

Chapter 1

Information Sources

Abstract An *information source* or *source* is a mathematical model for a physical entity that produces a succession of symbols called “outputs” in a random manner. The symbols produced may be real numbers such as voltage measurements from a transducer, binary numbers as in computer data, two dimensional intensity fields as in a sequence of images, continuous or discontinuous waveforms, and so on. The space containing all of the possible output symbols is called the *alphabet* of the source and a source is essentially an assignment of a probability measure to events consisting of sets of sequences of symbols from the alphabet. It is useful, however, to explicitly treat the notion of time as a transformation of sequences produced by the source. Thus in addition to the common random process model we shall also consider modeling sources by dynamical systems as considered in ergodic theory. The material in this chapter is a distillation of [55, 58] and is intended to establish notation.

1.1 Probability Spaces and Random Variables

A measurable space (Ω, \mathcal{B}) is a pair consisting of a sample space Ω together with a σ -field \mathcal{B} of subsets of Ω (also called the event space). A σ -field or σ -algebra \mathcal{B} is a nonempty collection of subsets of Ω with the following properties:

$$\Omega \in \mathcal{B}. \quad (1.1)$$

$$\text{If } F \in \mathcal{B}, \text{ then } F^c = \{\omega : \omega \notin F\} \in \mathcal{B}. \quad (1.2)$$

$$\text{If } F_i \in \mathcal{B}; i = 1, 2, \dots, \text{ then } \bigcup_i F_i \in \mathcal{B}. \quad (1.3)$$

From de Morgan’s “laws” of elementary set theory it follows that also

$$\bigcap_{i=1}^{\infty} F_i = \left(\bigcup_{i=1}^{\infty} F_i^c \right)^c \in \mathcal{B}.$$

An event space is a collection of subsets of a sample space (called events by virtue of belonging to the event space) such that any countable sequence of set theoretic operations (union, intersection, complementation) on events produces other events. Note that there are two extremes: the largest possible σ -field of Ω is the collection of all subsets of Ω (sometimes called the *power set*) and the smallest possible σ -field is $\{\Omega, \emptyset\}$, the entire space together with the null set $\emptyset = \Omega^c$ (called the *trivial space*).

If instead of the closure under countable unions required by (1.3), we only require that the collection of subsets be closed under finite unions, then we say that the collection of subsets is a *field*.

While the concept of a field is simpler to work with, a σ -field possesses the additional important property that it contains all of the limits of sequences of sets in the collection. That is, if F_n , $n = 1, 2, \dots$ is an increasing sequence of sets in a σ -field, that is, if $F_{n-1} \subset F_n$ and if $F = \bigcup_{n=1}^{\infty} F_n$ (in which case we write $F_n \uparrow F$ or $\lim_{n \rightarrow \infty} F_n = F$), then also F is contained in the σ -field. In a similar fashion we can define decreasing sequences of sets: If F_n decreases to F in the sense that $F_{n+1} \subset F_n$ and $F = \bigcap_{n=1}^{\infty} F_n$, then we write $F_n \downarrow F$. If $F_n \in \mathcal{B}$ for all n , then $F \in \mathcal{B}$.

A *probability space* (Ω, \mathcal{B}, P) is a triple consisting of a sample space Ω , a σ -field \mathcal{B} of subsets of Ω , and a probability measure P which assigns a real number $P(F)$ to every member F of the σ -field \mathcal{B} so that the following conditions are satisfied:

- *Nonnegativity:*

$$P(F) \geq 0, \text{ all } F \in \mathcal{B}; \quad (1.4)$$

- *Normalization:*

$$P(\Omega) = 1; \quad (1.5)$$

- *Countable Additivity:*

If $F_i \in \mathcal{B}$, $i = 1, 2, \dots$ are disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i). \quad (1.6)$$

A set function P satisfying only (1.4) and (1.6) but not necessarily (1.5) is called a *measure* and the triple (Ω, \mathcal{B}, P) is called a *measure space*. Since the probability measure is defined on a σ -field, such countable unions of subsets of Ω in the σ -field are also events in the σ -field.

A standard result of basic probability theory is that if $G_n \downarrow \emptyset$ (the empty or null set), that is, if $G_{n+1} \subset G_n$ for all n and $\bigcap_{n=1}^{\infty} G_n = \emptyset$, then we have

- *Continuity at \emptyset :*

$$\lim_{n \rightarrow \infty} P(G_n) = 0. \quad (1.7)$$

similarly it follows that we have

- *Continuity from Below:*

$$\text{If } F_n \uparrow F, \text{ then } \lim_{n \rightarrow \infty} P(F_n) = P(F), \quad (1.8)$$

and

- *Continuity from Above:*

$$\text{If } F_n \downarrow F, \text{ then } \lim_{n \rightarrow \infty} P(F_n) = P(F). \quad (1.9)$$

Given a measurable space (Ω, \mathcal{B}) , a collection \mathcal{G} of members of \mathcal{B} is said to *generate* \mathcal{B} and we write $\sigma(\mathcal{G}) = \mathcal{B}$ if \mathcal{B} is the smallest σ -field that contains \mathcal{G} ; that is, if a σ -field contains all of the members of \mathcal{G} , then it must also contain all of the members of \mathcal{B} . The following is a fundamental approximation theorem of probability theory. A proof may be found in Corollary 1.5.3 of [55] or Corollary 1.5 of [58]. The result is most easily stated in terms of the symmetric difference Δ defined by

$$F \Delta G \equiv (F \cap G^c) \cup (F^c \cap G).$$

Theorem 1.1. *Given a probability space (Ω, \mathcal{B}, P) and a generating field \mathcal{F} , that is, \mathcal{F} is a field and $\mathcal{B} = \sigma(\mathcal{F})$, then given $F \in \mathcal{B}$ and $\epsilon > 0$, there exists an $F_0 \in \mathcal{F}$ such that $P(F \Delta F_0) \leq \epsilon$.*

Let (A, \mathcal{B}_A) denote another measurable space. We will also use $\mathcal{B}(A)$ as a synonym for \mathcal{B}_A . A *random variable* or *measurable function* defined on (Ω, \mathcal{B}) and taking values in (A, \mathcal{B}_A) is a mapping or function $f : \Omega \rightarrow A$ with the property that

$$\text{if } F \in \mathcal{B}_A, \text{ then } f^{-1}(F) = \{\omega : f(\omega) \in F\} \in \mathcal{B}. \quad (1.10)$$

The name “random variable” is commonly associated with the special case where A is the real line and \mathcal{B} the Borel field, the smallest σ -field containing all the intervals. Occasionally a more general sounding name such as “random object” is used for a measurable function to implicitly include random variables (A the real line), random vectors (A a Euclidean space), and random processes (A a sequence or waveform space). We will use the terms “random variable” in the more general sense. Usually A will either be a metric space or a product of metric spaces, in which case the σ -field will be a Borel field \mathcal{B}_A or $\mathcal{B}(A)$ of subsets of A . If A is a product of metric spaces, then \mathcal{B}_A will be taken as the corresponding product σ -field, that is, the σ -field generated by the rectangles.

A random variable is just a function or mapping with the property that inverse images of “output events” determined by the random variable are events in the original measurable space. This simple property ensures that the output of the random variable will inherit its own probability measure. For example, with the probability measure P_f defined by

$$P_f(B) = P(f^{-1}(B)) = P(\omega : f(\omega) \in B); B \in \mathcal{B}_A,$$

(A, \mathcal{B}_A, P_f) becomes a probability space since measurability of f and elementary set theory ensure that P_f is indeed a probability measure. The induced probability measure P_f is called the *distribution* of the random variable f . The measurable space (A, \mathcal{B}_A) or, simply, the sample space A , is called the alphabet of the random variable f . We shall occasionally also use the notation Pf^{-1} which is a mnemonic for the relation $Pf^{-1}(F) = P(f^{-1}(F))$ and which is less awkward when f itself is a function with a complicated name, e.g., $\Pi_{I-\mathcal{M}}$.

It is often convenient to abbreviate an English description the of a probability of an event to the pseudo mathematical form $\Pr(f \in F)$, which can be considered shorthand for $P_f(F) = P(f^{-1}(F))$ and can be read as “the probability that f is in F .”

If the alphabet A of a random variable f is not clear from context, then we shall refer to f as an *A-valued random variable*. If f is a measurable function from (Ω, \mathcal{B}) to (A, \mathcal{B}_A) , we will say that f is $\mathcal{B}/\mathcal{B}_A$ -measurable if the σ -fields might not be clear from context.

Given a probability space (Ω, \mathcal{B}, P) , a collection of subsets \mathcal{G} is a sub- σ -field if it is a σ -field and all its members are in \mathcal{B} . A random variable $f : \Omega \rightarrow A$ is said to be measurable with respect to a sub- σ -field \mathcal{G} if $f^{-1}(H) \in \mathcal{G}$ for all $H \in \mathcal{B}_A$.

Given a probability space (Ω, \mathcal{B}, P) and a sub- σ -field \mathcal{G} , for any event $H \in \mathcal{B}$ the conditional probability $m(H|\mathcal{G})$ is defined as any function, say g , which satisfies the two properties

$$g \text{ is measurable with respect to } \mathcal{G} \quad (1.11)$$

$$\int_G g h dP = m(G \cap H); \text{ all } G \in \mathcal{G}. \quad (1.12)$$

An important special case of conditional probability occurs when studying the distributions of random variables defined on an underlying probability space. Suppose that $X : \Omega \rightarrow A_X$ and $Y : \Omega \rightarrow A_Y$ are two random variables defined on (Ω, \mathcal{B}, P) with alphabets A_X and A_Y and σ -fields \mathcal{B}_{A_X} and \mathcal{B}_{A_Y} , respectively. Let P_{XY} denote the induced distribution on $(A_X \times A_Y, \mathcal{B}_{A_X} \times \mathcal{B}_{A_Y})$, that is, $P_{XY}(F \times G) = P(X \in F, Y \in G) = P(X^{-1}(F) \cap Y^{-1}(G))$. Let $\sigma(Y)$ denote the sub- σ -field of \mathcal{B} generated by Y , that is, $Y^{-1}(\mathcal{B}_{A_Y})$. Since the conditional probability $P(F|\sigma(Y))$ is real-valued and measurable with respect to $\sigma(Y)$, it can be written as

$g(Y(\omega))$, $\omega \in \Omega$, for some function $g(\gamma)$. (See, for example, Lemma 5.2.1 of [55] or Lemma 6.1 of [58].) Define $P(F|\gamma) = g(\gamma)$. For a fixed $F \in \mathcal{B}_{A_X}$ define the *conditional distribution* of F given $Y = \gamma$ by

$$P_{X|Y}(F|\gamma) = P(X^{-1}(F)|\gamma); \gamma \in \mathcal{B}_{A_Y}.$$

From the properties of conditional probability,

$$P_{XY}(F \times G) = \int_G P_{X|Y}(F|\gamma) dP_Y(\gamma); F \in \mathcal{B}_{A_X}, G \in \mathcal{B}_{A_Y}. \quad (1.13)$$

It is tempting to think that for a fixed γ , the set function defined by $P_{X|Y}(F|\gamma); F \in \mathcal{B}_{A_X}$ is actually a probability measure. This is not the case in general. When it does hold for a conditional probability measure, the conditional probability measure is said to be *regular*. This text will focus on standard alphabets for which regular conditional probabilities always exist.

1.2 Random Processes and Dynamical Systems

We now consider two mathematical models for a source: A random process and a dynamical system. The first is the familiar one in elementary courses, a source is just a random process or sequence of random variables. The second model is possibly less familiar — a random process can also be constructed from an abstract dynamical system consisting of a probability space together with a transformation on the space. The two models are connected by considering a time shift to be a transformation.

A *discrete time random process* or, simply, a *random process* is a sequence of random variables $\{X_n\}_{n \in \mathbb{T}}$ or $\{X_n; n \in \mathbb{T}\}$, where \mathbb{T} is an index set, defined on a common probability space (Ω, \mathcal{B}, P) . We define a *source* as a random process, although we could also use the alternative definition of a dynamical system to be introduced shortly. We usually assume that all of the random variables share a common alphabet, say A . The two most common index sets of interest are the set of all integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$, in which case the random process is referred to as a *two-sided* random process, and the set of all nonnegative integers $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$, in which case the random process is said to be *one-sided*. One-sided random processes will often prove to be far more difficult in theory, but they provide better models for physical random processes that must be “turned on” at some time or which have transient behavior.

Observe that since the alphabet A is general, we could also model continuous time random processes in the above fashion by letting A

consist of a family of waveforms defined on an interval, e.g., the random variable X_n could in fact be a continuous time waveform $X(t)$ for $t \in [nT, (n+1)T)$, where T is some fixed positive real number.

The above definition does not specify any structural properties of the index set \mathbb{T} . In particular, it does not exclude the possibility that \mathbb{T} be a finite set, in which case “random vector” would be a better name than “random process.” In fact, the two cases of $\mathbb{T} = \mathbb{Z}$ and $\mathbb{T} = \mathbb{Z}_+$ will be the only important examples for our purposes. Nonetheless, the general notation of \mathbb{T} will be retained in order to avoid having to state separate results for these two cases.

An abstract dynamical system consists of a probability space (Ω, \mathcal{B}, P) together with a measurable transformation $T : \Omega \rightarrow \Omega$ of Ω into itself. Measurability means that if $F \in \mathcal{B}$, then also $T^{-1}F = \{\omega : T\omega \in F\} \in \mathcal{B}$. The quadruple $(\Omega, \mathcal{B}, P, T)$ is called a *dynamical system* in ergodic theory. The interested reader can find excellent introductions to classical ergodic theory and dynamical system theory in the books of Halmos [73] and Sinai [170]. More complete treatments may be found in [16], [164], [149], [30], [191], [140], [46]. The term “dynamical systems” comes from the focus of the theory on the long term “dynamics” or “dynamical behavior” of repeated applications of the transformation T on the underlying measure space.

An alternative to modeling a random process as a sequence or family of random variables defined on a common probability space is to consider a single random variable together with a transformation defined on the underlying probability space. The outputs of the random process will then be values of the random variable taken on transformed points in the original space. The transformation will usually be related to shifting in time and hence this viewpoint will focus on the action of time itself. Suppose now that T is a measurable mapping of points of the sample space Ω into itself. It is easy to see that the cascade or composition of measurable functions is also measurable. Hence the transformation T^n defined as $T^2\omega = T(T\omega)$ and so on ($T^n\omega = T(T^{n-1}\omega)$) is a measurable function for all positive integers n . If f is an A -valued random variable defined on (Ω, \mathcal{B}) , then the functions $fT^n : \Omega \rightarrow A$ defined by $fT^n(\omega) = f(T^n\omega)$ for $\omega \in \Omega$ will also be random variables for all n in \mathbb{Z}_+ . Thus a dynamical system together with a random variable or measurable function f defines a one-sided random process $\{X_n\}_{n \in \mathbb{Z}_+}$ by $X_n(\omega) = f(T^n\omega)$. If it should be true that T is invertible, that is, T is one-to-one and its inverse T^{-1} is measurable, then one can define a two-sided random process by $X_n(\omega) = f(T^n\omega)$, all n in \mathbb{Z} .

The most common dynamical system for modeling random processes is that consisting of a sequence space Ω containing all one- or two-sided A -valued sequences together with the shift transformation T , that is, the transformation that maps a sequence $\{x_n\}$ into the sequence $\{x_{n+1}\}$ wherein each coordinate has been shifted to the left by one time unit.

Thus, for example, let $\Omega = A^{\mathbb{Z}^+} = \{\text{all } x = (x_0, x_1, \dots) \text{ with } x_i \in A \text{ for all } i\}$ and define $T : \Omega \rightarrow \Omega$ by $T(x_0, x_1, x_2, \dots) = (x_1, x_2, x_3, \dots)$. T is called the *shift* or *left shift* transformation on the one-sided sequence space. The shift for two-sided spaces is defined similarly. The sequence-space model of a random process is sometimes referred to as the Kolmogorov representation of a process.

The different models provide equivalent models for a given process — one emphasizing the sequence of outputs and the other emphasizing the action of a transformation on the underlying space in producing these outputs. In order to demonstrate in what sense the models are equivalent for given random processes, we next turn to the notion of the distribution of a random process.

1.3 Distributions

While in principle all probabilistic quantities associated with a random process can be determined from the underlying probability space, it is often more convenient to deal with the induced probability measures or distributions on the space of possible outputs of the random process. In particular, this allows us to compare different random processes without regard to the underlying probability spaces and thereby permits us to reasonably equate two random processes if their outputs have the same probabilistic structure, even if the underlying probability spaces are quite different.

We have already seen that each random variable X_n of the random process $\{X_n\}$ inherits a distribution because it is measurable. To describe a process, however, we need more than just probability measures on output values of separate individual random variables; we require probability measures on collections of random variables, that is, on sequences of outputs. In order to place probability measures on sequences of outputs of a random process, we first must construct the appropriate measurable spaces. A convenient technique for accomplishing this is to consider product spaces, spaces for sequences formed by concatenating spaces for individual outputs.

Let \mathbb{T} denote any finite or infinite set of integers. In particular, $\mathbb{T} = \mathbb{Z}(n) = \{0, 1, 2, \dots, n-1\}$, $\mathbb{T} = \mathbb{Z}$, or $\mathbb{T} = \mathbb{Z}_+$. Define $x^{\mathbb{T}} = \{x_i\}_{i \in \mathbb{T}}$. For example, $x^{\mathbb{Z}} = (\dots, x_{-1}, x_0, x_1, \dots)$ is a two-sided infinite sequence. When $\mathbb{T} = \mathbb{Z}(n)$ we abbreviate $x^{\mathbb{Z}(n)}$ to simply x^n . Given alphabets $A_i, i \in \mathbb{T}$, define the cartesian product space

$$\prod_{i \in \mathbb{T}} A_i = \{\text{all } x^{\mathbb{T}} : x_i \in A_i \text{ all } i \text{ in } \mathbb{T}\}.$$

In most cases all of the A_i will be replicas of a single alphabet A and the above product will be denoted simply by $A^{\mathbb{T}}$. Thus, for example, $A^{\{m, m+1, \dots, n\}}$ is the space of all possible outputs of the process from time m to time n ; $A^{\mathbb{Z}}$ is the sequence space of all possible outputs of a two-sided process. We shall abbreviate the notation for the space $A^{Z(n)}$, the space of all n dimensional vectors with coordinates in A , by A^n .

To obtain useful σ -fields of the above product spaces, we introduce the idea of a rectangle in a product space. A *rectangle* in $A^{\mathbb{T}}$ taking values in the coordinate σ -fields \mathcal{B}_i , $i \in \mathbb{J}$, is defined as any set of the form

$$B = \{x^{\mathbb{T}} \in A^{\mathbb{T}} : x_i \in B_i; \text{ all } i \text{ in } \mathbb{J}\}, \quad (1.14)$$

where \mathbb{J} is a finite subset of the index set \mathbb{T} and $B_i \in \mathcal{B}_i$ for all $i \in \mathbb{J}$. (Hence rectangles are sometimes referred to as finite dimensional rectangles.) A rectangle as in (1.14) can be written as a finite intersection of one-dimensional rectangles as

$$B = \bigcap_{i \in \mathbb{J}} \{x^{\mathbb{T}} \in A^{\mathbb{T}} : x_i \in B_i\} = \bigcap_{i \in \mathbb{J}} X_i^{-1}(B_i) \quad (1.15)$$

where here we consider X_i as the coordinate functions $X_i : A^{\mathbb{T}} \rightarrow A$ defined by $X_i(x^{\mathbb{T}}) = x_i$.

As rectangles in $A^{\mathbb{T}}$ are clearly fundamental events, they should be members of any useful σ -field of subsets of $A^{\mathbb{T}}$. Define the product σ -field $\mathcal{B}_A^{\mathbb{T}}$ as the smallest σ -field containing all of the rectangles, that is, the collection of sets that contains the clearly important class of rectangles and the minimum amount of other stuff required to make the collection a σ -field. To be more precise, given an index set \mathbb{T} of integers, let $RECT(\mathcal{B}_i, i \in \mathbb{T})$ denote the set of all rectangles in $A^{\mathbb{T}}$ taking coordinate values in sets in $\mathcal{B}_i, i \in \mathbb{T}$. We then define the product σ -field of $A^{\mathbb{T}}$ by

$$\mathcal{B}_A^{\mathbb{T}} = \sigma(RECT(\mathcal{B}_i, i \in \mathbb{T})). \quad (1.16)$$

Consider an index set \mathbb{T} and an A -valued random process $\{X_n\}_{n \in \mathbb{T}}$ defined on an underlying probability space (Ω, \mathcal{B}, P) . Given any index set $\mathbb{J} \subset \mathbb{T}$, measurability of the individual random variables X_n implies that of the random vectors $X^{\mathbb{J}} = \{X_n; n \in \mathbb{J}\}$. Thus the measurable space $(A^{\mathbb{J}}, \mathcal{B}_A^{\mathbb{J}})$ inherits a probability measure from the underlying space through the random variables $X^{\mathbb{J}}$. Thus in particular the measurable space $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}})$ inherits a probability measure from the underlying probability space and thereby determines a new probability space $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, P_{X^{\mathbb{T}}})$, where the induced probability measure is defined by

$$P_{X^{\mathbb{T}}}(F) = P((X^{\mathbb{T}})^{-1}(F)) = P(\omega : X^{\mathbb{T}}(\omega) \in F); F \in \mathcal{B}_A^{\mathbb{T}}. \quad (1.17)$$

Such probability measures induced on the outputs of random variables are referred to as *distributions* for the random variables, exactly as in the simpler case first treated. When $\mathbb{T} = \{m, m+1, \dots, m+n-1\}$, e.g., when we are treating $X_m^n = (X_n, \dots, X_{m+n-1})$ taking values in A^n , the distribution is referred to as an n -dimensional or n th order distribution and it describes the behavior of an n -dimensional random variable. If \mathbb{T} is the entire process index set, e.g., if $\mathbb{T} = \mathbb{Z}$ for a two-sided process or $\mathbb{T} = \mathbb{Z}_+$ for a one-sided process, then the induced probability measure is defined to be the distribution of the process. Thus, for example, a probability space (Ω, \mathcal{B}, P) together with a doubly infinite sequence of random variables $\{X_n\}_{n \in \mathbb{Z}}$ induces a new probability space $(A^{\mathbb{Z}}, \mathcal{B}_A^{\mathbb{Z}}, P_{X^{\mathbb{Z}}})$ and $P_{X^{\mathbb{Z}}}$ is the distribution of the process. For simplicity, let us now denote the process distribution simply by m . We shall call the probability space $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, m)$ induced in this way by a random process $\{X_n\}_{n \in \mathbb{Z}}$ the output space or sequence space of the random process.

Since the sequence space $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, m)$ of a random process $\{X_n\}_{n \in \mathbb{Z}}$ is a probability space, we can define random variables and hence also random processes on this space. One simple and useful such definition is that of a sampling or coordinate or projection function defined as follows: Given a product space $A^{\mathbb{T}}$, define the sampling functions $\Pi_n : A^{\mathbb{T}} \rightarrow A$ by

$$\Pi_n(x^{\mathbb{T}}) = x_n, x^{\mathbb{T}} \in A^{\mathbb{T}}; n \in \mathbb{T}. \quad (1.18)$$

The sampling function is named Π since it is also a projection. Observe that the distribution of the random process $\{\Pi_n\}_{n \in \mathbb{T}}$ defined on the probability space $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, m)$ is exactly the same as the distribution of the random process $\{X_n\}_{n \in \mathbb{T}}$ defined on the probability space (Ω, \mathcal{B}, P) . In fact, so far they are the same process since the $\{\Pi_n\}$ simply read off the values of the $\{X_n\}$.

What happens, however, if we no longer build the Π_n on the X_n , that is, we no longer first select ω from Ω according to P , then form the sequence $x^{\mathbb{T}} = X^{\mathbb{T}}(\omega) = \{X_n(\omega)\}_{n \in \mathbb{T}}$, and then define $\Pi_n(x^{\mathbb{T}}) = X_n(\omega)$? Instead we directly choose an x in $A^{\mathbb{T}}$ using the probability measure m and then view the sequence of coordinate values. In other words, we are considering two completely separate experiments, one described by the probability space (Ω, \mathcal{B}, P) and the random variables $\{X_n\}$ and the other described by the probability space $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, m)$ and the random variables $\{\Pi_n\}$. In these two separate experiments, the actual sequences selected may be completely different. Yet intuitively the processes should be the “same” in the sense that their statistical structures are identical, that is, they have the same distribution. We make this intuition formal by defining two processes to be *equivalent* if their process distributions are identical, that is, if the probability measures on the output sequence spaces are the same, regardless of the functional form of the random variables of the underlying probability spaces. In the same way, we con-

sider two random variables to be equivalent if their distributions are identical.

We have described above two equivalent processes or two equivalent models for the same random process, one defined as a sequence of random variables on a perhaps very complicated underlying probability space, the other defined as a probability measure directly on the measurable space of possible output sequences. The second model will be referred to as a *directly given* random process or as the *Kolmogorov* model for the random process.

Which model is “better” depends on the application. For example, a directly given model for a random process may focus on the random process itself and not its origin and hence may be simpler to deal with. If the random process is then coded or measurements are taken on the random process, then it may be better to model the encoded random process in terms of random variables defined on the original random process and not as a directly given random process. This model will then focus on the input process and the coding operation. We shall let convenience determine the most appropriate model.

We can now describe yet another model for the above random process, that is, another means of describing a random process with the same distribution. This time the model is in terms of a dynamical system. Given the probability space $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, m)$, define the (left) shift transformation $T : A^{\mathbb{T}} \rightarrow A^{\mathbb{T}}$ by

$$T(x^{\mathbb{T}}) = T(\{x_n\}_{n \in \mathbb{T}}) = y^{\mathbb{T}} = \{y_n\}_{n \in \mathbb{T}},$$

where

$$y_n = x_{n+1}, n \in \mathbb{T}.$$

Thus the n th coordinate of $y^{\mathbb{T}}$ is simply the $(n + 1)$ st coordinate of $x^{\mathbb{T}}$. (We assume that \mathbb{T} is closed under addition and hence if n and 1 are in \mathbb{T} , then so is $(n + 1)$.) If the alphabet of such a shift is not clear from context, we will occasionally denote the shift by T_A or $T_{A^{\mathbb{T}}}$. The shift can easily be shown to be measurable.

Consider next the dynamical system $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, P, T)$ and the random process formed by combining the dynamical system with the zero time sampling function Π_0 (we assume that 0 is a member of \mathbb{T}). If we define $Y_n(x) = \Pi_0(T^n x)$ for $x = x^{\mathbb{T}} \in A^{\mathbb{T}}$, or, in abbreviated form, $Y_n = \Pi_0 T^n$, then the random process $\{Y_n\}_{n \in \mathbb{T}}$ is equivalent to the processes developed above. Thus we have developed three different, but equivalent, means of producing the same random process. Each will be seen to have its uses.

The above development shows that a dynamical system is a more fundamental entity than a random process since we can always construct an equivalent model for a random process in terms of a dynamical system — use the directly given representation, shift transformation, and zero

time sampling function. Two important properties of dynamical systems or random processes can be defined at this point, the implications will be developed throughout the book. A dynamical system $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, P, T)$ is said to be *stationary* (with respect to T) if the distribution P is invariant with respect to P , that is,

$$P(T^{-1}F) = P(F), \text{ all } F \in \mathcal{B}_A^{\mathbb{T}}. \quad (1.19)$$

In other words, probabilities of process events are unchanged by shifting. The dynamical system is said to be *ergodic* if

$$\text{If } T^{-1}F = F, \text{ then } P(F) = 0 \text{ or } 1, \quad (1.20)$$

that is, all invariant events are trivial. Note that neither definition implies or excludes the other.

The shift transformation on a sequence space introduced above is the most important transformation that we shall encounter. It is not, however, the only important transformation. When dealing with transformations we will usually use the notation T to reflect the fact that it is often related to the action of a simple left shift of a sequence, yet it should be kept in mind that occasionally other operators will be considered and the theory to be developed will remain valid, even if T is not required to be a simple time shift. For example, we will also consider block shifts.

Most texts on ergodic theory deal with the case of an invertible transformation, that is, where T is a one-to-one transformation and the inverse mapping T^{-1} is measurable. This is the case for the shift on $A^{\mathbb{Z}}$, the two-sided shift. It is not the case, however, for the one-sided shift defined on $A^{\mathbb{Z}^+}$ and hence we will avoid use of this assumption. We will, however, often point out in the discussion what simplifications or special properties arise for invertible transformations.

Since random processes are considered equivalent if their distributions are the same, we shall adopt the notation $[A, m, X]$ for a random process $\{X_n; n \in \mathbb{T}\}$ with alphabet A and process distribution m , the index set \mathbb{T} usually being clear from context. We will occasionally abbreviate this to the more common notation $[A, m]$, but it is often convenient to note the name of the output random variables as there may be several, e.g., a random process may have an input X and output Y . By “the associated probability space” of a random process $[A, m, X]$ we shall mean the sequence probability space $(A^{\mathbb{T}}, \mathcal{B}_A^{\mathbb{T}}, m)$. It will often be convenient to consider the random process as a directly given random process, that is, to view X_n as the coordinate functions Π_n on the sequence space $A^{\mathbb{T}}$ rather than as being defined on some other abstract space. This will not always be the case, however, as often processes will be formed by coding or communicating other random processes. Context should render such bookkeeping details clear.

1.4 Standard Alphabets

A measurable space (A, \mathcal{B}_A) is a *standard space* if there exists a sequence of finite fields \mathcal{F}_n ; $n = 1, 2, \dots$ with the following properties:

- (1) $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ (the fields are increasing).
- (2) \mathcal{B}_A is the smallest σ -field containing all of the \mathcal{F}_n (the \mathcal{F}_n generate \mathcal{B}_A or $\mathcal{B}_A = \sigma(\bigcup_{n=1}^{\infty} \mathcal{F}_n)$).
- (3) An event $G_n \in \mathcal{F}_n$ is called an *atom* of the field if it is nonempty and its only subsets which are also field members are itself and the empty set. If $G_n \in \mathcal{F}_n$; $n = 1, 2, \dots$ are atoms and $G_{n+1} \subset G_n$ for all n , then

$$\bigcap_{n=1}^{\infty} G_n \neq \emptyset.$$

Standard spaces are important for several reasons: First, they are a general class of spaces for which two of the key results of probability hold: (1) the Kolmogorov extension theorem showing that a random process is completely described by its finite order distributions, and (2) the existence of regular conditional probability measures. Thus, in particular, the conditional probability measure $P_{X|Y}(F|\mathcal{Y})$ of (1.13) is regular if the alphabets A_X and A_Y are standard and hence for each fixed $\mathcal{Y} \in A_Y$ the set function $P_{X|Y}(F|\mathcal{Y})$; $F \in \mathcal{B}_{A_X}$ is a probability measure. In this case we can interpret $P_{X|Y}(F|\mathcal{Y})$ as $P(X \in F | Y = \mathcal{Y})$. Second, the ergodic decomposition theorem of ergodic theory holds for such spaces. The ergodic decomposition implies that any stationary process is equivalent to a mixture of stationary and ergodic processes; that is, a stationary nonergodic source can be viewed as a random selection of one of a family of stationary and ergodic sources. Third, the class is sufficiently general to include virtually all examples arising in applications, e.g., discrete spaces, the real line, Euclidean vector spaces, Polish spaces (complete separable metric spaces), etc. The reader is referred to [55] or [58] and the references cited therein for a detailed development of these properties and examples of standard spaces.

Standard spaces are not the most general space for which the Kolmogorov extension theorem, the existence of conditional probability, and the ergodic decomposition theorem all hold. These results also hold for perfect spaces which include standard spaces as a special case. (See, e.g., [161],[174],[155], [114].) We limit discussion to standard spaces, however, as they are easier to characterize and work with and they are sufficiently general to handle most cases encountered in applications. Although standard spaces are not the most general for which the required probability theory results hold, they are the most general for which all finitely additive normalized measures extend to countably additive prob-

ability measures, a property which greatly eases the proof of many of the desired results.

Throughout this book we shall assume that the alphabet A of the information source is a standard space.

1.5 Expectation

Let (Ω, \mathcal{B}, m) be a probability space, e.g., the probability space of a directly given random process with alphabet A , $(A^{\mathbb{T}}, B_A^{\mathbb{T}}, m)$. A real-valued random variable $f : \Omega \rightarrow \mathbb{R}$ will also be called a *measurement* since it is often formed by taking a mapping or function of some other set of more general random variables, e.g., the outputs of some random process which might not have real-valued outputs. Measurements made on such processes, however, will always be assumed to be real.

Suppose next we have a measurement f whose range space or *alphabet* $f(\Omega) \subset \mathbb{R}$ of possible values is finite. Then f is called a *discrete random variable* or *discrete measurement* or *digital measurement* or, in the common mathematical terminology, a *simple function*.

Given a discrete measurement f , suppose that its range space is $f(\Omega) = \{b_i, i = 1, \dots, N\}$, where the b_i are distinct. Define the sets $F_i = f^{-1}(b_i) = \{x : f(x) = b_i\}$, $i = 1, \dots, N$. Since f is measurable, the F_i are all members of \mathcal{B} . Since the b_i are distinct, the F_i are disjoint. Since every input point in Ω must map into some b_i , the union of the F_i equals Ω . Thus the collection $\{F_i; i = 1, 2, \dots, N\}$ forms a partition of Ω . We have therefore shown that any discrete measurement f can be expressed in the form

$$f(x) = \sum_{i=1}^M b_i 1_{F_i}(x), \quad (1.21)$$

where $b_i \in \mathbb{R}$, the $F_i \in \mathcal{B}$ form a partition of Ω , and 1_{F_i} is the indicator function of F_i , $i = 1, \dots, M$. Every simple function has a unique representation in this form with distinct b_i and $\{F_i\}$ a partition.

The *expectation* or *ensemble average* or *probabilistic average* or *mean* of a discrete measurement $f : \Omega \rightarrow \mathbb{R}$ as in (1.21) with respect to a probability measure m is defined by

$$E_m f = \sum_{i=1}^M b_i m(F_i). \quad (1.22)$$

An immediate consequence of the definition of expectation is the simple but useful fact that for any event F in the original probability space,

$$E_m 1_F = m(F),$$

that is, probabilities can be found from expectations of indicator functions.

Again let (Ω, \mathcal{B}, m) be a probability space and $f : \Omega \rightarrow \mathbb{R}$ a measurement, that is, a real-valued random variable or measurable real-valued function. Define the sequence of *quantizers* $q_n : \mathbb{R} \rightarrow \mathbb{R}$, $n = 1, 2, \dots$, as follows:

$$q_n(r) = \begin{cases} n & n \leq r \\ (k-1)2^{-n} & (k-1)2^{-n} \leq r < k2^{-n}, k = 1, 2, \dots, n2^n \\ -(k-1)2^{-n} & -k2^{-n} \leq r < -(k-1)2^{-n}; k = 1, 2, \dots, n2^n \\ -n & r < -n. \end{cases}$$

We now define expectation for general measurements in two steps. If $f \geq 0$, then define

$$E_m f = \lim_{n \rightarrow \infty} E_m(q_n(f)). \quad (1.23)$$

Since the q_n are discrete measurements on f , the $q_n(f)$ are discrete measurements on Ω ($q_n(f)(x) = q_n(f(x))$ is a simple function) and hence the individual expectations are well defined. Since the $q_n(f)$ are nondecreasing, so are the $E_m(q_n(f))$ and this sequence must either converge to a finite limit or grow without bound, in which case we say it converges to ∞ . In both cases the expectation $E_m f$ is well defined, although it may be infinite.

If f is an arbitrary real random variable, define its positive and negative parts $f^+(x) = \max(f(x), 0)$ and $f^-(x) = -\min(f(x), 0)$ so that $f(x) = f^+(x) - f^-(x)$ and set

$$E_m f = E_m f^+ - E_m f^- \quad (1.24)$$

provided this does not have the form $+\infty - \infty$, in which case the expectation does not exist. It can be shown that the expectation can also be evaluated for nonnegative measurements by the formula

$$E_m f = \sup_{\text{discrete } g: g \leq f} E_m g.$$

The expectation is also called an *integral* and is denoted by any of the following:

$$E_m f = \int f dm = \int f(x) dm(x) = \int f(x) m(dx).$$

The subscript m denoting the measure with respect to which the expectation is taken will occasionally be omitted if it is clear from context.

A measurement f is said to be *integrable* or *m-integrable* if $E_m f$ exists and is finite. A function is integrable if and only if its absolute value is

integrable. Define $L^1(m)$ to be the space of all m -integrable functions. Given any m -integrable f and an event B , define

$$\int_B f dm = \int f(x) 1_B(x) dm(x).$$

Two random variables f and g are said to be equal m -almost-everywhere or equal m -a.e. or equal with m -probability one if $m(f = g) = m(\{x : f(x) = g(x)\}) = 1$. The m - is dropped if it is clear from context.

Given a probability space (Ω, \mathcal{B}, m) , suppose that \mathcal{G} is a sub- σ -field of \mathcal{B} , that is, it is a σ -field of subsets of Ω and all those subsets are in \mathcal{B} ($\mathcal{G} \subset \mathcal{B}$). Let $f : \Omega \rightarrow \mathbb{R}$ be an integrable measurement. Then the *conditional expectation* $E(f|\mathcal{G})$ is described as any function, say $h(\omega)$, that satisfies the following two properties:

$$h(\omega) \text{ is measurable with respect to } \mathcal{G} \quad (1.25)$$

$$\int_G h dm = \int_G f dm; \text{ all } G \in \mathcal{G}. \quad (1.26)$$

If a regular conditional probability distribution given \mathcal{G} exists, e.g., if the space is standard, then one has a constructive definition of conditional expectation: $E(f|\mathcal{G})(\omega)$ is simply the expectation of f with respect to the conditional probability measure $m(\cdot|\mathcal{G})(\omega)$. Applying this to the example of two random variables X and Y with standard alphabets described in Section 1.2 we have from (1.26) that for integrable $f : A_X \times A_Y \rightarrow \mathbb{R}$

$$E(f) = \int f(x, y) dP_{XY}(x, y) = \int \left(\int f(x, y) dP_{X|Y}(x|y) \right) dP_Y(y). \quad (1.27)$$

In particular, for fixed y , $f(x, y)$ is an integrable (and measurable) function of x .

Equation (1.27) provides a generalization of (1.13) from rectangles to arbitrary events. For an arbitrary $F \in \mathcal{B}_{A_X \times A_Y}$ we have that

$$P_{XY}(F) = \int \left(\int 1_F(x, y) dP_{X|Y}(x|y) \right) dP_Y(y) = \int P_{X|Y}(F_y|y) dP_Y(y), \quad (1.28)$$

where $F_y = \{x : (x, y) \in F\}$ is called the *section* of F at y . If F is measurable, then so is F_y for all y . Alternatively, since $1_F(x, y)$ is measurable with respect to x for each fixed y , $F_y \in \mathcal{B}_{A_X}$ and the inner integral is just

$$\int_{x:(x,y) \in F} dP_{X|Y}(x|y) = P_{X|Y}(F_y|y).$$

1.6 Asymptotic Mean Stationarity

Recall that a dynamical system (or the associated source) $(\Omega, \mathcal{B}, P, T)$ is said to be stationary if $P(T^{-1}G) = P(G)$ for all $G \in \mathcal{B}$. It is said to be *asymptotically mean stationary* or, simply, AMS if the limit

$$\bar{P}(G) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} P(T^{-k}G) \quad (1.29)$$

exists for all $G \in \mathcal{B}$. The following theorems summarize several important properties of AMS sources. Details may be found in Chapter 6 of [55] or Chapter 7 of [58].

Theorem 1.2. *If a dynamical system $(\Omega, \mathcal{B}, P, T)$ is AMS, then \bar{P} defined in (1.29) is a probability measure and $(\Omega, \mathcal{B}, \bar{P}, T)$ is stationary. The distribution \bar{P} is called the stationary mean of P . If an event G is invariant in the sense that $T^{-1}G = G$, then*

$$P(G) = \bar{P}(G).$$

If a random variable g is invariant in the sense that $g(Tx) = g(x)$ with P probability 1, then

$$E_P g = E_{\bar{P}} g.$$

The stationary mean \bar{P} *asymptotically dominates* P in the sense that if $\bar{P}(G) = 0$, then

$$\limsup_{n \rightarrow \infty} P(T^{-n}G) = 0.$$

Theorem 1.3. *Given an AMS source $\{X_n\}$ let $\sigma(X_n, X_{n+1}, \dots)$ denote the σ -field generated by the random variables X_n, \dots , that is, the smallest σ -field with respect to which all these random variables are measurable. Define the tail σ -field \mathcal{F}_∞ by*

$$\mathcal{F}_\infty = \bigcap_{n=0}^{\infty} \sigma(X_n, \dots).$$

If $G \in \mathcal{F}_\infty$ and $\bar{P}(G) = 0$, then also $P(G) = 0$.

The tail σ -field can be thought of as events that are determinable by looking only at samples of the sequence in the arbitrarily distant future. The theorem states that the stationary mean *dominates* the original measure on such tail events in the sense that zero probability under the stationary mean implies zero probability under the original source.

1.7 Ergodic Properties

Two of the basic results of ergodic theory that will be called upon extensively are the pointwise or almost-everywhere ergodic theorem and the ergodic decomposition theorem. We quote these results along with some relevant notation for reference. Detailed developments may be found in Chapters 6–8 of [55] or Chapters 7–10 of [58]. The ergodic theorem states that AMS dynamical systems (and hence also sources) have convergent sample averages, and it characterizes the limits.

Theorem 1.4. *If a dynamical system $(\Omega, \mathcal{B}, m, T)$ is AMS with stationary mean \bar{m} and if $f \in L^1(\bar{m})$, then with probability one under m and \bar{m}*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} fT^i = E_{\bar{m}}(f|I),$$

where I is the sub- σ -field of invariant events, that is, events G for which $T^{-1}G = G$.

The basic idea of the ergodic decomposition is that any stationary source which is not ergodic can be represented as a mixture of stationary ergodic components or subsources.

Theorem 1.5. *Ergodic Decomposition Given the standard sequence space (Ω, \mathcal{B}) with shift T as previously, there exists a family of stationary ergodic measures $\{p_x; x \in \Omega\}$, called the ergodic decomposition, with the following properties:*

(a) $p_{Tx} = p_x$.

(b) For any stationary measure m ,

$$m(G) = \int p_x(G) dm(x); \text{ all } G \in \mathcal{B}.$$

(c) For any $g \in L^1(m)$

$$\int g dm = \int \left(\int g dp_x \right) dm(x).$$

It is important to note that the same collection of stationary ergodic components works for any stationary measure m . This is the strong form of the ergodic decomposition.

The final result of this section is a variation on the ergodic decomposition. To describe the result, we need to digress briefly to introduce a metric on spaces of probability measures. A thorough development can be found in Chapter 8 of [55] or Chapter 9 of [58]. We have a standard sequence measurable space (Ω, \mathcal{B}) and hence we can generate the σ -field \mathcal{B}

by a countable field $\mathcal{F} = \{F_n; n = 1, 2, \dots\}$. Given such a countable generating field, a *distributional distance* between two probability measures p and m on (Ω, \mathcal{B}) is defined by

$$d(p, m) = \sum_{n=1}^{\infty} 2^{-n} |p(F_n) - m(F_n)|.$$

Any choice of a countable generating field yields a distributional distance. Such a distance or metric yields a measurable space of probability measures as follows: Let Λ denote the space of all probability measures on the original measurable space (Ω, \mathcal{B}) . Let $\mathcal{B}(\Lambda)$ denote the σ -field of subsets of Λ generated by all open spheres using the distributional distance, that is, all sets of the form $\{p : d(p, m) \leq \epsilon\}$ for some $m \in \Lambda$ and some $\epsilon > 0$. We can now consider properties of functions that carry sequences in our original space into probability measures. The following is Theorem 8.5.1 of [55] and Theorem 10.1 of [58].

Theorem 1.6. *A Variation on the Ergodic Decomposition Fix a standard measurable space (Ω, \mathcal{B}) and a transformation $T : \Omega \rightarrow \Omega$. Then there are a standard measurable space (Λ, \mathcal{L}) , a family of stationary ergodic measures $\{m_\lambda; \lambda \in \Lambda\}$ on (Ω, \mathcal{B}) , and a measurable mapping $\psi : \Omega \rightarrow \Lambda$ such that*

- (a) ψ is invariant ($\psi(Tx) = \psi(x)$ all x);
- (b) if m is a stationary measure on (Ω, \mathcal{B}) and P_ψ is the induced distribution; that is, $P_\psi(G) = m(\psi^{-1}(G))$ for $G \in \mathcal{L}$ (which is well defined from (a)), then

$$m(F) = \int dm(x) m_{\psi(x)}(F) = \int dP_\psi(\lambda) m_\lambda(F), \text{ all } F \in \mathcal{B},$$

and if $f \in L^1(m)$, then so is $\int f dm_\lambda$ P_ψ -a.e. and

$$E_m f = \int dm(x) E_{m_{\psi(x)}} f = \int dP_\psi(\lambda) E_{m_\lambda} f.$$

Finally, for any event F , $m_\psi(F) = m(F|\psi)$, that is, given the ergodic decomposition and a stationary measure m , the ergodic component λ is a version of the conditional probability under m given $\psi = \lambda$.

The following corollary to the ergodic decomposition is Lemma 8.6.2 of [55] and Lemma 10.4 of [58]. It states that the conditional probability of a future event given the entire past is unchanged by knowing the ergodic component in effect. This is because the infinite past determines the ergodic component in effect.

Corollary 1.1. *Suppose that $\{X_n\}$ is a two-sided stationary process with distribution m and that $\{m_\lambda; \lambda \in \Lambda\}$ is the ergodic decomposition and ψ*

the ergodic component function. Then the mapping ψ is measurable with respect to $\sigma(X_{-1}, X_{-2}, \dots)$ and

$$m((X_0, X_1, \dots) \in F | X_{-1}, X_{-2}, \dots) = m_\psi((X_0, X_1, \dots) \in F | X_{-1}, X_{-2}, \dots); \text{ } m - \text{a.e.}$$



<http://www.springer.com/978-1-4419-7969-8>

Entropy and Information Theory

Gray, R.M.

2011, XXVII, 409 p., Hardcover

ISBN: 978-1-4419-7969-8