of a pdf we have

$$p_{\mathrm{y}}(y)|\Delta y| = p_{\mathrm{x}}(x)|\Delta x|, \tag{2.47}$$

which leads to Eq. (2.46).

**Example 2.1.** Let us consider random vectors that are related via the linear transform,

$$\mathbf{y} = A\mathbf{x}, \tag{2.48}$$

where $A$ is invertible. Compute the joint pdf of $\mathbf{y}$ in terms of $p_{\mathbf{x}}(\boldsymbol{x})$.

The Jacobian of the transformation is easily computed and given by

$$J(\mathbf{y}; \mathbf{x}) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = A.$$

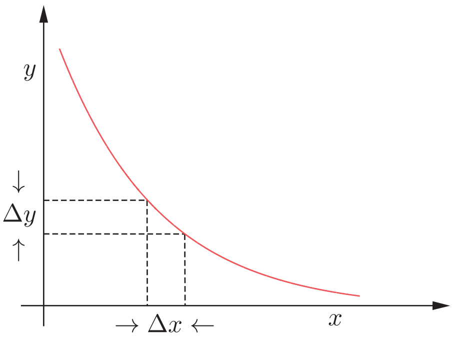

**FIGURE 2.1**

Note that by the definition of a pdf, $p_y(y)|\Delta y| = p_x(x)|\Delta x|$.

Hence,

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{p_{\mathbf{x}}(A^{-1}\mathbf{y})}{|\det(A)|}. \tag{2.49}$$

## 2.3 EXAMPLES OF DISTRIBUTIONS

In this section, some notable examples of distributions are provided. These are popular for modeling the random nature of variables met in a wide range of applications, and they will be used later in this book.

### 2.3.1 DISCRETE VARIABLES

#### *The Bernoulli distribution*

A random variable is said to be distributed according to a Bernoulli distribution if it is binary, $\mathcal{X} = \{0, 1\}$, with

$$P(\mathrm{x} = 1) = p, \;\; P(\mathrm{x} = 0) = 1 - p.$$

In a more compact way, we write $\mathrm{x} \sim \mathrm{Bern}(x|p)$ where

$$\boxed{P(x) = \mathrm{Bern}(x|p) := p^x(1-p)^{1-x}.} \tag{2.50}$$

Its mean value is equal to

$$\mathbb{E}[\mathrm{x}] = 1p + 0(1-p) = p \tag{2.51}$$

and its variance is equal to

$$\sigma_x^2 = (1-p)^2 p + p^2(1-p) = p(1-p). \tag{2.52}$$

#### *The Binomial distribution*

A random variable, x, is said to follow a binomial distribution with parameters $n, p$, and we write $\mathrm{x} \sim \mathrm{Bin}(x|n, p)$ if $\mathcal{X} = \{0, 1, \ldots, n\}$ and

$$\boxed{P(\mathrm{x}=k) := \mathrm{Bin}(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,\ldots,n,} \tag{2.53}$$

where by definition

$$\binom{n}{k} := \frac{n!}{(n-k)!k!}. \tag{2.54}$$

For example, this distribution models the times that heads occurs in $n$ successive trials, where $P(\text{Heads}) = p$. The binomial is a generalization of the Bernoulli distribution, which results if in Eq. (2.53) we set $n = 1$. The mean and variance of the binomial distribution are (Problem 2.1)

$$\mathbb{E}[\mathrm{x}] = np, \tag{2.55}$$

and

$$\sigma_x^2 = np(1-p). \tag{2.56}$$

Figure 2.2a shows the probability $P(k)$ as a function of $k$ for $p = 0.4$ and $n = 9$. Figure 2.2b shows the respective cumulative distribution. Observe that the latter has a staircase form, as is always the case for discrete variables.

### *The Multinomial distribution*

This is a generalization of the binomial distribution if the outcome of each experiment is not binary but can take one out of $K$ possible values. For example, instead of tossing a coin, a die with $K$ sides is thrown. Each one of the possible $K$ outcomes has probability $P_1, P_2, \ldots, P_K$, respectively, to occur, and we denote

$$\boldsymbol{P} = [P_1, P_2, \ldots, P_K]^T.$$

After $n$ experiments, assume that $x_1, x_2, \ldots, x_K$ times sides $\mathrm{x} = 1, \mathrm{x} = 2, \ldots, \mathrm{x} = K$ occurred, respectively. We say that the random (discrete) vector,

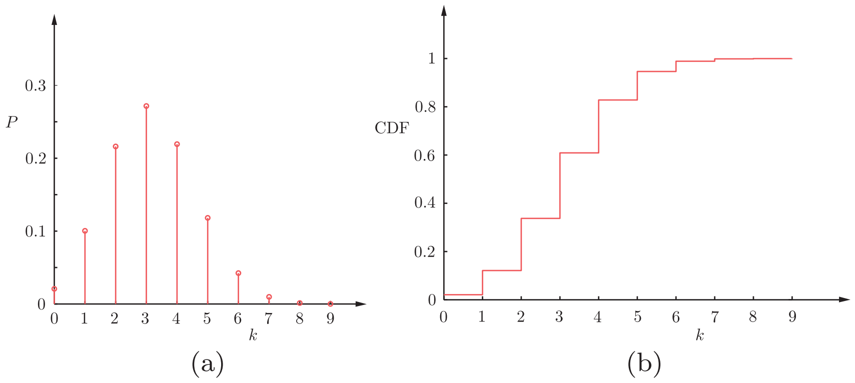

**FIGURE 2.2**

(a) The probability mass function (pmf) for the binomial distribution for $p = 0.4$ and $n = 9$. (b) The respective cumulative probability distribution (cdf). Since the random variable is discrete, the cdf has a staircase-like graph.

$$\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_K]^T, \tag{2.57}$$

follows a multinomial distribution, $\mathbf{x} \sim \mathrm{Mult}(\boldsymbol{x}|n, \boldsymbol{P})$, if

$$\boxed{P(\boldsymbol{x}) = \mathrm{Mult}(\boldsymbol{x}|n, \boldsymbol{P}) := \binom{n}{x_1, x_2, \ldots, x_K} \prod_{k=1}^{K} P_k^{x_k},} \tag{2.58}$$

where

$$\binom{n}{x_1, x_2, \ldots, x_K} := \frac{n!}{x_1! x_2! \ldots x_K!}.$$

Note that the variables, $\mathrm{x}_1, \ldots, \mathrm{x}_K$, are subject to the constraint

$$\sum_{k=1}^{K} \mathrm{x}_k = n,$$

and also

$$\sum_{k=1}^{K} P_K = 1.$$

The mean value, the variances, and the covariances are given by

$$\mathbb{E}[\mathbf{x}] = n\boldsymbol{P},\ \sigma_k^2 = nP_k(1 - P_k),\ k = 1, 2, \ldots, K,\ \mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j) = -nP_iP_j,\ i \neq j. \tag{2.59}$$

## 2.3.2 CONTINUOUS VARIABLES

### *The uniform distribution*

A random variable $\mathrm{x}$ is said to follow a *uniform* distribution in an interval $[a, b]$, and we write $\mathrm{x} \sim \mathcal{U}(a, b)$, with $a > -\infty$ and $b < +\infty$, if

$$\boxed{p(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}} \tag{2.60}$$

Figure 2.3 shows the respective graph. The mean value is equal to

$$\mathbb{E}[\mathrm{x}] = \frac{a+b}{2}, \tag{2.61}$$

and the variance is given by (Problem 2.2).

$$\sigma_x^2 = \frac{1}{12}(b-a)^2. \tag{2.62}$$

### *The Gaussian distribution*

The Gaussian or normal distribution is one among the most widely used distributions in all scientific disciplines. We say that a random variable, $\mathrm{x}$, is *Gaussian* or *normal* with parameters $\mu$ and $\sigma^2$, and we write $x \sim \mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(x|\mu, \sigma^2)$, if

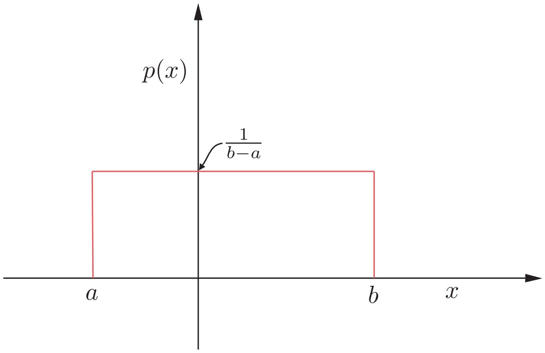

**FIGURE 2.3**

The pdf of a uniform distribution $\mathcal{U}(a, b)$.

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2.63}$$

It can be shown that the corresponding mean and variance are

$$\mathbb{E}[\mathrm{x}] = \mu \quad \text{and} \quad \sigma_x^2 = \sigma^2. \tag{2.64}$$

Indeed, by the definition of the mean value, we have that

$$\begin{aligned} \mathbb{E}[\mathrm{x}] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (y+\mu) \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y. \end{aligned} \tag{2.65}$$

Due to the symmetry of the exponential function, performing the integration involving $y$ gives zero and the only surviving term is due to $\mu$. Taking into account that a pdf integrates to one, we obtain the result.

To derive the variance, from the definition of the Gaussian pdf, we have that

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x = \sqrt{2\pi}\sigma. \tag{2.66}$$

Taking the derivative of both sides with respect to $\sigma$, we obtain

$$\int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{\sigma^3} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x = \sqrt{2\pi} \tag{2.67}$$

or

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x = \sigma^2, \tag{2.68}$$

which proves the claim.

Figure 2.4 shows the graph for two cases, $\mathcal{N}(x|1, 0.1)$ and $\mathcal{N}(x|1, 0.01)$. Both curves are symmetrically placed around the mean value $\mu = 1$. Observe that the smaller the variance is, the sharper around

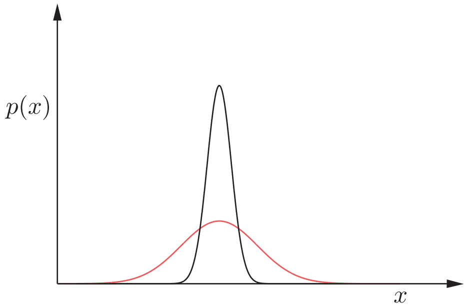

**FIGURE 2.4**

The graphs of two Gaussian pdfs for $\mu = 1$ and $\sigma^2 = 0.1$ (red) and $\sigma^2 = 0.01$ (gray).

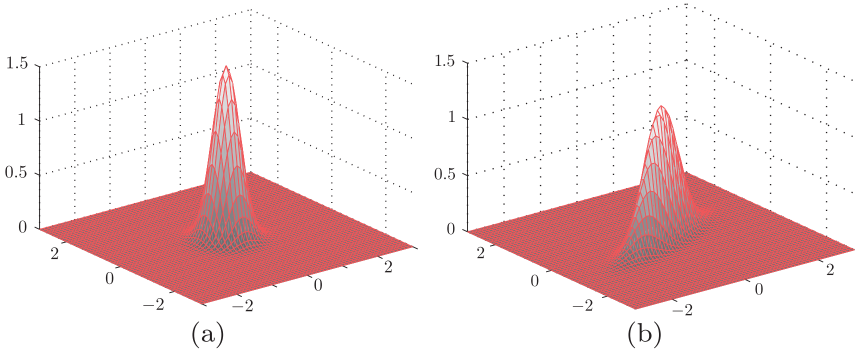

**FIGURE 2.5**

The graph of two two-dimensional Gaussian pdfs for $\boldsymbol{\mu} = \mathbf{0}$ and different covariance matrices. (a) The covariance matrix is diagonal with equal elements along the diagonal. (b) The corresponding covariance matrix is nondiagonal.

the mean value the pdf becomes. The generalization of the Gaussian to vector variables, $\mathbf{x} \in \mathbb{R}^l$, results in the so-called *multivariate Gaussian* or *normal* distribution, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ with parameters $\boldsymbol{\mu}$ and $\Sigma$, which is defined as

$$\boxed{p(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right): \quad \text{Gaussian pdf,}} \tag{2.69}$$

where $|\cdot|$ denotes the determinant of a matrix. It can be shown (Problem 2.3) that

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad \text{and} \quad \text{Cov}(\mathbf{x}) = \Sigma. \tag{2.70}$$

Figure 2.5 shows the two-dimensional normal pdf for two cases. Both share the same mean value, $\boldsymbol{\mu} = \mathbf{0}$, but they have different covariance matrices,

$$\Sigma_1 = \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 0.1 & 0.01 \\ 0.01 & 0.2 \end{bmatrix}. \tag{2.71}$$

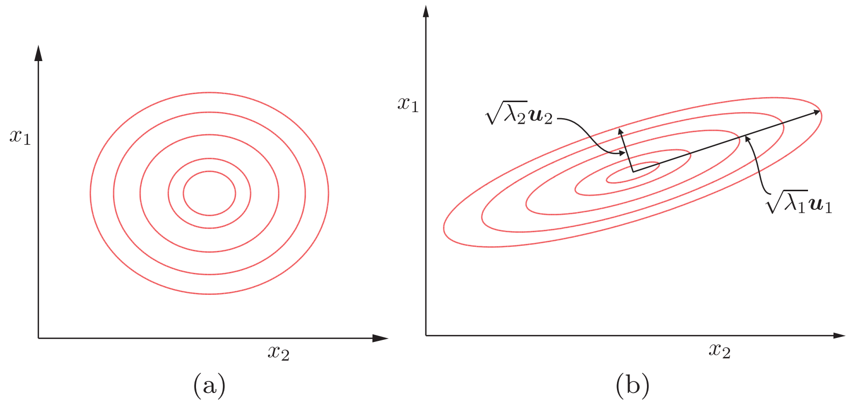
$x_1$
$x_2$
(a)
$x_1$
$\sqrt{\lambda_2}\boldsymbol{u}_2$
$\sqrt{\lambda_1}\boldsymbol{u}_1$
$x_2$
(b)

**FIGURE 2.6**

The isovalue contours for the two Gaussians of Figure 2.5. The contours for the Gaussian in Figure 2.5a are circles, while those corresponding to Figure 2.5b are ellipses. The major and minor axes of the ellipse are determined by the eigenvectors/eigenvalues of the respective covariance matrix, and they are proportional to $\sqrt{\lambda_1}c$ and $\sqrt{\lambda_2}c$, respectively. In the figure, they are shown for the case of $c = 1$. For the case of the diagonal matrix, with equal elements along the diagonal, all eigenvalues are equal, and the ellipse becomes a circle.

Figure 2.6 shows the corresponding isovalue contours for equal density values. In Figure 2.6a, the contours are circles, corresponding to the symmetric pdf in Figure 2.5a with covariance matrix $\Sigma_1$. The one shown in Figure 2.6b corresponds to the pdf in Figure 2.5b associated with $\Sigma_2$. Observe that, in general, the isovalue curves are ellipses/hyperellipsoids. They are centered at the mean value, and the orientation of the major axis as well their exact shape is controlled by the eigenstructure of the associated covariance matrix. Indeed, all points $\boldsymbol{x} \in \mathbb{R}^l$, which score the same density value, obey

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \text{constant} = c. \tag{2.72}$$

We know that the covariance matrix is *symmetric*, $\Sigma = \Sigma^T$. Thus, its eigenvalues are real and the corresponding eigenvectors can be chosen to form an orthonormal basis (Appendix A.2), which leads to its diagonalization,

$$\Sigma = U^T \Lambda U, \tag{2.73}$$

with

$$U := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l], \tag{2.74}$$

where $\boldsymbol{u}_i, i = 1, 2, \ldots, l$, are the orthonormal eigenvectors, and

$$\Lambda := \text{diag}\{\lambda_1, \ldots, \lambda_l\} \tag{2.75}$$

are the respective eigenvalues. We assume that $\Sigma$ is invertible, hence all eigenvalues are positive (being a positive definite it has positive eigenvalues, Appendix A.2). Due to the orthonormality of the eigenvectors, matrix $U$ is orthogonal as expressed in $UU^T = U^T U = I$. Thus, Eq. (2.72) can now be written as

$$\boldsymbol{y}^T \Lambda^{-1} \boldsymbol{y} = c, \tag{2.76}$$

where we have used the linear transformation

$$\boldsymbol{y} := U(\boldsymbol{x} - \boldsymbol{\mu}), \tag{2.77}$$

which corresponds to a rotation of the axes by $U$ and a translation of the origin to $\boldsymbol{\mu}$. Equation (2.76) can be written as

$$\frac{y_1^2}{\lambda_1} + \cdots + \frac{y_l^2}{\lambda_l} = c, \tag{2.78}$$

where it can be readily observed that it is an equation describing a (hyper)ellipsoid in the $\mathbb{R}^l$. From Eq. (2.77), it is easily seen that it is centered at $\boldsymbol{\mu}$ and that the major axes of the ellipsoid are parallel to $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l$ (plug in place of $\boldsymbol{x}$ the standard basis vectors, $[1, 0, \ldots, 0]^T$, etc.). The size of the respective axes are controlled by the corresponding eigenvalues. This is shown in Figure 2.6b. For the special case of a diagonal covariance with equal elements across the diagonal, all eigenvalues are equal to the value of the common diagonal element and the ellipsoid becomes a (hyper)sphere (circle).

The Gaussian pdf has a number of nice properties, which we are going to discover as we move on in this book. For the time being, note that if the covariance matrix is diagonal,

$$\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_l^2\},$$

that is, when the covariance of all the elements $\text{cov}(\mathrm{x}_i, \mathrm{x}_j) = 0, i, j = 1, 2, \ldots, l$, then the random variables comprising $\mathbf{x}$ are statistically *independent*. In general, this is not true. Uncorrelated variables are not necessarily independent; independence is a much stronger condition. This is true, however, if they follow a multivariate Gaussian. Indeed, if the covariance matrix is diagonal, then the multivariate Gaussian is written as

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right). \tag{2.79}$$

In other words,

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} p(x_i), \tag{2.80}$$

which is the condition for statistical independence.

### The central limit theorem

This is one of the most fundamental theorems in probability theory and statistics and it partly explains the popularity of the Gaussian distribution. Consider $N$ mutually *independent* random variables, each following its own distribution with mean values $\mu_i$ and variances $\sigma_i^2, i = 1, 2, \ldots, N$. Define a new random variable as their sum,

$$\mathrm{x} = \sum_{i=1}^{N} \mathrm{x}_i. \tag{2.81}$$

Then the mean and variance of the new variable are given by

$$\mu = \sum_{i=1}^{N} \mu_i, \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{N} \sigma_i^2. \tag{2.82}$$

It can be shown (e.g., [4, 6]) that as $N \longrightarrow \infty$ the distribution of the normalized variable

$$\mathrm{z} = \frac{\mathrm{x} - \mu}{\sigma} \tag{2.83}$$

tends to the *standard* normal distribution, and for the corresponding pdf we have

$$p(z) \xrightarrow[N \to \infty]{} \mathcal{N}(z|0, 1). \tag{2.84}$$

In practice, even summing up a relatively small number, $N$, of random variables, one can obtain a good approximation to a Gaussian. For example, if the individual pdfs are smooth enough and each random variable is *independent and identically distributed* (i.i.d.), a number $N$ between 5 and 10 can be sufficient.

### The exponential distribution

We say that a random variable follows an exponential distribution with parameter $\lambda > 0$, if

$$\boxed{p(x) = \begin{cases} \lambda \exp\left(-\lambda x\right), & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}} \tag{2.85}$$

The distribution has been used, for example, to model the time between arrivals of telephone calls or of a bus at a bus stop. The mean and variance can be easily computed by following simple integration rules, and they are

$$\mathbb{E}[\mathrm{x}] = \frac{1}{\lambda}, \quad \sigma_x^2 = \frac{1}{\lambda^2}. \tag{2.86}$$

### The beta distribution

We say that a random variable, $\mathrm{x} \in [0, 1]$, follows a beta distribution with positive parameters, $a, b$, and we write, $\mathrm{x} \sim \mathrm{Beta}(x|a, b, )$, if

$$\boxed{p(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & \text{if } 0 \leq x \leq 1, \\ 0, & \text{otherwise,} \end{cases}} \tag{2.87}$$

where $B(a, b)$ is the beta function, defined as

$$B(a, b) := \int_0^1 x^{a-1}(1-x)^{b-1}\, dx. \tag{2.88}$$

The mean and variance of the beta distribution are given by (Problem 2.4)

$$\mathbb{E}[\mathrm{x}] = \frac{a}{a+b}, \quad \sigma_x^2 = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2.89}$$

Moreover, it can be shown (Problem 2.5) that

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{2.90}$$

where $\Gamma(\cdot)$ is the gamma functions defined as

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x}\, dx. \tag{2.91}$$

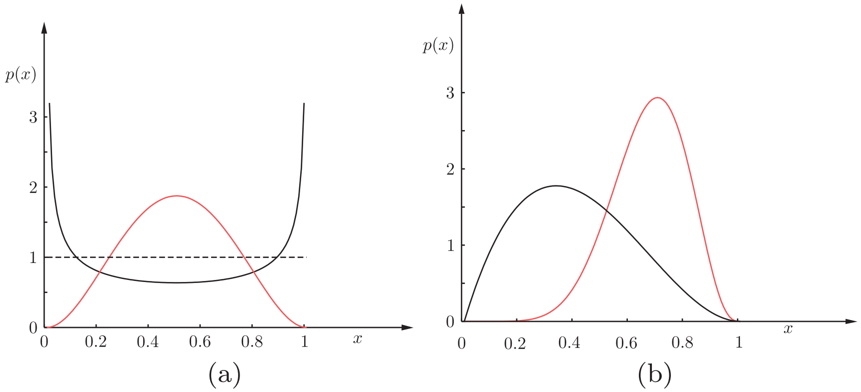

**FIGURE 2.7**

The graphs of the pdfs of the Beta distribution for different values of the parameters. (a) The dotted line corresponds to $a = 1$, $b = 1$, the gray line to $a = 0.5$, $b = 0.5$, and the red one to $a = 3$, $b = 3$. (b) The gray line corresponds to $a = 2$, $b = 3$, and the red one to $a = 8$, $b = 4$. For values $a = b$, the shape is symmetric around $1/2$. For $a < 1$, $b < 1$, it is convex. For $a > 1$, $b > 1$, it is zero at $x = 0$ and $x = 1$. For $a = 1 = b$, it becomes the uniform distribution. If $a < 1$, $p(x) \longrightarrow \infty$, $x \longrightarrow 0$ and if $b < 1$, $p(x) \longrightarrow \infty$, $x \longrightarrow 1$.

The beta distribution is very flexible and one can achieve various shapes by changing the parameters $a, b$. For example, if $a = b = 1$, the uniform distribution results. If $a = b$, the pdf has a symmetric graph around $1/2$. If $a > 1, b > 1$ then $p(x) \longrightarrow 0$ both at $x = 0$ and $x = 1$. If $a < 1$ and $b < 1$, it is convex with a unique minimum. If $a < 1$, it tends to $\infty$ as $x \longrightarrow 0$, and if $b < 1$, it tends to $\infty$ for $x \longrightarrow 1$. Figures 2.7a and b show the graph of the beta distribution for different values of the parameters.

### The gamma distribution

A random variable follows the gamma distribution with positive parameters $a, b$, and we write $\mathrm{x} \sim \mathrm{Gamma}(x|a, b)$ if

$$\boxed{p(x) = \begin{cases} \dfrac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, & x > 0, \\ 0, & \text{otherwise}. \end{cases}} \tag{2.92}$$

The mean and variance are given by

$$\mathbb{E}[\mathrm{x}] = \frac{a}{b}, \quad \sigma_x^2 = \frac{a}{b^2}. \tag{2.93}$$

The gamma distribution also takes various shapes by varying the parameters. For $a < 1$, it is strictly decreasing and $p(x) \longrightarrow \infty$ as $x \longrightarrow 0$ and $p(x) \longrightarrow 0$ as $x \longrightarrow \infty$. Figure 2.8 shows the resulting graphs for various values of the parameters.

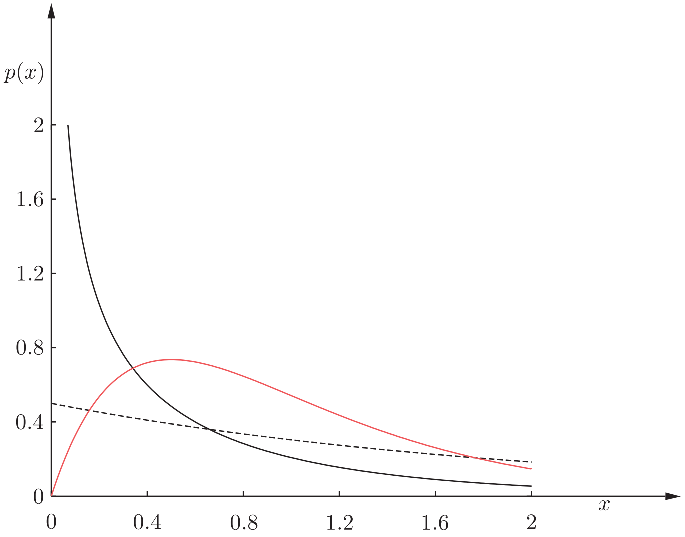

**FIGURE 2.8**

The pdf of the gamma distribution takes different shapes for the various values of the parameters: $a = 0.5$, $b = 1$ (full line gray), $a = 2$, $b = 0.5$ (red), $a = 1$, $b = 2$ (dotted).

*Remarks 2.1.*

- Setting in the gamma distribution $a$ to be an integer (usually $a = 2$), the *Erlang* distribution results. This distribution is being used to model waiting times in queueing systems.
- The *chi-squared* is also a special case of the gamma distribution, and it is obtained if we set $b = 1/2$ and $a = \nu/2$. The chi-squared distribution results if we sum up $\nu$ squared normal variables.

### *The Dirichlet distribution*

The Dirichlet distribution can be considered as the multivariate generalization of the beta distribution. Let $\mathbf{x} = [\mathrm{x}_1, \ldots, \mathrm{x}_K]^T$ be a random vector, with components such as

$$0 \leq \mathrm{x}_k \leq 1, \quad k = 1, 2, \ldots, K, \quad \text{and} \quad \sum_{k=1}^{K} \mathrm{x}_k = 1. \tag{2.94}$$

In other words, the random variables lie on $(K-1)$-dimensional *simplex*, Figure 2.9. We say that the random vector $\mathbf{x}$ follows a Dirichlet distribution with parameters $\boldsymbol{a} = [a_1, \ldots, a_K]^T$, and we write $\mathbf{x} \sim \text{Dir}(\boldsymbol{x}|\boldsymbol{a})$, if

$$\boxed{p(\boldsymbol{x}) = \text{Dir}(\boldsymbol{x}|\boldsymbol{a}) := \frac{\Gamma(\bar{a})}{\Gamma(a_1)\ldots\Gamma(a_K)} \prod_{k=1}^{K} x_k^{a_k - 1},} \tag{2.95}$$

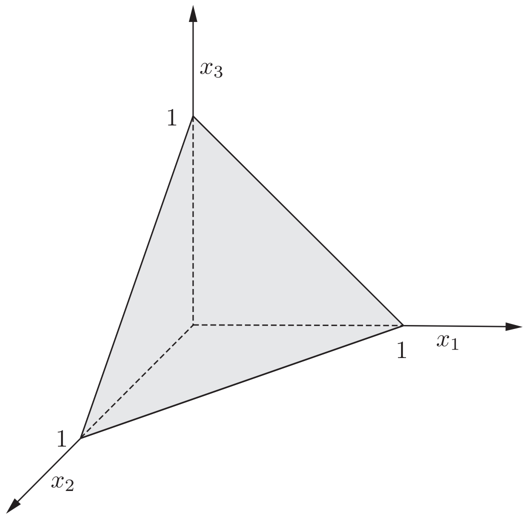

**FIGURE 2.9**

The 2-dimensional simplex in $\mathbb{R}^3$.

where

$$\bar{a} = \sum_{k=1}^{K} a_k. \tag{2.96}$$

The mean, variance, and covariances of the involved random variables are given by (Problem 2.7),

$$\mathbb{E}[\mathbf{x}] = \frac{1}{\bar{a}}\boldsymbol{a}, \quad \sigma_k^2 = \frac{a_k(\bar{a}-a_k)}{\bar{a}^2(\bar{a}+1)}, \quad \text{cov}(\mathrm{x}_i, \mathrm{x}_j) = -\frac{a_i a_j}{\bar{a}^2(\bar{a}+1)}. \tag{2.97}$$

Figure 2.10 shows the graph of the Dirichlet distribution for different values of the parameters, over the respective 2D-simplex.

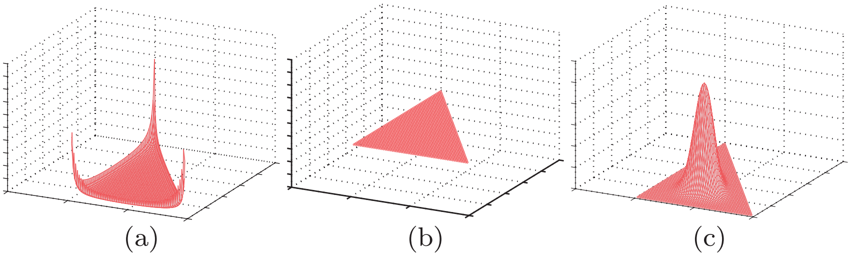

**FIGURE 2.10**

The Dirichlet distribution over the 2D-simplex for (a) (0.1,0.1,0.1), (b) (1,1,1), and (c) (10,10,10).

## 2.4 STOCHASTIC PROCESSES

The notion of a random variable has been introduced to describe the result of a random experiment whose outcome is a single value, as occurs, heads or tails in a coin-tossing experiment, or a value between one and six when throwing the die in a backgammon game.

In this section, the notion of a *stochastic process* is introduced to describe random experiments where the outcome of each experiment is a function or a sequence; in other words, the outcome of each experiment is an infinite number of values. In this book, we are only going to be concerned with stochastic processes associated with sequences. Thus, the result of a random experiment is a *sequence*, $u_n$ (or sometimes denoted as $u(n)$), $n \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers. Usually, $n$ is interpreted as a time index, and $u_n$ is called a *time series*, or in signal processing jargon, a *discrete-time signal*. In contrast, if the outcome is a function, $u(t)$, it is called a *continuous-time* signal. We are going to adopt the time interpretation of the free variable, $n$, for the rest of the chapter, without harming generality.

When discussing random variables, we used the notation $\mathrm{x}$ to denote the random variable, which assumes a value, $x$, from the sample space once an experiment is performed. Similarly, we are going to use $u_n$ to denote the specific sequence resulting from a single experiment and the roman font, $\mathrm{u}_n$, to denote the corresponding *discrete-time* random process, that is, the rule that assigns a specific sequence as the outcome of an experiment. A stochastic process can be considered as a family or *ensemble* of sequences. The individual sequences are known as *sample sequences* or simply as *realizations*.

For our notational convention, in general, we are going to reserve different symbols for processes and random variables. We have already used the symbol $\mathrm{u}$ and not $\mathrm{x}$; this is only for pedagogical reasons, just to make sure that the reader readily recognizes when the focus is on random variables and when it is on random processes. In signal processing jargon, a stochastic process is also known as a *random signal*. Figure 2.11 illustrates the fact that the outcome of an experiment involving a stochastic process is a sequence of values.

Note that fixing the time to a specific value, $n = n_0$, makes $\mathrm{u}_{n_0}$ a random variable. Indeed, for each random experiment we perform, a single value results at time instant $n_0$. From this perspective, a random process can be considered the collection of infinite random variables, $\{\mathrm{u}_n, n \in \mathbb{Z}\}$. So, is there a need to study a stochastic process separate from random variables/vectors? The answer is yes, and the reason is that we are going to allow certain time dependencies among the random variables, corresponding to different time instants, and study the respective effect on the time evolution of the random process. Stochastic processes will be considered in Chapter 5, where the underlying time dependencies will be exploited for computational simplifications, and in Chapter 13 in the context of Gaussian processes.

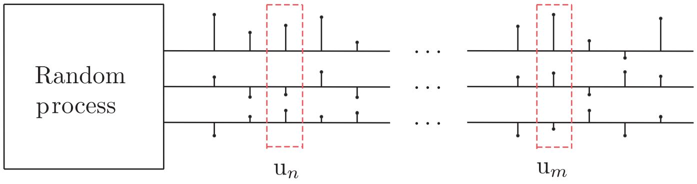

**FIGURE 2.11**

The outcome of each experiment, associated with a *discrete-time* stochastic process, is a *sequence* of values. For each one of the realizations, the corresponding values obtained at any instant (e.g., $n$ or $m$) comprise the outcomes of a corresponding random variable, $\mathrm{u}_n$ or $\mathrm{u}_m$, respectively.

### 2.4.1 FIRST AND SECOND ORDER STATISTICS

For a stochastic process to be fully described, one must know the joint pdfs (pmfs for discrete-valued random variables)

$$p(u_n, u_m, \ldots, u_r; n, m, \ldots, r), \tag{2.98}$$

for all possible combinations of random variables, $\mathrm{u}_n, \mathrm{u}_m, \ldots, \mathrm{u}_r$. Note that, in order to emphasize it, we have explicitly denoted the dependence of the joint pdfs on the involved time instants. However, from now on, this will be suppressed for notational convenience. Most often, in practice, and certainly in this book, the emphasis is on computing first and second order statistics only, based on $p(u_n)$ and $p(u_n, u_m)$. To this end, the following quantities are of particular interest:

*Mean at Time $n$*:

$$\mu_n := \mathbb{E}[\mathrm{u}_n] = \int_{-\infty}^{+\infty} u_n p(u_n) du_n. \tag{2.99}$$

*Autocovariance at Time Instants*, $n, m$:

$$\operatorname{cov}(n, m) := \mathbb{E}\left[\left(\mathrm{u}_n - \mathbb{E}[\mathrm{u}_n]\right)\left(\mathrm{u}_m - \mathbb{E}[\mathrm{u}_m]\right)\right]. \tag{2.100}$$

*Autocorrelation at Time Instants*, $n, m$:

$$r(n, m) := \mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m\right]. \tag{2.101}$$

We refer to these mean values as *ensemble* averages to stress that they convey statistical information over the ensemble of sequences that comprise the process.

The respective definitions for complex stochastic processes are

$$\operatorname{cov}(n, m) = \mathbb{E}\left[\left(\mathrm{u}_n - \mathbb{E}[\mathrm{u}_n]\right)\left(\mathrm{u}_m - \mathbb{E}[\mathrm{u}_m]\right)^*\right], \tag{2.102}$$

and

$$r(n, m) = \mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m^*\right]. \tag{2.103}$$

### 2.4.2 STATIONARITY AND ERGODICITY

**Definition 2.1** (*Strict-Sense Stationarity*)**.** A stochastic process, $\mathrm{u}_n$, is said to be *strict-sense stationary* (SSS) if its statistical properties are invariant to a shift of the origin, or if $\forall k \in \mathbb{Z}$

$$p(u_n, u_m, \ldots, u_r) = p(u_{n-k}, u_{m-k}, \ldots, u_{r-k}), \tag{2.104}$$

and for *any* possible combination of time instants, $n, m, \ldots, r \in \mathbb{Z}$.

In other words, the stochastic processes $\mathrm{u}_n$ and $\mathrm{u}_{n-k}$ are described by the same joint pdfs of all orders. A weaker version of stationarity is that of the *m*th order stationarity, where joint pdfs involving up to $m$ variables are invariant to the choice of the origin. For example, for a second order ($m = 2$) stationary process, we have that $p(u_n) = p(u_{n-k})$ and $p(u_n, u_r) = p(u_{n-k}, u_{r-k})$, $\forall n, r, k \in \mathbb{Z}$.

**Definition 2.2** (*Wide-Sense Stationarity*)**.** A stochastic process, $\mathrm{u}_n$, is said to be *wide-sense stationary* (WSS) if the mean value is constant over all time instants and the autocorrelation/autocovariance sequences depend on the difference of the involved time indices, or

$$\mu_n = \mu, \quad \text{and} \quad r(n, n-k) = r(k). \tag{2.105}$$

Note that WSS is a weaker version of the second order stationarity; in the latter case, all possible second order statistics are independent of the time origin. In the former, we only require the autocorrelation (autocovariance) and the mean value to be independent of the time origin. The reason we focus on these two quantities (statistics) is that they are of major importance in the study of linear systems and in the mean-square estimation, as we will see in Chapter 4.

Obviously, a strict-sense stationary process is also wide-sense stationary but, in general, not the other way around. For wide-sense stationary processes, the autocorrelation becomes a *sequence* with a *single* time index as the free parameter; thus its value, which measures a relation of the variables at two time instants, depends *solely on how much these time instants differ*, and not on their specific values.

From our basic statistics course, we know that given a random variable, x, its mean value can be approximated by the sample mean. Carrying out $N$ successive independent experiments, let $x_n, n = 1, 2, \ldots, N$, be the obtained values, known as *observations*. The *sample mean* is defined as

$$\hat{\mu}_N := \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{2.106}$$

For large enough values of $N$, we expect the sample mean to be close to the true mean value, $\mathbb{E}[\mathrm{x}]$. In a more formal way, this is guaranteed by the fact that $\hat{\mu}_N$ is associated with an *unbiased* and *consistent* estimator. We will discuss such issues in Chapter 3; however, we can refresh our memory at this point. Every time we repeat the $N$ random experiments, different samples result and hence a different estimate $\hat{\mu}_N$ is computed. Thus, the values of the estimates define a new random variable, $\hat{\mathrm{\mu}}_n$, known as the estimator. This is unbiased, because it can easily be shown that

$$\mathbb{E}[\hat{\mathrm{\mu}}_N] = \mathbb{E}[\mathrm{x}], \tag{2.107}$$

and it is consistent because its variance tends to zero as $N \longrightarrow +\infty$ (Problem 2.8). These two properties guarantee that, with high probability, for large values of $N$, $\hat{\mu}_N$ will be close to the true mean value.

To apply the concept of sample mean approximation to random processes, one must have at her/his disposal a number of $N$ realizations, and compute the sample mean at different time instants "across the process," using *different* realizations, representing the ensemble of sequences. Similarly, sample mean arguments can be used to approximate the autocovariance/autocorrelation sequences. However, this is a costly operation, since now each experiment results in an infinite number of values (a sequence of values). Moreover, it is common in practical applications that only one realization is available to the user.

To this end, we will now define a special type of stochastic processes, where the sample mean operation can be significantly simplified.

**Definition 2.3** (*Ergodicity*)**.** A stochastic process is said to be *ergodic* if the complete statistics can be determined by any one of the realizations.

In other words, if a process is ergodic, every single realization carries identical statistical information and it can describe the entire random process. Since from a single sequence only one set of pdfs can be obtained, we conclude that *every ergodic process is necessarily stationary*. A nonstationary process has infinite sets of pdfs, depending upon the choice of the origin. For example, there is only one mean value that can result from a single realization and be obtained as a (time) average, over the values of the sequence. Hence, the mean value of a stochastic process that is ergodic must be constant for all time instants, or independent of the time origin. The same is true for all higher order statistics.

A special type of ergodicity is that of the *second order* ergodicity. This means that only statistics up to a second order can be obtained from a single realization. Second order ergodic processes are necessarily wide-sense stationary. For second order ergodic processes, the following are true:

$$\mathbb{E}[\mathrm{u}_n] = \mu = \lim_{N\to\infty} \hat{\mu}_N, \tag{2.108}$$

where

$$\hat{\mu}_N := \frac{1}{2N+1} \sum_{n=-N}^{N} u_n.$$

Also,

$$\mathrm{cov}(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} (u_n - \mu)(u_{n-k} - \mu), \tag{2.109}$$

where both limits are in the mean-square sense; that is,

$$\lim_{N\to\infty} \mathbb{E}\left[ |\hat{\mu}_N - \mu|^2 \right] = 0,$$

and similarly for the autocovariance. Note that often, ergodicity is only required to be assumed for the computation of the mean and covariance and not for all possible second order statistics. In this case, we talk about *mean-ergodic* and *covariance-ergodic* processes.

*In summary, when ergodic processes are involved, ensemble averages "across the process" can be obtained as time averages "along the process"*; see Figure 2.12.

In practice, when only a finite number of samples from a realization is available, then the mean and covariance are approximated as the respective sample means.

An issue is to establish conditions under which a process is mean-ergodic or covariance-ergodic. Such conditions do exist, and the interested reader can find such information in more specialized books [6]. It turns out that the condition for mean-ergodicity relies on second order statistics and the condition for covariance-ergodicity on fourth order statistics.

It is very common in statistics as well as in machine learning and signal processing to subtract the mean value from the data during the *preprocessing stage*. In such a case, we say that the data are *centered*. The resulting new process has now *zero mean* value, and the covariance and autocorrelation sequences coincide. From now on, we will assume that the mean is known (or computed as a sample mean) and then subtracted. Such a treatment simplifies the analysis without harming generality.

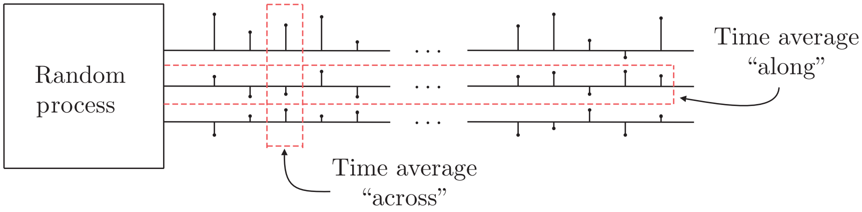

**FIGURE 2.12**

For ergodic processes, mean values for each time instant (time averaging "across" the process) are computed as time averages "along" the process.

**Example 2.2.** The goal of this example is to construct a process that is WSS yet not ergodic. Let a WSS process, $\mathrm{u}_n$,

$$\mathbb{E}[\mathrm{u}_n] = \mu,$$

and

$$\mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}] = r_u(k).$$

Define the process,

$$\mathrm{v}_n := \mathrm{a}\mathrm{u}_n, \tag{2.110}$$

where a is a random variable taking values in $\{0, 1\}$, with probabilities $P(0) = P(1) = 0.5$. Moreover, a and $\mathrm{u}_n$ are statistically independent. Then, we have that

$$\mathbb{E}[\mathrm{v}_n] = \mathbb{E}[\mathrm{a}\mathrm{u}_n] = \mathbb{E}[\mathrm{a}]\,\mathbb{E}[\mathrm{u}_n] = 0.5\mu, \tag{2.111}$$

and

$$\mathbb{E}[\mathrm{v}_n \mathrm{v}_{n-k}] = \mathbb{E}[\mathrm{a}^2]\,\mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}] = 0.5 r_u(k). \tag{2.112}$$

Thus, $\mathrm{v}_n$ is WSS. However, it is not covariance-ergodic. Indeed, some of the realizations will be equal to zero (when $a = 0$), and the mean value and autocorrelation, which will result from them as time averages, will be zero, which is different from the ensemble averages.

## 2.4.3 POWER SPECTRAL DENSITY

The Fourier transform is an indispensable tool for representing in a compact way, in the frequency domain, the variations that a function/sequence undergoes in terms of its free variable (e.g., time). Stochastic processes are inherently related to time. The question that is now raised is whether stochastic processes can be described in terms of a Fourier transform. The answer is affirmative, and the vehicle to achieve this is via the autocorrelation sequence for processes that are at least wide-sense stationary. Prior to providing the necessary definitions, it is useful to summarize some common properties of the autocorrelation sequence.

### *Properties of the autocorrelation sequence*

Let $\mathrm{u}_n$ be a wide-sense stationary process. Its autocorrelation sequence has the following properties, which are given for the more general complex-valued case:

- *Property I.*

$$r(k) = r^*(-k), \quad \forall k \in \mathbb{Z}. \tag{2.113}$$

  This property is a direct consequence of the invariance with respect to the choice of the origin. Indeed,

$$r(k) = \mathbb{E}[\mathrm{u}_n \mathrm{u}^*_{n-k}] = \mathbb{E}[\mathrm{u}_{n+k} \mathrm{u}^*_n] = r^*(-k).$$

- *Property II.*

$$r(0) = \mathbb{E}\left[|\mathrm{u}_n|^2\right]. \tag{2.114}$$

That is, the value of the autocorrelation at $k = 0$ is equal to the mean-square of the magnitude of the respective random variables. Interpreting the square of the magnitude of a variable as its energy, $r(0)$ can be interpreted as the corresponding (average) power.

- *Property III.*

$$r(0) \geq |r(k)|, \quad \forall k \neq 0. \tag{2.115}$$

The proof is provided in Problem 2.9. In other words, the correlation of the variables, corresponding to two different time instants, cannot be larger (in magnitude) than $r(0)$. As we will see in Chapter 4, this property is essentially the Cauchy-Schwartz inequality for the inner products (see also Appendix of Chapter 8).

- *Property IV*. The autocorrelation sequence of a stochastic process is *positive definite*. That is,

$$\boxed{\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* r(n,m) \geq 0, \quad \forall a_n \in \mathbb{C},\ n = 1, 2, \ldots, N,\ \forall N \in \mathbb{Z}.} \tag{2.116}$$

*Proof.* The proof is easily obtained by the definition of the autocorrelation,

$$0 \leq \mathbb{E}\left[\left|\sum_{n=1}^{N} a_n \mathrm{u}_n\right|^2\right] = \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* \,\mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m\right], \tag{2.117}$$

which proves the claim. Note that strictly speaking, we should say that it is semipositive definite. However, the "positive definite" name is the one that has survived in the literature. This property will be useful when introducing *Gaussian processes* in Chapter 13. □

- *Property V*. Let $\mathrm{u}_n$ and $\mathrm{v}_n$ be two WSS processes. Define the new process

$$\mathrm{z}_n = \mathrm{u}_n + \mathrm{v}_n.$$

Then,

$$r_z(k) = r_u(k) + r_v(k) + r_{uv}(k) + r_{vu}(k), \tag{2.118}$$

where the *cross-correlation* between two jointly WS stationary stochastic processes is defined as

$$\boxed{r_{uv}(k) := \mathbb{E}[\mathrm{u}_n \mathrm{v}_{n-k}^*],\ k \in \mathbb{Z} : \quad \text{Cross-correlation.}} \tag{2.119}$$

The proof is a direct consequence of the definition. Note that if the two processes are *uncorrelated*, as when $r_{uv}(k) = r_{vu}(k) = 0$, then

$$r_z(k) = r_u(k) + r_v(k).$$

Obviously, this is also true if the processes $\mathrm{u}_n$ and $\mathrm{v}_n$ are independent and of zero mean value, since then $\mathbb{E}[\mathrm{u}_n \mathrm{v}_{n-k}^*] = \mathbb{E}[\mathrm{u}_n]\,\mathbb{E}[\mathrm{v}_{n-k}^*] = 0$. It should be stressed here that uncorrelatedness is a weaker condition and it *does not* necessarily imply independence; the opposite is true for zero mean values.

- *Property VI.*

$$r_{uv}(k) = r_{vu}^*(-k) \tag{2.120}$$

The proof is similar to that of Property I.

- *Property VII.*

$$r_u(0)r_v(0) \geq |r_{uv}(k)|, \quad \forall k \in \mathbb{Z}. \tag{2.121}$$

The proof is also given in Problem 2.9.

### Power spectral density

**Definition 2.4.** Given a WSS stochastic process, $\mathrm{u}_n$, its *power spectral density* (PSD) (or simply the *power spectrum*) is defined as the Fourier transform of its autocorrelation sequence,

$$\boxed{S(\omega) := \sum_{k=-\infty}^{\infty} r(k) \exp(-j\omega k) : \quad \text{Power Spectral Density.}} \tag{2.122}$$

Using the Fourier transform properties, we can recover the autocorrelation sequence via the *inverse Fourier transform*, in the following manner:

$$\boxed{r(k) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} S(\omega) \exp(j\omega k)\ \mathrm{d}\omega.} \tag{2.123}$$

Due to the properties of the autocorrelation sequence, the PSD have some interesting and useful properties, from a practical point of view.

*Properties of the PSD*

- The PSD of a WSS stochastic process is a *real* and *nonnegative* function of $\omega$. Indeed, we have that

$$\begin{aligned} S(\omega) &= \sum_{k=-\infty}^{+\infty} r(k) \exp(-j\omega k) \\ &= r(0) + \sum_{k=-\infty}^{-1} r(k) \exp(-j\omega k) + \sum_{k=1}^{\infty} r(k) \exp(-j\omega k) \\ &= r(0) + \sum_{k=1}^{+\infty} r^*(k) \exp(j\omega k) + \sum_{k=1}^{\infty} r(k) \exp(-j\omega k) \\ &= r(0) + 2\sum_{k=1}^{+\infty} \text{Real}\left(r(k) \exp(-j\omega k)\right), \end{aligned} \tag{2.124}$$

  which proves the claim that PSD is a real number. In the proof, Property I of the autocorrelation sequence has been used. We defer the proof for the nonnegative part to the end of this section.
- The area under the graph of $S(\omega)$ is proportional to the power of the stochastic process, as expressed by

$$\mathbb{E}\left[|\mathrm{u}_n|^2\right] = r(0) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} S(\omega) d\omega, \tag{2.125}$$

  which is obtained from Eq. (2.123) if we set $k = 0$. We will come to the physical meaning of this property very soon.

### Transmission through a linear system

One of the most important tasks in signal processing and systems theory is the *linear filtering operation* on an *input* time series (*signal*) to generate another *output* sequence. The block diagram of the filtering

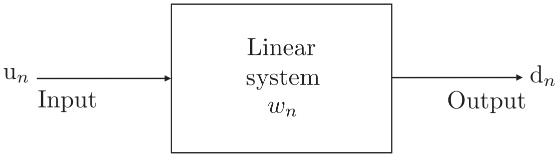

**FIGURE 2.13**

The linear system (filter) is excited by the input sequence (signal), $\mathrm{u}_n$, and provides the output sequence (signal), $\mathrm{d}_n$.

operation is shown in Figure 2.13. From the linear system theory and signal processing basics, it is established that for a class of linear systems known as *linear time invariant* (LTI), the input-output relation is given via the elegant *convolution* between the input sequence and the *impulse response* of the filter,

$$\boxed{\mathrm{d}_n = w_n * \mathrm{u}_n := \sum_{i=-\infty}^{+\infty} w_i^* \mathrm{u}_{n-i} : \quad \text{Convolution Sum},} \tag{2.126}$$

where $\ldots, w_0, w_1, w_2, \ldots$ are the parameters comprising the impulse response describing the filter [8]. In case the impulse response is of finite duration, for example, $w_0, w_1, \ldots, w_{l-1}$, and the rest of the values are zero, then the convolution can be written as

$$\mathrm{d}_n = \sum_{i=0}^{l-1} w_i^* \mathrm{u}_{n-i} = \boldsymbol{w}^H \mathbf{u}_n, \tag{2.127}$$

where

$$\boldsymbol{w} := [w_0, w_1, \ldots, w_{l-1}]^T, \tag{2.128}$$

and

$$\mathbf{u}_n := [\mathrm{u}_n, \mathrm{u}_{n-1}, \ldots, \mathrm{u}_{n-l+1}]^T \in \mathbb{R}^l. \tag{2.129}$$

The latter is known as the *input vector* of order $l$ and at time $n$. It is interesting to note that this is a random vector. However, its elements are part of the stochastic process at *successive* time instants. This gives the respective autocorrelation matrix certain properties and a rich structure, which will be studied and exploited in Chapter 4. As a matter of fact, this is the reason that we used different symbols to denote processes and general random vectors; thus, the reader can readily remember that when dealing with a process, the elements of the involved random vectors have this *extra structure*. Moreover, observe from Eq. (2.126) that if the impulse response of the system is zero for negative values of the time index, $n$, this guarantees *causality*. That is, the output depends only on the values of the input at the current and previous time instants only, and there is no dependence on future values. As a matter of fact, this is also a necessary condition for causality; that is, if the system is causal, then its impulse response is zero for negative time instants [8].

**Theorem 2.1.** *The power spectral density of the output,* $\mathrm{d}_n$*, of a linear time invariant system, when it is excited by a WSS stochastic process,* $\mathrm{u}_n$*, is given by*

$$\boxed{S_d(\omega) = |W(\omega)|^2 S_u(\omega),} \tag{2.130}$$

*where*

$$W(\omega) := \sum_{n=-\infty}^{+\infty} w_n \exp(-j\omega n). \tag{2.131}$$

*Proof.* First, it is shown (Problem 2.10) that

$$\boxed{r_d(k) = r_u(k) * w_k * w^*_{-k}.} \tag{2.132}$$

Then, taking the Fourier transform of both sides, we obtain Eq. (2.130). To this end, we used the well-known properties of the Fourier transform,

$$r_u(k) * w_k \longmapsto S_u(\omega)W(\omega), \quad \text{and} \quad w^*_{-k} \longmapsto W^*(\omega).$$

□

### Physical interpretation of the PSD

We are now ready to justify why the Fourier transform of the autocorrelation sequence was given the specific name of "power spectral density." We restrict our discussion to real processes, although similar arguments hold true for the more general complex case. Figure 2.14 shows the magnitude of the Fourier transform of the impulse response of a very special linear system. The Fourier transform is unity for any frequency in the range $|\omega - \omega_o| \leq \frac{\Delta\omega}{2}$ and zero otherwise. Such a system is known as *bandpass filter*. We assume that $\Delta\omega$ is very small. Then, using Eq. (2.130) and assuming that within the intervals $|\omega - \omega_o| \leq \frac{\Delta\omega}{2}$, $S_u(\omega) \approx S_u(\omega_o)$, we have that

$$S_d(\omega) = \begin{cases} S_u(\omega_o), & \text{if } |\omega - \omega_o| \leq \frac{\Delta\omega}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.133}$$

Hence,

$$\Delta P := \mathbb{E}\left[|\mathrm{d}_n|^2\right] = r_d(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_d(\omega)d\omega \approx S_u(\omega_o)\frac{\Delta\omega}{\pi}, \tag{2.134}$$

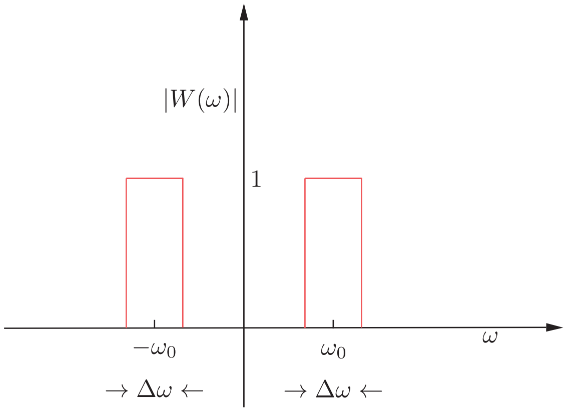

**FIGURE 2.14**

An ideal bandpass filter. The output contains frequencies only in the range of $|\omega - \omega_o| < \Delta\omega/2$.

due to the symmetry of the power spectral density ($S_u(\omega) = S_u(-\omega)$). Hence,

$$\frac{1}{\pi} S_u(\omega_o) = \frac{\Delta P}{\Delta \omega}. \tag{2.135}$$

In other words, the value $S_u(\omega_o)$ can be interpreted as the power density (power per frequency interval) in the frequency (spectrum) domain.

Moreover, this also establishes what was said before that the PSD is a *nonnegative* real function for any value of $\omega \in [-\pi, +\pi]$ (The PSD, being the Fourier transform of a sequence, is periodic with period $2\pi$, e.g., [8]).

*Remarks 2.2.*

- Note that for any WSS stochastic process, there is only one autocorrelation sequence that describes it. However, the converse is not true. A single autocorrelation sequence can correspond to more than one WSS process. Recall that the autocorrelation is the mean value of the product of random variables. However, many random variables can have the same mean value.
- We have shown that the Fourier transform, $S(\omega)$, of an autocorrelation sequence, $r(k)$, is nonnegative. Moreover, if a sequence, $r(k)$, has a nonnegative Fourier transform, then it is positive definite and we can *always* construct a WSS process that has $r(k)$ as its autocorrelation sequence (e.g., [6, pages 410,421]). Thus, the *necessary and sufficient condition for a sequence to be an autocorrelation sequence is the nonnegativity of its Fourier transform.*

**Example 2.3.** *White Noise Sequence.*

A stochastic process, $\eta_n$, is said to be *white noise* if the mean and its autocorrelation sequence satisfy

$$\boxed{\mathbb{E}[\eta_n] = 0 \quad \text{and} \quad r(k) = \begin{cases} \sigma_\eta^2, & \text{if } k = 0, \\ 0, & \text{if } k \neq 0. \end{cases} \quad : \quad \text{White Noise,}} \tag{2.136}$$

where $\sigma_\eta^2$ is its variance. In other words, all variables at different time instants are uncorrelated. If, in addition, they are independent, we say that it is *strictly white noise*. It is readily seen that its PSD is given by

$$S_\eta(\omega) = \sigma_\eta^2. \tag{2.137}$$

That is, it is constant, and this is the reason it is called white noise, analogous to the white light whose spectrum is equally spread over all the wavelengths.

### 2.4.4 AUTOREGRESSIVE MODELS

We have just seen an example of a stochastic process, namely white noise. We now turn our attention to generating WSS processes via appropriate modeling. In this way, we will introduce controlled correlation among the variables, corresponding to the various time instants. We focus on the real data case, to simplify the discussion.

*Autoregressive* processes are among the most popular and widely used models. An autoregressive process of order $l$, denoted as AR($l$), is defined via the following *difference equation*,

$$\boxed{\mathrm{u}_n + a_1\mathrm{u}_{n-1} + \cdots + a_l\mathrm{u}_{n-l} = \eta_n : \quad \text{Autoregressive Process,}} \tag{2.138}$$

where $\eta_n$ is a white noise process with variance $\sigma_\eta^2$.

As is always the case with any difference equation, one starts from some initial conditions and then generates samples recursively by plugging into the model the input sequence samples. The input samples here correspond to a white noise sequence and the initial conditions are set equal to zero, $u_{-1} = \ldots u_{-l} = 0$.

There is no need to mobilize mathematics to see that such a process is not stationary. Indeed, time instant $n = 0$ is distinctly different from all the rest, since it is the time in which initial conditions are applied. However, the effects of the initial conditions tend asymptotically to zero if all the roots of the corresponding characteristic polynomial,

$$z^l + a_1 z^{l-1} + \cdots + a_l = 0,$$

have magnitude *less that unity* (the solution of the corresponding homogeneous equation, without input, tends to zero) [7]. Then, it can be shown that asymptotically, the AR($l$) becomes WSS. This is the assumption that is usually adopted in practice, which will be the case for the rest of this section. Note that the mean value of the process is zero (try it).

The goal now becomes to compute the corresponding autocorrelation sequence, $r(k), k \in \mathbb{Z}$. Multiplying both sides in Eq. (2.138) with $\mathrm{u}_{n-k}, k > 0$, and taking the expectation, we obtain

$$\sum_{i=0}^{l} a_i\, \mathbb{E}[\mathrm{u}_{n-i}\mathrm{u}_{n-k}] = \mathbb{E}[\eta_n \mathrm{u}_{n-k}], \quad k > 0,$$

where $a_0 := 1$, or

$$\sum_{i=0}^{l} a_i r(k-i) = 0. \tag{2.139}$$

We have used the fact that $\mathbb{E}[\eta_n \mathrm{u}_{n-k}]$, $k > 0$ is zero. Indeed, $\mathrm{u}_{n-k}$ depends recursively on $\eta_{n-k}, \eta_{n-k-1} \ldots$, which are all uncorrelated to $\eta_n$, since this is a white noise process. Note that Eq. (2.139) is a difference equation, which can be solved provided we have the initial conditions. To this end, multiply Eq. (2.138) by $\mathrm{u}_n$ and take expectations, which results in

$$\sum_{i=0}^{l} a_i r(i) = \sigma_\eta^2, \tag{2.140}$$

since $\mathrm{u}_n$ recursively depends on $\eta_n$, which contributes the $\sigma_\eta^2$ term, and $\eta_{n-1}, \ldots$, which result to zeros. Combining Eqs. (2.140) with (2.139) the following *linear* system of equations results

$$\begin{bmatrix} r(0) & r(1) & \ldots & r(l) \\ r(1) & r(0) & \ldots & r(l-1) \\ \vdots & \vdots & \vdots & \vdots \\ r(l) & r(l-1) & \ldots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} \sigma_\eta^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.141}$$

These are known as the *Yule-Walker equations*, whose solution results in the values, $r(0), \ldots, r(l)$, which are then used as the initial conditions to solve the difference equation in (2.139) and obtain $r(k)$, $\forall k \in \mathbb{Z}$.

Observe the special structure of the matrix in the linear system. This type of matrix is known as *Toeplitz*, and this is the property that will be exploited to solve efficiently such systems, which result when the autocorrelation matrix of a WSS process is involved; see Chapter 4.

Besides the autoregressive models, other types of stochastic models have been suggested and used. The *autoregressive-moving average* (ARMA) model of order $(l, m)$ is defined by the difference equation,

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \ldots + a_l \mathrm{u}_{n-l} = b_1 \eta_n + \ldots + b_m \eta_{n-m}, \tag{2.142}$$

and the *moving average* model of order $m$, denoted as MA($m$), is defined as

$$\mathrm{u}_n = b_1 \eta_n + \cdots + b_m \eta_{n-m}. \tag{2.143}$$

Note that the AR(l) and the MA($m$) models can be considered as special cases of the ARMA($l, m$). For a more theoretical treatment of the topic, see [1].

**Example 2.4.** Consider the AR(1) process,

$$\mathrm{u}_n + a\mathrm{u}_{n-1} = \eta_n.$$

Following the general methodology explained before, we have

$$r(k) + ar(k-1) = 0, \quad k = 1, 2, \ldots$$

$$r(0) + ar(1) = \sigma_\eta^2.$$

Taking the first equation for $k = 1$ together with the second one readily results in

$$r(0) = \frac{\sigma_\eta^2}{1 - a^2}.$$

Plugging this value into the difference equation, we recursively obtain

$$r(k) = (-a)^{|k|} \frac{\sigma_\eta^2}{1 - a^2}, \; k = 0, \pm 1, \pm 2, \ldots, \tag{2.144}$$

where we used the property, $r(k) = r(-k)$. Observe that if $|a| > 1$, $r(0) < 0$, is meaningless. Also, $|a| < 1$ guarantees that the root of the characteristic polynomial ($z_* = -a$) is smaller than one. Moreover, $|a| < 1$ guarantees that $r(k) \longrightarrow 0$ as $k \longrightarrow \infty$. This is in line with common sense, since variables that are far away must be uncorrelated.

Figure 2.15 shows the time evolution of two AR(1) processes (after the processes have converged to be stationary) together with the respective autocorrelation sequences, for two cases, corresponding to $a = -0.9$ and $a = -0.4$. Observe that the larger the magnitude of $a$, the smoother the realization becomes and time variations are *slower*. This is natural, since nearby samples are highly correlated and so, on average, they tend to have similar values. The opposite is true for small values of $a$. For comparison purposes, Figure 2.16a is the case of $a = 0$, which corresponds to a white noise. Figure 2.16b shows the power spectral densities corresponding to the two cases of Figure 2.15. Observe that the faster the autocorrelation approaches zero, the more spread out the PSD is, and vice versa.

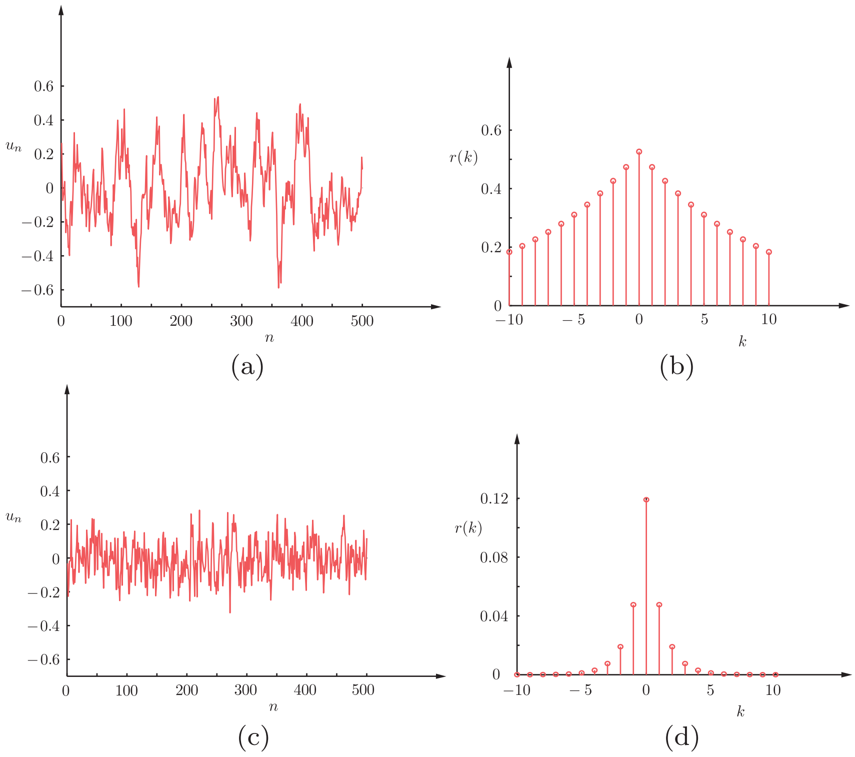

**FIGURE 2.15**

(a) The time evolution of a realization of the AR(1) with $a = -0.9$ and (b) the respective autocorrelation sequence. (c) The time evolution of a realization of the AR(1) with $a = -0.4$ and (d) the corresponding autocorrelation sequence.

## 2.5 INFORMATION THEORY

So far in this chapter, we have looked at some basic definitions and properties concerning probability theory and stochastic processes. In the same vein, we will now focus on the basic definitions and notions related to *information theory*. Although information theory was originally developed in the context of communications and coding disciplines, its application and use has now been adopted in a wide range of areas, including machine learning. Notions from information theory are used for establishing cost functions for optimization in parameter estimation problems, and concepts from information theory are employed to estimate unknown probability distributions in the context of constrained optimization tasks. We will discuss such methods later in this book.

The father of information theory is *Claude Elwood Shannon* (1916-2001), an American mathematician and electrical engineer. He founded information theory with the landmark paper “A mathematical theory of communication,” published in the Bell System Technical Journal in 1948. However, he is

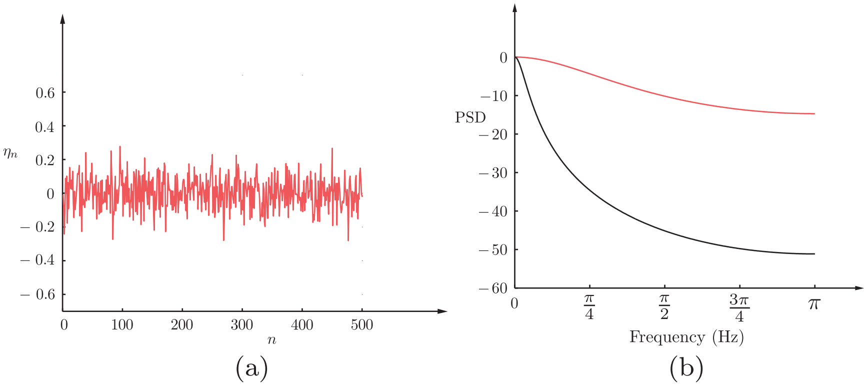

**FIGURE 2.16**

(a) The time evolution of a realization from a white noise process. (b) The power spectral densities in dBs, for the two AR(1) sequences of Figure 2.15. The red one corresponds to $a = -0.4$ and the gray one to $a = -0.9$. The smaller the magnitude of $a$, the closer the process is to a white noise, and its power spectral density tends to increase the power with which high frequencies participate. Since the PSD is the Fourier transform of the autocorrelation sequence, observe that the broader a sequence is in time, the narrower its Fourier transform becomes, and vice versa.

also credited with founding digital circuit design theory in 1937, when, as a 21-year-old master's degree student at the Massachusetts Institute of Technology (MIT), he wrote his thesis demonstrating that electrical applications of Boolean algebra could construct and resolve any logical, numerical relationship. So he is also credited as a father of digital computers. Shannon, while working for the national defense during World War II, contributed to the field of cryptography, converting it from an art to a rigorous scientific field.

As is the case for probability, the notion of information is part of our everyday vocabulary. In this context, an event carries information if it is either unknown to us, or if the probability of its occurrence is very low and, in spite of that, it happens. For example, if one tells us that the sun shines bright during summer days in the Sahara desert, we could consider such a statement rather dull and useless. On the contrary, if somebody gives us news about snow in the Sahara during summer, that statement carries a lot of information and can possibly ignite a discussion concerning the climate change.

Thus, trying to formalize the notion of information from a mathematical point of view, it is reasonable to define it in terms of the negative logarithm of the probability of an event. If the event is certain to occur, it carries zero information content; however, if its probability of occurrence is low, then its information content has a large positive value.

## 2.5.1 DISCRETE RANDOM VARIABLES

### *Information*

Given a discrete random variable, x, which takes values in the set $\mathcal{X}$, the *information* associated with any value $x \in \mathcal{X}$ is denoted as $I(x)$ and it is defined as

$$\boxed{I(x) = -\log P(x): \quad \text{Information Associated with } \mathrm{x} = x \in \mathcal{X}.} \tag{2.145}$$

Any base for the logarithm can be used. If the natural logarithm is chosen, information is measured in terms of *nats* (natural units). If the base 2 logarithm is employed, information is measured in terms of *bits* (binary digits). Employing the logarithmic function to define information is also in line with common sense reasoning that the information content of two statistically independent events should be the sum of the information conveyed by each one of them individually; $I(x, y) = -\ln P(x, y) = -\ln P(x) - \ln P(y)$.

**Example 2.5.** We are given a binary random variable $\mathrm{x} \in \mathcal{X} = \{0, 1\}$, and assume that $P(1) = P(0) = 0.5$. We can consider this random variable as a source that generates and emits two possible values. The information content of each one of the two equiprobable events is

$$I(0) = I(1) = -\log_2 0.5 = 1 \text{ bit}.$$

Let us now consider another source of random events, which generates *code words* comprising $k$ binary variables together. The output of this source can be seen as a random vector with binary-valued elements, $\mathbf{x} = [\mathrm{x}_1, \ldots, \mathrm{x}_k]^T$. The corresponding probability space, $\mathcal{X}$, comprises $K = 2^k$ elements. If all possible values have the same probability, $1/K$, then the information content of each possible event is equal to

$$I(\boldsymbol{x}_i) = -\log_2 \frac{1}{K} = k \text{ bits}.$$

We observe that in the case where the number of possible events is larger, the information content of each individual one (assuming equiprobable events) becomes larger. This is also in line with common sense reasoning, since if the source can emit a large number of (equiprobable) events, the occurrence of any one of them carries more information than a source that can only emit a few possible events.

### Mutual and conditional information

Besides marginal probabilities, we have already been introduced to the concept of conditional probability. This leads to the definition of mutual information.

Given two discrete random variables, $\mathrm{x} \in \mathcal{X}$ and $\mathrm{y} \in \mathcal{Y}$, the information content provided by the occurrence of the event $\mathrm{y} = y$ about the event $\mathrm{x} = x$ is measured by the *mutual information*, denoted as $I(x; y)$ and defined by

$$\boxed{I(x, y) := \log \frac{P(x|y)}{P(x)} : \quad \text{Mutual Information.}} \tag{2.146}$$

Note that if the two variables are statistically independent, then their mutual information is zero; this is most reasonable, since observing $y$ says nothing about $x$. On the contrary, if by observing $y$ it is certain that $x$ will occur, as when $P(x|y) = 1$, then the mutual information becomes $I(x, y) = I(x)$, which is again in line with common reasoning. Mobilizing our now familiar product rule, we can see that

$$I(x, y) = I(y, x).$$

The *conditional information* of x given y is defined as

$$\boxed{I(x|y) = -\log P(x|y) : \quad \text{Conditional Information.}} \tag{2.147}$$

It is straightforward to show that

$$I(x, y) = I(x) - I(x|y). \tag{2.148}$$

**Example 2.6.** In a communications channel, the source transmits binary symbols, x, with probability $P(0) = P(1) = 1/2$. The channel is noisy, so the received symbols, y, may have changed polarity, due to noise, with the following probabilities:

$$P(\mathrm{y}=0|\mathrm{x}=0) = 1-p,$$

$$P(\mathrm{y}=1|\mathrm{x}=0) = p,$$

$$P(\mathrm{y}=1|\mathrm{x}=1) = 1-q,$$

$$P(\mathrm{y}=0|\mathrm{x}=1) = q.$$

This example illustrates in its simplest form the effect of a *communications channel*. Transmitted bits are hit by noise and what the receiver receives is the noisy (possibly wrong) information. The task of the receiver is to decide, upon reception of a sequence of symbols, which was the originally transmitted one.

The goal of our example is to determine the mutual information about the occurrence of $\mathrm{x} = 0$ and $\mathrm{x} = 1$ once $\mathrm{y} = 0$ has been observed. To this end, we first need to compute the marginal probabilities,

$$P(\mathrm{y}=0) = P(\mathrm{y}=0|\mathrm{x}=0)P(\mathrm{x}=0) + P(\mathrm{y}=0|\mathrm{x}=1)P(\mathrm{x}=1) = \frac{1}{2}(1-p+q),$$

and similarly,

$$P(\mathrm{y}=1) = \frac{1}{2}(1-q+p).$$

Thus, the mutual information is

$$\begin{aligned} I(0,0) &= \log_2 \frac{P(\mathrm{x}=0|\mathrm{y}=0)}{P(\mathrm{x}=0)} = \log_2 \frac{P(\mathrm{y}=0|\mathrm{x}=0)}{P(\mathrm{y}=0)} \\ &= \log_2 \frac{2(1-p)}{1-p+q}, \end{aligned}$$

and

$$I(1,0) = \log_2 \frac{2q}{1-p+q}.$$

Let us now consider that $p = q = 0$. Then $I(0,0) = 1$ bit, which is equal to $I(\mathrm{x}=0)$, since the output specifies the input with certainty. If on the other hand $p = q = 1/2$, then $I(0,0) = 0$ bits, since the noise can randomly change polarity with equal probability. If now $p = q = 1/4$, then $I(0,0) = \log_2 \frac{3}{2} = 0.587$ bits and $I(1,0) = -1$ bit. Observe that the mutual information can take negative values, too.

### Entropy and average mutual information

Given a discrete random variable, $\mathrm{x} \in \mathcal{X}$, its *entropy* is defined as the average information over all possible outcomes,

$$\boxed{H(\mathrm{x}) := -\sum_{x \in \mathcal{X}} P(x) \log P(x) : \quad \text{Entropy of x.}} \tag{2.149}$$

Note that if $P(x) = 0$, $P(x)\log P(x) = 0$, by taking into consideration that $\lim_{x \to 0} x \log x = 0$.

In a similar way, the *average mutual information* between two random variables, x, y, is defined as

$$I(\mathrm{x}, \mathrm{y}) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) I(x; y)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log \frac{P(x|y)P(y)}{P(x)P(y)}$$

or

$$\boxed{I(\mathrm{x}, \mathrm{y}) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} : \quad \text{Average Mutual Information.}} \tag{2.150}$$

It can be shown that

$$I(\mathrm{x}, \mathrm{y}) \geq 0,$$

and it is zero if x and y are statistically independent (Problem 2.12).

In comparison, the *conditional entropy* of x given y is defined as

$$\boxed{H(\mathrm{x}|\mathrm{y}) := -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log P(x|y) : \quad \text{Conditional Entropy.}} \tag{2.151}$$

It is readily shown, by taking into account the probability product rule, that

$$I(\mathrm{x}, \mathrm{y}) = H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y}). \tag{2.152}$$

**Lemma 2.1.** *The entropy of a random variable,* $\mathrm{x} \in \mathcal{X}$*, takes its maximum value if all possible values,* $x \in \mathcal{X}$*, are equiprobable.*

*Proof.* The proof is given in Problem 2.14. □

In other words, the entropy can be considered as a measure of randomness of a source that emits symbols randomly. The maximum value is associated with the maximum uncertainty of what is going to be emitted, since the maximum value occurs if all symbols are equiprobable. The smallest value of the entropy is equal to zero, which corresponds to the case where all events have zero probability with the exception of one, whose probability to occur is equal to one.

**Example 2.7.** Consider a binary source that transmits the values 1 or 0 with probabilities $p$ and $1 - p$, respectively. Then the entropy of the associated random variable is

$$H(\mathrm{x}) = -p \log_2 p - (1 - p) \log_2 (1 - p).$$

Figure 2.17 shows the graph for various values of $p \in [0, 1]$. Observe that the maximum value occurs for $p = 1/2$.

### 2.5.2 CONTINUOUS RANDOM VARIABLES

All the definitions given before can be generalized to the case of continuous random variables. However, this generalization must be made with caution. Recall that the probability of occurrence of any single value of a random variable that takes values in an interval in the real axis is zero. Hence, the corresponding information content is infinite.

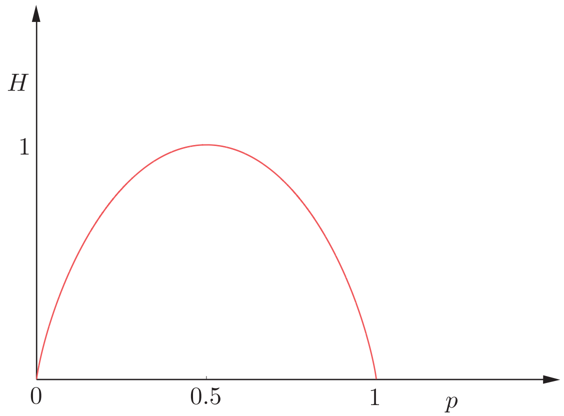

**FIGURE 2.17**

The maximum value of the entropy for a binary random variable occurs if the two possible events have equal probability, $p = 1/2$.

To define the entropy of a continuous variable, x, we first *discretize* it and form the corresponding discrete variable, $\mathrm{x}_\Delta$,

$$\mathrm{x}_\Delta := n\Delta, \text{ if } (n-1)\Delta < \mathrm{x} \leq n\Delta, \tag{2.153}$$

where $\Delta > 0$. Then,

$$P(\mathrm{x}_\Delta = n\Delta) = P(n\Delta - \Delta < \mathrm{x} \leq n\Delta) = \int_{(n-1)\Delta}^{n\Delta} p(x)\,\mathrm{d}x = \Delta\bar{p}(n\Delta), \tag{2.154}$$

where $\bar{p}(n\Delta)$ is a number between the maximum and the minimum value of $p(x), x \in (n\Delta - \Delta, n\Delta]$ (such a number exists by the mean value theorem). Then we can write,

$$H(\mathrm{x}_\Delta) = -\sum_{n=-\infty}^{+\infty} \Delta\bar{p}(n\Delta)\log\left(\Delta\bar{p}(n\Delta)\right), \tag{2.155}$$

and since

$$\sum_{n=-\infty}^{+\infty} \Delta\bar{p}(n\Delta) = \int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x = 1,$$

we obtain

$$H(\mathrm{x}_\Delta) = -\log\Delta - \sum_{n=-\infty}^{+\infty} \Delta\bar{p}(n\Delta)\log\left(\bar{p}(n\Delta)\right). \tag{2.156}$$

Note that $\mathrm{x}_\Delta \longrightarrow \mathrm{x}$ as $\Delta \longrightarrow 0$. However, if we take the limit in Eq. (2.156), then $-\log\Delta$ goes to infinity. This is the crucial difference compared to the discrete variables.

The entropy for a continuous random variable, x, is defined as the limit

$$H(\mathrm{x}) := \lim_{\Delta\to 0}\left(H(\mathrm{x}_\Delta) + \log\Delta\right),$$

or

$$\boxed{H(\mathrm{x}) = -\int_{-\infty}^{+\infty} p(x) \log p(x)\, \mathrm{d}x : \quad \text{Entropy.}} \tag{2.157}$$

This is the reason that the entropy of a continuous variable is also called *differential entropy*.

Note that the entropy is still a measure of randomness (uncertainty) of the distribution describing x. This is demonstrated via the following example.

**Example 2.8.** We are given a random variable $\mathrm{x} \in [a, b]$. Of all the possible pdfs that can describe this variable, find the one that maximizes the entropy.

This task translates to the following constrained optimization task:

$$\text{maximize with respect to } p: \quad H = -\int_a^b p(x) \ln p(x) dx,$$

$$\text{subject to:} \quad \int_a^b p(x) dx = 1.$$

The constraint guarantees that the function to result is indeed a pdf. Using calculus of variations to perform the optimization (Problem 2.15), it turns out that

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

In other words, the result is the uniform distribution, which is indeed the most random one since it gives no preference to any particular subinterval of $[a, b]$.

We will come to this method of estimating pdfs in Section 12.4.1. This elegant method for estimating pdfs comes from Jaynes [3, 4], and it is known as the *maximum entropy method*. In its more general form, more constraints are involved to fit the needs of the specific problem.

### Average mutual information and conditional information

Given two continuous random variables, the average mutual information is defined as

$$\boxed{I(\mathrm{x}, \mathrm{y}) := \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}\, \mathrm{d}x\, \mathrm{d}y} \tag{2.158}$$

and the conditional entropy of x given y

$$\boxed{H(\mathrm{x}|\mathrm{y}) := \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \log p(x|y)\, \mathrm{d}x\, \mathrm{d}y.} \tag{2.159}$$

Using standard arguments and the product rule, it is easy to show that

$$I(\mathrm{x}; \mathrm{y}) = H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y}) = H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x}). \tag{2.160}$$

### Relative entropy or Kullback-Leibler divergence

The *relative entropy* or *Kullback-Leibler divergence* is a quantity that has been developed within the context of information theory for measuring similarity between two pdfs. It is widely used in machine learning optimization tasks when pdfs are involved; see Chapter 12. Given two pdfs, $p(\cdot)$ and $q(\cdot)$, their Kullback-Leibler divergence, denoted as $\text{KL}(p||q)$, is defined as

$$\boxed{\text{KL}(p||q) := \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} \, dx : \quad \text{Kullback-Leibler Divergence.}} \tag{2.161}$$

Note that

$$I(\mathrm{x}, \mathrm{y}) = \text{KL}\left(p(x, y)||p(x)p(y)\right).$$

The Kullback-Leibler divergence is *not* symmetric, i.e., $\text{KL}(p||q) \neq \text{KL}(q||p)$ and it can be shown that it is a nonnegative quantity (the proof is similar to the proof that the mutual information is nonnegative; see Problem 12.16 of Chapter 12). Moreover, it is zero if and only if $p = q$.

Note that all we have said concerning entropy and mutual information is readily generalized to the case of random vectors.

## 2.6 STOCHASTIC CONVERGENCE

We will close this memory-refreshing tour of the theory of probability and related concepts with some definitions concerning convergence of sequences of random variables.

Let a sequence of random variables,

$$\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_n \ldots$$

We can consider this sequence as a discrete-time stochastic process. Due to the randomness, a realization of this process, as shown by

$$x_0, x_1, \ldots, x_n \ldots,$$

may converge or may not. Thus, the notion of convergence of random variables has to be treated carefully, and different interpretations have been developed.

Recall from our basic calculus that a sequence of numbers, $x_n$, converges to a value, $x$, if $\forall \epsilon > 0$ there exists a number, $n(\epsilon)$, such that

$$|x_n - x| < \epsilon, \quad \forall n \geq n(\epsilon). \tag{2.162}$$

### *Convergence everywhere*

We say that a random sequence *converges everywhere* if every realization, $x_n$, of the random process converges to a value $x$, according to the definition given in Eq. (2.162). Note that every realization converges to a different value, which itself can be considered as the outcome of a random variable x, and we write

$$\mathrm{x}_n \xrightarrow[n\to\infty]{} \mathrm{x}. \tag{2.163}$$

It is common to denote a realization (outcome) of a random process as $\mathrm{x}_n(\zeta)$, where $\zeta$ denotes a specific experiment.

### Convergence almost everywhere

A weaker version of convergence, compared to the previous one, is the *convergence almost everywhere*. Let the set of outcomes $\zeta$ such as

$$\lim \mathrm{x}_n(\zeta) = \mathrm{x}(\zeta), \quad n \longrightarrow \infty.$$

We say that the sequence $\mathrm{x}_n$ converges almost everywhere, if

$$P(\mathrm{x}_n \longrightarrow \mathrm{x}) = 1, \quad n \longrightarrow \infty. \tag{2.164}$$

Note that $\{\mathrm{x}_n \longrightarrow \mathrm{x}\}$ denotes the event comprising *all* the outcomes such as $\lim \mathrm{x}_n(\zeta) = \mathrm{x}(\zeta)$. The difference with the convergence everywhere is that now it is allowed to a finite or countably infinite number of realizations (that is, to a set of zero probability) not to converge. Often, this type of convergence is referred to as *almost sure* convergence or convergence *with probability 1*.

### Convergence in the mean-square sense

We say that a random sequence, $\mathrm{x}_n$, converges to the random variable, x, in the *mean-square* (MS) *sense*, if

$$\mathbb{E}\left[|\mathrm{x}_n - \mathrm{x}|^2\right] \longrightarrow 0, \quad n \longrightarrow \infty. \tag{2.165}$$

### Convergence in probability

Given a random sequence, $\mathrm{x}_n$, a random variable, x, and a nonnegative number $\epsilon$, then $\{|\mathrm{x}_n - \mathrm{x}| > \epsilon\}$ is an event. We define the new sequence of numbers, $P\left(\{|\mathrm{x}_n - \mathrm{x}| > \epsilon\}\right)$. We say that $\mathrm{x}_n$ converges to x *in probability* if the constructed sequence of numbers tends to zero,

$$P\big(\{|\mathrm{x}_n - \mathrm{x}| > \epsilon\}\big) \longrightarrow 0, \; n \longrightarrow \infty, \; \forall \epsilon > 0. \tag{2.166}$$

### Convergence in distribution

Given a random sequence, $\mathrm{x}_n$, and a random variable, x, let $F_n(x)$ and $F(x)$ be the cdfs, respectively. We say that $\mathrm{x}_n$ converges to x *in distribution*, if

$$F_n(x) \longrightarrow F(x), \quad n \longrightarrow \infty, \tag{2.167}$$

for *every* point $x$ of continuity of $F(x)$.

It can be shown that if a random sequence converges either almost everywhere or in the MS sense then it necessarily converges in probability, and if it converges in probability then it necessarily converges in distribution. The converse arguments are not necessarily true. In other words, the weakest version of convergence is that of convergence in distribution.

## PROBLEMS

**2.1** Derive the mean and variance for the binomial distribution.

**2.2** Derive the mean and variance for the uniform distribution.

**2.3** Derive the mean and covariance matrix of the multivariate Gaussian.

**2.4** Show that the mean and variance of the beta distribution with parameters $a$ and $b$ are given by

$$\mathbb{E}[\mathrm{x}] = \frac{a}{a+b}$$

and

$$\sigma_x^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

*Hint.* Use the property $\Gamma(a+1) = a\Gamma(a)$.

**2.5** Show that the normalizing constant in the beta distribution with parameters $a, b$ is given by

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}.$$

**2.6** Show that the mean and variance of the gamma pdf

$$\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \quad a, b, x > 0$$

are given by

$$\mathbb{E}[\mathrm{x}] = \frac{a}{b},$$

$$\sigma_x^2 = \frac{a}{b^2}.$$

**2.7** Show that the mean and variance of a Dirichlet pdf with $K$ variables $\mathrm{x}_k, k = 1, 2, \ldots, K$ and parameters $a_k, k = 1, 2, \ldots, K$, are given by

$$\mathbb{E}[\mathrm{x}_k] = \frac{a_k}{\overline{a}}, \; k = 1, 2, \ldots, K$$

$$\sigma_k^2 = \frac{a_k(\overline{a} - a_k)}{\overline{a}^2(1+\overline{a})}, \; k = 1, 2, \ldots, K,$$

$$\text{cov}[\mathrm{x}_i \mathrm{x}_j] = -\frac{a_i a_j}{\overline{a}^2(1+\overline{a})}, \; i \neq j,$$

where $\overline{a} = \sum_{k=1}^{K} a_k$.

**2.8** Show that the sample mean, using $N$ i.i.d. drawn samples, is an unbiased estimator with variance that tends to zero asymptotically, as $N \longrightarrow \infty$.

**2.9** Show that for WSS processes

$$r(0) \geq |r(k)|, \quad \forall k \in \mathbb{Z},$$

and that for jointly WSS processes

$$r_u(0) r_v(0) \geq |r_{uv}(k)|, \quad \forall k \in \mathbb{Z}.$$

**2.10** Show that the autocorrelation of the output of a linear system, with impulse response $w_n, n \in \mathbb{Z}$, is related to the autocorrelation of the input WSS process, via

$$r_d(k) = r_u(k) * w_k * w_{-k}^*.$$

**2.11** Show that

$$\ln x \leq x - 1.$$

**2.12** Show that

$$I(\mathrm{x}, \mathrm{y}) \geq 0.$$

*Hint.* Use the inequality of Problem 2.11.

**2.13** Show that if $a_i, b_i, i = 1, 2, \ldots, M$, are positive numbers, such as

$$\sum_{i=1}^{M} a_i = 1 \quad \text{and} \quad \sum_{i=1}^{M} b_i \leq 1,$$

then

$$-\sum_{i=1}^{M} a_i \ln a_i \leq -\sum_{i=1}^{M} a_i \ln b_i.$$

**2.14** Show that the maximum value of the entropy of a random variable occurs if all possible outcomes are equiprobable.

**2.15** Show that from all the pdfs that describe a random variable in an interval $[a, b]$, the uniform one maximizes the entropy.

## REFERENCES

[1] P.J. Brockwell, R.A. Davis, Time Series: Theory and Methods, second ed., Springer, New York, 1991.

[2] R.T. Cox, Probability, frequency and reasonable expectation. Am. J. Phys. 14 (1) (1946) 1-13.

[3] E.T. Jaynes, Information theory and statistical mechanics. Phys. Rev. 106 (4) (1957) 620-630.

[4] E.T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, 2003.

[5] A.N. Kolmogorov, Foundations of the Theory of Probability, second ed., Chelsea Publishing Company, New York, 1956.

[6] A. Papoulis, S.U. Pillai, Probability, Random Variables and Stochastic Processes, fourth ed., McGraw Hill, New York, 2002.

[7] M.B. Priestly, Spectral Analysis and Time Series, Academic Press, New York, 1981.

[8] J. Proakis, D. Manolakis, Digital Signal Processing, second ed., MacMillan, New York, 1992.