

# CHAPTER 1

---

## INTRODUCTION

---

All of us have experienced the annoyance of having to wait in line. Unfortunately, this phenomenon continues to be common in congested, urbanized, “high-tech” societies. We wait in line in our cars in traffic jams or at toll booths; we wait on hold for an operator to pick up our telephone calls; we wait in line at supermarkets to check out; we wait in line at fast-food restaurants; and we wait in line at banks and post offices. We, as customers, do not generally like these waits, and the managers of the establishments at which we wait also do not like us to wait, since it may cost them business. Why then is there waiting?

The answer is simple: There is more demand for service than there is facility for service available. Why is this so? There may be many reasons; for example, there may be a shortage of available servers, it may be infeasible economically for a business to provide the level of service necessary to prevent waiting, or there may be a space limit to the amount of service that can be provided. Generally these limitations can be removed with the expenditure of capital, and to know how much service should then be made available, one would need to know answers to such questions as, “How long must a customer wait?” and “How many people will form in the line?” Queueing theory attempts (and in many cases succeeds) to answer these questions through detailed mathematical analysis. The word “queue” is in more common usage

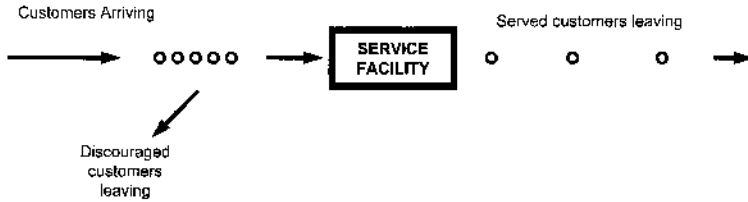


Figure 1.1 A typical queueing process.

in Great Britain and other countries than in the United States, but it is rapidly gaining acceptance in this country, although it must be admitted that it is just as displeasing to spend time in a queue as in a waiting line.

### 1.1 Description of the Queueing Problem

A queueing system can be described as customers arriving for service, waiting for service if it is not immediate, and if having waited for service, leaving the system after being served. The term “customer” is used in a general sense and does not imply necessarily a human customer. For example, a customer could be a ball bearing waiting to be polished, an airplane waiting in line to take off, or a computer program waiting to be run. Such a basic system can be schematically shown as in Figure 1.1. Although any queueing system may be diagrammed in this manner, it should be clear that a reasonably accurate representation of such a system would require a detailed characterization of the underlying processes.

Queueing theory was developed to provide models to predict the behavior of systems that attempt to provide service for randomly arising demands; not unnaturally, then, the earliest problems studied were those of telephone traffic congestion. The pioneer investigator was the Danish mathematician A. K. Erlang, who, in 1909, published “The Theory of Probabilities and Telephone Conversations.” In later works he observed that a telephone system was generally characterized by either (1) Poisson input, exponential holding (service) times, and multiple channels (servers), or (2) Poisson input, constant holding times, and a single channel. Erlang was also responsible for the notion of stationary equilibrium, for the introduction of the so-called balance-of-state equations, and for the first consideration of the optimization of a queueing system.

Work on the application of the theory to telephony continued after Erlang. In 1927, E. C. Molina published his paper “Application of the Theory of Probability to Telephone Trunking Problems,” which was followed one year later by Thornton Fry’s book *Probability and Its Engineering Uses*, which expanded much of Erlang’s earlier work. In the early 1930s, Felix Pollaczek did some further pioneering work on Poisson input, arbitrary output, and single- and multiple-channel problems. Additional work was done at that time in Russia by Kolmogorov and Khintchine, in France by Crommelin, and in Sweden by Palm. The work in queueing theory picked

up momentum rather slowly in its early days, but accelerated in the 1950s, and there has been a great deal of work in the area since then.

There are many valuable applications of the theory, most of which have been well documented in the literature of probability, operations research, management science, and industrial engineering. Some examples are traffic flow (vehicles, aircraft, people, communications), scheduling (patients in hospitals, jobs on machines, programs on a computer), and facility design (banks, post offices, amusement parks, fast-food restaurants).

Queueing theory originated as a very practical subject, but much of the literature up to the middle 1980s was of little direct practical value. However, queueing theorists have once again become concerned about the application of the sophisticated theory that has largely arisen since the close of World War II. The emphasis in the literature on the exact solution of queueing problems with clever mathematical tricks is now becoming secondary to model building and the direct use of these techniques in management decisionmaking. Most real problems do not correspond exactly to a mathematical model, and increasing attention is being paid to complex computational analysis, approximate solutions, sensitivity analyses, and the like. The development of the practice of queueing theory must not be restricted by a lack of closed-form solutions, and problem solvers must be able to put the developed theory to good use. These points should be kept in mind by the reader, and we attempt to illustrate them whenever possible throughout this text.

## 1.2 Characteristics of Queueing Processes

In most cases, six basic characteristics of queueing processes provide an adequate description of a queueing system: (1) arrival pattern of customers, (2) service pattern of servers, (3) queue discipline, (4) system capacity, (5) number of service channels, and (6) number of service stages.

### 1.2.1 Arrival Pattern of Customers

In usual queueing situations, the process of arrivals is stochastic, and it is thus necessary to know the probability distribution describing the times between successive customer arrivals (interarrival times). It is also necessary to know whether customers can arrive simultaneously (batch or bulk arrivals), and if so, the probability distribution describing the size of the batch.

It is also necessary to know the reaction of a customer upon entering the system. A customer may decide to wait no matter how long the queue becomes, or, on the other hand, if the queue is too long, the customer may decide not to enter the system. If a customer decides not to enter the queue upon arrival, the customer is said to have *balked*. A customer may enter the queue, but after a time lose patience and decide to leave. In this case, the customer is said to have *renege*d. In the event that there are two or more parallel waiting lines, customers may switch from one to another, that is,

*jockey* for position. These three situations are all examples of queues with *impatient customers*.

One final factor to be considered regarding the arrival pattern is the manner in which the pattern changes with time. An arrival pattern that does not change with time (i.e., the probability distribution describing the input process is time-independent) is called a *stationary* arrival pattern. One that is not time-independent is called *nonstationary*.

### 1.2.2 Service Patterns

Much of the previous discussion concerning the arrival pattern is appropriate in discussing service. Most importantly, a probability distribution is needed to describe the sequence of customer service times. Service may also be single or batch. One generally thinks of one customer being served at a time by a given server, but there are many situations where customers may be served simultaneously by the same server, such as a computer with parallel processing, sightseers on a guided tour, or people boarding a train.

The service process may depend on the number of customers waiting for service. A server may work faster if the queue is building up or, on the contrary, may get flustered and become less efficient. The situation in which service depends on the number of customers waiting is referred to as *state-dependent* service. Although this term was not used in discussing arrival patterns, the problems of customer impatience can be looked upon as ones of state-dependent arrivals, since the arrival behavior depends on the amount of congestion in the system.

Service, like arrivals, can be stationary or nonstationary with respect to time. For example, learning may take place, so that service becomes more efficient as experience is gained. The dependence on time is not to be confused with dependence on state. The former does not depend on the number of customers in the system, but rather on how long it has been in operation. The latter does not depend on how long the system has been in operation, but only on the state of the system at a given time, that is, on how many customers are currently in the system. Of course, a queueing system can be both nonstationary and state-dependent.

Even if the service rate is high, it is very likely that some customers will be delayed by waiting in the line. In general, customers arrive and depart at irregular intervals; hence the queue length will assume no definitive pattern unless arrivals and service are deterministic. Thus it follows that a probability distribution for queue lengths will be the result of two separate processes—arrivals and services—which are generally, though not universally, assumed mutually independent.

### 1.2.3 Queue Discipline

Queue discipline refers to the manner in which customers are selected for service when a queue has formed. The most common discipline that can be observed in everyday life is first come, first served (FCFS). However, this is certainly not the only possible queue discipline. Some others in common usage are last come, first

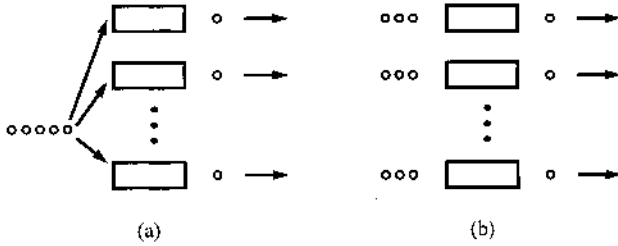


Figure 1.2 Multichannel queueing systems.

served (LCFS), which is applicable to many inventory systems when there is no obsolescence of stored units, as it is easier to reach the nearest items, which are the last in; selection for service in random order independent of the time of arrival to the queue (RSS); and a variety of *priority* schemes, where customers are given priorities upon entering the system, the ones with higher priorities to be selected for service ahead of those with lower priorities, regardless of their time of arrival to the system.

There are two general situations in priority disciplines. In the first, which is called *preemptive*, the customer with the highest priority is allowed to enter service immediately even if a customer with lower priority is already in service when the higher-priority customer enters the system; that is, the lower-priority customer in service is preempted, its service stopped, to be resumed again after the higher-priority customer is served. There are two possible additional variations: the preempted customer's service when resumed can either continue from the point of preemption or start anew. In the second general priority situation, called the *nonpreemptive* case, the highest-priority customer goes to the head of the queue but cannot get into service until the customer presently in service is completed, even though this customer has a lower priority.

### 1.2.4 System Capacity

In some queueing processes there is a physical limitation to the amount of waiting room, so that when the line reaches a certain length, no further customers are allowed to enter until space becomes available as the result of a service completion. These are referred to as finite queueing situations; that is, there is a finite limit to the maximum system size. A queue with limited waiting room can be viewed as one with forced balking where a customer is forced to balk if it arrives when the queue size is at its limit. This is a simple case, since it is known exactly under what circumstances arriving customers must balk.

### 1.2.5 Number of Service Channels

As we shortly explain in more detail, it is generally preferable to design multiserver queueing systems to be fed by a single line. Thus, when we specify the number of service channels, we are typically referring to the number of parallel service stations

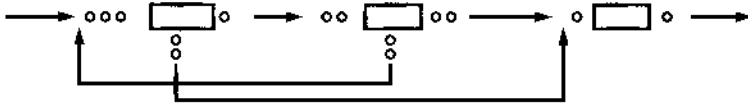


Figure 1.3 A multistage queueing system with feedback.

that can serve customers simultaneously. Figure 1.1 depicts an illustrative single-channel system, while Figure 1.2 shows two variations of multichannel systems. The two multichannel systems differ in that the first has a single queue, while the second allows a queue for each channel. A hair-styling salon with many chairs is an example of the first type of multichannel system (assuming no customer is waiting for any particular stylist), while a supermarket or fast-food restaurant might fit the second description. It is generally assumed that the service mechanisms of parallel channels operate independently of each other.

### 1.2.6 Stages of Service

A queueing system may have only a single stage of service, as in the hair-styling salon, or it may have several stages. An example of a multistage queueing system would be a physical examination procedure, where each patient must proceed through several stages, such as medical history; ear, nose, and throat examination; blood tests; electrocardiogram; eye examination; and so on. In some multistage queueing processes recycling or feedback may occur. Recycling is common in manufacturing processes, where quality control inspections are performed after certain stages, and parts that do not meet quality standards are sent back for reprocessing. Similarly, a telecommunications network may process messages through a randomly selected sequence of nodes, with the possibility that some messages will require rerouting on occasion through the same stage. A multistage queueing system with some feedback is depicted in Figure 1.3.

The six characteristics of queueing systems discussed in this section are generally sufficient to completely describe a process under study. Clearly, a wide variety of queueing systems can be encountered. Before performing any mathematical analyses, however, it is absolutely necessary to describe adequately the process being modeled. Knowledge of the basic six characteristics is essential in this task.

It is extremely important to use the correct model or at least the model that best describes the real situation being studied. A great deal of thought is often required in this *model selection procedure*. For example, let us reconsider the supermarket mentioned previously. Suppose there are  $c$  checkout counters. If customers choose a checkout counter on a purely random basis (without regard to the queue length in front of each counter) and never switch lines (no jockeying), then we truly have  $c$  independent single-channel models. If, on the other hand, there is a single waiting line and when a checker becomes idle, the customer at the head of the line (or with the lowest number if numbers are given out) enters service, we have a  $c$ -channel model. Neither, of course, is generally the case in most supermarkets. What usually happens

is that queues form in front of each counter, but new customers enter the queue that is the shortest (or has shopping carts that are lightly loaded). Also, there is a great deal of jockeying between lines. Now the question becomes which choice of models ( $c$  independent single channels or a single  $c$ -channel) is more appropriate. If there were complete jockeying, the single  $c$ -channel model would be quite appropriate, since even though in reality there are  $c$  lines, there is little difference, when jockeying is present, between these two cases. This is so because no servers will be idle as long as customers are waiting for service, which would not be the case with  $c$  truly independent single channels. As jockeying is rather easy to accomplish in supermarkets, the  $c$ -channel model will be more appropriate and realistic than the  $c$ -single-channels model, which one might have been tempted to choose initially prior to giving much thought to the process. Thus it is important not to jump to hasty conclusions but to select carefully the most appropriate model.

### 1.3 Notation

As a shorthand for describing queuing processes, a notation has evolved, due for the most part to Kendall (1953), which is now rather standard throughout the queuing literature. A queuing process is described by a series of symbols and slashes such as  $A/B/X/Y/Z$ , where  $A$  indicates in some way the interarrival-time distribution,  $B$  the service pattern as described by the probability distribution for service time,  $X$  the number of parallel service channels,  $Y$  the restriction on system capacity, and  $Z$  the queue discipline (Appendix 1 contains a dictionary of symbols used throughout this text). Some standard symbols for these characteristics are presented in Table 1.1. For example, the notation  $M/D/2/\infty/FCFS$  indicates a queuing process with exponential interarrival times, deterministic service times, two parallel servers, no restriction on the maximum number allowed in the system, and first-come, first-served queue discipline.

In many situations only the first three symbols are used. Current practice is to omit the service-capacity symbol if no restriction is imposed ( $Y = \infty$ ) and to omit the queue discipline if it is first come, first served ( $Z = FCFS$ ). Thus  $M/D/2$  would be a queuing system with exponential input, deterministic service, two servers, no limit on system capacity, and first-come, first-served discipline.

The symbols in Table 1.1 are, for the most part, self-explanatory; however, a few require further comment. The symbol  $G$  represents a general probability distribution; that is, no assumption is made as to the precise form of the distribution. Results in these cases are applicable to any probability distribution. These general-time distributions, however, are required to represent independent and identically distributed random variables.

It may also appear strange that the symbol  $M$  is used for exponential. The use of the symbol  $E$ , as one might expect, would be too easily confused with  $E_k$ , which is used for the type- $k$  Erlang distribution (a gamma with an integer shape parameter). So  $M$  is used instead; it stands for the Markovian or memoryless property of the exponential, which is developed in some detail in Section 1.9.

Table 1.1 Queuing Notation  $A/B/X/Y/Z$

Characteristic	Symbol	Explanation
Interarrival-time distribution ( $A$ )	$M$	Exponential
	$D$	Deterministic
Service-time distribution ( $B$ )	$E_k$	Erlang type $k(k = 1, 2, \dots)$
	$H_k$	Mixture of $k$ exponentials
	$PH$	Phase type
	$G$	General
# of parallel servers ( $X$ )	$1, 2, \dots, \infty$	
Max. system capacity ( $Y$ )	$1, 2, \dots, \infty$	
Queue discipline ( $Z$ )	FCFS	First come, first served
	LCFS	Last come, first served
	RSS	Random selection for service
	PR	Priority
	GD	General discipline

The reader may have noticed that the list of symbols is not complete. For example, there is no indication of a symbol to represent bulk arrivals, to represent series queues, to denote any state dependence, and so on. If a suitable notation does exist for any previously unmentioned model, it is indicated when that particular model is brought up in the text. However, there still remain models for which no symbolism has either been developed or accepted as standard, and this is generally true for those models less frequently analyzed in the literature.

### 1.4 Measuring System Performance

Up to now the concentration has been on the physical description of queueing processes. What, then, might one like to know about the effectiveness of a queueing system? Generally there are three types of system responses of interest: (1) some measure of the waiting time that a typical customer might be forced to endure; (2) an indication of the manner in which customers may accumulate; and (3) a measure of the idle time of the servers. Since most queueing systems have stochastic elements, these measures are often random variables and their probability distributions, or at the very least their expected values, are desired.

There are two types of customer waiting times, the time a customer spends in the queue and the total time a customer spends in the system (queue plus service). Depending on the system being studied, one may be of more interest than the other. For example, if we are studying an amusement park, it is the time waiting in the queue that makes the customer unhappy. On the other hand, if we are dealing with

machines that require repair, then it is the total down time (queue wait plus repair time) that we wish to keep as small as possible. Correspondingly, there are two customer accumulation measures as well: the number of customers in the queue and the total number of customers in the system. The former would be of interest if we desire to determine a design for waiting space (say, the number of seats to have for customers waiting in a hair-styling salon), while the latter may be of interest for knowing how many of our machines may be unavailable for use. Idle-service measures can include the percentage of time any particular server may be idle, or the time the entire system is devoid of customers.

The task of the queueing analyst is generally one of two things. He or she is either to determine the values of appropriate measures of effectiveness for a given process, or to design an "optimal" (according to some criterion) system. To do the former, one must relate waiting delays, queue lengths, and such to the given properties of the input stream and the service procedures. On the other hand, for the design of a system the analyst might want to balance customer waiting time against the idle time of servers according to some inherent cost structure. If the costs of waiting and idle service can be obtained directly, they can be used to determine the optimum number of channels to maintain and the service rates at which to operate these channels. Also, to design the waiting facility it is necessary to have information regarding the possible size of the queue to plan for waiting room. There may also be a space cost that should be considered along with customer-waiting and idle-server costs to obtain the optimal system design. In any case, the analyst will strive to solve this problem by analytical means; however, if these fail, he or she must resort to simulation. Ultimately, the issue generally comes down to a trade-off of better customer service versus the expense of providing more service capability, that is, determining the increase in investment of service for a corresponding decrease in customer delay.

## 1.5 Some General Results

We present some general results and relationships for  $G/G/1$  and  $G/G/c$  queues in this section, prior to specific model development. These results will prove useful in many of the following sections and chapters, as well as providing some insight at this early stage.

Denoting the average rate of customers entering the queueing system as  $\lambda$  and the average rate of serving customers as  $\mu$ , a measure of traffic congestion for  $c$ -server systems is  $\rho \equiv \lambda/c\mu$  (often called *traffic intensity*). When  $\rho > 1$  ( $\lambda > c\mu$ ), the average number of arrivals into the system exceeds the maximum average service rate of the system, and we would expect, as time goes on, the queue to get bigger and bigger, unless, at some point, customers were not allowed to join. If we are interested in steady-state conditions (the state of the system after it has been in operation a long time), when  $\rho > 1$ , the queue size never settles down (assuming customers are not prevented from entering the system) and there is no steady state. It turns out that for steady-state results to exist,  $\rho$  must be strictly less than 1 (again, assuming no denial of customer entry). When  $\rho = 1$ , unless arrivals and service are deterministic

and perfectly scheduled, no steady state exists, since randomness will prevent the queue from ever emptying out and allowing the servers to catch up, thus causing the queue to grow without bound. Therefore, if one knows the average arrival rate and average service rate, the minimum number of parallel servers required to guarantee a steady-state solution can be calculated immediately by finding the smallest  $c$  such that  $\lambda/c\mu < 1$ .

What we most often desire in solving queueing models is to find the probability distribution for the total number of customers in the system at time  $t$ ,  $N(t)$ , which is made up of those waiting in queue,  $N_q(t)$ , plus those in service,  $N_s(t)$ . Let  $p_n(t) = \Pr\{N(t) = n\}$ , and  $p_n = \Pr\{N = n\}$  in the steady state. Considering  $c$ -server queues in steady state, two expected-value measures of major interest are the mean number in the system,

$$L = E[N] = \sum_{n=0}^{\infty} np_n,$$

and the expected number in queue,

$$L_q = E[N_q] = \sum_{n=c+1}^{\infty} (n - c)p_n.$$

### 1.5.1 Little's Formulas

One of the most powerful relationships in queueing theory was developed by John D. C. Little in the early 1960s (see Little, 1961, for the original proof—a host of papers refining the proof followed in the ensuing decades). Little related the steady-state mean system sizes to the steady-state average customer waiting times as follows. Letting  $T_q$  represent the time a customer (transaction) spends waiting in the queue prior to entering service and  $T$  represent the total time a customer spends in the system ( $T = T_q + S$ , where  $S$  is the service time, and  $T$ ,  $T_q$ , and  $S$  are random variables), two often used measures of system performance with respect to customers are  $W_q = E[T_q]$  and  $W = E[T]$ , the mean waiting time in queue and the mean waiting time in the system, respectively. Little's formulas are

$$L = \lambda W \tag{1.1a}$$

and

$$L_q = \lambda W_q. \tag{1.1b}$$

Thus it is necessary to find only one of the four expected-value measures, in view of Little's formulas and the fact that  $E[T] = E[T_q] + E[S]$ , or, equivalently,  $W = W_q + 1/\mu$ , where  $\mu$ , as before, is the mean service rate.

Although the following does not constitute a proof, we illustrate the concept of Little's formulas by considering a *sample path* of one *busy period* (time from when

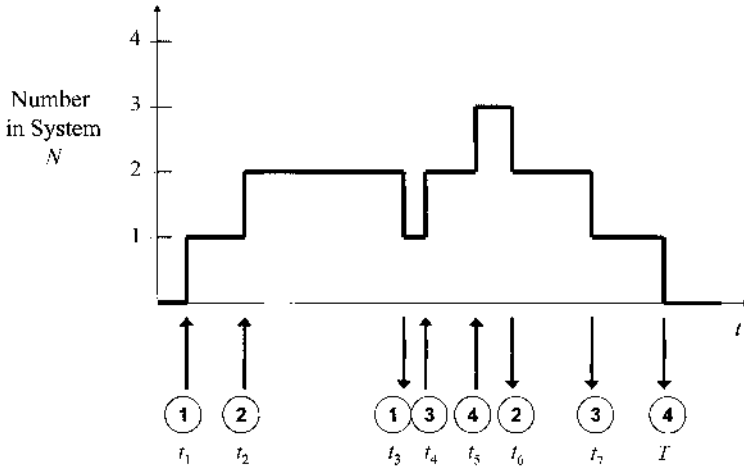


Figure 1.4 Busy-period sample path.

a customer enters an empty system until it next empties out again). Consider the illustration in Figure 1.4, where the number of customers (say,  $N_c$ ) that arrive over the time period  $(0, T)$  is 4.

The calculations for  $L$  and  $W$  are

$$\begin{aligned}
 L &= [1(t_2 - t_1) + 2(t_3 - t_2) + 1(t_4 - t_3) + 2(t_5 - t_4) \\
 &\quad + 3(t_6 - t_5) + 2(t_7 - t_6) + 1(T - t_7)]/T \\
 &= (\text{area under curve})/T \\
 &= (T + t_7 + t_6 - t_5 - t_4 + t_3 - t_2 - t_1)/T
 \end{aligned} \tag{1.2a}$$

and

$$\begin{aligned}
 W &= [(t_3 - t_1) + (t_6 - t_2) + (t_7 - t_4) + (T - t_5)]/4 \\
 &= (T + t_7 + t_6 - t_5 - t_4 + t_3 - t_2 - t_1)/4 \\
 &= (\text{area under curve})/N_c.
 \end{aligned} \tag{1.2b}$$

Thus we see from (1.2a) and (1.2b) that the area under curve is  $LT = WN_c$ , which yields  $L = WN_c/T$ . The fraction  $N_c/T$  is the number of customers arriving over the time  $T$  and is, for this period, the arrival rate  $\lambda$ , so that  $L = \lambda W$ . A similar argument would hold for a picture of the number in the queue  $N_q$  over the period  $(0, T)$ , yielding  $L_q = \lambda W_q$ . While this is not a proof (since it needs to be shown that these relationships hold in the limit over many busy periods as time goes to infinity), one can see the idea behind the relationships.

An interesting result that can be derived from Little's formulas [(1.1a) and (1.1b)] and the relation between  $W$  and  $W_q$  is

$$L - L_q = \lambda(W - W_q) = \lambda(1/\mu) = \lambda/\mu. \tag{1.3}$$

Table 1.2 Summary of General Results for  $G/G/c$  Queues

$\rho = \lambda/c\mu$	Traffic intensity; offered work load rate to a server
$L = \lambda W$	Little's formula
$L_q = \lambda W_q$	Little's formula
$W = W_q + 1/\mu$	Expected-value argument
$p_b = \lambda/c\mu = \rho$	Busy probability for an arbitrary server
$r = \lambda/\mu$	Expected number of customers in service; offered work load rate
$L = L_q + r$	Combined result—(1.3)
$p_0 = 1 - \rho$	$G/G/1$ empty-system probability
$L = L_q + (1 - p_0)$	Combined result for $G/G/1$

But  $L - L_q = E[N] - E[N_q] = E[N - N_q] = E[N_s]$ , so that the expected number of customers in service in the steady state is  $\lambda/\mu$ , which we will denote by  $r$ . Note for a single-server system that  $r = \rho$  and it also follows from simple algebra that

$$L - L_q = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} (n-1)p_n = \sum_{n=1}^{\infty} p_n = 1 - p_0.$$

From this, we can easily derive the probability that any given server is busy in a multiserver system in the steady state. We denote this probability by  $p_b$ . Since we have just shown that the expected number present in service at any instant in the steady state is  $r$ , it follows from the symmetry of the  $c$  servers that the expected number present at one server is  $r/c$ . Then, by a simple expected-value argument, we can show that  $p_b = \rho$ , since

$$r/c = \rho = 0 \cdot (1 - p_b) + 1 \cdot p_b.$$

For a single-server queue ( $G/G/1$ ), the probability of the system being idle ( $N = 0$ ) is the same as the probability of a server being idle. Thus  $p_0 = 1 - p_b$  in this case, and  $p_0 = 1 - \rho = 1 - r = 1 - \lambda/\mu$ . The quantity  $r = \lambda/\mu$ , the expected number of customers in service, has another interesting connotation. It is sometimes also referred to as the *offered load*, since, on average, each customer requires  $1/\mu$  time units of service and the average number of customers arriving per unit time is  $\lambda$ , so that the product  $\lambda(1/\mu)$  is the amount of work arriving to the system per unit time. Dividing this by the number of servers  $c$  (which yields  $\rho$ ) gives the average amount of work coming to each server per unit time.

Table 1.2 summarizes the results of this section.

### 1.6 Simple Data Bookkeeping for Queues

At this point, it might be useful to use a table format to show how the random events of arrivals and service completions interact for a sample single-server system

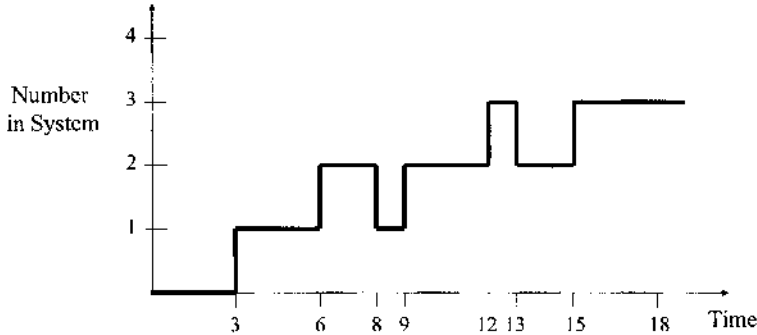


Figure 1.5 Sample path for queuing process.

to form a queue. In the following, we begin at time 0 with a first arrival and then update the system state when events (arrivals or departures) occur—thus the name *event-oriented bookkeeping* is used for this sort of table.

Consider the elementary case of a constant rate of arrivals to a single channel that possesses a constant service rate. (Figure 1.5 is an illustration of this with interarrival times of 3 and serve times of 5.) These regularly spaced arrivals are to be served first come, first served (FCFS). Let it also be assumed that at time  $t = 0$  there are no customers waiting and that the channel is empty. Let  $\lambda$  be defined as the number of arrivals per unit time, and  $1/\lambda$  then will be the constant time between successive arrivals. The particular unit of time (minutes, hours, etc.) is up to the choice of the analyst. However, consistency must be adhered to once the unit is chosen so that the same basic unit is used throughout the analysis. Similarly, if  $\mu$  is to be the rate of service in terms of completions per unit time when the server is busy, then  $1/\mu$  is the constant service time. We would like to calculate the number in the system at an arbitrary time  $t$ , say,  $n(t)$ , and the time the  $n$ th arriving customer must wait in the queue to obtain service, say,  $W_q^{(n)}$ . From these, it then becomes easy to compute the major measures of effectiveness. Under the assumption that as soon as a service is completed another is begun, the number in the system (including the customer in service) at time  $t$  is determined by the equation

$$n(t) = \{\text{number of arrivals in } (0, t]\} - \{\text{number of services completed in } (0, t]\}. \quad (1.4)$$

It should be pointed out that there are usually three waiting times of interest—the time spent by the  $n$ th customer waiting for service (or line delay), which we write here as  $W_q^{(n)}$ ; the time the  $n$ th customer spent in the system, which we shall call  $W^{(n)}$ ; and what is called the virtual line wait  $V(t)$ , namely, the wait a fictitious arrival would have to endure if it arrived at time  $t$ . The reader is cautioned that various authors are not consistent and each of these quantities is sometimes referred to simply as the waiting time.

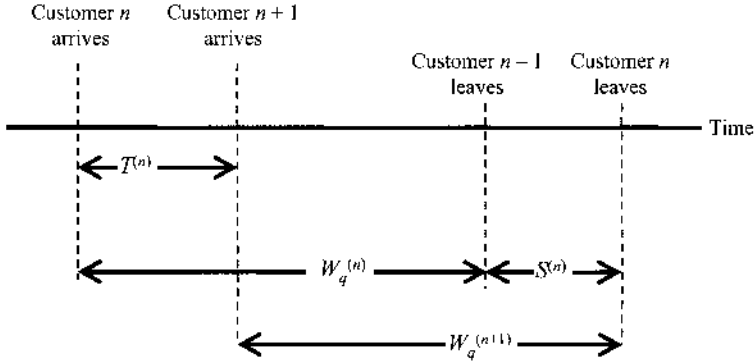


Figure 1.6 Successive  $G/G/1$  waiting times.

To find the waiting times in queue until service begins, we observe that the line waits  $W_q^{(n)}$  and  $W_q^{(n-1)}$  of two successive customers in *any* single-server queue (deterministic or otherwise) are related by the simple recurrence relation

$$W_q^{(n+1)} = \begin{cases} W_q^{(n)} + S^{(n)} - T^{(n)} & (W_q^{(n)} + S^{(n)} - T^{(n)} > 0), \\ 0 & (W_q^{(n)} + S^{(n)} - T^{(n)} \leq 0). \end{cases} \quad (1.5)$$

where  $S^{(n)}$  is the service time of the  $n$ th customer and  $T^{(n)}$  is the interarrival time between the  $n$ th and  $(n + 1)$ st customers. This can be seen by a simple diagram as shown in Figure 1.6. (This is an important general relation that is also utilized in later portions of the text.)

Bookkeeping has to do with updating the system status when events occur, recording items of interest, and calculating measures of effectiveness. Event-oriented bookkeeping updates the system state only when events (arrivals or departures) occur. Since there is not necessarily an event every basic time unit, in next-event bookkeeping the master clock is increased by a variable amount each time, rather than a fixed amount as it would be in time-oriented bookkeeping. The event-oriented approach will be illustrated here by an example, using the arrival and service data given in Table 1.3.

We see from simple averaging calculations for columns (5) and (6) in Table 1.4 that the mean line delay of the 12 customers was  $40/12 = \frac{10}{3}$ , while their mean system waiting time turned out to be  $70/12 = \frac{35}{6}$ . Furthermore, we observe that we can estimate the mean arrival rate as  $\frac{12}{31}$  customers per unit time, since there were 12 arrivals over the 31-time-unit observation horizon. Thus the application of Little's law to these numbers tells us that the average system size  $L$  over the full time horizon was

$$L = \lambda W = \frac{70/12}{31/12} = \frac{70}{31}.$$

The mean queue size can be computed similarly.

Table 1.3 Input Data

$i$	1	2	3	4	5	6	7	8	9	10	11	12
Interarrival time between customers $i + 1$ and $i$	2	1	3	1	1	4	2	5	1	4	2	—
Customer $i$ service time	1	3	6	2	1	1	4	2	5	1	1	3

Table 1.4 Event-Oriented Bookkeeping

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Master Clock Time	Arrival/Departure Customer $i$	Time Arrival $i$ Enters Service	Time Arrival $i$ Leaves Service	Time in Queue	Time in System	No. in Queue Just After Master Clock Time	No. in System Just After Master Clock Time
0	1-A	0	1	0	1	0	1
1	1-D					0	0
2	2-A	2	5	0	3	0	1
3	3-A	5	11	2	8	1	2
5	2-D					0	1
6	4-A	11	13	5	7	1	2
7	5-A	13	14	6	7	2	3
8	6-A	14	15	6	7	3	4
11	3-D					2	3
12	7-A	15	19	3	7	3	4
13	4-D					2	3
14	8-A;5-D	19	21	5	7	2	3
15	6-D					1	2
19	9-A;7-D	21	26	2	7	1	2
20	10-A	26	27	6	7	2	3
21	8-D					1	2
24	11-A	27	28	3	4	2	3
26	12-A;9-D	28	31	2	5	2	3
27	10-D					1	2
28	11-D					0	1
31	12-D					0	0

## 1.7 Poisson Process and the Exponential Distribution

The most common stochastic queueing models assume that interarrival times and service times obey the exponential distribution or, equivalently, that the arrival rate and service rate follow a Poisson distribution. In this section we will derive the Poisson distribution and show that assuming the number of occurrences in some time interval to be a Poisson random variable is equivalent to assuming the time between successive occurrences to be an exponentially distributed random variable.

We consider an arrival counting process  $\{N(t), t \geq 0\}$ , where  $N(t)$  denotes the total number of arrivals up to time  $t$ , with  $N(0) = 0$ , and which satisfies the following three assumptions:

- (i) The probability that an arrival occurs between time  $t$  and time  $t + \Delta t$  is equal to  $\lambda \Delta t + o(\Delta t)$ . We write this as  $\Pr\{\text{arrival occurs between } t \text{ and } t + \Delta t\} = \lambda \Delta t + o(\Delta t)$ , where  $\lambda$  is a constant independent of  $N(t)$ ,  $\Delta t$  is an incremental element, and  $o(\Delta t)$  denotes a quantity that becomes negligible when compared to  $\Delta t$  as  $\Delta t \rightarrow 0$ ; that is,

$$\lim_{\Delta t \rightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

- (ii)  $\Pr\{\text{more than one arrival between } t \text{ and } t + \Delta t\} = o(\Delta t)$ .

- (iii) The numbers of arrivals in nonoverlapping intervals are statistically independent; that is, the process has independent increments.

We wish to calculate  $p_n(t)$ , the probability of  $n$  arrivals in a time interval of length  $t$ ,  $n$  being an integer  $\geq 0$ . We will do this by first developing differential-difference equations for the arrival process. For  $n \geq 1$  we have

$$\begin{aligned} p_n(t + \Delta t) &= \Pr\{n \text{ arrivals in } t \text{ and none in } \Delta t\} \\ &\quad + \Pr\{n - 1 \text{ arrivals in } t \text{ and one in } \Delta t\} \\ &\quad + \Pr\{n - 2 \text{ arrivals in } t \text{ and two in } \Delta t\} + \dots \\ &\quad - \Pr\{\text{no arrivals in } t \text{ and } n \text{ in } \Delta t\}. \end{aligned} \quad (1.6)$$

Using assumptions i, ii, and iii, (1.6) becomes

$$p_n(t + \Delta t) = p_n(t)[1 - \lambda \Delta t - o(\Delta t)] + p_{n-1}(t)[\lambda \Delta t + o(\Delta t)] + o(\Delta t), \quad (1.7)$$

where the last term,  $o(\Delta t)$ , represents the terms  $\Pr\{n - j \text{ arrivals in } t \text{ and } j \text{ in } \Delta t; 2 \leq j \leq n\}$ .

For the case  $n = 0$ , we have

$$p_0(t + \Delta t) = p_0(t)[1 - \lambda \Delta t - o(\Delta t)]. \quad (1.8)$$

Rewriting (1.7) and (1.8) and combining all  $o(\Delta t)$  terms, we have

$$p_0(t + \Delta t) - p_0(t) = -\lambda \Delta t p_0(t) + o(\Delta t) \quad (1.9)$$

and

$$p_n(t + \Delta t) - p_n(t) = -\lambda \Delta t p_n(t) + \lambda \Delta t p_{n-1}(t) + o(\Delta t) \quad (n \geq 1). \quad (1.10)$$

We divide (1.9) and (1.10) by  $\Delta t$ , take the limit as  $\Delta t \rightarrow 0$ , and obtain the differential-difference equations

$$\begin{aligned} \lim_{\Delta t \rightarrow 0} \left[ \frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + \frac{o(\Delta t)}{\Delta t} \right], \\ \lim_{\Delta t \rightarrow 0} \left[ \frac{p_n(t + \Delta t) - p_n(t)}{\Delta t} = -\lambda p_n(t) + \lambda p_{n-1}(t) + \frac{o(\Delta t)}{\Delta t} \right] \quad (n \geq 1), \end{aligned}$$

which reduce to

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) \quad (1.11)$$

and

$$\frac{dp_n(t)}{dt} = -\lambda p_n(t) + \lambda p_{n-1}(t) \quad (n \geq 1). \quad (1.12)$$

We now have an infinite set of linear, first-order ordinary differential equations to solve. Equation (1.11) clearly has the general solution  $p_0(t) = Ce^{-\lambda t}$ , where the constant  $C$  is easily determined to be equal to 1, since  $p_0(0) = 1$ . Next, let  $n = 1$  in (1.12), and we find that

$$\frac{dp_1(t)}{dt} = -\lambda p_1(t) + \lambda p_0(t),$$

or

$$\frac{dp_1(t)}{dt} + \lambda p_1(t) = \lambda p_0(t) = \lambda e^{-\lambda t}.$$

The solution to this equation is

$$p_1(t) = Ce^{-\lambda t} + \lambda t e^{-\lambda t}.$$

Use of the boundary condition  $p_n(0) = 0$  for all  $n > 0$  yields  $C = 0$  and gives

$$p_1(t) = \lambda t e^{-\lambda t}.$$

Continuing sequentially to  $n = 2, 3, \dots$  in (1.12) and proceeding similarly, we find that

$$p_2(t) = \frac{(\lambda t)^2}{2} e^{-\lambda t}, \quad p_3(t) = \frac{(\lambda t)^3}{3!} e^{-\lambda t}, \quad \dots \quad (1.13)$$

From (1.13), we conjecture that the general formula is

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \quad (1.14)$$

It is left as an exercise (Problem 1.13) to use mathematical induction to verify (1.14), which is the well-known formula for a Poisson probability distribution with mean

$\lambda t$ . Thus if we consider the random variable defined as the number of arrivals to a queueing system by time  $t$ , this random variable has the Poisson distribution given by (1.14) with a mean of  $\lambda t$  arrivals, or a mean arrival rate (arrivals per unit time) of  $\lambda$ .

Poisson processes have a number of interesting additional properties. One of the most important is that the numbers of occurrences in intervals of equal width are identically distributed (stationary increments). In particular, for  $t > s$ , the difference  $N(t) - N(s)$  is identically distributed as  $N(t+h) - N(s+h)$ , with probability function

$$p_n(t-s) = \frac{[\lambda(t-s)]^n}{n!} e^{-\lambda(t-s)}.$$

This can easily be seen by the following argument. Since the Poisson has independent increments (assumption iii), there is no loss of generality if  $N(s)$  and  $N(s+h)$  are assumed to be zero. Then if the Poisson derivation is carried out for both  $N(t)$  and  $N(t+h)$  under assumptions i, ii, and iii, the foregoing formula results for each (see Problem 1.16).

We now show that if the arrival process is Poisson, an associated random variable defined as the time between successive arrivals (interarrival time) follows the exponential distribution. Let  $T$  be the random variable "time between successive arrivals"; then

$$\Pr\{T \geq t\} = \Pr\{\text{no arrivals in time } t\} = p_0(t) = e^{-\lambda t}.$$

Therefore we see that the cumulative distribution function of  $T$  can be written as

$$A(t) = \Pr\{T \leq t\} = 1 - e^{-\lambda t},$$

with corresponding density function

$$a(t) = \frac{dA(t)}{dt} = \lambda e^{-\lambda t}.$$

Thus  $T$  has the exponential distribution with mean  $1/\lambda$ . We would intuitively expect the *mean* time between arrivals to be  $1/\lambda$  if the *mean* arrival rate is  $\lambda$ . Our analysis substantiates this. It can also be shown that if the interarrival times are independent and have the same exponential distribution, then the arrival rate follows the Poisson distribution, and a proof of this assertion follows.

To begin, let the cumulative distribution function (CDF) of the arrival counting process,  $\Pr\{N(t) \leq n\}$ , be denoted by  $P_n(t)$ . Then it follows that

$$\begin{aligned} p_n(t) &= \Pr\{N(t) = n\} \\ &= P_n(t) - P_{n-1}(t). \end{aligned}$$

But

$$P_n(t) = \Pr\{(\text{sum of } n+1 \text{ interarrival times}) > t\}.$$

However, the sum of independent and identically distributed exponential random variables has an Erlang distribution (which is a special type of gamma distribution); hence

$$P_n(t) = \int_t^\infty \frac{\lambda(\lambda x)^n}{n!} e^{-\lambda x} dx. \tag{1.15}$$

The transformation of variables  $u = x - t$  gives

$$\begin{aligned} P_n(t) &= \int_0^\infty \frac{\lambda^{n+1}(u+t)^n}{n!} e^{-\lambda t} e^{-\lambda u} du \\ &= \int_0^\infty \frac{\lambda^{n+1} e^{-\lambda t} e^{-\lambda u}}{n!} \sum_{i=0}^n u^{n-i} t^i \frac{n!}{(n-i)! i!} du, \end{aligned}$$

from the binomial theorem. The summation and integral may be switched to give

$$P_n(t) = \sum_{i=0}^n \frac{\lambda^{n+1} e^{-\lambda t} t^i}{(n-i)! i!} \int_0^\infty e^{-\lambda u} u^{n-i} du.$$

But the integral in the above equation is essentially the well-known gamma function and equals  $(n-i)!/\lambda^{n-i+1}$ . So

$$P_n(t) = \sum_{i=0}^n \frac{(\lambda t)^i e^{-\lambda t}}{i!},$$

which is clearly recognizable as the CDF of the Poisson process.

The Poisson-exponential arrival process derived here is sometimes referred to as completely random arrivals. Although the reader might think that completely random would allude to some sort of haphazard arrival process or a uniform distribution for interarrival times, when encountered in queueing literature it specifically refers to the Poisson-arrival-rate-exponential-interarrival-time pattern. This can be explained in light of the following characteristic of a Poisson process. Given that  $k$  arrivals have occurred in an interval  $[0, T]$ , the  $k$  times  $\tau_1 < \tau_2 < \dots < \tau_k$  at which the arrivals occurred are distributed as the order statistics of  $k$  uniform random variables on  $[0, T]$ . Note that it is not the interarrival times, but rather the times at which the arrivals occurred, that are uniformly distributed. This can be proved as follows.

The differential element of the conditional probability density may be written as

$$\begin{aligned} f_\tau(t|k) dt &\equiv f(t_1, t_2, \dots, t_k | k \text{ arrivals in } [0, T]) dt_1 dt_2 \dots dt_k \\ &\approx \Pr\{t_1 \leq \tau_1 \leq t_1 + dt_1, \dots, t_k \leq \tau_k \leq t_k + dt_k | k \text{ arrivals in } [0, T]\}. \end{aligned}$$

Using the definition of conditional probability gives

$$\begin{aligned} f_\tau(t|k) dt &= \frac{\Pr\{t_1 \leq \tau_1 \leq t_1 + dt_1, \dots, t_k \leq \tau_k \leq t_k + dt_k \text{ and } k \text{ arrivals in } [0, T]\}}{\Pr\{k \text{ arrivals in } [0, T]\}} \end{aligned}$$

The numerator of the right-hand side above can be found by making direct use of the Poisson probability function and its properties, since we wish to find the probability that exactly one event occurs in each of the  $k$  time intervals  $(t_i, t_i + dt_i)$ , and no events occur elsewhere, that is, in  $T - dt_1 - dt_2 - \dots - dt_k$ . Therefore, since the probability of  $k$  occurrences in a time  $t$  is Poisson, we have

$$\begin{aligned} f_\tau(t|k) dt &\approx \frac{\lambda dt_1 e^{-\lambda dt_1} \lambda dt_2 e^{-\lambda dt_2} \dots \lambda dt_k e^{-\lambda dt_k} e^{-\lambda(T-dt_1-dt_2-\dots-dt_k)}}{(\lambda T)^k e^{-\lambda T}/k!} \\ &= \frac{k!}{T^k} dt_1 dt_2 \dots dt_k. \end{aligned}$$

Hence

$$f_\tau(t_1, t_2, \dots, t_k | k \text{ arrivals in } [0, T]) = \frac{k!}{T^k}, \quad (1.16)$$

which is identical to the joint density of the order statistics of  $k$  random variables uniform on  $[0, T]$ .

One important consequence of the uniform property of the Poisson process is that the outcomes of random observations of a stochastic process  $X(t)$  have the same probabilities as if the scans were taken at Poisson-selected points. When  $X(t)$  is a queue, this property is called *PASTA*, for “Poisson arrivals see time averages.”

Making similar assumptions to those above for arrivals, one could utilize the same type of process to describe the service pattern. If we change the three assumptions in the beginning of this section slightly by using the word service instead of arrival and condition the probability statements by requiring the system to be nonempty, we obtain a Poisson service rate and an exponential service-time distribution for describing the service pattern. In the following section, we prove an important property of the exponential distribution which aids in a relatively simple analysis of queueing problems when arrival and service patterns exhibit the Poisson-exponential characteristics as derived in this section.

## 1.8 Markovian Property of the Exponential Distribution

We will now prove the Markovian or memoryless property of the exponential distribution. To explain this property in words, suppose service times are exponentially distributed. This property states that the probability that a customer currently in service has  $t$  units of remaining service is independent of how long it has already been in service. Thus we wish to prove that

$$\Pr\{T \leq t_1 | T \geq t_0\} = \Pr\{0 \leq T \leq t_1 - t_0\}. \quad (1.17)$$

The proof is relatively straightforward and proceeds as follows. From the definition of conditional probability we have

$$\begin{aligned} \Pr\{T \leq t_1 | T \geq t_0\} &= \frac{\Pr\{(T \leq t_1) \text{ and } (T \geq t_0)\}}{\Pr\{T \geq t_0\}} \\ &= \frac{e^{-\lambda t_0} - e^{-\lambda t_1}}{e^{-\lambda t_0}} \\ &= 1 - e^{-\lambda(t_1 - t_0)} = \Pr\{0 \leq T \leq t_1 - t_0\}. \end{aligned}$$

It is also true that the exponential distribution is the only continuous distribution that exhibits this memoryless property. (The only other distribution to exhibit this property is the geometric, which is the discrete analog of the exponential.) The proof of this assertion rests on the fact that the only continuous function solution of the equation

$$g(s + t) = g(s) + g(t)$$

is the linear form

$$g(y) = Cy, \tag{1.18}$$

where  $C$  is an arbitrary constant. This rather intuitive result turns out not to be a trivial matter to prove. However, the proof is well documented in the literature (e.g., see Parzen, 1962), and, for the purposes of this text, this additional detail is not necessary. Then, under the assumption of this result, we proceed as follows. We wish to show that if the memorylessness of (1.17) holds, then it follows that

$$\Pr\{T \leq t\} = F(t) = 1 - e^{Ct}.$$

Now subtract both sides of (1.17) from 1 and denote the complementary CDF by  $\tilde{F}$ . Thus

$$\tilde{F}(t_1 | T \geq t_0) = \tilde{F}(t_1 - t_0). \tag{1.19}$$

From the laws of conditional probability, we can write (1.19) as

$$\frac{\tilde{F}(t_1 \text{ and } T \geq t_0)}{\tilde{F}(t_0)} = \frac{\tilde{F}(t_1)}{\tilde{F}(t_0)} = \tilde{F}(t_1 - t_0),$$

or

$$\tilde{F}(t_1) = \tilde{F}(t_0)\tilde{F}(t_1 - t_0).$$

Letting  $t = t_1 - t_0$  yields

$$\tilde{F}(t + t_0) = \tilde{F}(t_0)\tilde{F}(t),$$

and when natural logarithms are taken of both sides, it is found that

$$\ln \tilde{F}(t + t_0) = \ln \tilde{F}(t_0) - \ln \tilde{F}(t).$$

It thus follows from (1.18) that

$$\ln \tilde{F}(t) = Ct, \quad \text{or} \quad \tilde{F}(t) = e^{Ct}.$$

Thus  $F(t) = 1 - e^{Ct}$ .

There are many possible and well-known generalizations of the Poisson process, most of which have rather obvious applications to queues and are taken up in greater detail later in the text. The most obvious of the generalizations is a truncation of the infinite domain, that is, omission of some of the nonnegative integers from the range of possible values. This is done whenever the removed values are either theoretically meaningless or practically unobservable. An example of this occurs in the  $M/M/c/c$  queue and gives rise to *Erlang's loss* or the *Erlang-B formula*. The only change to be made here, with caution, is the rescaling of the respective probabilities because the Poisson terms no longer sum to one.

Another generalization arises if we go back to the axiomatic derivation and no longer permit  $\lambda$  to be a constant independent of time. If instead the functional relationship is denoted by  $\lambda(t)$ , then the probability of one occurrence in a small time increment is rewritten as  $\lambda(t)\Delta t + o(\Delta t)$ , and it turns out that the resulting distribution of the counting process is the so-called nonhomogeneous Poisson given by

$$p_n(t) = e^{-m(t)} \frac{[m(t)]^n}{n!}, \quad m(t) = \int_0^t \lambda(s) ds \quad (n \geq 0).$$

A third, and very common, generalization occurs when one relaxes the Poisson assumption that more than one occurrence in  $\Delta t$  has probability  $o(\Delta t)$ . Instead, let

$$\Pr\{i \text{ occurrences in } (t, t + \Delta t)\} = \lambda_i \Delta t + o(\Delta t) \quad (i = 1, 2, \dots, n)$$

with

$$\sum_{i=1}^n \lambda_i = \lambda.$$

It should be immediately clear now that this is equivalent to allowing the event of  $i$  simultaneous occurrences in  $\Delta t$  with probability  $\lambda_i \Delta t + o(\Delta t)$ , and each individual stream of occurrences of the same batch size ( $i$ ) itself forms a Poisson process. If these substreams are denoted by  $N_i(t)$ , then it should also be clear that the total process is  $N(t) = \sum_i iN_i(t)$ , with probability function

$$\begin{aligned} p_n(t) &= \Pr\{n \text{ occurrences in } [0, t]\} \\ &= \sum_{i=0}^n e^{-\lambda t} \frac{(\lambda t)^i}{i!} c_n^{(i)} \quad (c_0^{(0)} \equiv 1), \end{aligned}$$

where  $c_n^{(i)}$  is the probability that  $i$  occurrences give a grand total of  $n$  (i.e., the probability associated with the  $i$ -fold convolution of the batch-size probabilities  $\{\lambda_i/\lambda\}$ ). The process  $N(t)$  is known as the *multiple Poisson* and clearly also has the stationary and independent-increment properties.

The foregoing probability function  $p_n(t)$  has an alternative derivation as a compound distribution, since it admits of a random sum interpretation as follows. Consider the process  $N(t)$  to be defined by

$$N(t) = \sum_{n=1}^{M(t)} Y_n,$$

where  $M(t)$  is a regular Poisson process and  $\{Y_n\}$  is a sequence of independent and identically distributed (IID) discrete random variables with probabilities

$$c_j = \Pr\{Y_n = j\} \quad (\text{for all } n) \\ = \frac{\lambda_j}{\sum_i \lambda_i};$$

that is, occurrences happen according to a Poisson process  $\{M(t)\}$  but are not necessarily singles in that their size is  $j$  with probability  $c_j$ . Then, by the laws of probability,

$$\Pr\{N(t) = m\} = \sum_{k=0}^m \left[ \Pr\{M(t) = k\} \cdot \Pr \left\{ \sum_{n=1}^k Y_n = m \right\} \right] \\ = \sum_{k=0}^m e^{-\lambda t} \frac{(\lambda t)^k}{k!} c_m^{(k)}.$$

In other words, the compound approach looks at the process as one Poisson stream with a randomly varying batch size, whereas the multiple approach looks at the process as the sum of Poisson streams, each with a constant batch size.

Poisson-exponential streams are special cases of a larger class of problems called *renewal processes*. An ordinary renewal process arises from any sequence of non-negative IID random variables. Many of the properties that we have derived for Poisson-exponential sequences can also be derived in a renewal context. Some other specific results will be needed later in the text, particularly when the input is an arbitrary  $G$  stream, but these will be derived as needed by direct probabilistic arguments. The reader particularly interested in renewal theory is referred to Chapter 7 of Ross (2007), Chapter 3 of Resnick (1992), Chapter 9 of Çinlar (1975), or Chapter 5 of Heyman and Sobel (1982).

In subsequent chapters of the book, the Poisson process and its associated characteristics will play a key role in the development of many queueing models. This is true not only because of the many mathematically agreeable properties of the Poisson-exponential, but also because many real-life situations in fact do obey the appropriate requirements. Though it may seem at first glance that the demands of exponential interoccurrence times are rather stringent, this is not the case.

A strong argument in favor of exponential inputs is the one that often occurs in the context of reliability. It is the result of the well-known fact that the limit of a binomial distribution is Poisson, which says that if a mechanism consists of many parts, each of which can fail with only a small probability, and if the failures for the different parts are mutually independent and identical, then the total flow of failures can be considered Poisson. Another view that favors the exponential comes from the theory of extreme values. Here the exponential appears quite frequently as the limiting distribution of the (normalized) *first-order statistic* of random samples drawn from continuous populations (see Problem 1.24 for one such example).

There is also an additional argument that comes out of information theory. It is that the exponential distribution is the one that provides the least information,

where information content or negative entropy of the distribution  $f(x)$  is defined as  $\int f(x) \log f(x) dx$ . It can easily be shown that the exponential distribution has least information or highest entropy, and is therefore the most random law that can be used, and thus certainly provides a reasonably conservative approach. We treat the topic of choosing the appropriate probability model in more detail in Chapters 6 and 8.

## 1.9 Stochastic Processes and Markov Chains

A stochastic process is the mathematical abstraction of an empirical process whose development is governed by probabilistic laws (the Poisson process is one example). From the point of view of the mathematical theory of probability, a stochastic process is best defined as a family of random variables,  $\{X(t), t \in T\}$ , defined over some index set or parameter space  $T$ . The set  $T$  is sometimes also called the time range, and  $X(t)$  denotes the state of the process at time  $t$ . Depending on the nature of the time range, the process is classified as a discrete-parameter or continuous-parameter process as follows:

1. If  $T$  is a countable sequence, for example,  $T = \{0, \pm 1, \pm 2, \dots\}$  or  $T = \{0, 1, 2, \dots\}$ , then the stochastic process  $\{X(t), t \in T\}$  is said to be a discrete-parameter process defined on the index set  $T$ .
2. If  $T$  is an interval or an algebraic combination of intervals, for example,  $T = \{t: -\infty < t < +\infty\}$  or  $T = \{t: 0 < t < +\infty\}$ , then the stochastic process  $\{X(t), t \in T\}$  is called a continuous-parameter process defined on the index set  $T$ .

### 1.9.1 Markov Process

A discrete-parameter stochastic process  $\{X(t), t = 0, 1, 2, \dots\}$  or a continuous-parameter stochastic process  $\{X(t), t > 0\}$  is said to be a Markov process if, for any set of  $n$  time points  $t_1 < t_2 < \dots < t_n$  in the index set or time range of the process, the conditional distribution of  $X(t_n)$ , given the values of  $X(t_1), X(t_2), X(t_3), \dots, X(t_{n-1})$ , depends only on the immediately preceding value,  $X(t_{n-1})$ ; more precisely, for any real numbers  $x_1, x_2, \dots, x_n$ ,

$$\begin{aligned} \Pr\{X(t_n) \leq x_n | X(t_1) = x_1, \dots, X(t_{n-1}) = x_{n-1}\} \\ = \Pr\{X(t_n) \leq x_n | X(t_{n-1}) = x_{n-1}\}. \end{aligned}$$

In nonmathematical language one says that, given the "present" condition of the process, the "future" is independent of the "past," and the process is thus "memoryless."

Markov processes are classified according to:

1. the nature of the index set of the process (whether discrete or continuous parameter), and
2. the nature of state space of the process (whether discrete or continuous parameter).

Table 1.5 Classification of Markov Processes

State Space	Type of Parameter	
	Discrete	Continuous
Discrete	(Discrete-parameter) Markov chain	Continuous-parameter Markov chain
Continuous	Discrete-parameter Markov process	Continuous-parameter Markov process

A real number  $x$  is said to be a state of a stochastic process  $\{X(t), t \in T\}$  if there exists a time point  $t$  in  $T$  such that the  $\Pr\{x - h < X(t) < x + h\}$  is positive for every  $h > 0$ . The set of possible states constitutes the state space of the process. If the state space is discrete, the Markov process is generally called a Markov chain, although some authors reserve the term “chain” for only those Markov processes with both discrete state space and discrete-parameter space. In this text, we shall say that a discrete-parameter Markov process with discrete state space is a plain Markov chain, and that a continuous-parameter Markov process with discrete state space is a continuous-time Markov chain. (Multivariate extensions can be nicely formulated for vector states  $\mathbf{x}$ .)

A Markov chain is finite if the state space is finite; otherwise it is a denumerable or infinite Markov chain. Since a discrete-parameter process is observed at a countable number of time points, let the successive observations be denoted by  $X_0, X_1, X_2, \dots, X_n, \dots$  where  $X_n$  is the random variable whose value represents the state of the system at the  $n$ th time point. An arbitrary sequence of random variables  $\{X_n\}$  is thus a Markov chain if each random variable  $X_n$  is discrete and the following holds: For any integer  $m > 2$  and any set of  $m$  points  $n_1 < n_2 < \dots < n_m$ , the conditional distribution of  $X_{n_m}$ , given values of  $X_{n_1}, X_{n_2}, \dots, X_{n_{m-1}}$  depends only on  $X_{n_{m-1}}$ , the immediately preceding value; that is,

$$\begin{aligned} \Pr\{X_{n_m} = x_{n_m} | X_{n_1} = x_{n_1}, \dots, X_{n_{m-1}} = x_{n_{m-1}}\} \\ = \Pr\{X_{n_m} = x_{n_m} | X_{n_{m-1}} = x_{n_{m-1}}\}. \end{aligned}$$

When a Markov process has a continuous-state space and discrete-parameter space, we call it a discrete-parameter Markov process. If both the state space and parameter space are continuous, it is called a continuous-parameter Markov process. Table 1.5 summarizes our classification scheme for Markov processes.

An important generalization of the Markov chain, which is very useful in queuing, is the semi-Markov process (SMP) or Markov renewal process (MRP). The state transitions in an SMP form a discrete Markov chain, but the times between successive transitions are random variables. If these random variables are distributed exponentially for a continuous-parameter case or geometrically for a discrete-parameter case,

with mean dependent on the current state only, the SMP reduces to a Markov process because of the memoryless property of these random variables.

### 1.9.2 Discrete-Parameter Markov Chains

Consider a sequence of random variables,  $\{X_n, n = 0, 1, 2, \dots | X_0 = 0, 1, 2, \dots\}$ , which forms a Markov chain with discrete-parameter space; that is, for all  $n$ ,

$$\Pr\{X_n = j | X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}\} = \Pr\{X_n = j | X_{n-1} = i_{n-1}\}.$$

If the value of the random variable  $X_n$  is  $j$ , then the system is said to be in state  $j$  after  $n$  steps or transitions. The conditional probabilities  $\Pr\{X_n = j | X_{n-1} = i\}$  are called the *single-step transition probabilities* or just *transition probabilities*. If these probabilities are independent of  $n$ , then the chain is said to be *homogeneous* and the probabilities  $\Pr\{X_n = j | X_{n-1} = i\}$  can be written as  $p_{ij}$ . The matrix formed by placing  $p_{ij}$  in the  $(i, j)$  location is known as the *transition matrix* or *chain matrix* (call it  $\mathbf{P}$ ). For homogeneous chains, the  $m$ -step transition probabilities

$$\Pr\{X_{n+m} = j | X_n = i\} = p_{ij}^{(m)}$$

are also independent of  $n$ . The unconditional probability of state  $j$  at the  $n$ th trial will be written as

$$\Pr\{X_n = j\} = \pi_j^{(n)},$$

so that the initial distribution is given by  $\pi_j^{(0)}$ .

From the basic laws of probability, one can easily show that the matrix formed by the elements  $\{p_{ij}^{(m)}\}$ , say,  $\mathbf{P}^{(m)}$ , can be found simply by multiplying  $\mathbf{P}^{(m-k)}$  by  $\mathbf{P}^{(k)}$  for any value of  $k$ ,  $0 < k < m$ . This is the matrix equivalent of the well-known Chapman-Kolmogorov (CK) equations for this Markov process, namely,

$$p_{ij}^{(m)} = \sum_r p_{ir}^{(m-k)} p_{rj}^{(k)} \quad (0 < k < m)$$

or, in matrix notation,

$$\mathbf{P}^{(m)} = \mathbf{P}^{(m-k)} \mathbf{P}^{(k)}. \quad (1.20)$$

Letting  $k = m - 1$  in (1.20) gives

$$\mathbf{P}^{(m)} = \mathbf{P} \cdot \mathbf{P}^{(m-1)}, \quad (1.21)$$

and, continuing this procedure recursively, we can easily show that

$$\mathbf{P}^{(m)} = \mathbf{P} \cdot \mathbf{P} \cdots \mathbf{P} = \mathbf{P}^m.$$

Hence  $\mathbf{P}^{(m)}$  can be obtained by multiplying the matrix  $\mathbf{P}$  by itself  $m$  times.

Often one is interested in the probabilities of being in state  $j$  after  $m$  transitions regardless of the starting state. If the vector  $\pi^{(m)}$  is created from the probabilities  $\{\pi_j^{(m)}\}$ , then

$$\pi^{(m)} = \pi^{(m-1)} \mathbf{P}, \quad (1.22)$$

which, when used recursively, gives

$$\pi^{(m)} = \pi^{(0)} \mathbf{P}^m \tag{1.23}$$

for the initial state vector  $\pi^{(0)}$ .

Defining the matrix  $\mathbf{Q}$  as  $\mathbf{P} - \mathbf{I}$ , where  $\mathbf{I}$  is the identity matrix, and subtracting  $\pi^{(m-1)}$  from both sides of (1.22) gives

$$\pi^{(m)} - \pi^{(m-1)} = \pi^{(m-1)} \mathbf{Q}. \tag{1.24}$$

Note that  $\mathbf{P}$  is always a *stochastic matrix* (i.e., its rows sum to one), while  $\mathbf{Q}$  has rows that sum to zero.

### 1.9.3 Continuous-Parameter Markov Chains

We now consider a continuous-parameter Markov chain  $\{X(t), t \in T\}$  for  $T = \{t: 0 \leq t < \infty\}$ . Consider any times  $s > t > u \geq 0$  and states  $i$  and  $j$ ; then

$$p_{ij}(u, s) = \sum_r p_{ir}(u, t) p_{rj}(t, s), \tag{1.25}$$

where  $p_{ij}(u, s)$  is the probability of moving from state  $i$  to  $j$  in the time beginning at  $u$  and ending at  $s$ , and the summation is over all states of the chain. This result should be fairly intuitive and says that the chain can reach state  $j$  at time  $s$  by starting from state  $i$  at time  $u$  and stopping off at time  $t$  at any other possible state  $r$ . This is the CK equation for the continuous process [analogous to (1.20) for the discrete process]. In matrix notation, (1.25) becomes

$$\mathbf{P}(u, s) = \mathbf{P}(u, t) \mathbf{P}(t, s).$$

Letting  $u = 0$  and  $s = t + \Delta t$  in (1.25) gives

$$p_{ij}(0, t + \Delta t) = \sum_r p_{ir}(0, t) p_{rj}(t, t + \Delta t).$$

Defining  $p_i(0)$  as the probability that the chain starts in state  $i$  at time 0 and  $p_j(t)$  as the unconditional probability that the chain is in state  $j$  at time  $t$  regardless of starting state, we multiply the above equation by  $p_i(0)$  and sum over all states  $i$  to get

$$\sum_i p_i(0) p_{ij}(0, t + \Delta t) = \sum_r \sum_i p_{ir}(0, t) p_i(0) p_{rj}(t, t + \Delta t),$$

or

$$p_j(t + \Delta t) = \sum_r p_r(t) p_{rj}(t, t + \Delta t). \tag{1.26}$$

For the Poisson process treated earlier,

$$p_{rj}(t, t + \Delta t) = \begin{cases} \lambda \Delta t + o(\Delta t) & (r = j - 1, j \geq 1), \\ 1 - \lambda \Delta t + o(\Delta t) & (r = j), \\ o(\Delta t) & (\text{elsewhere}). \end{cases}$$

Substituting this into (1.26) gives

$$p_j(t + \Delta t) = [\lambda \Delta t + o(\Delta t)]p_{j-1}(t) + [1 - \lambda \Delta t + o(\Delta t)]p_j(t) + o(\Delta t) \quad (j \geq 1),$$

which is (1.7). What we did in Section 1.7 was to derive the CK equation for the Poisson process from scratch, appealing to the same basic probability arguments that yield the general CK equations (1.25) and (1.26).

There is an additional theory that takes one from the CK equation to two differential equations, which are called the forward and backward equations. If the transition probability functions  $p(u, s)$  of the chain have the additional properties that there exist continuous functions  $q_i(t)$  and  $q_{ij}(t)$  such that

$$\Pr\{\text{a change of state in } (t, t + \Delta t)\} = 1 - p_{ii}(t, t + \Delta t) = q_i(t)\Delta t + o(\Delta t), \quad (1.27a)$$

$$p_{ij}(t, t + \Delta t) = q_{ij}(t)\Delta t + o(\Delta t), \quad (1.27b)$$

then under some mild regularity conditions (e.g., see Ross, 1996), (1.25) leads to (see Problem 1.20)

$$\frac{\partial}{\partial t} p_{ij}(u, t) = -q_j(t)p_{ij}(u, t) + \sum_{r \neq j} p_{ir}(u, t)q_{rj}(t) \quad (1.28a)$$

and

$$\frac{\partial}{\partial u} p_{ij}(u, t) = q_i(u)p_{ij}(u, t) - \sum_{r \neq i} q_{ir}(u)p_{rj}(u, t). \quad (1.28b)$$

These two differential equations are known, respectively, as Kolmogorov's forward and backward equations. Consider (1.28a) further. Let  $u = 0$ , and assume a homogeneous process so that  $q_i(t) = q_i$  and  $q_{ij}(t) = q_{ij}$  for all  $t$ . We then obtain

$$\frac{dp_{ij}(0, t)}{dt} = -q_j p_{ij}(0, t) + \sum_{r \neq j} p_{ir}(0, t)q_{rj}.$$

Multiplying both sides of the above equation by  $p_i(0)$  and summing over all  $i$  yields

$$\frac{dp_j(t)}{dt} = -q_j p_j(t) + \sum_{r \neq j} p_r(t)q_{rj}.$$

In matrix notation, this equation can be written as

$$p'(t) = p(t)Q, \quad (1.29)$$

where  $\mathbf{p}(t)$  is the vector  $(p_0(t), p_1(t), p_2(t), \dots)$ ,  $\mathbf{p}'(t)$  is the vector of its derivatives, and

$$\mathbf{Q} = \begin{pmatrix} -q_0 & q_{01} & q_{02} & q_{03} & \dots \\ q_{10} & -q_1 & q_{12} & q_{13} & \dots \\ q_{20} & q_{21} & -q_2 & q_{23} & \dots \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

Note from (1.27b) that  $q_i = \sum_{j \neq i} q_{ij}$ , since

$$\sum_j p_{ij}(t, t + \Delta t) = 1,$$

which implies that

$$1 - q_i \Delta t + o(\Delta t) + \sum_{j \neq i} [q_{ij} \Delta t + o(\Delta t)] = 1,$$

or

$$-q_i \Delta t + o(\Delta t) = - \sum_{j \neq i} [q_{ij} \Delta t + o(\Delta t)],$$

so that  $q_i = \sum_{j \neq i} q_{ij}$ .

The matrix  $\mathbf{Q}$  can also be looked at as

$$\mathbf{Q} = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{P}(t, t + \Delta t) - \mathbf{I}}{\Delta t},$$

where

$$\mathbf{P}(t, t + \Delta t) = \{p_{ij}(t, t + \Delta t)\}.$$

Thus  $\mathbf{Q}$  plays a similar role for continuous-parameter Markov chains to the one  $\mathbf{Q} = \mathbf{P} - \mathbf{I}$  played for discrete-parameter Markov chains. The matrix  $\mathbf{Q}$  is often called the *intensity matrix* (or *infinitesimal generator*) of the continuous-parameter Markov chain.

Again referring to the Poisson process, we can use (1.29) to get (1.12) directly by noting that  $q_j = \lambda$ ,  $q_{rj} = \lambda$  for  $r = j - 1$  and  $j > 1$ , and  $q_{ij} = 0$  elsewhere, so that

$$\frac{dp_j(t)}{dt} = -\lambda p_j(t) + \lambda p_{j-1}(t).$$

Equation (1.11) can be obtained similarly.

Since the state space of any queue is composed of nonnegative integers (representing the number of customers present), a large percentage of queueing problems can be categorized as continuous-parameter (time) Markov chains. Many such models have the additional *birth-death* property that the net change across an infinitesimal time interval can never be other than  $-1, 0$ , or  $+1$ , and that

$$\Pr\{\text{increase } n \rightarrow n + 1 \text{ in } (t, t + \Delta t)\} = \lambda_n \Delta t + o(\Delta t) \quad (n \geq 0),$$

$$\Pr\{\text{decrease } n \rightarrow n - 1 \text{ in } (t, t + \Delta t)\} = \mu_n \Delta t + o(\Delta t) \quad (n \geq 1).$$

Hence

$$\Pr\{\text{no change in } (t, t + \Delta t)\} = 1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t)$$

and

$$\begin{aligned} q_{n,n+1} &= \lambda_n, \\ q_{n,n-1} &= \mu_n \quad (\mu_n \neq 0), \\ q_{rj} &= 0 \quad (\text{elsewhere}), \\ q_n &= \lambda_n + \mu_n \quad (q_0 = \lambda_0), \end{aligned}$$

so that it is possible to get either a forward or a backward Kolmogorov equation.

Substituting for  $q_i$  and  $q_{ij}$ , the matrix  $Q$  is

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

and using the Kolmogorov forward equation (1.29), we obtain a set of differential-difference equations for the birth-death process, namely,

$$\begin{aligned} \frac{dp_j(t)}{dt} &= -(\lambda_j + \mu_j)p_j(t) + \lambda_{j-1}p_{j-1}(t) + \mu_{j+1}p_{j+1}(t) \quad (j \geq 1), \\ \frac{dp_0(t)}{dt} &= -\lambda_0p_0(t) + \mu_1p_1(t). \end{aligned} \quad (1.30)$$

Note that setting  $\mu = 0$  and  $\lambda_i = \lambda$  for all  $i$  also gives the Poisson equations (1.11) and (1.12) of the previous section. The Poisson process is thus often called the *pure birth* process (see Problem 1.21).

### 1.9.4 Imbedded Markov Chains

In many of the situations in this text requiring the use of a continuous-parameter queueing model, we can often get quite satisfactory results by instead looking at the state of the queue only at certain selected times, leading to an *imbedded* discrete-parameter Markov chain. To illustrate this, consider the birth-death process *only* at transition times; that is, we create an imbedded stochastic process that **turns out** to be a discrete-parameter Markov chain. The transition probability matrix  $P = \{p_{ij}\}$ , where  $p_{ij}$  is the probability that, given the process is in state  $i$  and a transition occurs, it goes next to state  $j$ , can be shown to be (see Problem 1.23)

$$p_{ij} = \begin{cases} \frac{\lambda_i}{\lambda_i + \mu_i} & (j = i + 1, i \geq 1), \\ \frac{\mu_i}{\lambda_i + \mu_i} & (j = i - 1, i \geq 1), \\ 1 & (i = 0, j = 1), \\ 0 & \text{otherwise.} \end{cases} \quad (1.31)$$

Another way to view a continuous-time Markov chain is as a process that traverses from state to state in continuous time, the holding times being exponential (required to satisfy the Markov property) random variables with mean  $1/q_i$  [in the birth-death example,  $1/q_i = 1/(\lambda_i + \mu_i)$ ] for holding in state  $i$ . At times of transition, we have a discrete-parameter Markov chain with transition probabilities  $p_{ij} = q_{ij}/q_i$  [in the birth-death example,  $p_{ij}$  is given by (1.31)].

### 1.9.5 Long-Run Behavior of Markov Processes

We are often interested in the behavior of a Markov process after a long period of time, particularly in whether its behavior “settles down” probabilistically. We discuss three related concepts having to do with long-run behavior, namely, *limiting distributions*, *stationary distributions*, and *ergodicity*.

Consider a discrete-parameter Markov chain, and suppose that

$$\lim_{m \rightarrow \infty} p_{ij}^{(m)} = \pi_j \quad (\text{for all } i);$$

that is, after a long time, the probability that the process is in state  $j$  given that it started in state  $i$  is independent of the starting state  $i$ . This means that  $\mathbf{P}^m$  approaches a limit as  $m$  goes to infinity, namely, that all rows of  $\mathbf{P}^m$  become identical. We call the  $\{\pi_j\}$ , the *limiting* or *steady-state* probabilities of the Markov chain.

Consider now the unconditional state probabilities after  $m$  steps, as given by  $\pi^{(m)} = \pi^{(0)}\mathbf{P}^{(m)}$ ; that is,

$$\pi_j^{(m)} = \sum_i \pi_i^{(0)} p_{ij}^{(m)}.$$

Then

$$\begin{aligned} \lim_{m \rightarrow \infty} \pi_j^{(m)} &= \lim_{m \rightarrow \infty} \sum_i \pi_i^{(0)} p_{ij}^{(m)} = \sum_i \pi_i^{(0)} \lim_{m \rightarrow \infty} p_{ij}^{(m)} \\ &= \sum_i \pi_i^{(0)} \pi_j = \pi_j \sum_i \pi_i^{(0)} = \pi_j, \end{aligned}$$

and hence  $\pi_j^{(m)}$  goes to the same limit as  $p_{ij}^{(m)}$  and is independent of starting-state probabilities and the time parameter  $m$ . When these unconditional limiting probabilities exist, they can be found as follows. From (1.22),

$$\lim_{m \rightarrow \infty} \pi^{(m)} = \lim_{m \rightarrow \infty} \pi^{(m-1)}\mathbf{P}.$$

Letting  $\pi = (\pi_0, \pi_1, \dots)$  represent the limiting probability vector, we have

$$\lim_{m \rightarrow \infty} \pi^{(m)} = \lim_{m \rightarrow \infty} \pi^{(m-1)} = \pi,$$

so that

$$\boxed{\pi = \pi\mathbf{P}, \quad \text{or} \quad \mathbf{0} = \pi\mathbf{Q}.} \tag{1.32}$$

From this, together with the *boundary condition* that  $\sum_j \pi_j = 1$ , we can obtain the  $\{\pi_j\}$ . These well-known equations are called the *stationary equations* of the Markov chain, and their solution is called the *stationary distribution*. The equations (1.32) will play a major role in the solution to some of the more advanced queueing models treated in Chapter 5. Note that the boundary condition may be written in vector notation as  $\pi e = 1$ , for  $e$  a column vector with all elements equal to one.

It is possible in some cases to get solutions to (1.32) even when no limiting distribution exists. Thus when a limiting distribution exists, this implies a solution to (1.32), and the resulting stationary distribution is the limiting distribution. But the converse is not true; that is, a solution to (1.32) does not imply the existence of a limiting distribution. We illustrate this with the following examples.

### ■ EXAMPLE 1.1

Consider a degenerate discrete-parameter stochastic process that sequentially alternates between two states,  $-1$  and  $+1$ . If we call  $-1$  state 0 and  $+1$  state 1, the transition probability matrix is

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

We can solve (1.32) and get the stationary distribution vector by

$$(\pi_0, \pi_1) = (\pi_0, \pi_1) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

which gives  $\pi_0 = \pi_1$ . Using the fact that  $\pi_0 + \pi_1 = 1$  (for a valid probability distribution), we have the stationary probability distribution  $\pi_0 = \pi_1 = \frac{1}{2}$ . But is this also the limiting distribution? Intuitively we would think not, since the process keeps alternating forever and at any time in the future, the probability that it will be found in a particular state (say,  $-1$ ) is either one or zero, depending on the particular value of  $m$  chosen. This is easily verified by successive multiplication of the matrix  $P$ , yielding

$$P^{(m)} = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & m \text{ even,} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, & m \text{ odd.} \end{cases}$$

Thus even though a solution to  $\pi = \pi P$  exists, there is no  $\lim_{m \rightarrow \infty} p_{ij}^{(m)}$ . Furthermore,  $\lim_{m \rightarrow \infty} p_{ij}^{(m)}$  does not exist unless  $\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$ , that is, unless the initial probability distribution is set equal to the stationary distribution. For this  $\pi^{(0)}$  (and only this  $\pi^{(0)}$ ),

$$\pi^{(1)} = (\frac{1}{2}, \frac{1}{2}) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = (\frac{1}{2}, \frac{1}{2}),$$

and hence

$$\pi^{(m)} =: \left(\frac{1}{2}, \frac{1}{2}\right) \quad \text{for all } m.$$

In the previous example, what we have done by using  $\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$ , the stationary solution, is to make this stochastic process strictly stationary. A strictly stationary stochastic process is defined as follows: For all  $k$  and  $h$ , the joint probability distribution of  $X(t_1), X(t_2), \dots, X(t_k)$  (called a *finite-dimensional distribution* of order  $k$ ) is equal to the joint probability distribution of  $X(t_1 + h), X(t_2 + h), \dots, X(t_k + h)$ . Note that stationary processes therefore possess time-independent distribution functions. The solution to  $\pi = \pi P$  does not imply strict stationarity, but strict stationarity does imply that  $\pi^{(m)}$  is time-independent. The process in Example 1.1 is not in general strictly stationary, but can be made so by using an initial probability vector equal to the stationary vector  $(\frac{1}{2}, \frac{1}{2})$ . While strict stationarity guarantees the time independence of  $\pi^{(m)}$ , the process still does not possess a steady state, since  $p_{ij}^{(m)}$  never becomes independent of the starting state  $i$  or the time parameter  $m$ .

■ **EXAMPLE 1.2**

Consider now a discrete-parameter Markov chain similar to the one of Example 1.1 but with two possible states, 0 and 1, and transition probability matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Note that at each transition, there is a probability  $\frac{1}{2}$  of staying in the current state and a probability  $\frac{1}{2}$  that the chain will change state. Again, it is easy to get the stationary distribution by  $\pi = \pi P$  and  $\pi e = 1$ , which yields  $\pi_0 = \pi_1 = \frac{1}{2}$ .

Now let us see if a limiting distribution exists. Successive multiplication of  $P$  yields

$$P^{(m)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad \text{for all } m.$$

Thus  $\pi = (\frac{1}{2}, \frac{1}{2})$  is the limiting (steady-state) distribution vector, namely,

$$\begin{aligned} \lim_{m \rightarrow \infty} p_{i0}^{(m)} &= \frac{1}{2} & (i = 0, 1), \\ \lim_{m \rightarrow \infty} p_{i1}^{(m)} &= \frac{1}{2} & (i = 0, 1). \end{aligned}$$

This process is also not, in general, stationary unless  $\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$ , but it does, as we have shown, possess a steady state, since  $\lim_{m \rightarrow \infty} p_{ij}^{(m)} = \pi_j$ , regardless of which  $\pi^{(0)}$  is used.

Let us complicate the example slightly by changing  $P$  to

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}.$$

The stationary solution is still  $\pi = (\frac{1}{2}, \frac{1}{2})$ , which is easy to verify by solving  $\pi = \pi P$  and  $\pi e = 1$ . We must now see whether  $P^{(m)}$  goes to a limit; that is, whether  $\lim_{m \rightarrow \infty} p_{ij}^{(m)}$  exists. Successive multiplication of  $P$  here by itself yields

$$P \cdot P = P^2 = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{5}{9} & \frac{4}{9} \\ \frac{4}{9} & \frac{5}{9} \end{pmatrix}.$$

Now

$$P \cdot P^2 = P^3 = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{5}{9} & \frac{4}{9} \\ \frac{4}{9} & \frac{5}{9} \end{pmatrix} = \begin{pmatrix} \frac{13}{27} & \frac{14}{27} \\ \frac{14}{27} & \frac{13}{27} \end{pmatrix},$$

and continuing on with this procedure, we see that  $P^{(m)}$  converges to a matrix all of whose entries are  $\frac{1}{2}$ . Thus the stationary distribution is indeed the steady-state distribution. Once again, we can make the process completely stationary by letting  $\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$ , since then  $\pi^{(0)}P = (\frac{1}{2}, \frac{1}{2})$  and  $\pi^{(m)} = (\frac{1}{2}, \frac{1}{2})$  for all  $m$ . If we were to use any other  $\pi^{(0)}$ , we would find that  $\pi^{(m)}$  equals  $(\frac{1}{2}, \frac{1}{2})$  only in the limit.

For continuous-parameter processes, the stationary solution can be obtained from

$$0 = pQ, \tag{1.33}$$

where  $Q$  is the intensity matrix as defined in (1.29). The same concepts of stationarity and steady state apply for the continuous-parameter case, with  $t$  replacing  $m$  in the limiting process. Direct determination of the existence of a steady-state solution is more difficult here, since it would involve obtaining the solution to the differential equations of (1.29) and then taking the  $\lim_{t \rightarrow \infty} p'(t)$ . If it exists, it will, of course, equal the  $p$  obtained from (1.33). This is a considerably more difficult task than successive multiplication of the transition probability matrix  $P$  required for the discrete-parameter case.

At the end of this section, we shall present some theorems that will allow us to determine the existence of a limiting distribution. Once we know a limiting distribution exists, it can be obtained from  $\pi = \pi P$  or  $0 = pQ$ , and the respective boundary conditions  $\pi e = 1$  or  $pe = 1$ .

### 1.9.6 Ergodicity

Closely associated with the concepts of limiting and stationary distributions is the idea of *ergodicity*, which has to do with the information contained in one infinitely long sample path of a process (e.g., see Papoulis, 1991). Ergodicity is important in that it deals with the problem of determining measures of a stochastic process  $X(t)$  from a single realization, as is often done in analyzing simulation output. That is  $X(t)$  is ergodic in the most general sense if, with probability one, all its "measures" can be determined or well approximated from a single realization,  $x_0(t)$ , of the process. Since statistical measures of the process are usually expressed as time averages, this

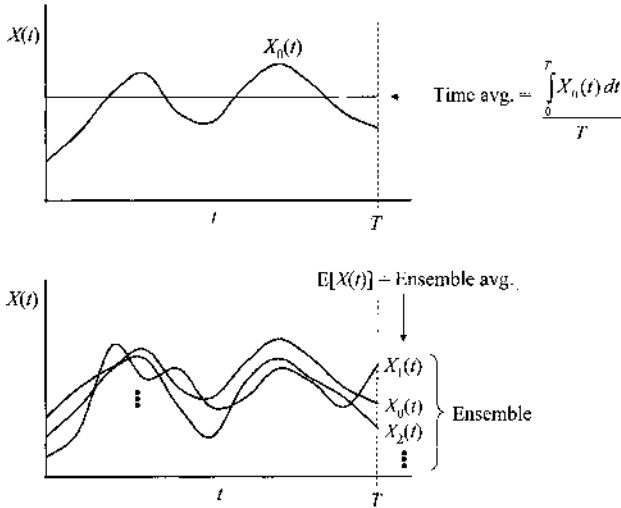


Figure 1.7 Ergodicity.

is often stated as follows:  $X(t)$  is ergodic if time averages equal ensemble averages, that is, expected values (see Figure 1.7). One may not always be interested in all, but sometimes only in certain measures (or moments) of a process. We can then define ergodicity with respect to these moments, and a process might thus be ergodic for certain moments, but not for others. However, in queueing theory, we are typically interested in fully ergodic processes, that is, processes that are ergodic with respect to all moments.

When dealing with stationary processes, statistical averages are independent of time; hence concern with respect to ergodicity centers only on convergence of time averages (e.g., see Karlin and Taylor, 1975, pp. 474 ff., or Heyman and Sobel, 1982, pp. 366 ff.). In queueing theory, the processes of interest are in general not stationary, and thus our interest in ergodicity involves the convergence of both time and ensemble averages.

Mathematically now, the time average of the square of a realization of a process, for example, would be written as

$$\overline{x_T^2} = \frac{1}{T} \int_0^T [x_0(t)]^2 dt. \tag{1.34}$$

This is then the second moment of the sample function  $x_0(t)$ . The ensemble average at time  $t$  would be denoted by

$$E\{[X(t)]^2\} = \lim_{n \rightarrow \infty} \frac{\sum_{i=0}^n \{x_i(t)\}^2}{n} = m_2(t) \tag{1.35}$$

and is the population second moment. To then say that the process is ergodic with respect to its second moment is equivalent to requiring that

$$\lim_{T \rightarrow \infty} \overline{x_T^2} = \lim_{t \rightarrow \infty} m_2(t) = \overline{x^2} < \infty. \quad (1.36)$$

We thus say that a process is ergodic if this property holds for all its moments (ergodic in distribution function), and thus that each moment possesses a limit (becomes free of "time"). Note that ergodicity means that moments become time-independent, but not necessarily independent of the initial state. This independence of the process from its actual starting state (existence of a limiting distribution) is a somewhat stronger condition than ergodicity, though the two often go together. Clearly, processes need not be ergodic with respect to any specific number of moments, and the conditions under which any moment is ergodic are the subject of a large class of theorems commonly referred to as ergodic theorems, some of which will be presented later in this section.

We now go back to our previous examples to illustrate the concept of ergodicity. We shall show that the process of Example 1.1 is not ergodic by showing that the relationship (1.36) does not hold for the first moment (this is all that is required to prove nonergodicity, since for a process to be ergodic it must be ergodic in *all* its moments).

Consider the equivalent of (1.34) for the first moment of our alternating process. The average value of the process after  $m - 1$  transitions [assume it starts in state +1, i.e.,  $\pi^{(0)} = (1, 0)$ ] is

$$\begin{aligned} \overline{x_m} &= \frac{1}{m} \sum_{i=0}^{m-1} x_i \\ &= \begin{cases} 0 & (m \text{ even}), \\ 1/m & (m \text{ odd}). \end{cases} \end{aligned}$$

The limit of  $\overline{x_m}$  as  $m$  goes to infinity is clearly zero.

Since each path must be identical, the ensemble average after  $m$  transitions is

$$m_1(m) = E[X_m] = \begin{cases} +1 & (m \text{ even}), \\ -1 & (m \text{ odd}). \end{cases}$$

The limit of the ensemble average as  $m$  goes to infinity does not equal zero; in fact, it does not even exist. Hence the process is not ergodic in its first moment and is thus not ergodic in any way.

If we use  $\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$ , however, then the limit of  $m_1(m)$  as  $m$  goes to infinity is also zero and the process is indeed ergodic even though, as we have previously shown, no limiting distribution exists. We note that stationary processes [e.g., our problem when  $\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$ ] are ergodic. But we shall see shortly that ergodic processes need not be stationary.

Now let us see if the process of Example 1.2 is ergodic. We first consider the time average of the  $k$ th power of the process. We need to calculate

$$\lim_{m \rightarrow \infty} \overline{x_m^k} = \lim_{m \rightarrow \infty} \frac{\sum_{i=0}^{m-1} x_i^k}{m}.$$

Now  $x_i^k$ , the value of the  $k$ th power of the process at the  $i$ th transition, is an observation from a random variable taking on values 0 or 1, each with probability  $\frac{1}{2}$ . Thus, ignoring the starting state (it washes out in the limit, since  $x_0^k$  is either zero or one and  $\lim_{m \rightarrow \infty} x/m = 0$ ), we have that  $\lim_{m \rightarrow \infty} \overline{x_m^k}$  is the limit of the average value of  $m - 1$  IID Bernoulli random variables with parameter  $\frac{1}{2}$ . By the strong law of large numbers, this quantity goes to the Bernoulli mean of  $\frac{1}{2}$ .

Let us now consider the ensemble average at time  $m$ ,  $E[X_m^k] = m_k(m)$ , which is the  $k$ th population moment of a Bernoulli random variable with parameter  $\frac{1}{2}$  and can be calculated as

$$E[X_m^k] = \sum_{i=0}^1 x_i^k p_i = 0^k \left(\frac{1}{2}\right) + 1^k \left(\frac{1}{2}\right) = \frac{1}{2} \quad (k = 1, 2, 3, \dots).$$

The limit of  $E[X_m^k]$  as  $m$  goes to infinity is also  $\frac{1}{2}$ ; hence the process is ergodic. Note that ergodicity of this process holds regardless of  $\pi^{(0)}$ , so that even though in general the process is not stationary, it is nevertheless ergodic.

It is generally not easy to show ergodicity or nonergodicity by direct methods, as in the foregoing examples, and the theorems to follow will aid in this task. Prior to stating the key theorems that will enable us to determine when a solution to the stationary equation exists, if a limiting distribution exists, and whether or not the process is ergodic, we first present some definitions required for characterizing discrete-parameter Markov chains.

Two states,  $i$  and  $j$ , are said to *communicate* ( $i \leftrightarrow j$ ) if  $i$  is accessible from  $j$  ( $j \rightarrow i$ ) and  $j$  is accessible from  $i$  ( $i \rightarrow j$ ). A chain is said to be *irreducible* if all of its states communicate, that is, if there exists an  $n$  such that  $p_{ij}^{(n)} > 0$  for all pairs  $(i, j)$ .

The period of a return state  $k$  of a chain is defined as the greatest common divisor (GCD) of the set of integers  $\{n\}$  for which  $p_{kk}^{(n)} > 0$ . A state is said to be aperiodic if this GCD is 1, that is, if it has period 1. A chain is said to be aperiodic if each of its states is aperiodic. For example, if a process can return to a certain state in 5 or 7 time units, this state is aperiodic, since possible return times are 5, 7, 10, 12, 14, 15, 17, 19, 20, 21, 22, 24, 25, 26, 27, . . . , and this sequence has greatest divisor equal to one. On the other hand, if the return times are 3 and 6, then we get a sequence of possible returns of 3, 6, 9, 12, . . . , meaning that the state is periodic with period equal to three.

Define  $f_{jj}^{(n)}$  as the probability that a chain starting at state  $j$  returns for the first time to  $j$  in  $n$  transitions. Hence the probability that the chain ever returns to  $j$  is

$$f_{jj} = \sum_{n=1}^{\infty} f_{jj}^{(n)}.$$

If  $f_{jj} = 1$ , then  $j$  is said to be a *recurrent state*; if  $f_{jj} < 1$ , then  $j$  is a *transient state*. When  $f_{jj} = 1$ ,

$$m_{jj} = \sum_{n=1}^{\infty} n f_{jj}^{(n)}$$

is the *mean recurrence time*. If  $m_{jj} < \infty$ , then  $j$  is known as a *positive recurrent state*, while if  $m_{jj} = \infty$ , then  $j$  is a *null recurrent state*.

To illustrate, consider the following transition probability matrix, with six states numbered  $0 \rightarrow 5$ :

$$\begin{pmatrix} \frac{1}{6} & 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & 0 & \frac{2}{5} & \frac{1}{5} & 0 & \frac{1}{5} \\ \frac{1}{4} & 0 & 0 & \frac{3}{4} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

Now state 1 is recurrent, since if the process starts there, it stays there forever. State 5 is transient (can visit a few times and then never return), as is state 2. States 0, 3, 4 are recurrent—once one is entered the process moves among these forever. These states form a recurrent *class*.

As another example, consider the imbedded birth-death transition probability matrix given by (1.31). We see that it is irreducible, may be recurrent (e.g., if  $\mu_i \geq \lambda_i$  for all  $i \geq$  some value  $n$ ), and has period two.

Define  $f_{ij}^{(n)}$ ,  $i \neq j$ , as the probability that the first passage from state  $i$  to state  $j$  occurs in exactly  $n$  steps. Then the probability that state  $j$  is ever reached from  $i$  is

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}.$$

The expected value of the sequence  $\{f_{ij}^{(n)}, n = 1, 2, \dots\}$  of first-passage probabilities for a fixed pair  $(i, j)$ ,  $i \neq j$ , is denoted by  $m_{ij}$  and is called the *mean first-passage time*; that is,

$$m_{ij} = \sum_{n=1}^{\infty} n f_{ij}^{(n)} \quad (i \neq j).$$

There are an extensive number of theorems in the literature which permit one to determine the presence of recurrence in a Markov chain and to calculate the mean recurrence time whenever appropriate (e.g., see Çinlar, 1975). The following theorems, which we state without proofs, tie much of the theory together and relate the concepts of ergodicity, limiting probabilities, and stationary probabilities.

### Theorem 1.1.

- (a) *In an irreducible and positive recurrent discrete-parameter Markov chain, a nondegenerate solution to the stationary equations*

$$\pi = \pi P, \quad \pi e = 1$$

always exists, where the solution vector  $\pi = \{\pi_j\}$  is such that  $\pi_j = 1/m_{jj}$ .

Furthermore, when all moments of this stationary distribution are finite, we have the following:

- (b) If the starting probability vector  $\pi^{(0)}$  is set equal to the stationary probability vector  $\pi$ , the above chain becomes a stationary stochastic process and hence ergodic.
- (c) If the above chain is aperiodic as well as irreducible and positive recurrent, the process is ergodic and has a limiting probability distribution equal to the stationary distribution.

Note that existence of a limiting distribution is the strongest condition, ergodicity is somewhat weaker, and a nondegenerate solution to the stationary equations is the weakest of the three conditions.

**Theorem 1.2.** *An irreducible, aperiodic chain is positive recurrent if there exists a nonnegative solution of the system*

$$\sum_{j=0}^{\infty} p_{ij}x_j \leq x_i - 1 \quad (i \neq 0)$$

such that

$$\sum_{j=0}^{\infty} p_{0j}x_j < \infty.$$

**Theorem 1.3.** *For a continuous-parameter Markov chain, the imbedded chain (at points of transition) need not be aperiodic as long as the mean holding times in all states are bounded, for Theorem 1.1 to be valid.*

The reason aperiodicity is not required for the imbedded Markov chain is that since the process is continuous in time, the time between transitions varies continuously, and as long as it is bounded, the transition times “wash out” any periodicity that may come from the imbedded process.

Returning to the birth–death process and considering its imbedded Markov chain, we see that the matrix  $P$  as given by (1.31) is irreducible and positive recurrent as long as certain conditions hold (e.g.,  $\mu_i > \lambda_i$  for all  $i \geq n$ ). Furthermore, if the mean state holding times,  $1/(\lambda_i + \mu_i)$ , are bounded ( $\lambda_i$  and/or  $\mu_i \geq \epsilon > 0$  for all  $i$ ), then Theorem 1.3 is satisfied and the process is ergodic and possesses a limiting (steady-state) probability distribution. Such processes are sometimes also referred to as achieving *statistical equilibrium* (see Cooper, 1981).

When dealing with Markov chains, some authors (e.g., Feller, 1957, pp. 353ff.; Heyman and Sobel, 1982, pp. 230ff.) define a positive recurrent aperiodic state as ergodic, and thus the necessary conditions for Theorem 1.1 to hold, that is, irreducibility and positive recurrence, can be stated as requiring all states of a Markov chain to be ergodic. While it is certainly true that such chains are ergodic as defined

Table 1.6 Long-Run Behavior Concepts

Concept	Definition	Relationship/Interpretation
Stationary distribution (SD)	Solution to $\pi = \pi P, \pi e = 1$ ( $\mathbf{0} = pQ, pe = 1$ ).	If all states communicate, $\pi_j$ can be interpreted as the long-run percentage of time the process spends in state $j$ .
Stationary stochastic process (SSP)	State distributions invariant over time.	If all states communicate, a solution to $\pi = \pi P$ exists and the process is ergodic.
Ergodic process (EP)	Time averages equal ensemble averages in limit.	A solution to $\pi = \pi P$ exists. EPs become SSPs in limit as time goes to infinity.
Limiting or steady-state or statistical equilibrium distribution	Limiting state probability distribution independent of time and initial state.	Process is also ergodic. Stationarity and independence of starting state achieved in limit. The limiting distribution is the SD.

by (1.36) and the text immediately following it, we believe it is somewhat confusing to define ergodicity as synonymous with the existence of limiting distributions, since it is a somewhat weaker condition, as we have previously mentioned.

To summarize the long-run behavior concepts of Markov chains studied in this section, we recap the preceding definitions in Table 1.6.

## 1.10 Introduction to the QtsPlus Software

Today, spreadsheets are an indispensable tool for engineers and operations research specialists. Several papers have discussed the application of spreadsheets in the various operations research disciplines, such as optimization and queuing theory (Bodily, 1986; Leon et al., 1996; Grossman, 1999). To facilitate learning queueing theory, a collection of spreadsheet queueing models, collectively known as *QtsPlus*, is available with this textbook. We now introduce QtsPlus and some of its capabilities. See Appendix E for instructions to install and run the software.

We illustrate how to use QtsPlus to solve Example 1.2. Follow the instructions in Appendix E to start the software. Once it is active, select the **Basic** model category from the drop-down list provided, then select the **Discrete-Time Markov Chain** model from the list of models available under the **Basic** category. Once the

## MARKOV CHAIN

To start a new problem, enter number of states and press button.

Number of States:

Enter transition probabilities, then press "Solve" button.

Solve

Enter P Matrix below

0.333333	0.666667
0.666667	0.333333

Transition probability matrix  
 $\begin{pmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{pmatrix}$

Answer

Solution

p0	0.5
p1	0.5

Figure 1.8 QtsPlus solution to Example 1.2.

workbook, called marchain.xls, is open, enter 2 into the input field: **Number of States**. If a pop-up message box appears asking,

This will cause existing model parameters to be discarded. Do you wish to continue?

Press the **Yes** button to set up a new **P** matrix. Using Excel formulas, fill the respective cells of the **P** matrix on the worksheet with the initial parameters as shown below.

$$\begin{aligned} &= 1/3 &= 2/3 \\ &= 2/3 &= 1/3 \end{aligned}$$

Press the **Solve** button. The answer appears on the left side of the worksheet (Figure 1.8) and coincides with the stationary solution  $\pi = (\frac{1}{2}, \frac{1}{2})$  previously calculated.

## PROBLEMS

- 1.1. Discuss the following queuing situations in terms of the characteristics given in Section 1.2.
- Aircraft landing at an airport.
  - Supermarket checkout procedures.
  - Post-office or bank customer windows.
  - Toll booths on a bridge or highway.
  - Gasoline station with several pump islands.
  - Automatic car wash facility.
  - Telephone calls coming into a customer information system.
  - Appointment patients coming into a doctor's office.
  - Tourists wishing a guided tour of the White House.
  - Electronic components on an assembly line consisting of three operations and one inspection at end of line.
  - Processing of programs coming from a number of independent sources on a local area network into a central computer.

- 1.2. Give three examples of a queuing situation other than those in Problem 1.1, and discuss them in terms of the basic characteristics of Section 1.2.
- 1.3. The Carry Out Curry House, a fast-food Indian restaurant, must decide on how many parallel service channels to provide. They estimate that, during the rush hour, the average number of arrivals per hour will be approximately 40. They also estimate that, on average, a server will take about 5.5 min to serve a typical customer. Using only this information, how many service channels will you recommend they install?
- 1.4. Fluffy Air, a small local feeder airline, needs to know how many slots to provide for telephone callers to be placed on hold. They plan to have enough answerers so that the average waiting time on hold for a caller will be 75 seconds during the busiest period of the day. They estimate the average call-in rate to be 3 per minute. What would you advise?
- 1.5. The Outfront BBQ Rib Haven does carry out only. During peak periods, two servers are on duty. The owner notices that during these periods, the servers are almost never idle. She estimates the percent time idle of each server to be 1 percent. Ideally, the percent idle time would be 10 percent to allow time for important breaks).
- (a) If the owner decides to add a third server during these times, how much idle time would each server have then?
- (b) Suppose that by adding the third server, the pressure on the servers is reduced, so they can work more carefully, but their service output rate is reduced by 20 percent. What now is the percent time each would be idle?
- (c) Suppose, instead, the owner decides to hire an aid (at a much lower salary) who serves as a gofer for the two servers, rather than hiring another full server. This allows the two servers to decrease their average service time by 20 percent (relative to the original service rate). What now is the percent idle time of each of the two servers?
- 1.6. The Happidaiz frozen-yogurt stand does a thriving business on warm summer evenings. Even so, there is only a single person on duty at all times. It is known that the service time (dishing out the yogurt and collecting the money) is normally distributed with mean 2.5 min and standard deviation 0.5 min. (Although the normal distribution allows for negative values, the standard deviation with respect to the mean is small so that negative values are more than 4 standard deviations below the mean and the probability of negative values is essentially zero.) You arrive on a particular evening to get your favorite crunchy chocolate yogurt cone and find 8 people ahead of you. Estimate the average time until you get the first lick. What is the probability that you will have to wait more than  $\frac{1}{2}$  hour? [*Hint:* Remember that the sum of normal random variables is itself normally distributed.]

- 1.7. A certain football league consists of 32 teams. Each team has 67 active players. There is a draft each year for teams to acquire new players. Each team acquires 7 new players per year in the draft. The number of active players on each team must always be 67. Thus each team must cut some existing players each year to make room for the new players.
- (a) Assuming that a football player can only join a team by being selected in the draft, estimate the average career length of a football player in the league.
- (b) Now suppose that a player can join a team in one of two ways: (1) by being selected in the draft, as before, or (2) by signing directly with a team outside the draft. Suppose further that the average career length of a football player is known to be 3.5 years. Under the same assumptions as before, estimate the average number of players who enter the league each year *without being drafted*.
- 1.8. The following table gives enrollment statistics for undergraduates at a university. From this data, estimate the average length of time that an undergraduate is enrolled at the university (this average should include students who graduate as well as students who transfer or drop out).

Year	New Students		Total Enrollment
	First-yr Students	Transfer Students	
1	1,820	2,050	16,800
2	1,950	2,280	16,700
3	1,770	2,220	17,100
4	1,860	2,140	16,400
5	1,920	2,250	17,000

- 1.9. You are selling your home. You observe that at any given time there are typically about 50 homes for sale in your area. New homes go on the market at a rate of about 5 per week. About how long will it take to sell your home? What assumptions are made to arrive at your answer?
- 1.10. Table 1.7 gives observations regarding customers at a single-server FCFS queue.
- (a) Compute the average time in the queue and the average time in the system.
- (b) Calculate the average system waiting time of those customers who had to wait for service (i.e., exclude those who were immediately taken into service). Calculate the average length of the queue, the average number in the system, and the fraction of idle time of the server.
- 1.11. Items arrive at an initially unoccupied inspection station at a uniform rate of one every 5 min. With the time of the first arrival set equal to 5, the chronological times for inspection completion of the first 10 items were observed to be 7, 17, 23, 29, 35, 38, 39, 44, 46, and 60, respectively. By

Table 1.7 Data for Problem 1.10

Customer	Interarrival Time	Service Time
1	1	3
2	9	7
3	6	9
4	4	9
5	7	10
6	9	4
7	5	8
8	8	5
9	4	5
10	10	3
11	6	6
12	12	3
13	6	5
14	8	4
15	9	9
16	5	9
17	7	8
18	8	6
19	8	8
20	7	3

manual simulation of the operation for 60 min, using these data, develop sample results for the mean number in system and the percentage idle time experienced.

- 1.12.** Table 1.8 lists the arrival times and service durations for customers in a FCFS single-server queue. From this data, compute  $L_q$  (the time-average number in queue) and  $L_q^{(A)}$  (the average number in queue as seen by arriving customers). For  $L_q$ , use a time horizon of  $[0, 15.27]$ , where 15.27 is the time that the last customer exits the system. Assume the system is empty at  $t = 0$ .
- 1.13.** Derive (1.13) of Section 1.7 by the sequential use of (1.12); then employ mathematical induction to prove (1.14).
- 1.14.** Given the probability function found for the Poisson process in (1.14), find its moment generating function,  $M_{N(t)}(\theta)$ , that is, the expected value of  $e^{\theta N(t)}$ . Then use this MGF to show that the mean and variance both equal  $\lambda t$ .

Table 1.8 Data for Problem 1.12

Arrival Time (min)	Service Duration (min)
1	2.22
2	1.76
3	2.13
4	0.14
5	0.76
6	0.70
7	0.47
8	0.22
9	0.18
10	2.41
11	0.41
12	0.46
13	1.37
14	0.27
15	0.27

- 1.15.** Derive the Poisson process by using the third assumption that the numbers of arrivals in nonoverlapping intervals are statistically independent and then applying the binomial distribution.
- 1.16.** Prove that the Poisson process has stationary increments.
- 1.17.** By the use of the arguments of Sections 1.7 and 1.8, find the distribution of the counting process associated with IID Erlang interoccurrence times.
- 1.18.** Assume that arrivals can occur singly or in batches of two, with the batch size following the probability distribution

$$f(1) = p, f(2) = 1 - p \quad (0 < p < 1)$$

and with the time between successive batches following the exponential probability distribution

$$a(t) = \lambda e^{-\lambda t} \quad (t > 0).$$

Show that the probability distribution for the number of arrivals in time  $t$  is the compound Poisson distribution given by

$$p_n(t) = e^{-\lambda t} \sum_{k=0}^{\lfloor n/2 \rfloor} \frac{p^{n-2k} (1-p)^k (\lambda t)^{n-k}}{(n-2k)! k!}.$$

where  $\lfloor n/2 \rfloor$  is the greatest integer  $\leq n/2$ .

- 1.19. (a) You are given two Poisson processes with intensities  $\lambda_1$  and  $\lambda_2$ . Find the probability that there is an occurrence of the first stream before the second, starting at time  $t = 0$ .
- (b) A queueing system is being observed. We see that all  $m$  identical, exponential servers are busy, with  $n$  more customers waiting, and decide to shut off the arrival stream. On average, how long will it take for the system to empty completely?
- 1.20. Verify (1.28a) and (1.28b) by using (1.27b) in (1.25). [Hint: To obtain (1.28a), let  $s = t + \Delta t$ . What is required to obtain (1.28b)?]
- 1.21. Derive the Poisson equations (1.11) and (1.12) by considering the Poisson process as a pure birth process and by using the Kolmogorov forward equation (1.29) with the proper  $q_i$  and  $q_{ij}$ .
- 1.22. (a) Compute the stationary probability distribution for Examples 1.1 and 1.2.  
 (b) Compute the stationary probability distribution for a Markov chain with the following single-step transition probability matrix:

$$\begin{pmatrix} .25 & .20 & .12 & .43 \\ .25 & .20 & .12 & .43 \\ 0 & .25 & .20 & .55 \\ 0 & 0 & .25 & .75 \end{pmatrix}$$

- 1.23. Derive the imbedded transition probabilities for the birth-death process as given by (1.31). [Hint: Use results of Problem 1.19(a) or the law of conditional probability:  $\Pr\{A|B, C\} = \Pr\{A, B|C\}/\Pr\{B|C\}$ .]
- 1.24. Consider the first-order statistic (call it  $T'_{(1)}$ ) of a random sample of size  $n$  drawn from a uniform (0, 1) population. Its CDF is found to be

$$F(t) = 1 - (1 - t)^n.$$

Show that the random variable  $nT'_{(1)}$  converges in law to an exponential as  $n \rightarrow \infty$ .

- 1.25. Suppose you have learned that an  $M/G/1/K$  has a blocking probability of  $p_k = .1$ , that  $\lambda = \mu = 1$ , and that  $L = 5$ . Find  $\lambda_{\text{eff}}$ ,  $W$ ,  $W_q$ ,  $p_0$ , and  $\rho_{\text{eff}}$ .
- 1.26. In choosing the proper distributions to represent interarrival and service times, the *coefficient of variation* (CV) can often be useful. The CV is defined as the ratio of the standard deviation to the mean and provides a measure of the relative spread of a distribution. For example, service consisting of routine tasks should have a relatively small spread around the mean ( $\text{CV} \leq 1$ ), whereas service consisting of diverse tasks (some quick, some time consuming) should have a relatively large spread around the mean ( $\text{CV} \geq 1$ ). The exponential distribution, widely used in queue modeling,

has  $CV = 1$  (standard deviation = mean). Two other distributions often employed in queueing are the *Erlang* distribution and the *mixed-exponential* distribution (the *hyperexponential* distribution is a special case). The Erlang is a two-parameter distribution, having a type or shape parameter  $k$  (an integer  $\geq 1$ ) and a scale parameter, which we shall denote by  $\beta$ . The mean of the Erlang is the product  $k\beta$ , and its standard deviation is the product  $\beta\sqrt{k}$ . The CV for an Erlang is then  $1/\sqrt{k} \leq 1$ . When  $k = 1$ , the Erlang reduces to the exponential distribution. The mixed-exponential distribution function is a convex linear combination of exponential distributions, mixed according to some probability distribution (e.g., we select from one exponential population, mean  $\mu_1$ , with probability  $p$ , and from a second exponential, mean  $\mu_2$ , with probability  $1 - p$ ). The CV for a mixed-exponential distribution can be shown to be always  $> 1$ . Using the software, solve the following problems:

- (a) Data taken on a server who provides espresso to customers at the Betterbean Boutique show that the mean time to serve a customer is 2.25 min with a standard deviation of 1.6 min. What is the probability that service takes more than 5 min? [*Hint*: Find the closest integer value for  $k$  and then solve for  $\beta$ .]
- (b) Data collected at a small post office in the rural town of Arlingrock reveal that the clerk has two types of customers—those who desire to purchase stamps only and those who require other more complicated functions. The distributions of service times for each type of customer can be well approximated by the exponential distribution; the stamp-only customers take on average 1.06 min, while the nonstamp customers take on average 3.8 min. What is the probability that a stamp customer requires more than 5 min? What is the probability that a nonstamp customer takes more than 5 min? If 15% of all arrivals are stamp-only customers, what is the probability that the next customer in line requires more than 5 min?

