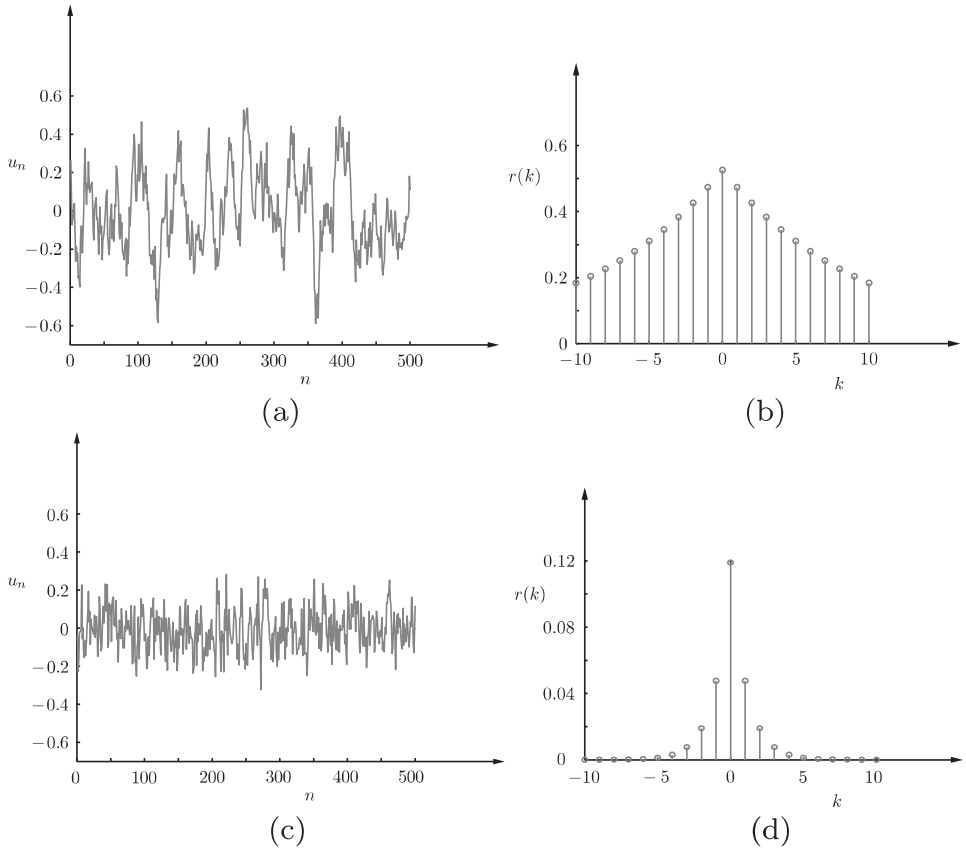

FIGURE 2.15

(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

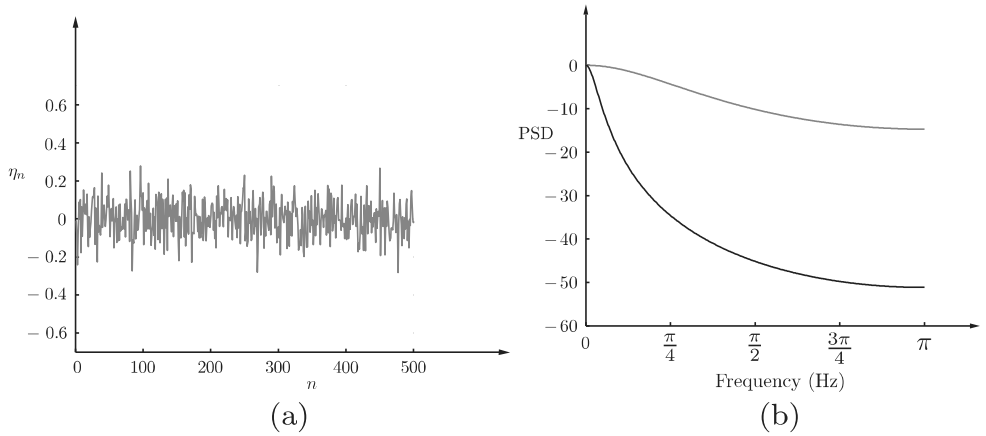

FIGURE 2.16

(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

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

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.

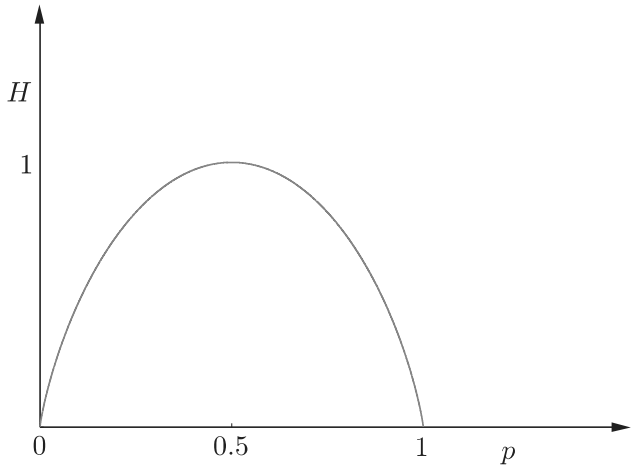

FIGURE 2.17

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, x_Δ,

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

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:

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

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

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 $\mathrm{KL}(p||q)$, is defined as

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

Note that

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

The Kullback-Leibler divergence is *not* symmetric, i.e., $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$ 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 ζ denotes a specific experiment.