A

Probability and Random Processes

In this appendix we will look at some of the concepts relating to probability and random processes that are important in the study of systems. Our coverage will be highly selective and somewhat superficial, but enough to use probability and random processes as a tool in understanding data compression systems.

A.1 Probability

There are several different ways of defining and thinking about probability. Each approach has some merit; perhaps the best approach is the one that provides the most insight into the problem being studied.

A.1.1 Frequency of Occurrence

The most common way that most people think about probability is in terms of outcomes, or sets of outcomes, of an experiment. Let us suppose we conduct an experiment E that has N possible outcomes. We conduct the experiment n_T times. If the outcome ω_i occurs n_i times, we say that the frequency of occurrence of the outcome ω_i is $\frac{n_i}{n_T}$. We can then define the probability of occurrence of the outcome ω_i as

$$P(\omega_i) = \lim_{n_T \to \infty} \frac{n_i}{n_T}.$$

In practice we do not have the ability to repeat an experiment an infinite number of times, so we often use the frequency of occurrence as an approximation to the probability. To make this more concrete consider a specific experiment. Suppose we turn on a television 1,000,000 times. Of these times, 800,000 times we turn the television on during a commercial and 200,000 times we turn it on and we don't get a commercial. We could say the frequency of occurrence, or the estimate of the probability, of turning on a television set in the middle of a commercial is 0.8. Our experiment E here is the turning on a television set, and the outcomes are *commercial* and *no commercial.* We could have been more careful with noting what was on when we turned on the television set and noticed whether the program was a news program (2000 times), a newslike program (20,000 times), a comedy program (40,000 times), an adventure program (18,000 times), a variety show (20,000 times), a talk show (90,000 times), or a movie (10,000 times), and whether the commercial was for products or services. In this case the outcomes would be *product commercial, service commercial, comedy, adventure, news, pseudonews, variety, talk show*, and *movie*. We could then define an *event* as a set of outcomes. The event *commercial* would consist of the outcomes *product commercial, service commercial*; the event *no commercial* would consist of the outcomes *comedy, adventure, news, pseudonews, variety, talk show, movie*. We could also define other events such as *programs that may contain news*. This set would contain the outcomes *news, pseudonews*, and *talk shows*, and the frequency of occurrence of this set is 0.112.

Formally, when we define an experiment E, associated with the experiment we also define a *sample space* S that consists of the *outcomes* $\{\omega_i\}$. We can then combine these outcomes into sets that are called *events*, and assign probabilities to these events. The largest subset of S (event) is S itself, and the probability of the event S is simply the probability that the experiment will have an outcome. The way we have defined things, this probability is one; that is, $P(S) = 1$.

A.1.2 A Measure of Belief

Sometimes the idea that the probability of an event is obtained through the repetitions of an experiment runs into trouble. What, for example, is the probability of your getting from Logan Airport to a specific address in Boston in a specified period of time? The answer depends on a lot of different factors, including your knowledge of the area, the time of day, the condition of your transport, and so on. You cannot conduct an experiment and get your answer because the moment you conduct the experiment, the conditions have changed, and the answer will now be different. We deal with this situation by defining a priori and a posteriori probabilities. The a priori probability is what you think or believe the probability to be before certain information is received or certain events take place; the a posteriori probability is the probability after you have received further information. Probability is no longer as rigidly defined as in the frequency of occurrence approach but is a somewhat more fluid quantity, the value of which changes with changing experience. For this approach to be useful we have to have a way of describing how the probability evolves with changing information. This is done through the use of *Bayes' rule*, named after the person who first described it. If $P(A)$ is the a priori probability of the event A and $P(A|B)$ is the a posteriori probability of the event A given that the event B has occurred, then

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{A.1}$$

where $P(A, B)$ is the probability of the event A *and* the event B occurring. Similarly,

$$P(B|A) = \frac{P(A, B)}{P(A)}. \tag{A.2}$$

Combining (A.1) and (A.2) we get

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{A.3}$$

If the events A and B do not provide any information about each other, it would be reasonable to assume that

$$P(A|B) = P(A)$$

and therefore from (A.1),

$$P(A, B) = P(A)P(B). \tag{A.4}$$

Whenever (A.4) is satisfied, the events A and B are said to be *statistically independent*, or simply *independent*.

Example A.1.1:

A very common channel model used in digital communication is the *binary symmetric channel*. In this model the input is a random experiment with outcomes 0 and 1. The output of the channel is another random event with two outcomes 0 and 1. Obviously, the two outcomes are connected in some way. To see how, let us first define some events:

A: Input is 0
B: Input is 1
C: Output is 0
D: Output is 1

Let's suppose the input is equally likely to be a 1 or a 0. So $P(A) = P(B) = 0.5$. If the channel was perfect, that is, you got out of the channel what you put in, then we would have

$$P(C|A) = P(D|B) = 1$$

and

$$P(C|B) = P(D|A) = 0.$$

With most real channels this system is seldom encountered, and generally there is a small probability ϵ that the transmitted bit will be received in error. In this case, our probabilities would be

$$\begin{aligned} P(C|A) &= P(D|B) = 1 - \epsilon \\ P(C|B) &= P(D|A) = \epsilon. \end{aligned}$$

How do we interpret $P(C)$ and $P(D)$? These are simply the probability that at any given time the output is a 0 or a 1. How would we go about computing these probabilities given the available information? Using (A.1) we can obtain $P(A, C)$ and $P(B, C)$ from $P(C|A)$, $P(C|B)$, $P(A)$, and $P(B)$. These are the probabilities that the input is 0 and the output is 1, and the input is 1 and the output is 1. The event C—that is, the output is 1—will occur only when one of the two *joint* events occurs, therefore,

$$P(C) = P(A, C) + P(B, C).$$

Similarly,

$$P(D) = P(A, D) + P(B, D).$$

Numerically, this comes out to be

$$P(C) = P(D) = 0.5.$$ ♦

A.1.3 The Axiomatic Approach

Finally, there is an approach that simply defines probability as a measure, without much regard for physical interpretation. We are very familiar with measures in our daily lives. We talk about getting a 9-foot cable or a pound of cheese. Just as length and width measure the extent of certain physical quantities, probability measures the extent of an abstract quantity, a set. The thing that probability measures is the "size" of the event set. The probability measure follows similar rules to those followed by other measures. Just as the length of a physical object is always greater than or equal to zero, the probability of an event is always greater than or equal to zero. If we measure the length of two objects that have no overlap, then the combined length of the two objects is simply the sum of the lengths of the individual objects. In a similar manner the probability of the union of two events that do not have any outcomes in common is simply the sum of the probability of the individual events. So as to keep this definition of probability in line with the other definitions, we normalize this quantity by assigning the largest set, which is the sample space S, the size of 1. Thus, the probability of an event always lies between 0 and 1. Formally, we can write these rules down as the three *axioms* of probability.

Given a sample space S:

- *Axiom 1:* If A is an event in S, then $P(A) \geq 0$.
- *Axiom 2:* The probability of the sample space is 1; that is, $P(S) = 1$.
- *Axiom 3:* If A and B are two events in S and $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$.

Given these three axioms we can come up with all the other rules we need. For example, suppose A^c is the complement of A. What is the probability of A^c? We can get the answer by using Axiom 2 and Axiom 3. We know that

$$A^c \cup A = S$$

and Axiom 2 tells us that $P(S) = 1$, therefore,

$$P(A^c \cup A) = 1. \tag{A.5}$$

We also know that $A^c \cap A = \phi$, therefore, from Axiom 3

$$P(A^c \cup A) = P(A^c) + P(A). \tag{A.6}$$

Combining equations (A.5) and (A.6), we get

$$P(A^c) = 1 - P(A). \tag{A.7}$$

Similarly, we can use the three axioms to obtain the probability of $A \cup B$ when $A \cap B \neq \phi$ as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{A.8}$$

In all of the above we have been using two events A and B. We can easily extend these rules to more events.

Example A.1.2:

Find $P(A \cup B \cup C)$ when $A \cap B = A \cap C = \phi$, and $B \cup C \neq \phi$.

Let

$$D = B \cup C.$$

Then

$$A \cap C = \phi, \quad A \cap B = \phi \qquad \Rightarrow \qquad A \cap D = \phi.$$

Therefore, from Axiom 3,

$$P(A \cup D) = P(A) + P(D)$$

and using (A.8)

$$P(D) = P(B) + P(C) - P(B \cap C).$$

Combining everything, we get

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(B \cap C).$$ ♦

The axiomatic approach is especially useful when an experiment does not have discrete outcomes. For example, if we are looking at the voltage on a telephone line, the probability of any specific value of the voltage is zero because there are an uncountably infinite number of different values that the voltage can take, and we can assign nonzero values to only a countably infinite number. Using the axiomatic approach, we can view the sample space as the range of voltages, and events as subsets of this range.

We have given three different interpretations of probability, and in the process described some rules by which probabilities can be manipulated. The rules described here (such as

Bayes' rule, the three axioms, and the other rules we came up with) work the same way regardless of which interpretation you hold dear. The purpose of providing you with three different interpretations was to provide you with a variety of perspectives with which to view a given situation. For example, if someone says that the probability of a head when you flip a coin is 0.5, you might interpret that number in terms of repeated experiments (*if I flipped the coin 1000 times, I would expect to get 500 heads*). However, if someone tells you that the probability of your getting killed while crossing a particular street is 0.1, you might wish to interpret this information in a more subjective manner. The idea is to use the interpretation that gives you the most insight into a particular problem, while remembering that your interpretation will not change the mathematics of the situation.

Now that have expended a lot of verbiage to say what probability is, let's spend a few lines saying what it is not. Probability does not imply certainty. When we say that the probability of an event is one, this does not mean that event *will* happen. On the other hand, when we say that the probability of an event is zero, that does not mean that event *won't* happen. Remember, mathematics only models reality, it is *not* reality.

A.2 Random Variables

When we are trying to mathematically describe an experiment and its outcomes, it is much more convenient if the outcomes are numbers. A simple way to do this is to define a mapping or function that assigns a number to each outcome. This mapping or function is called a *random variable*. To put that more formally: Let S be a sample space with outcomes $\{\omega_i\}$. Then the random variable X is a mapping

$$X : S \rightarrow \mathcal{R} \tag{A.9}$$

where $\mathcal{R}$ denotes the real number line. Another way of saying the same thing is

$$X(\omega) = x; \qquad \omega \in S, x \in \mathcal{R}. \tag{A.10}$$

The random variable is generally represented by an uppercase letter, and this is the convention we will follow. The value that the random variable takes on is called the *realization* of the random variable and is represented by a lowercase letter.

Example A.2.1:

Let's take our television example and rewrite it in terms of a random variable X:

$$\begin{aligned} X(product\ commercial) &= 0 \\ X(service\ commercial) &= 1 \\ X(news) &= 2 \\ X(pseudonews) &= 3 \\ X(talk\ show) &= 4 \end{aligned}$$

$$\begin{aligned} X(variety) &= 5 \\ X(comedy) &= 6 \\ X(adventure) &= 7 \\ X(movie) &= 8 \end{aligned}$$

Now, instead of talking about the probability of certain programs, we can talk about the probability of the random variable X taking on certain values or ranges of values. For example, $P(X(\omega) \leq 1)$ is the probability of seeing a commercial when the television is turned on (generally, we drop the argument and simply write this as $P(X \leq 1)$). Similarly, the $P(programs\ that\ may\ contain\ news)$ could be written as $P(1 < X \leq 4)$, which is substantially less cumbersome. ♦

A.3 Distribution Functions

Defining the random variable in the way that we did allows us to define a special probability $P(X \leq x)$. This probability is called the *cumulative distribution function (cdf)* and is denoted by $F_X(x)$, where the random variable is the subscript and the realization is the argument. One of the primary uses of probability is the modeling of physical processes, and we will find the cumulative distribution function very useful when we try to describe or model different random processes. We will see more on this later.

For now, let us look at some of the properties of the *cdf*:

Property 1: $0 \leq F_X(x) \leq 1$. This follows from the definition of the *cdf*.

Property 2: The *cdf* is a monotonically nondecreasing function. That is,

$$x_1 \geq x_2 \quad \Rightarrow \quad F_X(x_1) \geq F_X(x_2).$$

To show this simply write the *cdf* as the sum of two probabilities:

$$\begin{aligned} F_X(x_1) &= P(X \leq x_1) = P(X \leq x_2) + P(x_2 < X \leq x_1) \\ &= F_X(x_2) + P(x_1 < X \leq x_2) \geq F_X(x_2) \end{aligned}$$

Property 3:

$$\lim_{n \to \infty} F_X(x) = 1.$$

Property 4:

$$\lim_{n \to -\infty} F_X(x) = 0.$$

Property 5: If we define

$$F_X(x^-) = P(X < x)$$

then

$$P(X = x) = F_X(x) - F_X(x^-).$$

Example A.3.1:

Assuming that the frequency of occurrence was an accurate estimate of the probabilities, let us obtain the *cdf* for our television example:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 0.4 & 0 \le x < 1 \\ 0.8 & 1 \le x < 2 \\ 0.802 & 2 \le x < 3 \\ 0.822 & 3 \le x < 4 \\ 0.912 & 4 \le x < 5 \\ 0.932 & 5 \le x < 6 \\ 0.972 & 6 \le x < 7 \\ 0.99 & 7 \le x < 8 \\ 1.00 & 8 \le x \end{cases}$$

♦

Notice a few things about this *cdf*. First, the *cdf* consists of step functions. This is characteristic of discrete random variables. Second, the function is continuous from the right. This is due to the way the *cdf* is defined.

The *cdf* is somewhat different when the random variable is a continuous random variable. For example, if we sampled a speech signal and then took differences of the samples, the resulting random process would have a *cdf* that would look something like this:

$$F_X(x) = \begin{cases} \frac{1}{2}e^{2x} & x \le 0 \\ 1 - \frac{1}{2}e^{-2x} & x > 0. \end{cases}$$

The thing to notice in this case is that because $F_X(x)$ is continuous

$$P(X = x) = F_X(x) - F_X(x^-) = 0.$$

We can also have processes that have distributions that are continuous over some ranges and discrete over others.

Along with the cumulative distribution function, another function that also comes in very handy is the *probability density function (pdf)*. The *pdf* corresponding to the *cdf* $F_X(x)$ is written as $f_X(x)$. For continuous *cdf*s, the *pdf* is simply the derivative of the *cdf*. For the discrete random variables, taking the derivative of the *cdf* would introduce delta functions, which have problems of their own. So in the discrete case, we obtain the *pdf* through differencing. It is somewhat awkward to have different procedures for obtaining the same function for different types of random variables. It is possible to define a rigorous unified procedure for getting the *pdf* from the *cdf* for all kinds of random variables. However, in order to do so, we need some familiarity with measure theory, which is beyond the scope of this appendix. Let us look at some examples of *pdf*s.

Example A.3.2:

For our television scenario:

$$f_X(x) = \begin{cases} 0.4 & \text{if } X = 0 \\ 0.4 & \text{if } X = 1 \\ 0.002 & \text{if } X = 2 \\ 0.02 & \text{if } X = 3 \\ 0.09 & \text{if } X = 4 \\ 0.02 & \text{if } X = 5 \\ 0.04 & \text{if } X = 6 \\ 0.018 & \text{if } X = 7 \\ 0.01 & \text{if } X = 8 \\ 0 & \text{otherwise} \end{cases}$$

♦

Example A.3.3:

For our speech example, the *pdf* is given by

$$f_X(x) = \frac{1}{2}e^{-2|x|}.$$

♦

A.4 Expectation

When dealing with random processes, we often deal with average quantities, like the signal power and noise power in communication systems, and the mean time between failures in various design problems. To obtain these average quantities, we use something called an *expectation operator*. Formally, the expectation operator $E[\,]$ is defined as follows: The *expected value* of a random variable X is given by

$$E[X] = \sum_i x_i P(X = x_i) \tag{A.11}$$

when X is a discrete random variable with realizations $\{x_i\}$ and by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{A.12}$$

where $f_X(x)$ is the *pdf* of X.

The expected value is very much like the average value and, if the frequency of occurrence is an accurate estimate of the probability, is identical to the average value. Consider the following example:

Example A.4.1:

Suppose in a class of 10 students the grades on the first test were

$$10, 9, 8, 8, 7, 7, 7, 6, 6, 2$$

The average value is $\frac{70}{10}$, or 7. Now let's use the frequency of occurrence approach to estimate the probabilities of the various grades. (Notice in this case the random variable is an identity mapping, i.e., $X(\omega) = \omega$.) The probability estimate of the various values the random variable can take on are

$$\begin{aligned} P(10) &= P(9) = P(2) = 0.1, \quad P(8) = P(6) = 0.2, \quad P(7) = 0.3, \\ &P(6) = P(5) = P(4) = P(3) = P(1) = P(0) = 0 \end{aligned}$$

The expected value is therefore given by

$$\begin{aligned} E[X] &= (0)(0) + (0)(1) + (0.1)(2) + (0)(3) + (0)(4) + (0)(5) + (0.2)(6) \\ &+ (0.3)(7) + (0.2)(8) + (0.1)(9) + (0.1)(10) = 7. \end{aligned}$$

♦

It seems that the expected value and the average value *are* exactly the same! But we have made a rather major assumption about the accuracy of our probability estimate. In general the relative frequency is not exactly the same as the probability, and the average expected values are different. To emphasize this difference and similarity, the expected value is sometimes referred to as the *statistical average*, while our everyday average value is referred to as the *sample average*.

We said at the beginning of this section that we are often interested in things such as signal power. The average signal power is often defined as the average of the signal squared. If we say that the random variable is the signal value, then this means that we have to find the expected value of the square of the random variable. There are two ways of doing this. We could define a new random variable $Y = X^2$, then find $f_Y(y)$ and use (A.12) to find $E[Y]$. An easier approach is to use the *fundamental theorem of expectation*, which is

$$E[g(X)] = \sum_i g(x_i)P(X = x_i) \tag{A.13}$$

for the discrete case, and

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx \tag{A.14}$$

for the continuous case.

The expected value, because of the way it is defined, is a linear operator. That is,

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y], \qquad \alpha \text{ and } \beta \text{ are constants.}$$

You are invited to verify this for yourself.

There are several functions $g()$ whose expectations are used so often that they have been given special names.

A.4.1 Mean

The simplest and most obvious function is the identity mapping $g(X) = X$. The expected value $E(X)$ is referred to as the *mean* and is symbolically referred to as μ_X. If we take a random variable X and add a constant value to it, the mean of the new random process is simply the old mean plus the constant. Let

$$Y = X + a$$

where a is a constant value. Then

$$\mu_Y = E[Y] = E[X+a] = E[X] + E[a] = \mu_X + a.$$

A.4.2 Second Moment

If the random variable X is an electrical signal, the total power in this signal is given by $E[X^2]$, which is why we are often interested in it. This value is called the *second moment* of the random variable.

A.4.3 Variance

If X is a random variable with mean μ_X, then the quantity $E[(X-\mu_X)^2]$ is called the *variance* and is denoted by σ_X^2. The square root of this value is called the *standard deviation* and is denoted by σ. The variance and the standard deviation can be viewed as a measure of the "spread" of the random variable. We can show that

$$\sigma_X^2 = E[X^2] - \mu_X^2.$$

If $E[X^2]$ is the total power in a signal, then the variance is also referred to as the total AC power.

A.5 Types of Distribution

There are several specific distributions that are very useful when describing or modeling various processes.

A.5.1 Uniform Distribution

This is the distribution of ignorance. If we want to model data about which we know nothing except its range, this is the distribution of choice. This is not to say that there are not times when the uniform distribution is a good match for the data. The *pdf* of the uniform distribution is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq X \leq b \\ 0 & \text{otherwise.} \end{cases} \quad \text{(A.15)}$$

The mean of the uniform distribution can be obtained as

$$\mu_X = \int_a^b x \frac{1}{b-a} dx = \frac{b+a}{2}.$$

Similarly, the variance of the uniform distribution can be obtained as

$$\sigma_X^2 = \frac{(b-a)^2}{12}.$$

Details are left as an exercise.

A.5.2 Gaussian Distribution

This is the distribution of choice in terms of mathematical tractability. Because of its form, it is especially useful with the squared error distortion measure. The probability density function for a random variable with a Gaussian distribution, and mean μ and variance σ^2, is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2} \tag{A.16}$$

where the mean of the distribution is μ and the variance is σ^2.

A.5.3 Laplacian Distribution

Many sources that we will deal with will have probability density functions that are quite peaked at zero. For example, speech consists mainly of silence; therefore, samples of speech will be zero or close to zero with high probability. Image pixels themselves do not have any attraction to small values. However, there is a high degree of correlation among pixels. Therefore, a large number of the pixel-to-pixel differences will have values close to zero. In these situations, a Gaussian distribution is not a very close match to the data. A closer match is the Laplacian distribution, which has a *pdf* that is peaked at zero. The density function for a zero mean random variable with Laplacian distribution and variance σ^2 is

$$f_X(x) = \frac{1}{\sqrt{2\sigma^2}} \exp \frac{-\sqrt{2}\,|x|}{\sigma}. \tag{A.17}$$

A.5.4 Gamma Distribution

A distribution with a *pdf* that is even more peaked, though considerably less tractable than the Laplacian distribution, is the Gamma distribution. The density function for a Gamma distributed random variable with zero mean and variance σ^2 is given by

$$f_X(x) = \frac{\sqrt[4]{3}}{\sqrt{8\pi\sigma\,|x|}} \exp \frac{-\sqrt{3}\,|x|}{2\sigma}. \tag{A.18}$$

A.6 Stochastic Process

We are often interested in experiments whose outcomes are a function of time. For example, we might be interested in designing a system that encodes speech. The outcomes are particular patterns of speech that will be encountered by the speech coder. We can mathematically describe this situation by extending our definition of a random variable. Instead of the random variable mapping an outcome of an experiment to a number, we map it to a function of time. Let S be a sample space with outcomes $\{\omega_i\}$. Then the random or stochastic process X is a mapping

$$X : S \rightarrow \mathcal{F} \tag{A.19}$$

where $\mathcal{F}$ denotes the set of functions on the real number line. In other words,

$$X(\omega) = x(t); \qquad \omega \in S, \ x \in \mathcal{F}, \ -\infty < t < \infty. \tag{A.20}$$

The functions $x(t)$ are called the *realizations* of the random process, and the collection of functions $\{x_\omega(t)\}$ indexed by the outcomes ω is called the *ensemble* of the stochastic process. We can define the mean and variance of the ensemble as

$$\mu(t) = E[X(t)] \tag{A.21}$$

$$\sigma^2(t) = E[(X(t) - \mu(t))^2]. \tag{A.22}$$

If we sample the ensemble at some time t_0, we get a set of numbers $\{x_\omega(t_0)\}$ indexed by the outcomes ω, which by definition is a random variable. By sampling the ensemble at different times t_i, we get different random variables $\{x_\omega(t_i)\}$. For simplicity we often drop the ω and t and simply refer to these random variables as $\{x_i\}$.

Associated with each of these random variables, we will have a distribution function. We can also define a joint distribution function for two or more of these random variables: Given a set of random variables $\{x_1, x_2, \ldots, x_N\}$, the *joint* cumulative distribution function is defined as

$$F_{X_1X_2\cdots X_N}(x_1, x_2, \ldots, x_N) = P(X_1 < x_1, X_2 < x_2, \ldots, X_N < x_N) \tag{A.23}$$

Unless it is clear from the context what we are talking about, we will refer to the *cdf* of the individual random variables X_i as the *marginal cdf* of X_i.

We can also define the joint probability density function for these random variables $f_{X_1X_2\cdots X_N}(x_1, x_2, \ldots, x_N)$ in the same manner as we defined the *pdf* in the case of the single random variable. We can classify the relationships between these random variables in a number of different ways. In the following we define some relationships between two random variables. The concepts are easily extended to more than two random variables.

Two random variables X_1 and X_2 are said to be *independent* if their joint distribution function can be written as the product of the marginal distribution functions of each random variable; that is,

$$F_{X_1X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2). \tag{A.24}$$

This also implies that

$$f_{X_1X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2). \tag{A.25}$$

If all the random variables $X_1, X_2, \ldots$ are independent and they have the same distribution, they are said to be *independent, identically distributed* (*iid*).

Two random variables X_1 and X_2 are said to be *orthogonal* if

$$E[X_1 X_2] = 0. \tag{A.26}$$

Two random variables X_1 and X_2 are said to be *uncorrelated* if

$$E[(X_1 - \mu_1)(X_2 - \mu_2)] = 0 \tag{A.27}$$

where $\mu_1 = E[X_1]$ and $\mu_2 = E[X_2]$.

The *autocorrelation function* of a random process is defined as

$$R_{xx}(t_i, t_2) = E[X_1 X_2]. \tag{A.28}$$

For a given value of N, suppose we sample the stochastic process at N times $\{t_i\}$ to get the N random variables $\{X_i\}$ with *cdf* $F_{X_1 X_2 \ldots X_N}(x_1, x_2, \ldots, x_N)$, and another N times $\{t_i + T\}$ to get the random variables $\{X_i'\}$ with *cdf* $F_{X_1' X_2' \ldots X_N'}(x_1', x_2', \ldots, x_N')$. If

$$F_{X_1 X_2 \ldots X_N}(x_1, x_2, \ldots, x_N) = F_{X_1' X_2' \ldots X_N'}(x_1', x_2', \ldots, x_N') \tag{A.29}$$

for all N and T, the process is said to be *stationary*.

The assumption of stationarity is a rather important assumption because it is a statement that the statistical characteristics of the process under investigation do not change with time. Thus, if we design a system for an input based on the statistical characteristics of the input today, the system will still be useful tomorrow because the input will not change its characteristics. The assumption of stationarity is also a very strong assumption, and we can usually make do quite well with a weaker condition, *wide sense* or *weak sense* stationarity.

A stochastic process is said to be wide sense or weak sense stationary if it satisfies the following conditions:

1. The mean is constant; that is, $\mu(t) = \mu$ for all t.
2. The variance is finite.
3. The autocorrelation function $R_{xx}(t_1, t_2)$ is a function only of the difference between t_1 and t_2, and not of the individual values of t_1 and t_2; that is,

$$R_{xx}(t_1, t_2) = R_{xx}(t_1 - t_2) = R_{xx}(t_2 - t_1). \tag{A.30}$$

Further Reading

1. The classic books on probability are the two-volume set *An Introduction to Probability Theory and Its Applications,* by W. Feller [171].
2. A commonly used text for an introductory course on probability and random processes is *Probability, Random Variables, and Stochastic Processes*, by A. Papoulis [172].

A.7 Projects and Problems

1. If $A \cap B \neq \phi$, show that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

2. Show that expectation is a linear operator in both the discrete and the continuous case.
3. If a is a constant, show that $E[a] = a$.
4. Show that for a random variable X,

$$\sigma_X^2 = E[X^2] - \mu_X^2.$$

5. Show that the variance of the uniform distribution is given by

$$\sigma_X^2 = \frac{(b-a)^2}{12}.$$