Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond
by Jianbo Gao, Yinhe Cao, Wen-wen Tung and Jing Hu
Copyright © 2007 John Wiley & Sons. Inc.

CHAPTER 3

BASICS OF PROBABILITY THEORY AND STOCHASTIC PROCESSES

In this chapter we first review the most important definitions and results of elementary probability theory. Then we describe a number of commonly used distributions and discuss their main properties. To facilitate applications, we shall also explain how random variables of a specific distribution can be simulated. Finally, we briefly discuss random processes. Our main purpose here is to equip readers with an intuitive understanding. Therefore, occasionally, mathematical rigor will be sacrificed.

Most readers might think that the materials covered in this chapter have been so well developed in many excellent textbooks that they might not have much to do with research of considerable current interest. To counter such a belief, when appropriate, we will comment on where such elementary materials have shed new light on complex problems. We particularly remind research-oriented readers of the literature search using the concept of the power-law network, described in Sec. 3.4.

3.1 BASIC ELEMENTS OF PROBABILITY THEORY

3.1.1 Probability system

Probability theory is concerned with the description of random events occurring in the real world. To better appreciate the basic elements of the theory, let us first examine the main features of an experiment with random outcomes. There are three main features:

1. A set of possible experimental outcomes (obtained under precisely controlled identical experimental conditions).
2. A grouping of these outcomes into classes called results.
3. The relative frequency of these results in many independent trials of the experiment.

The relative frequency f_c of a result is merely the number of times the result is observed divided by the number of times the experiment is performed; as the number of experimental trials increases, we expect f_c to be close to a constant value. As an example, suppose a fair coin is tossed a thousand times; then the relative frequency for either head or tail to be observed will be close to 1/2. As another example, if a fair die is thrown many times, the relative frequency with which one of the numbers $\{1, 2, 3, 4, 5, 6\}$ is observed will be close to 1/6. There are situations in which one may want to group the six numbers into only two classes: odd numbers $\{1, 3, 5\}$ and even numbers $\{2, 4, 6\}$ or small numbers $\{1, 2, 3\}$ and large numbers $\{4, 5, 6\}$. Then there are only two classes or results of the die-throwing experiment. One can also construct complicated experiments by combining simple ones, such as first tossing a coin, then throwing a die.

Abstracting from the above considerations, one obtains three basic elements that constitute the probability theory:

1. A sample space S, which is a collection of objects, corresponding to the set of mutually exclusive, exhaustive outcomes of an experiment. Each point ω in S is called a sample point.
2. A family of events E, denoted as $\{A, B, C, \cdots\}$, in which each event is a set of sample points $\{\omega\}$. An event corresponds to a class or result of a real-world experiment.
3. A probability measure P which assigns each event, say A, a real nonnegative number $P(A)$, which corresponds to the relative frequency in the experimental situation. This assignment must satisfy three properties (axioms):
   - For any event A, $0 \le P(A) \le 1$.
   - $P(S) = 1$.
   - If A and B are mutually exclusive events, then
     $P(A \cup B) = P(A) + P(B)$.
     More generally, for any sequence of mutually exclusive events $E_1, E_2, \cdots$ (that is, events satisfying $E_iE_j = \Phi$, $i \neq j$, where Φ is the empty set),

$$P\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} P(E_i) .$$

The triplet (S, E, P), along with the three axioms, forms a probability system.

Let us now present two other definitions. The first is the conditional probability of event A given that event B occurred, $P(A|B)$. It is defined as

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{3.1}$$

whenever $P(B) \neq 0$. The introduction of the conditional event B forces us to shift attention from the original sample space S to a new sample space defined by the event B. In terms of real-world applications, this amounts to renormalizing the problem by magnifying the probabilities associated with the conditional events by dividing by the term $P(B)$ as given above.

The second important notion is that of statistical independence of events. Two events A and B are said to be statistically independent if and only if

$$P(AB) = P(A)P(B). \tag{3.2}$$

For three events A, B, C, we require that each pair of them satisfies Eq. (3.2) and in addition

$$P(ABC) = P(A)P(B)P(C). \tag{3.3}$$

It is worth emphasizing that Eq. (3.3) alone is not sufficient to define statistical independence of three events A, B, C. To see this, let us consider a situation schematically shown in Fig. 3.1, where

$$P(A) = P(B) = P(C) = 1/5,$$

and the events A, B, C intersect with each other on a common subset O,

$$P(AB) = P(BC) = P(CA) = P(ABC) = P(O) = p.$$

If one chooses $p = 1/25$, then Eq. (3.2) is satisfied but Eq. (3.3) is not. If one chooses $p = 1/125$, then Eq. (3.3) is satisfied but Eq. (3.2) is not. In fact, Eqs. (3.2) and (3.3) cannot be simultaneously satisfied in this situation.

It is easy to see that the notion of statistical independence can be readily extended to any $n > 2$ events.

3.1.2 Random variables

Up to now, we have associated probabilities directly with random events. Often it is more convenient to represent a random event by a number, called a random variable, and talk about the probability of a random variable. In fact, in throwing a die, we naturally have numbers as outcomes. When tossing a coin, we can denote head by 1 and tail by -1. Such a mapping is particularly convenient, noticing that the summation of a sequence of 1 and -1 gives the number of net wins (e.g., head) or losses (e.g., tail) in a total of N experiments. The summation of a *sequence of* random variables is called a random walk process. We will talk more about it later.

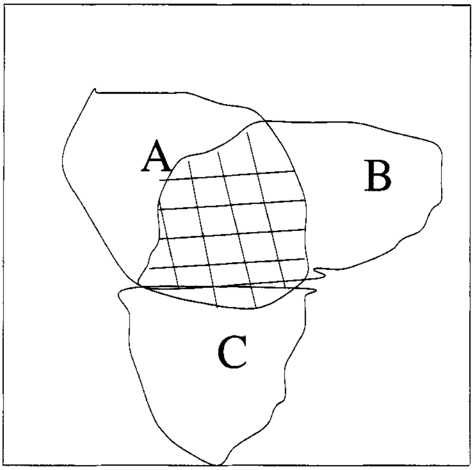

Figure 3.1. A situation where $AB = BC = AC = ABC$.

Let us denote a random variable by X, whose value depends on the outcome, ω, of a random experiment. The event $(X \leq x)$ is thus equivalent to $\{\omega : X(\omega) \leq x\}$. Let us denote the probability of $(X \leq x)$ by $F_X(x)$, which is called the cumulative distribution function (CDF),

$$F_X(x) = P(X \leq x). \tag{3.4}$$

The important properties of $F_X(x)$ include

$$F_X(x) \geq 0,$$

$$F_X(\infty) = 1,$$

$$F_X(-\infty) = 0,$$

$$P(a < X \leq b) = F_X(b) - F_X(a) \geq 0 \quad \text{for} \quad a \leq b.$$

Therefore, $F_X(x)$ is a nonnegative, monotonically nondecreasing function with limits 0 and 1 at $-\infty$ and ∞, respectively. Conventionally, $F_X(x)$ is assumed to be continuous from the right. When the derivative of $F_X(x)$ exists, it is often more convenient to work with its derivative, called the probability density function (PDF),

$$f_X(x) = \frac{dF_X(x)}{dx}. \tag{3.5}$$

"Inverting" Eq. (3.5) yields

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy. \tag{3.6}$$

Therefore,

$$P(a < X \leq b) = \int_a^b f_X(y)dy.$$

Since probability has to be nonnegative, letting $a \to b$, we see that the last equation implies that $f_X(x) \geq 0$.

Given a probability system (S, E, P), we can also define many random variables in the same sample space. In real-world applications, such a situation amounts to describing a random event by a high-dimensional vector instead of a scalar. Such an extension is necessary if the outcome of an experiment cannot be characterized by just one number. For ease of illustration below, let us mainly consider the case of two random variables X and Y defined for some probability system (S, E, P). The joint CDF is then defined by

$$F_{XY}(x, y) = P(X \leq x, Y \leq y), \tag{3.7}$$

which is merely the probability that X takes on a value not larger than x and at the same time Y takes on a value not larger than y; that is, it is the sum of the probabilities associated with all sample points in the intersection of the two events $\{\omega : X(\omega) \leq x\}$ and $\{\omega : Y(\omega) \leq y\}$. Associated with this function is a joint PDF defined as

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}. \tag{3.8}$$

Given a joint PDF, the "marginal" density function for one of the variables is given by integrating over all possible values of the second variable. For example,

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{XY}(x, y)dy. \tag{3.9}$$

We are now in a position to define the notion of independence between random variables. Two random variables X and Y are said to be independent if and only if their joint PDF factors into the product of the one-dimensional PDFs:

$$f_{XY}(x, y) = f_X(x)f_Y(y).$$

This is very much like the definition for two independent events as given in Eq. (3.2). However, for three or more random variables, the definition is essentially the same as for two, namely, $X_1, X_, \cdots, X_n$ are said to be independent random variables if and only if

$$f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n). \tag{3.10}$$

This is simpler than the definition required for multiple events to be independent. Loosely speaking, this simplification comes from the fact that PDFs characterize "elementary" events of an experiment with random outcomes. Mathematically, this is because Eq. (3.10) implies that the following equation is true:

$$f_{X_{i1} X_{i2} \cdots X_{ik}}(x_{i1}, x_{i2}, \cdots, x_{ik}) = f_{X_{i1}}(x_{i1})f_{X_{i2}}(x_{i2}) \cdots f_{X_{ik}}(x_{ik}),$$

where $k \leq n$ and $i1, i2, \cdots, ik$ are all different. This is easily proven by resorting to the definition of a marginal distribution and integrating both sides of Eq. (3.10) with respect to the random variables one wishes to remove.

With more than one random variable, we can now define the conditional distributions and densities. For example, we can ask for the CDF of the random variable X conditioned on some given value of the random variable Y, which is merely the probability $P[X \leq x|Y = y]$. Similarly, the conditional PDF on X, given Y, is defined as

$$f_{X|Y}(x|y) = \frac{d}{dx}P[X \leq x|Y = y] = \frac{f_{XY}(x,y)}{f_Y(y)},$$

much as the definition for the conditional probability of events.

We can also define one random variable Y in terms of a second random variable X. In this case, Y is referred to as a function of the random variable X. In its most general form, we have

$$Y = g(X),$$

where $g(\cdot)$ is some given function of its argument. Thus, once the value of X is determined, the value of Y can be computed. Since the value of X depends upon the sample point ω, we can write $Y = g(X(\omega)) = Y(\omega)$. Therefore,

$$F_Y(y) = P(Y \leq y) = P[\{\omega : g(X(\omega)) \leq y\}].$$

In general, the computation of this last equation may be complicated. It is clear that a random variable may be a function of many random variables rather than just one. For example,

$$Y = \sum_{i=1}^{n} X_i.$$

This simple equation describes a random walk process, which we will discuss in depth later.

Before ending this section, let us comment on statistical independence and vector description of a real-world problem. A complicated experiment or situation typically has multiple features; each of them could be described by a number. If the features identified are all different, then the dependence among the random variables describing each feature would be quite minor. This is a desired situation. If, however, the dependence among the random variables is quite high, then the number of independent features identified would be considerably smaller than the number of variables designed for the problem. It could be that the problem under study might be less complicated than we thought or that some important features have not been identified.

3.1.3 Expectation

An important class of measures associated with the CDF and the PDF for a random variable is expectations or mean values. They deal with a special type of integral of the PDF. Since PDF is the derivative of CDF, when CDF is not differentiable, a difficulty arises when one tries to define PDF. This difficulty can be resolved by the use of impulse functions. Alternatively (and preferably), one resorts to Stieltjes integrals. A Stieltjes integral is defined in terms of a nondecreasing function $F(x)$ and a continuous function $\phi(x)$; in addition, two sets of points $\{t_k\}$ and $\{\zeta_k\}$ such that $t_{k-1} < \zeta_k \leq t_k$ are defined and a limit is considered where max $|t_k - t_{k-1}| \to 0$. With these, consider the sum

$$\sum_k \phi(\zeta_k)[F(t_k) - F(t_{k-1})].$$

This sum tends to a limit as the intervals shrink to zero independent of the sets $\{t_k\}$ and $\{\zeta_k\}$, and the limit is referred to as the Stieltjes integral of ϕ with respect to F. This integral is written as

$$\int \phi(x)dF(x).$$

When $F(x)$ is differentiable or if impulse functions are allowed in PDF, we can write

$$dF(x) = f_X(x)dx.$$

Therefore,

$$\int \phi(x)dF(x) = \int \phi(x)f_X(x)dx.$$

We are now ready to define expectations.

The mean or average of X is given by

$$E[X] = \overline{X} = \int_{-\infty}^{\infty} xdF(x) = \int_{-\infty}^{\infty} xf_X(x)dx.$$

The nth moment of X is defined as

$$E(X^n) = \int_{-\infty}^{\infty} x^n dF_X(x) = \int_{-\infty}^{\infty} x^n f_X(x)dx.$$

Similarly, the nth central moment of X is defined as

$$E[(x - \overline{X})^n] = \int_{-\infty}^{\infty} (x - \overline{X})^n f_X(x)dx.$$

The variance of X is defined by

$$Var(X) = \sigma_X^2 = E(X^2) - [E(X)]^2,$$

which is the second central moment. Its square root σ_X is called the standard deviation, and the coefficient of variation is defined by

$$C_X = \sigma_X/\overline{X}.$$

Let us now consider the case where the random variable Y is a function of the random variable X, $Y = g(X)$. We have

$$E_Y[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

Since $f_Y(y)dy$ and $f_X(x)dx$ describe the same probability for the set of sample points of an experiment, they have to be equal; hence,

$$E_Y[Y] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

We can similarly define expectations for multiple random variables. For example,

$$E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f_{XY}(x,y) dx dy,$$

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y) dx dy.$$

It is easy to prove that

$$E[X+Y] = E[X] + E[Y]$$

without any condition. On the other hand, in order to have

$$E[XY] = E[X]E[Y],$$

the random variables X and Y have to be independent. Extending the above two equations, we have

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$$

without any condition and

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

when X and Y are independent.

3.1.4 Characteristic function, moment generating function, Laplace transform, and probability generating function

In computations using probability theory, one has to evaluate integrals of PDFs. To simplify computations, one can resort to a number of functions. Here we review four closely related functions, called the characteristic function, the moment generating function, the Laplace transform, and the probability generating function. Their usefulness will be illustrated in the next section when we talk about various types of distributions.

The characteristic function of a random variable X, denoted by $\Phi_X(u)$, is given by

$$\Phi_X(u) = E[e^{juX}] = \int_{-\infty}^{\infty} e^{jux} f_X(x)dx,$$

where $j = \sqrt{-1}$ and u is an arbitrary real variable. Clearly,

$$|\Phi_X(u)| \leq \int_{-\infty}^{\infty} |e^{jux}||f_X(x)dx| = 1.$$

The equality is achieved at $u = 0$. Expanding e^{jux} in terms of its power series, we have

$$e^{jux} = 1 + jux + \frac{(jux)^2}{2!} + \cdots.$$

We thus have

$$\Phi_X(u) = 1 + ju\overline{X} + \frac{(ju)^2}{2!}\overline{X^2} + \cdots.$$

Differentiating the above equation n times on both sides with respect to u, we have

$$\left.\frac{d^n\Phi_X(u)}{du^n}\right|_{u=0} = j^n\overline{X^n}.$$

For simplicity, we shall define

$$g^{(n)}(x_0) = \left.\frac{d^n g(x)}{dx^n}\right|_{x=x_0}.$$

Therefore,

$$\Phi_X^{(n)}(0) = j^n\overline{X^n}.$$

The moment generating function $M_X(v)$ is obtained by replacing ju by a real variable v in the characteristic function,

$$M_X(v) = E[e^{vX}] = \int_{-\infty}^{\infty} e^{vx} f_X(x)dx.$$

Similarly, we have

$$M_X^{(n)}(0) = \overline{X^n}.$$

The Laplace transform $L_X(s)$ of the PDF is defined as

$$L_X(s) = E[e^{-sX}] = \int_{-\infty}^{\infty} e^{-sx} f_X(x)dx.$$

Similarly, we can prove that

$$L_X^{(n)}(0) = (-1)^n\overline{X^n}.$$

It is clear that the three functions $\Phi_X(u), M_X(v), L_X(s)$ are closely related. In particular, we have

$$\Phi_X(js) = M_X(-s) = L_X(s).$$

In the case of a discrete random variable defined by

$$g_k = P[X = k],$$

we can define the probability generating function $G(z)$ as follows:

$$G(z) = E[z^X] = \sum_k z^k g_k,$$

where z is a complex variable. In fact, $G(z)$ is just the z-transform of the discrete sequence g_k. When $|z| \leq 1$, we have, as with the continuous transforms,

$$|G(z)| \leq \sum_k |z^k||g_k| \leq 1.$$

It is easy to see that the first derivative evaluated at $z = 1$ yields the mean of X

$$G^{(1)}(1) = \overline{X},$$

and the second derivative yields

$$G^{(2)}(1) = \overline{X^2} - \overline{X}.$$

Alternatively, one may obtain $\overline{X^2}$ by differentiating $zG^{(1)}(z)$ and then taking $z = 1$:

$$\overline{X^2} = \frac{d}{dz} zG^{(1)}(z)\bigg|_{z=1}.$$

The foregoing discussions make it clear that in many situations, these functions may greatly simplify the calculations of expectations of a given distribution. This will be illustrated in the next section. Another important application of these functions is related to the following property. Let X and Y be independent random variables with characteristic functions $\Phi_X(u)$ and $\Phi_Y(u)$, respectively. Let $Z = X + Y$. Then the characteristic function for Z is

$$\Phi_Z(u) = \Phi_X(u) \cdot \Phi_Y(u).$$

This property is of crucial importance for the discussion of stable laws and Levy motions in Chapter 7. The proof is left as an exercise.

3.2 COMMONLY USED DISTRIBUTIONS

In this section, we describe a number of commonly used distributions and discuss the relations among them.

(1) Exponential and related distributions. The CDF for the exponential distribution is given by

$$F_X(x) = 1 - e^{-\lambda x}, \quad x \geq 0. \tag{3.11}$$

Its PDF is

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0. \tag{3.12}$$

The exponential distribution has the remarkable memoryless property, which is the defining property of the Markov process. It can be shown as follows:

$$\begin{aligned} P(X \leq x + x_0 | X > x_0) &= \frac{P(x_0 < X \leq x + x_0)}{P(X > x_0)} \\ &= \frac{P(X \leq x + x_0) - P(X \leq x_0)}{P(X > x_0)}. \end{aligned}$$

Using Eq. (3.11) we then have

$$P(X \leq x + x_0 | X > x_0) = \frac{1 - e^{-\lambda(x+x_0)} - (1 - e^{-\lambda x_0})}{1 - (1 - e^{-\lambda x_0})} = 1 - e^{-\lambda x}.$$

Since the conditional distribution does not depend on x_0, the distribution is memoryless. To understand this property, let us imagine that we are waiting for a bus to come to a bus stop. Assume that the time interval between two successive arrivals follows an exponential distribution, with a mean, say, of 15 min. Then, even if we have waited for 20 minutes, on average, we will have to wait another 15 min to get on a bus; the time that we have waited does not have any effect. In Sec. 3.3.2, when discussing continuous-time Markov chains, we will prove that the exponential distribution is the only continuous distribution that has the memoryless property.

The mean and variance of an exponentially distributed random variable can be directly calculated using the definition. Instead of doing so, here let us find its Laplace transform:

$$L_X(s) = \int_0^\infty e^{-sx} \lambda e^{-\lambda x} dx = \frac{\lambda}{s + \lambda}. \tag{3.13}$$

From the Laplace transform, we find

$$\begin{aligned} \overline{X} &= -L_X^{(1)}(0) = \frac{1}{\lambda}, \\ \overline{X^2} &= L_X^{(2)}(0) = \frac{2}{\lambda^2}, \\ \sigma_X^2 &= \overline{X^2} - (\overline{X})^2 = \frac{1}{\lambda^2}. \end{aligned}$$

Therefore, its mean and standard deviation are both $1/\lambda$.

Now let us consider a sum of k exponentially distributed random variables

$$Y = X_1 + X_2 + \cdots X_k,$$

where X_i are independent. Usually $X_1, \cdots, X_k$ are called identically and independently distributed (iid) random variables. The PDF for Y is given by the convolution of the densities on each of the X_i's, since they are independently distributed. The Laplace transform of the PDF for Y is therefore given by

$$L_Y(s) = [L_{X_i}(s)]^k = \left(\frac{\lambda}{s+\lambda}\right)^k.$$

Inversion of the above equation gives the Erlang distribution:

$$f_Y(y) = \frac{\lambda(\lambda y)^{k-1}}{(k-1)!}e^{-\lambda y}, \quad y \geq 0, \tag{3.14}$$

which is a special case of the gamma distribution:

$$f(y) = \frac{1}{\Gamma(t)}\lambda^t y^{t-1}e^{-\lambda y}, \quad y \geq 0, \tag{3.15}$$

where parameters $\lambda, t > 0$, and $\Gamma(t)$ is the gamma function:

$$\Gamma(t) = \int_0^\infty y^{t-1}e^{-y}dy.$$

The relations $\Gamma(t) = (t-1)\Gamma(t-1)$ for any t, $\Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(n) = (n-1)!$, where n is an integer, can be used to evaluate the gamma distribution.

The discrete version of the exponential distribution is called the geometrical distribution, given by

$$P(X = k) = f(k) = p(1-p)^{k-1}, \tag{3.16}$$

where $0 < p < 1$. Often $f(k)$ is called the mass function. This distribution arises in the following way. Suppose that independent Bernoulli trials (with parameter p) are performed at times $1, 2, \cdots$. Let X be the time that elapses before the first success; X is called the waiting time. Then $P(X > k) = (1-p)^k$ and thus

$$P(X = k) = P(X > k-1) - P(X > k) = p(1-p)^{k-1}.$$

It is easy to prove that the mean and variance are $1/p$ and $(1-p)/p^2$, respectively. Readers are encouraged to prove, preferably at this point, that the geometrical distribution is the unique discrete distribution that has the memoryless property.

(2) Normal and related distributions. The normal distribution is perhaps the most important continuous distribution. It has two parameters, μ and σ^2. Its density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < X < \infty. \tag{3.17}$$

It is denoted by $N(\mu, \sigma^2)$. If $\mu = 0$ and $\sigma^2 = 1$, then

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \quad -\infty < X < \infty$$

is the density of the standard normal distribution. The characteristic function for $N(\mu, \sigma^2)$ is

$$\Phi_X(u) = e^{j\mu u - \frac{\sigma^2 u^2}{2}}. \tag{3.18}$$

The normal distribution arises in many ways. In particular, it can be obtained as a continuous limit of the binomial distribution, as will be discussed shortly. This result is a special case of the central limit theorem, which says that in many cases the sum of a large number of independent (or at least not too dependent) random variables is approximately normally distributed.

Now let us consider the distribution for $x = \sum_{i=1}^{d} x_i^2$, where $x_i \sim N(0,1)$ and x_i's are independent. The distribution for x is called the chi-squared distribution, with degree d. It can be obtained by setting $\lambda = 1/2$ and $t = d/2$ in the gamma distribution (Eq. 3.15)).

Next let us consider $X = \sqrt{X_1^2 + X_2^2}$, where $X_i \sim N(0, \sigma^2)$, $i = 1, 2$, and X_1 and X_2 are independent. The distribution for X is called the Rayleigh distribution given by

$$f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \quad x \geq 0. \tag{3.19}$$

Finally, let

$$Y = e^X, \quad \text{where } X \sim N(\mu, \sigma^2).$$

The distribution for Y is called the log-normal distribution, given by

$$f(y) = \frac{1}{\sigma y \sqrt{2\pi}} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}}. \tag{3.20}$$

The log-normal distribution has been found to accurately describe, among many other things, switching times in multistable visual perceptions and many different types of network traffic processes, including variable bit rate (VBR) video traffic, the call duration and dwell time in a wireless network, data connections and messages, and page size, page request frequency, and user's think time in the World Wide Web (WWW) traffic. One mechanism for the log-normal distribution is that the random variable Y is a multiplication of a sequence of iid random variables, $Y = X_1 X_2 \cdots X_n$, where each X_i may represent one event in a complex chain of events. This could correspond to propagation of a signal in a complex system. By taking the logarithm and then applying the central limit theorem, one readily sees that Y follows the log-normal distribution. We shall have more to say on this when discussing multiplicative cascade multifractal models.

(3) Binomial and related distributions. Consider the Bernoulli trial, where a coin is tossed. Denote $Y = 1$ when a head turns up and $Y = 0$ when a tail turns up. Assume the probability for a head to turn up to be $0 < p < 1$. Toss the coin n times, and let $X = Y_1 + Y_2 + \cdots + Y_n$. The total number X of heads takes values in the set $\{0, 1, 2, \cdots, n\}$ and is a discrete random variable. Then X has the binomial distribution

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \tag{3.21}$$

where $q = 1 - p$. It is easy to prove that

$$E(X) = np, \ \ \mathrm{var}(X) = npq.$$

Now let us consider two special cases:

- $n \to \infty$ and $p \to 0$ in such a way that $E(X) = np$ approaches a nonzero constant λ. Then, for $k = 0, 1, 2, \cdots$,

  $$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{k!}\left(\frac{np}{1-p}\right)^k \left(1 - \frac{np}{n}\right)^n \to \frac{\lambda^k}{k!} e^{-\lambda},$$

  which is the Poisson distribution.

  The probability generating function for the Poisson distribution is

  $$G(z) = E[z^K] = \sum_{k=0}^{\infty} z^k P(X = k) = \sum_{k=0}^{\infty} z^k e^{-\lambda} \frac{\lambda^k}{k!} = e^{\lambda(z-1)}.$$

  Therefore,

  $$E[X] = G^{(1)}(1) = \lambda,$$
  $$\sigma_X^2 = G^{(2)}(1) + G^{(1)}(1) - [G^{(1)}(1)]^2 = \lambda.$$

  These results can, of course, also be obtained from the mean and variance for the binomial distribution, noting that $np = \lambda$, $npq \to np = \lambda$.

- If $npq \gg 1$, then

  $$P(X = k) = \binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}}$$

  for k in the $\sqrt{npq}$ neighborhood of np. This result is called the DeMoivre-Laplace theorem. It says that when n is large, X follows a normal distribution with mean np and variance npq. This result can be understood by noting that X is the sum of n Bernoulli random variables and then applying the central limit theorem.

  One can ask a different question: what is the distribution when k is far away from the mean? The answer is given by Bernstein's inequality:

  $$P\left(\frac{1}{n}X \geq p + \epsilon\right) \leq e^{-\frac{1}{4}n\epsilon^2} \ \ for \ \ \epsilon > 0.$$

  Such an inequality is known as large-deviation estimation.

(4) Heavy-tailed distributions. The distributions discussed up to now all have finite variance and are thus called thin-tailed. Do distributions with infinite variance exist? The answer is, of course, yes. As an example, let us derive the Cauchy distribution.

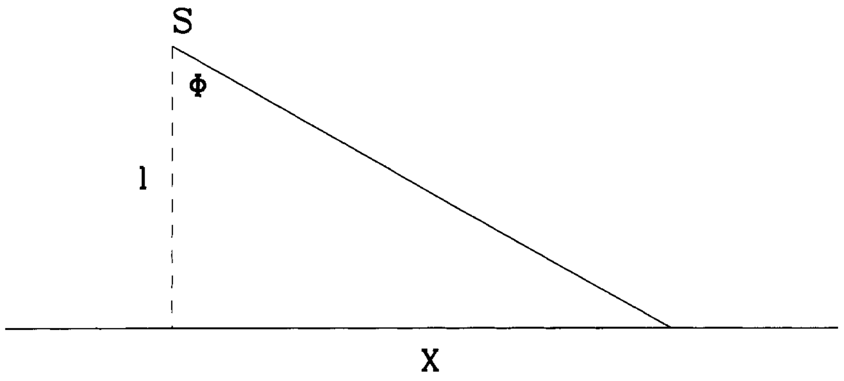

Figure 3.2. Schematic for deriving the Cauchy distribution.

Suppose there is a point light source S emitting photons. S is at the distance l away from a plane. See Fig. 3.2. Let the angle Φ be uniformly distributed on $(-\pi/2, \pi/2)$. Due to symmetry, we only wish to know the distribution for the random variable X. Since X and Φ are related through the relation $x = l \tan \phi$, using equality $f(x) = f(\phi)\left|\frac{d\phi}{dx}\right|$, we obtain the distribution for X:

$$f(x) = \frac{l}{\pi(l^2 + x^2)}. \tag{3.22}$$

This is the Cauchy distribution. It does not have any finite moments. For large x, the distribution can be written as

$$f(x) \sim x^{-2}, \quad x \to \infty.$$

This is a special case of the so-called heavy-tailed distribution, commonly expressed as

$$f(x) \sim x^{-\alpha-1}, \quad x \to \infty.$$

Equivalently, one may write

$$P[X \geq x] \sim x^{-\alpha}, \quad x \to \infty. \tag{3.23}$$

This expression is in fact more popular, since it emphasizes the tail of the distribution. It is easy to prove that when $\alpha < 2$, the variance and all moments higher than second-order moments do not exist. Furthermore, when $\alpha \leq 1$, the mean also diverges.

When the power-law relation extends to the entire range of the allowable x, we have the Pareto distribution:

$$P[X \geq x] = \left(\frac{b}{x}\right)^{\alpha}, \quad x \geq b > 0, \ \alpha > 0, \tag{3.24}$$

where α and b are called the shape and location parameters, respectively. In the discrete case, the Pareto distribution is called the Zipf distribution.

Some readers might find it mind-boggling to believe that power-law-like heavy-tailed distributions with infinite variance can be very relevant in practice. To help those readers, we note that Pareto- or Zipf-like distributions have been found to be ubiquitous. See the Bibliographic Notes at the end of the chapter. See also Sec. 3.4 about an application of the concept of power-law networks.

At this point, it is appropriate to pose a fundamental question. Its answer will be revealed in Chapter 7.

- What is the proper theoretical framework to study the sum of large numbers of iid random variables with infinite variance?

(5) Simulation of random variables. Numerically simulating random variables of various distributions is an important exercise and can provide great help with one's research. Here we explain how to obtain different distributions from uniform $[0,1]$ random variables.

As we explained earlier, the value of a random variable U represents an outcome of a random experiment. Such an outcome can certainly be represented by the value of another random variable X. The probability of an event of the experiment is then either $dF_U(u) = f_U(u)du$ or $dF_X(x) = f_X(x)dx$, where $F_U(u)$ and $F_X(x)$ are the CDFs for the U and X, while $f_U(u)$ and $f_X(x)$ are the PDFs. Now suppose U is a uniform $[0,1]$ random variable, while X is a random variable defined on the interval $[a,b]$, and a can be $-\infty$, while b can be ∞. Then we have

$$\int_a^X dF_X(x) = \int_0^U du.$$

Since $F_X(x)$ is monotonically nondecreasing, its inverse function exists. We then have

$$X = F_X^{-1}(U). \tag{3.25}$$

When there is no closed-form formula for F_X^{-1}, we can solve the problem numerically. Below we consider a few simple examples.

Example 1: Exponential distribution with parameter λ. Since $F_X(x) = 1 - e^{-\lambda x}$,

$$X = -\frac{1}{\lambda}\ln(1-U) = -\frac{1}{\lambda}\ln U', \tag{3.26}$$

where $U' = 1 - U$. It is easy to see that U' is also a uniform $[0,1]$ random variable.

Example 2: Rayleigh distribution. In this case, $F_X(x) = 1 - e^{-\frac{x^2}{2\sigma^2}}$. Therefore,

$$X = \sqrt{-2\sigma^2 \ln U}. \tag{3.27}$$

Example 3: Standard normal distribution. Several methods are available for generating normal random variables. One simple (but not very efficient) way is to use the central limit theorem. That is, we form

$$X = U_1 + U_2 + \cdots + U_n,$$

where $U_i, i = 1, 2, \cdots, n$ are all uniform $[0, 1]$ random variables. When n is large, X is approximately normally distributed. In practice, $n = 12$ is already quite acceptable. Larger n gives more accurate results, but the simulation takes longer. The following method is more efficient.

First, we note (see exercise 3) that if x and y are $N(0, \sigma)$ and independent, if we form

$$x\cos(\omega t) + y\sin(\omega t) = r\cos(\omega t - \varphi), \quad |\varphi| < \pi,$$

then the random variables r and φ are independent, φ is uniform in the interval $[-\pi, \pi]$, and r has a Rayleigh distribution. Now it is easy to see that X_1 and X_2 given by the following equations are $N(0, \sigma)$:

$$\begin{aligned} X_1 &= \sqrt{-2\sigma^2 \ln U}\cos(2U_1 - 1)\pi, \\ X_2 &= \sqrt{-2\sigma^2 \ln U}\sin(2U_2 - 1)\pi, \end{aligned} \tag{3.28}$$

where U, U_1, and U_2 are all uniform $[0, 1]$ random variables. This is called the Box-Muller method.

Example 4: Pareto distribution. From Eq. (3.24), we have

$$F_X(x) = 1 - \left(\frac{b}{x}\right)^{\alpha}, \quad 0 < b \le x.$$

Therefore,

$$X = bU^{-\frac{1}{\alpha}}. \tag{3.29}$$

3.3 STOCHASTIC PROCESSES

In this section, we first present basic definitions of stochastic processes. Then we discuss Markov processes, with an emphasis on their correlation structures, to facilitate comparison between Markov processes and fractal processes with long-range correlations in later chapters.

3.3.1 Basic definitions

Given a probability system (S, E, P), which consists of a sample space S, a set E of events $\{A, B, \cdots\}$, and a probability measure P, a stochastic process is defined as follows: For each sample point $\omega \in S$, we assign a time function $X(t, \omega)$. This family of functions forms a stochastic process. Note that for each allowable parameter t (often interpreted as a time index), $X(t, \omega)$ is a random variable. When ω is fixed, $X(t, \omega)$ is a function of t. Examples of stochastic processes are abundant. The closing price of a given security on the New York Stock Exchange and motions of dust in sky are familiar examples of stochastic processes. A stochastic process is also called a random process. For simplicity, we denote $X(t, \omega)$ by $X(t)$.

How can a random process be specified? For this purpose, we define, for each allowable t, a CDF $F_X(x, t)$, which is given by

$$F_X(x,t) = P[X(t) \leq x].$$

Further, we define for each of n allowable t, $\{t_1, t_2, \cdots, t_n\}$, a joint CDF, given by

$$\begin{aligned} &F_{X_1X_2\cdots X_n}(x_1, x_2, \cdots, x_n;\ t_1, t_2, \cdots, t_n) \\ &= P[X(t_1) \leq x_1, X(t_2) \leq x_2, \cdots, X(t_n) \leq x_n]. \end{aligned}$$

For simplicity, we use the vector notation $F_{\mathbf{X}}(\mathbf{x}, \mathbf{t})$ to denote this function. Of course, in general, specifying the joint CDF is a formidable task.

A stochastic process is said to be stationary if all $F_{\mathbf{X}}(\mathbf{x}, \mathbf{t})$ are invariant to shifts in time; that is, for any given constant τ, the following holds:

$$F_{\mathbf{X}}(\mathbf{x}, \mathbf{t} + \tau) = F_{\mathbf{X}}(\mathbf{x}, \mathbf{t}),$$

where $\mathbf{t} + \tau$ denotes the vector $(t_1 + \tau, t_2 + \tau, \cdots, t_n + \tau)$.

Next, let us define the PDF for a stochastic process. It is given by

$$f_{\mathbf{X}}(\mathbf{x}, \mathbf{t}) = \frac{\partial F_{\mathbf{X}}(\mathbf{x}, \mathbf{t})}{\partial \mathbf{X}} = \frac{\partial^n F_{X_1X_2\cdots X_n}(x_1, x_2, \cdots, x_n;\ t_1, t_2, \cdots, t_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

Using the first-order PDF, we can define the mean of a stochastic process:

$$\overline{X(t)} = E[X(t)] = \int_{-\infty}^{\infty} x f_X(x;\ t) dx.$$

Using the second-order PDF, we can define the autocorrelation of $X(t)$,

$$R_{XX}(t_1, t_2) = E[X(t_1)X(t_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_{X_1X_2}(x_1, x_2;\ t_1, t_2) dx_1 dx_2,$$

and the covariance function of $X(t)$,

$$C_{XX}(t_1, t_2) = R_{XX}(t_1, t_2) - \overline{X(t_1)X(t_2)}.$$

In the case of a stationary random process, we have

$$\overline{X(t)} = \overline{X} = \text{constant}, \tag{3.30}$$

$$R_{XX}(t_1, t_2) = R_{XX}(t_2 - t_1). \tag{3.31}$$

That is, R_{XX} is a function only of the time difference $\tau = t_2 - t_1$. $R_{XX}(\tau)$ is an even function. It attains the maximum value at $\tau = 0$, which is $E\{x(t)^2\}$. A random process is said to be wide-sense stationary if Eqs. (3.30) and (3.31) hold.

Clearly, all stationary processes are wide-sense stationary, but the converse is not true.

The concept of autocorrelation can be readily extended to the concept of cross-correlation, $R_{uv}(t_1, t_2)$, between two random processes $u(t)$ and $v(t)$. It is defined by

$$R_{uv}(t_1, t_2) = E\{u(t_1)v(t_2)\}. \tag{3.32}$$

It is often convenient to study a random process in the frequency domain. However, we shall postpone treatment of this material until Chapter 4.

3.3.2 Markov processes

A Markov process is a random process whose past has no influence on the future if its present is specified. This means the following: If $t_{n-1} < t_n$, then

$$p[x(t_n) \leq x_n | x(t), t \leq t_{n-1}] = p[x(t_n) \leq x_n | x(t_{n-1})].$$

From this, it follows that if

$$t_1 < t_2 < \cdots < t_n,$$

then

$$p[x(t_n) \leq x_n | x(t_{n-1}), \cdots, x(t_1)] = p[x(t_n) \leq x_n | x(t_{n-1})]. \tag{3.33}$$

From Eq. (3.33), it follows that

$$f(x_n | x_{n-1}, \cdots, x_1) = f(x_n | x_{n-1}).$$

Applying the chain rule, we thus obtain

$$f(x_1, \cdots, x_n) = f(x_n | x_{n-1}) f(x_{n-1} | x_{n-2}) \cdots f(x_2 | x_1) f(x_1). \tag{3.34}$$

Example 1: If x_n satisfies the recursion equation

$$x_{n+1} - g(x_n, n) = \eta_n,$$

where $g(x_n, n)$ is a function of x_n and n, η_n is a white noise of zero mean. Then x_n is a Markov process, since x_{n+1} is determined solely by x_n and η_n, and hence is independent of x_k for $k < n$.

Example 2: Let us consider a first-order autoregressive process denoted by AR(1). It is a special case of example 1 and is given by

$$x_{n+1} - \overline{x} = a(x_n - \overline{x}) + \eta_n, \tag{3.35}$$

where the constant coefficient a satisfies $0 \neq |a| < 1$.

We first find the variance σ^2 of the process. Taking the square of both sides of Eq. (3.35) and then taking the expectation, we have

$$R(0) = \sigma^2 = \sigma_n^2/(1 - a^2),$$

where σ_n^2 is the variance of the noise η_n. Next, we compute the autocorrelation function. We have

$$\begin{aligned} R(1) = E[(x_{n+1} - \overline{x})(x_n - \overline{x})] &= E[(x_n - \overline{x})(a(x_n - \overline{x}) + \eta_n)] \\ &= aE[(x_n - \overline{x})^2] = a\sigma^2 = aR(0). \end{aligned}$$

Similarly, we can obtain

$$R(2) = a^2\sigma^2 = a^2R(0).$$

Similar expressions can be obtained for negative time lag m. By induction, we then see that the autocorrelation function of x_n is given by

$$R(m) = \sigma^2 a^{|m|}. \tag{3.36}$$

It decays exponentially.

Example 3: In example 1, if $x_0 = 0$ and $g(x_n, n) = x_n$, then we have the random walk process given by

$$x_n = x_{n-1} + \eta_n = \eta_1 + \eta_2 + \cdots + \eta_n.$$

Since the variance of x_n is proportional to n, this process is not stationary. In general, a Markov process may not be stationary.

Below, we shall only consider homogeneous Markov processes where the conditional density $f(x_n|x_{n-1})$ does not depend on n. However, the first-order density $f(x_n)$ may depend on n. In general, a homogeneous process is not stationary. In many cases, it tends toward a stationary process as $n \to \infty$. To further simplify the matter, we shall only consider Markov processes with a finite or countably infinite number of states. Such processes are usually called Markov chains.

3.3.2.1 Homogeneous discrete-time (DT) Markov chains Let the N states be denoted by $E_1, E_2, \cdots, E_N$, or simply $1, 2, \cdots, N$. A DT Markov chain is specified by the transition probability matrix $\mathbf{P} = [p_{ij}]$, where

$$p_{ij} = P[X_n = j|X_{n-1} = i], \tag{3.37}$$

and its initial probability vector is

$$\vec{\pi}_0 = [\pi_0(1), \pi_0(2), \cdots, \pi_0(N)], \tag{3.38}$$

where the subscript 0 denotes the time step 0. Being a probability, p_{ij} has to be nonnegative. Furthermore, by the axioms of probability, the sum of each row of the matrix $\mathbf{P}$ has to be 1:

$$\sum_{j=1}^{N} p_{ij} = 1.$$

Now we can compute the probabilities of the system at any time step. In particular, at the time step 1, we have

$$\vec{\pi}_1 = \vec{\pi}_0 \mathbf{P}.$$

In general, we have

$$\vec{\pi}_k = \vec{\pi}_{k-1} \mathbf{P}.$$

When $\vec{\pi}_*$ satisfies

$$\vec{\pi}_* = \vec{\pi}_* \mathbf{P}$$

it is called the stationary distribution of the Markov chain.

Recall that when a vector $\vec{B}$ satisfies

$$\vec{B}\mathbf{P} = \lambda \vec{B}$$

it is called a left eigenvector of the matrix $\mathbf{P}$, while the number λ is called the eigenvalue. Therefore, $\vec{\pi}_*$ is the left eigenvector of $\mathbf{P}$ associated with the eigenvalue 1. Now let us assume that $\mathbf{P}$ has N distinct eigenvalues denoted by $1, \lambda_2, \lambda_3, \cdots, \lambda_N$. Their corresponding left eigenvectors are

$$\vec{\pi}_*, \vec{B_2}, \vec{B_3}, \cdots, \vec{B_N}.$$

These eigenvectors are linearly independent (but not necessarily orthogonal, since the matrix $\mathbf{P}$ is typically not symmetric). Expressing $\vec{\pi}_0$ as a linear combination of these eigenvectors,

$$\vec{\pi}_0 = \alpha_1 \vec{\pi}_* + \alpha_2 \vec{B_2} + \cdots + \alpha_N \vec{B_N},$$

where α_i are coefficients, we then have

$$\vec{\pi}_n = \vec{\pi}_0 \mathbf{P}^n = \alpha_1 \vec{\pi}_* + \alpha_2 \lambda_2^n \vec{B_2} + \cdots + \alpha_N \lambda_N^n \vec{B_N}. \tag{3.39}$$

We assume that when $n \to \infty$, all the exponential terms die out, and

$$\vec{\pi}_n \to \vec{\pi}_*.$$

Therefore, the coefficient α_1 has to be 1. The speed of convergence to the stationary distribution, of course, depends on the magnitudes of the eigenvalues $\lambda_2, \lambda_3, \cdots, \lambda_N$. Eq. (3.39) makes it clear that in general, a Markov process is not stationary. However, when $n \to \infty$, it may become stationary.

Example 4: Consider a two-state Markov chain. Its transition probability matrix is given by

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

and the initial probability vector is given by

$$\vec{\pi}_0 = [\pi_0(0), 1 - \pi_0(0)].$$

It is easy to find that the two eigenvalues are given by

$$\lambda_1 = 1, \quad \lambda_2 = 1 - p - q$$

and their corresponding eigenvectors are

$$\left[\frac{q}{p+q}, \frac{p}{p+q}\right], \quad [1, -1].$$

We have normalized the first eigenvector, since it corresponds to the stationary distribution $\vec{\pi}_*$. Also notice that the two eigenvectors are linearly independent but not orthogonal. Decomposing $\vec{\pi}_0$ in terms of the two eigenvectors, we have

$$[\pi_0(0), 1 - \pi_0(0)] = \left[\frac{q}{p+q}, \frac{p}{p+q}\right] + c[1, -1],$$

where $c = \pi_0(0) - q/(p+q)$. Therefore,

$$\vec{\pi}_n = \vec{\pi}_* + (\vec{\pi}_0 - \vec{\pi}_*)(1 - p - q)^n$$

when $p+q \neq 2$. When $p+q = 2$, the binary Markov chain is simply a deterministic system with period 2.

Now let us examine the time the system spends in a given state. We assert that the number of time units the system spends in the same state is geometrically distributed. To prove this, let us focus on an arbitrary state E_i. Assume that at time step k the system is at E_i. The system will remain in this state at the next time step with probability p_{ii}; similarly, it will leave this state at the next time step with probability $1 - p_{ii}$. If indeed it does remain in this state at the next time step, then the probability of its remaining for an additional time step is again p_{ii}, while the probability of leaving at this second time step is $1 - p_{ii}$. Since these probabilities are independent, we thus have

- P[system remains in E_i for exactly m additional time units given that it has entered E_i] $= (1 - p_{ii})p_{ii}^m$.

This is the geometrical distribution, as we have claimed. Note that the geometrical distribution is the unique discrete memoryless distribution.

3.3.2.2 *Homogeneous continuous-time (CT) Markov chains*

Simplifying the Markov property to a state space with a finite or infinitely countable number of states, we obtain the definition for a homogeneous CT Markov chain:

Definition: The random process $X(t)$ forms a CT Markov chain if for all integers n and for any sequence $t_1, t_2, \cdots, t_{n+1}$ such that $t_1 < t_2 < \cdots < t_{n+1}$ we have

$$\begin{aligned} P[X(t_{n+1}) &= j | X(t_1) = i_1, X(t_2) = i_2, \cdots, X(t_n) = i_n] \\ &= P[X(t_{n+1}) = j | X(t_n) = i_n]. \end{aligned} \tag{3.40}$$

The main body of a homogeneous CT Markov chain is described by differential equations. In developing signal processing tools based on random fractals, we do not have much use for these equations. Hence, we will not present the general theory here. Below, we shall discuss the memoryless property of CT Markov chains and then describe the Poisson process.

We have proven that a DT Markov chain has geometrically distributed state times (also called sojourn times). We will now prove that a CT Markov chain has exponentially distributed sojourn times. Let τ_i be a random variable representing the time the process spends in state E_i. Since the influences of the past trajectory of the process on its future development are completely specified by giving the current state of the process, we need not specify how *long* the process has been in the current state. This means that the remaining time in E_i must have a distribution that depends only upon i and not upon how long the process has been in E_i. We may write this statement as

$$P(\tau_i > s+t|\tau_i > s) = h(t),$$

where $h(t)$ is a function only of the additional time t (and not of the expended time s). The above equation can be rewritten as

$$P(\tau_i > s+t|\tau_i > s) = \frac{P(\tau_i > s+t,\ \tau_i > s)}{P(\tau_i > s)} = \frac{P(\tau_i > s+t)}{P(\tau_i > s)}.$$

The last step follows since the event $\tau_i > s+t$ implies the event $\tau_i > s$. We can rewrite the last equation as

$$P(\tau_i > s+t) = P(\tau_i > s)\, h(t). \tag{3.41}$$

Setting $s = 0$ and observing that $P(\tau_i > 0) = 1$, we immediately have

$$h(t) = P(\tau_i > t).$$

Substituting this last equation in Eq. (3.41), we obtain

$$P(\tau_i > s+t) = P(\tau_i > s)\, P(\tau_i > t) \tag{3.42}$$

for $s, t \geq 0$. We now show that the only continuous distribution that satisfies Eq. (3.42) is the exponential distribution. First, we have, by definition, the following general relationship:

$$\frac{d}{dt}P(\tau_i > t) = \frac{d}{dt}[1 - P(\tau_i \leq t)] = -f_{\tau_i}(t), \tag{3.43}$$

where $f_{\tau_i}(t)$ is the PDF for τ_i. Differentiating Eq. (3.42) with respect to s, we have

$$\frac{dP(\tau_i > s+t)}{ds} = -f_{\tau_i}(s)P(\tau_i > t).$$

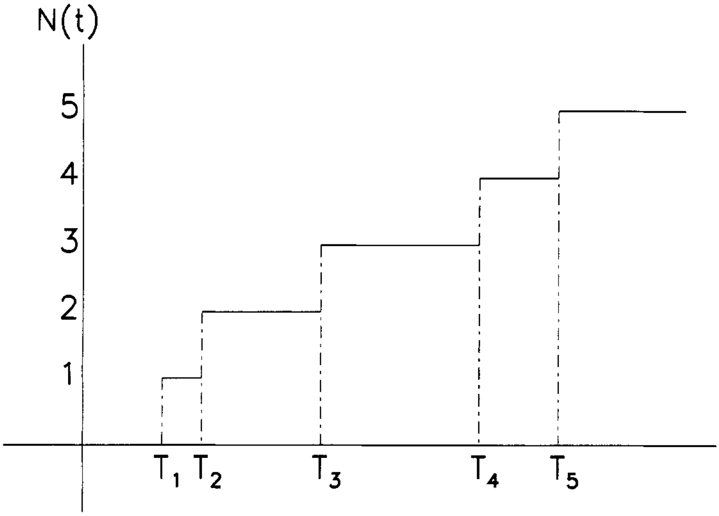

Figure 3.3. A typical realization of a Poisson process $N(t)$.

Dividing both sides by $P(\tau_i > t)$ and setting $s = 0$, we have

$$\frac{dP(\tau_i > t)}{P(\tau_i > t)} = -f_{\tau_i}(0)ds.$$

Integrating s from 0 to t, we have

$$P(\tau_i > t) = e^{-f_{\tau_i}(0)t}.$$

Hence, the PDF is given by

$$f_{\tau_i}(t) = f_{\tau_i}(0)e^{-f_{\tau_i}(0)t}, \quad t \geq 0.$$

This is the exponential sojourn time that we have claimed.

Poisson process: Suppose we are observing customers arriving at a store, light bulbs burning out, or discharging of a neuron. In such situations, we are most interested in how many events occur in a given time interval and the distribution of the time interval between successive events. The Poisson process is the simplest model used to describe such phenomena. It forms one of the most important classes of stochastic processes and finds applications in areas of science and engineering as diverse as physics, biology, and teletraffic modeling. Let us now give the formal definition of the Poisson process (see also Fig. 3.3).

Definition: A Poisson process having rate λ is a sequence of events such that

1. $N(0) = 0$.

2. The process has independent increments. That is, for $t_1 < t_2 < \cdots < t_n$, the random variables $N(t_2 - t_1)$, $N(t_3 - t_2)$, $\cdots$, $N(t_n - t_{n-1})$ are all independent.
3. The number of events in any interval of length t is Poisson distributed with mean λt. That is, $\forall s, t \geq 0$,

$$P[N(t+s) - N(s) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \quad n = 0, 1, \cdots .$$

Let us now examine the distribution for the time interval between two successive events. This time interval is called interarrival time in teletraffic, interspike interval in neuroscience, etc. Denote this time interval by U. It is obvious that the event $U > t$ is equivalent to the event $N(t) = 0$. Therefore,

$$P(U > t) = P(N(t) = 0) = e^{-\lambda t}.$$

We thus see that this time interval has an exponential distribution. In fact, a Poisson process can also be defined via interevent intervals, as shown below.

Suppose that $\{U_j,\ j = 1, 2, \cdots,\}$ are iid exponential random variables with rate λ, that is,

$$P(U_j > t) = e^{-\lambda t}.$$

Let $T_0 = 0$ and $T_n = T_{n-1} + U_n$ for $n \geq 1$. We think of T_n as the time of the occurrence of some random event and U_n as the time between successive occurrences. We have found out that T_n follows an Erlang distribution (Eq. 3.14)). We define a counting process $N = \{N(t)\}$ as follows. For $n = 1, 2, \cdots$,

$$N(t) < n \text{ if and only if } t < T_n.$$

In other words, $N(t) < n$ if the time of the nth occurrence is after t. Therefore,

$$P(N(t) \leq n) = P(N(t) < n+1) = P(T_{n+1} > t) = \int_t^\infty \frac{\lambda(\lambda u)^n}{(n)!} e^{-\lambda u} du.$$

Repeatedly integrating by parts shows that for $n = 0, 1, 2, \cdots$ we have

$$P(N(t) \leq n) = \sum_{k=0}^{n} \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \quad n = 0, 1, \cdots .$$

That is, $N(t)$ has a Poisson distribution with mean λt.

3.4 SPECIAL TOPIC: HOW TO FIND RELEVANT INFORMATION FOR A NEW FIELD QUICKLY

Nowadays phrases such as *interdisciplinary* or *multidisciplinary research* have become increasingly popular. When one is involved in such activities, often one finds it necessary to familiarize oneself with a completely new topic as quickly as possible. One can, of course, try to get started with a Google search. Sometimes information found that way may not be very specific. The authors' trick is to search the ISI Web of Knowledge. This little trick could be regarded as an effective application of the concept of power-law or scale-free networks, where the distribution of the number of links to a node in a network or graph follows a power law. In such a network, a few nodes have an extremely large number of links, while most others only have a few. The few nodes with a large number of links can be considered hubs. A citation network consists of nodes, which are published papers, and links, which are the citations of those papers in other papers. It has been found that such a network is a power-law network, with a few famous papers being cited thousands of times, while many unimportant papers have few or no citations. Of course, a paper with no citations may not be a bad paper. It could become a classic paper after many years. To get started in a new field as quickly as possible, it is wise to find some of the best papers in the field as well as a few important and recent papers on a topic that one is particularly interested in.

The url for the ISI Web of Knowledge is http://portal.isiknowledge.com/portal.cgi/ If one clicks on the first item, *Web of Science*, then one finds a page with multiple choices. One is *GENERAL SEARCH*. If one clicks on it, one again finds a page with multiple choices. Now, suppose we want to find out the characteristics of noise in nanotubes. Nanotubes are very appealing as bio-sensors. As a sensor, a noise level has to be very low. So this is a topic of considerable current interest. Now if one inputs

```
1/f noise in carbon nanotubes
```

to the box *TOPIC*, then, by March 6, 2007, one finds that the earliest paper was written by Collins et al.:

> Collins PG, Fuhrer MS, Zettl A
> 1/f noise in carbon nanotubes
> APPLIED PHYSICS LETTERS 76 (7): 894-896 FEB 14 2000

It has been cited 40 times. Its 40 citations form a small network. It contains a paper with 65 citations:

> Sinnott SB, Andrews R
> Carbon nanotubes: Synthesis, properties, and applications
> CRITICAL REVIEWS IN SOLID STATE AND MATERIALS SCIENCES 26 (3): 145-249 2001

This paper can be considered the hub of the small network. With these two papers, one can be well on the way to research on the noise character of carbon nanotubes. Interestingly, the noise in nanotubes is a $1/f^{\beta}$ process, which we will study in depth in later chapters.

Research-oriented readers are strongly encouraged to use this little trick.

3.5 BIBLIOGRAPHIC NOTES

Readers new to probability theory may find the small book [257] by two masters, Khinchin and Gnedenko, very helpful. Feller's classic [136] is always a pleasure to read. For systematic textbooks, we refer to [205,335]. Readers wanting to know more about large-deviations theory are referred to [143,205]. Readers interested in lognormality in network traffic are referred to [30,54,63,311], while those interested in lognormality in switching times in ambiguous perceptions are referred to [496]. An entertaining as well as insightful paper to read about lognormal distribution and many other topics is [316]. Finally, readers interested in Zipf's law and power-law networks may want to browse an interesting website created by Dr. Wentian Li: http://www.nslij-genetics.org/wli/, and an excellent review article by Albert and Barabasi [7].

3.6 EXERCISES

1. Let z be $N(0,1)$. Prove that $x = a + bz$ is $N(a, b^2)$ and that $w = e^{a+bz}$ has a log-normal distribution.

2. Prove that the geometrical distribution is the unique discrete memoryless distribution.

3. Let x and y be $N(0, \sigma)$ and independent. Define r and φ by the following equation:

$$x\cos(\omega t) + y\sin(\omega t) = r\cos(\omega t - \varphi), \quad |\varphi| < \pi.$$

   Prove that the random variables r and φ are independent, φ is uniform in the interval $[-\pi, \pi]$, and r has a Rayleigh distribution.

4. Simulate 10^4 exponentially distributed random variables with parameter λ. Then estimate PDF and CDF. Plot PDF and CDF, both in linear scale and semilog scale. Estimate λ as the slope of the semilog plots. Is the estimated λ the same as that used in the simulation?

5. Simulate 10^4 Pareto-distributed random variables with parameters α =0.5, 1, 1.5, 2, 2.5, and 3. Estimate PDF and CDF using simple histograms. Then plot them out in log-log scale. Estimate α as the slope of the log-log plots. Are the estimated α the same as those used in the simulations?

6. This is the same as exercise 5, but with a different way of estimating PDF. The procedure in exercise 5 can be termed equal-linear-bin. An alternative is to use the equal-log-bin procedure: before forming the histogram, take the log of the random variables first; then plot out the estimated PDF or CDF using a semi-log plot. Determine if this procedure improves the accuracy in estimating α. If it does, explain why.

7. Simulate a number of AR(1) processes described by Eq. (3.35) for $a = 0.1$, 0.2, $\cdots$, 0.9. Numerically compute the autocorrelation function for these processes and compare the result with the theory.

8. Construct a simple random walk process $Y = X_1 + X_2 + \cdots + X_n$, where $X_i \sim N(0, 1)$ are iid. Let $n = 1000$. Analytically find the mean and variance of Y. If you are asked to compute the mean and variance of Y numerically, how many realizations of Y would be needed? Compare your answer with the results of simulations.

9. Let X and Y be independent random variables with characteristic functions $\Phi_X(u)$ and $\Phi_Y(u)$, respectively. Let $Z = X + Y$. Prove that the characteristic function for Z is $\Phi_Z(u) = \Phi_X(u) \cdot \Phi_Y(u)$.