CHAPTER

11

Markov Chains

In general, the random variables within the family defining a stochastic process are not independent, and in fact can be statistically dependent in very complex ways. In this chapter we introduce the class of Markov random processes that have a simple form of dependence and that are quite useful in modeling many problems found in practice. We concentrate on integer-valued Markov processes, which are called Markov chains.

- Section 11.1 introduces Markov processes and the special case of Markov chains.
- Section 11.2 considers discrete-time Markov chains and examines the behavior of their state probabilities over time.
- Section 11.3 discusses structural properties of discrete-time Markov chains that determine their long-term behavior and limiting state probabilities.
- Section 11.4 introduces continuous-time Markov chains and considers the transient as well as long-term behavior of their state probabilities.
- Section 11.5 considers time-reversed Markov chains and develops interesting properties of reversible Markov chains that look the same going forwards and backwards in time.
- Finally, Section 11.6 introduces methods for simulating discrete-time and continuous-time Markov chains.

11.1 MARKOV PROCESSES

A random process $X(t)$ is a **Markov process** if the future of the process given the present is independent of the past, that is, if for arbitrary times $t_1 < t_2 < \cdots < t_k < t_{k+1}$,

$$\begin{aligned} P[X(t_{k+1}) = x_{k+1} \,|\, X(t_k) = x_k, \ldots, X(t_1) = x_1] \\ = P[X(t_{k+1}) = x_{k+1} \,|\, X(t_k) = x_k] \end{aligned} \tag{11.1}$$

if $X(t)$ is discrete-valued, and

$$\begin{aligned} P[a < X(t_{k+1}) \leq b \,|\, X(t_k) = x_k, \ldots, X(t_1) = x_1] \\ = P[a < X(t_{k+1}) \leq b \,|\, X(t_k) = x_k] \end{aligned} \tag{11.2a}$$

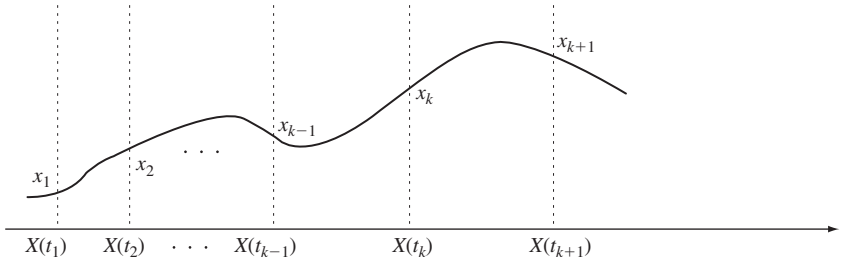

FIGURE 11.1
Markov property: Given $X(t_k)$, $X(t_{k+1})$ is independent of samples prior to t_k.

if $X(t)$ is continuous-valued. If the samples of $X(t)$ are jointly continuous, then Eq. (11.2a) is equivalent to

$$f_{X(t_{k+1})}(x_{k+1} \mid X(t_k) = x_k, \ldots, X(t_1) = x_1) = f_{X(t_{k+1})}(x_{k+1} \mid X(t_k) = x_k). \quad (11.2b)$$

We refer to Eqs. (11.1) and (11.2) as the **Markov property**. In the above expression t_k is the "present," t_{k+1} is the "future," and $t_1, \ldots, t_{k-1}$ is the "past," as shown in Fig. 11.1. Thus in Markov processes, pmf's and pdf's that are conditioned on several time instants always reduce to a pmf/pdf that is conditioned only on the most recent time instant. For this reason we refer to the value of $X(t)$ at time t as the **state** of the process at time t.

Example 11.1 Sum Process

Consider the sum process discussed in Section 9.3:

$$S_n = X_1 + X_2 + \cdots + X_n = S_{n-1} + X_n,$$

where the X_i's are an iid sequence of random variables and where $S_0 = 0$. S_n is a Markov process since

$$\begin{aligned} P[S_{n+1} = s_{n+1} \mid S_n = s_n, \ldots, S_1 = s_1] &= P[X_{n+1} = s_{n+1} - s_n] \\ &= P[S_{n+1} = s_{n+1} \mid S_n = s_n]. \end{aligned}$$

The binomial counting process and the random walk processes introduced in Section 9.3 are sum processes and therefore Markov processes.

Example 11.2 Moving Average

Consider the moving average of a Bernoulli sequence:

$$Y_n = \frac{1}{2}(X_n + X_{n-1}),$$

where the X_i are an independent Bernoulli sequence with $p = 1/2$. We now show that Y_n is not a Markov process.

The pmf of Y_n is

$$P[Y_n = 0] = P[X_n = 0, X_{n-1} = 0] = \frac{1}{4},$$

$$P\left[Y_n = \frac{1}{2}\right] = P[X_n = 0, X_{n-1} = 1] + P[X_n = 1, X_{n-1} = 0] = \frac{1}{2},$$

and

$$P[Y_n = 1] = P[X_n = 1, X_{n-1} = 1] = \frac{1}{4}.$$

Now consider the following conditional probability for two consecutive values of Y_n:

$$P\left[Y_n = 1 \,|\, Y_{n-1} = \frac{1}{2}\right] = \frac{P[Y_n = 1, Y_{n-1} = 1/2]}{P[Y_{n-1} = 1/2]}$$

$$= \frac{P[X_n = 1, X_{n-1} = 1, X_{n-2} = 0]}{1/2} = \frac{(1/2)^3}{1/2} = \frac{1}{4}.$$

Now suppose we have additional knowledge about the past:

$$P\left[Y_n = 1 \,|\, Y_{n-1} = \frac{1}{2}, Y_{n-2} = 1\right] = \frac{P[Y_n = 1, Y_{n-1} = 1/2, Y_{n-2} = 1]}{P[Y_{n-1} = 1/2, Y_{n-2} = 1]} = 0,$$

since no sequence of X_n's leads to the sequence $1, 1/2, 1$. Thus

$$P\left[Y_n = 1 \,|\, Y_{n-1} = \frac{1}{2}, Y_{n-2} = 1\right] \neq P\left[Y_n = 1 \,|\, Y_{n-1} = \frac{1}{2}\right],$$

and the process is not Markov.

Example 11.3 Poisson Process

The Poisson process is a continuous-time Markov process since

$$P[N(t_{k+1}) = j \,|\, N(t_k) = i, N(t_{k-1}) = x_{k-1}, \ldots, N(t_1) = x_1]$$

$$= P[j - i \text{ events in } t_{k+1} - t_k \text{ seconds}]$$

$$= P[N(t_{k+1}) = j \,|\, N(t_k) = i].$$

Example 11.4 Random Telegraph

The random telegraph signal of Example 9.24 is a continuous-time Markov process since

$$P[X(t_{k+1}) = a \,|\, X(t_k) = b, \ldots, X(t_1) = x_1]$$

$$= P[\text{even (odd) number of jumps in } t_{k+1}$$

$$- t_k \text{ seconds if } a = b (a \neq b)]$$

$$= P[X(t_{k+1}) = a \,|\, X(t_k) = b].$$

Example 11.5 Wiener Process

The Wiener process, from Section 9.5, is a Markov process. Since it satisfies the independent increments property (Eq. 9.52), we have that:

$$f_{X(t_{k+1})}(x_{k+1}|\,X(t_k)=x_k,\ldots,X(t_1)=x_1)=f_{X(t_{k+1}-t_k)}(x_{k+1}-x_k)$$
$$=\frac{\exp\left\{-\frac{1}{2}\left[\frac{(x_{k+1}-x_k)^2}{\alpha(t_{k+1}-t_k)}\right]\right\}}{\sqrt{(2\pi\alpha)(t_{k+1}-t_k)}}.$$

The Wiener process is Gaussian and so it provides an example of a Gaussian Markov process.

An integer-valued Markov random process is called a **Markov chain**.[1] In the remainder of this chapter we concentrate on Markov chains.

If $X(t)$ is a Markov chain, then the joint pmf for three arbitrary time instants is

$$\begin{aligned}P[X(t_3)=x_3,X(t_2)=x_2,X(t_1)=x_1]&\\ =P[X(t_3)=x_3|X(t_2)=x_2,X(t_1)=x_1]P[X(t_2)=x_2,X(t_1)=x_1]&\\ =P[X(t_3)=x_3|X(t_2)=x_2]P[X(t_2)=x_2,X(t_1)=x_1]&\\ =P[X(t_3)=x_3|X(t_2)=x_2]P[X(t_2)=x_2|X(t_1)=x_1]P[X(t_1)=x_1],&\end{aligned}$$

where we have used the definition of conditional probability and the Markov property. In general, the joint pmf for $k+1$ arbitrary time instants is

$$\begin{aligned}P[X(t_{k+1})=x_{k+1},X(t_k)=x_k,\ldots,X(t_1)=x_1]&\\ =P[X(t_{k+1})=x_{k+1}|X(t_k)=x_k]&\\ P[X(t_k)=x_k|X(t_{k-1})=x_{k-1}]\ldots P[X(t_1)=x_1]&\\ =\left\{\prod_{j=1}^{k}P[X(t_{j+1})=x_{j+1}|X(t_j)=x_j]\right\}P[X(t_1)=x_1]&\end{aligned}\qquad(11.3)$$

Thus the *joint pmf of $X(t)$ at arbitrary time instants is given by the product of the pmf of the initial time instant and the probabilities for the subsequent state transitions*. Clearly, the state transition probabilities determine the statistical behavior of a Markov chain.

11.2 DISCRETE-TIME MARKOV CHAINS

Let X_n be a discrete-time integer-valued Markov chain that starts at $n=0$ with pmf

$$p_j(0)\triangleq P[X_0=j]\qquad j=0,1,2,\ldots.\qquad(11.4)$$

[1]See Cox and Miller [6] for a discussion of continuous-valued Markov processes.

We will assume that X_n takes on values from a countable set of integers, usually $\{0, 1, 2, \dots\}$. We say that the Markov chain is finite state if X_n takes on values from a finite set.

From Eq. (11.3), the joint pmf for the first $n + 1$ values of the process is

$$\begin{aligned} &P[X_n = i_n, \dots, X_0 = i_0] \\ &\quad = P[X_n = i_n | X_{n-1} = i_{n-1}] \dots P[X_1 = i_1 | X_0 = i_0] P[X_0 = i_0]. \end{aligned} \tag{11.5}$$

Thus the joint pmf for a particular sequence is simply the product of the probability for the initial state and the probabilities for the subsequent one-step state transitions.

We will assume that the one-step state transition probabilities are fixed and do not change with time, that is,

$$P[X_{n+1} = j | X_n = i] = p_{ij} \qquad \text{for all } n. \tag{11.6}$$

X_n is said to have **homogeneous transition probabilities**. The joint pmf for $X_n, \dots, X_0$ is then given by

$$P[X_n = i_n, \dots, X_0 = i_0] = p_{i_{n-1}, i_n} \dots p_{i_0, i_1} p_{i_0}(0). \tag{11.7}$$

Thus X_n is completely specified by the *initial pmf* $p_i(0)$ and the *matrix of one-step transition probabilities* P:

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ . & . & . & \\ p_{i0} & p_{i1} & \cdots & \\ . & . & \cdots & \end{bmatrix}. \tag{11.8}$$

We will call P the **transition probability matrix**. Note that each row of P must add to one since

$$1 = \sum_j P[X_{n+1} = j | X_n = i] = \sum_j p_{ij}. \tag{11.9}$$

If the Markov chain is finite state, then the matrix P will be an $n \times n$ nonnegative square with rows that add up to 1.

Example 11.6 Two-State Markov Chain for Speech Activity

A Markov model for packet speech assumes that if the nth packet contains silence, then the probability of silence in the next packet is $1 - \alpha$ and the probability of speech activity is α. Similarly, if the nth packet contains speech activity, then the probability of speech activity in the next packet is $1 - \beta$ and the probability of silence is β.

Let X_n be the indicator function for speech activity in a packet at time n, then X_n is a two-state Markov chain with the state transition diagram shown in Fig. 11.2(a), and transition probability matrix

$$P = \begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix}. \tag{11.10}$$

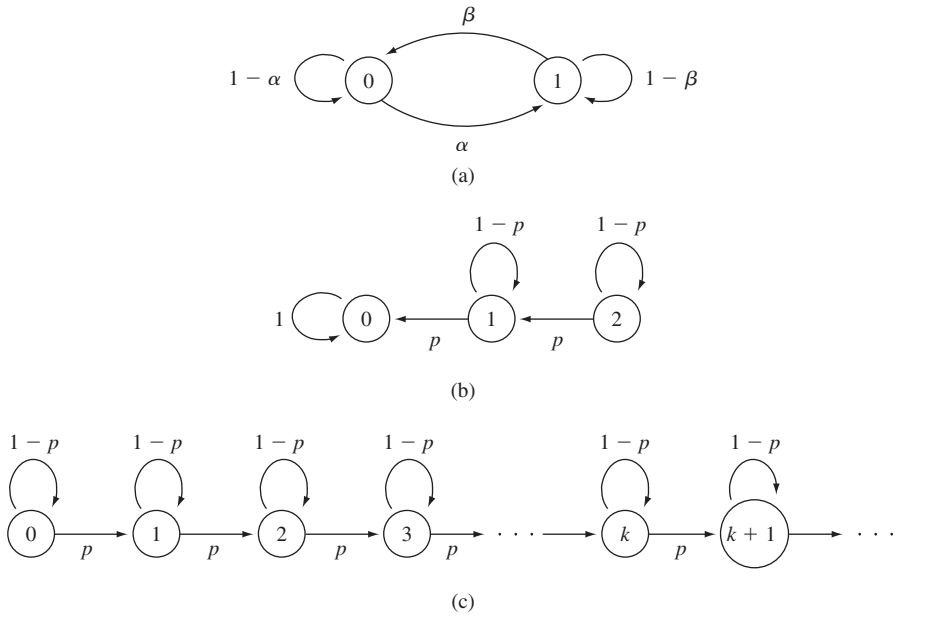

FIGURE 11.2
(a) State transition diagram for two-state Markov chain. (b) State transition diagram for Markov chain for light bulb inventory. (c) State transition diagram for binomial counting process.

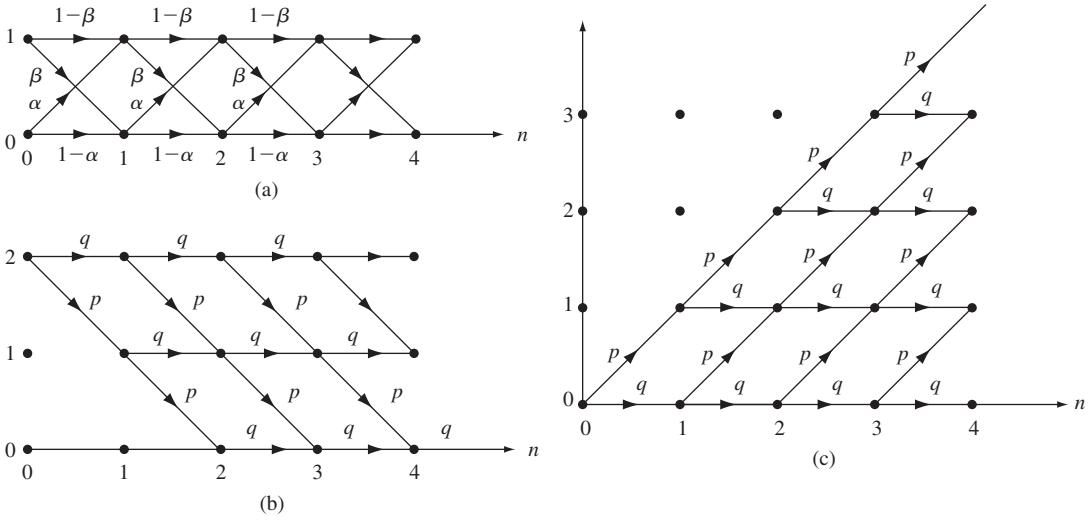

FIGURE 11.3
Trellis diagrams for Markov chain examples.

The sample functions of X_n can be viewed as traversing the **trellis diagram** in Fig. 11.3(a) which shows the possible values of the process over time. At any give time, the process occupies the "state" that corresponds to its value. The sample function is realized as the process steps from one state at a given time instant to a state in the next time instant. The transitions are determined according to the transition probability matrix.

Example 11.7

On day 0 a house has two new light bulbs in reserve. The probability that the house will need a single new light bulb during day n is p, and the probability that it will not need any is $q = 1 - p$. Let Y_n be the number of new light bulbs left in the house at the end of day n. Y_n is a Markov chain with state transition diagram shown in Fig. 11.2(b), and transition probability matrix

$$P = \begin{bmatrix} 1 & 0 & 0 \\ p & q & 0 \\ 0 & p & q \end{bmatrix}.$$

The trellis diagram for this process in Fig. 11.3(b) shows that, unless $q = 1$, the transition probabilities bias the process towards the "trapping" state $Y_n = 0$. Thus the sample functions of Y_n are nonincreasing functions of n.

Example 11.8 Binomial Counting Process

Let S_n be the binomial counting process introduced in Example 9.15. In one step, S_n can either stay the same or increase by one. The state transition diagram is shown in Fig. 11.2(c), and the transition probability matrix is given by

$$P = \begin{bmatrix} 1-p & p & 0 & 0 & \cdots \\ 0 & 1-p & p & 0 & \cdots \\ 0 & 0 & 1-p & p & \cdots \\ \cdot & \cdot & \cdots & & \cdots \end{bmatrix}.$$

The trellis diagram for binomial process in Fig. 11.3(c) shows that, unless $q = 1$, the transition probabilities bias the process towards steady growth over time. The sample functions of S_n are nondecreasing functions of n.

11.2.1 The *n*-Step Transition Probabilities

To evaluate the joint pmf for arbitrary time instants (see Eq. 11.3), we need to know the transition probabilities for an arbitrary number of steps. Let $P(n) = \{p_{ij}(n)\}$ *be the matrix of n-step transition probabilities*, where

$$p_{ij}(n) = P[X_{n+k} = j \,|\, X_k = i] \qquad n \geq 0, i, j \geq 0. \tag{11.11}$$

Note that $P[X_{n+k} = j|X_k = i] = P[X_n = j|X_0 = i]$ for all $n \geq 0$ and $k \geq 0$, since the transition probabilities do not depend on time.

First, consider the two-step transition probabilities. The probability of going from state i at $t = 0$, passing through state k at $t = 1$, and ending at state j at $t = 2$ is

$$\begin{aligned} P[X_2 = j, X_1 = k \,|\, X_0 = i] &= \frac{P[X_2 = j, X_1 = k, X_0 = i]}{P[X_0 = i]} \\ &= \frac{P[X_2 = j \,|\, X_1 = k]P[X_1 = k \,|\, X_0 = i]P[X_0 = i]}{P[X_0 = i]} \\ &= P[X_2 = j \,|\, X_1 = k]P[X_1 = k \,|\, X_0 = i] \\ &= p_{ik}(1)p_{kj}(1). \end{aligned}$$

Note that $p_{ik}(1)$ and $p_{kj}(1)$ are components of P, the one-step transition probability matrix. We obtain $p_{ij}(2)$, the probability of going from i at $t = 0$ to j at $t = 2$, by summing over all possible intermediate states k:

$$p_{ij}(2) = \sum_k p_{ik}(1)p_{kj}(1) \qquad \text{for all } i, j. \tag{11.12a}$$

Equation (11.12a) states that the ij entry of $P(2)$ is obtained by multiplying the ith row of $P(1)$ by the jth column of $P(1)$. In other words, $P(2)$ is obtained by multiplying the one-step transition probability matrices:

$$P(2) = P(1)P(1) = P^2. \tag{11.12b}$$

Now consider the probability of going from state i at $t = 0$, passing through state k at $t = m$, and ending at state j at time $t = m + n$. Following the same procedure as above we obtain the **Chapman–Kolmogorov equations**:

$$p_{ij}(m + n) = \sum_k p_{ik}(m)p_{kj}(n) \quad \text{for all } n, m \geq 0 \text{ all } i, j. \tag{11.13a}$$

Therefore the matrix of $n + m$ step transition probabilities $P(n + m) = \{p_{ij}(n + m)\}$ is obtained by the following matrix multiplication:

$$P(n + m) = P(n)P(m). \tag{11.13b}$$

It is easy to show by an induction argument that this implies that:

$$P(n) = P^n. \tag{11.14}$$

When the Markov chain has finite state, we can use computer programs to calculate the powers of P numerically.

11.2.2 The State Probabilities

Now consider the state probabilities at time n. Let $\mathbf{p}(n) = \{p_j(n)\}$ denote the row vector of **state probabilities** at time n. The probability $p_j(n)$ is related to $\mathbf{p}(n - 1)$ by

$$\begin{aligned} p_j(n) &= \sum_i P[X_n = j \,|\, X_{n-1} = i]P[X_{n-1} = i] \\ &= \sum_i p_{ij}p_i(n - 1). \end{aligned} \tag{11.15a}$$

Equation (11.15a) states that $\mathbf{p}(n)$ is obtained by multiplying the row vector $\mathbf{p}(n-1)$ by the matrix P:

$$\mathbf{p}(n) = \mathbf{p}(n-1)P. \tag{11.15b}$$

Similarly, $p_j(n)$ is related to $\mathbf{p}(0)$ by

$$\begin{aligned} p_j(n) &= \sum_i P[X_n = j | X_0 = i]P[X_0 = i] \\ &= \sum_i p_{ij}(n)p_i(0), \end{aligned} \tag{11.16a}$$

and in matrix notation

$$\mathbf{p}(n) = \mathbf{p}(0)P(n) = \mathbf{p}(0)P^n \qquad n = 1, 2, \ldots. \tag{11.16b}$$

Thus the *state pmf at time n is obtained by multiplying the initial state pmf by* P^n.

Example 11.9

To find the n-step transition probability in Example 11.7, note that

$$\begin{aligned} p_{22}(n) &= P[\text{no new light bulbs needed in } n \text{ days}] = q^n \\ p_{21}(n) &= P[1 \text{ light bulb needed in } n \text{ days}] = npq^{n-1} \\ p_{20}(n) &= 1 - p_{22}(n) - p_{21}(n). \end{aligned}$$

The other terms in $P(n)$ are found in similar fashion, thus

$$P(n) = \begin{bmatrix} 1 & 0 & 0 \\ 1-q^n & q^n & 0 \\ 1-q^n-npq^{n-1} & npq^{n-1} & q^n \end{bmatrix}.$$

Note that if $q < 1$ then, as $n \to \infty$,

$$P(n) \to \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

As a result, the state pmf $\mathbf{p}(n) = (p_0(n), p_1(n), p_2(n))$ approaches

$$\begin{aligned} \mathbf{p}(n) &= (p_0(0), p_1(0), p_2(0))P(n) \\ &= (0, 0, 1)P(n) \\ &\to (0, 0, 1)\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = (1, 0, 0), \end{aligned}$$

where $(p_0(0), p_1(0), p_2(0))$ is the row vector of initial state probabilities and $(p_0(0), p_1(0), p_2(0)) = (0, 0, 1)$ since we start with two light bulbs. As time progresses, $p_0(n) \to 1$. In words, the above equation states that we eventually run out of light bulbs!

Example 11.10

Let $\alpha = 1/10$ and $\beta = 1/5$ in Example 11.6. Find $P(n)$ for $n = 2, 4, 8$, and 16.

$$P^2 = \begin{bmatrix} .9 & .1 \\ .2 & .8 \end{bmatrix}^2 = \begin{bmatrix} .83 & .17 \\ .34 & .66 \end{bmatrix}$$

$$P^4 = \begin{bmatrix} .83 & .17 \\ .34 & .66 \end{bmatrix}^2 = \begin{bmatrix} .7467 & .2533 \\ .5066 & .4934 \end{bmatrix}$$

and similarly

$$P^8 = \begin{bmatrix} .6859 & .3141 \\ .6282 & .3718 \end{bmatrix} \qquad P^{16} = \begin{bmatrix} .6678 & .3322 \\ .6644 & .3356 \end{bmatrix}.$$

There is a clear trend here: It appears that as $n \to \infty$,

$$P^n \to \begin{bmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{bmatrix}.$$

We can use matrix diagonalization methods from linear algebra to find P^n [Anton, p. 246]. First we find that the eigenvalues of P are 1 and $1 - \alpha - \beta$ from:

$$0 = \det(P - \lambda I) = \begin{vmatrix} 1 - \alpha - \lambda & \alpha \\ \beta & 1 - \beta - \lambda \end{vmatrix} = (1 - \alpha - \lambda)(1 - \beta - \lambda) - \alpha\beta$$
$$= (1 - \lambda)(1 - \alpha - \beta - \lambda).$$

The corresponding eigenvectors are:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \mathbf{e}_2 = \begin{bmatrix} \alpha \\ -\beta \end{bmatrix}$$

so the matrix with eigenvectors as columns is:

$$\mathbf{E} = [\mathbf{e}_1\, \mathbf{e}_2] = \begin{bmatrix} 1 & \alpha \\ 1 & -\beta \end{bmatrix}.$$

We then have that:

$$P = \mathbf{E}\Lambda\mathbf{E}^{-1} = \frac{1}{\alpha + \beta}\begin{bmatrix} 1 & \alpha \\ 1 & -\beta \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 - \alpha - \beta \end{bmatrix}\begin{bmatrix} \beta & \alpha \\ 1 & -1 \end{bmatrix}.$$

The payoff is in the calculation of P^n:

$$P^n = (\mathbf{E}\Lambda\mathbf{E}^{-1})(\mathbf{E}\Lambda\mathbf{E}^{-1}) \ldots (\mathbf{E}\Lambda\mathbf{E}^{-1}) = \mathbf{E}\Lambda(\mathbf{E}^{-1}\mathbf{E})\Lambda \ldots \Lambda(\mathbf{E}^{-1}\mathbf{E})\Lambda\mathbf{E}^{-1}$$
$$= \mathbf{E}\Lambda\Lambda \ldots \Lambda\mathbf{E}^{-1} = \mathbf{E}\Lambda^n\mathbf{E}^{-1}$$
$$= \frac{1}{\alpha + \beta}\begin{bmatrix} 1 & \alpha \\ 1 & -\beta \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & (1 - \alpha - \beta)^n \end{bmatrix}\begin{bmatrix} \beta & \alpha \\ 1 & -1 \end{bmatrix}$$
$$= \begin{bmatrix} \dfrac{\beta}{\alpha + \beta} & \dfrac{\alpha}{\alpha + \beta} \\ \dfrac{\beta}{\alpha + \beta} & \dfrac{\alpha}{\alpha + \beta} \end{bmatrix} + \frac{(1 - \alpha - \beta)^n}{\alpha + \beta}\begin{bmatrix} \alpha & \alpha \\ -\beta & \beta \end{bmatrix}.$$

As long as $|1 - \alpha - \beta| < 1$, the second term goes to zero as $n \to \infty$ and so

$$P^n \to \begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}.$$

Note that all the rows are the same in the limiting matrix.

Example 11.11

Let the initial state probabilities in Example 11.10 be

$$P[X_0 = 0] = p_0(0) \quad \text{and} \quad P[X_0 = 1] = 1 - p_0(0).$$

Find the state probabilities as $n \to \infty$.

The state probability vector at time n is:

$$\begin{aligned} \mathbf{p}(n) &= (p_0(0), 1 - p_0(0)P^n \\ &= (p_0(0), 1 - p_0(0)\begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} + \frac{(1-\alpha-\beta)^n}{\alpha+\beta}(p_0(0), 1 - p_0(0)\begin{bmatrix} \alpha & \alpha \\ -\beta & \beta \end{bmatrix}. \end{aligned}$$

As $n \to \infty$, we have that

$$\mathbf{p}(n) \to (p_0(0), 1 - p_0(0))\begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{2}{3}, \frac{1}{3} \end{bmatrix}.$$

We see that the state probabilities do not depend on the initial state probabilities as $n \to \infty$.

Example 11.12 Google PageRank

A Web surfer browses pages in a five-page Web universe shown in Fig. 11.4(a). The surfer selects the next page to view by selecting with equal probability from the pages pointed to by the current page. If a page has no outgoing link (e.g., page 2), then the surfer selects any of the pages in the universe with equal probability. Find the probability that the surfer views page i.

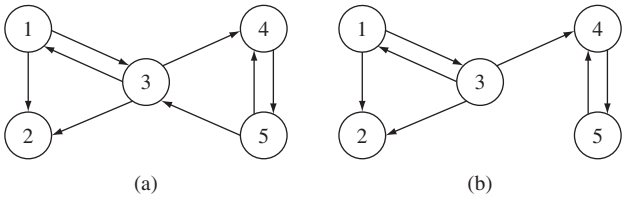

FIGURE 11.4
State-transition diagrams for PageRank examples.

The viewing behavior can be modeled by a Markov chain where the state represents the page currently viewed. If the current page points to k pages, then the next page is selected from that group with probability $1/k$. If the current page does not point to any pages, then the next page can be any of the 5 pages with probability 1/5. The transition probability for the Markov chain is:

$$P = [p_{ij}] = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1/2 & 0 & 1/2 \end{bmatrix}.$$

We can obtain the limiting state probabilities numerically by letting Octave calculate a high power of P, say P^{50}. We then obtain a 5×5 matrix in which all the rows are equal:

$$\mathbf{p}(n) \to (0.12195, 0.18293, 0.25610, 0.12195, 0.31707).$$

In the next subsection we will show an easier way of finding the steady state pmf.

The random surfer model forms the basis for the PageRank algorithm that was introduced by Google to rank the importance of a page in the Web. The rank of a page is given by the steady state probability of the page in the Markov chain model. The size of the state space in this Markov chain is in the billions of pages!

11.2.3 Steady State Probabilities

Example 11.11 is typical of Markov chains that settle into stationary behavior after the process has been running for a long time. As $n \to \infty$, the n-step transition probability matrix approaches a matrix in which all the rows are equal to the same pmf, that is,

$$p_{ij}(n) \to \pi_j \quad \text{for all } i. \tag{11.17a}$$

We can express the above in matrix notation as:

$$P^n \to \mathbf{1}\boldsymbol{\pi} \tag{11.17b}$$

where $\mathbf{1}$ is a column vector of all 1's, that is, $\mathbf{1}^{\mathrm{T}} = (1, 1, \dots)$ and $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$. From Eq. (11.16a), the convergence of P^n implies the convergence of the state pmf's:

$$p_j(n) = \sum_i p_{ij}(n) p_i(0) \to \sum_i \pi_j p_i(0) = \pi_j. \tag{11.18}$$

We say that the system reaches "equilibrium" or "steady state."

We can find the pmf $\boldsymbol{\pi} \triangleq \{\pi_j\}$ in Eq. (11.18) (when it exists) by noting that as $n \to \infty$, $p_j(n) \to \pi_j$ and $p_i(n-1) \to \pi_i$, so Eq. (11.15) approaches

$$\pi_j = \sum_i p_{ij} \pi_i, \tag{11.19a}$$

which in matrix notation is

$$\boldsymbol{\pi} = \boldsymbol{\pi} P. \tag{11.19b}$$

Equation (11.19b) is underdetermined and requires the normalization equation:

$$\sum_i \pi_i = 1. \tag{11.19c}$$

We refer to $\boldsymbol{\pi}$ as the **stationary state pmf** of the Markov chain. If we start the Markov chain with initial state pmf $\mathbf{p}(0) = \boldsymbol{\pi}$, then by Eqs. (11.16b) and (11.19b) we have that the state probability vector

$$\mathbf{p}(n) = \boldsymbol{\pi} P^n = \pi \qquad \text{for all } n.$$

The resulting process is a stationary random process as defined in Section 9.6, since the probability of the sequence of states $i_0, i_1, \ldots, i_n$ starting at time k is, by Eq. (11.5),

$$\begin{aligned} P[X_{n+k} = i_n, \ldots, X_k = i_0] &= P[X_{n+k} = i_n \,|\, X_{n+k-1} = i_{n-1}] \ldots P[X_{1+k} = i_1 \,|\, X_k = i_0] P[X_k = i_0] \\ &= P[X_{n+k} = i_n \,|\, X_{n+k-1} = i_{n-1}] \ldots P[X_{1+k} = i_1 \,|\, X_k = i_0] \pi_{i_0} \\ &= p_{i_{n-1},i_n} \cdots p_{i_0,i_1} \pi_{i_0}, \end{aligned}$$

which is independent of the initial time k. Thus the probabilities are independent of the choice of time origin, and the process is stationary.

Example 11.13

Find the stationary state pmf in Example 11.6.

Equation (11.19a) gives

$$\begin{aligned} \pi_0 &= (1 - \alpha)\pi_0 + \beta\pi_1 \\ \pi_1 &= \alpha\pi_0 + (1 - \beta)\pi_1, \end{aligned}$$

which imply that $\alpha\pi_0 = \beta\pi_1 = \beta(1 - \pi_0)$ since $\pi_0 + \pi_1 = 1$. Thus

$$\pi_0 = \frac{\beta}{\alpha + \beta} = \frac{2}{3} \qquad \pi_1 = \frac{\alpha}{\alpha + \beta} = \frac{1}{3}.$$

In this section we have shown the typical behavior of many Markov chains where the n-step transition probabilities and the state probabilities converge to constants that are independent of the initial conditions. These constant probabilities are found by solving the set of linear equations (11.19). It is worth noting, however, that not all Markov chains settle into stationary behavior where the process "forgets" the initial conditions. For example, the binomial counting process (Example 9.15) with $p > 0$ grows steadily so that for any fixed j, $p_j(n) \to 0$ as $n \to \infty$. The following example shows two atypical situations where the initial conditions determine the behavior for all time.

Example 11.14 Two-State Process with Atypical Behavior

Consider the two-state process with state transition diagram shown in Fig. 11.2(a). In Example 11.10 we found that the two-state process settles into steady state behavior so long as $|1 - \alpha - \beta| < 1$. Let's see what happens when this condition is not satisfied.

Consider first the case where $\alpha = \beta = 1$, and suppose that we start the process in state 0, that is, $p_0(0) = 1$. The state probabilities at time n are:

$$\mathbf{p}(n) = (p_0(0), 1 - p_0(0)P^n = (1, 0)\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^n.$$

The process in this case alternates between state 0 at even time instants and state 1 at odd time instants. P^n does not converge, and instead alternates assuming the values P and $P^2 = I$. The state probability vector alternates between the values $(1, 0)$ and $(0, 1)$ so it does not exhibit convergence.

Now consider the case $\alpha = \beta = 0$, and suppose again that we start the process in state 0, that is, $p_0(0) = 1$. The state probabilities at time n are:

$$\mathbf{p}(n) = (1, 0)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^n = (1, 0) \text{ for all } n.$$

In this case, the process remains fixed at state 0, which was selected at the initial time instant. Note that the process would have remained fixed at state 1 if state 1 had been selected initially. The state probability vector remains fixed at $(1, 0)$ if the initial state was 0 or $(0, 1)$ if the initial state was 1. In this case, both P^n and $\mathbf{p}(n)$ converge immediately but to values that are determined by the initial condition.

The previous example demonstrates that we need to identify the conditions under which the state probability of Markov chains will converge to a stationary pmf that is found from Eq. (11.19). This is the topic of the next section.

11.3 CLASSES OF STATES, RECURRENCE PROPERTIES, AND LIMITING PROBABILITIES

In this section we take a closer look at the relation between the behavior of a Markov chain and its transition probability matrix. First we see that the states of a discrete-time Markov chain can be divided into one or more separate classes and that these classes can be of several types. We then show that the long-term behavior of a Markov chain is related to the types of its state classes. Figure 11.5 summarizes the types of classes to which a state can belong and identifies the associated long-term behavior.

11.3.1 Classes of States

We say that **state *j* is accessible from state *i*** if for some $n \geq 0$, $p_{ij}(n) > 0$, that is, if there is a sequence of transitions from i to j that has nonzero probability. We say that **states *i* and *j* communicate** if they are accessible to each other; we then write $i \leftrightarrow j$. Note that a state communicates with itself since $p_{ii}(0) = 1$.

If state i communicates with state j and state j communicates with state k, that is, $i \leftrightarrow j$ and $j \leftrightarrow k$, then state i communicates with k. To see this, note that $i \leftrightarrow j$ implies that there is a nonzero probability path from i to j and $j \leftrightarrow k$ implies that there is a subsequent nonzero probability path from j to k. The combined paths form a nonzero probability path from i to k. A nonzero probability path in the reverse direction exists for the same reasons.

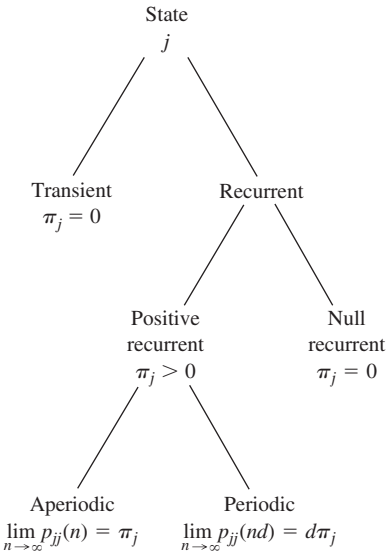

FIGURE 11.5
Classification of states and associated long-term behavior. The proportion of time spent in state j is denoted by π_j.

We say that two states belong to the same **class** if they communicate with each other. Note that two different classes of states must be disjoint since having a state in common would imply that the states from both classes communicate with each other. Thus *the states of a Markov chain consist of one or more disjoint communication classes.* A Markov chain that consists of a single class is said to be **irreducible**.

Example 11.15

Figure 11.6(a) shows the state transition diagram for a Markov chain with three classes: $\{0\}, \{1, 2\}$, and $\{3\}$.

Example 11.16

Figure 11.6(b) shows the state transition diagram for a Markov chain with one class: $\{0, 1, 2, 3\}$. Thus the chain is irreducible.

Example 11.17 Binomial Counting Process

Figure 11.6(c) shows the state transition diagram for a binomial counting process. It can be seen that the classes are: $\{0\}, \{1\}, \{2\}, \ldots$.

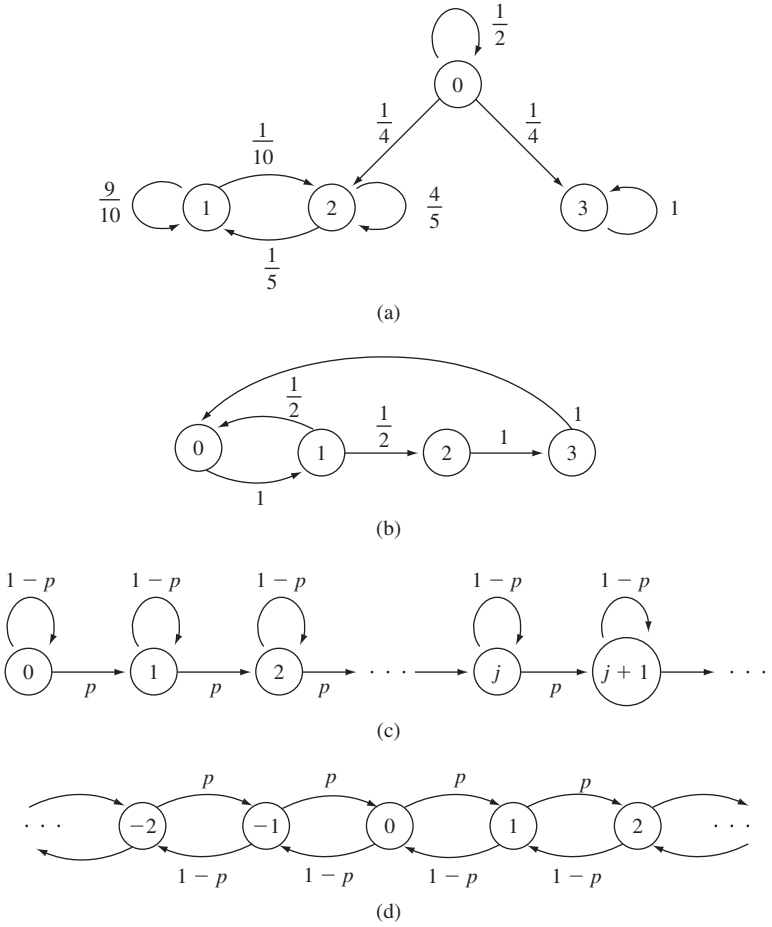

FIGURE 11.6
(a) A three-class Markov chain. (b) A periodic Markov chain. (c) A binomial counting process. (d) The random walk process.

Example 11.18 Random Walk

Figure 11.6(d) shows the state transition diagram for the random walk process. If $p > 0$, then the process has only one class, $\{0, \pm 1, \pm 2, \dots\}$, so it is irreducible.

11.3.2 Recurrence Properties

Suppose we start a Markov chain in state i. State i is said to be **recurrent** if the process returns to the state with probability one, that is,

$$f_i = P[\text{ever returning to state } i] = 1. \tag{11.20a}$$

State i is said to be **transient** if

$$f_i < 1. \tag{11.20b}$$

If we start the Markov chain in a recurrent state i, then the state reoccurs an infinite number of times. If we start the Markov chain in a transient state, the state does not reoccur after some finite number of returns. Each reoccurrence of the state can be viewed as a failure in a Bernoulli trial. The probability of failure is f_i. Thus the number of returns to state i terminating with a success (no return) is a geometric random variable with mean $(1 - f_i)^{-1}$. If $f_i < 1$, then the probability of an infinite number of successes is zero. Therefore a transient state reoccurs only a finite number of times.

Let X_n denote the Markov chain with initial state i, $X_0 = i$. Let $I_i(X)$ be the indicator function for state i, that is, $I_i(X)$ is equal to 1 if $X = i$ and equal to 0 otherwise. The expected number of returns to state i is then

$$E\left[\sum_{n=1}^{\infty} I_i(X_n) \,|\, X_0 = i\right] = \sum_{n=1}^{\infty} E[I_i(X_n) \,|\, X_0 = i] = \sum_{n=1}^{\infty} p_{ii}(n) \tag{11.21}$$

since by Example 4.16

$$E[I_i(X_n) \,|\, X_0 = i] = P[X_n = i \,|\, X_0 = i] = p_{ii}(n).$$

A state is recurrent if and only if it reoccurs an infinite number of times, thus from Eq. (11.21) *state i is recurrent if and only if*

$$\sum_{n=1}^{\infty} p_{ii}(n) = \infty. \tag{11.22}$$

Similarly, *state i is transient if and only if*

$$\sum_{n=1}^{\infty} p_{ii}(n) < \infty. \tag{11.23}$$

Example 11.19

In Example 11.15 (Fig. 11.6a), state 0 is transient since $p_{00}(n) = (1/2)^n$, so

$$\sum_{n=1}^{\infty} p_{00}(n) = \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \cdots = 1 < \infty.$$

On the other hand, if the process were started in state 1, we would have the two-state process discussed in Example 11.10. For such a process we found that

$$p_{11}(n) = \frac{\beta + \alpha(1 - \alpha - \beta)^n}{\alpha + \beta} = \frac{1/2 + 1/4(7/10)^n}{3/4}$$

so that

$$\sum_{n=1}^{\infty} p_{11}(n) = \sum_{n=1}^{\infty}\left(\frac{2}{3} + \frac{(7/10)^n}{3}\right) = \infty.$$

Therefore state 1 is recurrent.

Example 11.20 Binomial Counting Process

In the binomial counting process all the states are transient since $p_{ii}(n) = (1 - p)^n$ so that for $p > 0$,

$$\sum_{n=1}^{\infty} p_{ii}(n) = \sum_{n=1}^{\infty} (1 - p)^n = \frac{1 - p}{p} < \infty.$$

Example 11.21 Random Walk

Consider state zero in the random walk process in Fig. 11.6(d). The state reoccurs in $2n$ steps if and only if n +1s and n −1s occur during the $2n$ steps. This occurs with probability

$$p_{00}(2n) = \binom{2n}{n} p^n (1 - p)^n.$$

Stirling's formula for $n!$ can be used to show that

$$\binom{2n}{n} p^n (1 - p)^n \sim \frac{(4p(1 - p))^n}{\sqrt{\pi n}},$$

where $a_n \sim b_n$ when $\lim_{n\to\infty} a_n/b_n = 1$.

Thus Eq. (11.21) for state 0 is

$$\sum_{n=1}^{\infty} p_{00}(2n) \sim \sum_{n=1}^{\infty} \frac{(4p(1 - p))^n}{\sqrt{\pi n}}.$$

If $p = 1/2$, then $4p(1 - p) = 1$ and the series diverges. It then follows that state 0 is recurrent. If $p \neq 1/2$, then $(4p(1 - p)) < 1$, and the above series converges. This implies that state 0 is transient. Thus when $p = 1/2$, the random walk process maintains a precarious balance about 0. As soon as $p \neq 1/2$, a positive or negative drift is introduced and the process grows towards $\pm\infty$.

Recurrence and transience are class properties: If a state i is recurrent, then all states in its class are recurrent; if a state is transient, then all the states in its class are transient. If state i is recurrent, then all states in its class will be visited eventually as the process forever returns to state i over and over again. Indeed all other states in its class will appear an infinite number of times.

To show the recurrence class property, let i be a recurrent state and let j be another state in the class, then $i \leftrightarrow j$, and there are probabilities $p_{ji}(m) > 0$ and $p_{ij}(l) > 0$ that corresponds to nonzero probability paths that lead from j to i in m steps, and back from i to j in l steps. We can identify many nonzero probability paths that go from j to j by splicing the above two paths to recurrent paths for state i: go from j to i using the above path; then from i to i using an n-step recurrent path; then back from i to j using the above path. The probabilities for these paths provide a lower bound to the recurrence probabilities for j:

$$\sum_k p_{jj}(k) > \sum_n p_{ji}(m) p_{ii}(n) p_{ij}(l) = p_{ji}(m) p_{ij}(l) \sum_n p_{ii}(n) = \infty,$$

since state i is recurrent. This implies that state j is also recurrent. Now suppose that state i is transient, and let j be another state in its class. State j cannot be recurrent, for this would imply that i is recurrent, in contradiction to our assumption. Therefore j must be transient.

If a Markov chain is irreducible then either all its states are transient or all its states are recurrent. If the Markov chain has a finite state space, it is impossible for all of its states to be transient. At least some of the states must occur an infinite number of times as time progresses, implying that all states are recurrent. Therefore, *the states of a finite-state, irreducible Markov chain are all recurrent*. If the state space is countably infinite, then all the states can be transient. The random walk with $p \neq 1/2$ provides an example of such a Markov chain.

The structure of the state transition diagram and the associated nonzero transition probabilities can impose periodicity in the realizations of a discrete-time Markov chain. We say that state i has **period** d if it can only reoccur at times that are multiples of d, that is, $p_{ii}(n) = 0$ whenever n is not a multiple of d, where d is the largest integer with this property. We say that state i is **aperiodic** if it has period $d = 1$.

Periodicity is a class property, that is, all states in a class have the same period. An irreducible Markov chain is said to be **aperiodic** if the states in its single class have period one. An irreducible Markov chain is said to be **periodic** if its states have period $d > 1$.

To show that periodicity is a class property, suppose that state i has period d and let j be another state in the same class. Since $i \leftrightarrow j$, there are probabilities $p_{ji}(m) > 0$ and $p_{ij}(l) > 0$ that corresponds to paths that lead from j to i in m steps, and back from i to j in l steps. We can create a path from j to j by splicing the m-step path for j to i with the l-step path from i to j; this path has length $m + l$ and probability $p_{ji}(m)p_{ij}(l) > 0$. The length $m + l$ must be divisible by d', the period of state j. Now create multiple paths from j to j by attaching the above two paths to nonzero probability paths that go from i to i in n steps. These paths have length $m + l + n$ and probability $p_{ji}(m)p_{ii}(n)p_{ij}(l) > 0$. All these paths go from j to j so $m + n + l$ must be divisible by d'. We already showed that $m + l$ is divisible by d', so we have that n must also be divisible by d'. But n can be the length of any path that goes from i to i, and so d, the period of state i, is the largest value that divides all such n. This implies that d' must divide d. By reversing the roles of state i and state j, the same series of arguments imply that d must divide d'. Thus $d = d'$ and state i and state j have the same period.

Example 11.22 Two-State Process with Atypical Behavior

Characterize the two "atypical" Markov chains in Example 11.14.

In the case where $\alpha = \beta = 1$, Fig. 11.2(a) shows that we have a single communication class with period $d = 2$. This explains why the process alternates between state 0 at even time instants and state 1 at odd time instants

In the case $\alpha = \beta = 0$, we have two communication classes: $\{0\}$ and $\{1\}$. The selection of the initial state at $t = 0$ effectively picks one of the two classes, and the process remains in that class forever.

Example 11.23

In Example 11.15 (Fig 11.6a), all the states have the property that $p_{ii}(n) > 0$ for $n = 1, 2, \ldots$. Therefore all three classes in the Markov chain have period 1.

Example 11.24

In the Markov chain in Fig 11.6(b), the states 0 and 1 can reoccur at time $2, 4, 6, \ldots$ and states 2 and 3 at times $4, 6, 8, \ldots$. Therefore the Markov chain has period 2.

Example 11.25

In the random walk process in Fig 11.6(d), a state reoccurs when the number of successes $(+1\text{s})$ equals the number of failures (-1s). This can only happen after an even number of steps. The process therefore has period 2.

Figure 11.7(a) summarizes the possible structures that can be encountered for Markov chains. In the case of irreducible finite-state Markov chains, all states in the single class must be recurrent and the class can either be aperiodic or periodic. If a finite-state

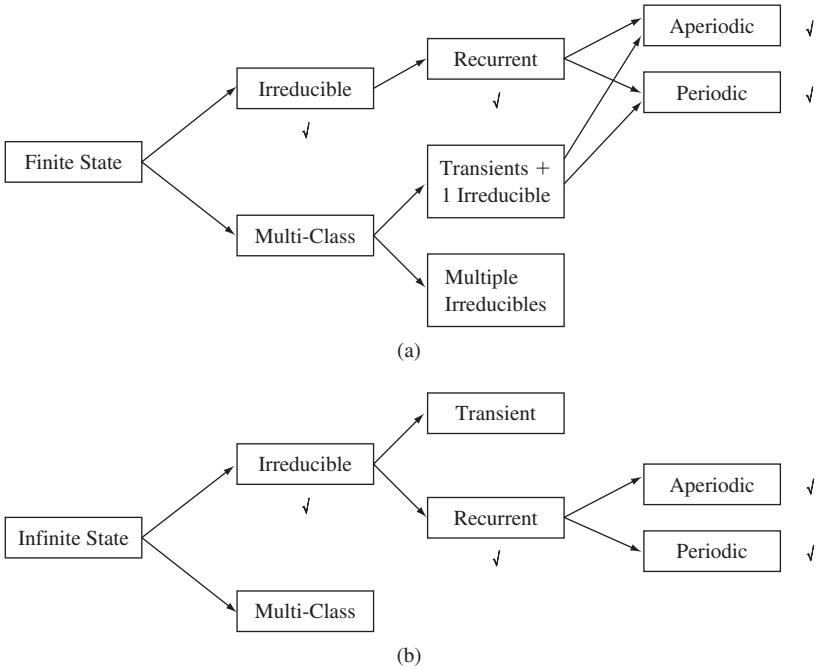

FIGURE 11.7
Possible structures for Markov chains.

Markov chain consists of multiple transient classes and a single irreducible class, then the chain will eventually settle in the states of the irreducible class. Thus in the long-run the behavior is the same as that of an irreducible chain. A finite-state Markov chain with multiple irreducible classes will eventually enter and remain thereafter in one of the irreducible classes. Over the long run, the chain will exhibit the behavior of an irreducible Markov chain with the given class of states. Thus the case of multi-irreducible classes can be viewed as a two stage random experiment in which the first stage involves selecting one of the irreducible classes.

Figure 11.7(b) summarizes the possible structures for Markov chains with infinite state space. The major difference from the finite case is that an irreducible class can have all of its states be transient. Consequently when a chain has multiple classes it is now possible for the chain to enter and remain in a class that is either transient or recurrent.

11.3.3 Limiting Probabilities

If all the states in a Markov chain are transient, then all the state probabilities approach zero as $n \to \infty$. If a Markov chain has some transient classes and some recurrent classes, as in Fig. 11.6(a), then eventually the process enters and remains thereafter in one of the recurrent classes. Therefore we can concentrate on individual recurrent classes when studying the limiting probabilities of a chain. For this reason we assume in this section that we are dealing with an irreducible Markov chain.

Suppose we start a Markov chain in a *recurrent* state i at time $n = 0$. Let $T_i(1), T_i(1) + T_i(2), \ldots$ be the times when the process returns to state i, where $T_i(k)$ is the time that elapses between the $(k-1)$th and kth returns (see Fig. 11.8). The T_i form an iid sequence since each return time is independent of previous return times.

The proportion of time spent in state i after k returns to i is

$$\text{proportion of time in state } i = \frac{k}{T_i(1) + T_i(2) + \cdots + T_i(k)}. \tag{11.24}$$

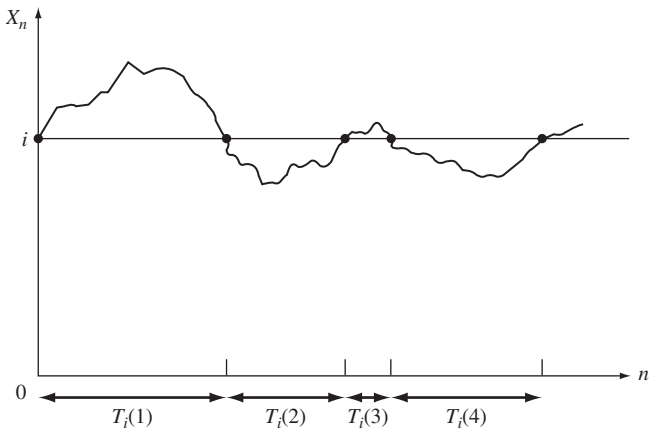

FIGURE 11.8
Recurrence times for state i.

Since the state is recurrent, the process returns to state i an infinite number of times. Thus the law of large numbers implies that, with probability one, the reciprocal of the above expression approaches the **mean recurrence time** $E[T_i]$ so the long-term proportion of time spent in state i approaches

$$\text{proportion of time in state } i \rightarrow \frac{1}{E[T_i]} = \pi_i, \tag{11.25}$$

where π_i *is the long-term proportion of time spent in state i.*

If $E[T_i] < \infty$, then we say that state i is **positive recurrent**. Equation (11.25) then implies that

$$\pi_i > 0 \qquad \text{if state } i \text{ is positive recurrent.}$$

If $E[T_i] = \infty$, then we say that state i is **null recurrent**. Equation (11.25) then implies that

$$\pi_i = 0 \qquad \text{if state } i \text{ is null recurrent.}$$

It can be shown that positive and null recurrence are class properties.

Positive recurrent, aperiodic states are called **ergodic**. Once a Markov chain enters an ergodic state, then the process will remain in the state's class forever. Furthermore the process will visit all states in the class sufficiently frequently that the long-term proportion of time in a given state will be governed by Eq. (11.25) and approach a nonzero value. Thus the process will reveal its underlying state probabilities through time averages. Given our previous discussion on ergodicity in Chapter 9, it is not surprising that an **ergodic Markov chain** is defined as an irreducible, aperiodic, positive recurrent Markov chain.

Example 11.26

The process in Fig. 11.6(b) returns to state 0 in two steps with probability 1/2 and in four steps with probability 1/2. Therefore the mean recurrence time for state 0 is

$$E[T_0] = \frac{1}{2}(2) + \frac{1}{2}(4) = 3.$$

Therefore state 0 is positive recurrent and the long-term proportion of time spent in state 0 is

$$\pi_0 = \frac{1}{3}.$$

Example 11.27 Random Walk

In Example 11.21 it was shown that the random walk process is recurrent if $p = 1/2$. However, the mean recurrence time can be shown to be infinite when $p = 1/2$ (Feller, 1968, p. 314). Thus all the states in the chain are null recurrent.

The π_j's in Eq. (11.25) satisfy the equations that define the stationary state pmf:

$$\pi_j = \sum_i \pi_i P_{ij} \qquad \text{for all } j \tag{11.26a}$$

and

$$1 = \sum_i \pi_i. \tag{11.26b}$$

To see this, note that since π_i is the proportion of time spent in state i, then $\pi_i P_{ij}$ is the proportion of time in which state j follows i. If we sum over all i, we then obtain the long-term proportion of time in state j, π_j.

Example 11.28

The stationary state pmf for the periodic Markov chain in Fig. 11.6(b) is found from Eqs. (11.26a) and (11.26b):

$$\begin{aligned}
\pi_0 &= \frac{1}{2}\pi_1 + \pi_3 \\
\pi_1 &= \pi_0 \\
\pi_2 &= \frac{1}{2}\pi_1 \\
\pi_3 &= \pi_2.
\end{aligned}$$

These equations imply that $\pi_1 = \pi_0$ and $\pi_2 = \pi_3 = \pi_0/2$. Since the probabilities must add to one, we obtain

$$\pi_1 = \pi_0 = \frac{1}{3} \qquad \text{and} \qquad \pi_2 = \pi_3 = \frac{1}{6}.$$

Note that $\pi_0 = 1/3$ was obtained for the mean recurrence time in Example 11.26.

In Section 11.2 we found that for certain Markov chains, the n-step transition matrix approaches a fixed matrix of equal rows as $n \to \infty$ (see Eq. 11.17). We also saw that the rows of this limiting matrix consisted of a pmf that satisfied Eqs. (11.26a) and (11.26b). We are now ready to state under what conditions this occurs.

Theorem 1[2]

For an irreducible aperiodic Markov chain exactly one of the following assertions holds:

(i) All states are transient or all states are null recurrent; $p_{ij}(n) \to 0$ as $n \to \infty$ for all i and j and there exists no stationary pmf

(ii) All states are positive recurrent, so

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j \quad \text{for all } j \tag{11.27}$$

where $\{\pi_j, j = 1, 2, 3, \dots\}$ is the unique stationary pmf solution to Eq. (11.26ab).

---

[2]A proof to Theorem 1 is given by [Ross, pp. 108–110].

Theorem 1 states that for ergodic Markov chains, the n-step transition probabilities approach constant values given by the steady state pmf. Note that Eq. (11.27) can be written in matrix form as shown in Eq. (11.17b). From Eq. (11.18), it then follows that the state probabilities approach steady state values that are independent of the initial conditions. These steady state probabilities correspond to the stationary probabilities obtained by solving Eq. (11.26ab), and thus correspond to the long-term proportion of time spent in a given state. Theorem 1 also states that if the irreducible Markov chain is transient or null recurrent, then a stationary pmf solution to Eq. (11.26ab) does not exist. This implies that when we do find a solution, and the chain is irreducible and aperiodic, then the Markov chain must be positive recurrent and hence ergodic.

Example 11.29 Age of a Device

Consider a Markov Chain that counts the age of a device in service at the end of each day. At the end of each day, the device either increases its age by 1 (with probability a) or fails and returns to the "1" state (with probability $1 - a$). A failed device is replaced at the beginning of the next day and the age counting processes is resumed. Determine whether the Markov chain has a stationary distribution.

The state transition diagram for the Markov chain is shown in Fig. 11.9. If $a > 0$, then every state i can access any state $i + 1$, and consequently any state i can access any state $j > i$. In addition every state i can access state 1. This implies that there is a nonzero probability path between any two states, and so the Markov chain is irreducible. State 1 can reoccur in intervals of $1, 2, 3, 4, \ldots$, and so state 1 has period 1. Therefore all the states have period 1 and the Markov chain is aperiodic.

The equations for the stationary probabilities are:

$$\pi_1 = (1 - a)\pi_1 + (1 - a)\pi_2 + \cdots = (1 - a)(\pi_1 + \pi_2 + \cdots) = 1 - a$$

$$\pi_{i+1} = a\pi_i \quad \text{for} \quad i \geq 1.$$

By a simple induction argument we can show that:

$$\pi_i = (1 - a)a^{i-1} \quad \text{for} \quad i \geq 1.$$

Therefore the Markov chain is positive recurrent and has this stationary pmf.

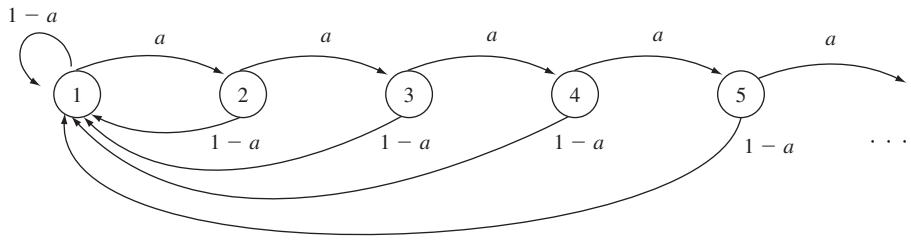

FIGURE 11.9
Age of a device.

Example 11.30 Google PageRank Algorithm

In Example 11.12 we showed the basic approach for ranking Web pages according to an associated Markov chain. The approach included a strategy to deal with the case where users become trapped in a page with no outgoing links, i.e., page 2 in Fig. 11.4(a). The approach, however, is not sufficient to ensure that the Markov chain is irreducible and aperiodic. For example, in Fig. 11.4(b) users can also become trapped in the periodic class $\{4, 5\}$. This poses a problem for the rank algorithm which uses the power of the transition probability matrix to obtain the stationary pmf. To deal with this problem, the PageRank algorithm also assumes that each time a new page is selected, the procedure in Example 11.12 is used with probability α, but otherwise (with probability $1 - \alpha$) any of all possible Web pages is selected with equal probability. The value $\alpha = 0.85$ is usually cited as appropriate. The modified ranking method then has a transition probability matrix that is aperiodic and irreducible and the conditions of Theorem 1 are satisfied.

For the example in Fig. 11.4(b) we have:

$$P = (0.85)\begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1/2 & 0 & 1/2 \end{bmatrix} + (0.15)\begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

$$= \begin{bmatrix} 0.0300 & 0.4550 & 0.4550 & 0.0300 & 0.0300 \\ 0.2000 & 0.2000 & 0.2000 & 0.2000 & 0.2000 \\ 0.3133 & 0.3133 & 0.0300 & 0.3133 & 0.0300 \\ 0.0300 & 0.0300 & 0.0300 & 0.0300 & 0.8800 \\ 0.0300 & 0.0300 & 0.4550 & 0.0300 & 0.4550 \end{bmatrix}.$$

The matrix P has a stationary state pmf given by:

$$\mathbf{p}(n) = (0.13175, 0.18772, 0.24642, 0.13173, 0.30239).$$

See [Langville] for more details on the PageRank algorithm.

For periodic processes, we have the following result.

Theorem 2

For an irreducible, periodic, and positive recurrent Markov chain with period d,

$$\lim_{n\to\infty} p_{jj}(nd) = d\pi_j \text{ for all } j \tag{11.28}$$

where π_j is the unique nonnegative solution of Eqs. (11.26a) and (11.26b).

As before, π_j represents the proportion of time spent in state j. However, the fact that state j is constrained to occur at multiples of d steps implies that the probability of occurrence of the state j is d times greater at the allowable times and zero elsewhere.

Example 11.31

In Examples 11.26 and 11.28 we found that the long-term proportion of time spent in state 0 is $\pi_0 = 1/3$. If we start in state 0, then only even states can occur at even time instants. Thus at these even time instants the probability of state 0 is 2/3 and of state 2 is 1/3. At odd time instants, the probabilities of states 0 and 2 are zero.

Theorems 1 and 2 only address the most important cases of irreducible, periodic and aperiodic Markov chains indicated by the checkmarks in Fig. 11.7. The following example considers a case not covered by Theorems 1 and 2.

Example 11.32 Markov Chain with Multiple Irreducible Classes

Does the Markov chain in Fig. 11.6(a) have a unique stationary pmf?

The equations for the stationary probabilities are:

$$\begin{aligned} p_0 &= 1/2p_0 \\ p_1 &= 9/10p_1 + 1/5p_2 \\ p_2 &= 1/4p_0 + 1/10p_1 + 4/5p_2 \\ p_3 &= 1/4p_0 + p_3. \end{aligned}$$

The first equation implies that $p_0 = 0$, which reduces the fourth equation to $p_3 = p_3$, which imposes no constraints on p_3. The middle two equations are equivalent and both imply that $p_1 = 2p_2$. The normalization condition requires that $1 = p_1 + p_2 + p_3 = 3p_2 + p_3$. Therefore the equations are underdetermined and there are many solutions with the form: $(0, 2p_2, p_2, 1 - 3p_2)$ where $0 \le p_2 \le 1/3$.

Now let's approach the problem according to its three classes: $\{0\}$, $\{1, 2\}$, and $\{3\}$. The first class is transient and the other two classes are recurrent. Suppose the initial state is 3, then the process remains in that state forever. The stationary pmf for class $\{3\}$ by itself is $(0, 0, 0, 1)$. If the initial state is 1 or 2, then the process remains in this class forever; the stationary pmf for this class in isolation is $(0, 2/3, 1/3, 0)$. Finally if the initial state is 0, then the process will eventually leave and enter one of the other two classes with equal probability. In the general case, if the initial state is selected according to the pmf $(p_0(0), p_1(0), p_2(0), p_3(0))$ then the class $\{1, 2\}$ will be entered with probability $1/2\, p_0(0) + p_1(0) + p_2(0)$, and class $\{3\}$ will be entered with probability $1/2\, p_0(0) + p_3(0)$. The stationary pmf would then have the form:

$$\begin{aligned} &\{1/2\, p_0(0) + p_1(0) + p_2(0)\}(0, 2/3, 1/3, 0) + \{1/2\, p_0(0) + p_3(0)\}\ (0, 0, 0, 1) \\ &\quad = \gamma(0, 2/3, 1/3, 0) + (1 - \gamma)(0, 0, 0, 1) \\ &\quad = (0, 2\gamma/3, \gamma/3, 1 - \gamma). \end{aligned}$$

If we let $\gamma/3 = p_2$ we see that this solution has the form we derived before.

For example, suppose the initial pmf was $(0, 1/3, 1/6, 1/2)$, then this pmf satisfies the condition for a stationary pmf and the repeated multiplication by P will yield the same pmf. In this sense this multiclass Markov chain has a stationary pmf. Note however that the relative frequencies of the states depend on which irreducible class is actually entered. Thus if we record long-term average frequencies we will observe either $(0, 2/3, 1/3, 0)$ or $(0, 0, 0, 1)$. The stationary pmf

does not correspond to either of these two pmf's; instead the stationary pmf gives us the expected value of the two pmf's:

$$(0, 1/3, 1/6, 1/2) = 1/2(0, 2/3, 1/3, 0) + 1/2(0, 0, 0, 1)$$

where 1/2 is the probability of entering the two irreducible classes for this choice of initial pmf.

Example 11.32 illustrates the behavior of multiclass finite-state Markov chain. In these chains the process will eventually enter and remain forever in one of its recurrent classes. Each recurrent class can be considered as a separate irreducible Markov chain with its own stationary pmf. The multiclass Markov chain will then have stationary pmf's that depend on the stationary pmf's of its constituent recurrent classes according to the initial state probabilities. These multiclass Markov chains are not ergodic since the relative frequencies of the states do not correspond to the stationary pmf.

If a multiclass chain has infinite state space, then the situation discussed above can occur as a special case: the process initially works its way through transient classes and eventually settles in one of a number of ergodic classes. However, in general, it is possible for some or all of the classes to be transient and/or null recurrent. In such case the process may never settle into stationary behavior.

11.4 CONTINUOUS-TIME MARKOV CHAINS

In Section 11.2 we saw that the transition probability matrix determines the behavior of a discrete-time Markov chain. In this section we see that the same is true for continuous-time Markov chains.

The joint pmf for $k + 1$ arbitrary time instants of a Markov chain is given by Eq. (11.3):

$$\begin{aligned} P[X(t_{k+1}) = x_{k+1}, X(t_k) = x_k, \ldots, X(t_1) = x_1] \\ = P[X(t_{k+1}) = x_{k+1} \mid X(t_k) = x_k] \cdots \\ \times P[X(t_2) = x_2 \mid X(t_1) = x_1] P[X(t_1) = x_1]. \end{aligned} \quad (11.29)$$

This result holds regardless of whether the process is discrete-time or continuous-time. In the continuous-time case, Eq. (11.29) requires that we know the transition probabilities from an arbitrary time s to an arbitrary time $s + t$:

$$P[X(s + t) = j \mid X(s) = i] \qquad t \geq 0.$$

We assume here that the transition probabilities depend only on the difference between the two times:

$$P[X(s + t) = j \mid X(s) = i] = P[X(t) = j \mid X(0) = i] = p_{ij}(t) \qquad t \geq 0, \text{ all } s. \quad (11.30)$$

We say that $X(t)$ has **homogeneous transition probabilities**.

Let $P(t) = \{p_{ij}(t)\}$ *denote the matrix of transition probabilities in an interval of length t.* Since $p_{ii}(0) = 1$ and $p_{ij}(0) = 0$ for $i \neq j$, we have

$$P(0) = I, \tag{11.31}$$

where I is the identity matrix.

Example 11.33 Poisson Process

For the Poisson process, the transition probabilities satisfy

$$\begin{aligned} p_{ij}(t) &= P[j - i \text{ events in } t \text{ seconds}] \\ &= p_{0,\,j-i}(t) \\ &= \frac{(\alpha t)^{j-i}}{(j-i)!} e^{-\alpha t} \qquad j \geq i. \end{aligned}$$

Therefore

$$P(t) = \begin{bmatrix} e^{-\alpha t} & \alpha t e^{-\alpha t} & (\alpha t)^2 e^{-\alpha t}/2! & \cdot & \cdots \\ 0 & e^{-\alpha t} & \alpha t e^{-\alpha t} & (\alpha t)^2 e^{-\alpha t}/2! & \cdots \\ 0 & 0 & e^{-\alpha t} & \alpha t e^{-\alpha t} & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

As t approaches zero, $e^{-\alpha t} \approx 1 - \alpha t$. Thus for a small time interval δ,

$$P(\delta) \approx \begin{bmatrix} 1 - \alpha\delta & \alpha\delta & 0 & \cdots \\ 0 & 1 - \alpha\delta & \alpha\delta & \cdots \\ 0 & 0 & 1 - \alpha\delta & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix},$$

where all terms of order δ^2 or higher have been neglected. Thus the probability of more than one transition in a very short time interval is negligible. Note that this is consistent with the assumptions made in deriving the Poisson process in Section 9.4.

Example 11.34 Random Telegraph

In the random telegraph example, the process $X(t)$ changes with each occurrence of an event in a Poisson process. From Eqs. (9.40) and (9.41) we see that the transition probabilities are as follows:

$$\begin{aligned} P[X(t) = a \,|\, X(0) = a] &= \frac{1}{2}\{1 + e^{-2\alpha t}\} \\ P[X(t) = a \,|\, X(0) = b] &= \frac{1}{2}\{1 - e^{-2\alpha t}\} \qquad \text{if } a \neq b. \end{aligned}$$

Thus the transition probability matrix is

$$P(t) = \begin{bmatrix} 1/2\{1 + e^{-2\alpha t}\} & 1/2\{1 - e^{-2\alpha t}\} \\ 1/2\{1 - e^{-2\alpha t}\} & 1/2\{1 + e^{-2\alpha t}\} \end{bmatrix}.$$

11.4.1 State Occupancy Times

Since the random telegraph signal changes polarity with each occurrence of an event in a Poisson process, it follows that the time spent in each state is an exponential random variable. It turns out that this is a property of the **state occupancy time** for all continuous-time Markov chains, that is: $X(t)$ *remains at a given value* (*state*) *for an exponentially distributed random time.* To see why, let T_i be the time spent in a state i. The probability of spending more than t seconds in this state is then

$$P[T_i > t].$$

Now suppose that the process has already been in state i for s seconds; then the probability of spending t more seconds in this state is

$$P[T_i > t + s \,|\, T_i > s] = P[T_i > t + s \,|\, X(s') = i, 0 \leq s' \leq s],$$

since the $\{T_i > s\}$ implies that the system has been in state i during the time interval $(0, s)$. The Markov property implies that if $X(s) = i$, then the past is irrelevant and we can view the system as being restarted in state i at time s:

$$P[T_i > t + s \,|\, T_i > s] = P[T_i > t]. \tag{11.32}$$

Only the exponential random variable satisfies this memoryless property (see Section 4.4). Thus the time spent in state i is an exponential random variable with some mean $1/v_i$:

$$P[T_i > t] = e^{-v_i t}. \tag{11.33}$$

The *mean state occupancy time* $1/v_i$ will usually be different for each state.

The above result provides us with another way of looking at continuous-time Markov chains. Each time a state, say i, is entered, an exponentially distributed state occupancy time T_i is selected. When the time is up, the next state j is selected according to a *discrete-time* Markov chain, with transition probabilities $\widetilde{q}_{ij}$. Then the new state occupancy time is selected according to T_j, and so on.[3] We call $\widetilde{q}_{ij}$ an **embedded Markov chain**. We will see in the last part of this section that the properties of the continuous-time Markov chain depends on the class properties of its embedded chain.

Example 11.35

The random telegraph signal in Example 11.34 spends an exponentially distributed time with mean $1/\alpha$ in each state. When a transition occurs, the transition is always from the present state to the only other state, thus the embedded Markov chain is

$$\widetilde{q}_{00} = 0 \qquad \widetilde{q}_{01} = 1$$
$$\widetilde{q}_{10} = 1 \qquad \widetilde{q}_{11} = 0.$$

---

[3]This view of Markov chains is useful in setting up computer simulation models of Markov chain processes.

11.4.2 Transition Rates and Time-Dependent State Probabilities

Consider the transition probabilities in a very short time interval of duration δ seconds. The probability that the process remains in state i during the interval is

$$\begin{aligned} P[T_i > \delta] &= e^{-v_i\delta} \\ &= 1 - \frac{v_i\delta}{1!} + \frac{v_i^2\delta^2}{2!} - \cdots \\ &= 1 - v_i\delta + o(\delta), \end{aligned}$$

where $o(\delta)$ denotes terms that become negligible relative to δ as δ approaches zero.[4] The exponential distributions of the state occupancy times imply that the probability of two or more transitions in an interval of duration δ is $o(\delta)$. Thus for small δ, $p_{ii}(\delta)$ is approximately equal to the probability that the process remains in state i for δ seconds:

$$\begin{aligned} p_{ii}(\delta) &= P[T_i > \delta] + o(\delta) \\ &= 1 - v_i\delta + o(\delta) \end{aligned}$$

or equivalently,

$$1 - p_{ii}(\delta) = v_i\delta + o(\delta). \tag{11.34}$$

We call v_i the *rate at which the process $X(t)$ leaves state i.*

Once the process leaves state i, it will enter state j with probability $\widetilde{q}_{ij}$, where $\widetilde{q}_{ij}$ is the transition probability of the embedded Markov chain. Thus

$$\begin{aligned} p_{ij}(\delta) &= (1 - p_{ii}(\delta))\widetilde{q}_{ij} \\ &= v_i\widetilde{q}_{ij}\delta + o(\delta) \\ &= \gamma_{ij}\delta + o(\delta). \end{aligned} \tag{11.35a}$$

We call $\gamma_{ij} = v_i\widetilde{q}_{ij}$ the *rate at which the process $X(t)$ enters state j from state i.* For completeness, we define $\gamma_{ii} = -v_i$, so that by Eq. (11.34),

$$p_{ii}(\delta) - 1 = \gamma_{ii}\delta + o(\delta). \tag{11.35b}$$

If we divide both sides of Eqs. (11.35a) and (11.35b) by δ and take the limit $\delta \to 0$, we obtain

$$\lim_{\delta\to 0}\frac{p_{ij}(\delta)}{\delta} = \gamma_{ij} \qquad i \neq j \tag{11.36a}$$

and

$$\lim_{\delta\to 0}\frac{p_{ii}(\delta) - 1}{\delta} = \gamma_{ii}, \tag{11.36b}$$

since

$$\lim_{\delta\to 0}\frac{o(\delta)}{\delta} = 0,$$

because $o(\delta)$ is of order higher than δ.

[4] A function $g(h)$ is said to be $o(h)$ if $\lim_{h\to 0} g(h)/h = 0$, that is, $g(h)$ goes to zero faster than h does.

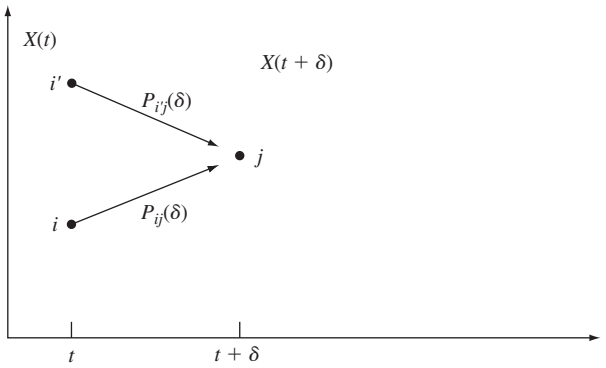

FIGURE 11.10
Transitions into state j.

We are now ready to develop a set of equations for finding the state probabilities at time t, which will be denoted by

$$p_j(t) \triangleq P[X(t) = j].$$

For $\delta > 0$, we have (see Fig. 11.10)

$$\begin{aligned} p_j(t + \delta) &= P[X(t + \delta) = j] \\ &= \sum_i P[X(t + \delta) = j \,|\, X(t) = i] P[X(t) = i] \\ &= \sum_i p_{ij}(\delta) p_i(t). \end{aligned} \tag{11.37}$$

If we subtract $p_j(t)$ from both sides, we obtain

$$p_j(t + \delta) - p_j(t) = \sum_{i \neq j} p_{ij}(\delta) p_i(t) + (p_{jj}(\delta) - 1) p_j(t). \tag{11.38}$$

If we divide by δ, apply Eqs. (11.36a) and (11.36b) and let $\delta \to 0$, we obtain

$$p_j'(t) = \sum_i \gamma_{ij} p_i(t). \tag{11.39}$$

Equation (11.39) is a form of the **Chapman–Kolmogorov equations** for continuous-time Markov chains. To find $p_j(t)$ we need to solve this system of differential equations with initial conditions specified by the initial state pmf $\{p_j(0), j = 0, 1, \dots\}$.

Note that if we solve Eq. (11.39) under the assumption that the state at time zero was i, that is, with initial condition $p_i(0) = 1$ and $p_j(0) = 0$ for all $j \neq i$, then the solution is actually $p_{ij}(t)$, the ij component of $P(t)$. Thus Eq. (11.39) can also be used to find the transition probability matrix.

Example 11.36 A Simple Queueing System

A queueing system alternates between two states. In state 0, the system is idle and waiting for a customer to arrive. This idle time is an exponential random variable with mean $1/\alpha$. In state 1, the system is busy servicing a customer. The time in the busy state is an exponential random variable with mean $1/\beta$. Find the state probabilities $p_0(t)$ and $p_1(t)$ in terms of the initial state probabilities $p_0(0)$ and $p_1(0)$.

The system moves from state 0 to state 1 at a rate α, and from state 1 to state 0 at a rate β:

$$\gamma_{00} = -\alpha \qquad \gamma_{01} = \alpha$$
$$\gamma_{10} = \beta \qquad \gamma_{11} = -\beta.$$

Equation (11.39) then gives

$$p_0'(t) = -\alpha p_0(t) + \beta p_1(t)$$
$$p_1'(t) = \alpha p_0(t) - \beta p_1(t).$$

Since $p_0(t) + p_1(t) = 1$, the first equation becomes

$$p_0'(t) = -\alpha p_0(t) + \beta(1 - p_0(t)),$$

which is a first-order differential equation:

$$p_0'(t) + (\alpha + \beta)p_0(t) = \beta \qquad p_0(0) = p_0.$$

The general solution of this equation is

$$p_0(t) = \frac{\beta}{\alpha + \beta} + Ce^{-(\alpha+\beta)t}.$$

We obtain C by setting $t = 0$ and solving in terms of $p_0(0)$; then we find

$$p_0(t) = \frac{\beta}{\alpha + \beta} + \left(p_0(0) - \frac{\beta}{\alpha + \beta}\right)e^{-(\alpha+\beta)t}$$

and

$$p_1(t) = \frac{\alpha}{\alpha + \beta} + \left(p_1(0) - \frac{\alpha}{\alpha + \beta}\right)e^{-(\alpha+\beta)t}.$$

Note that as $t \to \infty$,

$$p_0(t) \to \frac{\beta}{\alpha + \beta} \qquad \text{and} \qquad p_1(t) \to \frac{\alpha}{\alpha + \beta}.$$

Thus as $t \to \infty$, the state probabilities approach constant values that are independent of the initial state probabilities.

Example 11.37 The Poisson Process

Find the state probabilities for the Poisson process.

The Poisson process moves only from state i to state $i + 1$ at a rate α.

Thus

$$\gamma_{ii} = -\alpha \qquad \text{and} \qquad \gamma_{i,i+1} = \alpha.$$

Equation (11.39) then gives

$$p_0'(t) = -\alpha p_0(t) \qquad \text{for } j = 0$$
$$p_j'(t) = -\alpha p_j(t) + \alpha p_{j-1}(t) \qquad \text{for } j \geq 1.$$

The initial condition for the Poisson process is $p_0(0) = 1$, so the solution for the $j = 0$ equation is

$$p_0(t) = e^{-\alpha t}.$$

The equation for $j = 1$ is

$$p_1'(t) = -\alpha p_1(t) + \alpha e^{-\alpha t} \qquad p_1(0) = 0,$$

which is also a first-order differential equation for which the solution is

$$p_1(t) = \frac{\alpha t}{1!} e^{-\alpha t}.$$

It can be shown by an induction argument that the solution of the state j equation is

$$p_j(t) = \frac{(\alpha t)^j}{j!} e^{-\alpha t}.$$

For any fixed time t, the sum of $\{p_j(t)\}$ is one. Note however, that for any j, $p_j(t) \to 0$ as $t \to \infty$. Figure 11.11 shows how the pmf drifts to higher values as time progresses. Thus for the Poisson process, the probability of any finite state approaches zero as $t \to \infty$. This is consistent with the fact that the process grows steadily with time.

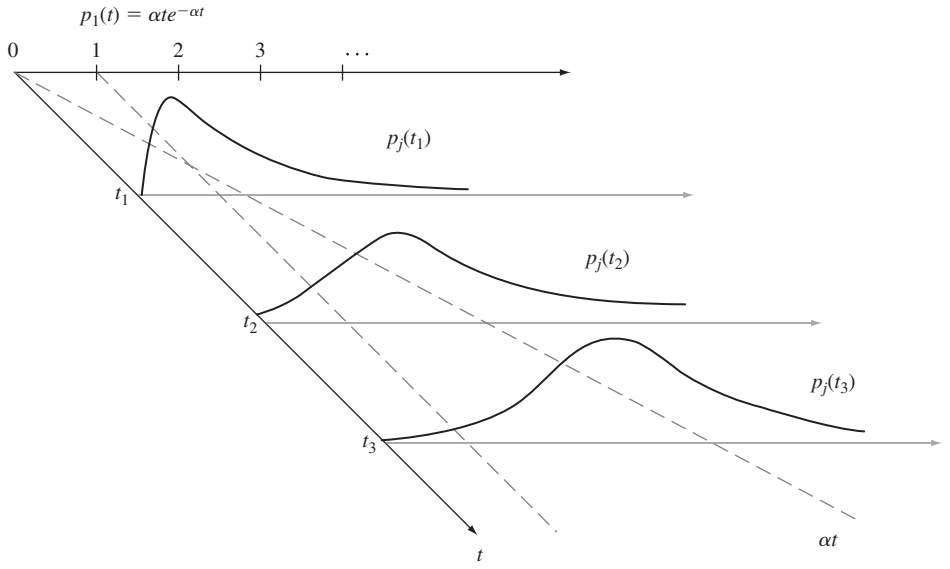

FIGURE 11.11
State pmf of Poisson process vs. time.

11.4.3 Steady State Probabilities and Global Balance Equations

As $t \to \infty$, the state probabilities in the two-state queueing system in Example 11.36 converge to a pmf that does not depend on the initial conditions. This is typical of systems that reach "equilibrium" or "steady state." For such a system, $p_j(t) \to p_j$ and $p_j'(t) \to 0$, so Eq. (11.39) becomes

$$0 = \sum_i \gamma_{ij} p_i \qquad \text{for all } j, \tag{11.40a}$$

or equivalently, recalling that $\gamma_{jj} = -v_j$,

$$v_j p_j = \sum_{i \neq j} \gamma_{ij} p_i \qquad \text{for all } j, \tag{11.40b}$$

where

$$\sum_j p_j = 1. \tag{11.40c}$$

Equation (11.40b) can be rewritten as follows:

$$p_j \left(\sum_{i \neq j} \gamma_{ji} \right) = \sum_{i \neq j} \gamma_{ij} p_i \tag{11.40d}$$

since

$$v_j = \sum_{i \neq j} \gamma_{ji}.$$

The system of linear equations given by Eq. (11.40b) or (11.40d) are called the **global balance equations**. These equations state that at equilibrium, the rate of probability flow out of state j, namely $v_j p_j$, is equal to the rate of flow into state j, as shown in Fig. 11.12. By solving this set of linear equations we can obtain the stationary state pmf of the system (when it exists).[5]

We refer to $\mathbf{p} = \{p_i\}$ as the **stationary state pmf** of the Markov chain. Since $\mathbf{p}$ satisfies Eq. (11.39), if we start the Markov chain with initial state pmf given by $\mathbf{p}$, then the state probabilities will be

$$p_i(t) = p_i \qquad \text{for all } t.$$

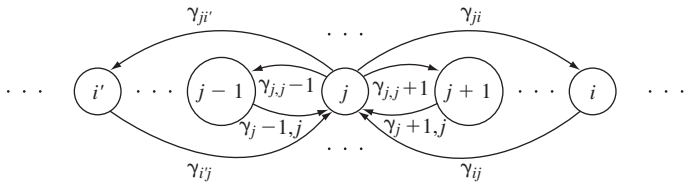

FIGURE 11.12
Global balance of probability flows.

[5]The last part of this section discusses conditions under which the stationary pmf exists.

The resulting process is a stationary random process as defined in Section 9.6 since the probability of the sequence of states $i_0, i_1, \ldots, i_n$ at times $t < t_1 + t < \cdots < t_n + t$ is, by Eq. (11.29),

$$\begin{aligned} P[X(t) = i_0, X(t_1 + t) = i_1, \ldots, X(t_n + t) = i_n] \\ = P[X(t_n + t) = i_n \,|\, X(t_{n-1} + t) = i_{n-1}] \cdots \\ \times P[X(t_1 + t) = i_1 \,|\, X(t) = i_0] P[X(t) = i_0]. \end{aligned}$$

The transition probabilities depend only on the difference between the associated times. Thus the above joint probability depends on the choice of origin only through $P[X(t) = i_0]$. But $P[X(t) = i_0] = p_{i_0}$ for all t. Therefore we conclude that the above joint probability is independent of the choice of time origin and thus that the process is stationary.

Example 11.38

Find the stationary state pmf for the two-state queueing system discussed in Example 11.36.

Equation (11.40b) for this system gives

$$\alpha p_0 = \beta p_1 \qquad \text{and} \qquad \beta p_1 = \alpha p_0.$$

Noting that $p_0 + p_1 = 1$, we obtain

$$p_0 = \frac{\beta}{\alpha + \beta} \qquad \text{and} \qquad p_1 = \frac{\alpha}{\alpha + \beta}.$$

Example 11.39 The M/M/1 Single-Server Queueing System

Consider a queueing system in which customers are served one at a time in order of arrival. The time between customer arrivals is exponentially distributed with rate λ, and the time required to service a customer is exponentially distributed with rate μ. Find the steady state pmf for the number of customers in the system.

The state transition rates are as follows. Customers arrive at a rate λ, so

$$\gamma_{i,i+1} = \lambda \qquad i = 0, 1, 2, \ldots.$$

When the system is nonempty, customers depart at the rate μ. Thus

$$\gamma_{i,i-1} = \mu \qquad i = 1, 2, 3, \ldots.$$

The transition rate diagram is shown in Fig. 11.13. The global balance equations are

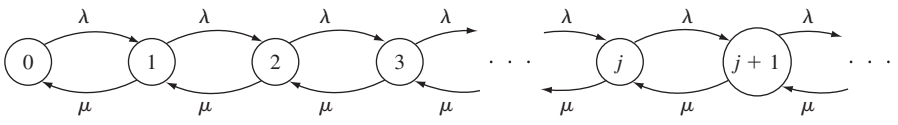

FIGURE 11.13
Transition rate diagram for M/M/1 queueing system.

$$\lambda p_0 = \mu p_1 \qquad \text{for } j = 0 \tag{11.41a}$$

$$(\lambda + \mu)p_j = \lambda p_{j-1} + \mu p_{j+1} \qquad \text{for } j = 1, 2, \ldots. \tag{11.41b}$$

We can rewrite Eq. (11.41b) as follows:

$$\lambda p_j - \mu p_{j+1} = \lambda p_{j-1} - \mu p_j \qquad \text{for } j = 1, 2, \ldots,$$

which implies that

$$\lambda p_{j-1} - \mu p_j = \text{constant} \qquad \text{for } j = 1, 2, \ldots. \tag{11.42}$$

Equation (11.42) with $j = 1$ and Eq. (11.41a) together imply that

$$\text{constant} = \lambda p_0 - \mu p_1 = 0.$$

Thus Eq. (11.42) becomes

$$\lambda p_{j-1} = \mu p_j,$$

or equivalently,

$$p_j = \rho p_{j-1} \qquad j = 1, 2, \ldots$$

and by a simple induction argument

$$p_j = \rho^j p_0,$$

where $\rho = \lambda/\mu$. We obtain p_0 by noting that the sum of the probabilities must be one:

$$1 = \sum_{j=0}^{\infty} p_j = (1 + \rho + \rho^2 + \cdots)p_0 = \frac{1}{1 - \rho} p_0,$$

where the series converges if and only if $\rho < 1$.

Thus

$$p_j = (1 - \rho)\rho^j \qquad j = 0, 1, 2, \ldots. \tag{11.43}$$

This queueing system is discussed in detail in Section 12.3.

The condition for the existence of a steady state solution has a simple explanation. The condition $\rho < 1$ is equivalent to

$$\lambda < \mu,$$

that is, the rate at which customers arrive must be less than the rate at which the system can process them. Otherwise the queue builds up without limit as time progresses.

Example 11.40 A Birth-and-Death Process

A **birth-and-death process** is a Markov chain in which only transitions between adjacent states occur as shown in Fig. 11.14. The single-server queueing system discussed in Example 11.39 is an example of a birth-and-death process.

The global balance equations for a general birth-and-death process are

$$\lambda_0 p_0 = \mu_1 p_1 \qquad j = 0 \tag{11.44a}$$

$$\lambda_j p_j - \mu_{j+1} p_{j+1} = \lambda_{j-1} p_{j-1} - \mu_j p_j \qquad j = 1, 2, \ldots. \tag{11.44b}$$

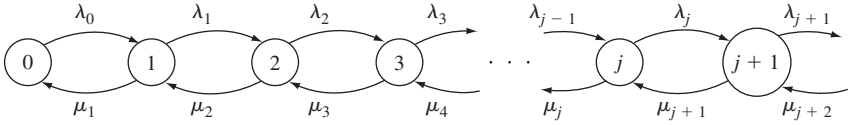

FIGURE 11.14
Transition rate diagram for general birth-and-death process.

As in the previous example, it then follows that

$$p_j = r_j p_{j-1} \qquad j = 1, 2, \ldots$$

and

$$p_j = r_j r_{j-1} \ldots r_1 p_0 \qquad j = 1, 2, \ldots, \tag{11.45}$$

where $r_j = (\lambda_{j-1})/\mu_j$. If we define

$$R_j = r_j r_{j-1} \ldots r_1 \qquad \text{and} \qquad R_0 = 1,$$

then p_0 is found from

$$1 = \left(\sum_{j=0}^{\infty} R_j \right) p_0.$$

If the series in the above equation converges, then the stationary pmf is given by

$$p_j = \frac{R_j}{\sum_{i=0}^{\infty} R_i}. \tag{11.46}$$

If the series does not converge, then a stationary pmf does not exist, and $p_j = 0$ for all j. In Chapter 12, we will see that many useful queueing systems can be modeled by birth-and-death processes.

11.4.4 Limiting Probabilities for Continuous-Time Markov Chains

We saw above that a continuous-time Markov chain $X(t)$ can be viewed as consisting of a sequence of states determined by some discrete-time Markov chain X_n with transition probabilities $\widetilde{q}_{ij}$ and a corresponding sequence of exponentially distributed state occupancy times. In this section we use this approach to investigate the limiting probabilities of continuous-time Markov chains.

First we consider the construction of stationary solutions for $X(t)$ from the steady state solutions of X_n. Suppose that the embedded Markov chain X_n is irreducible and positive recurrent, so that Eq. (11.25) holds. Let $N_i(n)$ denote the number of times state i occurs in the first n transitions, and let $T_i(j)$ denote the occupancy time the jth time state i occurs. The proportion of time spent by $X(t)$ in state i after the first n transitions is

$$\frac{\text{time spent in state } i}{\text{time spent in all states}} = \frac{\sum_{j=1}^{N_i(n)} T_i(j)}{\sum_i \sum_{j=1}^{N_i(n)} T_i(j)}$$

$$= \frac{\frac{N_i(n)}{n}\frac{1}{N_i(n)}\sum_{j=1}^{N_i(n)} T_i(j)}{\sum_i \frac{N_i(n)}{n}\frac{1}{N_i(n)}\sum_{j=1}^{N_i(n)} T_i(j)}. \tag{11.47}$$

As $n \to \infty$, by Eqs. (11.25) and (11.26ab), with probability one,

$$\frac{N_i(n)}{n} \to \pi_i, \tag{11.48}$$

the stationary pmf of the embedded Markov chain. In addition, we also have that $N_i(n) \to \infty$ as $n \to \infty$, so that by the strong law of large numbers, with probability one,

$$\frac{1}{N_i(n)} \sum_{j=1}^{N_i(n)} T_i(j) \to E[T_i] = 1/v_i, \tag{11.49}$$

where we have used the fact that the state occupancy time in state i has mean $1/v_i$. Similarly the denominator in Eq. (11.47) must approach $\left(\sum \pi_j/v_j\right)$. Equations (11.48) and (11.49) when applied to Eq. (11.47) imply that if $\sum \pi_j/v_j < \infty$, with probability one, the long-term proportion of time spent in state i approaches

$$p_i = \frac{\pi_i/v_i}{\sum_j \pi_j/v_j} = c\pi_i/v_i, \tag{11.50}$$

where π_j is the unique pmf solution to

$$\pi_j = \sum_i \pi_i \widetilde{q}_{ij} \qquad \text{for all } j \tag{11.51}$$

and c is a normalization constant.

We obtain the global balance equation, Eq. (11.40b), by substituting $\pi_i = v_i p_i/c$ from Eq. (11.50) and $\widetilde{q}_{ij} = \gamma_{ij}/v_i$ into Eq. (11.51):

$$v_j p_j = \sum_{i \neq j} p_i \gamma_{ij} \qquad \text{for all } j.$$

Thus the p_i's are the unique solution of the global balance equations.

We have proved the following result:

Theorem 3

Assume a time-continuous Markov chain, for which the embedded Markov chain is irreducible and positive recurrent with stationary pmf $\{\pi_j\}$ and $\sum_j \pi_j/v_j < \infty$, then the following assertions hold:

(i) $\lim_{t\to\infty} p_j(t) = p_j$ for all j;

(ii) The solution $\{p_i\}$ is unique and satisfies Eqs. (11.40bc);

(iii) For each j, p_j is the long-term proportion of time spent in state j.

Now assume that we know that the Markov chain is irreducible and that we have a solution $\{p_j\}$ to the global balance equations (11.40bc):

$$p_j v_j = \sum_{i\neq j} p_i \gamma_{ij}.$$

Substituting Eq. (11.50) into the above equation

$$c\pi_j = \left(\frac{c\pi_j}{v_j}\right) v_j = \sum_{i\neq j}\left(\frac{c\pi_i}{v_i}\right)\gamma_{ij} = c\sum_{i\neq j}\pi_i\left(\frac{\gamma_{ij}}{v_i}\right) = c\sum_{i\neq j}\pi_i \tilde{q}_{ij}$$

implies that the following choice of $\{\pi_j\}$ gives a solution for the stationary pmf of the embedded Markov chain:

$$\pi_j = \frac{p_j v_j}{\sum_i p_i v_i}.$$

Note that we must require that the denominator be finite. From Theorem 1 in Section 11.4, if there is a stationary pmf then it is unique and positive recurrent. Furthermore the construction of $\{\pi_j\}$ from the $\{p_j\}$ ensures that p_j is the long-term proportion of time in state j as well as the limiting state probability for $X(t)$.

We have shown the following theorem:

Theorem 4

Assume a time-continuous Markov chain, for which the embedded Markov chain is irreducible. Suppose that $\{p_j\}$ is a solution to the global balance equations (11.40bc), and that $\sum_j \pi_j v_j < \infty$, then the following assertions hold:

(i) The solution $\{p_i\}$ is unique;

(ii) $\lim_{t\to\infty} p_j(t) = p_j$ for all j;

(iii) For each j, p_j is the long-term proportion of time spent in state j;

(iv) The embedded Markov chain is positive recurrent.

Example 11.41

In the two-state system in Example 11.36,

$$[\widetilde{q}_{ij}] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The equation $\boldsymbol{\pi} = \boldsymbol{\pi}[\widetilde{q}_{ij}]$ implies that

$$\pi_0 = \pi_1 = \frac{1}{2}.$$

In addition, $v_0 = \alpha$ and $v_1 = \beta$. Thus

$$p_0 = \frac{1/2(1/\alpha)}{1/2(1/\alpha + 1/\beta)} = \frac{\beta}{\alpha + \beta}$$

and

$$p_1 = \frac{\alpha}{\alpha + \beta}.$$

---

*11.5 TIME-REVERSED MARKOV CHAINS

We now consider the random process that results when we play a Markov chain backwards in time. We will see that the resulting process is also a Markov chain and so develop another method for obtaining the stationary probabilities of the forward and reverse processes. The insights gained by looking at the reverse process prove useful in developing certain results in queueing theory in Chapter 12.

Let X_n be a stationary ergodic Markov chain[6] with one-step transition probability matrix $P = \{p_{ij}\}$ and stationary state pmf $\{\pi_j\}$. Consider the dependence of X_{n-1}, the "future" in the reverse process, on $X_n, X_{n+1}, \ldots, X_{n+k}$, the "present and past":

$$\begin{aligned} P[X_{n-1} = j \,|\, X_n = i, X_{n+1} = i_1, \ldots, X_{n+k} = i_k] &= \frac{P[X_{n-1} = j, X_n = i, X_{n+1} = i_1, \ldots, X_{n+k} = i_k]}{P[X_n = i, X_{n+1} = i_1, \ldots, X_{n+k} = i_k]} \\ &= \frac{\pi_j p_{ji} p_{i,i_1} \cdots p_{i_{k-1},i_k}}{\pi_i p_{i,i_1} \cdots p_{i_{k-1},i_k}} \\ &= \frac{\pi_j p_{ji}}{\pi_i} \\ &= P[X_{n-1} = j \,|\, X_n = i]. \end{aligned} \tag{11.52}$$

[6]That is, let it be an irreducible, aperiodic, stationary Markov chain.

The above equations show that *the **time-reversed process** is also a Markov chain with one-step transition probabilities*

$$P[X_{n-1} = j \,|\, X_n = i] = q_{ij} = \frac{\pi_j p_{ji}}{\pi_i}. \tag{11.53}$$

Since X_n is irreducible and aperiodic, its stationary state probabilities π_j represent the proportion of time that the state is in state j. This proportion of time does not depend on whether one goes forward or backward in time, so π_j must also be the stationary pmf for the reverse process. Thus *the forward and reverse process must have the same stationary pmf*.

Example 11.42

Suppose that a new light bulb is put in use at day $n = 0$, and suppose that each time a light bulb fails it is replaced the next day. Let X_n be the age of the light bulb (in days) at the end of day n. If a_i is the probability that the lifetime L of a light bulb is i days, then the probability that the light bulb fails on day j given that it has not failed up to then is

$$b_j = \frac{P[L = j]}{P[L \geq j]} = \frac{a_j}{\sum_{k=j}^{\infty} a_k} \qquad j = 1, 2, \ldots.$$

Thus the transition probabilities for X_n are

$$\begin{aligned} p_{i,i+1} &= 1 - b_i \qquad i = 1, 2, \ldots \\ p_{i1} &= b_i \qquad\quad i = 1, 2, \ldots \\ p_{ij} &= 0 \qquad\quad \text{otherwise.} \end{aligned}$$

Figure 11.15(a) shows the state transition diagram of X_n, and Fig. 11.16(a) shows a typical sample function that consists of a sawtooth-shaped function that increases linearly and then falls abruptly to one when a light bulb fails.

Figure 11.16(b) shows a sample function of the reverse process from which we deduce that the state transition diagram must be as shown in Fig. 11.15(b). The transition probabilities for the reverse process are obtained from Eq. (11.53):

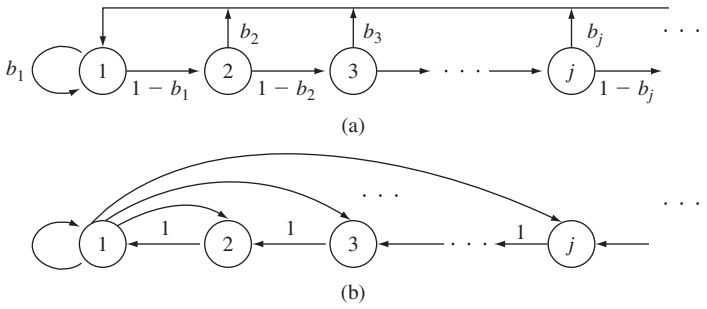

FIGURE 11.15
(a) Transition diagram for age of a renewal process. (b) Transition diagram for time-reversed process.

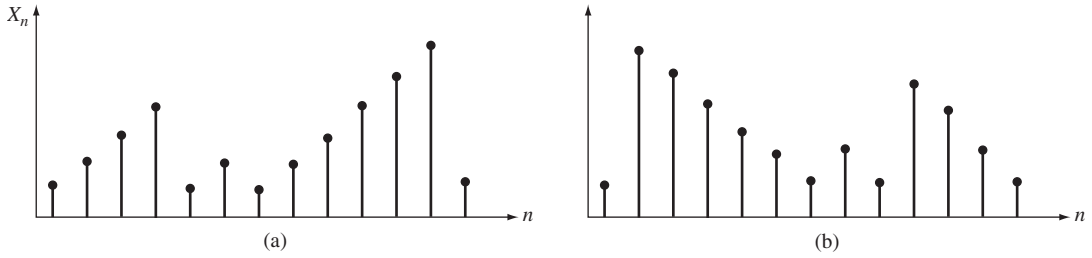

FIGURE 11.16
(a) Age of light bulb in use at time n. (b) Time-reversed process of X_n.

$$q_{i,i-1} = \frac{\pi_{i-1}}{\pi_i}(1 - b_{i-1}) \qquad i = 2, 3, 4, \ldots$$
$$q_{1,i} = \frac{\pi_i}{\pi_1} b_i \qquad i = 1, 2, \ldots$$
$$q_{i,j} = 0 \qquad \text{otherwise.}$$

For now we defer the problem of finding the stationary state probabilities π_j.

Example 11.42 shows that Eq. (11.53) provides us with conditions that must be satisfied by the stationary probabilities π_j. Suppose we were able to *guess* a pmf $\{\pi_j\}$ so that Eq. (11.53) holds, that is,

$$\pi_i q_{ij} = \pi_j p_{ji} \qquad \text{for all } i, j. \tag{11.54}$$

It then follows that $\{\pi_j\}$ is the stationary pmf. To see this, sum Eq. (11.54) over all j, then

$$\sum_j \pi_j p_{ji} = \pi_i \sum_j q_{ij} = \pi_i \qquad \text{for all } i. \tag{11.55}$$

But Eq. (11.55) is the condition for π_j to be the stationary pmf for the forward process, thus π_j is the stationary pmf. Equation (11.54) thus provides us with another method for finding the stationary pmf of a discrete-time Markov chain: *If we can guess a set of transition probabilities $q_{i,j}$ for the reverse process and a pmf π_j so that Eq. (11.54) is satisfied, then it follows that the π_j is the stationary pmf for the Markov chain and the $q_{i,j}$ are the transition probabilities for the reverse process.*

Example 11.43

The sample function of the reverse process in Example 11.42 suggests that for $i > 1$, the process moves from state i to state $i - 1$ with probability one; that is,

$$q_{i,i-1} = \frac{\pi_{i-1}(1 - b_{i-1})}{\pi_i} = 1,$$

which implies that

$$\begin{aligned}\pi_i &= (1 - b_{i-1})\pi_{i-1} \qquad i = 2, 3, \ldots \\ &= (1 - b_{i-1})(1 - b_{i-2}) \cdots (1 - b_1)\pi_1. \end{aligned} \tag{11.56}$$

However, from Example 11.42 for $i \geq 2$,

$$(1 - b_{i-1}) = 1 - \frac{a_{i-1}}{\sum_{k=i-1}^{\infty} a_k} = \frac{\sum_{k=i}^{\infty} a_k}{\sum_{k=i-1}^{\infty} a_k},$$

so in Eq. (11.56), the denominator of $(1 - b_{i-1})$ cancels the numerator of $(1 - b_{i-2})$, the denominator of $(1 - b_{i-2})$ cancels the numerator of $(1 - b_{i-3})$, and so on. Thus

$$\pi_i = \left\{ \sum_{k=i}^{\infty} a_k \right\} \pi_1 = P[L \geq i]\pi_1 \qquad i = 2, 3, \ldots .$$

We obtain π_1 by using the fact that the probabilities sum to one:

$$1 = \pi_1 \sum_{i=1}^{\infty} P[L \geq i] = \pi_1 E[L],$$

where we have used Eq. (4.29) for $E[L]$. Thus

$$\pi_i = \frac{P[L \geq i]}{E[L]} \qquad i = 1, 2, \ldots . \tag{11.57}$$

11.5.1 Time-Reversible Markov Chains

A stationary ergodic Markov chain is said to be **reversible** if the one-step transition probability matrix of the forward and reverse processes are the same, that is, if

$$q_{ij} = p_{ij} \qquad \text{for all } i, j. \tag{11.58}$$

Equations (11.53) and (11.58) together imply that a Markov chain is reversible if and only if

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \text{for all } i, j. \tag{11.59}$$

Since π_i and π_j are the long-term proportion of transitions out of states i and j, respectively, Eq. (11.59) implies that a chain is reversible if the proportion of transitions from i to j is equal to the proportion of transitions from j to i.

Example 11.44 Discrete-Time Birth-and-Death Process

Figure 11.17 shows the state transition diagram for a discrete-time birth-and-death process with transition probabilities

$$\begin{aligned} p_{00} &= 0 \qquad & p_{01} &= 1 = a_0 & \\ p_{i,i+1} &= a_i \qquad & p_{i,i-1} &= 1 - a_i \qquad & i = 1, 2, \ldots \\ p_{ij} &= 0 \qquad & \text{otherwise.} & & \end{aligned}$$

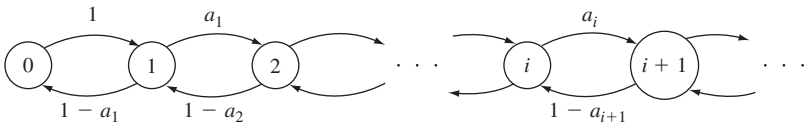

FIGURE 11.17
Transition diagram for a discrete-time birth-and-death process.

For any sample path, the number of transitions from i to $i+1$ can differ by at most 1 from the number of transitions from $i+1$ to i since the only way to return to i is through $i+1$. Thus the long-term proportion of transitions from i to $i+1$ is equal to that from $i+1$ to i. Since these are the only possible transitions, it follows that birth-and-death processes are reversible.

Equation (11.59) implies that

$$a_j\pi_j = (1-a_{j+1})\pi_{j+1} \qquad j = 0, 1, 2, \ldots,$$

which allows us to write all the π_j's in terms of π_0:

$$\pi_j = \left(\frac{a_{j-1}}{1-a_j}\right)\cdots\left(\frac{a_0}{1-a_1}\right)\pi_0 = \frac{a_{j-1}\cdots a_0}{(1-a_j)\cdots(1-a_1)}\pi_0 \triangleq R_j\pi_0. \tag{11.60}$$

The probability π_0 is found from $$1 = \pi_0\sum_{j=0}^{\infty} R_j. \tag{11.61}$$

The series in Eq. (11.61) must converge in order for π_j to exist.

11.5.2 Time-Reversible Continuous-Time Markov Chains

Now consider a stationary, continuous-time Markov chain played backward in time. If $X(t) = i$ (i.e., the process is in state i at time t), then the probability that the reverse process remains in state i for an additional s seconds is

$$\begin{aligned} P[X(t') = i, \quad t-s \le t' \le t \,|\, X(t) = i] &= \frac{P[X(t-s) = i, T_i > s]}{P[X(t) = i]} \\ &= \frac{P[X(t-s) = i]P[T_i > s]}{P[X(t) = i]} \\ &= P[T_i > s] = e^{-v_i s}, \end{aligned} \tag{11.62}$$

where $P[X(t-s) = i] = P[X(t) = i]$ because $X(t)$ is a stationary process, and where T_i is the time spent in state i for the forward process. Thus *the reverse process also spends an exponentially distributed amount of time with rate v_i in state i.*

The jumps in the forward process $X(t)$ are determined by the embedded Markov chain $\widetilde{q}_{ij}$, so the jumps in the reverse process are determined by the discrete-time Markov chain corresponding to the time-reversed embedded Markov chain given by Eq. (11.53):

$$q_{ij} = \frac{\pi_j\widetilde{q}_{ji}}{\pi_i}. \tag{11.63}$$

It follows that the transition rates for the time-reversed continuous-time process are given by

$$\begin{aligned}\gamma'_{ij} &= v_i q_{ij} = \frac{\pi_j v_i \widetilde{q}_{ji}}{\pi_i} \\ &= \frac{v_i \pi_j \gamma_{ji}}{\pi_i v_j} = \frac{p_j \gamma_{ji}}{p_i},\end{aligned} \tag{11.64}$$

where we used the fact that $\widetilde{q}_{ji} = \gamma_{ji}/v_j$ and $p_j = c\pi_j/v_j$. In comparing Eq. (11.64) to Eq. (11.53), note that the transition rates γ'_{ij} have simply replaced the transition probabilities q_{ij} in going from the discrete-time to the continuous-time case.

The discussion that led to Eq. (11.54) provides us with another method for determining the stationary pmf p_j of $X(t)$. *If we can guess a set of transition rates $\gamma'_{i,j}$ and a pmf p_j such that*

$$p_i \gamma'_{i,j} = p_j \gamma_{j,i} \qquad \text{for all } i, j \tag{11.65a}$$

and

$$\sum_{j \neq i} \gamma_{i,j} = \sum_{j \neq i} \gamma'_{i,j} \qquad \text{for all } i, \tag{11.65b}$$

then p_j is the stationary pmf for $X(t)$ and $\gamma'_{i,j}$ are the transition rates for the reverse process.

Since the state occupancy times in the forward and reverse processes are exponential random variables with the same mean, the continuous-time Markov chain $X(t)$ is reversible if and only if its embedded Markov chain is reversible. Equation (11.59) implies that the following condition must be satisfied:

$$\pi_i \widetilde{q}_{ij} = \pi_j \widetilde{q}_{ji} \qquad \text{for all } i, j, \tag{11.66}$$

where π_j is the stationary pmf of the embedded Markov chain. Recall from Eq. (11.50) that $\pi_j = cv_j p_j$, where p_j is the stationary pmf of $X(t)$. Substituting into Eq. (11.66), we obtain

$$p_i v_i \widetilde{q}_{ij} = p_j v_j \widetilde{q}_{ji},$$

which is equivalent to

$$p_i \gamma_{ij} = p_j \gamma_{ji}. \tag{11.67}$$

Thus we conclude that $X(t)$ *is reversible if and only if Eq. (11.67) is satisfied*. As in the discrete-time case, Eq. (11.67) can be interpreted as stating that the rate at which $X(t)$ goes from state i to state j is equal to the rate at which $X(t)$ goes from state j to state i.

Example 11.45 Continuous-Time Birth-and-Death Process

Consider the general continuous-time birth-and-death process introduced in Example 11.40. The embedded Markov chain in this process is a discrete-time birth-and-death process of the type discussed in Example 11.44. It therefore follows that all continuous-time birth-and-death processes are time-reversible.

In Chapter 12 we will see that the time reversibility of certain Markov chains implies some remarkable properties about the departure processes of queueing systems.

11.6 NUMERICAL TECHNIQUES FOR MARKOV CHAINS

In this section we present several numerical techniques that are useful in the analysis of Markov chains. The first part of the section presents methods for finding the stationary as well as transient solutions for the state probabilities of Markov chains. The second part of the section addresses the simulation of discrete-time and continuous-time Markov chains.

11.6.1 Stationary Probabilities of Markov Chains

The most basic calculation with *finite-state discrete-time Markov chains* involves finding their stationary state probabilities. To do so, we consider the equation:

$$\boldsymbol{\pi} = \boldsymbol{\pi} P \quad \text{or equivalently} \quad \mathbf{0} = \boldsymbol{\pi}(P - I). \tag{11.68a}$$

In general the above set of linear equations is undetermined. To see this, note that the sum of the columns of the matrix $P - I$ is zero. Therefore we need the normalization equation: $\pi_1 + \pi_2 + \cdots + \pi_K = 1$. We can incorporate this equation by replacing one of the columns of $P - I$ with the all 1's column vector. Let Q be the matrix that results when we replace the first column of $P - I$; the system of linear equations becomes:

$$\mathbf{b} = \boldsymbol{\pi} Q, \tag{11.68b}$$

where $\mathbf{b}$ is a row vector with 1 in the first entry and zeros elsewhere. If the Markov chain is irreducible, then a unique stationary pmf exists and is obtained by inverting the above equation.

Example 11.46 Google PageRank

Find the stationary pmf for the PageRank algorithm in Example 11.30.

After we take $P - I$ from the example and replace the first column with all 1's we obtain:

$$Q = \begin{bmatrix} 1 & 0.4550 & 0.4550 & 0.0300 & 0.0300 \\ 1 & -0.8000 & 0.2000 & 0.2000 & 0.2000 \\ 1 & 0.3133 & -0.9700 & 0.3133 & 0.0300 \\ 1 & 0.0300 & 0.0300 & -0.9700 & 0.8800 \\ 1 & 0.0300 & 0.4550 & 0.0300 & -0.5450 \end{bmatrix}.$$

We then invert Q to obtain the pmf:

$$\boldsymbol{\pi} = (0.13175, 0.18772, 0.24642, 0.13172, 0.30239).$$

The Octave commands for the above procedure are given below:

```
> Q=[1 0.455 0.455 0.03 0.03
> 1 −.8 .2 .2 .2
> 1 0.3133 −.97 0.3133 0.03
> 1 0.03 0.03 −0.97 0.88
```

```
> 1 0.03 0.455 0.03 −.545];
> b=[1 0 0 0 0];
> p=b*inv(Q)
p =
  0.13175 0.18772 0.24642 0.13172 0.30239
```

In the case of infinite-state Markov chains, we can apply matrix inversion by truncating the state space at some value where the state probabilities become negligible. Another method, discussed in the next chapter, involves the application of the probability generating function for the state of the system.

To find the stationary pmf for *finite-state continuous-time Markov chains*, we need to find a pmf that satisfies Eq. (11.40a) as well as the normalization condition:

$$\mathbf{0} = \mathbf{p}\Gamma \quad \text{and} \quad 1 = \mathbf{pe} \tag{11.69a}$$

where

$$\Gamma = \begin{bmatrix} -\nu_0 & \gamma_{01} & \gamma_{02} & \gamma_{03} \\ \gamma_{10} & -\nu_1 & \cdots & \gamma_{1K-1} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{K-1\,0} & \gamma_{K-1\,1} & \cdots & -\nu_{K-1} \end{bmatrix} \quad \text{and} \quad \mathbf{e} = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix}. \tag{11.69b}$$

The columns of Γ sum to zero, so as before we need to replace a column of Γ with $\mathbf{e}$. We obtain $\mathbf{p}$ by multiplying $\mathbf{b}$ by the inverse of the resulting matrix.

Example 11.47 Cartridge Inventory

An office orders laser printer cartridges in batches of four cartridges. Suppose that each cartridge lasts for an exponentially distributed time with mean 1 month. Assume that a new batch of four cartridges becomes available as soon as the last cartridge in a batch runs out. Find the stationary pmf for $N(t)$, the number of cartridges available at time t.

$N(t)$ takes on values from the set $\{1, 2, 3, 4\}$ and follows a periodic sequence of values $4 \to 3 \to 2 \to 1 \to 4 \ldots$. The rate out of each state is 1 and the rate into each state from the previous state is also 1. Therefore the transition rate matrix and the modified global balance equations are:

$$\Gamma = \begin{bmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \qquad \mathbf{b} = \mathbf{p}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & 0 & 1 & -1 \end{bmatrix}.$$

It is easy to show that the $\mathbf{p} = (1/4, 1/4, 1/4, 1/4)$. In a more complicated case we would use numerical inversion to solve for $\mathbf{p}$.

11.6.2 Time-Dependent Probabilities of Markov Chains

We now consider finding the time-dependent probabilities of a *finite-state discrete-time Markov chain* as given by Eq. (8.16b). Example 11.9 described the general approach for finding P^n. First, however, we note a few facts about the transition probability matrix P.

A **stochastic matrix** is defined as a nonnegative matrix for which the elements of each row add to one. Thus P is a stochastic matrix. A stochastic matrix always has $\lambda = 1$ as an eigenvalue and $\mathbf{e}^T = (1, \ldots, 1)$ as a right eigenvector: $1\mathbf{e} = P\mathbf{e}$. This follows from the fact that all the row elements of P add to one. On the other hand, the stationary pmf $\boldsymbol{\pi}$ is a left eigenvector for the $\lambda = 1$ eigenvalue of P: $1\boldsymbol{\pi} = \boldsymbol{\pi}P$. It can be shown [Gallager, pp. 116–117] that if P corresponds to an aperiodic irreducible Markov chain, then $\lambda = 1$ is the largest eigenvalue and the magnitude of all other eigenvalues are less than 1.

Let P correspond to an aperiodic irreducible Markov chain. Proceeding as in Example 11.19, to find P^n we first find the eigenvalues $1 = \lambda_1 > |\lambda_2| > \ldots > |\lambda_K|$ and right eigenvectors of P: $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K$. Letting $\mathbf{E}$ be the matrix with eigenvectors as columns, we then have that:

$$\begin{aligned} P^n &= \mathbf{E}\Lambda^n\mathbf{E}^{-1} \\ &= \mathbf{E}\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & \lambda_2^n & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & \lambda_K^n \end{bmatrix}\mathbf{E}^{-1}. \end{aligned} \tag{11.70}$$

Note how all but the 1-1 entry in the diagonal matrix approach zero as n increases. Note as well that the first column of $\mathbf{E}$ is the all 1's vector. This implies that the first row of $\mathbf{E}^{-1}$ contains the stationary pmf $\boldsymbol{\pi}$. In Octave the eigenvalues and eigenvectors of P are obtained using the `eig(P)` function, which was discussed previously in Section 10.7. In practice it is simpler and more convenient to use the command `P^n`.

Next we consider finding the time-dependent probabilities of a *finite-state continuous-time Markov chain* that are the solution to Eq. (11.39):

$$\mathbf{p}'(t) = [p_j'(t)] = \sum_{i=1}^{K} p_i(t)\gamma_{ij} = \mathbf{p}(t)\Gamma \quad \text{subject to} \quad \mathbf{p}(0) = (p_i(0), \ldots, p_K(0)). \tag{11.71}$$

We are now dealing with first-order vector differential equations. Electrical engineering students encounter this equation in an introductory linear systems course. The solution is given by:

$$\mathbf{p}(t) = \mathbf{p}(0)P(t) = \mathbf{p}(0)e^{\Gamma t} \tag{11.72a}$$

where $P(t) = e^{\Gamma t}$ is the matrix of transition probabilities in an interval of length t seconds, and where the exponential matrix function is defined by:

$$P(t) = [p_{ij}(t)] = e^{\Gamma t} = \sum_{j=0}^{\infty} \frac{(\Gamma t)^j}{j!}. \tag{11.72b}$$

Furthermore, using matrix diagonalization the exponential matrix can be evaluated as:

$$P(t) = \mathbf{E}[e^{\lambda_i t}]\mathbf{E}^{-1} \tag{11.72c}$$

where $\mathbf{E}$ is a matrix whose columns are the eigenvectors of Γ and the middle matrix is a diagonal matrix with exponential functions as its elements. [Gallager, p. 194] shows

that if the Markov chain is finite state and irreducible, then Γ has an eigenvalue $\lambda = 0$ which has right eigenvector $\mathbf{e}^T = (1, 1,...,1)$. Γ also has a left eigenvector $\mathbf{p}$ corresponding to $\lambda = 0$ which is the unique stationary state pmf. Furthermore the remaining eigenvalues of Γ have negative real parts. This implies that all but the $\lambda = 0$ exponential terms in the diagonal matrix decay to zero as t increases. If we let $\lambda = 0$ occupy the 1-1 entry in the diagonal matrix, then as $t \to \infty$, $P(t)$ approaches the product of the $\mathbf{e}$ and the first row of $\mathbf{E}^{-1}$.

Example 11.48 Cartridge Inventory

Find the state probabilities for $N(t)$ in Example 11.47 if $N(0) = 4$.

We use the `eig(Γ)` function to obtain the eigenvalues and eigenvectors of Γ and the associated matrices, $\mathbf{E}$, Λ, and $\mathbf{E}^{-1}$:

$$\mathbf{E} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{j} & -j & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -j & j & -1 \end{bmatrix} \quad \Lambda = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -1-j & 0 & 0 \\ 0 & 0 & -1+j & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix} \quad \mathbf{E}^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{j} & -1 & j \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

Note that two of the eigenvalues and their corresponding eigenvectors are complex. The state probabilities are given by:

$$\begin{aligned} \mathbf{p}(t) &= \mathbf{p}(0)\mathbf{E}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{-(1+j)t} & 0 & 0 \\ 0 & 0 & e^{-(1-j)t} & 0 \\ 0 & 0 & 0 & e^{-2t} \end{bmatrix}\mathbf{E}^{-1} \\ &= (0,0,0,1)\frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -j & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -j & j & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{-(1+j)t} & 0 & 0 \\ 0 & 0 & e^{-(1-j)t} & 0 \\ 0 & 0 & 0 & e^{-2t} \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \end{bmatrix} \\ &= \frac{1}{4}(1,-j,j,-1)\begin{bmatrix} 1 & 1 & 1 & 1 \\ e^{-(1+j)t} & -je^{-(1+j)t} & -e^{-(1+j)t} & je^{-(1+j)t} \\ e^{-(1-j)t} & je^{-(1-j)t} & -e^{-(1-j)t} & -je^{-(1-j)t} \\ e^{-2t} & -e^{-2t} & e^{-2t} & -e^{-2t} \end{bmatrix} \\ &= \frac{1}{4}(1 - 2e^{-t}\sin t - e^{-2t}, 1 - 2e^{-t}\cos t + e^{-2t}, 1 + 2e^{-t}\sin t - e^{-2t}, 1 + 2e^{-t}\cos t + e^{-2t}). \end{aligned}$$

Figure 11.18 shows the four state probabilities vs. time. It can be seen that all of the probability mass is initially in state 4 and that the mass first transfers to state 3, then state 2, and finally to state 1. Eventually all state probabilities approach the steady state value of 1/4.

11.6.3 Simulation of Markov Chains

We simulate a Markov chain by emulating its underlying random experiments. We begin by selecting the initial state according to an initial state pmf. We then generate the sequence of states by producing outcomes according to the associated transition

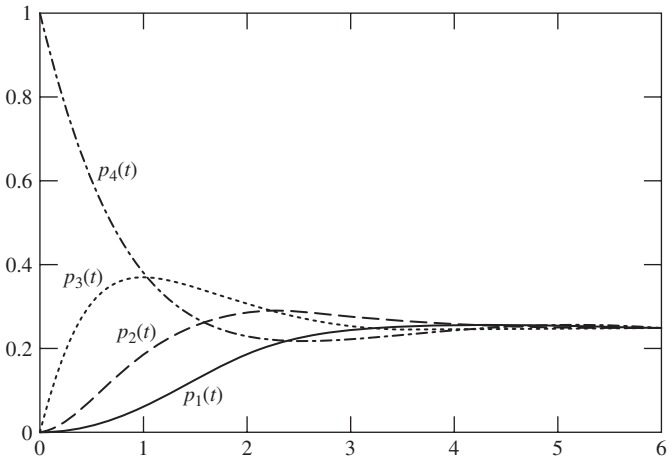

FIGURE 11.18
Time-dependent probabilities in cartridge inventory.

probabilities. In the case of continuous-time Markov chains we also need to generate a state occupancy time after each state transition has been determined. Figure 11.19 shows the inputs and outputs of generic modules for generating realizations of a Markov chain.

Discrete-Time Markov Chains The module for generating a sequence of states for a Markov chain requires the following inputs: i. The state space; ii. The matrix of state transition probabilities; iii. The initial state probability mass function; and iv. The number of steps in the simulation sequence. The module operates as follows:

1. Generates the initial state according to π_0.
2. Repeatedly generates the next state according to the transition probabilities of the current state.
3. Stops when the required number of steps has been simulated.

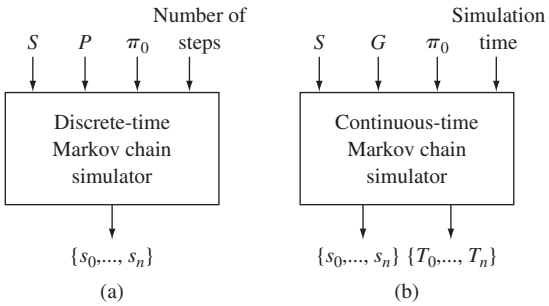

FIGURE 11.19
Generic modules for simulating Markov chains.

Example 11.49 Discrete-Time Markov Chain

Develop a program to generate Markov chains with the state transition diagram as shown in Fig. 11.20(a). Note that the Markov chain is similar to that of a birth-death process except that transitions from a state to itself are allowed. Use the program to simulate 1000 time steps in a data multiplexer where in each time unit a packet is received with probability a, and/or a packet transmitted from its buffer with probability b. Assume the data multiplexer is initially empty.

For this example we wrote the function `Discrete_MC(Nmax,P,IC,L)`. The state space is $\{0, 1, \ldots, N_{max}\}$. Since Octave uses indices from 1 onwards, the array state ranges from 1 to $N_{max} + 1$. For the Markov chains under consideration we need to specify only three probabilities for the transition probabilities for each state. Therefore P is an $N_{max} + 1$ row by 3 column matrix. The initial state pmf is a $N_{max} + 1$ by 1 vector. The output of the function is a vector of states of size L.

The Markov chain for the data multiplexer has the following transition probabilities. If $N = 0$, that is, the system is empty, the next state is either $N = 1$ with probability a, or $N = 0$ with probability $1 - a$, that is: $p_{00} = 1 - a$, $p_{01} = a$. If $N = n > 0$, the next state is $n + 1$ with probability $(1 - b)a$; n with probability ab; or $n - 1$ with probability $b(1 - a)$, that is: $p_{n\,n+1} = (1 - b)a$, $p_{n\,n} = ab$, $p_{n-1\,n} = (1 - a)b$. If $N = N_{max}$, the next state is $N_{max} - 1$ with probability $(1 - a)b$; or N_{max} with probability $1 - b(1 - a)$, since the system is not allowed to grow beyond N_{max}.

The code below prepares the inputs and then calls the function `Discrete_MC(S, P,IC,N)`. The basic step in the function involves generating a discrete random variable that determines whether the chain increases by 1, decreases by 1, or remains the same.

```
Nmax=50;
P=zeros(Nmax+1,3);
a=0.45;
b=0.50;
P(1,:)=[0,1-a,a];
r=[(1-a)*b,a*b+(1-a)*(1-b),(1-b)*a];
for n=2:Nmax;
    P(n,:)=r;
end
```

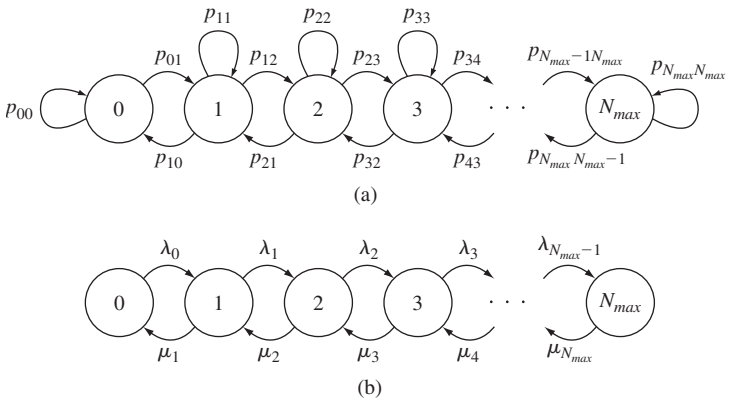

FIGURE 11.20
Generic Markov chains: (a) discrete-time; (b) birth-death continuous-time.

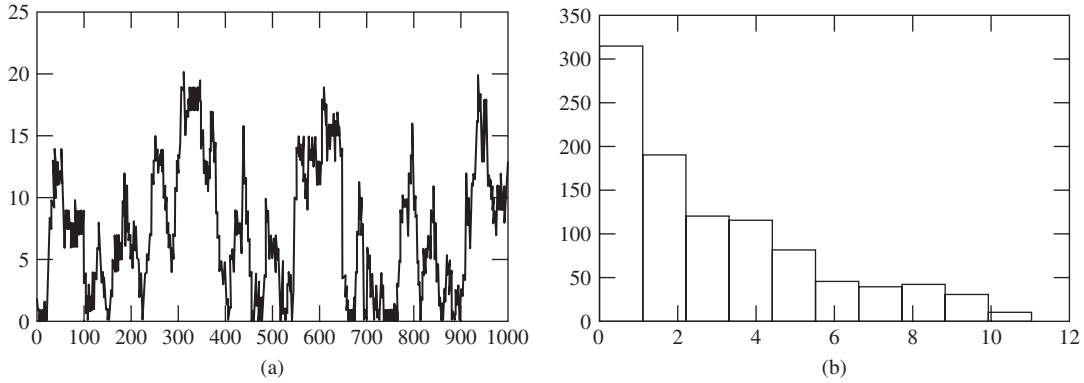

FIGURE 11.21
(a) Simulation of discrete-time data multiplexer; (b) histogram of number of packets in data multiplexer.

```
P(Nmax+1,:)=[(1-a)*b,1-(1-a)*b,0];
IC=zeros(Nmax+1,1);
IC(1,1)=1;
L=1000
Seq=Discrete_MC(Nmax,P,IC,L);
plot(Seq-1)

function stseq = Discrete_MC(Nmax,P,IC,L)
stseq=zeros(1,L);
s=[1:Nmax+1];
step=[-1,0,1];
InitSt=discrete_rnd(1,s,IC);
stseq(1)=InitSt;
for n=2:L+1;
    nextst=stseq(n-1)+discrete_rnd(1,step,P(stseq(n-1),:));
    stseq(n)=nextst;
end
```

Figure 11.21(a) shows a graph of a 1000-step realization of the Markov chain. The parameters in the simulation are $a = 0.45$ and $b = 0.5$. The latter parameter implies that a packet requires two time units on average of service before it departs the system. During the two time units that it takes to service the above packet, $2 \times (0.45) = 0.9$ packets arrive on average. This is an example of a "heavy traffic" situation which is characterized by the sporadic but sustained buildups of packets seen in the simulation. Figure 11.21(b) shows the histogram of the state occurrences in the simulation. It can be seen that the probability mass is concentrated at the lower state values.

Continuous-Time Markov Chains The module for generating a sequence of states for a continuous-time Markov chain requires the following inputs: i. The state space; ii. The

matrix of state transition rates; iii. The initial state probability mass function; and iv. The duration of the simulation. The module operates as follows:

1. Generates the initial state according to π_0.
2. Repeatedly generates the next state using the transition probabilities from the current state, and the state occupancy times for the new state.
3. Stops when the elapsed time has been simulated.

Example 11.50 Continuous-Time Birth-Death Process

Develop a program to generate continuous-time Markov chains with the state transition diagram shown in Figure 11.20(b). Use the program to simulate 1000 seconds of an M/M/1 queueing system. Assume the system is initially empty.

For this example we wrote the function `Continuous_MC(S,G,IC,T)`, given below. The module uses the embedded Markov chain approach and sequentially generates next state and occupancy time pairs. The transition probabilities for the embedded Markov chain are $\{\widetilde{q}_{jj-1} = \mu_j/(\lambda_j + \mu_j), \widetilde{q}_{jj+1} = \lambda_j/(\lambda_j + \mu_j)\}$ and the mean occupancy times are exponential random variables with mean $\{1/(\lambda_j + \mu_j)\}$. The basic step involves generating a binary random variable that determines whether the chain increases or decreases by 1, and then determines the occupancy time in the resulting state.

```
function [stseq,OccTime,n] = Continuous_MC(Nmax,G,IC,T)
Taggr=-1;
L=T*(G(Nmax-1,1)+G(Nmax-1,2));  % Estimate max number of state transitions.
stseq=zeros(1,L);
OccTime=zeros(1,L);
Q=zeros(1,2);
s=[1:Nmax+1];
step=[-1,1];
InitSt=discrete_rnd(1,s,IC);
stseq(1)=InitSt;
n=1;
OccTime(n)=exponential_rnd(G(stseq(n),1)+G(stseq(n),2));
Taggr=OccTime(n);
while (Taggr < T);
    n=n+1;
    Q(stseq(n-1),:)=[G(stseq(n-1),1),G(stseq(n-1),2)]/(G(stseq(n-
1),1)+G(stseq(n-1),2));
    nextst=stseq(n-1)+discrete_rnd(1,step,Q(stseq(n-1),:));
    stseq(n)=nextst;
    OccTime(n)=exponential_rnd((G(stseq(n),1)+G(stseq(n),2)));
    Taggr=Taggr+OccTime(n);
End
```

Figure 11.22 shows a graph of a realization of the Markov chain. The simulated queueing system has an arrival rate of $\lambda = 0.9$ jobs/second and a mean job service time of $\mu = 1$ second. Therefore the system is operating in heavy traffic and experiences surges in job backlogs. The

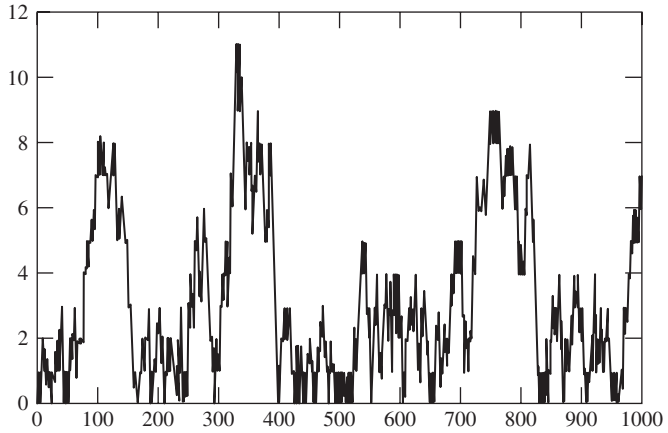

FIGURE 11.22
Simulation of M/M/1 continuous-time Markov chain.

calculation of the proportion of time that the system spends in each state is more complicated than for discrete-time systems because the occupancy times must be taken into account. These calculations will be addressed in the next chapter.

SUMMARY

- A random process is said to be Markov if the future of the process, given the present, is independent of the past.
- A Markov chain is an integer-valued Markov process.
- The joint pmf for a Markov chain at several time instants is equal to the product of the probability of the state at the first time instant and the probabilities of the subsequent state transitions (Eq. 11.3).
- For discrete-time Markov chains: (1) the n-step transition probability matrix $P(n)$ is equal to P^n, where P is the one-step transition probability; (2) the state probability after n steps $\mathbf{p}(n)$ is equal to $\mathbf{p}(0)P^n$, where $\mathbf{p}(0)$ is the initial state probability; and (3) P^n approaches a constant matrix as $n \to \infty$ for Markov chains that settle into steady state.
- The states of a discrete-time Markov chain can be divided into disjoint classes. The long-term behavior of a Markov chain is determined by the properties of its classes. In particular, for ergodic Markov chains the stationary state probabilities represent the long-term proportion of time spent in each state.
- A continuous-time Markov chain can be viewed as consisting of a discrete-time embedded Markov chain that determines the state transitions and of exponentially distributed state occupancy times.
- For continuous-time Markov chains: (1) the state probabilities and the transition probability matrix can be found by solving Eq. (11.39); (2) the steady state

probabilities can be found by solving the global balance equation, Eq. (11.40b) or (11.40c).

- A continuous-time Markov chain has a steady state if its embedded Markov chain is irreducible and positive recurrent with unique stationary pmf given by the solution of the global balance equations.
- The time-reversed version of a Markov chain is also a Markov chain. A discrete-time (continuous-time) irreducible, stationary ergodic Markov chain is reversible if the transition probability matrix (transition rate matrix) for the forward and reverse processes is the same.
- Matrix numerical methods can be used to find the time-dependent and the stationary probabilities of Markov chains.

CHECKLIST OF IMPORTANT TERMS

Accessible state
Birth-and-death process
Chapman–Kolmogorov equations
Class of states
Embedded Markov chain
Ergodic Markov chain
Global balance equations
Homogeneous transition probabilities
Irreducible Markov chain
Markov chain
Markov process
Markov property
Mean recurrence time
Null recurrent state
Period of a state/class
Positive recurrent state
Recurrent state/class
Reversible Markov chain
State
State occupancy time
State probabilities
Stationary state pmf
Stochastic matrix
Time-reversed Markov chain
Transient state/class
Transition probability matrix
Trellis diagram

ANNOTATED REFERENCES

References [1] and [2] contain very good discussions of discrete-time Markov chains. Feller has a rich set of classic examples that are a pleasure to read. Reference [3] gives a concise but quite complete introduction to Markov chains. Reference [4] provides an introduction to discrete-time and continuous-time Markov chains at about the same level as this chapter. References [6] and [7] give a more rigorous and complete coverage of Markov chains and processes.

1. K. L. Chung, *Elementary Probability Theory with Stochastic Processes*, Springer-Verlag, New York, 1975.
2. W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, Wiley, New York, 1968.
3. Y. A. Rozanov, *Probability Theory: A Concise Course*. Dover Publications, New York, 1969.
4. S. M. Ross, *Introduction to Probability Models*, Academic Press, Orlando, FL, 2003.

5. S. M. Ross, *Stochastic Processes,* Wiley, New York, 1983.
6. D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, Chapman and Hall, London, 1972.
7. R. G. Gallager, *Discrete Stochastic Processes*, Kluwer Academic Press, Boston, 1996.
8. J. Kohlas, *Stochastic Methods of Operations Research*, Cambridge University Press, London, 1982.
9. H. Anton, *Elementary Linear Algebra*, Wiley, New York, 1981.
10. A. M. Langville and C. D. Meyer, *Google's PageRank and Beyond*, Princeton University Press, Princeton, NJ, 2006.

PROBLEMS

Section 11.1: Markov Processes

11.1. Let M_n denote the sequence of sample means from an iid random process X_n:

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

(a) Is M_n a Markov process?

(b) If the answer to part a is yes, find the following state transition pdf:

$$f_{M_n}(x \mid M_{n-1} = y).$$

11.2. An urn initially contains five black balls and five white balls. The following experiment is repeated indefinitely: A ball is drawn from the urn; if the ball is white, it is put back in the urn, otherwise it is left out. Let X_n be the number of black balls remaining in the urn after n draws from the urn.

(a) Is X_n a Markov process? If so, find the appropriate transition probabilities and the corresponding trellis diagram.

(b) Do the transition probabilities depend on n?

(c) Repeat part a if the urn initially has K black balls and K white balls.

11.3. An urn initially contains two black balls and two white balls. The following experiment is repeated indefinitely: A ball is drawn from the urn; with probability a, the color of the ball is changed to the other color and is then put back in the urn, otherwise it is put back without change. Let X_n be the number of black balls in the urn after n draws from the urn.

(a) Is X_n a Markov process? If so, find the appropriate transition probabilities.

(b) Do the transition probabilities depend on n?

(c) Repeat part a if $a = 1$. What changes?

(d) Repeat parts a and c if the urn contains K black balls and K white balls.

11.4. Michael and Marisa initially have four pens each. Out of the total of eight pens, half are good and half are dry. The following experiment is repeated indefinitely: Michael and Marisa exchange a randomly selected pen from their set. Let X_n be the number of good pens in Marisa's set after n draws.

(a) Is X_n a Markov process? If so, find the appropriate transition probabilities.

(b) Do the transition probabilities depend on n?

(c) Repeat part a if Michael and Marisa initially have a total of K good pens and K dry pens.

11.5. Does a Markov process have independent increments? *Hint:* Use the process in Problem 11.2 to support your answer.

11.6. Let X_n be the Bernoulli iid process, and let Y_n be given by

$$Y_n = X_n + X_{n-1}.$$

It was shown in Example 11.2 that Y_n is not a Markov process. Consider the vector process defined by $\mathbf{Z}_n = (X_n, X_{n-1})$.

(a) Show that $\mathbf{Z}_n$ is a Markov process.

(b) Find the state transition diagram for $\mathbf{Z}_n$.

11.7. **(a)** Show that the following autoregressive process is a Markov process:

$$Y_n = rY_{n-1} + X_n \qquad Y_0 = 0,$$

where X_n is an iid process.

(b) Find the transition pdf if X_n is an iid Gaussian sequence.

11.8. The amount of water in an aquifer at year end is a random variable X_n. The amount of water drawn from the aquifer in a year is a random variable D_n and the amount restored by rainfall is W_n.

(a) Find a set of equations to describe the total amount of water X_n in the aquifer over time.

(b) Under what conditions is X_n a Markov process?

Section 11.2: Discrete-Time Markov Chains

11.9. Let X_n be an iid integer-valued random process. Show that X_n is a Markov process and give its one-step transition probability matrix.

11.10. An information source generates iid bits for X_n for which $P[0] = a = 1 - P[1]$.

(a) Suppose that X_n is transmitted over a binary symmetric channel with error probability ε. Find the probabilities of the outputs of the channel.

(b) Suppose that X_n is transmitted over K consecutive identical and independent binary symmetric channels. Does the sequence of channel outputs form a Markov chain?

(c) Find the K-step transition probabilities that relate the input bits from the source to the outputs of the Kth channel.

(d) What are the probabilities of the outputs of the Kth channel as $K \to \infty$?

11.11. Each time unit a data multiplexer receives a packet with probability a, and/or transmits a packet from its buffer with probability b. Assume that the multiplexer can hold at most N packets. Let X_n be the number of packets in the multiplexer at time n.

(a) Show that the system can be modeled by a Markov chain.

(b) Find the transition probability matrix P.

(c) Find the stationary pmf.

11.12. Let X_n be the Markov chain defined for the urn experiment in Problem 11.2.

(a) Find the one-step transition probability matrix P for X_n.

(b) Find the two-step transition probability matrix P^2 by matrix multiplication. Check your answer by computing $p_{54}(2)$ and comparing it to the corresponding entry in P^2.

(c) What happens to X_n as n approaches infinity? Use your answer to guess the limit of P^n as $n \to \infty$.

11.13. Let X_n be the Markov chain defined in Problem 11.3.

(a) Find the one-step transition probability matrix P for X_n with $a = 1/10$.

(b) Find P^2, P^4, and P^8 by matrix multiplication.

(c) What happens to X_n as n approaches infinity?

(d) Repeat parts a, b, and c if $a = 1$.

11.14. In the Ehrenfest model of heat exchange, two containers hold a total of ρ particles [Feller, pp. 121]. Each time instant a particle is selected at random and moved to the other container. Let X_n be the number of particles in the first container.

(a) Show that this model is the same as in Problem 11.3(d).

(b) Use the state transition diagram to explain why the model exhibits a "central force."

(c) Show that the stationary pmf is given by a binomial pmf with parameters ρ and 1/2. Give an intuitive explanation for this result.

11.15. Let X_n be the pen-exchange Markov chain defined in Problem 11.4.

(a) Find P.

(b) Use Octave or a numerical program to find P^2, P^4, and P^8 by matrix multiplication.

(c) What happens to X_n as n approaches infinity?

11.16. In the Bernoulli–Laplace model for diffusion, a total of 2ρ particles are distributed between two containers, and half of the particles are black and half are white [Feller, 1968, pp. 378]. Each time instant a particle is selected at random from each container and moved to the other container. Let X_n be the number of white particles in the first container.

(a) Show that this model is the same as in Problem 11.4(c).

(b) Show that the stationary pmf is given by:

$$\pi_j = \binom{\rho}{j}^2 \Big/ \binom{2\rho}{\rho} \text{ for } j = 0, 1, \ldots, \rho.$$

11.17. The vector process $\mathbf{Z}_n$ in Problem 11.6 has four possible states, so in effect it is equivalent to a Markov chain with states $\{0, 1, 2, 3\}$.

(a) Find the one-step transition probability matrix P.

(b) Find P^2 and check your answer by computing the probability of going from state $(0, 1)$ to state $(0, 1)$ in two steps.

(c) Show that $P^n = P^2$ for all $n > 2$. Give an intuitive justification for why this is true for this random process.

(d) Find the steady state probabilities for the process.

11.18. Consider a sequence of Bernoulli trials with probability of success p and let X_n denote the number of consecutive successes in a streak up to time n.

(a) Show that X_n is a Markov chain.

(b) Find the one-step transition probability and draw the corresponding state transition diagram.

(c) Find the stationary pmf assuming $p < 1$.

11.19. Two gamblers play the following game. A fair coin is flipped; if the outcome is heads, player A pays player B \$1, and if the outcome is tails player B pays player A \$1. The game is continued until one of the players goes broke. Suppose that initially player A has \$1 and player B has \$2, so a total of \$3 is up for grabs. Let X_n denote the number of dollars held by player A after n trials.

(a) Show that X_n is a Markov chain.

(b) Sketch the state transition diagram for X_n and give the one-step transition probability matrix P.

(c) Use the state transition diagram to help you show that for n even (i.e., $n = 2k$),

$$p_{ii}(n) = \left(\frac{1}{2}\right)^n \quad \text{for } i = 1, 2 \text{ and } p_{10}(n) = \frac{2}{3}\left(1 - \left(\frac{1}{4}\right)^k\right) = p_{23}(n).$$

(d) Find the n-step transition probability matrix for n even using part c.

(e) Find the limit of P^n as $n \to \infty$.

(f) Find the probability that player A eventually wins.

11.20. A certain part of a machine can be in two states: working or undergoing repair. A working part fails during the course of a day with probability a. A part undergoing repair is put into working order during the course of a day with probability b. Let X_n be the state of the part.

(a) Show that X_n is a two-state Markov chain and give its one-step transition probability matrix P.

(b) Find the n-step transition probability matrix P^n.

(c) Find the steady state probability for each of the two states.

11.21. A machine consists of two parts that fail and are repaired independently. A working part fails during any given day with probability a. A part that is not working is repaired by the next day with probability b. Let X_n be the number of working parts in day n.

(a) Show that X_n is a three-state Markov chain and give its one-step transition probability matrix P.

(b) Show that the steady state pmf π is binomial with parameter $p = b/(a + b)$.

(c) What do you expect is the steady state pmf for a machine that consists of n parts?

11.22. A stochastic matrix is defined as a nonnegative matrix for which the elements of each row add to one.

(a) Show that the transition probability matrix P for a Markov chain is a stochastic matrix.

(b) Show that if P and Q are stochastic matrices, then PQ is also a stochastic matrix.

(c) Show that if P is a stochastic matrix, then P^n is also a stochastic matrix.

11.23. Show that if P^k has identical rows, then P^j has identical rows for all $j \geq k$.

11.24. Prove Eq. (11.14) by induction.

Section 11.3: Classes of States, Recurrence Properties, and Limiting Probabilities

11.25. **(a)** Sketch the state-transition diagrams for the Markov chains with the following transition probability matrices.

(b) Specify the classes of the Markov chains and classify them as recurrent or transient.

(c) Use Octave to calculate the first few powers of each matrix. Note any interesting behavior.

(i) $\begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}$ **(ii)** $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ **(iii)** $\begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}$

(iv) $\begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$ **(v)** $\begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/4 & 1/4 \\ 0 & 1/4 & 1/4 & 1/2 \end{bmatrix}$

11.26. Characterize the long-term behavior of the Markov chains in Problem 11.25. Find the long-term proportion of time spent in each state. Find the stationary pmf where applicable and determine whether it is unique.

11.27. Consider a three-state Markov chain. Select transition probabilities and sketch the associated transition diagram to produce the following attributes:

(a) X_n is irreducible.

(b) X_n is has one transient class and one recurrent class.

(c) X_n is has two recurrent classes.

11.28. **(a)** Find the transition probability matrices for the Markov chains with the state transition diagrams shown in Fig. P11.1.

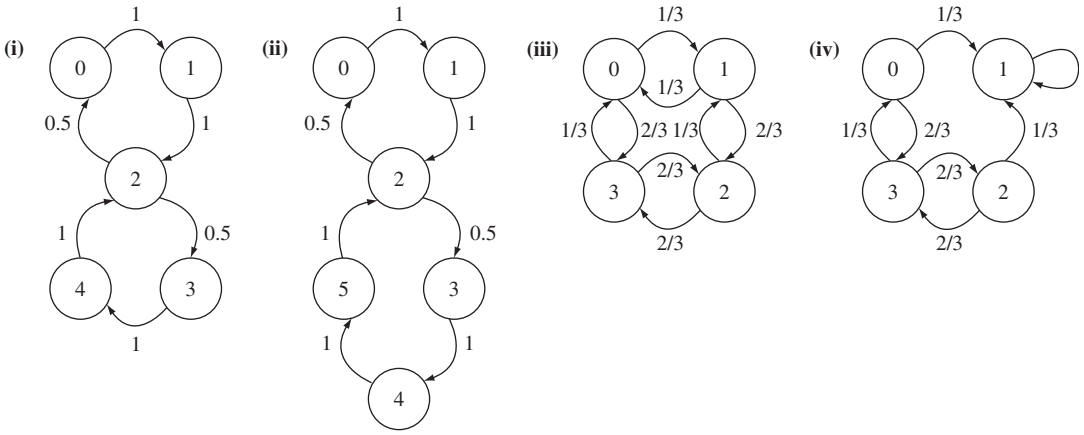

FIGURE P11.1

(b) Specify the classes of the Markov chains and classify them as recurrent or transient; periodic or aperiodic.

(c) Characterize the long-term behavior of the Markov chains and find the long-term proportion of time spent in each state, and the stationary pmf where applicable.

(d) Use Octave to evaluate P^n for $n = 1, 2, 3, 4, 5$. Explain any interesting results you may find.

11.29. **(a)** Apply the PageRank modeling procedure to the Markov chains in Problem 11.28 to find the transition probability matrix.

(b) Find the PageRank value for each node.

11.30. Consider a random walk in the set $\{0, 1, \ldots, M\}$ with transition probabilities

$$p_{01} = 1, p_{M,M-1} = 1, \text{and} \quad p_{i,i-1} = q \qquad p_{i,i+1} = p \text{ for } \quad i = 1, \ldots, M-1.$$

(a) Sketch the state transition diagram.

(b) Find the long-term proportion of time spent in each state, and the limit of $p_{ii}(n)$ as $n \to \infty$. Evaluate the special case when $p = 1/2$.

11.31. Repeat Problem 11.30 if the random walk is modified so that

$$p_{01} = p, p_{00} = q, p_{M,M-1} = q, \text{and } p_{M,M} = p.$$

11.32. For a finite-state, irreducible Markov chain, explain why none of the states can have zero probability.

11.33. Suppose that state i belongs to a recurrent class of a finite-state Markov chain and that $p_{ii}(1) > 0$. Show that i belongs to a class which is aperiodic.

11.34. Prove that positive and null recurrence are class properties.

11.35. In this problem we develop expressions for recurrence probabilities and expectations. Let $a_n = f_{ii}(n)$ be the probability that a first return to state i from state i occurs after n steps; and let $b_n = p_{ii}(n)$ be the probability of a return to state i from state i after n steps.

(a) Show that: $b_n = \sum_{j=0}^{n} b_j a_{n-j}$ where $b_0 = 1, a_0 = 0$. *Hint:* Use conditional probability.

(b) Let $A(z)$ and $B(z)$ be the generating functions of $\{a_n\}$ and $\{b_n\}$ as defined in Eq. (4.84). Explain why the series converge for $|z| < 1$, and show that $B(z) = \dfrac{1}{1 - A(z)}$.

(c) Show that $f_i = \lim_{z \to 1} A(z)$.

(d) Show that state i is recurrent if and only if $\lim_{z \to 1} B(z) = \infty$.

11.36. Consider a Markov chain with state space $\{0, 1, 2, \dots\}$ and the following transition probabilities:

$$p_{0j} = f_j \text{ and } p_{jj-1} = 1 \text{ where } 1 = f_1 + f_2 + \cdots + f_j + \cdots.$$

(a) Sketch the state transition diagram.

(b) Determine whether the Markov chain is irreducible.

(c) Determine whether state 0 is transient, or null/positive recurrent.

(d) Find an expression for the stationary pmf, if it exists.

(e) Provide specific answers to parts c and d if $\{f_i\}$ is given by the following pmfs: **(i)** geometric; **(ii)** Zipf. (See Eq. (3.51).)

11.37. Consider a Markov chain with state space $\{1, 2, \dots\}$ and the following transition probabilities:

$$p_{jj+1} = a_j \text{ and } p_{j1} = 1 - a_j \text{ where } 0 < a_j < 1.$$

(a) Sketch the state transition diagram.

(b) Determine whether the Markov chain is irreducible.

(c) Determine whether state 1 is transient, or null/positive recurrent.

(d) Find an expression for the stationary pmf, if it exists.

(e) Provide specific answers to parts c and d if:

(i) $a_j = 1/2$ all j **(ii)** $a_j = (j-1)/j$ **(iii)** $a_j = 1/j$

(iv) $a_j = (1/2)^j$ **(v)** $a_j = 1 - (1/2)^j$.

11.38. Let X_n and Y_n be two ergodic Markov chains with the same state space but different transition probability matrices, P_1 and P_2, respectively, and different stationary pmf's.

(a) A new process is constructed as follows. A coin is flipped and if the outcome is heads, P_1 is used to generate the entire sequence; but if the outcome is tails, P_2 is used instead. Is the resulting process Markov and does it have a stationary pmf? Is it ergodic?

(b) Repeat part a if the process is constructed as follows. A coin is flipped before every time instant and the associated transition probability matrix is used to determine the next state.

(c) Repeat part a if the state for odd (even) time instants is determined according to P_1 (P_2).

11.39. Find the probability of state 1 for the processes in Problem 11.38(a–c) if X_n and Y_n are two processes from Problem 11.37(e) with two different geometric pmfs in (i) and (iv).

11.40. Construct a multiclass infinite-state Markov chain that has the following attributes:

(a) One class is transient and one class is null recurrent.

(b) One class is null recurrent and one class is positive recurrent.

Section 11.4: Continuous-Time Markov Chains

11.41. Consider the simple queueing system discussed in Example 11.36.

(a) Use the results in Example 11.36 to find the state transition probability matrix.

(b) Find the following probabilities:

$$P[X(1.5) = 1, X(3) = 1 \,|\, X(0) = 0]$$
$$P[X(1.5) = 1, X(3) = 1].$$

11.42. A rechargeable battery in a depot is in one of three states: fully charged, in use, or recharging. Assume the mean time in each of these states is: $1/\lambda$; 1 hour; 3 hours. Batteries are not put into use unless they are fully charged.

(a) Find a Markov model for the battery states and sketch the state transition diagram.

(b) Find the stationary pmf. Explain how the pmf varies with λ.

11.43. Suppose that the depot in Problem 11.42 has two batteries. Define the state at time t by $\{N_F(t), N_U(t), N_C(t)\}$, that is, by the number of batteries in each state.

(a) Sketch the state transition diagram for a six-state Markov chain for the system.

(b) Find the stationary pmf and evaluate it for various values of λ.

11.44. Rolo, a Chihuahua, spends most of the daytime sleeping in the kitchen. When a person enters the kitchen, Rolo greets him or her and wags her tail for an average time of one minute. At the end of this period Rolo is fed with probability 1/4, patted briefly with probability 5/8, or taken for a walk with probability 1/8. If fed, Rolo spends an average of two minutes eating. The walks take 15 minutes on average. After eating, being patted, or walking, she returns to sleep. Assume that people enter the kitchen on average every hour.

(a) Find a Markov chain model with four states: {sleep, greet, eat, walk}. Specify the transition rate matrix.

(b) Find the steady state probabilities.

11.45. A critical part of a machine has an exponentially distributed lifetime with parameter $\alpha = 1$. Suppose that $n = 4$ spare parts are initially in stock, and let $N(t)$ be the number of spares left at time t.

(a) Find $p_{ij}(t) = P[N(s + t) = j \,|\, N(s) = i]$.

(b) Find the transition probability matrix.

(c) Find $p_j(t)$.

(d) Plot $p_j(t)$ versus time for $j = 0, 1, 2, 3, 4$.

(e) Give the general solution for $p_j(t)$ for arbitrary $\alpha > 0$ and n.

11.46. A shop has $n = 3$ machines and one technician to repair them. A machine remains in the working state for an exponentially distributed time with parameter $\mu = 1/3$. The technician works on one machine at a time, and it takes him an exponentially distributed time of rate $\alpha = 1$ to repair each machine. Let $X(t)$ be the number of working machines at time t.

(a) Show that if $X(t) = k$, then the time until the next machine breakdown is an exponentially distributed random variable with rate $k\mu$.

(b) Find the transition rate matrix $[\gamma_{ij}]$ and sketch the transition rate diagram for $X(t)$.

(c) Write the global balance equations and find the steady state probabilities for $X(t)$.

(d) Redo parts b and c if the number of technicians is increased to 2.

(e) Find the steady state probabilities for arbitrary values of n, α, and μ.

11.47. A speaker alternates between periods of speech activity and periods of silence. Suppose that the former are exponentially distributed with mean $1/\alpha = 200$ ms and the latter exponentially distributed with mean $1/\beta = 400$ ms. Consider a group of $n = 4$ independent speakers and let $N(t)$ denote the number of speakers in speech activity at time t.

(a) Find the transition rate diagram and the transition rate matrix for this system.

(b) Write the global balance equations and show that the steady state pmf is given by a binomial distribution. Why is this solution not surprising?

(c) Find the steady state probabilities for arbitrary values of n, α, and β.

11.48. A continuous-time Markov chain $X(t)$ can be approximated by a sampled-time discrete-time Markov chain $X_n = X(n\delta)$ where the sampling interval is δ seconds.

(a) Find the transition probabilities for X_n if $X(t)$ is the M/M/1 queue in Example 11.39.

(b) Find the stationary pmf for part a. Compare to the answer in the example.

11.49. Consider the single-server queueing system in Example 11.39. Suppose that at most K customers can be in the system at any time. Let $N(t)$ be the number of customers in the system at time t. Find the steady state probabilities for $N(t)$.

11.50. **(a)** Find the embedded Markov chain for the process described in Example 11.39.

(b) Find the stationary pmf of the embedded Markov chain.

(c) Characterize the long-term probabilities of the process using Eq. (11.50).

11.51. Repeat Problem 11.50 for the process described in Example 11.40.

11.52. Suppose that the embedded Markov chain for the process $N(t)$ is given by the discrete-time Markov chain in Problem 11.36 with $\{f_i\}$ given by a geometric pmf. Find the steady state probabilities of $N(t)$, if they exist, in the following cases:

(a) The occupancy times of all states are exponentially distributed with mean 1.

(b) The occupancy time of state j is exponentially distributed with mean j.

(c) The occupancy time of state j is exponentially distributed with mean 2^j.

*Section 11.5: Time-Reversed Markov Chains

11.53. N balls are distributed in two urns. At time n, a ball is selected at random, removed from its present urn, and placed in the other urn. Let X_n denote the number of balls in urn 1.

(a) Find the transition probabilities for X_n.

(b) Argue that the process is time reversible and then obtain the steady state probabilities for X_n.

11.54. A point moves in the unit circle in jumps of $\pm 90°$. Suppose that the process is initially at $0°$, and that the probability of $+90°$ is p.

(a) Find the transition probabilities for the resulting Markov chain and obtain the steady state probabilities.

(b) Is the process reversible? Why or why not?

11.55. Find the transition probabilities for the time-reversed version of the random walk discussed in Problem 11.31. Is the process reversible?

11.56. Is the Markov chain in Problem 11.16 time reversible?

11.57. Is the Markov chain in Problem 11.17 time reversible?

11.58. **(a)** Specify the time-reversed version of the process defined in Problem 11.49. Is the process reversible?

(b) Find the steady state probabilities of the process using Eq. (11.67).

11.59. Use the results of Example 11.42 to find the stationary pmf of the Markov chains in Problem 11.37(i).

11.60. Determine whether the simple queueing system in Example 11.36 is reversible.

11.61. Determine whether the machine repair model in Problem 11.46 is reversible.

11.62. **(a)** Is the speech activity model in Problem 11.47 reversible?

(b) Is the model reversible if $\alpha = \beta$?

*Section 11.6: Numerical Techniques for Markov Chains

11.63. Consider the urn experiment in Problem 11.2.

(a) Use matrix diagonalization to find an expression for the state pmf as a function of time. Plot the state pmf vs. time.

(b) Run a simulation for this urn experiment 100 times and build a histogram of the number of steps that take place until the last black ball is removed.

(c) Derive the pmf for the number of steps that elapse until the last black ball is removed. Compare the theoretical pmf with the observed histogram in part b.

11.64. Consider the Bernoulli–Laplace diffusion model from Problem 11.16 with $\rho = 5$.

(a) Use matrix diagonalization to obtain an expression for the time-dependent state pmf. Plot the state pmf vs. time for different initial conditions.

(b) Write a simulation for the model and make several observations of 200-step sample functions. Is the process ergodic? Is it necessary to perform multiple realizations of the process, or does it suffice to collect statistics from one long realization?

(c) Compare histograms of the state occupancy and compare to the theoretical result for: 5 separate realizations of 200 steps; 1 realization of 1000 steps.

(d) Use the `autocov` function in Octave to estimate the covariance function of the process.

11.65. Consider the data multiplexer in Problem 11.11.

(a) Derive the transition probabilities for the multiplexer assuming a maximum state of $N = 100$. Find the steady state pmf for the following parameters: $b = 0.5$ and $a = 0.1$, $a = 0.25$, $a = 0.50$.

(b) Simulate the data multiplexer for each of the cases in part a. Run the simulation for 1000 steps.

(c) For each realization record a histogram of the length of idle periods (when the system remains continuously empty) and the length of the busy periods (when the system remains continuously nonempty). Which of the three choices of parameters above correspond to "heavy traffic"; "light traffic?"

11.66. Consider the gamblers' experiment in Problem 11.19 with player A beginning with \$6 and player B with \$3.

(a) Find the transition probability P and obtain an expression for P^n. What is the probability that player A wins? What is the average time until player A wins (when he wins)?

(b) Simulate 500 trials of the experiment. Find the relative frequency of player A winning and compare to the theoretical result.

(c) Find the mean time until player A wins; until player B wins. Compare to the theoretical results.

11.67. Consider the residual lifetime process in Problem 11.36. Assume a machine state of 100.

(a) Simulate 1000 steps of the process with a geometric random variable with mean 5. Record histograms of the state pmf and obtain the autocovariance of the realization.

(b) Repeat part a with a Zipf random variable of mean 5. Compare the histogram and autocovariance to those found in part a.

11.68. Consider the age process in Problem 11.37. Assume a machine state of 100.

(a) Simulate 1000 steps of the process with $a_j = (j-1)/j$. Does the process behave as expected?

(b) Repeat part a with $a_j = 1 - (1/2)^j$.

11.69. Consider the battery experiment in Problem 11.43.

(a) Use matrix diagonalization to obtain the time-dependent state transition probabilities for $\lambda = 0.1, 1, 10$. What are the steady state probabilities? What are the corresponding embedded state probabilities?

(b) Simulate 500 hours of operation and observe the histogram of the embedded state occupancies. Compare to the theoretical results.

11.70. Consider the machine repair model in Problem 11.46. Assume $n = 10$ machines, $\mu = 1/10$ average working time, and $\alpha = 1$.

(a) Obtain the time-dependent state transition probabilities for 1 and 2 technicians. What are the steady state probabilities? What are the corresponding embedded state probabilities?

(b) Simulate 1000 hours of operation and observe the histogram of the embedded state occupancies. Compare to the theoretical results.

11.71. Use the simulator developed in Example 11.49 to simulate a sampled-time approximation to the birth-death process shown in Figure 11.20(b). Simulate 200 seconds of an M/M/1 queue in which jobs arrive at rate $\lambda = 0.9$ jobs per second and jobs complete processing at a rate of 1 job every second. Assume the system is initially empty. Show the realizations of the sampled process and measure the proportion of time spent in each state. Compare these to the theoretical values.

Problems Requiring Cumulative Knowledge

11.72. **(a)** The Markov chain in Fig. 11.6(b) is started in state 0 at time 0. Find the n-step transition probability matrix for even and odd numbers of steps. What happens as $n \to \infty$?

(b) Let X_n be an irreducible, periodic, positive recurrent Markov chain in steady state. Is X_n a cyclostationary random process?

11.73. Let X_n be an ergodic Markov chain. Let $I_j(n)$ be the indicator function for state j at time n, that is, $I_j(n)$ is 1 if the state at time n is j, and 0 otherwise. What is the limiting value of the time average of $I_j(n)$? Is this result an ergodic theorem?

11.74. Let $X(t)$ be a continuous-time model for speech activity, in which a speaker is active (state 1) for an exponentially distributed time with rate α and is silent (state 0) for an exponentially distributed time with rate β. Assume all active and silence durations are independent random variables.

(a) Find a two-state Markov chain for $X(t)$.

(b) Find $p_0(t)$ and $p_1(t)$.

(c) Find the autocorrelation function of $X(t)$.

(d) If $X(t)$ is asymptotically wide-sense stationary, find its power spectral density.

(e) Suppose we have n independent speakers, and let $N(t)$ be the total number of speakers active at time t. Find the autocorrelation function of $N(t)$, and its power spectral density if it is asymptotically wide-sense stationary.

11.75. Let X_n be a continuous-valued discrete-time Markov process.

(a) Find the expression for the joint pdf corresponding to Eq. (11.5).

(b) Find the expression for the two-step transition pdf corresponding to Eq. (11.12a).

11.76. Consider the aquifer in Problem 11.8.

(a) Find a recursive equation for the amount of water in the aquifer X_{n+1} in year $n + 1$ in terms of the amount of water in year n, the amount withdrawn from use D_n, and the amount restored by rainfall W_n. Note that the amount of water must be nonnegative.

(b) Find an integral expression relating the steady state pdf of X to the pdf's of W and D. Assume that W and D are independent and Gaussian random variables. Propose possible approaches to solving these equations.

(c) Write a computer simulation to investigate the distribution of X as a function of W and D assuming: W_n and D_n are iid random variables with the same mean; D_n is iid random variable, but W_n is independent with a slowly varying mean (with period 100 years) that is equal to that of D_n when averaged over the entire period.

CHAPTER

12

Introduction to Queueing Theory

In many applications, scarce resources such as computers and communication systems are shared among a community of users. Users place demands for these resources at random times, and they require use of these resources for time periods whose durations are random. Inevitably requests for the resource arrive while the resource is occupied, and a mechanism to provide an orderly access to the resource is required. The most common access control mechanism is to file user requests in a waiting line or "queue" such as might be formed at a bank by customers waiting to be served. Resource sharing can also take place in systems of very large scale, e.g., peer-to-peers networks, where the "queues" are not as readily apparent.

Queueing theory deals with the study of waiting lines and resource sharing. The random nature of the demand behavior of customers implies that probabilistic measures such as average delay, average throughput, and delay percentiles are required to assess the performance of such systems. Queueing theory provides us with the probability tools needed to evaluate these measures.

This chapter is organized as follows:

- Section 12.1 introduces the basic structure of a queueing system.
- Section 12.2 develops Little's formula which provides a fundamental relationship that is applicable in most queueing systems.
- In Section 12.3 we examine the M/M/1 queue and use it to develop many of the basic insights into queueing systems.
- Sections 12.4 and 12.5 develop multiserver systems and finite-source systems which can both be represented by Markov chains.
- Sections 12.6 and 12.7 develop M/G/1 queues which require more complex modeling.
- Section 12.8 and 12.9 presents Burke's and Jackson's theorems which allow us to model networks of queues.
- Finally Section 12.10 considers the simulation of queueing systems.

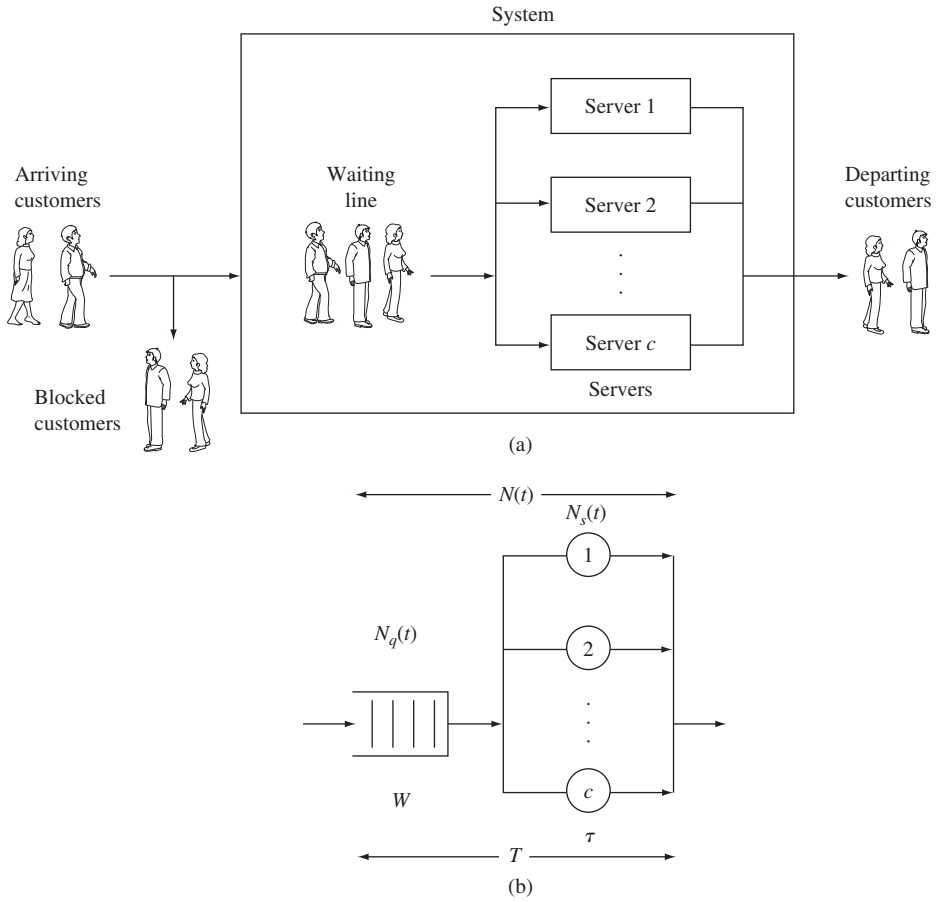

FIGURE 12.1
(a) Elements of a queueing system. (b) Elements of a queueing system model: $N(t)$, number in system; $N_q(t)$, number in queue; $N_s(t)$, number in service; W, waiting time in queue; τ, service time; and T, total time in the system.

12.1 THE ELEMENTS OF A QUEUEING SYSTEM

Figure 12.1(a) shows a typical queueing system and Fig. 12.1(b) shows the elements of a queueing system model. Customers from some population arrive at the system at the random *arrival times* $S_1, S_2, S_3, \ldots, S_i, \ldots$, where S_i denotes the arrival time of the ith customer. We denote the customer **arrival rate** by λ.

The queueing system has one or more identical servers, as shown in Fig. 12.1(a). The ith customer arrives at the system seeking a service that will require τ_i seconds of **service time** from one server. If all the servers are busy, then the arriving customer joins a queue where he remains until a server becomes available. Sometimes, only a limited number of waiting spaces are available so customers that arrive when there is no room are turned away. Such customers are called "blocked" and we will denote the rate at which customers are turned away by λ_b.

The **queue** or **service discipline** specifies the order in which customers are selected from the queue and allowed into service. For example, some common queueing disciplines are *first come*, *first served*, and *last come*, *first served.* The queueing discipline affects the **waiting time** W_i that elapses from the arrival time of the ith customer until the time when it enters service. The **total delay** T_i of the ith customer in the system is the sum of its waiting time and service time:

$$T_i = W_i + \tau_i. \tag{12.1}$$

From the customer's point of view, the performance of the system is given by the statistics of the waiting time W and the total delay T, and the proportion of customers that are blocked, λ_b/λ. From the point of view of resource allocation, the performance of the system is measured by the proportion of time that each server is utilized and the rate at which customers are serviced by the system, $\lambda_d = \lambda - \lambda_b$. These quantities are a function of $N(t)$, the number of customers in the system at time t, and $N_q(t)$, the number of customers in queue at time t.

The notation *a/b/m/K* is used to describe a queueing system, where a specifies the type of arrival process, b denotes the service time distribution, m specifies the number of servers, and K denotes the maximum number of customers allowed in the system at any time. If a is given by M, then the arrival process is Poisson and the interarrival times are independent, identically distributed (iid) exponential random variables. If b is given by M, then the service times are iid exponential random variables. If b is given by D, then the service times are constant, that is, deterministic. If b is given by G, then the service times are iid according to some general distribution. For example, in this chapter we deal with M/M/1, M/M/1/K, M/M/c, M/M/c/c, M/D/1, and M/G/1 queues.

Queueing system models find many applications in electrical and computer engineering. The "servers" in Fig. 12.1 can represent a variety of resources that perform "work." For example, in communication networks, the server can represent a communications line that transmits packets of information. In computer systems, the servers could represent processes in a computer that each handles Web queries from a particular client. Modern distributed applications combine these communications and computing resources into vast networks of interacting queueing systems.

12.2 LITTLE'S FORMULA

We now develop **Little's formula**, which states that, for systems that reach steady state, the average number of customers in a system is equal to the product of the average arrival rate and the average time spent in the system:

$$E[N] = \lambda E[T]. \tag{12.2}$$

This formula is valid under very general conditions, so it is applicable in an amazing number of situations.

Consider the queueing system shown in Fig. 12.2. The system begins empty at time $t = 0$, and the customer arrival times are denoted by $S_1, S_2, \ldots$. Let $A(t)$ be the number of customer arrivals up to time t. The ith customer spends time T_i in the system and then

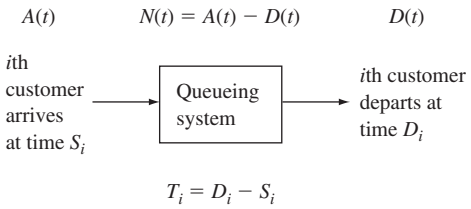

FIGURE 12.2
Time in system is departure time minus arrival time. Number in system at time t is number of arrivals minus number of departures.

departs at time $D_i = S_i + T_i$. We will let $D(t)$ be the number of customer departures up to time t. The **number of customers in the system** at time t is the number of arrivals that have not yet left the system:

$$N(t) = A(t) - D(t). \tag{12.3}$$

Figure 12.3 shows a possible sample path for $A(t)$, $D(t)$, and $N(t)$ in a queueing system with "first come, first served" service discipline.

Consider the time average of the number of customers in the system $N(t)$ during the interval $(0, t]$:

$$\langle N \rangle_t = \frac{1}{t} \int_0^t N(t')\, dt'. \tag{12.4}$$

In Fig. 12.3, $N(t)$ is the region between $A(t)$ and $D(t)$, so the above integral is given by the area of the enclosed region up to time t. It can be seen that each customer who has

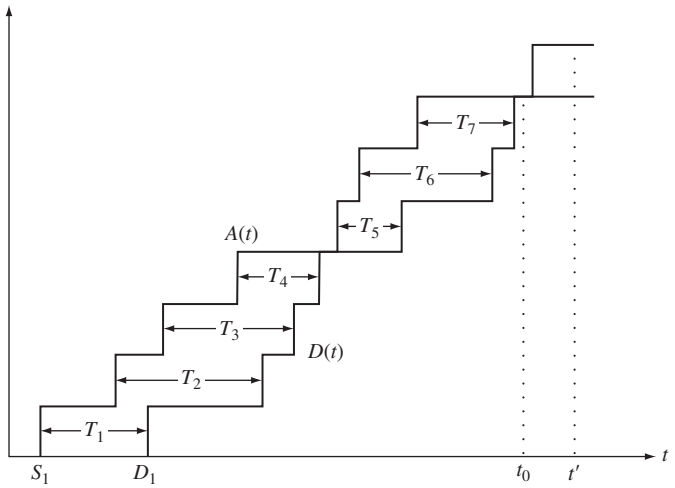

FIGURE 12.3
Total time spent by the first seven customers is the area in $A(t) - D(t)$ up to time t_0.

departed the system by time t contributes T_i to the integral, and thus the integral is simply the total time all customers have spent in the system up to time t.

Consider, for now, a time instant $t = t_0$ for which $N(t) = 0$ as in Fig. 12.3, then the integral is exactly given by the sum of the T_i of the first $A(t)$ customers:

$$\langle N \rangle_t = \frac{1}{t} \sum_{i=1}^{A(t)} T_i. \tag{12.5}$$

The average arrival rate up to time t is given by

$$\langle \lambda \rangle_t = \frac{A(t)}{t}. \tag{12.6}$$

If we solve Eq. (12.6) for t and substitute into Eq. (12.5), we obtain

$$\langle N \rangle_t = \langle \lambda \rangle_t \frac{1}{A(t)} \sum_{i=1}^{A(t)} T_i. \tag{12.7}$$

Let $\langle T \rangle_t$ be the average of the times spent in the system by the first $A(t)$ customers, then

$$\langle T \rangle_t = \frac{1}{A(t)} \sum_{i=1}^{A(t)} T_i. \tag{12.8}$$

Comparing Eqs. (12.7) and (12.8), we conclude that

$$\langle N \rangle_t = \langle \lambda \rangle_t \langle T \rangle_t. \tag{12.9}$$

Finally, we assume that as $t \to \infty$, with probability one, the above time averages converge to the expected value of the corresponding steady state random processes, that is,

$$\begin{aligned} \langle N \rangle_t &\to E[N] \\ \langle \lambda \rangle_t &\to \lambda \\ \langle T \rangle_t &\to E[T]. \end{aligned} \tag{12.10}$$

Equations (12.9) and (12.10) then imply Little's formula:

$$E[N] = \lambda E[T]. \tag{12.11}$$

The restriction of t to instants t_0 where $N(t_0) = 0$ is not necessary. The time average of $N(t)$ up to an arbitrary time t' as shown in Fig. 12.3 is given by the average up to time t_0 plus a contribution from the interval from t_0 to t'. If $E[N] < \infty$, then as t becomes large, this contribution becomes negligible.

The assumption of first come, first served service discipline is not necessary. It turns out that Little's formula holds for many service disciplines. See Problem 12.2 for examples. In addition, Little's formula holds for systems with an arbitrary number of servers.

Up to this point we have implicitly assumed that the "system" is the entire queueing system, so N is the number in the queueing system and T is the time spent in the

queueing system. However, Little's formula is so general that it applies to many interpretations of "system." Examples 12.1 and 12.2 show other designations for "system."

Example 12.1 Mean Number in Queue

Let $N_q(t)$ be the number of customers waiting in queue for the server to become available, and let the random variable W denote the waiting time. If we designate the queue to be the "system," then Little's formula becomes

$$E[N_q] = \lambda E[W]. \tag{12.12}$$

Example 12.2 Server Utilization

Let $N_s(t)$ be the number of customers that are being served at time t, and let τ denote the service time. If we designate the set of servers to be the "system," then Little's formula becomes

$$E[N_s] = \lambda E[\tau]. \tag{12.13}$$

$E[N_s]$ is the average number of busy servers for a system in steady state.

For single-server systems, $N_s(t)$ can only be 0 or 1, so $E[N_s]$ represents the proportion of time that the server is busy. If $p_0 = P[N(t) = 0]$ denotes the steady state probability that the system is empty, then we must have that

$$1 - p_0 = E[N_s] = \lambda E[\tau] \tag{12.14}$$

or

$$p_0 = 1 - \lambda E[\tau], \tag{12.15}$$

since $1 - p_0$ is the proportion of time that the server is busy. For this reason, the **utilization of a single-server system** is defined by

$$\rho = \lambda E[\tau]. \tag{12.16}$$

We similarly define **utilization of a *c*-server system** by

$$\rho = \frac{\lambda E[\tau]}{c}. \tag{12.17}$$

From Eq. (12.13), ρ represents the average fraction of busy servers.

12.3 THE M/M/1 QUEUE

Consider a single-server system in which customers arrive according to a Poisson process of rate λ so the **interarrival times** are iid exponential random variables with mean $1/\lambda$. Assume that the service times are iid exponential random variables with mean $1/\mu$, and that the interarrival and service times are independent. In addition, assume that the system can accommodate an unlimited number of customers. The resulting system is an M/M/1 queueing system. In this section we find the steady state pmf of $N(t)$, the number of customers in the system, and the pdf of T, the total customer delay in the system.

12.3.1 Distribution of Number in the System

The number of customers $N(t)$ in an M/M/1 system is a continuous-time Markov chain. To see why, suppose we are given that $N(t) = k$, and consider the next possible change in the number in the system. The time until the next arrival is an exponential random variable that is independent of the service times of customers already in the system. The memoryless property of the exponential random variable implies that this interarrival time is independent of the present and past history of $N(t)$. If the system is nonempty (i.e., $N(t) > 0$) the time until the next departure is also an exponential random variable. The memoryless property implies that the time until the next departure is independent of the time already spent in service. Thus if we know that $N(t) = k$, then the past history of the system is irrelevant as far as the probabilities of future states are concerned. This is the property required of a Markov chain.

To find the transition rates for $N(t)$, consider the probabilities of the various ways in which $N(t)$ can change.

(i) Since $A(t)$, the number of arrivals in an interval of length t, is a Poisson process, the probability of one arrival in an interval of length δ is

$$P[A(\delta) = 1] = \frac{\lambda\delta}{1!}e^{-\lambda\delta} = \lambda\delta\left\{1 - \frac{\lambda\delta}{1!} + \frac{(\lambda\delta)^2}{2!} - \cdots\right\}$$
$$= \lambda\delta + o(\delta). \tag{12.18}$$

(ii) Similarly, the probability of more than one arrival is

$$P[A(\delta) \geq 2] = o(\delta). \tag{12.19}$$

(iii) Since the service time is an exponential random variable τ, the time a customer has spent in service is independent of how much longer he will remain in service because of the memoryless property of τ. In particular, the probability of a customer in service completing his service in the next δ seconds is

$$P[\tau \leq \delta] = 1 - e^{-\mu\delta} = \mu\delta + o(\delta). \tag{12.20}$$

(iv) Since service times and the arrival process are independent, the probability of one arrival and one departure in an interval of length δ is

$$P[A(\delta) = 1, \tau \leq \delta] = P[A(\delta) = 1]P[\tau \leq \delta] = o(\delta) \tag{12.21}$$

from Eqs. (12.18) and (12.20). Similarly, the probability of any change that involves more than a single arrival or a single departure is $o(\delta)$.

Properties (i) through (iv) imply that $N(t)$ has the transition rate diagram shown in Fig. 12.4. The global balance equations for the steady state probabilities are

$$\lambda p_0 = \mu p_1$$
$$(\lambda + \mu)p_j = \lambda p_{j-1} + \mu p_{j+1} \qquad j = 1, 2, \ldots. \tag{12.22}$$

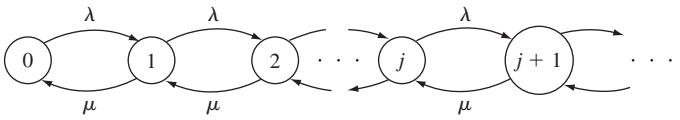

FIGURE 12.4
Transition rate diagram for M/M/1 system.

In Example 11.39, we saw that a steady state solution exists when $\rho = \lambda/\mu < 1$:

$$P[N(t) = j] = (1 - \rho)\rho^j \qquad j = 0, 1, 2, \ldots. \tag{12.23}$$

The condition $\rho = \lambda/\mu < 1$ must be met if the system is to be stable in the sense that $N(t)$ does not grow without bound. Since μ is the maximum rate at which the server can process customers, the condition $\rho < 1$ is equivalent to

$$\text{Arrival rate} = \lambda < \mu = \text{Maximum service rate}. \tag{12.24}$$

If the inequality is violated, we have customers arriving at the system faster than they can be processed and sent out. This is an unstable situation in which the number in the queue will grow steadily without bound.

The mean number of customers in the system is given by

$$E[N] = \sum_{j=0}^{\infty} jP[N(t) = j] = \frac{\rho}{1 - \rho}, \tag{12.25}$$

where we have used the fact that N has a geometric distribution (see Table 3.1).

The mean total customer delay in the system is found from Eq. (12.25) and Little's formula:

$$\begin{aligned} E[T] &= \frac{E[N]}{\lambda} = \frac{\rho/\lambda}{1 - \rho} \\ &= \frac{1/\mu}{1 - \rho} = \frac{E[\tau]}{1 - \rho} = \frac{1}{\mu - \lambda}. \end{aligned} \tag{12.26}$$

The mean waiting time in queue is given by the mean of the total time in the system minus the service time:

$$\begin{aligned} E[W] &= E[T] - E[\tau] \\ &= \frac{E[\tau]}{1 - \rho} - E[\tau] \\ &= \frac{\rho}{1 - \rho}E[\tau]. \end{aligned} \tag{12.27}$$

Little's formula then gives the mean number in queue:

$$\begin{aligned} E[N_q] &= \lambda E[W] \\ &= \frac{\rho^2}{1 - \rho}. \end{aligned} \tag{12.28}$$

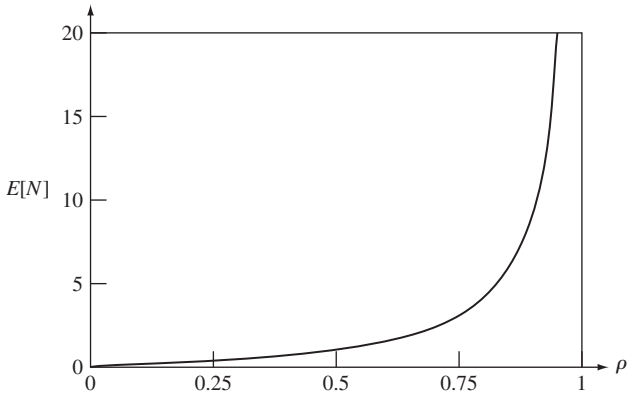

FIGURE 12.5
Mean number of customers in the system versus utilization for M/M/1 queue.

The server utilization (defined in Example 12.2) is given by

$$1 - p_0 = 1 - (1 - \rho) = \rho = \frac{\lambda}{\mu}. \tag{12.29}$$

Figures 12.5 and 12.6 show $E[N]$ and $E[T]$ versus ρ. It can be seen that as ρ approaches one, the mean number in the system and the system delay become arbitrarily large.

Example 12.3

A router receives packets from a group of users and transmits them over a single transmission line. Suppose that packets arrive according to a Poisson process at a rate of one packet every 4 ms, and suppose that packet transmission times are exponentially distributed with mean 3 ms.

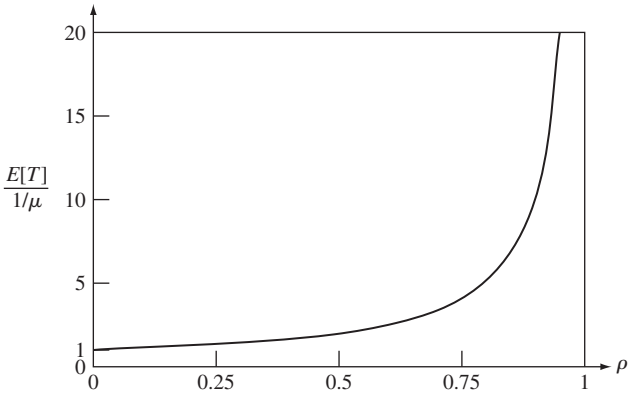

FIGURE 12.6
Mean total customer delay versus utilization for M/M/1 system. The delay is expressed in multiples of mean service times.

Find the mean number of packets in the system and the mean total delay in the system. What percentage increase in arrival rate results in a doubling of the above mean total delay?

The arrival rate is 1/4 packets/ms and the mean service time is 3 ms. The utilization is therefore

$$\rho = \frac{1}{4}(3) = \frac{3}{4}.$$

The mean number of packets in the system is then

$$E[N] = \frac{\rho}{1-\rho} = 3.$$

The mean time in the system is

$$E[T] = \frac{E[N]}{\lambda} = \frac{3}{1/4} = 12 \text{ ms}.$$

The mean time in the system will be doubled to 24 ms when

$$24 = \frac{E[\tau]}{1-\rho'} = \frac{3}{1-\rho'}.$$

The resulting utilization is $\rho' = 7/8$ and the corresponding arrival rate is $\lambda' = \rho'\mu = 7/24$. The original arrival rate was 6/24. Thus an increase in arrival rate of 1/6 = 17% leads to a 100% increase in mean system delay.

The point of this example is that *the onset of congestion is swift.* The mean delay increases rapidly once the utilization increases beyond a certain point.

Example 12.4 Concentration and Effect of Scale

A large processor handles transactions at a rate of $K\mu$ transactions per second. Suppose transactions arrive according to a Poisson process of rate $K\lambda$ transactions/second, and that transactions require an exponentially distributed amount of processing time. Suppose that a proposal is made to eliminate the large processor and to replace it with K processors, each with a processing rate of μ transactions per second and an arrival rate of λ. Compare the mean delay performance of the existing and the proposed systems.

The large processor system is an M/M/1 queue with arrival rate $K\lambda$, service rate $K\mu$, and utilization $\rho = K\lambda/K\mu = \lambda/\mu$. The mean delay is given by Eq. (12.26):

$$E[T] = \frac{E[\tau]}{1-\rho} = \frac{1/K\mu}{1-\rho}.$$

Each of the small processors is an M/M/1 system with arrival rate λ, service rate μ, and utilization $\rho = \lambda/\mu$. The mean delay is

$$E[T'] = \frac{E[\tau']}{1-\rho} = \frac{1/\mu}{1-\rho} = KE[T].$$

Thus, the system with the single large processor with processing rate $K\mu$ has a smaller mean delay than the system with K small processors each of rate μ. In other words, the concentration of customer demand into a single system results in significant delay performance improvement.

12.3.2 Delay Distribution in M/M/1 System and Arriving Customer's Distribution

Let N_a denote the number of customers found in the system by a customer arrival. We call $P[N_a = k]$ the **arriving customer's distribution**. We now show that if arrivals are Poisson and independent of the system state and customer service times, then the arriving customer's distribution is equal to the steady state distribution for the number in the system. A customer that arrives at time $t + \delta$ finds k in the system if $N(t) = k$, thus

$$\begin{aligned} P[N_a(t) = k] &= \lim_{\delta \to 0} P[N(t) = k \,|\, A(t+\delta) - A(t) = 1] \\ &= \lim_{\delta \to 0} \frac{P[N(t) = k, A(t+\delta) - A(t) = 1]}{P[A(t+\delta) - A(t) = 1]} \\ &= \lim_{\delta \to 0} \frac{P[A(t+\delta) - A(t) = 1 \,|\, N(t) = k] P[N(t) = k]}{P[A(t+\delta) - A(t) = 1]}, \end{aligned}$$

where we have used the definition of conditional probability. The probability of an arrival in the interval $(t, t + \delta]$ is independent of $N(t)$, thus

$$\begin{aligned} P[N_a(t) = k] &= \lim_{\delta \to 0} \frac{P[A(t+\delta) - A(t) = 1] P[N(t) = k]}{P[A(t+\delta) - A(t) = 1]} \\ &= P[N(t) = k]. \end{aligned}$$

Thus the probability that $N_a = k$ is simply the proportion of time during which the system has k customers in the system. For the M/M/1 queueing system under consideration we have

$$P[N_a = k] = P[N(t) = k] = (1 - \rho)\rho^k. \tag{12.30}$$

We are now ready to compute the distribution for the total time T that a customer spends in an M/M/1 system. Suppose that an arriving customer finds k in the system, that is, $N_a = k$. If the service discipline is "first come, first served," then T is the residual service time of the customer found in service, the service times of the $k - 1$ customers found in queue, and the service time of the arriving customer. The memoryless property of the exponential service time implies that the residual service time of the customer found in service has the same distribution as a full service time. Thus T is the sum of $k + 1$ iid exponential random variables. In Example 7.5 we saw that this sum has the gamma pdf

$$f_T(x \,|\, N_a = k) = \frac{(\mu x)^k}{k!} \mu e^{-\mu x} \qquad x > 0. \tag{12.31}$$

The pdf of T is found by averaging over the probability of an arriving customer finding k messages in the system, $P[N_a = k]$. Thus the pdf of T is

$$\begin{aligned} f_T(x) &= \sum_{k=0}^{\infty} \frac{(\mu x)^k}{k!} \mu e^{-\mu x} P[N(t) = k] \\ &= \sum_{k=0}^{\infty} \frac{(\mu x)^k}{k!} \mu e^{-\mu x} (1 - \rho)\rho^k \end{aligned}$$

$$= (1 - \rho)\mu e^{-\mu x} \sum_{k=0}^{\infty} \frac{(\mu\rho x)^k}{k!}$$
$$= (1 - \rho)\mu e^{-\mu x} e^{\mu\rho x}$$
$$= (\mu - \lambda)e^{-(\mu-\lambda)x} \qquad x > 0. \tag{12.32}$$

Thus T is an exponential random variable with mean $1/(\mu - \lambda)$. Note that this is in agreement with Eq. (12.26) for the mean of T obtained through Little's formula.

We can similarly show that the pdf for the waiting time is

$$f_W(x) = (1 - \rho)\delta(x) + \lambda(1 - \rho)e^{-\mu(1-\rho)x} \qquad x > 0. \tag{12.33}$$

Example 12.5

Find the 95% percentile of the total delay.

The pth percentile of T is that value of x for which

$$p = P[T \leq x]$$
$$= \int_0^x (\mu - \lambda)e^{-(\mu-\lambda)y}\, dy = 1 - e^{-(\mu-\lambda)x},$$

which yields

$$x = \frac{1}{\mu - \lambda} \ln \frac{1}{1 - p} = -E[T] \ln(1 - p). \tag{12.34}$$

The 95% percentile is obtained by substituting $p = .95$ above. The result is $x = 3.0\, E[T]$.

12.3.3 The M/M/1 System with Finite Capacity

Real systems can only accommodate a finite number of customers, but the assumption of infinite capacity is convenient when the probability of having a full system is negligible. Consider the M/M/1/K queueing system that is identical to the M/M/1 system with the exception that it can only hold a maximum of K customers in the system. Customers that arrive when the system is full are turned away.

The process $N(t)$ for this system is a continuous-time Markov chain that takes on values from the set $\{0, 1, \ldots, K\}$ with transition rate diagram as shown in Fig. 12.7. It can be seen that the arrival rate *into* the system is now zero when $N(t) = K$. The transition rates from the other states are the same as for the M/M/1 system.

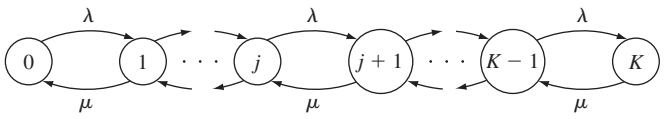

FIGURE 12.7
Transition rate diagram for M/M/1/K system.

The global balance equations are now

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda + \mu)p_j &= \lambda p_{j-1} + \mu p_{j+1} \qquad j = 1, 2, \ldots, K-1 \\ \mu p_K &= \lambda p_{K-1}. \end{aligned} \tag{12.35}$$

Let $\rho = \lambda/\mu$. It can be readily shown (see Problem 12.14) that the steady state probabilities are

$$P[N = j] = \frac{(1-\rho)\rho^j}{1-\rho^{K+1}} \qquad j = 0, 1, 2, \ldots, K \tag{12.36}$$

for $\rho < 1$ or $\rho > 1$. When $\rho = 1$ all the states are equiprobable. Figure 12.8 shows the steady state probabilities for various values of ρ.

The mean number of customers in the system is given by

$$\begin{aligned} E[N] &= \sum_{j=0}^{K} jP[N(t) = j] \\ &= \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} & \text{for } \rho \neq 1 \\ \dfrac{K}{2} & \text{for } \rho = 1. \end{cases} \end{aligned} \tag{12.37}$$

The mean total time spent by customers in the system is found from Eq. (12.37) by using Little's formula with λ_a, the rate of arrivals that actually enter the system. The proportion of time when the system turns away customers is $P[N(t) = K] = p_K$. Thus the system turns away customers at the rate

$$\lambda_b = \lambda p_K, \tag{12.38}$$

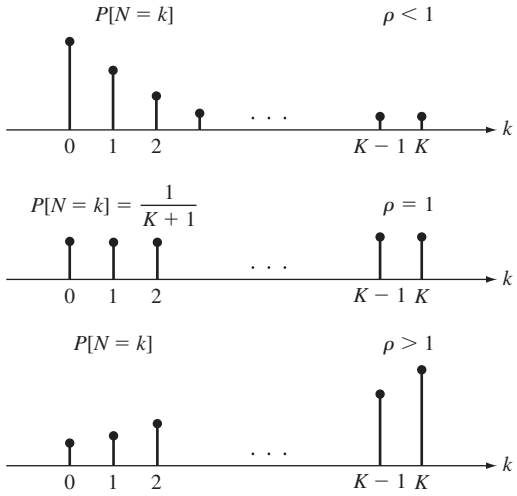

FIGURE 12.8
Typical pmf's for $N(t)$ of M/M/1/K system.

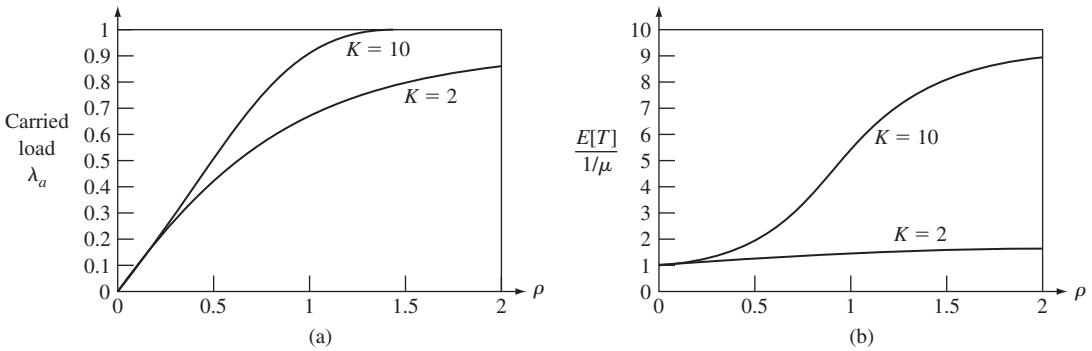

FIGURE 12.9
(a) Carried load versus offered load for M/M/1/K system with $K = 2$ and $K = 10$. (b) Mean customer delay versus offered load in M/M/1/K system with $K = 2$ and $K = 10$.

and the actual arrival rate *into* the system is

$$\lambda_a = \lambda(1 - p_K). \tag{12.39}$$

Applying Little's formula to Eq. (12.37) we obtain

$$E[T] = \frac{E[N]}{\lambda_a} = \frac{E[N]}{\lambda(1 - p_K)}. \tag{12.40}$$

In finite-capacity systems, it is necessary to distinguish between the traffic load offered to a system and the actual load carried by the system. The **offered load**, or **traffic intensity**, is a measure of the demand made on the system and is defined as

$$\lambda \frac{\text{customers}}{\text{second}} \times E[\tau] \frac{\text{seconds of service}}{\text{customer}}. \tag{12.41}$$

The **carried load** is the actual demand met by the system:

$$\lambda_a \frac{\text{customers}}{\text{second}} \times E[\tau] \frac{\text{seconds of service}}{\text{customer}}. \tag{12.42}$$

Example 12.6 Mean Delay and Carried Load Versus *K*

Figure 12.9(a) gives a comparison of the carried load versus the offered load ρ for two values of K. It can be seen that increasing the capacity K results in an increase in carried load since more customers are allowed into the system. Figure 12.9(b) gives the corresponding values for the mean delay. We see that increasing K results in increased delays, again because more customers are allowed into the system.

Example 12.7

Suppose that an M/M/1 model is used for a system that has capacity K, and that the probability of rejecting customers is approximated by $P[N = K]$. Compare this approximation to the exact probability given by the M/M/1/K model.

For the M/M/1 system the above probability is given by

$$P[N = K] = (1 - \rho)\rho^K.$$

For $\rho < 1$, the probability of rejecting a customer in the M/M/1/K system is

$$P[N' = K] = \frac{(1 - \rho)\rho^K}{1 - \rho^{K+1}} = (1 - \rho)\rho^K\{1 + \rho^{K+1} + (\rho^{K+1})^2 + \cdots\}.$$

For $\rho < 1$ and K large, $P[N = k] \simeq P[N' = K]$. For $\rho > 1$, the M/M/1 approximation breaks down and gives a negative probability.

12.4 MULTI SERVER SYSTEMS: M/M/c, M/M/c/c, AND M/M/∞

We now modify the M/M/1 system to consider queueing systems with multiple servers. In particular, we consider systems with iid exponential interarrival times and iid exponential service times. As in the case of the M/M/1 system, the resulting systems can be modeled by continuous-time Markov chains.

12.4.1 Distribution of Number in the M/M/c System

The transition rate diagram for an M/M/c system is shown in Fig. 12.10. As before, arrivals occur at a rate λ. The difference now is that the departure rate is $k\mu$ when k servers are busy. To see why, suppose that k of the servers are busy, then the time until the next departure is given by

$$X = \min(\tau_1, \tau_2, \ldots, \tau_k),$$

where τ_i are iid exponential random variables with parameter μ. The complementary cdf of this random variable is

$$\begin{aligned} P[X > t] &= P[\min(\tau_1, \tau_2, \ldots, \tau_k) > t] \\ &= P[\tau_1 > t, \tau_2 > t, \ldots, \tau_k > t] \\ &= P[\tau_1 > t]P[\tau_2 > t] \ldots P[\tau_k > t] \\ &= e^{-\mu t}e^{-\mu t} \ldots e^{-\mu t} \\ &= e^{-k\mu t}. \end{aligned} \tag{12.43}$$

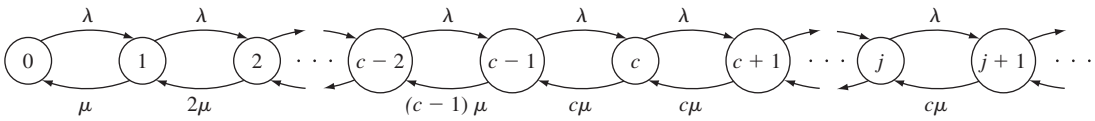

FIGURE 12.10
Transition rate diagram for M/M/c system.

Thus the time until the next departure is an exponential random variable with mean $1/k\mu$. So when k servers are busy, customers depart at rate $k\mu$. When the number of customers in the system is greater than c, all c servers are busy and the departure rate is $c\mu$.

We obtain the steady state probabilities for the M/M/c system from the general solution for birth-and-death processes found in Example 11.40. The probabilities of the first c states are obtained from the following recursion (see Eq. 11.45):

$$p_j = \frac{\lambda}{j\mu} p_{j-1} \qquad j = 1, \ldots, c,$$

which leads to

$$p_j = \frac{a^j}{j!} p_0 \qquad j = 0, 1, \ldots, c, \tag{12.44}$$

where

$$a = \frac{\lambda}{\mu}. \tag{12.45}$$

The probabilities for states equal to or greater than c are obtained from the following recursion:

$$p_j = \frac{\lambda}{c\mu} p_{j-1} \qquad j = c, c+1, c+2, \ldots,$$

which leads to

$$p_j = \rho^{j-c} p_c \qquad j = c, c+1, c+2, \ldots \tag{12.46a}$$

$$= \frac{\rho^{j-c} a^c}{c!} p_0, \tag{12.46b}$$

where we have used Eq. (12.44) with $j = c$ and where

$$\rho = \frac{\lambda}{c\mu}. \tag{12.47}$$

Finally p_0 is obtained from the normalization condition:

$$1 = \sum_{j=0}^{\infty} p_j = p_0 \left\{ \sum_{j=0}^{c-1} \frac{a^j}{j!} + \frac{a^c}{c!} \sum_{j=c}^{\infty} \rho^{j-c} \right\}.$$

The system is stable and has a steady state if the term inside the brackets is finite. This is the case if the second series converges, which in turn requires that $\rho < 1$, or equivalently,

$$\lambda < c\mu. \tag{12.48}$$

In other words, the system is stable if the customer arrival rate is less than the total rate at which the c servers can process customers. The final form for p_0 is

$$p_0 = \left\{ \sum_{j=0}^{c-1} \frac{a^j}{j!} + \frac{a^c}{c!} \frac{1}{1-\rho} \right\}^{-1}. \tag{12.49}$$

The probability that an arriving customer finds all servers busy and has to wait in queue is an important parameter of the M/M/c system:

$$P[W > 0] = P[N \geq c] = \sum_{j=c}^{\infty} \rho^{j-c} p_c = \frac{p_c}{1-\rho}. \tag{12.50}$$

This probability is called the **Erlang *C* formula** and is denoted by $C(c, a)$:

$$C(c, a) = \frac{p_c}{1-\rho} = P[W > 0]. \tag{12.51}$$

The mean number of customers in queue is given by

$$\begin{aligned} E[N_q] &= \sum_{j=c}^{\infty} (j-c)\rho^{j-c} p_c = p_c \sum_{j'=0}^{\infty} j' \rho^{j'} \\ &= \frac{\rho}{(1-\rho)^2} p_c \\ &= \frac{\rho}{1-\rho} C(c, a). \end{aligned} \tag{12.52}$$

The mean waiting time is found from Little's formula:

$$\begin{aligned} E[W] &= \frac{E[N_q]}{\lambda} \\ &= \frac{1/\mu}{c(1-\rho)} C(c, a). \end{aligned} \tag{12.53}$$

The mean total time in the system is

$$E[T] = E[W] + E[\tau] = E[W] + \frac{1}{\mu}. \tag{12.54}$$

Finally, the mean number in the system is found from Little's formula:

$$E[N] = \lambda E[T] = E[N_q] + a, \tag{12.55}$$

where we have used Equation (12.54).

Example 12.8

A company has two 1 Megabit/second lines connecting two of its sites. Suppose that packets for these lines arrive according to a Poisson process at a rate of 150 packets per second, and that packets are exponentially distributed with mean 10 kbits. When both lines are busy, the system queues the packets and transmits them on the first available line. Find the probability that a packet has to wait in queue.

First we need to compute p_0. The system parameters are $c = 2$, $\lambda = 150$ packets/sec, $1/\mu = 10$ kbit/1 Mbit/s $= 10$ ms, $a = \lambda/\mu = 1.5$ and $\rho = \lambda/c\mu = 3/4$. Therefore:

$$p_0 = \left\{ 1 + 1.5 + \frac{(1.5)^2}{2!} \frac{1}{1-3/4} \right\}^{-1} = \frac{1}{7}.$$

The probability of having to wait is then

$$C(2, 1.5) = \frac{(1.5)^2}{2!} p_0 \frac{1}{1 - \rho} = \frac{9}{14}.$$

Example 12.9 M/M/1 Versus M/M/*c*

Compare the mean delay and mean waiting time performance of the two systems shown in Fig. 12.11. Note that both systems have the same processing rate.

For the M/M/1 system, $\rho = \lambda/\mu = (1/2)/1 = 1/2$, so the mean waiting time is

$$E[W] = \frac{\rho/\mu}{1 - \rho} = 1 \text{ s},$$

and the mean total delay is

$$E[T] = \frac{1/\mu}{1 - \rho} = 2 \text{ s}.$$

For the M/M/2 system, $a = \lambda/\mu' = 1$, and $\rho = \lambda/2\mu' = 1/2$. The probability of an empty system is

$$p_0 = \left\{1 + a + \frac{a^2/2}{1 - 1/2}\right\}^{-1} = \frac{1}{3}.$$

The Erlang *C* formula is

$$C(2, 1) = \frac{a^2/2}{1 - \rho} p_0 = \frac{1}{3}.$$

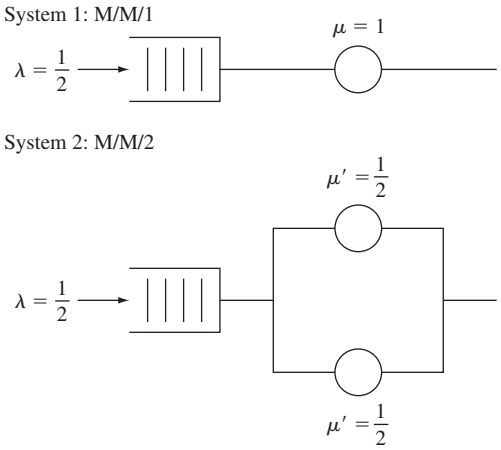

FIGURE 12.11
M/M/1 and M/M/2 systems with the same arrival rate and the same maximum processing rate.

The mean waiting time is then

$$E[W'] = \frac{1/\mu'}{2(1-\rho)}C(2,1) = \frac{2}{3},$$

and the mean delay is

$$E[T'] = \frac{2}{3} + \frac{1}{\mu'} = \frac{8}{3}.$$

Thus the M/M/1 system has a smaller total delay but a larger waiting time than the M/M/2. In general, increasing the number of servers decreases the waiting time but increases the total delay.

12.4.2 Waiting Time Distribution for M/M/c

Before we compute the pdf of the waiting time, consider the conditional probability that there are $j - c > 0$ customers in queue given that all servers are busy (i.e., $N(t) \geq c$):

$$P[N(t) = j \,|\, N(t) \geq c] = \frac{P[N(t) = j, N(t) \geq c]}{P[N(t) \geq c]} = \frac{P[N(t) = j]}{P[N(t) \geq c]} \qquad j \geq c$$

$$= \frac{\rho^{j-c}p_c}{p_c/(1-\rho)} = (1-\rho)\rho^{j-c} \qquad j \geq c. \tag{12.56}$$

This geometric pmf suggests that when all the servers are busy, the M/M/c system behaves like an M/M/1 system. We use this fact to compute the cdf of W.

Suppose that a customer arrives when there are k customers in queue. There must be $k+1$ service completions before our customer enters service. From Eq. (12.43), each service completion is exponentially distributed with rate $c\mu$. Thus the waiting time for our customer is the sum of $k+1$ iid exponential random variables with parameter $c\mu$, which we know is a gamma random variable with parameter $c\mu$:

$$f_W(x \,|\, N = c + k) = \frac{(c\mu x)^k}{k!}c\mu e^{-c\mu x}. \tag{12.57}$$

The cdf for W given that $W > 0$, or equivalently $N \geq c$, is obtained by combining Eqs. (12.56) and (12.57):

$$F_W(x \,|\, W > 0) = \sum_{k=0}^{\infty} F_W(x \,|\, N = c + k)P[N = c + k \,|\, N \geq c]$$

$$= \sum_{k=0}^{\infty} \int_0^x \frac{(c\mu y)^k}{k!}c\mu e^{-c\mu y}\, dy(1-\rho)\rho^k$$

$$= (1-\rho)\int_0^x \sum_{k=0}^{\infty} \frac{(c\mu y)^k}{k!}\rho^k c\mu e^{-c\mu y}\, dy$$

$$= (1-\rho)c\mu \int_0^x e^{-c\mu(1-\rho)y}\, dy$$

$$= 1 - e^{-c\mu(1-\rho)x}.$$

The cdf of W is then

$$\begin{aligned} P[W \le x] &= P[W = 0] + F_W(x \mid W > 0)P[W > 0] \qquad x > 0 \\ &= (1 - C(c, a)) + (1 - e^{-c\mu(1-\rho)x})C(c, a) \\ &= 1 - C(c, a)e^{-c\mu(1-\rho)x}. \end{aligned} \tag{12.58}$$

Since $T = W + \tau$, where W and τ are independent random variables, it is easy to show that if $a \neq c - 1$, the cdf of T is

$$P[T \le x] = 1 + \frac{a - c + P[W = 0]}{c - 1 - a} e^{-\mu x} + \frac{C(c, a)}{c - 1 - a} e^{-c\mu(1-\rho)x}. \tag{12.59}$$

Example 12.10

What is the probability that a packet has to wait more than one minute in the system discussed in Example 12.8?

In Example 12.8 we found that $p_0 = 1/7$ and that the probability of having to wait is

$$C(2, 1.5) = \frac{9}{14}.$$

The probability of having to wait more than one minute is

$$\begin{aligned} P[W > 1] &= 1 - P[W \le 1] \\ &= C(c, a)e^{-c\mu(1-\rho)1} = \frac{9}{14} e^{-200(1/4)(0.040)} \\ &= \frac{9}{14} e^{-2} = 0.3045. \end{aligned}$$

12.4.3 The M/M/*c*/*c* Queueing System

The M/M/c/c queueing system has c servers but no waiting room. Customers that arrive when all servers are busy are turned away. The transition rate diagram for this system is shown in Fig. 12.12, where it can be seen that the arrival rate is zero when $N(t) = c$.

The steady state probabilities for this system have the same form as those for states $0, \ldots, c$ in the M/M/c system:

$$p_j = \frac{a^j}{j!} p_0 \qquad j = 0, \ldots, c, \tag{12.60}$$

where

$$a = \frac{\lambda}{\mu} \tag{12.61}$$

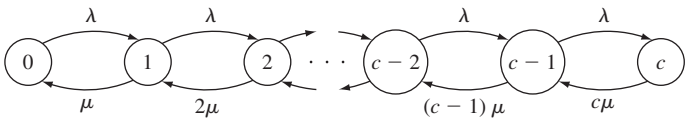

FIGURE 12.12
Transition rate diagram for M/M/c/c system.

is the offered load and

$$p_0 = \left\{ \sum_{j=0}^{c} \frac{a^j}{j!} \right\}^{-1}. \tag{12.62}$$

The **Erlang *B* formula** is defined as the probability that all servers are busy:

$$B(c, a) = P[N = c] = p_c = \frac{a^c/c!}{1 + a + a^2/2! + \cdots + a^c/c!}. \tag{12.63}$$

The actual arrival rate *into* the system is then

$$\lambda_a = \lambda(1 - B(c, a)). \tag{12.64}$$

The average number in the system is obtained from Little's formula:

$$E[N] = \lambda_a E[\tau] = \frac{\lambda}{\mu}(1 - B(c, a)). \tag{12.65}$$

Note that $E[N]$ is also equal to the carried load as defined by Eq. (12.42).

The Erlang *B* formula depends only on the arrival rate λ, the mean service time $E[\tau] = 1/\mu$, and the number of servers c. It turns out that Eq. (12.63) also gives the probability of blocking for M/G/c/c systems (see Ross, 1983).

Example 12.11

A company has five 1 Megabit per second lines to carry videoconferences between two company sites. Suppose that each videoconference requires 1 Mbps and lasts for an average of 1 hour. Assume that requests for videoconferences arrive according to a Poisson process with rate 3 calls per hour. Find the probability that a call request is blocked due to lack of lines.

The offered load is $a = \lambda/\mu = 3$ calls/hr $\times$ 1 hr/call $= 3$. The blocking probability is then:

$$B(5, 3) = \frac{3^5/5!}{1 + 3 + 9/2 + 27/6 + 81/24 + 243/120} = 0.11.$$

The M/M/∞ Queueing System

Consider a system with Poisson arrivals and exponential service times, and suppose that the number of servers is so large that arriving customers always find a server available. In effect we have a system with an infinite number of servers. If we allow c to approach infinity for the M/M/c/c system, we obtain the M/M/∞ system with the transition rate diagram shown in Fig. 12.13.

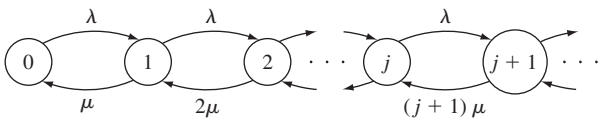

FIGURE 12.13
Transition rate diagram for M/M/∞ system.

The steady state probabilities are also found by letting c approach infinity in the equations for the M/M/c/c system:

$$p_j = \frac{a^j}{j!}e^{-a} \qquad j = 0, 1, 2, \ldots, \tag{12.66}$$

where $a = \lambda/\mu$. Thus the number of customers in the system is a Poisson random variable. The mean number of customers in the system is

$$E[N] = a.$$

Example 12.12

Subscribers connect to a university's online catalog at a rate of 4 subscribers per minute. Sessions have an average duration of 5 minutes. Find the probability that there are more than 25 users online.

The offered load is $a = \lambda/\mu = 4$ subscribers/minute $\times$ 5 minutes/subscriber $= 20$. The pmf for the number of users connected is a Poisson random variable with mean 20. The probability that there are more than 25 in the system is:

$$P[N > 25] = 1 - \sum_{j=0}^{25} \frac{25^j}{j!}e^{-25} = 0.888$$

where we used the Octave function `poisson_cdf(25,20)`.

12.5 FINITE-SOURCE QUEUEING SYSTEMS

Consider a single-server queueing system that serves K sources as shown in Fig. 12.14(a). Each source can be in one of two states: In the first state, the source is preparing a request for service from the server; in the second state, the source has generated a request that is either waiting in queue or being served. For example, the sources could represent K machines and the server could represent a repairman who repairs machines when they break down. In another example, the K sources could represent clients that generate queries for a Web server as shown in Fig. 12.14(b).

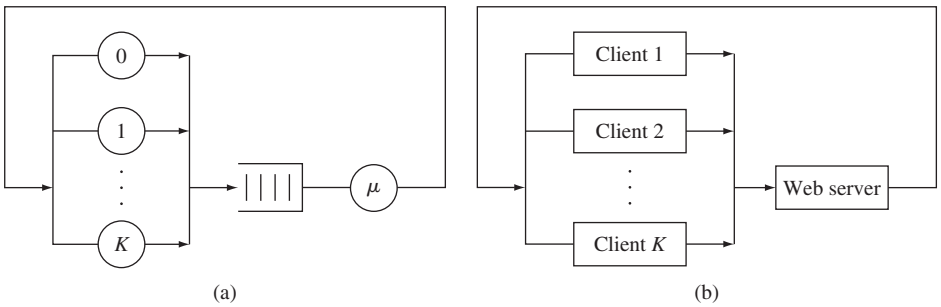

FIGURE 12.14
(a) A finite-source single-server system. (b) A multi-user computer system.

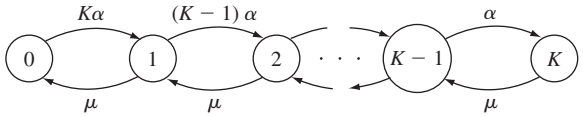

FIGURE 12.15
Transition rate diagram for a finite-source single-server system.

Let $N(t)$ be the number of requests in the system. We assume that each source spends an exponentially distributed amount of time with mean $1/\alpha$ preparing each service request. Thus when idle, a source generates a request for service in the interval $(t, t + \delta)$ with probability $\alpha\delta + o(\delta)$. If the state of the system is $N(t) = k$, then the number of idle sources is $K - k$, so the rate at which service requests are generated is $(K - k)\alpha$. We also assume that the time required to service each request is an exponentially distributed amount of time with mean $1/\mu$. $N(t)$ is then the continuous-time Markov chain with the transition rate diagram shown in Fig. 12.15.

The steady state probabilities are found using the results obtained in Example 11.40:

$$p_k = \frac{K!}{(K-k)!}\left(\frac{\alpha}{\mu}\right)^k p_0 \qquad k = 0, 1, \ldots, K, \tag{12.67}$$

where

$$p_0 = \left\{\sum_{k=0}^{K} \frac{K!}{(K-k)!}\left(\frac{\alpha}{\mu}\right)^k\right\}^{-1}. \tag{12.68}$$

We first compute the mean arrival rate λ and the mean delay $E[T]$ indirectly. In the last part of the section we show how they can be calculated directly. The server utilization ρ is the proportion of time when the system is busy, thus

$$\rho = 1 - p_0, \tag{12.69}$$

where p_0 is given by Eq. (12.68). The mean arrival rate to the queue can then be found from Little's formula with "system" defined as the server:

$$\lambda E[\tau] = \rho = 1 - p_0,$$

which implies

$$\lambda = \frac{\rho}{E[\tau]} = \mu\rho = \mu(1 - p_0). \tag{12.70}$$

A source takes an average time of $1/\alpha$ to generate a request and then spends time $E[T]$ having it serviced in the queueing system. Thus each source generates a request at the rate $(1/\alpha + E[T])^{-1}$ requests per second. Since the actual arrival rate must equal the rate at which the K sources generate requests, we have

$$\lambda = \frac{K}{1/\alpha + E[T]}. \tag{12.71}$$

The mean delay in the system for each request is found by solving for $E[T]$:

$$E[T] = \frac{K}{\lambda} - \frac{1}{\alpha}. \tag{12.72}$$

Finally, we can apply Little's formula to Eq. (12.72) to obtain the mean number in the system:

$$E[N] = \lambda E[T] = K - \frac{\lambda}{\alpha}. \tag{12.73}$$

Note that this implies that λ/α is the mean number of idle sources. The mean waiting time is obtained by subtracting the mean service time from $E[T]$:

$$E[W] = E[T] - \frac{1}{\mu}. \tag{12.74}$$

The proportion of time that a source spends waiting for the completion of a service request is the ratio of the time spent in the system to the mean cycle time:

$$P[\text{source busy}] = \frac{E[T]}{E[T] + 1/\alpha}. \tag{12.75}$$

Example 12.13 Web Server System

Some Web server designs place a limit K on the number of clients that can interact with it at any given time. The set of K clients generate queries to the Web server as follows. Each client spends an exponentially distributed "think" time preparing a transaction request, and the server takes an exponentially distributed time processing each request. The "throughput" of the server is defined as the rate at which it completes transactions. The response time is the total time a transaction spends in the server. Find expressions for the throughput and response time for two extreme cases: K small and K large.

When K is sufficiently small, there is no waiting in queue, so

$$E[T] \simeq \frac{1}{\mu} \qquad \text{for } K \text{ small}, \tag{12.76}$$

and by Eq. (12.71),

$$\lambda = \frac{K}{1/\alpha + 1/\mu} \qquad \text{for } K \text{ small}. \tag{12.77}$$

Thus λ grows linearly with K. As K increases, the server eventually becomes fully utilized, and then answers queries at its maximum rate, namely μ transactions per second. Thus

$$\lambda \simeq \mu \qquad \text{for } K \text{ large}, \tag{12.78}$$

and Eq. (12.72) becomes

$$E[T] = \frac{K}{\mu} - \frac{1}{\alpha} \qquad \text{for } K \text{ large}. \tag{12.79}$$

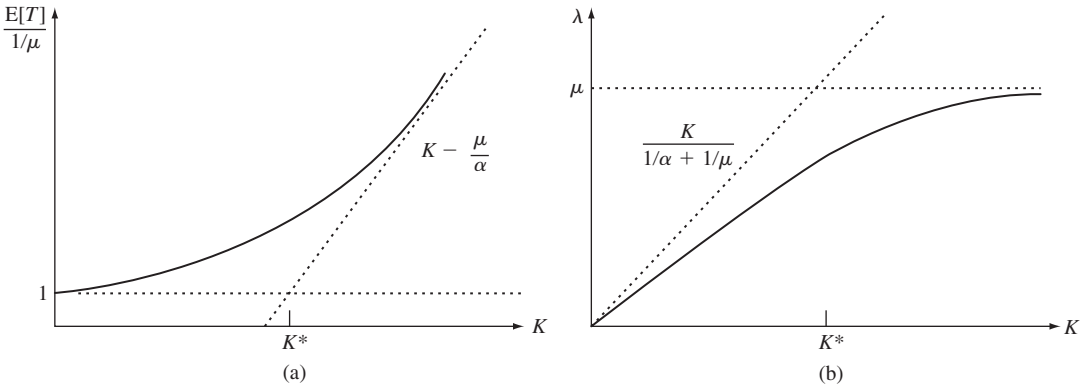

FIGURE 12.16
Delay and throughput for finite-source system as a function of number of sources. Dashed lines show small-K and large-K asymptotes.

These asymptotic expressions for the throughput and response time are shown in Fig. 12.16(a) and (b). The value of K where the two asymptotes for $E[T]$ intersect is called the *system saturation point,*

$$K^* = \frac{1/\mu + 1/\alpha}{1/\mu}. \tag{12.80}$$

When K becomes larger than K^*, the queries from different clients are certain to interfere with one another and the response time increases accordingly.

12.5.1 *Arriving Customer's Distribution

In the above discussion, we found λ, $E[N]$, and $E[T]$ in a roundabout way (see Eqs. 12.70, 12.71, and 12.72). To calculate $E[T]$ directly, we argue as follows. If we assume a first-come, first-served service discipline, then a customer who arrives when there are $N_a = k$ requests in the queueing system spends a total time in the system equal to the sum of 1 residual service time, $k - 1$ service times, and the customer's own service time. Since all of these times are iid exponential random variables with mean $1/\mu$, the mean time in the system for our request is

$$E[T \mid N_a = k] = \frac{k + 1}{\mu}.$$

The mean time in the system is then found by averaging over N_a:

$$E[T] = \frac{1}{\mu} \sum_{k=0}^{K-1} (k + 1) P[N_a = k]. \tag{12.81}$$

The difficulty with the above equation is that arrivals are not Poisson—remember that the arrival rate is $(K - N(t))\alpha$, and thus depends on the state of the system. Consequently, the distribution of states seen by an arriving customer is not the same as

$P[N = k]$, the proportion of time that there are k requests in the queueing system. For example, a service request cannot be generated when all sources have requests in the system, that is, $N(t) = K$, so $P[N_a = K] = 0$. However, $P[N = K]$ is nonzero since it is possible for all sources to have requests in the queueing system simultaneously.

To find $P[N_a = k]$ we need to find the long-term proportion of time that arriving customers find k customers in the system. Since $p_k = P[N(t) = k]$ is the long-term proportion of time the system is in state k, then in a very long time interval of duration T' approximately $p_k T'$ seconds are spent in state k. The arrival rate when $N(t) = k$ is $(K - k)\alpha$ requests per second, so the number of arrivals that find k requests is approximately

$$(K - k)\alpha \text{ customers/second} \times p_k T' \text{ seconds in state } k. \tag{12.82}$$

The total number of arrivals in time T' is obtained by summing over all states:

$$\sum_{j=0}^{K} (K - j)\alpha p_j T'. \tag{12.83}$$

Thus *the proportion of arrivals that find k requests in the system is*

$$\begin{aligned} P[N_a = k] &= \frac{(K - k)\alpha p_k T'}{\sum_{j=0}^{K} (K - j)\alpha p_j T'} = \frac{(K - k)p_k}{\sum_{j=0}^{K} (K - j)p_j} \\ &= \frac{(K - k)[K!/(K - k)!](\alpha/\mu)^k p_0}{\sum_{j=0}^{K} (K - j)[K!/(K - j)!](\alpha/\mu)^j p_0} \\ &= \frac{[(K - 1)!/(K - k - 1)!](\alpha/\mu)^k}{\sum_{j=0}^{K-1} [(K - 1)!/(K - j - 1)!](\alpha/\mu)^j} \qquad 0 \le k \le K - 1. \end{aligned} \tag{12.84}$$

If we compare Eq. (12.84) with Eq. (12.67), we see that Eq. (12.84) is the steady state probability of having k customers in a system with $K - 1$ sources. In other words, *a source when placing a request "sees" a queueing system that behaves as if the source were not present at all!*

We leave it up to you in Problem 12.37 to show that Eqs. (12.84) and (12.81) give $E[T]$ as given in Eq. (12.72). Indeed, this same approach can be used to find the pdf of T.

12.6 M/G/1 QUEUEING SYSTEMS

We now consider single-server queueing systems in which the arrivals follow a Poisson process but in which the service times need not be exponentially distributed. We assume that the service times are independent, identically distributed random variables with general pdf $f_\tau(x)$. The resulting queueing system is denoted by M/G/1.

The number of customers $N(t)$ in an M/G/1 system is a continuous-time random process. Recall that the "state" of the system is the information about the past history

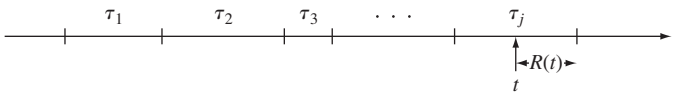

FIGURE 12.17
Sequence of service times and a residual service time.

of the system that is relevant to the probabilities of future events. In the preceding sections, customer interarrival times and service times were exponential distributions, so $N(t)$ was always the state of the system. This is no longer the case for M/G/1 systems. For example, if service times are constant, then knowledge about when a customer began service specifies the customer's future departure time. Thus the state of an M/G/1 system at time t is specified by $N(t)$ together with the remaining ("residual") service time of the customer being served at time t.

In this section we present a simple approach based on Little's formula that gives the mean waiting time and mean delay in an M/G/1 system. We also use this simple approach to find the mean waiting times in M/G/1 systems that have priority classes.

12.6.1 The Residual Service Time

Suppose that an arriving customer finds the server busy, and consider the residual time of the customer found in service. Let $\tau_1, \tau_2, \ldots$ be the iid sequence of service times of customers in this M/G/1 system, and suppose we divide the positive time axis into segments of length $\tau_1, \tau_2, \ldots$ as shown in Fig. 12.17. We can then view customers who arrive when the server is busy as picking a point at random on this time axis. The **residual service time** is then the remainder of time in the segment that is intercepted as shown in Fig. 12.17.

In Example 7.21 we showed that the long-term proportion of time that the residual service time exceeds x is given by

$$\frac{1}{E[\tau]}\int_x^\infty (1 - F_\tau(y))\, dy. \tag{12.85}$$

Since the arrival times of Poisson customers are independent of the system state, Eq. (12.85) is also the probability that the residual service time R of a customer found in service exceeds x, that is,

$$P[R > x] = \frac{1}{E[\tau]}\int_x^\infty (1 - F_\tau(y))\, dy. \tag{12.86}$$

The pdf of R is then

$$f_R(x) = -\frac{d}{dx}P[R > x] = \frac{1 - F_\tau(x)}{E[\tau]}. \tag{12.87}$$

The mean residual time is

$$E[R] = \int_0^\infty x\frac{1 - F_\tau(x)}{P[\tau]}\, dx.$$

Integrating by parts with $u = (1 - F_\tau(x))/E[\tau]$ and $dv = x\,dx$, we obtain

$$E[R] = (1 - F_\tau(x))\frac{x^2}{2E[\tau]}\bigg|_0^\infty + \frac{1}{2E[\tau]}\int_0^\infty x^2 f_\tau(x)\,dx$$
$$= \frac{E[\tau^2]}{2E[\tau]}. \tag{12.88}$$

Example 12.14

Compare the residual service times of two systems with exponential service times of mean m and constant service times of mean m, respectively.

For an exponential service time of mean m, the second moment is $2m^2$, thus the mean residual service time is, from Eq. (12.88),

$$E[R_{\text{exp}}] = \frac{2m^2}{2m} = m.$$

Thus the mean residual time is the same as the full service time of a customer. This is consistent with the memoryless property of the exponential random variable.

The second moment of a constant random variable of value m is m^2. Thus the mean residual service time is

$$E[R_{\text{const}}] = \frac{m^2}{2m} = \frac{m}{2},$$

which is what one would expect; on the average we expect to wait half a service time.

12.6.2 Mean Delay in M/G/1 Systems

Consider the time W spent by a customer waiting for service in an M/G/1 system. If the service discipline is first come, first served, then W is the sum of the residual service time R' of the customer (if any) found in service and the $N_q(t) = k - 1$ service times of the customers (if any) found in queue. Thus the mean waiting time is then

$$E[W] = E[R'] + E[N_q(t)]E[\tau], \tag{12.89}$$

since the service times are iid with mean $E[\tau]$ (see Eq. 7.13). From Little's formula we have that $E[N_q(t)] = \lambda E[W]$, so

$$E[W] = E[R'] + \lambda E[W]E[\tau] = E[R'] + \rho E[W]. \tag{12.90}$$

The residual service time R' encountered by an arriving customer is zero when the system is found empty, and R, as defined in the previous section, when a customer is found in service. Thus

$$E[R'] = 0P[N(t) = 0] + E[R](1 - P[N(t) = 0])$$
$$= \frac{E[\tau^2]}{2E[\tau]}\lambda E[\tau]$$
$$= \frac{\lambda E[\tau^2]}{2}, \tag{12.91}$$

where we have used Eq. (12.88) for $E[R]$ and Eq. (12.14) for the fact that $1 - P[N(t) = 0] = \rho = \lambda E[\tau]$.

The **mean waiting time $E[W]$** of a customer in an M/G/1 system is found by substituting Eq. (12.91) into Eq. (12.90) and solving for $E[W]$:

$$E[W] = \frac{\lambda E[\tau^2]}{2(1 - \rho)}. \tag{12.92}$$

We can obtain another expression for $E[W]$ by noting that $E[\tau^2] = \sigma_\tau^2 + E[\tau]^2$:

$$\begin{aligned} E[W] &= \frac{\lambda(\sigma_\tau^2 + E[\tau]^2)}{2(1 - \rho)} = \lambda E[\tau]^2 \frac{(1 + C_\tau^2)}{2(1 - \rho)} \\ &= \frac{\rho(1 + C_\tau^2)}{2(1 - \rho)} E[\tau], \end{aligned} \tag{12.93}$$

where $C_\tau^2 = \sigma_\tau^2 / E[\tau]^2$ is the coefficient of variation of the service time. Equation (12.93) is called the **Pollaczek–Khinchin mean value formula**.

The **mean delay $E[T]$** is found by adding the mean service time to $E[W]$:

$$E[T] = E[\tau] + E[\tau] \frac{\rho(1 + C_\tau^2)}{2(1 - \rho)}. \tag{12.94}$$

From Eqs. (12.93) and (12.94) we can see that the mean waiting time and mean delay time are affected not only by the mean service time and the server utilization but also by the coefficient of variation of the service time. Thus the degree of randomness of the service times as measured by C_τ^2 affects these delays.[1]

Example 12.15

Compare $E[W]$ for the M/M/1 and M/D/1 systems. The second moments of the exponential and constant random variables were found in Example 12.14. The exponential service time has a coefficient of variation equal to one. Thus Eq. (12.93) implies

$$E[W_{\text{M/M/1}}] = \frac{\rho}{(1 - \rho)} E[\tau]. \tag{12.95}$$

The constant service time has zero variance, so its coefficient of variation is zero. Thus

$$E[W_{\text{M/D/1}}] = \frac{\rho}{2(1 - \rho)} E[\tau]. \tag{12.96}$$

Thus we see that the waiting time in an M/D/1 is half that in an M/M/1 system.

---

[1]On the other hand, it is rather surprising that only the first two moments of the distribution of the service time affect $E[W]$ and $E[T]$.

12.6.3 Mean Delay in M/G/1 Systems with Priority Service Discipline

Consider a queueing system that handles K priority classes of customers. Type k customers arrive according to a Poisson process of rate λ_k and have service times with pdf $f_{\tau_k}(x)$ and mean $E[\tau_k]$. A separate queue is kept for each priority class, and each time the server becomes available it selects the next customer from the highest-priority nonempty queue. This service discipline is often referred to as **"head-of-line priority service."** We assume that customers cannot be preempted once their service has begun.

The server utilization from type k customers is

$$\rho_k = \lambda_k E[\tau_k].$$

We assume that the total server utilization is less than 1:

$$\rho = \rho_1 + \cdots + \rho_K < 1. \tag{12.97}$$

If this is not the case, one or more of the lower-priority queues become unstable, that is, grow without bound.

Consider the mean waiting time W_1 of the highest-priority (type 1) customer. If an arriving type 1 customer finds $N_{q_1}(t) = k_1$ type 1 customers in queue and if the service discipline is first come, first served within each class, then W_1 is the sum of the residual service time R'' of the customer (if any) found in service and the $N_{q_1}(t) = k_1$ service times of the type 1 customers (if any) found in queue. Thus

$$E[W_1] = E[R''] + E[N_{q_1}]E[\tau_1].$$

Following the same development that followed Eq. (12.89) in the previous section, we arrive at the following expression for the **mean waiting time for type 1 customers**:

$$E[W_1] = \frac{E[R'']}{1 - \rho_1}. \tag{12.98}$$

If an arriving type 2 customer finds $N_{q_1}(t) = k_1$ type 1 and $N_{q_2}(t) = k_2$ type 2 customers waiting in queue, then W_2 is the sum of the residual service time R'' of the customer (if any) found in service, the k_1 service times of the type 1 customers (if any) found in queue, the service times of the k_2 type 2 customers found in queue, *and* the service times of the higher-priority type 1 customers who arrive while our customer is waiting in queue. Thus

$$E[W_2] = E[R''] + E[N_{q_1}]E[\tau_1] + E[N_{q_2}]E[\tau_2] + E[M_1]E[\tau_1], \tag{12.99}$$

where M_1 denotes the number of type 1 arrivals during our customer's waiting time. By Little's formula we have $E[N_{q_1}] = \lambda_1 E[W_1]$ and $E[N_{q_2}] = \lambda_2 E[W_2]$. In addition, the mean number of type 1 arrivals during $E[W_2]$ seconds is $E[M_1] = \lambda_1 E[W_2]$. Substituting these expressions in Eq. (12.99) gives

$$E[W_2] = E[R''] + \rho_1 E[W_1] + \rho_2 E[W_2] + \rho_1 E[W_2].$$

Solving for $E[W_2]$,

$$E[W_2] = \frac{E[R''] + \rho_1 E[W_1]}{1 - \rho_1 - \rho_2} = \frac{E[R'']}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}, \tag{12.100}$$

where we have used Eq. (12.98) for $E[W_1]$.

If there are more than two classes of customers, the above method can be used to show that the mean waiting time for a type k customer is

$$E[W_k] = \frac{E[R'']}{(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_k)}. \tag{12.101}$$

The customer found in service by an arriving customer can be of any type, so R'' is the residual service time of customers of all types:

$$E[R''] = \frac{\lambda E[\tau^2]}{2}, \tag{12.102}$$

where λ is the total arrival rate,

$$\lambda = \lambda_1 + \cdots + \lambda_K, \tag{12.103}$$

and $E[\tau^2]$ is the second moment of the service time of customers of all types. The fraction of customers who are type k is λ_k/λ, thus

$$E[\tau^2] = \frac{\lambda_1}{\lambda} E[\tau_1^2] + \cdots + \frac{\lambda_K}{\lambda} E[\tau_K^2]. \tag{12.104}$$

We finally arrive at the following expression for the **mean waiting time for type *k* customers**:

$$E[W_k] = \frac{\sum_{j=1}^{K} \lambda_j E[\tau_j^2]}{2(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_k)}. \tag{12.105}$$

The **mean delay for type *k* customers** is then

$$E[T_k] = E[W_k] + E[\tau_k]. \tag{12.106}$$

Equation (12.105) reveals the effect of the priority classes on one another. Class k customers are affected by lower-priority customers only through the residual-service-time term in the numerator. On the other hand, if the server utilization of the first $k - 1$ classes exceeds one, then the queue for class k customers is unstable.

Example 12.16

A computer handles two types of jobs. Type 1 jobs require a constant service time of 1 ms, and type 2 jobs require an exponentially distributed amount of time with mean 10 ms. Find the mean waiting time if the system operates as follows: (1) an ordinary M/G/1 system and (2) a two-priority M/G/1 system with priority given to type 1 jobs. Assume that the arrival rates of the two classes are Poisson with the same rate.

The first two moments of the service time are

$$E[\tau] = \frac{1}{2}E[\tau_1] + \frac{1}{2}E[\tau_2] = 5.5$$
$$E[\tau^2] = \frac{1}{2}E[\tau_1^2] + \frac{1}{2}E[\tau_2^2] = \frac{1}{2}(1^2 + 2(10^2)) = 100.5.$$

The traffic intensity for each class and the total traffic intensity are

$$\rho_1 = 1\frac{\lambda}{2}, \qquad \rho_2 = 10\frac{\lambda}{2}, \quad \text{and}$$
$$\rho = \lambda E[\tau] = 5.5\lambda,$$

where λ is the total arrival rate. The mean residual service time is then

$$E[R] = \frac{\lambda E[\tau^2]}{2} = 50.25\lambda.$$

From Eq. (12.92), the mean waiting time for an M/G/1 system is

$$E[W] = \frac{E[R]}{1 - \rho} = \frac{50.25\lambda}{1 - 5.5\lambda}. \tag{12.107}$$

For the priority system we have

$$E[W_1] = \frac{E[R]}{1 - \rho_1} = \frac{50.25\lambda}{1 - 0.5\lambda} \tag{12.108}$$

and

$$E[W_2] = \frac{E[R]}{(1 - \rho_1)(1 - \rho)} = \frac{50.25\lambda}{(1 - 0.5\lambda)(1 - 5.5\lambda)}. \tag{12.109}$$

Comparison of Eqs. (12.108) and (12.109) with Eq. (12.107) shows that the waiting time of type 1 customers is improved by a factor of $(1 - \rho)/(1 - \rho_1)$ and that of type 2 is worsened by the factor $1/(1 - \rho_1)$.

The overall mean waiting for the priority system is

$$E[W_p] = \frac{1}{2}E[W_1] + \frac{1}{2}E[W_2] = \frac{1}{2}\left(\frac{E[R]}{1 - \rho_1}\right)\left(1 + \frac{1}{1 - \rho}\right)$$
$$= \left(\frac{1 - \rho/2}{1 - \rho_1}\right)\left(\frac{E[R]}{1 - \rho}\right)$$
$$= \frac{1 - 2.75\lambda}{1 - 0.5\lambda}E[W],$$

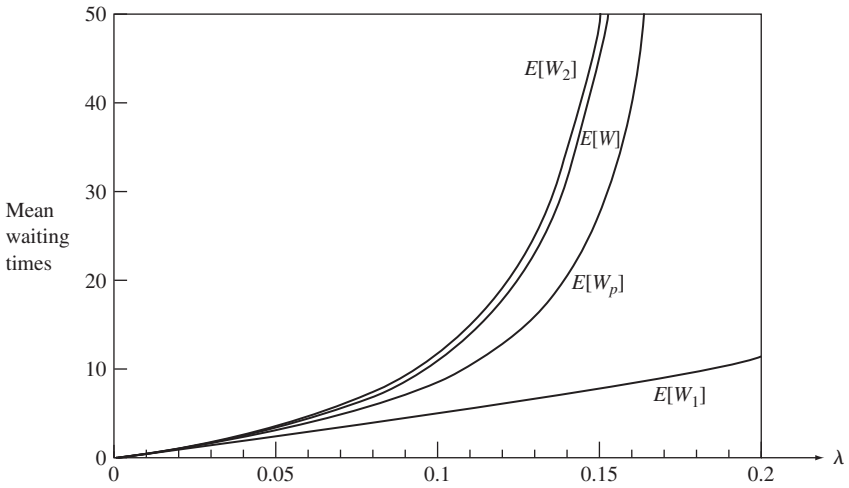

FIGURE 12.18
Relative mean waiting times in priority and nonpriority M/G/1 systems: $E[W]$, mean waiting time in M/G/1 system; $E[W_1]$, $E[W_2]$, mean waiting time for type 1 and type 2 customers in priority system; $E[W_p]$, overall mean waiting time in priority system.

where $E[W]$ is the mean waiting time of the M/G/1 system without priorities. Figure 12.18 shows $E[W]$, $E[W_p]$, $E[W_1]$, and $E[W_2]$. It can be seen that the discipline "short-job type first" used here improves the average waiting time. The graphs for $E[W_1]$ and $E[W_2]$ also show that at $\lambda = 2/11$ the lower-priority queue becomes unstable but the higher-priority remains stable up to $\lambda = 2$.

12.7 M/G/1 ANALYSIS USING EMBEDDED MARKOV CHAINS

In the previous section we noted that the state of an M/G/1 queueing system is given by the number of customers in the system $N(t)$ and the residual service time of the customer in service. Suppose we observe $N(t)$ at the instants when the residual service time becomes zero (i.e., at the instants D_j when the jth service completion occurs); then all of the information relevant to the probability of future events is embodied in $N_j = N(D_j)$, *the number of customers left behind by the jth departing customer*. We will show that the sequence N_j is a discrete-time Markov chain and that the steady state pmf at customer departure instants is equal to the steady state pmf of the system at arbitrary time instants. Thus we can find the steady state pmf of $N(t)$ if we can find the steady state pmf for the chain N_j.

12.7.1 The Embedded Markov Chain

First we show that the sequence $N_j = N(D_j)$ is a Markov chain. Consider the relation between N_j and N_{j-1}. If $N_{j-1} \geq 1$, then a customer enters service immediately at time D_j, as shown in Fig. 12.19(a), and N_j equals N_{j-1}, minus the customer that is served in

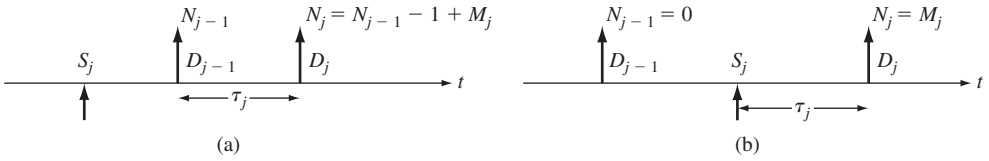

FIGURE 12.19
(a) Customer $j - 1$ leaves the system nonempty at time D_{j-1}. (b) Customer $j - 1$ leaves the system empty at time D_{j-1}.

between, plus the number of customers M_j that arrive during the service time of the jth customer:

$$N_j = N_{j-1} - 1 + M_j \qquad \text{if } N_{j-1} \geq 1. \tag{12.110a}$$

If $N_{j-1} = 0$, then as shown in Fig. 12.19(b), there are no departures until the jth customer arrives and completes his service; N_j then is the number of customers who arrive during this service time:

$$N_j = M_j \qquad \text{if } N_{j-1} = 0. \tag{12.110b}$$

Thus we see that N_j depends on the past only through N_{j-1} and M_j. The M_j form an iid sequence because the service times are iid and because of the memoryless property of Poisson arrivals. Thus N_j depends on the past of the system only through N_{j-1}. We therefore conclude that the sequence N_j is a Markov chain.

Next we need to show that the steady state pmf of $N(t)$ is the same as the steady state pmf of N_j. We do so in two steps: first, we show that in M/G/1 systems, the distribution of customers found by arriving customers is the same as that left behind by departing customers; second, we show that in M/G/1 systems, the distribution of customers found by arriving customers is the same as the steady state distribution of $N(t)$. It then follows that the steady state pmf's of $N(t)$ and N_j are the same.

First we need to show that *for systems in which customers arrive one at a time and depart one at a time (i.e., M/G/1 systems) the distribution found by arriving customers is the same as that left behind by departing customers*. Let $U_n(t)$ be the number of times the system goes from n to $n + 1$ in the interval $(0, t)$; then $U_n(t)$ is the number of times an arriving customer finds n customers in the system. Similarly, let $V_n(t)$ be the number of times that the system goes from $n + 1$ to n; then $V_n(t)$ is the number of times a departing customer leaves n. Note that the transition n to $n + 1$ cannot reoccur until after the number in the system drops to n once more (i.e., until after the transition $n + 1$ to n reoccurs). Thus $U_n(t)$ and $V_n(t)$ can differ by at most 1. As t becomes large, both of these transitions occur a large number of times, so the rate of transitions from n to $n + 1$ equals the rate from $n + 1$ to n. Thus the rate at which customer arrivals find n in the system equals the rate at which departures leave n in the system. It then follows that the probability that an arrival finds n in the system is equal to the probability that a departure leaves n behind.

Since the arrivals in an M/G/1 system are Poisson and independent of the customer service times, the customer arrival times are independent of the state of the system. Thus the probability that an arrival finds n customers in the system is equal to the proportion of time the system has n customers, that is, the steady state probability $P[N(t) = n]$. Thus *the distribution of states seen by arriving customers is the same as the steady state distribution*.

By combining the results from the two previous paragraphs, we have that for an M/G/1 system, the pmf of N_j, the state at customer departure points, is the same as the steady state pmf of $N(t)$. In the next section, we find the generating function of N_j and thus of $N(t)$.

12.7.2 The Number of Customers in an M/G/1 System

We now find the generating function for the steady state pmf of N_j. The transition probabilities for N_j can be deduced from Eqs. (12.110a) and (12.110b):

$$p_{ik} = P[N_j = k \,|\, N_{j-1} = i] = P[M_j = k - i + 1] \qquad i > 0 \tag{12.111a}$$

$$p_{0k} = P[N_j = k \,|\, N_{j-1} = 0] = P[M_j = k]. \tag{12.111b}$$

Note that $p_{ik} = 0$ for $k - i + 1 < 0$. The probability that there are $N_j = k$ customers in the system at time j is

$$\begin{aligned} P[N_j = k] &= \sum_{i=0}^{\infty} P[N_{j-1} = i] p_{ik} \\ &= P[N_{j-1} = 0] P[M_j = k] \\ &\quad + \sum_{i=1}^{k+1} P[N_{j-1} = i] P[M_j = k + 1 - i] \qquad (12.112a) \\ &= P[N_{j-1} = 0] P[M_j = k] \\ &\quad + \sum_{i=1}^{\infty} P[N_{j-1} = i] P[M_j = k + 1 - i], \qquad (12.112b) \end{aligned}$$

where we have used the fact that $P[M_j = k + 1 - i] = 0$ for $i > k + 1$.

If the process N_j reaches a steady state as $j \to \infty$, then $P[N_j = k] \to P[N_d = k]$ and the above equation becomes

$$\begin{aligned} P[N_d = k] &= P[N_d = 0] P[M = k] \\ &\quad + \sum_{i=1}^{\infty} P[N_d = i] P[M = k + 1 - i], \end{aligned} \tag{12.113}$$

where N_d denotes the number of customers left behind by a departing customer.

Since the steady state pmf of N_j is equal to that of $N(t)$, Eq. (12.113) also holds for the steady state pmf of $N(t)$. Equation (12.113) is readily solved for the generating function of $N(t)$ by using the probability generating function. The generating functions for N and for M are given by

$$G_N(z) = \sum_{k=0}^{\infty} P[N = k]z^k \qquad \text{and} \qquad G_M(z) = \sum_{k=0}^{\infty} P[M = k]z^k.$$

We multiply both sides of Eq. (12.113) (with N_d replaced by N) by z^k and sum from 0 to infinity:

$$\sum_{k=0}^{\infty} P[N = k]z^k = \sum_{k=0}^{\infty} P[N = 0]P[M = k]z^k + \sum_{k=0}^{\infty}\sum_{i=1}^{\infty} P[N = i]P[M = k + 1 - i]z^k. \quad (12.114)$$

The generating functions for N and M are immediately recognizable in the first two summations:

$$G_N(z) = P[N = 0]G_M(z) + z^{-1}\sum_{i=1}^{\infty} P[N = i]z^i \sum_{k=0}^{\infty} P[M = k + 1 - i]z^{k+1-i}.$$

The first summation is the generating function for N with the $i = 0$ term missing. Let $k' = k + 1 - i$ in the second summation and note that $P[M = k'] = 0$ for $k' < 0$, then

$$\begin{aligned} G_N(z) &= P[N = 0]G_M(z) + z^{-1}\{G_N(z) - P[N = 0]\}\left\{\sum_{k'=0}^{\infty} P[M = k']z^{k'}\right\} \\ &= P[N = 0]G_M(z) + z^{-1}(G_N(z) - P[N = 0])G_M(z). \end{aligned} \quad (12.115)$$

The generating function for N is found by solving for $G_N(z)$:

$$G_N(z) = \frac{P[N = 0](z - 1)G_M(z)}{z - G_M(z)}. \quad (12.116)$$

We can find $P[N = 0]$ by noting that as $z \to 1$, we must have

$$G_N(z) = \sum_{k=0}^{\infty} P[N = k]z^k \to 1. \quad (12.117)$$

When we take the limit $z \to 1$ in Eq. (12.116) we obtain zero for the numerator and the denominator. By applying L'Hopital's rule, we obtain

$$1 = P[N = 0]\left.\frac{G_M(z) + (z - 1)G'_M(z)}{1 - G'_M(z)}\right|_{z=1} = \frac{P[N = 0]}{1 - E[M]}. \quad (12.118)$$

Thus

$$P[N = 0] = 1 - E[M] \quad (12.119)$$

and

$$G_N(z) = \frac{(1 - E[M])(z - 1)G_M(z)}{z - G_M(z)}. \quad (12.120)$$

Note from Eq. (12.119) that we must have $E[M] < 1$ since $P[N = 0] \geq 0$. This stability condition makes sense since it implies that on the average less than one customer should arrive during the time it takes to service a customer.

We now determine $G_M(z)$, the generating function for the number of arrivals during a service time:

$$\begin{aligned} G_M(z) &= \sum_{k=0}^{\infty} P[M = k] z^k \\ &= \sum_{k=0}^{\infty} \int_0^{\infty} P[M = k \,|\, \tau = t] f_\tau(t)\, dt\, z^k. \end{aligned} \tag{12.121a}$$

Noting that the number of arrivals in t seconds is a Poisson random variable,

$$\begin{aligned} G_M(z) &= \sum_{k=0}^{\infty} \int_0^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} f_\tau(t)\, dt\, z^k \\ &= \int_0^{\infty} e^{-\lambda t} f_\tau(t) \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} z^k\, dt \\ &= \int_0^{\infty} e^{-\lambda t} f_\tau(t) e^{\lambda t z}\, dt \\ &= \int_0^{\infty} e^{-\lambda(1-z)t} f_\tau(t)\, dt \\ &= \hat{\tau}(\lambda(1 - z)), \end{aligned} \tag{12.121b}$$

where $\hat{\tau}(s)$ is the Laplace transform of the pdf of τ:

$$\hat{\tau}(s) = \int_0^{\infty} e^{-st} f_\tau(t)\, dt. \tag{12.122}$$

We can obtain the moments of M by taking derivatives of $G_M(z)$:

$$\begin{aligned} E[M] &= \frac{d}{dz} G_M(z)\bigg|_{z=1} = \frac{d}{du}\hat{\tau}(u) \frac{d}{dz}\lambda(1 - z)\bigg|_{z=1} \\ &= \hat{\tau}'(\lambda(1 - z))(-\lambda)|_{z=1} \\ &= -\lambda\hat{\tau}'(0) = \lambda E[\tau] = \rho, \end{aligned} \tag{12.123}$$

where we used the chain rule in the second equality. Similarly,

$$E[M(M - 1)] = \lambda^2\hat{\tau}''(0) = \lambda^2 E[\tau^2].$$

Thus

$$\begin{aligned} \sigma_M^2 &= E[M^2] - E[M]^2 = \lambda^2 E[\tau^2] + \lambda E[\tau] - (\lambda E[\tau])^2 \\ &= \lambda^2\sigma_\tau^2 + \lambda E[\tau]. \end{aligned} \tag{12.124}$$

If we substitute Eqs. (12.123) and (12.121b) into Eq. (12.120), we obtain the **Pollaczek–Khinchin transform equation**,

$$G_N(z) = \frac{(1-\rho)(z-1)\hat{\tau}(\lambda(1-z))}{z - \hat{\tau}(\lambda(1-z))}. \tag{12.125}$$

Note that $G_N(z)$ depends on the utilization ρ, the arrival rate λ, and the Laplace transform of the service time pdf.

Example 12.17 M/M/1 System

Use the Pollaczek–Khinchin transform formula to find the pmf for $N(t)$ for an M/M/1 system.

The Laplace transform for the pdf of an exponential service of mean $1/\mu$ is

$$\hat{\tau}(s) = \frac{\mu}{s+\mu}.$$

Thus the Pollaczek–Khinchin transform formula is

$$\begin{aligned} G_N(z) &= \frac{(1-\rho)(z-1)[\mu/(\lambda(1-z)+\mu)]}{z - [\mu/(\lambda(1-z)+\mu)]} \\ &= \frac{(1-\rho)(z-1)\mu}{(\lambda - \lambda z + \mu)z - \mu} = \frac{1-\rho}{1-\rho z}, \end{aligned}$$

where we canceled the $z-1$ term from the numerator and denominator and noted that $\rho = \lambda/\mu$. By expanding $G_N(z)$ in a power series, we have

$$G_N(z) = \sum_{k=0}^{\infty}(1-\rho)\rho^k z^k = \sum_{k=0}^{\infty} P[N=k]z^k,$$

which implies that the steady state pmf is

$$P[N=k] = (1-\rho)\rho^k \qquad k = 0, 1, 2, \ldots,$$

which is in agreement with our previous results for the M/M/1 system.

Example 12.18 M/H_2/1 System

Find the pmf for the number of customers in an M/G/1 system that has arrivals of rate λ and where the service times are hyperexponential random variables of degree two, as shown in Fig. 12.20. In other words, with probability 1/9 the service time is exponentially distributed with mean $1/\lambda$, and with probability 8/9 the service time is exponentially distributed with mean $1/2\lambda$.

In order to find $\hat{\tau}(s)$ we note that the pdf of τ is

$$f_\tau(x) = \frac{1}{9}\lambda e^{-\lambda x} + \frac{8}{9}2\lambda e^{-2\lambda x} \qquad x > 0.$$

Thus the mean service time is

$$E[\tau] = \frac{1}{9\lambda} + \frac{8}{9(2\lambda)} = \frac{5}{9\lambda},$$

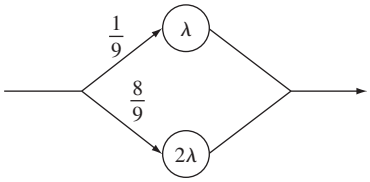

FIGURE 12.20
A hyperexponential service time results if we select an exponential service time of rate λ with probability 1/9 and an exponential service time of rate 2λ with probability 8/9.

and the server utilization is $\rho = \lambda E[\tau] = 5/9$. The Laplace transform of $f_\tau(x)$ is

$$\hat{\tau}(s) = \frac{1}{9}\frac{\lambda}{s+\lambda} + \frac{8}{9}\frac{2\lambda}{s+2\lambda} = \frac{18\lambda^2 + 17\lambda s}{9(s+\lambda)(s+2\lambda)}.$$

Substitution of $\hat{\tau}(\lambda(1-z))$ into Eq. (12.125) gives

$$\begin{aligned} G_N(z) &= \frac{(1-\rho)(z-1)(18\lambda^2 + 17\lambda^2(1-z))}{9(\lambda - \lambda z + \lambda)(\lambda - \lambda z + 2\lambda)z - (18\lambda^2 + 17\lambda^2(1-z))} \\ &= \frac{(1-\rho)(z-1)(35-17z)}{9(2-z)(3-z)z - (35-17z)}, \end{aligned}$$

where we have canceled λ^2 from the numerator and denominator. If we factor the denominator we obtain

$$\begin{aligned} G_N(z) &= \frac{(1-\rho)(35-17z)(z-1)}{9(z-1)(z-7/3)(z-5/3)} \\ &= (1-\rho)\left\{\frac{1/3}{1-3z/7} + \frac{2/3}{1-3z/5}\right\}, \end{aligned}$$

where we have carried out a partial fraction expansion. Finally we note that since $G_N(z)$ converges for $|z| < 1$, we can expand $G_N(z)$ as follows:

$$G_N(z) = (1-\rho)\left\{\frac{1}{3}\sum_{k=0}^{\infty}\left(\frac{3}{7}\right)^k z^k + \frac{2}{3}\sum_{k=0}^{\infty}\left(\frac{3}{5}\right)^k z^k\right\}.$$

Since the coefficient of z^k is $P[N = k]$, we finally have that

$$P[N = k] = \frac{4}{27}\left(\frac{3}{7}\right)^k + \frac{8}{27}\left(\frac{3}{5}\right)^k \qquad k = 0, 1, \ldots,$$

where we used the fact that $\rho = 5/9$.

12.7.3 Delay and Waiting Time Distribution in an M/G/1 System

We now find the delay and waiting time distributions for an M/G/1 system with first-come, first-served service discipline. If a customer spends T_j seconds in the queueing system, then the number of customers N_d it leaves behind in the system is the number of customers that arrive during these T seconds, since customers are served in order of arrival. An expression for the generating function for N_d is found by proceeding as in Eq. (12.121a):

$$\begin{aligned} G_{N_d}(z) &= \sum_{k=0}^{\infty} \int_0^{\infty} P[N_d = k \,|\, T = t] f_T(t)\, dt\, z^k \\ &= \hat{T}(\lambda(1 - z)), \end{aligned} \tag{12.126}$$

where $\hat{T}(s)$ is the Laplace transform of the pdf of T, the total delay in the system. Since the steady state distributions of $N_d(t)$ and $N(t)$ are equal, we have that $G_N(z) = G_{N_d}(z)$ and thus combining Eqs. (12.125) and (12.126):

$$\hat{T}(\lambda(1 - z)) = \frac{(1 - \rho)(z - 1)\hat{\tau}(\lambda(1 - z))}{z - \hat{\tau}(\lambda(1 - z))}. \tag{12.127}$$

If we let $s = \lambda(1 - z)$, Eq. (12.127) yields an expression for $\hat{T}(s)$:

$$\hat{T}(s) = \frac{(1 - \rho)s\hat{\tau}(s)}{s - \lambda + \lambda\hat{\tau}(s)}. \tag{12.128}$$

The pdf of T is found from the inverse transform of $\hat{T}(s)$ either analytically or numerically.

Since $T = W + \tau$, where W and τ are independent random variables, we have that

$$\hat{T}(s) = \hat{W}(s)\hat{\tau}(s). \tag{12.129}$$

Equations (12.128) and (12.129) can then be solved for the Laplace transform of the waiting time pdf:

$$\hat{W}(s) = \frac{(1 - \rho)s}{s - \lambda + \lambda\hat{\tau}(s)}. \tag{12.130}$$

Equations (12.128) and (12.130) are also referred to as the **Pollaczek–Khinchin transform equations**.

Example 12.19 M/M/1

Find the pdf's of W and T for an M/M/1 system. Substituting $\hat{\tau}(s) = \mu/(s + \mu)$ into Eq. (12.128) gives

$$\hat{T}(s) = \frac{(1 - \rho)s\mu}{(s + \mu)(s - \lambda) + \lambda\mu} = \frac{(1 - \rho)\mu}{s - (\lambda - \mu)}, \tag{12.131}$$

which is readily inverted to obtain

$$f_T(x) = \mu(1 - \rho)e^{-\mu(1-\rho)x} \qquad x > 0. \tag{12.132}$$

Similarly, Eq. (9.130) gives

$$\hat{W}(s) = \frac{(1 - \rho)s}{s - \lambda + \lambda\mu/(s + \mu)} = (1 - \rho)\frac{s + \mu}{s + \mu - \lambda}.$$

In order to invert this expression, the numerator polynomial must have order lower than that of the denominator polynomial. We achieve this by dividing the denominator into the numerator:

$$\hat{W}(s) = (1 - \rho)\frac{s + \mu - \lambda + \lambda}{s + \mu - \lambda} = (1 - \rho)\left\{1 + \frac{\lambda}{s + \mu - \lambda}\right\}. \tag{12.133}$$

We then obtain

$$f_W(x) = (1 - \rho)\delta(x) + \lambda(1 - \rho)e^{-\mu(1-\rho)x} \qquad x > 0. \tag{12.134}$$

The delta function at zero corresponds to the fact that a customer has zero wait with probability $(1 - \rho)$. Equations (12.132) and (12.134) were previously obtained as Eqs. (12.32) and (12.33) in Section 12.3 by a different method.

Example 12.20 $M/H_2/1$

Find the pdf of the waiting time in the $M/H_2/1$ system discussed in Example 12.18.

Substitution of $\hat{\tau}(s)$ from Example 12.18 into Eq. (12.130) gives

$$\begin{aligned}
\hat{W}(s) &= \frac{9s(1 - \rho)(s + \lambda)(s + 2\lambda)}{9(s - \lambda)(s + \lambda)(s + 2\lambda) + \lambda(18\lambda^2 + 17\lambda s)} \\
&= \frac{(1 - \rho)(s + \lambda)(s + 2\lambda)}{s^2 + 2\lambda s + 8\lambda^2/9} \\
&= (1 - \rho)\frac{9s^2 + 27\lambda s + 18\lambda^2}{9s^2 + 18\lambda s + 8\lambda^2} \\
&= (1 - \rho)\left\{1 + \frac{9\lambda s + 10\lambda^2}{9s^2 + 18\lambda s + 8\lambda^2}\right\} \\
&= (1 - \rho)\left\{1 + \frac{2\lambda/3}{s + 2\lambda/3} + \frac{\lambda/3}{s + 4\lambda/3}\right\},
\end{aligned}$$

where we have followed the same sequence of steps as in Example 12.18 and then done a partial fraction expansion.

The inverse Laplace transform then yields

$$f_W(x) = \frac{4}{9}\left\{\delta(x) + \frac{2\lambda}{3}e^{-2\lambda x/3} + \frac{1}{4}\frac{4\lambda}{3}e^{-4\lambda x/3}\right\} \qquad x > 0.$$

Examples 12.18 and 12.19 demonstrate that the Pollaczek–Khinchin transform equations can be used to obtain closed-form expressions for the pmf of $N(t)$ and the pdf's of W and T when the Laplace transform of the service time pdf is a rational function of s, that is, a ratio of polynomials in s. This result is particularly important because it can be shown that the Laplace transform of any service time pdf can be approximated arbitrarily closely by a rational function of s. Thus in principle we can obtain exact expressions for the pmf of $N(t)$ and pdf's of W and T.

In addition it should be noted that *the Pollaczek–Khinchin transform expressions can always be inverted numerically* using fast Fourier transform methods such as those discussed in Section 7.6. This numerical approach does not require that the Laplace transform of the pdf be a rational function of s.

12.8 BURKE'S THEOREM: DEPARTURES FROM M/M/c SYSTEMS

In many problems, a customer requires service from several service stations before a task is completed. These problems require that we consider a *network* of queueing systems. In such networks, the departures from some queues become the arrivals to other queues. This is the reason why we are interested in the statistical properties of the departure process from a queue.

Consider two queues in tandem as shown in Fig. 12.21, where the departures from the first queue become the arrivals at the second queue. Assume that the arrivals to the first queue are Poisson with rate λ and that the service time at queue 1 is exponentially distributed with rate $\mu_1 > \lambda$. Assume that the service time in queue 2 is also exponentially distributed with rate $\mu_2 > \lambda$.

The state of this system is specified by the number of customers in the two queues, $(N_1(t), N_2(t))$. This state vector forms a Markov process with the transition rate diagram shown in Fig. 12.22, and global balance equations:

$$\lambda P[N_1 = 0, N_2 = 0] = \mu_2 P[N_1 = 0, N_2 = 1] \tag{12.135a}$$

$$\begin{aligned}(\lambda + \mu_1)P[N_1 = n, N_2 = 0] &= \mu_2 P[N_1 = n, N_2 = 1] \\ &\quad + \lambda P[N_1 = n - 1, N_2 = 0] \qquad n > 0\end{aligned} \tag{12.135b}$$

$$\begin{aligned}(\lambda + \mu_2)P[N_1 = 0, N_2 = m] &= \mu_2 P[N_1 = 0, N_2 = m + 1] \\ &\quad + \mu_1 P[N_1 = 1, N_2 = m - 1] \qquad m > 0\end{aligned} \tag{12.135c}$$

$$\begin{aligned}(\lambda + \mu_1 + \mu_2)P[N_1 = n, N_2 = m] &= \mu_2 P[N_1 = n, N_2 = m + 1] \\ &\quad + \mu_1 P[N_1 = n + 1, N_2 = m - 1] \\ &\quad + \lambda P[N_1 = n - 1, N_2 = m] \\ &\quad n > 0, m > 0.\end{aligned} \tag{12.135d}$$

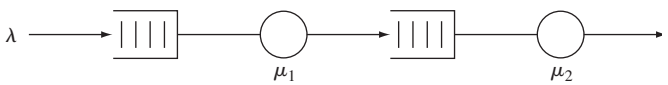

FIGURE 12.21
Two tandem exponential queues with Poisson input.

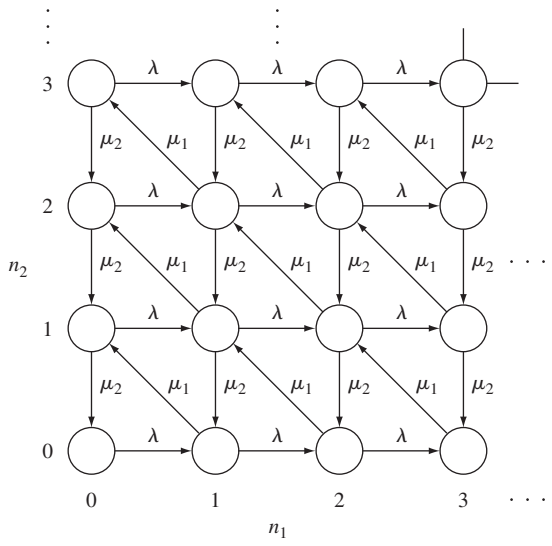

FIGURE 12.22
Transition rate diagram for two tandem exponential queues with Poisson input.

It is easy to verify that the following joint state pmf satisfies Eqs. (12.135a) through (12.135d):

$$P[N_1 = n, N_2 = m] = (1 - \rho_1)\rho_1^n(1 - \rho_2)\rho_2^m \qquad n \geq 0, m \geq 0, \quad (12.136)$$

where $\rho_i = \lambda/\mu_i$. We know that the first queue is an M/M/1 system, so

$$P[N_1 = n] = (1 - \rho_1)\rho_1^n \qquad n = 0, 1, \ldots. \quad (12.137)$$

By summing Eq. (12.136) over all n, we obtain the marginal state pmf of the second queue:

$$P[N_2 = m] = (1 - \rho_2)\rho_2^m \qquad m \geq 0. \quad (12.138)$$

Equations (12.136) through (12.138) imply that

$$P[N_1 = n, N_2 = m] = P[N_1 = n]P[N_2 = m] \qquad \text{for all } n, m. \quad (12.139)$$

In words, *the number of customers at queue 1 and the number at queue 2 at the same time instant are independent random variables.* Furthermore, *the steady state pmf at the second queue is that of an M/M/1 system with Poisson arrival rate λ and exponential service time* μ_2.

We say that a network of queues has a **product-form solution** when the joint pmf of the vector of numbers of customers at the various queues is equal to the product of the marginal pmf's of the number in the individual queues. We now discuss Burke's theorem, which states the fundamental result underlying the product-form solution in Eq. (12.139).

Burke's Theorem

Consider an M/M/1, M/M/c, or M/M/∞ queueing system at steady state with arrival rate λ, then

1. The departure process is Poisson with rate λ.
2. At each time t, the number of customers in the system $N(t)$ is independent of the sequence of departure times prior to t.

The product-form solution for the two tandem queues follows from Burke's theorem. Queue 1 is an M/M/1 queue, so from part 1 of the theorem the departures from queue 1 form a Poisson process. Thus the arrivals to queue 2 are a Poisson process, so the second queue is also an M/M/1 system with steady state pmf given by Eq. (12.138). It remains to show that the numbers of customers in the two queues at the same time instant are independent random variables.

The arrivals to queue 2 prior to time t are the departures from queue 1 prior to time t. By part 2 of Burke's theorem the departures from queue 1, and hence the arrivals to queue 2, prior to time t are independent of $N_1(t)$. Since $N_2(t)$ is determined by the sequence of arrivals from queue 1 prior to time t and the independent sequence of service times, it then follows that $N_1(t)$ and $N_2(t)$ are independent. Equation (12.139) then follows. Note that Burke's theorem does not state that $N_1(t)$ and $N_2(t)$ are independent random processes. This would require that $N_1(t_1)$ and $N_2(t_2)$ be independent random variables for all t_1 and t_2. This is clearly not the case.

Burke's theorem implies that the generalization of Eq. (12.139) holds for the tandem combination of any number of M/M/1, M/M/c, or M/M/∞ queues. Indeed, the result holds for any "feedforward" network of queues in which a customer cannot visit any queue more than once.

Example 12.21

Find the joint state pmf for the network of queues shown in Fig. 12.23, where queue 1 is driven by a Poisson process of rate λ_1, where the departures from queue 1 are randomly routed to queues 2 and 3, and where queue 3 also has an additional independent Poisson arrival stream of rate λ_2.

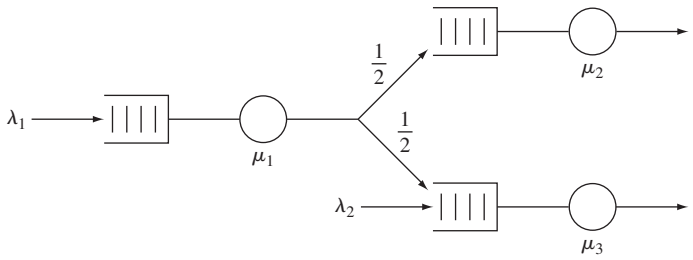

FIGURE 12.23
A feedforward network of queues.

From Burke's theorem $N_1(t)$ and $N_2(t)$ are independent, as are $N_1(t)$ and $N_3(t)$. Since the random split of a Poisson process yields independent Poisson processes, we have that the inputs to queues 2 and 3 are independent. The input to queue 2 is Poisson with rate $\lambda_1/2$. The input to queue 3 is Poisson of rate $\lambda_1/2 + \lambda_2$ since the merge of two independent Poisson processes is also Poisson. Thus

$$P[N_1(t) = k, N_2(t) = m, N_3(t) = n] = (1 - \rho_1)\rho_1^k(1 - \rho_2)\rho_2^m(1 - \rho_3)\rho_3^n \qquad k, m, n \geq 0,$$

where $\rho_1 = \lambda_1/\mu_1$, $\rho_2 = \lambda_1/2\mu_2$, and $\rho_3 = (\lambda_1/2 + \lambda_2)/\mu_3$, and where we have assumed that all of the queues are stable.

*12.8.1 Proof of Burke's Theorem Using Time Reversibility

Consider the sample path of an M/M/1, M/M/*c*, or M/M/∞ system as shown in Fig. 12.24(a). Note that the arrivals in the forward process correspond to the departures in the time-reversed process. In Section 11.5, we showed that birth-and-death Markov chains in steady state are time-reversible processes; that is, the sample functions of the process played backward in time have the same statistics as the forward process. Since M/M/1, M/M/*c*, and M/M/∞ systems are birth-and-death Markov chains, we

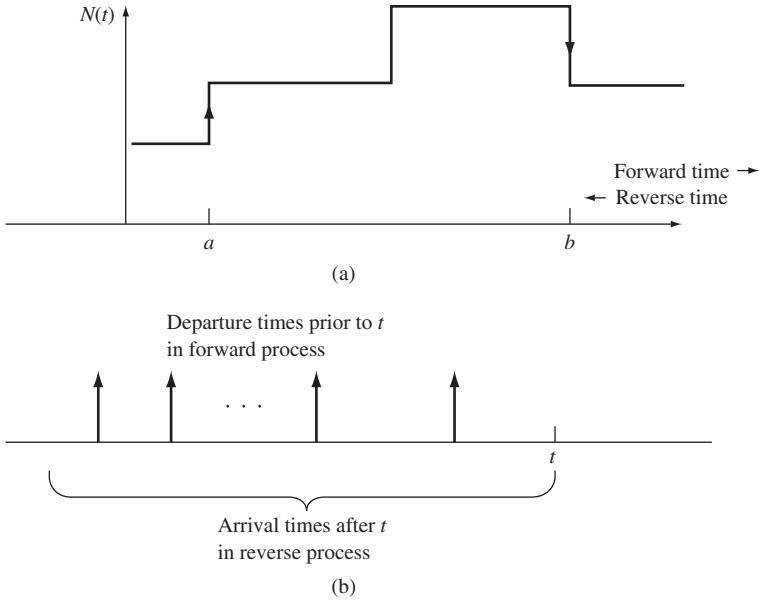

FIGURE 12.24
(a) Time instant *a* is an arrival time in the forward process and a departure time in the reverse process. Time instant *b* is a departure in the forward process and an arrival in the reverse process. (b) The departure times prior to time *t* in the forward process correspond exactly to the arrival times after time *t* in the reverse process.

have that their states are reversible processes. Thus the sample functions of these systems played backward in time correspond to the sample functions of queueing systems of the same type. It then follows that the arrival process of the time-reversed system is a Poisson process.

To prove part 1 of Burke's theorem, we note that the interdeparture times of the forward-time system are the interarrival times of the time-reversed system. Since the arrival process of the time-reversed system is Poisson, it then follows that the departure process of the forward system is also Poisson. Thus we have shown that the departure process of an M/M/1, M/M/c, or M/M/∞ system is Poisson.

To prove part 2 of Burke's theorem, fix a time t as shown in Fig. 12.24(b). The departures before time t from the forward system are the arrivals after time t in the reverse system. In the reverse system, the arrivals are Poisson and thus the arrival times after time t do not depend on $N(t)$. These arrival instants of the reverse process are exactly the departure instants before t in the forward process. It then follows that $N(t)$ and the departure instants prior to t are independent, so part 2 is proved.

12.9 NETWORKS OF QUEUES: JACKSON'S THEOREM

In many queueing networks, a customer is allowed to visit a particular queue more than once. Burke's theorem does not hold for such systems. In this section we discuss Jackson's theorem, which extends the product-form solution for the steady state pmf to a broader class of queueing networks.

If a customer is allowed to visit a queue more than once, then the arrival process at that queue will not be Poisson. For example, consider the simple M/M/1 queue with feedback shown in Fig. 12.25, where external customers arrive according to a Poisson process of rate λ and where departures are instantaneously fed back into the system with probability .9. If the arrival rate is much less than the departure rate, then we have that the net arrival process (i.e., external and feedback arrivals) typically consists of isolated external arrivals followed by a burst of feedback arrivals. Thus the arrival process does not have independent increments and so it is not Poisson.

12.9.1 Open Networks of Queues

Consider a network of K queues in which customers arrive from outside the network to queue k according to independent Poisson processes of rate α_k. We assume that the service time of a customer in queue k is exponentially distributed with rate μ_k and independent of all other service times and arrival processes. We also suppose that queue

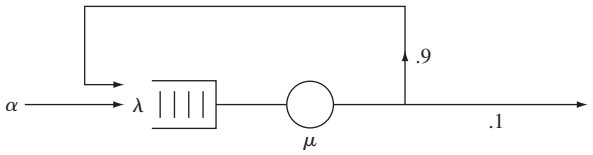

FIGURE 12.25
A queue with feedback.

k has c_k servers. After completion of service in queue k, a customer proceeds to queue i with probability P_{ki} and exits the network with probability

$$1 - \sum_{i=1}^{K} P_{ki}.$$

The total arrival rate λ_k into queue k is the sum of the external arrival rate and the internal arrival rates:

$$\lambda_k = \alpha_k + \sum_{j=1}^{K} \lambda_j P_{jk} \qquad k = 1, \ldots, K. \tag{12.140}$$

It can be shown that Eq. (12.140) has a unique solution if no customer remains in the network indefinitely. We call such networks **open queueing networks**.

The vector of the number of customers in all the queues,

$$\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_K(t)),$$

is a Markov process. Jackson's theorem gives the steady state pmf for $\mathbf{N}(t)$.

Jackson's Theorem

If $\lambda_k < c_k\mu_k$, then for any possible state $\mathbf{n} = (n_1, n_2, \ldots, n_K)$,

$$P[\mathbf{N}(t) = \mathbf{n}] = P[N_1 = n_1]P[N_2 = n_2] \ldots P[N_K = n_K], \tag{12.141}$$

where $P[N_k = n_k]$ is the steady state pmf of an M/M/c_k system with arrival rate λ_k and service rate μ_k.

Jackson's theorem states that the numbers of customers in the queues at time t are *independent* random variables. In addition, it states that the steady state probabilities of the individual queues are those of an M/M/c_k system. This is an amazing result because in general the input process to a queue is not Poisson, as was demonstrated in the simple queue with feedback discussed in the beginning of this section.

Example 12.22

Messages arrive at a concentrator according to a Poisson process of rate α. The time required to transmit a message and receive an acknowledgment is exponentially distributed with mean $1/\mu$. Suppose that a message needs to be retransmitted with probability p. Find the steady state pmf for the number of messages in the concentrator.

The overall system can be represented by the simple queue with feedback shown in Fig. 12.25. The net arrival rate into the queue is $\lambda = \alpha + \lambda p$, that is,

$$\lambda = \frac{\alpha}{1 - p}.$$

Thus, the pmf for the number of messages in the concentrator is

$$P[N = n] = (1 - \rho)\rho^n \qquad n = 0, 1, \ldots,$$

where $\rho = \lambda/\mu = \alpha/(1 - p)\mu$.

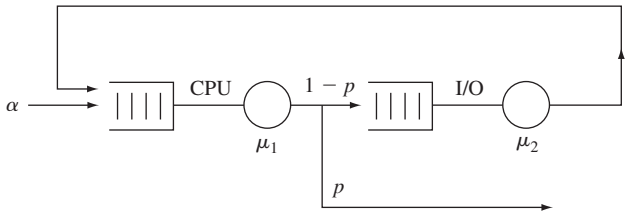

FIGURE 12.26
An open queueing network model for a computer system.

Example 12.23

New programs arrive at a CPU according to a Poisson process of rate α as shown in Fig. 12.26. A program spends an exponentially distributed execution time of mean $1/\mu_1$ in the CPU. At the end of this service time, the program execution is complete with probability p or it requires retrieving additional information from secondary storage with probability $1 - p$. Suppose that the retrieval of information from secondary storage requires an exponentially distributed amount of time with mean $1/\mu_2$. Find the mean time that each program spends in the system.

The net arrival rates into the two queues are

$$\lambda_1 = \alpha + \lambda_2 \qquad \text{and} \qquad \lambda_2 = (1 - p)\lambda_1.$$

Thus

$$\lambda_1 = \frac{\alpha}{p} \qquad \text{and} \qquad \lambda_2 = \frac{(1 - p)\alpha}{p}.$$

Each queue behaves like an M/M/1 system, so

$$E[N_1] = \frac{\rho_1}{1 - \rho_1} \qquad \text{and} \qquad E[N_2] = \frac{\rho_2}{1 - \rho_2},$$

where $\rho_1 = \lambda_1/\mu_1$ and $\rho_2 = \lambda_2/\mu_2$. Little's formula then gives the mean for the total time spent in the system:

$$E[T] = \frac{E[N_1 + N_2]}{\alpha} = \frac{1}{\alpha}\left[\frac{\rho_1}{1 - \rho_1} + \frac{\rho_2}{1 - \rho_2}\right].$$

*12.9.2 Proof of Jackson's Theorem

Jackson's theorem can be proved by writing the global balance equations for the queueing network and verifying that the solution is given by Eq. (12.141). We present an alternative proof of the theorem using a result from time-reversed Markov chains. For notational simplicity we consider only the case of a network of single-server queues.

Let $\mathbf{n}$ and $\mathbf{n}'$ be two possible states of the network, and let $v_{\mathbf{n},\mathbf{n}'}$ denote the transition rate from $\mathbf{n}$ to $\mathbf{n}'$. In Section 11.5, we found that if we can guess a state pmf $P[\mathbf{n}]$ and a set of transition rates $\hat{v}_{\mathbf{n}',\mathbf{n}}$ for the reverse process such that (Eq. 11.65)

$$P[\mathbf{n}]v_{\mathbf{n},\mathbf{n}'} = P[\mathbf{n}']\hat{v}_{\mathrm{n}',\mathrm{n}} \tag{12.142a}$$

and such that the total rate out of state $\mathbf{n}$ is the same in the forward and reverse processes (Eq. 11.64 summed over j)

$$\sum_{\mathbf{m}} v_{\mathbf{n},\mathbf{m}} = \sum_{\mathbf{m}} \hat{v}_{\mathbf{n},\mathbf{m}}, \tag{12.142b}$$

then $P[\mathbf{n}]$ is the steady state pmf of the process.

For the case under consideration our guess for the pmf is

$$P[\mathbf{n}] = \prod_{j=1}^{K} (1 - \rho_j)\rho_j^{n_j}, \tag{12.143}$$

so the proof reduces to finding a consistent set of transition rates for the reverse process that satisfy Eqs. (12.142a) and (12.142b). Noting that $v_{\mathbf{n},\mathbf{n}'}$ is known and that $P[\mathbf{n}]$ and $P[\mathbf{n}']$ are specified by Eq. (12.143), Eq. (12.142a) can be solved for the transition rates of the reverse process:

$$\hat{v}_{\mathbf{n}',\mathbf{n}} = \frac{P[\mathbf{n}]v_{\mathbf{n},\mathbf{n}'}}{P[\mathbf{n}']}. \tag{12.144}$$

Let $\mathbf{n} = (n_1, \ldots, n_k)$ denote a state for the network, and let $\mathbf{e}_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is located in the kth component. Only three types of transitions in the state of the queueing network have nonzero probabilities. In the first type of transition, an external arrival to queue k takes the state from $\mathbf{n}$ to $\mathbf{n} + \mathbf{e}_k$. In the second type of transition, a departure from queue k exits the network and takes the state from $\mathbf{n}$ to $\mathbf{n} - \mathbf{e}_k$, where $n_k > 0$. In the third type of transition, a customer leaves queue k and proceeds to queue j, thus taking the state from $\mathbf{n}$ to $\mathbf{n} - \mathbf{e}_k + \mathbf{e}_j$, where $n_k > 0$. Table 12.1 shows three types of transitions and their corresponding rates for the forward process.

A consistent set of transition rates for the reverse process is obtained by solving Eq. (12.144) for the three types of transitions possible. For example, if we let $\mathbf{n}' = \mathbf{n} + \mathbf{e}_k$, then the transition $\mathbf{n} \rightarrow \mathbf{n} + \mathbf{e}_k$ in the forward process corresponds to the transition $\mathbf{n} + \mathbf{e}_k \rightarrow \mathbf{n}$ in the reverse process. Equation (12.144) gives

$$\begin{aligned}\hat{v}_{\mathbf{n}',\mathbf{n}} &= \frac{\alpha_k \prod_{j=1}^{K} (1 - \rho_j)\rho_j^{n_j}}{\rho_k \prod_{j=1}^{K} (1 - \rho_j)\rho_j^{n_j}} \\ &= \frac{\alpha_k}{\rho_k} = \frac{\alpha_k}{\lambda_k/\mu_k} = \frac{\alpha_k \mu_k}{\lambda_k}.\end{aligned}$$

The other reverse process transition rates are found in similar manner. Table 12.1 shows the results for the transition rates of the reverse process that are implied by Eq. (12.144).

The proof that the pmf in Eq. (12.143) gives the steady state pmf of the network of queues is completed by showing that the total transition rate out of any state $\mathbf{n}$ is the same in the forward and in the reverse process, that is, Eq. (12.142b) holds. In the

TABLE 12.1 Allowable transitions in Jackson network and their corresponding rates in the forward and reverse processes

Forward Process		
Transition	*Rate*	
$\mathbf{n} \to \mathbf{n} + \mathbf{e}_k$	α_k	all k
$\mathbf{n} \to \mathbf{n} - \mathbf{e}_k$	$\mu_k\left(1 - \sum_{j=1}^{K} P_{kj}\right)$	all k: $n_k > 0$
$\mathbf{n} \to \mathbf{n} - \mathbf{e}_k + \mathbf{e}_j$	$\mu_k P_{kj}$	all k: $n_k > 0$, all j
Reverse Process		
Transition	*Rate*	
$\mathbf{n} \to \mathbf{n} + \mathbf{e}_k$	$\lambda_k\left(1 - \sum_{j} P_{kj}\right)$	all k
$\mathbf{n} \to \mathbf{n} - \mathbf{e}_k$	$\dfrac{\alpha_k \mu_k}{\lambda_k}$	all k: $n_k > 0$
$\mathbf{n} \to \mathbf{n} - \mathbf{e}_k + \mathbf{e}_j$	$\dfrac{\lambda_j P_{jk} \mu_k}{\lambda_k}$	all k: $n_k > 0$, all j

forward process, the total transition rate out of state $\mathbf{n}$ is obtained by adding the entries for the forward process in Table 12.1:

$$\sum_{\mathbf{m}} v_{\mathbf{n},\mathbf{m}} = \sum_{k} \alpha_k + \sum_{k:\, n_k > 0} \mu_k. \tag{12.145a}$$

For the reverse process, we have from Table 12.1 that

$$\sum_{\mathbf{m}} \hat{v}_{\mathbf{n},\mathbf{m}} = \sum_{k} \lambda_k \left(1 - \sum_{j} P_{kj}\right) + \sum_{k:\, n_k > 0} \left\{ \frac{\alpha_k \mu_k}{\lambda_k} + \sum_{j} \frac{\lambda_j P_{jk} \mu_k}{\lambda_k} \right\}. \tag{12.145b}$$

We need to show that the right-hand sides of Eqs. (12.145a) and (12.145b) are equal. First, note that Eq. (12.140) implies that

$$\lambda_k - \alpha_k = \sum_{j=1}^{K} \lambda_j P_{jk}.$$

The right-hand side of Eq. (12.145b) then becomes

$$\begin{aligned}
&\left(\sum_{k} \lambda_k - \sum_{j} \sum_{k} \lambda_k P_{kj} \right) + \sum_{k:\, n_k > 0} \left\{ \frac{\alpha_k \mu_k}{\lambda_k} + \frac{\mu_k}{\lambda_k} \sum_{j} \lambda_j P_{jk} \right\} \\
&\quad = \sum_{k} \lambda_k - \sum_{j} (\lambda_j - \alpha_j) + \sum_{k:\, n_k > 0} \left\{ \frac{\alpha_k \mu_k}{\lambda_k} + \frac{\mu_k}{\lambda_k} (\lambda_k - \alpha_k) \right\} \\
&\quad = \sum_{k} \alpha_k + \sum_{k:\, n_k > 0} \mu_k.
\end{aligned}$$

Thus the right-hand sides of Eqs. (12.145a) and (12.145b) are equal and thus Eq. (12.143) is the steady state pmf of the network of queues. This completes the proof of Jackson's theorem for a network of single-server queues.

12.9.3 Closed Networks of Queues

In some problems, a *fixed* number of customers, say I, circulate endlessly in a **closed network of queues**. For example, some computer system models assume that at any time a fixed number of processes use the CPU and input/output (I/O) resources of a computer as shown in Fig. 12.27. We now consider queueing networks that are identical to the previously discussed **open networks** except that the external arrival rates are zero and the networks always contain a fixed number of customers I. We show that the steady state pmf for such systems is product form but that the states of the queues are no longer independent.

The net arrival rate into queue k is now given by

$$\lambda_k = \sum_{j=1}^{K} \lambda_j P_{jk} \qquad k = 1, \ldots, K. \tag{12.146}$$

Note that these equations have the same form as the set of equations that define the stationary pmf for a discrete-time Markov chain with transition probabilities P_{jk}. The only difference is that the sum of the λ_k's need not be one. Thus the solution vector to Eq. (12.146) must be proportional to the stationary pmf $\{\pi_j\}$ corresponding to $\{P_{jk}\}$:

$$\lambda_k = \lambda(I)\pi_k, \tag{12.147}$$

where

$$\pi_k = \sum_{j=1}^{K} \pi_j P_{jk} \tag{12.148}$$

and where $\lambda(I)$ is a constant that depends on I, the number of customers in the queueing network. If we sum both sides of Eq. (12.147) over k, we see that $\lambda(I)$ is the sum of the arrival rates in all the queues in the network, and $\pi_k = \lambda_k/\lambda(I)$ is the fraction of total arrivals to queue k.

Theorem

Let $\lambda_k = \lambda(I)\pi_k$ be a solution to Eq. (12.146), and let $\mathbf{n} = (n_1, n_2, \ldots, n_K)$ be any state of the network for which $n_1, \ldots, n_K \geq 0$ and

$$n_1 + n_2 + \cdots + n_K = I, \tag{12.149}$$

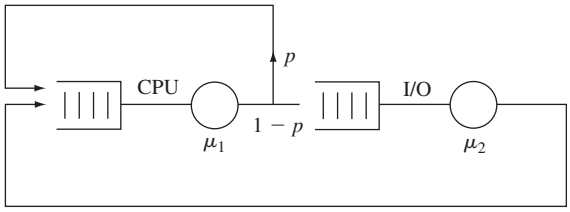

FIGURE 12.27
A closed queueing network model for a computer system.

then

$$P[\mathbf{N}(t) = \mathbf{n}] = \frac{P[N_1 = n_1]P[N_2 = n_2]\dots P[N_K = n_K]}{S(I)}, \tag{12.150}$$

where $P[N_k = n_k]$ is the steady state pmf of an M/M/c_k system with arrival rate λ_k and service rate μ_k, and where $S(I)$ is the normalization constant given by

$$S(I) = \sum_{\mathbf{n}:n_1+\cdots+n_K=I} P[N_1 = n_1]P[N_2 = n_2]\dots P[N_K = n_K]. \tag{12.151}$$

Equation (12.150) states that $P[\mathbf{N}(t) = \mathbf{n}]$ has a product form. However, $P[\mathbf{N}(t) = \mathbf{n}]$ is no longer equal to the product of the marginal pmf's because of the normalization constant $S(I)$. This constant arises because the fact that there are always I customers in the network implies that the allowable states $\mathbf{n}$ must satisfy Eq. (12.149). The theorem can be proved by taking the approach used to prove Jackson's theorem above.

Example 12.24

Suppose that the computer system in Example 12.23 is operated so that there are always I programs in the system. The resulting network of queues is shown in Fig. 12.27. Note that the feedback loop around the CPU signifies the completion of one job and its instantaneous replacement by another one. Find the steady state pmf of the system. Find the rate at which programs are completed.

The stationary probabilities associated with Eq. (9.146) are found by solving

$$\pi_1 = p\pi_1 + \pi_2, \qquad \pi_2 = (1 - p)\pi_1, \qquad \text{and} \qquad \pi_1 + \pi_2 = 1.$$

The stationary probabilities are then

$$\pi_1 = \frac{1}{2 - p} \qquad \text{and} \qquad \pi_2 = \frac{1 - p}{2 - p} \tag{12.152}$$

and the arrival rates are

$$\lambda_1 = \lambda(I)\pi_1 = \frac{\lambda(I)}{2 - p} \qquad \text{and} \qquad \lambda_2 = \frac{(1 - p)\lambda(I)}{2 - p}. \tag{12.153}$$

The stationary pmf for the network is then

$$P[N_1 = i, N_2 = I - i] = \frac{(1 - \rho_1)\rho_1^i(1 - \rho_2)\rho_2^{I-i}}{S(I)} \qquad 0 \le i \le I, \tag{12.154}$$

where $\rho_1 = \lambda_1/\mu_1$ and $\rho_2 = \lambda_2/\mu_2$, and where we have used the fact that if $N_1 = i$ then $N_2 = I - i$. The normalization constant is then

$$\begin{aligned} S(I) &= (1 - \rho_1)(1 - \rho_2)\sum_{i=0}^{I}\rho_1^i\rho_2^{I-i} \\ &= (1 - \rho_1)(1 - \rho_2)\rho_2^I\frac{1 - (\rho_1/\rho_2)^{I+1}}{1 - (\rho_1/\rho_2)}. \end{aligned} \tag{12.155}$$

Substitution of Eq. (12.155) into Eq. (12.154) gives

$$P[N_1 = i, N_2 = I - i] = \frac{1 - \beta}{1 - \beta^{I+1}}\beta^i \qquad 0 \le i \le I, \tag{12.156}$$

where

$$\beta = \frac{\rho_1}{\rho_2} = \frac{\pi_1\mu_2}{\pi_2\mu_1} = \frac{\mu_2}{(1 - p)\mu_1}. \tag{12.157}$$

Note that the form of Eq. (12.156) suggests that queue 1 behaves like an M/M/1/K queue. The apparent load to this queue is β, which is proportional to the ratio of I/O to CPU service rates and inversely proportional to the probability of having to go to I/O.

The rate at which programs are completed is $p\lambda_1$. We find λ_1 from the relation between server utilization and probability of an empty system:

$$1 - \lambda_1/\mu_1 = P[N_1 = 0] = \frac{1 - \beta}{1 - \beta^{I+1}},$$

which implies that

$$p\lambda_1 = p\mu_1\frac{\beta(1 - \beta^I)}{1 - \beta^{I+1}}.$$

Example 12.25

A transmitter (queue 1 in Fig. 12.28) has two permits for message transmission. As long as the transmitter has a permit ($N_1 > 0$), it generates messages with exponential interarrival times of rate λ. The messages enter the transmission system and require an exponential service time at station 2. As soon as a message arrives at the other side of the transmission system, the corresponding permit is sent back via station 3. Thus the transmitter can have at most two messages outstanding in the network at any given time. Find the steady state pmf for the network of queues. Find the rate at which messages enter the transmission system.

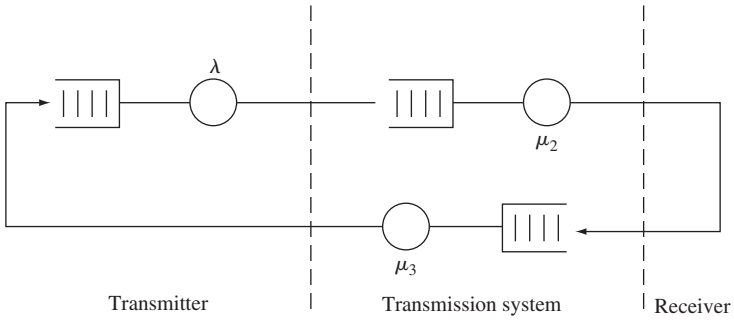

FIGURE 12.28
A closed queueing network model for a message transmission system.

We can view the two permits as two customers circulating the queueing network. Since $P_{1,2} = P_{2,3} = P_{3,1} = 1$, we have that $\pi_1 = \pi_2 = \pi_3 = 1/3$ and thus

$$\lambda_1 = \lambda_2 = \lambda_3 = \frac{\lambda(2)}{3}.$$

The steady state pmf for the network is

$$P[N_1 = i, N_2 = j, N_3 = 2 - i - j] = \frac{(1 - \rho_1)\rho_1^i(1 - \rho_2)\rho_2^j(1 - \rho_3)\rho_3^{I-i-j}}{S(2)}$$

$$\text{for } 0 \le i \le 2, 0 \le j \le 2 - i,$$

where $\rho_1 = \lambda(2)/3\lambda$ and $\rho_2 = \rho_3 = \lambda(2)/3\mu$. The normalization constant $S(2)$ is obtained by summing the above joint pmf over all possible states and equating the result to one. There are six possible network states: $(2, 0, 0)$, $(0, 2, 0)$, $(0, 0, 2)$, $(1, 1, 0)$, $(1, 0, 1)$, $(0, 1, 1)$. Thus the normalization constant is given by

$$\begin{aligned} S(2) &= (1 - \rho_1)(1 - \rho_2)(1 - \rho_3)\{\rho_1^2 + \rho_2^2 + \rho_3^2 + \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3\} \\ &= (1 - \rho_1)(1 - \rho_2)^2\{\rho_1^2 + 2\rho_2^2 + 2\rho_1\rho_2 + \rho_2^2\}, \end{aligned}$$

where we have used the fact that $\rho_2 = \rho_3$.

The rate at which messages enter the system is

$$\lambda_1 = \lambda(1 - P[N_1 = 0]),$$

where

$$\begin{aligned} P[N_1 = 0] &= P[\mathbf{N} = (0, 2, 0)] + P[\mathbf{N} = (0, 0, 2)] + P[\mathbf{N} = (0, 1, 1)] \\ &= \frac{3\rho_2^2}{\rho_1^2 + 2\rho_1\rho_2 + 3\rho_2^2} = \frac{3/\mu^2}{1/\lambda^2 + 2/\lambda\mu + 3/\mu^2}. \end{aligned}$$

12.9.4 Mean Value Analysis

Example 12.25 shows that the evaluation of the normalization constant is the fundamental difficulty with closed queueing networks. Fortunately, a method has been developed for obtaining certain average quantities of interest without having to evaluate this constant. This **mean value analysis** method is based on the following theorem.

Arrival Theorem

In a closed queueing network with I customers, the system as seen by a customer arrival to queue j is the steady state pmf of the same network with one fewer customer.

We have already encountered this result in the discussion of finite-source queueing systems in Section 12.5. We prove the result in the last part of this section. We now use the result to develop the mean value analysis method.

Let $E[N_j(I)]$ be the mean number of customers in the jth queue for a network that has I customers, let $E[T_j(I)]$ denote the mean time spent by a customer in queue j, and let $\lambda_j(I)$ denote the average customer arrival rate at queue j. The mean time spent by a customer in queue j is his service time plus the service times of the customers he finds in the queue upon arrival:

$$\begin{aligned} E[T_j(I)] &= E[\tau_j] + E[\tau_j] \times \text{mean number found upon arrival} \\ &= E[\tau_j] + E[\tau_j]E[N_j(I-1)] \\ &= \frac{1 + E[N_j(I-1)]}{\mu_j}, \end{aligned} \tag{12.158}$$

where $E[N_j(I-1)]$ is the mean number found upon arrival by the arrival theorem. By Little's formula, the mean number of customers in queue j when there are I in the network is

$$E[N_j(I)] = \lambda_j(I)E[T_j(I)] = \lambda(I)\pi_j E[T_j(I)]. \tag{12.159}$$

Since the sum of the customers in all queues is I in the previous equation, we have that

$$I = \sum_{j=1}^{K} E[N_j(I)] = \lambda(I)\sum_{j=1}^{K} \pi_j E[T_j(I)]. \tag{12.160}$$

Thus

$$\lambda(I) = \frac{I}{\sum_{j=1}^{K} \pi_j E[T_j(I)]}. \tag{12.161}$$

The *mean value analysis* method combines Eqs. (12.158) through (12.161) in the following way. First compute π_j by solving Eq. (12.148), then for $I = 0$:

$$E[N_j(0)] = 0 \qquad \text{for } j = 1, \ldots, K.$$

For $I = 1, 2, \ldots$:

$$E[T_j(I)] = \frac{1}{\mu_j} + \frac{E[N_j(I-1)]}{\mu_j} \qquad j = 1, \ldots, K \tag{12.158}$$

$$\lambda(I) = \frac{I}{\sum_{i=0}^{K} \pi_i E[T_j(I)]} \tag{12.161}$$

$$E[N_j(I)] = \lambda(I)\pi_j E[T_j(I)] \qquad j = 1, \ldots, K. \tag{12.159}$$

Thus the mean value algorithm begins with an empty system and by use of the above three equations builds up to a network with the desired number of customers. This method has considerably simplified the numerical solution of closed queueing networks and extended the range of network sizes that can be analyzed.

Example 12.26

In Example 12.24, let $\mu_1 = \mu_2 = 1$, and $p = 1/2$. Find the rate at which programs are completed if $I = 2$.

It was already indicated in Example 12.24 that the rate of program completion is $p\lambda_1(2) = p\pi_1\lambda(2)$. From Eq. (12.152), we have that $\pi_1 = 1/(2 - p) = 2/3$. Thus we only need to find $\lambda(2)$, the total arrival rate of the network with $I = 2$.

Starting the mean value method with $I = 1$, we have

$$E[T_1(1)] = \frac{1}{\mu_1} = 1 \qquad E[T_2(1)] = \frac{1}{\mu_2} = 1$$

$$\lambda(1) = \frac{1}{\pi_1 T_1(1) + \pi_2 T_2(1)} = 1$$

$$E[N_1(1)] = \lambda(1)\pi_1 E[T_1(1)] = \frac{2}{3}$$

$$E[N_2(1)] = \lambda(1)\pi_2 E[T_2(1)] = \frac{1}{3}.$$

Continuing with $I = 2$, we have

$$E[T_1(2)] = \frac{1}{\mu_1} + \frac{E[N_1(1)]}{\mu_1} = \frac{5}{3}$$

$$E[T_2(2)] = \frac{1}{\mu_2} + \frac{E[N_2(1)]}{\mu_2} = \frac{4}{3}$$

$$\lambda(2) = \frac{2}{\pi_1 E[T_1(2)] + \pi_2 E[T_2(2)]} = \frac{9}{7}.$$

Thus the program completion rate is

$$p\pi_1\lambda(2) = \frac{3}{7}.$$

You should verify that this is consistent with the results of Example 12.24.

Example 12.27

In Example 12.25, let $1/\lambda = a$ and $\mu = 1$. Find the rate at which messages enter the system when $I = 2$.

We previously found that $\pi_1 = \pi_2 = \pi_3 = 1/3$ and

$$\lambda_1(I) = \lambda_2(I) = \lambda_3(I) = \frac{\lambda(I)}{3}.$$

Starting the mean value method with $I = 1$, we have

$$E[T_1(1)] = a \qquad E[T_2(1)] = E[T_3(1)] = 1$$

$$\lambda(1) = \frac{1}{\pi_1 E[T_1(1)] + \pi_2 E[T_2(1)] + \pi_3 E[T_3(1)]} = \frac{3}{a + 2}$$

$$E[N_1(1)] = \lambda(1)\pi_1 E[T_1(1)] = \frac{a}{a+2}$$
$$E[N_2(1)] = \lambda(1)\pi_2 E[T_2(1)] = \frac{1}{a+2} = E[N_3(1)].$$

Continuing with $I = 2$, we have

$$E[T_1(2)] = a + aE[N_1(1)] = \frac{2a^2 + 2a}{a+2}$$
$$E[T_2(2)] = 1 + 1E[N_2(1)] = \frac{a+3}{a+2} = E[T_3(2)]$$
$$\lambda_2(2) = \frac{2}{(1/3)\{(2a^2+2a)/(a+2) + [2(a+3)/(a+2)]\}}$$
$$= \frac{3(a+2)}{a^2+2a+3}.$$

Finally, messages enter the transmission network at a rate $\lambda_1(2) = \lambda(2)/3$, so

$$\lambda_1(2) = \frac{a+2}{a^2+2a+3}.$$

You should verify that this is consistent with the results obtained in Example 12.25.

*12.9.5 Proof of the Arrival Theorem

Consider the instant when a customer leaves queue j and is proceeding to queue k. We are interested in the pmf of the system state at these arrival instants. Suppose that at this instant, with the customer removed from the system, the customer sees the network in state $\mathbf{n} = (n_1, \ldots, n_K)$. This occurs only when the network state goes from the state $\mathbf{n}' = (n_1, \ldots, n_j + 1, \ldots, n_K)$ to the state $\mathbf{n}'' = (n_1, \ldots, n_j, \ldots, n_k + 1, \ldots, n_K)$. Thus:

$$\begin{aligned}
&P[\text{customer sees } \mathbf{n} \mid \text{customer goes from } j \text{ to } k] \\
&= \frac{P[\text{customer sees } \mathbf{n}, \text{ customer goes from } j \text{ to } k]}{P[\text{customer goes from } j \text{ to } k]} \\
&= \frac{P[\text{customer goes from } j \text{ to } k \mid \text{state is } \mathbf{n}']P[\mathbf{N}(I) = \mathbf{n}']}{P[\text{customer goes from } j \text{ to } k]} \\
&= \frac{\mu_j P_{jk} P[\mathbf{N}(I) = \mathbf{n}']}{\mu_j P_{jk} P[N_j(I) > 0]} \\
&= \frac{P[\mathbf{N}(I) = \mathbf{n}']}{P[N_j(I) > 0]}.
\end{aligned} \tag{12.162}$$

To simplify the notation, let us assume that we are dealing with a network of M/M/1 queues, then

$$P[\mathbf{N}(I) = \mathbf{n}'] = \frac{P[N_1 = n_1] \dots P[N_j = n_j + 1] \dots P[N_K = n_K]}{S(I)}$$
$$= \rho_j \prod_{m=1}^{K} \frac{\rho_m^{n_m}}{S'(I)}, \tag{12.163}$$

where $S'(I)$ absorbs all the constants associated with the $P[N_m = n_m]$:

$$S'(I) = \sum_{\mathbf{n}: n_1 + \cdots + n_K = I} \prod_{m=1}^{K} \rho_m^{n_m}. \tag{12.164}$$

Next, consider the probability that queue j is not empty:

$$P[N_j(I) > 0] = \sum_{\mathbf{n}: n_1 + \cdots + n_K = I-1} P[N_1 = n_1] \dots P[N_j = n_j + 1] \dots P[N_K = n_k]$$
$$= \sum_{\mathbf{n}: n_1 + \cdots + n_K = I-1} \rho_j \frac{\prod_{m=1}^{K} \rho_m^{n_m}}{S'(I)}$$
$$= \frac{\rho_j}{S'(I)} \sum_{\mathbf{n}: n_1 + \cdots + n_K = I-1} \prod_{m=1}^{K} \rho_m^{n_m}$$
$$= \frac{\rho_j S'(I-1)}{S'(I)}, \tag{12.165}$$

where we have noted that the above summation is the normalization constant for a network with $I - 1$ customers $S'(I - 1)$.

Finally, we substitute Eqs. (12.165) and (12.163) into Eq. (12.162):

$$P[\text{customer sees } \mathbf{n} \mid \text{customer goes from } j \text{ to } k]$$
$$= \frac{\rho_j \prod_{m=1}^{K} \rho_m^{n_m} / S'(I)}{[\rho_j S'(I-1)]/S'(I)}$$
$$= \prod_{m=1}^{K} \frac{\rho_m^{n_m}}{S'(I-1)}$$
$$= P[\mathbf{N}(I-1) = \mathbf{n}],$$

which is the steady state probability for $\mathbf{n}$ in a network with $I - 1$ customers. This completes the proof of the arrival theorem.

12.10 SIMULATION AND DATA ANALYSIS OF QUEUEING SYSTEMS

In this section we present a basic introduction to the simulation of queueing systems. Analytical methods are valuable due to the ease with which they allow us to explore the issues and tradeoffs in a given model. Numerical techniques can supplement analytical methods and provide additional detailed information, especially when transient and dynamic behavior is of interest. However, in many situations analytical and numerical methods are not sufficient and simulation provides us with a flexible means to investigate the behavior of complex systems. In this section we introduce the basic approaches available for simulating queueing systems. Throughout our discussion we emphasize the need for *careful design of the simulation experiment* as well as the need for *careful application of statistical methods* on the observations to draw valid conclusions.

12.10.1 Approaches to Simulation

The dynamics of a queueing system are represented by one or more random processes, so the usual considerations in simulating random processes apply. A very basic option is whether a *single realization* or *multiple realizations* of the random process are used.

Multiple realizations that are statistically independent allow us to use the standard statistical methods introduced in Chapter 8 to analyze iid random variables, for example, to obtain confidence intervals and fit distributions. A single realization of a random process allows us a more restricted set of statistical tools and frequently leads to methods that attempt to provide a set of observations that are iid so that we can use standard tools. In some real experimental situations we may only have one realization of the process to work with and so we may have no choice. However in computer simulation with proper design, we can usually conduct multiple replications of an experiment to produce independent observations.[4] In general, we recommend a pragmatic approach that uses some replication when possible.

A simulation study based on a single realization usually involves *assumptions about stationarity and ergodicity* so that the behavior of the process over time reveals its ensemble averages and probabilities. Examples of such processes are processes with stationary independent increments and processes that involve ergodic Markov chains. Both of these classes of processes involve initial transient behavior and so we must decide whether to keep or discard the observations obtained during the initial portion of the simulation. If we decide to discard, then we need to somehow *identify when the transient phase is over* and the process has reached steady state. This is not an easy task, as discussed extensively in [Pawlikowski], and there are a variety of criteria that can be applied for declaring that a system has reached steady state. We note that the use of replicated simulations can help characterize the transient phase of a process. (See Problem 12.67.)

[4]Care should be taken to ensure that the seed in the random number generator is different in each replication.

The design of a simulation must take into account the behavior and parameters that we are interested in measuring and observing. Seemingly easy questions such as determining state probabilities are not so straightforward. We could be interested in the long-term proportion of time the system spends in state, or the states seen by arriving customers, or even the state left behind by a departing customer. We have seen that these quantities need not be the same. The design of the simulation can ease or make difficult the measurement of a particular parameter.

In the remainder of the section we are interested in the parameters of the system when it is in steady state, usually either the mean number of customers in the system or the long-term proportion of time the system has a certain number of customers. We cover the following approaches to simulating a queueing system.

- Simulation through independent replication;
- Time-sampled process: $\{N(k\delta)\}$;
- Embedded Markov chain and state occupancies: $\{N(t_k), T_k\}$;
- Replication through regenerative cycles.

12.10.2 Simulation through Independent Replications

Simulation through independent replications involves simulating a process R times to obtain a set of R independent observations $\{X(t, \zeta_1), X(t, \zeta_2), \ldots, X(t, \zeta_R)\}$. We use a function of the observations to estimate a parameter θ of the random process:

$$\hat{\Theta}(\mathbf{X}_R) = g(X(t, \zeta_1), X(t, \zeta_2), \ldots, X(t, \zeta_R)).$$

For example, to estimate the mean of the process at time t we use:

$$\overline{X}_R(t) = \frac{1}{R}\sum_{r=1}^{R} X(t, \zeta_r). \tag{12.166}$$

To estimate the variance of the process at time t we use:

$$\hat{\sigma}_R^2(t) = \frac{1}{R-1}\sum_{r=1}^{R} (X(t, \zeta_r) - \overline{X}_R(t))^2. \tag{12.167}$$

By design the observations are independent random variables. In order to proceed, we also need to assume that the observations are Gaussian random variables. The usual approach of taking the sum of a sufficiently large number of variables and using the central limit theorem applies. We can also use a statistical test to check that the samples are close to Gaussian distributed. Once we have Gaussian observations, we can provide the confidence intervals from Eq. (8.58):

$$(\overline{X}_R - t_{\alpha/2,n-1}\hat{\sigma}_R/\sqrt{n}, \overline{X}_R + t_{\alpha/2,n-1}\hat{\sigma}_R/\sqrt{n}). \tag{12.168}$$

Equation (12.168) is used widely to provide approximate confidence intervals.

We note that the sample mean and variance estimators in Eqs. (12.166) and (12.167) and the associated confidence intervals allow us to identify time dependencies in the behavior of the random process. In particular, in the next example, we use them to identify the transient phase of a random process that has a steady state.

When the random process is a continuous function of time, the estimator can take the form of an integral. For example, for a Markov chain process we can estimate either the time average of the process or the proportion of time in state j in the rth replication by an integral over an interval of time T:

$$\hat{N}_r = \frac{1}{T}\int_0^T N(t, \zeta_r)\, dt \qquad \text{and} \qquad \hat{p}_j^{(r)} = \frac{1}{T}\int_0^T I_j(N(t, \zeta_r))\, dt. \tag{12.169}$$

$\{\hat{N}_r\}$ and $\{\hat{p}_j^{(r)}\}$ provide the independent random variables that can be used to obtain a confidence interval for the time average of $N(t)$ and the proportion of time that $N(t) = j$.

12.10.3 Time-Sampled Process Simulation

A simple approach to simulating continuous-time queueing systems is to use **time-sampled process simulation**. The time axis is divided into small intervals of length δ and a discrete-time process is simulated. The following example demonstrates the approach.

Example 12.28 Transient of M/M/1 Queue Using Sampled-Time Approximation

Investigate the transient behavior of $N(t)$, the number of customers in an M/M/1 queueing system, using a sampled-time approach. Assume the system is initially empty. Generate 2000 steps of $\delta = 0.1$ seconds with $\mu = 1$ job/second and run two cases: $\lambda = 0.5$ and $\lambda = 0.9$ jobs/second. Replicate the simulation 20 times and plot the sample mean of the process across the 20 replications (Eq. 12.166). Find the covariance function for each realization and plot the average of the covariance functions across the 20 replications.

The sampled-time approximation involves simulating a system in small steps of δ seconds. For a birth-death process (such as the M/M/1 queue) in state $j > 0$, three outcomes can occur in δ seconds: (1.) no arrival and no departure occur with probability $1 - (\lambda_j + \mu_j)\delta$; (2.) one arrival occurs with probability $\lambda_j\delta$; (3.) one departure occurs with probability $\mu_j\delta$. We can adjust for the $j = 0$ state by letting $\mu_0 = 0$, and the $j = N_{max}$ state by letting $\lambda_{Nmax} = 0$. Note that the state-transition diagram of this sampled-time queueing system has the structure of the discrete-time Markov chain in Example 11.49. We use the code for that example to generate 20 realizations of 2000 steps of $N(k\delta)$, which corresponds to 200 seconds of time.

Figure 12.29(a) shows the sample mean of 20 realizations of $N(t)$. Note that this sample mean averages over 20 processes that can each exhibit a lot of variation, see Figs. 11.20 and 11.21. Consequently the averaged realizations still exhibit quite a bit of variation. The lower curve corresponds to $\rho = 0.5$, which can be seen to reach and vary about the true mean of $E[N] = 1$ after about 100 steps (10 seconds). The higher curve corresponds to $\rho = 0.9$, which is a much higher utilization. The true mean in this case is $E[N] = 9$ and it can be seen that the average of the realizations does not reach the area of the mean until about 1400 steps. Thus we see that the transient period increases dramatically as the utilization approaches 1.

Figure 12.29(b) shows the sample mean of the normalized covariance functions of the 20 realizations of $N(t)$. For $\rho = 0.5$, the autocovariance does not reach 0 until about 200 steps. Furthermore, for $\rho = 0.9$, the autocovariance is approximately 0.6 after 200 steps. This much longer sustained correlation is another indicator of the increase in transient time as utilization is increased.

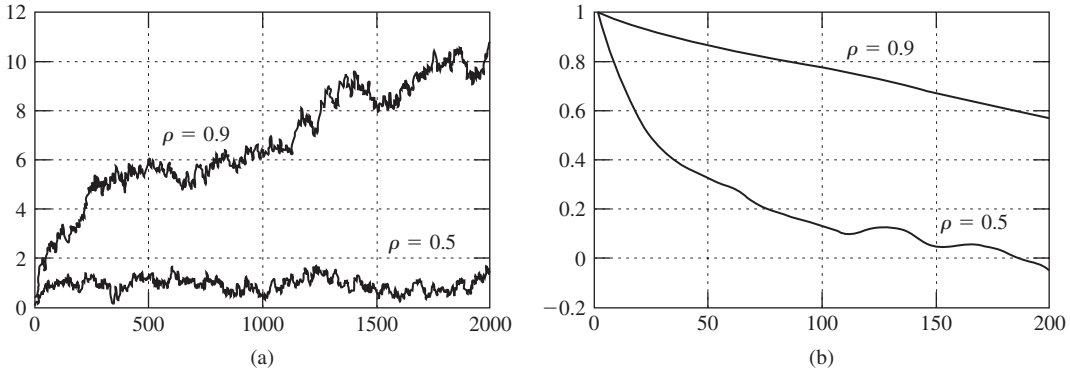

FIGURE 12.29
(a) Transient of M/M/1 queue using sampled-time approach, $\rho = 0.5, 0.9$; (b) normalized covariance of M/M/1 queue, $\rho = 0.5, 0.9$.

In order to approximate the queueing process accurately, the time-sampled approach requires that we use a small step size. In addition to possibly increasing the amount of computation required to perform the simulation size, a small step size has the effect of making more adjacent samples highly correlated. This is clearly evident in the observed autocovariance function in the above example.

The correlation of samples poses a problem in estimating parameters of a queueing process from a single realization. Suppose we are interested in estimating the mean of $\{N(k\delta)\}$ from a *single realization* of the process:

$$\overline{N}_n = \frac{1}{n}\sum_{k=1}^{n} N(k\delta). \tag{12.170}$$

The terms in the series $\{N(k\delta)\}$ are correlated, so from Eq. (9.108), assuming that the process is wide sense stationary, the variance of the sample mean is then larger than it would be for iid samples:

$$\text{VAR}[\overline{N}_n] = \frac{1}{n}\left[C_N(0) + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)C_N(k) \right], \tag{12.171}$$

where $C_N(k)$ is the covariance function of $N(t)$. Only $C_N(0)$, which corresponds to the variance of N, would be present if the observations were uncorrelated. Example 12.28 demonstrated how $N(t)$ in queueing systems can maintain significant correlation for significant periods of time. The example also illustrated how the process $N(t)$ becomes more correlated as the utilization increases. As discussed in Examples 9.49 and 9.50, the net effect is that the convergence of the sample mean to $E[N]$ is slower than if the samples were independent. This larger variance can be taken into account by gathering estimates for the covariance terms $C_N(k)$ and using Eq. (12.168) in the calculation of confidence intervals. (See [Law, p. 556] for a discussion on such confidence intervals).

The relative frequencies of the states provide estimates for the long-range proportion of time spent in each state:

$$\hat{p}_j = \frac{1}{n}\sum_{k=1}^{n} I_j(k\delta) \tag{12.172}$$

where I_j is the indicator function for the event $\{N(k\delta)\} = j\}$. Relative frequencies are a special case of sample means so the same cautions regarding the variance of the estimates and convergence rates apply.

The **method of batch means**, introduced in Section 8.4, provides an approach to dealing with the correlation among samples. A long simulation run is divided into multiple segments that are sufficiently long that the samples from different segments have low correlation. The parameter estimates from different segments, e.g., sample mean or relative frequencies, are treated as independent random variables and the standard statistical tools are applied to the batch means and batch relative frequencies.

Example 12.29 Confidence Intervals Using Batch Means

Use the method of batch means to estimate the mean of the M/M/1 queue when $\lambda = 0.5$ and $\mu = 1$ job per second. Each realization should consist of 8 batches of 600 steps. Replicate each simulation five times.

Five replications of 5000-step realizations were carried out. The first 200 samples from each realization were discarded to remove bias from the initial transient. The remaining 4800 samples in each realization were divided into 8 batches. Table 12.2(a) shows the means for each of the resulting 40 batches. For each realization the sample mean and sample standard deviation for the 8 batch means were calculated and are shown in Table 12.2(b). Confidence intervals were then calculated for each realization. For a 90% confidence level $(a = 10\%)$, $t_{a/2} = 1.8946$ and $\delta = t_{a/2}\sigma/\sqrt{8}$. The upper and lower limits of the confidence interval for the mean of the process are given in the two rightmost columns of Table 12.2(b). Every confidence interval contains the value 1, which is the expected value of the M/M/1 queue when $\rho = 1/2$.

TABLE 12.2a Sequence of batch means for five replications

r/b	1	2	3	4	5	6	7	8
1	0.84500	0.70667	0.51500	4.57167	0.30500	3.56000	1.75167	0.91167
2	0.83000	0.66000	0.97667	1.21833	1.14667	1.16333	2.39833	0.61000
3	0.96000	0.55333	0.89833	0.62500	0.31000	3.39167	0.86167	0.43333
4	2.73333	1.06167	0.62167	0.45667	2.17333	1.30000	0.57667	0.88167
5	1.14000	0.85667	0.82500	1.07167	0.67833	1.02167	1.08833	1.44667

TABLE 12.2b Confidence interval for mean for each of five replications

r/b	Mean	σ	δ	Lower	Upper
1	1.6458	1.57547	1.05532	0.59052	2.7011
2	1.1254	0.56347	0.37744	0.74798	1.5029
3	1.0042	0.99199	0.66448	0.33969	1.6686
4	1.2256	0.81934	0.54883	0.67679	1.7745
5	1.0160	0.23455	0.15711	0.85893	1.1732

TABLE 12.2c Sequence of batch confidence intervals across five replications

r/b	1	2	3	4	5	6	7	8
Mean	1.3017	0.7677	0.7673	1.5887	0.9227	2.0873	1.3353	0.8567
σ	0.8099	0.1972	0.1931	1.6965	0.7796	1.2727	0.7356	0.3846
δ	0.7721	0.1880	0.1841	1.6174	0.7432	1.2134	0.7013	0.3667
Upper	2.0738	0.9557	0.9515	3.2061	1.6659	3.3007	2.0366	1.2234
Lower	0.5296	0.5796	0.5832	−0.0287	0.1795	0.8739	0.6340	0.4900

Table 12.2(c) gives the 90% confidence interval that is calculated for the batch means *across* different replications. These batches are truly independent and will not be affected by correlation effects. It is important to determine whether any evidence of bias exists in the earlier batches due to the initial transient phase. It can be seen that the second and third columns do not include the value 1 by a small margin.

We also calculated a 90% confidence interval for the combined 40 batches and obtained $(1.2034 - 0.24575, 1.2034 + 0.24575) = (0.95765, 1.449)$. Finally, we calculated a 90% confidence interval based on the sample means of the 5 realizations, and obtained $(1.2034 - 0.25096, 1.2034 + 0.25096) = (0.95244, 1.4544)$. Note that the latter 5 realizations are truly independent and constitute a pure application (no batching) of the replication method.

12.10.4 Simulation Using Embedded Markov Chains

Many queueing systems have natural embedding points that lead to discrete-time Markov chains. We saw in Chapter 11 that queueing systems that are modeled by continuous-time Markov chains can be defined in terms of an embedded Markov chain and exponentially distributed state occupancy times. In this chapter we saw that the distribution of the steady state number of customers in an M/G/1 system can also be observed through an embedded Markov chain. In this section we discuss **simulation based on embedded Markov chains**.

First, let $N(t)$ be the number of customers in a queueing system that is modeled by a continuous-time Markov chain. The transition rate matrix Γ for the process provides us with the transition probabilities of the embedded chain as well as the state occupancy times (see Eq. 11.35). In Example 11.50 we used this approach to generate realizations of an M/M/1 queue. The output of this simulation is a sequence of states $\{N_i\}$ and the corresponding state occupancy times $\{T_i\}$. The relative frequencies obtained from the sequence of states provide us with an estimate for the state probabilities $\{\pi_j\}$ of the embedded Markov chain. The occupancy times according to their corresponding state, e.g., $\{T_k(j), k = 1, \ldots, n_j\}$ for state j, can also provide us with an estimate for the state occupancy times. We can obtain an estimate for the mean of $N(t)$ directly:

$$\hat{N} = \frac{1}{T}\int_0^T N(t)\,dt = \frac{1}{T}\sum_{k=1}^{n} N_k T_k. \tag{12.173}$$

An estimate for long-term proportion of time in state j is obtained similarly:

$$\hat{p}_j = \frac{1}{T}\int_0^T I_j(t)\,dt = \frac{1}{T}\sum_{k=1}^{n_j} T_k(j). \tag{12.174}$$

If the Markov chains that model the system are ergodic, then the above estimates will converge to the correct steady state values.

Example 12.30 M/M/1 Steady State Probabilities Using Embedded Markov Chain

Use the embedded Markov chain approach to estimate the state probabilities in an M/M/1 system with $\lambda = 0.75$ and $\mu = 1$. Calculate the proportion of time spent in each state and obtain confidence intervals for these values by using replication.

The code in Example 11.50 can be modified to calculate Eq. (12.174) by accumulating the total time spent in each state as the simulator generates each new state and occupancy time. Each realization was 1800 seconds in duration, but no data was gathered during the first 300 seconds of the simulation. Eight pmf estimates were obtained and the sample mean and standard deviation as well as a 90% confidence interval for each state probability were computed using the eight independent estimates from the replication. The results are shown in Fig. 12.30. It can be seen that there is generally good agreement between the theoretical pmf and the confidence intervals.

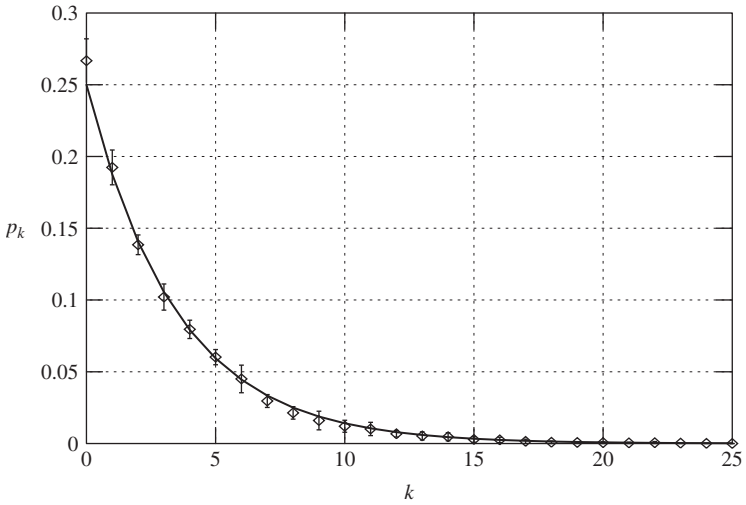

FIGURE 12.30
Confidence intervals for steady state M/M/1 pmf.

The following example shows that we can simulate an M/G/1 system using another type of embedded Markov chain.

Example 12.31 Simulating M/G/1 Using Embedded Markov Chains

Section 12.7 showed that the steady state distribution for the number of customers in an M/G/1 system is the same as the distribution for the number left behind by a customer departure. Furthermore, the number of customers left behind by the jth customer departure, N_j, forms a discrete-time Markov chain as follows:

$$N_j = (N_{j-1} - 1)^+ + M_j \tag{12.175}$$

where M_j is the number of arrivals during the service time of the jth customer and where

$$(x)^+ \triangleq \max(0, x).$$

Therefore we can obtain the steady state pmf for $N(t)$ in an M/G/1 system by finding the transition probability matrix associated with Eq. (12.175) and applying the methods developed in Section 11.6. We explore this approach further in the problems.

Next we introduce *Lindley's recursion for the waiting time in a G/G/1 system* as a final application of embedded Markov chain methods. Assume that the customer interarrival times and service times are independent random variables with arbitrary distributions. We focus on the waiting time experienced by an arriving customer and we show that the sequence of waiting times forms a Markov chain.

Let $a_1, a_2, \ldots$ denote the customer interarrival times and let $\tau_1, \tau_2, \ldots$ be their corresponding service times. Let W_n be the waiting time of the nth customer. Suppose the $(n + 1)$st customer arrives to a nonempty system, as shown in Fig. 12.31(a). Note that we must have:

$$W_n + \tau_n = a_{n+1} + W_{n+1}$$

in order for the arriving customer to find a nonempty system. It then follows that the waiting time for the $(n + 1)$st customer must be given by:

$$W_{n+1} = W_n + \tau_n - a_{n+1} \text{ if } W_n + \tau_n - a_{n+1} \geq 0. \tag{12.176a}$$

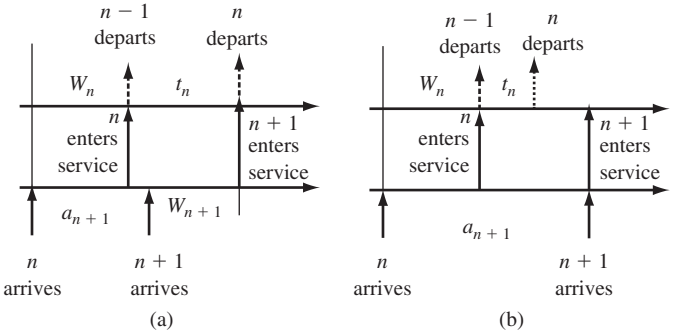

FIGURE 12.31
Customer arrivals and departures in G/G/1 queue.

On the other hand, the arriving customer finds an empty system (Fig. 12.31b) under the following conditions:

$$W_{n+1} = 0 \quad \text{if} \quad W_n + \tau_n - a_{n+1} < 0. \tag{12.176b}$$

Therefore we conclude that the sequence of waiting times is given by **Lindley's recursion**:

$$W_{n+1} = \max(0, W_n + \tau_n - a_{n+1}). \tag{12.177}$$

W_{n+1} depends on the past only through W_n and τ_n and a_{n+1}. Since τ_n and a_{n+1} are from iid sequences and are independent of each other, we conclude that W_{n+1} is a Markov process with stationary transition probabilities. Note that W_n assumes a continuum of values. We can generate the sequence of total delays experienced by the sequence of customers as follows: $T_n = W_n + a_n$.

Equation (12.177) can be used to derive an integral equation for the steady state waiting time of customers in a G/G/1 system [Kleinrock, p. 282]. The equation is similar to the Wiener–Hopf equation we encountered in Section 10.4 and usually requires transform methods to solve. However, *Eq. (12.177) is remarkably simple to use in simulations.*

Example 12.32 Estimating Waiting Time Distribution Using Lindley's Recursion

Estimate the distribution of the customer waiting times in an M/M/1 queue when $\lambda = 0.9$ and $\mu = 1$ job per second. Compare the empirical cdf of the observed total time in the system with the theoretical distribution.

Lindley's recursion can be readily implemented in Octave. Arrays of exponential interarrival times with $\lambda = 0.9$ and service times with $\mu = 1$ job per second are generated initially. Lindley's recursion is then used to compute the sequence of waiting times and total delays for the sequence of customers. The Octave function `empirical_cdf` is used to obtain the cdf of the observations. In the simulation a sequence of 2000 waiting and total times were collected and no data was deleted to allow for an initial transient period. Figure 12.32 compares the empirical cdf with the distribution for waiting time in an M/M/1 system with $\rho = 0.9$. A test such as the Kolmogorov–Smirnov test can be applied to assess goodness of fit of the empirical distribution to the hypothetical distribution.

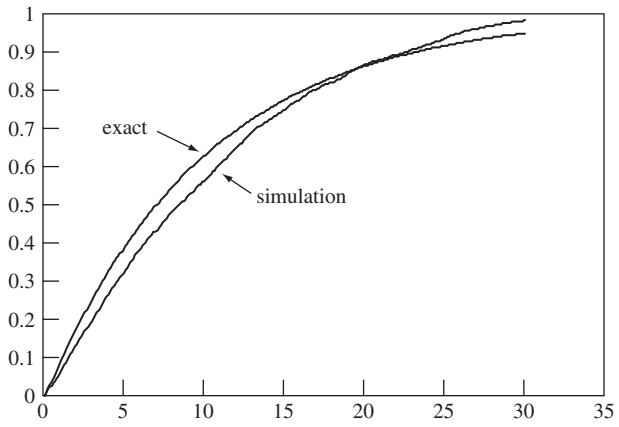

FIGURE 12.32
Empircial cdf of M/M/1 queue using Lindley's recursion, $\rho = 0.9$.

12.10.5 Replication through Regenerative Cycles

In Section 7.5 we considered renewal processes where time is divided into intervals according to an iid sequence of positive random variables $\{X_i\}$. We associated with each interval X_i a cost C_i. We then proved the following result Eq. (7.47):

$$\lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{M(t)} C_j = \frac{E[C]}{E[X]} \tag{12.178}$$

where $E[C]$ is the average cost per cycle and $E[X]$ is the mean cycle length.

The **regenerative method for simulation** involves finding renewal points in a queueing system where the process "restarts" itself so that its future is independent of the past. For example, in many queueing systems this renewal or regeneration occurs when a customer arrives to an empty system. Measurements taken during different cycles are then independent random variables. Thus in effect the regenerative method partitions a single simulation into a number of independent replications.

The long-term time average of $C(t)$ in Eq. (12.178) is given by the ratio of the sample mean for the measurements for C and the sample mean for X. For example, if we are interested in the probability that the system is in state j, then we let C_j be the time the system is in state j during the jth cycle:

$$C_i = \int_{R_{i-1}}^{R_i} I_j(t)\, dt = \sum_{k=1}^{n_i(j)} T_k^i(j) \tag{12.179}$$

where $n_i(j)$ is the number of times state j occurred during the ith cycle and $T_k^i(j)$ is the occupancy time of the jth occurrence of state j during the ith cycle. The corresponding estimate for the proportion of time in state j is:

$$\hat{p}_j = \frac{\frac{1}{n}\sum_{k=1}^{n_i(j)} T_k^i(j)}{\frac{1}{n}\sum_{i=1}^{n} X_i}. \tag{12.180}$$

On the other hand if we are interested in the mean of $N(t)$, we let

$$C_i = \int_{R_{i-1}}^{R_i} N(t)\, dt = \sum_{k=1}^{n_i} N_k^i T_k^i \tag{12.181}$$

where n_i is the number of states visited during the ith cycle. The corresponding estimate for the mean is:

$$\hat{N} = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n_i} N_k^i T_k^i}{\frac{1}{n}\sum_{i=1}^{n} X_i}. \tag{12.182}$$

The numerators and denominators in Eqs. (12.180) and (12.182) individually are strongly consistent estimators for their corresponding parameters. Therefore the estimators formed by taking their ratios in Eqs. (12.180) and (12.182) are also strongly consistent. Note, however, that the ratios provide biased estimates. We discuss confidence intervals after the following example.

Example 12.33 Regenerative Method for M/M/1 Simulation

Estimate the mean waiting time of customers in the system in Example 12.28 using the regenerative method to analyze the sequence of waiting times produced by Lindley's recursion.

Let a cycle consist of the time from when a customer arrives to an empty system until the next time a customer arrives to an empty system. We are interested in the average waiting time experienced by customers over a long period of time. Suppose we measure the number of customers serviced in a sequence of cycles $\{N_c(i)\}$, and the total of the waiting times of all customers in the cycle $\{W_{agg}(i)\}$. Each of these sequences is iid and so each one will converge to its respective mean. The ratio of the two expressions provides an estimate for the mean waiting time (see Problem 12.78):

$$\hat{W} = \frac{\frac{1}{n}\sum_{i=1}^{n} W_{agg}(i)}{\frac{1}{n}\sum_{i=1}^{n} N_c(i)}. \tag{12.183}$$

It is easy to prepare a simulation to gather $\{N_c(i)\}$, $\{W_{agg}(i)\}$, and the sequence of cycle durations $\{X_i\}$ using Lindley's recursion because each regeneration point is marked by arriving customers that have zero waiting time. The resulting sequences can be parsed according to their respective cycles and the above cycle statistics can then be gathered.

A simulation with 4000 customer arrivals to an M/M/1 systems with $\lambda = 0.9$ and $\mu = 1$ was conducted and the results in Table 12.3 were obtained. The 4000 arrivals produced 366 cycles. The ratio of the mean number of customers serviced in a cycle to the mean cycle duration gives the following estimate for the arrival rate:

$$\text{Arrival Rate Estimate} = 10.842/11.913 = 0.91,$$

which is close to $\lambda = 0.9$. The estimate for the mean waiting time obtained from the ratio in Eq. (12.183) was 8.80. From Eq. (9.27) the mean waiting time for this M/M/1 queue is $E[W] = 9$, which again is quite close.

TABLE 12.3 Per regenerative cycle statistics for M/M/1 queue

M/M/1 Mean Waiting Time	
L = 4000	TotCycle = 366
MeanCycle = 11.913	STDCycle = 41.374
MeanCount = 10.842	STDCount = 39.236
MeanCycleWait = 95.424	STDCycleWait = 612.20
MeanWait = 8.8017	

Of course the whole point of striving to get independent observations is to produce confidence intervals. In [Law, p. 559] an approximate confidence interval is developed for an estimator of the form in Eq. (12.183). The pair $(W_{agg}(i), N_c(i))$ form an iid sequence but in general $W_{agg}(i)$ and $N_c(i)$ are correlated. It can be shown that for large n the estimator in Eq. (12.183) is asymptotically Gaussian with mean $E[W]$ and variance:

$$\hat{\sigma}_{\hat{W}}^2(n) = \hat{\sigma}_{W_{agg}}^2(n) - 2\hat{W}(n)\hat{\sigma}_{W_{agg},N_c}^2 + (\hat{W}(n))^2\hat{\sigma}_{N_c}^2(n) \tag{12.184}$$

where $\hat{\sigma}_{W_{agg},N_c}^2$ is the estimator for the covariance of $W_{agg}(i)$ and $N_c(i)$. This result leads to the following confidence interval:

$$\left(\hat{W} - \frac{z_{1-\alpha/2}\hat{\sigma}_{\hat{W}}\sqrt{n}}{\hat{N}_c}, \hat{W} + \frac{z_{1-\alpha/2}\hat{\sigma}_{\hat{W}}\sqrt{n}}{\hat{N}_c}\right). \tag{12.185}$$

The required estimates for the variances and covariances of $W_{agg}(i)$, $N_c(i)$ can be made from the per-cycle statistics.

In practice the regenerative method is difficult to apply because the occurrence of regenerative instances is not controllable. For example, the busy periods of queueing systems under heavy traffic vary dramatically and so the occurrence of regeneration points can be quite unpredictable.

In conclusion, simulation straddles the space between theoretical models and the real world. The basic introduction to simulation methods for queueing systems provides an excellent opportunity to illustrate the role of statistical techniques in the application of probability models to real world problems. The presence of transient effects and correlations in the observed data provide an excellent opportunity to emphasize the need to apply probability models and statistical tools with care. But we should end this book on a positive note: the availability of plentiful and inexpensive computing allows us to extend the reach of our theoretical and simulation models into new frontiers!

SUMMARY

- A queueing system is specified by the arrival process, the service time distribution, the number of servers, the waiting room, and the queue discipline.
- Little's formula states that under very general conditions: The mean number in a system is equal to the product of the mean arrival rate and the mean time spent in the system.
- In M/M/1, M/M/1/K, M/M/c, M/M/c/c, and M/M/∞ queueing systems, the number of customers in the system is a continuous-time Markov chain. The steady state distribution for the number in the system is found by solving the global balance equations for the Markov chain. The waiting time and delay distribution when the service discipline is first come, first served is found by using the arriving customer's distribution.
- If the arrival process in a queueing system is a Poisson process and if the customer interarrival times are independent of the service times, then the arriving customer's distribution is the same as the steady state distribution of the queueing system.

- In M/G/1 queueing systems the arriving customer's distribution and the departing customer's distribution are both equal to the steady state distribution of the queueing system. The steady state distribution for the number of customers in an M/G/1 system can be found by embedding a discrete-time Markov chain at the customer departure instants.
- Burke's theorem states that the output process of M/M/1, M/M/c, and M/M/∞ systems at steady state are Poisson processes, and that the departure instants prior to time t are independent of the state of the system at time t. As a result, feedforward combinations of queueing systems with exponential service times have a product-form solution.
- Jackson's theorem states that for networks of queueing systems with exponential service times and external Poisson input processes, the joint state pmf is of product form. If the network of queues is open, the marginal state pmf of each queue is the same as that of a queue in isolation that has Poisson arrivals of the same rate. If the network of queues is closed, finding the joint state pmf requires finding a normalization constant. The mean value analysis method allows us to find the mean number in each queue, the mean time spent in each queue, and the arrival rate in each queue in a closed network of queues.
- Approaches to simulating queueing systems include replication, time sampling, and embedded Markov chains. The analysis of observations must deal with the effect of transient behavior as well as the correlation of observations.

CHECKLIST OF IMPORTANT TERMS

a/b/m/K
Arrival rate
Arriving customer's distribution
Burke's theorem
Carried load
Closed networks of queues
Departing customer's distribution
Erlang *B* formula
Erlang *C* formula
Finite-source queueing system
Head-of-line priority service
Interarrival times
Jackson's theorem
Lindley's recursion
Little's formula
Mean value analysis
Method of batch means
M/G/1 queueing system
M/M/c queueing system
M/M/c/c queueing system
M/M/1 queueing system
M/M/1/K queueing system
Offered load
Open networks of queues
Pollaczek–Khinchin mean value formula
Pollaczek–Khinchin transform equation
Product-form solution
Queue discipline
Regenerative method for simulation
Residual service time
Server utilization
Service discipline
Service time
Simulation based on embedded Markov chains
Simulation through independent replication
Time-sampled process simulation
Total delay
Traffic intensity
Waiting time

ANNOTATED REFERENCES

References [1] and [2] provide an introduction to queueing theory at a level slightly higher than that given here. Reference [2] is an invaluable source of classical queueing theory results in telephony problems. Reference [3] demonstrates the application of queueing theory to data communication networks. References [1–7] discuss techniques for simulating queueing systems and for analyzing the resulting data. [8–10] presents excellent discussions on reversible processes and M/G/c/c and M/G/∞.

1. L. Kleinrock, *Queueing Systems*, vol. 1, Wiley, New York, 1975.
2. R. B. Cooper, *Introduction to Queueing Theory*, 2nd ed., North Holland, 1981. Reprinted by CEE Press of the George Washington University.
3. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
4. A. M. Law and W. D. Kelton, *Simulation, Modeling, and Analysis*, 2nd ed., McGraw-Hill, New York, 1999.
5. J. Banks, J. S. Carson II, and B. L. Nelson, *Discrete-Event System Simulation*, Prentice-Hall, Upper Saddle River, NJ, 1996.
6. G. S. Fishman, *Discrete-Event Simulation: Modeling, Programming, and Analysis*, Springer-Verlag, New York, 2001.
7. S. M. Ross, *Stochastic Processes*, Wiley, New York, 1983.
8. M. Reiser and S. S. Lavenberg, "Mean-value analysis of closed multichain queueing networks," *J. Assoc. Comput. Mach.* 27: 313–322, 1980.
9. S. S. Lavenberg, *Computer Performance Modeling Handbook*, Academic Press, New York, 1983.
10. K. Pawlikowski, "Steady-state simulation of queueing processes: survey of problems and solutions," *ACM Computing Surveys*, Vol. 22, No. 2, pp. 123–170, 1990.

PROBLEMS

Sections 12.1 and 12.2: The Elements of a Queueing Network and Little's Formula

12.1. Describe the following queueing systems: M/M/1, M/D/1/K, M/G/3, D/M/2, G/D/1, D/D/2.

12.2. Suppose that a queueing system is empty at time $t = 0$, let the arrival times of the first six customers be 1, 3, 4, 7, 8, 15, and let their respective service times be 3.5, 4, 2, 1, 1.5, 4. Find S_i, τ_i, D_i, W_i, and T_i for $i = 1, \ldots, 5$; sketch $N(t)$ versus t; and check Little's formula by computing $\langle N \rangle_t$, $\langle \lambda \rangle_t$, and $\langle T \rangle_t$ for each of the following three service disciplines:

- **(a)** First come, first served.
- **(b)** Last come, first served.
- **(c)** Shortest job first (assume that the precise service time of each job is known before it enters service).

12.3. A data communication line delivers a block of information every 10 μs. A decoder checks each block for errors and corrects the errors if necessary. It takes 1 μs to determine whether a block has any errors. If the block has one error, it takes 5 μs to correct it, and if it has more than one error it takes 20 μs to correct the error. Blocks wait in a queue when the decoder falls behind. Suppose that the decoder is initially empty and that the numbers of errors in the first ten blocks are 0, 1, 3, 1, 0, 4, 0, 1, 0, 0.

(a) Plot the number of blocks in the decoder as a function of time.

(b) Find the mean number of blocks in the decoder.

(c) What percentage of the time is the decoder empty?

12.4. Three queues are arranged in a loop as shown in Fig. P12.1. Assume that the mean service time in queue i is $m_i = 1/\mu_i$.

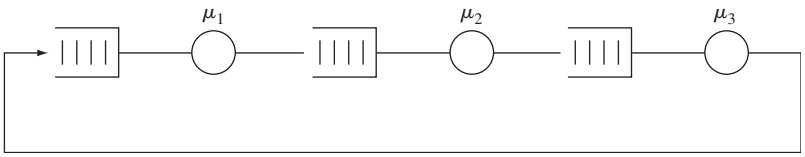

FIGURE P12.1

(a) Suppose the queue has a single customer circulating in the loop. Find the mean time $E[T]$ it takes the customer to cycle around the loop. Deduce from $E[T]$ the mean arrival rate λ at each of the queues. Verify that Little's formula holds for these two quantities.

(b) If there are N customers circulating in the loop, how are the mean arrival rate and the mean cycle time related?

12.5. A very popular barbershop is always full. The shop has two barbers and three chairs for waiting, and as soon as a customer completes his service and leaves the shop, another enters the shop. Assume the mean service time is m.

(a) Use Little's formula to relate the arrival rate and the mean time spent in the shop.

(b) Use Little's formula to relate the arrival rate and the mean time spent in service.

(c) Use the above formulas to find an expression for the mean time spent in the system in terms of the mean service time.

12.6. In Problem 12.3, suppose that the probabilities of zero, one, and more than one errors are p_0, p_1, and p_2, respectively. Use Little's formula to find the mean number of blocks in the decoder.

12.7. A communication network receives messages from R sources with mean arrival rates $\lambda_1, \ldots, \lambda_R$. On the average there are $E[N_i]$ messages from source i in the network.

(a) Use Little's formula to find the average time $E[T_i]$ spent by type i customers in the network.

(b) Let λ denote the total arrival rate into the network. Use Little's formula to find an expression for the mean time $E[T]$ spent by customers (of all types) in the network in terms of the $E[N_i]$.

(c) Combine the results of part a and part b to obtain an expression for $E[T]$ in terms of $E[T_i]$. Derive the same expression using $A(t)$ the arrival processes for each type.

Section 12.3: The M/M/1 Queue

12.8. **(a)** Find $P[N \geq n]$ for an M/M/1 system.

(b) What is the maximum allowable arrival rate in a system with service rate μ, if we require that $P[N \geq 10] = 10^{-3}$?

12.9. A decision to purchase one of two machines is to be made. Machine 1 has a processing rate of μ transactions/hour and it costs B dollars/hour to operate whether idle or not; machine 2 is twice as fast but costs twice as much to operate. Suppose that transactions arrive at the system according to a Poisson process of rate λ and that the transaction processing times are exponentially distributed. The total cost of the system is the operation cost plus a cost of A dollars for each hour a customer has to wait.

(a) Find expressions for the total cost per hour for each of the systems. Plot this cost versus the arrival rate.

(b) If $A = B/10$, for what range of arrival rates is machine 1 cheaper? Repeat for $A = 10B$.

12.10. Consider an M/M/1 queueing system in which each customer arrival brings in a profit of \$5 but in which each unit time of delay costs the system \$1. Find the range of arrival rates for which the system makes a net profit.

12.11. Consider an M/M/1 queueing system with arrival rate λ customers/second.

(a) Find the service rate required so that the average queue is five customers (i.e., $E[N_q] = 5$).

(b) Find the service rate required so that the queue that forms from time to time has mean 5 (i.e., $E[N_q \mid N_q > 0] = 5$).

(c) Which of the two criteria, $E[N_q]$ or $E[N_q \mid N_q > 0]$, do you consider the more appropriate?

12.12. Show that the pth percentile of the waiting time for an M/M/1 system is given by

$$x = \frac{1/\mu}{1 - \rho} \ln\left(\frac{\rho}{1 - p}\right).$$

12.13. Consider an M/M/1 queueing system with service rate two customers per second.

(a) Find the maximum allowable arrival rate if 90% of customers should not have a delay of more than 3 seconds.

(b) Find the maximum allowable arrival rate if 90% of customers should not have to wait for service for more than 2 seconds. *Hint:* Use the result from Problem 12.12, and then find λ by trial and error.

12.14. Verify Eq. (12.36) for the steady state pmf of an M/M/1/K system.

12.15. Consider an M/M/1/2 queueing system in which each customer accepted into the system brings in a profit of \$5 and each customer rejected results in a loss of \$1. Find the arrival rate at which the system breaks even.

12.16. For an M/M/1/K system show that

$$P[N = k \mid N < K] = \frac{P[N = k]}{1 - P[N = K]} \qquad 0 \leq k < K.$$

Why does this probability represent the proportion of arriving customers who actually enter the system and find exactly k customers in the system?

12.17. **(a)** Use the matrix exponential method of Eq. (11.72) to find the transient solution for the state pmfs for an M/M/1/5 queue under the following conditions:

(i) $\rho = 0.5$ and $N(0) = 0$, $N(0) = 2$, $N(0) = 5$;

(ii) $\rho = 1$ and $N(0) = 0$, $N(0) = 2$, $N(0) = 5$.

(b) Plot $E[N(t)]$ vs. t for the cases considered in part a.

12.18. Suppose that two types of customers arrive at a queueing system according to independent Poisson process of rate $\lambda/2$. Both types of customers require exponentially distributed service times of rate μ. Type 1 customers are always accepted into the system, but type 2 customers are turned away when the total number of customers in the system exceeds K.

(a) Sketch the transition rate diagram for $N(t)$, the total number of customers in the system.

(b) Find the steady state pmf of $N(t)$.

12.19. Consider the queueing system in Problem 12.18 with $K = 5$ and with a maximum system occupancy of 10 customers. In this problem we use the matrix exponential method of Eq. (11.72) to explore how the system adjusts to sudden increases in load.

(a) Find the transient state pmf for the system with $\lambda = 1/2$ and $\mu = 1$, assuming that initially there are 5 customers in the system.

(b) Suppose that at time 20, the λ increases to 1. Find the transient state pmf after this surge in traffic.

Section 12.4: Multiserver Systems: M/M/*c*, M/M/*c*/*c*, and M/M/∞

12.20. Find $P[N \geq c + k]$ for an M/M/c system.

12.21. Customers arrive at a shop according to a Poisson process of rate 12 customers per hour. The shop has two clerks to attend to the customers. Suppose that it takes a clerk an exponentially distributed amount of time with mean 5 minutes to service one customer.

(a) What is the probability that an arriving customer must wait to be served?

(b) Find the mean number of customers in the system and the mean time spent in the system.

(c) Find the probability that there are more than 4 customers in the system.

12.22. Little's formula applied to the servers implies that the mean number of busy servers is $\lambda E[\tau]$. Verify this by explicit calculation of the mean number of busy servers in an M/M/c system.

12.23. Inquiries arrive at an information center according to a Poisson process of rate 10 inquiries per second. It takes a server 1/2 second to answer each query.

(a) How many servers are needed if we require that the mean total delay for each inquiry should not exceed 4 seconds, and 90% of all queries should wait less than 8 seconds?

(b) What is the resulting probability that all servers are busy? Idle?

12.24. Consider a queueing system in which the maximum processing rate is $c\mu$ customers per second. Let k be the number of customers in the system. When $k \geq c$, c customers are served at a rate μ each. When $0 < k \leq c$, these k customers are served at a rate $c\mu/k$ each. Assume Poisson arrivals of rate λ and exponentially distributed times.

(a) Find the transition rate diagram for this system.

(b) Find the steady state pmf for the number in the system.

(c) Find $E[W]$ and $E[T]$.

(d) For $c = 2$, compare $E[W]$ and $E[T]$ for this system to those of M/M/1 and M/M/2 systems of the same maximum processing rate.

12.25. **(a)** Suppose that the queueing system in Problem 12.24 models a Web server where c is the maximum number of clients allowed to place queries at the same time. Discuss the impact of the choice of the parameter c on queueing and total delay performance.

(b) Consider the fact that while connected to the Web server, clients spend their time in three states: sending the query, waiting for the response, and thinking after each response. How does this affect the choice of c? Should the system impose a time-out limit on the customer's connection time?

12.26. Show that the Erlang B formula satisfies the following recursive equation:

$$B(c, a) = \frac{aB(c-1, a)}{c + aB(c-1, a)},$$

where $a = \lambda E[\tau]$.

12.27. Consider an M/M/5/5 system in which the arrival rate is 10 customers per minute and the mean service time is 1/2 minute.

(a) Find the probability of blocking a customer. *Hint:* Use the result from the Problem 12.26.

(b) How many more servers are required to reduce the blocking probability to 10%?

12.28. A tool rental shop has four floor sanders. Customers for floor sanders arrive according to a Poisson process at a rate of one customer every two days. The average rental time is exponentially distributed with mean two days. If the shop has no floor sanders available, the customers go to the shop across the street.

(a) Find the proportion of customers that go to the shop across the street.

(b) What is the mean number of floor sanders rented out?

(c) What is the increase in lost customers if one of the sanders breaks down and is not replaced?

12.29. **(a)** Show that the Erlang C formula is related to the Erlang B formula by

$$C(c, a) = \frac{cB(c, a)}{c - a\{1 - B(c, a)\}} \qquad \text{for } c > a.$$

(b) Show that this implies that $C(c, a) > B(c, a)$.

12.30. Suppose that department A in a certain company has three private videoconference lines connecting two sites. Calls arrive according to a Poisson process of rate 1 call/hour, and have an exponentially distributed holding time of 2 hours. Calls that arrive when the three lines are busy are automatically redirected to public video lines. Suppose that department B also has three private videoconference lines connecting the same sites, and that it has the same arrival and service statistics.

(a) Find the proportion of calls that are redirected to public lines.

(b) Suppose we consolidate the videoconference traffic from the two departments and allow all calls to share the six lines. What proportion of calls are redirected to public lines?

12.31. A $c = 10$ server blocking system handles two streams of customers that each arrive at rate $\lambda/2$. Type 1 customers have a mean service time of 1 time unit, and Type 2 customers have a service time of 3 time units. Compare the blocking performance of a system that allows customers to access any available server against one that allocates half the servers to each class. Does scale matter? Does the answer change if $c = 100$?

12.32. Suppose we use $P[N = c]$ from an M/M/∞ system to approximate $B(c, a)$ in selecting the number of servers in an M/M/c/c system. Is the resulting design optimistic or pessimistic?

12.33. During the evening rush hour, users log onto a peer-to-peer network at a rate of 10 users per second. Each user stays connected to the network an average of 1 hour.

(a) What is the steady state pmf for the number of customers logged onto the peer-to-peer network?

(b) Is steady state ever achieved?

(c) Is it reasonable to assume a Gaussian distribution for the number of customers in the system?

Section 12.5: Finite-Source Queueing Systems

12.34. A computer is shared by 15 users as shown in Fig. 12.14(b). Suppose that the mean service time is 2 seconds and the mean think time is 30 seconds, and that both of these times are exponentially distributed.

(a) Find the mean delay and mean throughput of the system.

(b) What is the system saturation point K^* for this system?

(c) Repeat part a if 5 users are added to the system.

12.35. A Web server that has the maximum number of clients connected is modeled by the system in Figure 12.14(b). Suppose that the system can handle a query in 10 milliseconds and the users click new queries at a rate of 1 every 5 seconds.

(a) Find the value of K^* for this system.

(b) Find the pmf for the number of requests found in queue by arriving queries.

12.36. Find the transition rate diagram and steady state pmf for a two-server finite-source queueing system.

12.37. Verify that Eqs. (12.84) and (12.81) give $E[T]$ as given in Eq. (12.72).

12.38. Consider a c-server, finite-source queueing system that allows no queueing for service. Requests that arrive when all servers are busy are turned away, and the corresponding source immediately returns to the "think" state, and spends another exponentially distributed think time before submitting another request for service.

(a) Find the transition rate diagram and show that the steady state pmf for the state of the system is

$$P_K[N = j] = \frac{\binom{K}{j} p^j (1-p)^{K-j}}{\sum_{i=0}^{c} \binom{K}{i} p^i (1-p)^{K-i}} \qquad i = 0, \ldots, c,$$

where c is the number of servers, K is the number of sources, and

$$p = \frac{\alpha/\mu}{1 + \alpha/\mu}.$$

(b) Find the probability that all servers are busy.

(c) Use the fact that arriving customers "see" the steady state pmf of a system with one less source to show that the fraction of arrivals that are turned away is given by $P_{K-1}(c)$. The resulting expression is called the Engset formula.

12.39. A video-on-demand system is modeled as a $c = 10$ server system that handles video chunk requests from K clients. Suppose that the system is modeled by the Engset system from Problem 12.38. Suppose that users generate requests at a rate of one per second and the each server can meet the request within 100 ms. Find the number of clients that can be connected if the probability of turning away a request is 10%? 1%?

Section 12.6: M/G/1 Queueing Systems

12.40. Find the mean waiting time and mean delay in an M/G/1 system in which the service time is a k-Erlang random variable (see Table 4.1) with mean $1/\mu$. Compare the results to M/M/1 and M/D/1 systems.

12.41. A $k = 2$ hyperexponential random variable is obtained by selecting a service time at random from one of two exponential random variables as shown in Fig. P12.2. Find the mean delay in an M/G/1 system with this hyperexponential service time distribution.

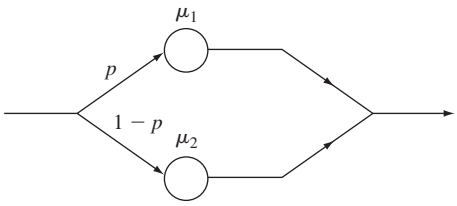

FIGURE P12.2

12.42. Customers arrive at a queueing system according to a Poisson process of rate λ. A fraction α of the customers require a fixed service time d, and a fraction $1 - \alpha$ require an exponential service time of mean $1/\mu$. Find the mean waiting time and mean delay in the resulting M/G/1 system.

12.43. Find the mean waiting time and mean delay in an M/G/1 system in which the service time consists of a fixed time d plus an exponentially distributed time of mean $1/\mu$.

12.44. Fixed-length messages arrive at a transmitter according to a Poisson process of rate λ. The time required to transmit a message and to receive an acknowledgment is d seconds. If a message is acknowledged as having been received correctly, then the transmitter proceeds with the next message. If the message is acknowledged as having been received in error, the transmitter retransmits the message. Assume that a message undergoes errors in transmission with probability p, and that transmission errors are independent.

(a) Find the mean and variance of the effective message service time.

(b) Find the mean message delay.

12.45. Packets at a router with a 1 Gigabit/second transmission line arrive at a rate of λ packets per second. Suppose that half the packets are 40 bytes long and half the packets are 1500 bytes long. Find the mean packet delay as a function of λ.

12.46. A file server receives requests at a rate of λ requests per second. The server can transmit files at a rate of 12.5 Megabytes per second. Suppose that file lengths have a Pareto distribution with mean 1 Megabyte.

(a) Find the average delay in meeting a file request.

(b) Discuss the effect of the Pareto distribution parameter on system performance.

12.47. Jobs arrive at a machine according to a Poisson process of rate λ. The service times for the jobs are exponentially distributed with mean $1/\mu$. The machine has a tendency to break down while it is serving customers; if a particular service time is t, then the probability that it will break down k times during this service time is a Poisson random variable with mean αt. It takes an exponentially distributed time with mean $1/\beta$ to repair the machine. Assume a machine is always working when it begins a job.

(a) Find the mean and variance of the total time required to complete a job. *Hint:* Use conditional expectation.

(b) Find the mean job delay for this system.

12.48. Consider a two-class nonpreemptive priority queueing system, and suppose that the lower-priority class is saturated (i.e., $\lambda_1 E[\tau_1] + \lambda_2 E[\tau_2] > 1$).

(a) Show that the rate of low-priority customers served by the system is $\lambda_2' = (1 - \lambda_1 E[\tau_1])/E[\tau_2]$. *Hint:* What proportion of time is the server busy with class two customers?

(b) Show that the mean waiting time for class 1 customers is

$$E[W_1] = \frac{(1/2)\lambda_1 E[\tau_1^2]}{1 - \lambda_1 E[\tau_1]} + \frac{E[\tau_2^2]}{2E[\tau_2]}.$$

12.49. Consider an M/G/1 system in which the server goes on vacations (becomes unavailable) whenever it empties the queue. If upon returning from vacation the system is still empty, the server takes another vacation, and so on until it finds customers in the system. Suppose that vacation times are independent of each other and of the other variables in the system. Show that the mean waiting time for customers in this system is

$$E[W] = \frac{(1/2)\lambda E[\tau^2]}{1 - \lambda E[\tau]} + \frac{E[V^2]}{2E[V]},$$

where V is the vacation time. *Hint:* Show that this system is equivalent to a nonpreemptive priority system and use the result of Problem 12.48.

12.50. Fixed-length packets arrive at a concentrator that feeds a synchronous transmission system. The packets arrive according to a Poisson process of rate λ, but the transmission system will only begin packet transmissions at times id, $i = 1, 2, \ldots$, where d is the transmission time for a single packet. Find the mean packet waiting time. *Hint:* Show that this is an M/D/1 queue with vacations as in Problem 12.49.

12.51. A queueing system handles two types of traffic. Type i traffic arrives according to a Poisson process and has exponentially distributed service times with mean $1/\mu_i$ for $i = 1, 2$. Suppose that type 1 customers are given nonpreemptive priority. Plot the overall and per-class mean waiting time versus λ if $\lambda_1 = \lambda_2 = \lambda$, $\mu_1 = 1$, $\mu_2 = 1/10$.

12.52. Consider a two-class priority M/G/1 system in which high-priority customer arrivals preempt low-priority customers who are found in service. Preempted low-priority customers are placed at the head of their queue, and they resume service when the server again becomes available to low-priority customers.

(a) What is the mean waiting time and the mean delay for the high-priority customers?

(b) Show that the time required to service all customers found by a type 2 arrival to the system is

$$\frac{R_2}{1 - \rho_1 - \rho_2},$$

where $\rho_j = \lambda_j E[\tau_j]$, and

$$R_2 = \frac{1}{2}\sum_{j=1}^{2} \lambda_j E[\tau_j^2].$$

(c) Show that the time required to service all type 1 customers who arrive during the time a type 2 customer spends in the system is $\rho_1 E[T_2]$.

(d) Use parts b and c to show that

$$E[T_2] = \frac{(1 - \rho_1 - \rho_2)/\mu_2 + R_2}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}.$$

12.53. Evaluate and plot the formulas developed in Problem 12.52 using the two traffic classes described in Problem 12.51.

Section 12.7: M/G/1 Analysis Using Embedded Markov Chain

12.54. The service time in an M/G/1 system has a $k = 2$ Erlang distribution with mean $1/\mu$ and $\lambda = \mu/2$.

(a) Find $G_N(z)$ and $P[N = j]$.

(b) Find $\hat{W}(s)$ and $\hat{T}(s)$ and the corresponding pdf's.

12.55. (a) In Problem 12.47, show that the Laplace transform of the pdf for the total time τ required to complete the service of a customer is

$$\hat{\tau}(s) = \frac{\mu(s + \beta)}{(s + \beta)(s + \mu) + \alpha s}.$$

Hint: Use conditional expectation in evaluating $E[e^{-s\tau}]$, and note that the number of breakdowns depends on the service time of the customer.

(b) Find $\hat{W}(s)$ and $\hat{T}(s)$ and the corresponding pdf's.

12.56. (a) Show that Eqs. (12.110a) and (12.110b) can be written as

$$N_j = N_{j-1} - U(N_{j-1}) + M_j, \tag{12.186}$$

where

$$U(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0. \end{cases}$$

(b) Take the expected value of both sides of Eq. (12.186) to obtain an expression for $P[N > 0]$.

(c) Square both sides of Eq. (12.186) and take the expected value to obtain the Pollaczek–Khinchin formula for $E[N]$.

12.57. (a) Show that for an M/D/1 system,

$$G_N(z) = \frac{(1 - \rho)(1 - z)}{1 - ze^{\rho(1-z)}}.$$

(b) Expand the denominator in a geometric series, and then identify the coefficient of z^k to obtain

$$P[N = k] = (1 - \rho)\sum_{j=0}^{k} \frac{(-j\rho)^{k-j-1}(-j\rho - k + 1)e^{j\rho}}{(k - j)!}.$$

12.58. (a) Show that Eq. (12.130) can be rewritten as

$$\hat{W}(s) = \frac{1 - \rho}{1 - \rho\hat{R}(s)}, \tag{12.87}$$

where

$$\hat{R}(s) = \frac{1 - \hat{\tau}(s)}{sE[\tau]}$$

is the Laplace transform of the pdf of the residual service time.

(b) Expand the denominator of Eq. (12.187) in a geometric series and invert the resulting transform expression to show that

$$f_W(x) = \sum_{k=0}^{\infty} (1 - \rho)\rho^k f^{(k)}(x), \tag{12.188}$$

where $f^{(k)}(x)$ is the kth-order convolution of the residual service time.

12.59. Approximate $f_W(x)$ for an M/D/1 system using the $k = 0, 1, 2$ terms of Eq. (12.188). Sketch the resulting pdf for $\rho = 1/2$.

Section 12.8: Burke's Theorem: Departures from M/M/*c* Systems

12.60. Consider the interdeparture times from a stable M/M/1 system in steady state.

(a) Show that if a departure leaves the system nonempty, then the time to the next departure is an exponential random variable with mean $1/\mu$.

(b) Show that if a departure leaves the system empty, then the time to the next departure is the sum of two independent exponential random variables of means $1/\lambda$ and $1/\mu$, respectively.

(c) Combine the results of parts a and b to show that the interdeparture times are exponential random variables with mean $1/\lambda$.

12.61. Find the joint pmf for the number of customers in the queues in the network shown in Fig. P12.3.

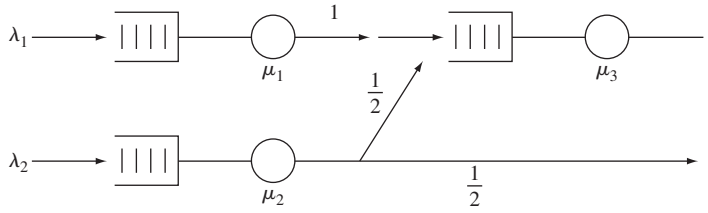

FIGURE P12.3

12.62. Write the balance equations for the feedforward network shown in Fig. P12.4 and verify that the joint state pmf is of product form.

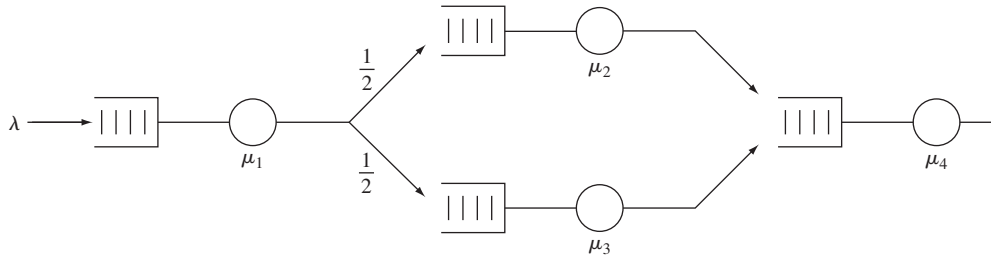

FIGURE P12.4

12.63. Verify that Eqs. (12.137) through (12.139) satisfy Eq. (12.135).

Section 12.9: Networks of Queues: Jackson's Theorem

12.64. Find the joint state pmf for the open network of queues shown in Fig. P12.5.

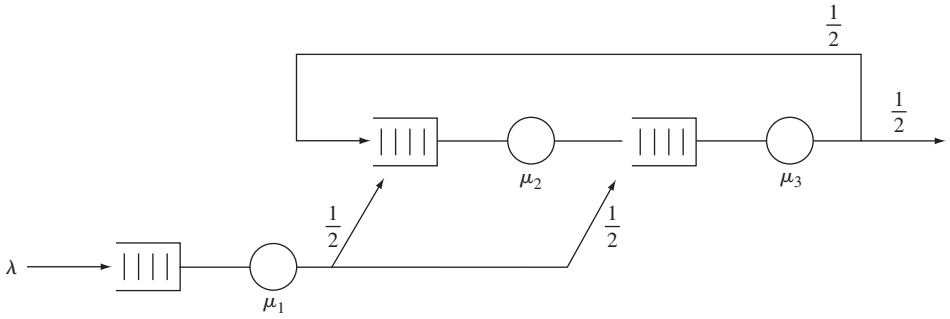

FIGURE P12.5

12.65. A computer system model has three programs circulating in the network of queues shown in Fig. P12.6.

(a) Find the joint state pmf of the system.

(b) Find the average program completion rate.

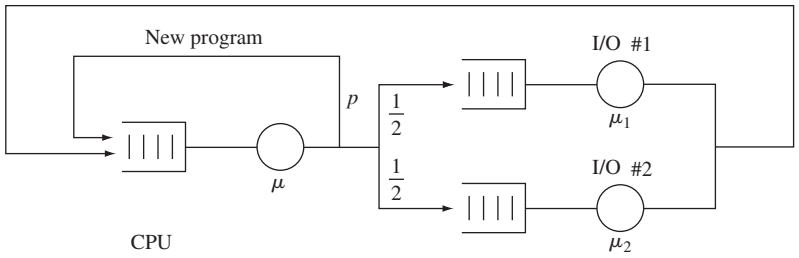

FIGURE P12.6

12.66. Use the mean value analysis algorithm to answer Problem 12.65, part b.

Section 12.10: Simulation and Data Analysis of Queueing Systems

12.67. **(a)** Repeat the experiment in Example 12.28 for an M/M/1 system with $\rho = 0.5, 0.7$, and 0.9. Use sample means for $N(t)$ based on 25 replications to characterize the transient behavior. Try out smoothing the sample means using a moving average filter over time. Give an estimate of the time to reach steady state in each of these systems.

(b) Now investigate the effect of initial condition on the duration of the transient phase. For each of the utilizations above compare the transient duration when the initial condition is: $N(0) = 0$; $N(0) = 5$; $N(0) = 10$.

12.68. For the experiment in Problem 12.67, calculate the sample covariance for each realization and then average over the 25 replications. Find the number of lags required for each value of r until the correlation drops to zero. Comment on the implications for the size of the batches if a method of batch means approach is to be used.

12.69. The correlation of $N(t)$ for an M/M/1 system has the following geometric upper bound [Fishman]:

$$\rho_j \leq \left[\frac{4\rho}{(1+\rho)^2}\right]^j \quad \text{for} \quad j = 0, 1, 2, \ldots.$$

Evaluate the ratio of the variance of the sample mean estimator for this process to that of an iid process when $\rho = 0.5, 0.75, 0.9, 0.99$.

12.70. Run the simulation for the experiment in Example 12.29 50 times. For each simulation produce a confidence interval using the method of batch means. Determine the fraction of the confidence intervals that covered the actual mean $E[N]$. Comment on the accuracy of the confidence intervals given by Eq. (12.168).

12.71. Develop a simulation model for an M/M/3 system with $\lambda = 2$ customers per second and $\mu = 1$ customer per second. Use the method of batch means as in Example 12.29 to estimate the probability that an arriving customer has to queue for service. Provide appropriate confidence intervals.

12.72. **(a)** Consider the simulation in Example 12.30 where the embedded Markov chain approach is used to estimate the steady state pmf. For $\rho = 0.5$ and $\rho = 0.9$, use different warm-up periods to investigate the effect of the initial transient on the pmf estimates.

(b) Double the number of replications and observe the impact on the confidence intervals.

12.73. Develop a simulation for an M/D/1 system with $\rho = 0.7$ using the embedded Markov chain in Eq. (12.172). Design the simulation to estimate the pmf for the number of customers in the system as well as the mean number in the system.

(a) Discuss what transient effects can be expected in this approach.

(b) Use the method of batch means to develop estimates for the mean number of customers in the system. Discuss the choice of batch size and warm-up period. Evaluate the confidence intervals produced by several realizations.

12.74. Use Lindley's recursion to estimate the waiting-time distribution for customers in an M/D/1 system with $\rho = 0.5$ and $\rho = 0.7$. Is there anything peculiar about the distribution?

12.75. Use Lindley's recursion to estimate the waiting-time distribution for customers in a D/M/1 system with $\rho = 0.5$ and $\rho = 0.7$.

12.76. Use Lindley's recursion to estimate the waiting-time distribution for customers in an M/G/1 system with $\rho = 0.5$ and $\rho = 0.7$ where the service-time distribution is Pareto with parameter $\alpha = 2.5$. Try a simulation with $\alpha = 1.5$. Does anything peculiar happen?

12.77. Repeat the experiment in Example 12.33, but use the method of batch means to provide confidence intervals for the mean waiting time.

12.78. Explain why the estimator in Eq. (12.183) will converge to the expected value of the waiting time.

12.79. Use the regenerative method to estimate the mean number in the system and the probability that the system is empty in an M/D/1 system. Evaluate the confidence interval provided by Eq. (12.185).

Problems Requiring Cumulative Knowledge

12.80. Consider an M/M/2/2 system in which one server is twice as fast as the other server.

(a) What definition of "state" of the system results in a continuous-time Markov chain?

(b) Find the steady state pmf for the system if customers arriving at an empty system are always routed to the faster server.

(c) Find the steady state pmf for the system if customers arriving at an empty system are equally likely to be routed to either server.

12.81. **(a)** Find the transient pmf, $P[N(t) = j]$, for an M/M/1/2 system which is in the empty state at time 0.

(b) Repeat part a if the system is full at time 0.

12.82. **(a)** In an M/G/1 system, why are the set of times when customers arrive to an empty system renewal instants?

(b) How would you apply the results from renewal theory in Section 7.5 to estimate the pmf for the number of customers in the system?

(c) How would you obtain a confidence interval for $P[N(t) = j]$?

12.83. Let $N(t)$ be a Poisson random process with parameter λ. Suppose that each time an event occurs, a coin is flipped and the outcome is recorded. Assume that the probability of heads depends on the time of the arrival and is denoted by $p(t)$. Let $N_1(t)$ and $N_2(t)$ denote the number of heads and tails recorded up to time t, respectively.

(a) Show that $N_1(t)$ and $N_2(t)$ are independent Poisson random variables with rates $p\lambda$ and $(1 - p)\lambda$, where

$$p = \frac{1}{t}\int_0^t p(t')\,dt'.$$

(b) Are $N_1(t)$ and $N_2(t)$ independent Poisson random processes? If so, how would you show this?

12.84. Consider an M/G/∞ system in which customers arrive at rate λ and in which the customer service times have distribution $F_X(x)$. Suppose that the system is empty at time 0. Let $N_1(t)$ be the number of customers who have completed their service by time t, and let $N_2(t)$ be the number of customers still in the system at time t.

(a) Use the result of Problem 12.83 to find the joint pmf of $N_1(t)$ and $N_2(t)$.

(b) What is the steady state pmf for the number of customers in an M/G/∞ system?

(c) Apply Little's formula to compute the average number of customers in the system. Is the result consistent with your result in part b?