Chapter 5

Imbedded Markov Chain Models

In the last chapter we used Markov process models for queueing systems with Poisson arrivals and exponential service times. To model a system as a Markov process, we should be able to give complete distribution characteristics of the process beyond time $t$, using what we know about the process at $t$ and changes that may occur after $t$, without referring back the events before $t$. When arrivals are Poisson and service times are exponential, because of the memoryless property of the exponential distribution we are able to use the Markov process as a model. If the arrival rate is $\lambda$ and service rate is $\mu$, at any time point $t$, time to next arrival has the exponential distribution with rate $\lambda$, and if a service is in progress, the remaining service time has the exponential distribution with rate $\mu$. If one or both of the arrival and service distributions are non-exponential, the memoryless property does not hold and a Markov model of the type discussed in the last chapter does not work. In this chapter we discuss a method by which a Markov model can be constructed, not for all $t$, but for specific time points on the time axis.
5.1 Imbedded Markov chains

In an M/G/1 queueing system customers arrive in a Poisson process and get served by a single server. We assume that service times of customers are independent and identically distributed with an unspecified general distribution. Let \( Q(t) \) be the number of customers in the system at time \( t \). For the complete description of the state of the system at time, we need the value of \( Q(t) \) as well as information on the remaining service time of the customer in service, if there is one being served at that time. Let \( R(t) \) be the remaining service time of such a customer. Now the vector \([Q(t), R(t)]\) is a vector Markov process since both of its components, viz, arrival and service times, are now completely specified. The earliest investigation to analyze this vector process by itself was by Cox (1955), who used information on \( R(t) \) as a supplementary variable in constructing the forward Kolmogorov equations developed in Chapter 3. Since it employs analysis techniques beyond the scope set for this text we shall not cover it here.

In two papers in the 1950’s Kendall (1951, 1953) developed a procedure to convert the queue length processes in M/G/1 and G/M/s into Markov chains. (In the queue G/M/s, the service time has the memoryless property. Therefore in the vector process \([Q(t), R(t)]\), \( R(t) \) now represents the time until a new arrival.) The strategy is to consider departure epochs in the queue M/G/1 and arrival epochs in the queue G/M/s. Let \( t_0 = 0, t_1, t_2, \ldots \) be the points of departure of customers in the M/G/1 queue and define \( Q(t_n + 0) = Q_n \). Thus \( Q_n \) is defined as the value of \( Q(t) \) soon after departure. At the points \( \{t_n, n = 0, 1, 2, \ldots\} \), \( R(t) \) is equal to zero and hence \( Q_n \) can be studied without reference to the random variable \( R(t) \). Because of Markov property of the Poisson distribution the process \( \{Q_n, n = 0, 1, 2, \ldots\} \) is a Markov chain with discrete parameter and state spaces. Because of the imbedded nature of the process it is known as an imbedded Markov chain. In the queue G/M/s, arrival
points generate the imbedded Markov chain. We discuss these two systems in the next two sections.

Imbedded Markov chains can also be used to analyze waiting times in the queue G/G/1. A limited exploration of that technique will be given in Chapter 7.

5.2 The queue M/G/1

Let customers arrive in a Poisson process with parameter $\lambda$ and get served by a single server. Let the service times of these customers be independent and identically distributed random variables $\{S_n, \ n = 1, 2, 3, \ldots\}$ with $P(S_n \leq x) = B(x)$, $x \geq 0$; $E(S_n) = b$; $V(S_n) = \sigma_s^2$. We assume that $S_n$ is the service time of the $n$th customer. Let $Q(t)$ be the number of customers in the system at time $t$ and identify $t_0 = 0, t_1, t_2, \ldots$ as the departure epochs of customers. As described above, at these points the remaining service times of customers are zero. Let $Q_n = Q(t_n + 0)$ be the number of customers in the system soon after the $n$th departure. We can show that $\{Q_n, \ n = 0, 1, 2, \ldots\}$ is a Markov chain as follows.

Let $X_n$ be the number of customers arriving during $S_n$. With the Poisson assumption for the arrival process we have

$$k_j = P(X_n = j) = \int_0^\infty P(X_n = j|S_n)P(t < S_n \leq t + dt)$$

$$= \int_0^\infty e^{-\lambda t}(\lambda t)^j/j!dB(t) \quad j = 0, 1, 2, \ldots \quad (5.2.1)$$

In writing $dB(t)$ in (5.2.1) we use the Stieltjes notation in order to accommodate discrete, continuous and mixed distributions.

Now consider the relationship between $Q_n$ and $Q_{n+1}$. We have

$$Q_{n+1} = \begin{cases} 
Q_n + X_{n+1} - 1 & \text{if } Q_n > 0 \\
X_{n+1} & \text{if } Q_n = 0
\end{cases} \quad (5.2.2)$$
The first expression for $Q_{n+1}$ is obvious. The second expression (i.e., $X_{n+1}$ if $Q_n = 0$) results from the fact that $T_{n+1}$ is the departure point of the customer who arrives after $t_n$. It is in fact $1 - 1 + X_{n+1}$.

As can be seen from (5.2.2), $Q_n + 1$ can be expressed in terms of $Q_n$ and a random variable $X_{n+1}$, which does not depend on any event before $t_n$. Since it is i.i.d., it does not depend on $Q_n$ either. The one-step dependence of a Markov chain holds. Hence \( \{Q_n, \ n = 0,1,2,\ldots\} \) is a Markov chain. Its parameter space is made up of departure points, and the state space $S$ is the number of customers in the system; $S = \{0,1,2,\ldots\}$. Because of the imbedded nature of the parameter space, it is known as an *imbedded Markov chain*.

Let

\[
P_{ij}^{(n)} = P(Q_n = j|Q_0 = i), \quad i, j \in S.
\]

and write $P_{ij}^{(1)} \equiv P_{ij}$.

From the relationship (5.2.2) and the definition of $k_j$ in (5.2.1) we can write

\[
P_{ij} = P(Q_{n+1} = j|Q_n = i)
\]

\[
= \begin{cases} 
P(i + X_{n+1} - 1 = j) & \text{if } i > 0 \\
P(X_{n+1} = j) & \text{if } i = 0
\end{cases}
\]

\[
= \begin{cases} 
k_{j-i+1} & \text{if } i > 0 \\
k_j & \text{if } i = 0
\end{cases}
\]

(5.2.4)

The transition probability matrix $P$ for the Markov chain is

\[
P = \begin{pmatrix}
0 & 1 & 2 & \ldots \\
0 & k_0 & k_1 & k_2 & \ldots \\
1 & k_0 & k_1 & k_2 & \ldots \\
3 & \vdots & \ddots & \ddots
\end{pmatrix}
\]

(5.2.5)

For the Markov Chain to be irreducible (the state space has a single equivalence class) the following two conditions must hold: $k_0 > 0$ and $k_0 + k_1 < 1$. It is
5.2. THE QUEUE M/G/1

easy to see that if \( k_0 = 0 \), with one or more customer arrivals for each departure, there is no possibility for the system to attain stability and the number in the system will only increase with time. If \( k_0 + k_1 = 1 \), the only two states \( \{0, 1\} \) are possible in the system. (If the system starts with \( i > 1 \) customers, once it gets 0 or 1 it will remain in \( \{0, 1\} \).

Further classification of states depends on \( E(X_n) \), the expected number of customers arriving during a service time.

Define the Laplace-Stieltjes transform of the service time distribution

\[
\psi(\theta) = \int_0^\infty e^{-\theta t} dB(t) \quad \Re(\theta) > 0
\]  
(5.2.6)

and the probability generating function of the number of customers arriving during a service time

\[
K(z) = \sum_{j=0}^{\infty} k_j z^j \quad |z| \leq 1
\]  
(5.2.7)

The following results follow from well known properties of Laplace-Stieltjes transforms and probability generating functions.

\[
E(S_n) = b = -\psi'(0)
\]
\[
E(S_n^2) = \psi''(0)
\]
\[
E(X_n) = K'(1)
\]
\[
E(X_n^2) = K''(1) + K'(1)
\]  
(5.2.8)

From (5.2.1) we get

\[
K(z) = \int_0^\infty e^{-\lambda t} \sum_{j=0}^{\infty} \frac{(\lambda z)^j}{j!} dB(t)
\]
\[
= \int_0^\infty e^{-(\lambda - \lambda z)t} dB(t)
\]
\[
= \psi(\lambda - \lambda z)
\]

Hence

\[
K'(z) = -\lambda \psi'(\lambda - \lambda z)
\]
\[ K'(1) = -\lambda \psi'(0) \]
\[ = \lambda b = \rho \quad (5.2.9) \]

Note that \( \lambda b = \text{(arrival rate)} \times \text{(mean service times)} \). This quantity is called the 
traffic intensity of the queueing system denoted by \( \rho \). The value of \( \rho \) determines 
whether the system is in equilibrium (attains steady state) when the time parameter \( n \) (of \( t_n \)) \( \to \infty \). It can be shown that when \( \rho < 1 \), the Markov chain 
is positive recurrent (i.e. the process returns to any state with probability one 
and the mean time for the return \( < \infty \)); when \( \rho = 1 \), the chain is null recurrent 
(i.e., the process returns to any state with probability one, but the mean time 
for the return \( = \infty \)); and when \( \rho > 1 \), the chain is transient (i.e., the process 
may not return to the finite states at all. Then the probability that the process 
will be found in one of the finite states is zero.) These derivations are beyond 
the scope of this text. Nevertheless, these properties are easy to comprehend if 
we understand the real significance of the value of the traffic intensity.

Recalling the result derived in (3.4.13), the \( n \)-step transition probabilities 
\( P_{ij}^{(n)} \) \( (i, j = 0, 1, 2, \ldots) \) of the Markov chain \( \{Q_n\} \) are obtained as elements of 
the \( n \)th power of the (one-step) transition probability matrix \( P \). In considering 
\( P^n \) in real systems the following three observations will be useful.

1. The result (3.4.13) holds regardless of the structure of the matrix.

2. As \( n \) in \( P^n \) increases, the non-zero elements cluster within submatrices 
representing recurrent equivalence classes.

3. In an aperiodic irreducible positive recurrent Markov chain as \( n \) in \( P^n \) 
increases the elements in each column tend toward an intermediate value.

The probability \( P_{ij}^{(n)} \) for \( j = 0, 1, 2 \ldots \) and finite \( n \) gives the time dependent 
behavior of the queue length process \( \{Q_n\} \). There are analytical techniques for 
deriving these probabilities. However they involve mathematical techniques beyond 
the scope of this text. For example see Takács (1962), who uses probability
generating functions to simplify recursive relations generated by the Chapman-Kolmogorov relations for $P_{ij}^{(n)}$. Prabhu and Bhat (1963) look at the transitions of $Q_n$ as some first passage problems and use combinatorial methods in solving them. (Also see Prabhu (1965)). In practice, however, with the increasing computer power for matrix operations, simple multiplications of $P$ to get its $n$th power seem to be the best course of action. When the state space is not finite, the observations given above can be used to limit it without losing significant amount of information.

**Limiting distribution**

The third observation given above stems from the property of aperiodic positive recurrent irreducible Markov chains which results in $\lim_{n \to \infty} P^n$ becoming a matrix with identical rows. Computationally this property can be validated by getting successive powers of $P^n$; as $n$ increases the elements in the columns of the matrix tend to a constant intermediate value. This behavior of the Markov chain is codified in the following theorem and the corollary, given without proof.

**Theorem 5.2.1**

(1) Let $i$ be a state belonging to an aperiodic recurrent equivalence class. Let $P_{ii}^{(n)}$ be the probability of the $n$-step transition $i \to i$, and $\mu_i$ be its mean recurrence time. Then $\lim_{n \to \infty} P_{ii}^{(n)}$ exists and is given by

$$\lim_{n \to \infty} P_{ii}^{(n)} = \frac{1}{\mu_i} = \pi_i, \text{ say.}$$

(2) Let $j$ be another state belonging to the same equivalence class and $P_{ji}^{(n)}$ be the probability of the $n$-step transition $j \to i$. Then

$$\lim_{n \to \infty} P_{ji}^{(n)} = \lim_{n \to \infty} P_{ii}^{(n)} = \pi_i$$

**Corollary 5.2.1** If $i$ is positive recurrent, $\pi_i > 0$ and if $i$ is null recurrent, $\pi_i = 0$. [See Karlin and Taylor (1975) for a proof of this theorem].
Note that we have the term *recurrence time* in the theorem to signify the number of steps a Markov chain takes to return for the first time to the starting state. See Section 3. for other definitions.

Theorem 5.2.1 applies to Markov chains whether their state space is finite or countably infinite.

For a state-space \( S = \{0, 1, 2, \ldots\} \) let \((\pi_0, \pi_1, \pi_2, \ldots)\) be the limiting probability vector where \( \pi_i = \lim_{n \to \infty} P_{ji}^{(n)}, i, j \in S \). Let \( \Pi \) be the matrix with identical rows \( \pi = (\pi_0, \pi_1, \pi_2, \ldots) \). Now, using Chapman-Kolmogorov relations we may write

\[
P^{(n)} = P^{n-1} P
\]

(see discussion leading up to (3.4.13)).

Applying Theorem 5.2.1 to \( P^{(n)} \) and \( P^{n-1} \), it is easy to write

\[
\Pi = \Pi P
\]

or

\[
\pi = \pi P
\quad (5.2.10)
\]

Furthermore, multiplying both sides of (5.2.10) repeatedly by \( P \), we can also establish that

\[
\pi P = \pi = \pi P^2
\]

\[
\pi P^{n-1} = \pi = \pi P^n
\quad (5.2.11)
\]

The last equation shows that if we use the limiting distribution as the intial distribution of the state of an irreducible aperiodic, and positive recurrent Markov chain, the state distribution after \( n \) transitions \((n = 1, 2, 3\ldots)\) is also given the same limiting distribution. Such a property is known as the *stationarity* of the distribution. The following theorem summarizes these results and provides a procedure by which the limiting distribution can be determined.
Theorem 5.2.2 (1) In an irreducible, aperiodic, and positive recurrent Markov chain, the limiting probabilities \( \{ \pi_i, \ i = 0, 1, 2, \ldots \} \) satisfy the equations

\[
\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij} \quad j = 0, 1, 2, \ldots
\]

\[
\sum_{j=0}^{\infty} \pi_j = 1 \quad (5.2.12)
\]

The limiting distribution is stationary.

(2) Any solution of the equations

\[
\sum_{i=0}^{\infty} x_i P_{ij} = x_j \quad j = 0, 1, 2, \ldots \quad (5.2.13)
\]

is a scalar multiple of \( \{ \pi_i, \ i = 0, 1, 2, \ldots \} \) provided \( \sum |x_i| < \infty \).

Thus the limiting distribution of the Markov chain can be obtained by solving the set of simultaneous equations (5.2.12) and normalizing the solution using the second equation \( \sum_{j=0}^{\infty} \pi_j = 1 \). It should be noted that because the two sums of the Markov chain are equal to 1, (5.2.12) by itself yields a solution only up to a multiplicative constant. The normalizing condition is, therefore, essential in the determination of the limiting distribution.

With this background on the general theory of Markov chains, we are now in a position to determine the limiting distribution of the imbedded Markov chain of the M/G/1 queue.

Let \( \boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots) \) be the limiting distribution of the imbedded chain. Using the transition probability matrix (5.2.5) in the equation \( \boldsymbol{\pi} \boldsymbol{P} = \boldsymbol{\pi} \) (which is (5.2.12)), we have

\[
k_0 \pi_0 + k_0 \pi_1 = \pi_0
\]

\[
k_1 \pi_0 + k_1 \pi_1 + k_0 \pi_2 = \pi_1
\]

\[
k_2 \pi_0 + k_2 \pi_1 + k_1 \pi_2 + k_0 \pi_3 = \pi_2
\]

\[
\vdots
\]

\[
(5.2.14)
\]
A convenient way of solving these equations computationally is to define
\[ \nu_0 \equiv 1 \text{ and } \nu_i = \frac{\pi_i}{\pi_0} \]
and rewrite (5.2.14) in terms of \( \nu_i \) \((i = 1, 2, \ldots)\) as

\[
\begin{align*}
\nu_1 &= \frac{1 - k_0}{k_0} \\
\nu_2 &= \frac{1 - k_1}{k_0} \nu_1 - \frac{k_1}{k_0} \\
\vdots \\
\nu_j &= \frac{1 - k_1}{k_0} \nu_{j-1} - \frac{k_2}{k_0} \nu_{j-2} - \cdots - \frac{k_{j-1}}{k_0} \nu_1 - \frac{k_{j-1}}{k_0} \\
\vdots 
\end{align*}
\]

These equations can be solved recursively to determine \( \nu_i \) \((i = 1, 2, \ldots)\). The limiting probabilities \((\pi_0, \pi_1, \pi_2, \ldots)\) are known to be monotonic and concave, and therefore for larger values of \( n \) they become extremely small. Clearly \( \nu_i = \frac{\pi_i}{\pi_0} \) will also have the same properties and for computational purposes it is easy to establish a cutoff value for the size of the state space.

In order to recover \( \pi_i \)'s from \( \nu_i \)'s, we note that

\[
\begin{align*}
\sum_{i=0}^{\infty} \nu_i &= 1 + \sum_{i=1}^{\infty} \frac{\pi_i}{\pi_0} = \sum_{i=0}^{\infty} \frac{\pi_i}{\pi_0} = \frac{1}{\pi_0}
\end{align*}
\]

Here we have incorporated the normalizing condition \( \sum_0^{\infty} \pi_i = 1 \). Thus we get

\[
\pi_0 = (1 + \sum_{i=1}^{\infty} \nu_i)^{-1}
\]

and

\[
\pi_i = \frac{\nu_i}{1 + \sum_{i=1}^{\infty} \nu_i} \tag{5.2.16}
\]

Analytically the limiting distribution \((\pi_0, \pi_1, \pi_2, \ldots)\) can be determined by solving equations (5.2.14) using generating functions. Unfortunately deriving the explicit expressions for the probabilities require inverting the resulting probability generating function. However, we can obtain the mean and variance of
the distribution using standard techniques. Define
\[ \Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j \quad |z| \leq 1 \]
and
\[ K(z) = \sum_{j=0}^{\infty} k_j z^j \quad |z| \leq 1. \]

Multiplying equations (5.2.14) with appropriate powers of \( z \) and summing we get
\[
\Pi(z) = \pi_0 K(z) + \pi_1 K(z) + \pi_2 z K(z) + \ldots
\]
\[
= \pi_0 K(z) + \frac{K(z)}{z}(\pi_1 z + \pi_2 z^2 + \ldots)
\]
\[
= \pi_0 K(z) + \frac{K(z)}{z} [\Pi(z) - \pi_0]
\]

Rearranging terms,
\[
\Pi(z)[1 - \frac{K(z)}{z}] = \pi_0 K(z)[1 - \frac{1}{z}]
\]
\[
\Pi(z) = \frac{\pi_0 K(z)(z - 1)}{z - K(z)} \quad (5.2.17)
\]

The unknown quantity \( \pi_0 \) on the right hand side expression for \( \Pi(z) \) in (5.2.17) can be determined using the normalizing condition \( \sum_{j=0}^{\infty} \pi_j = 0 \). We must have
\[
\Pi(1) = \sum_{j=0}^{\infty} \pi_j = 1
\]

Letting \( z \to 1 \) in (5.2.17) we get (applying L’Hopital’s rule)
\[
1 = \frac{\lim_{z \to 1} \pi_0 [K(z) - (z - 1)K'(z)]}{\lim_{z \to 1} [1 - K'(z)]}
\]

Recalling that \( K(1) = 1 \) and \( K'(1) = \rho \), (from (5.2.9)) we have
\[
1 = \frac{\pi_0}{1 - \rho}
\]
\[
\pi_0 = 1 - \rho \quad \text{(5.2.18)}
\]

Thus we get
\[
\Pi(z) = \frac{(1 - \rho)(z - 1)K(z)}{z - K(z)} \quad (5.2.19)
\]
Explicit expressions for probabilities \( \{\pi_j, j = 0, 1, 2, \ldots\} \) can be obtained by expanding \( \Pi(z) \) in special cases. An alternative form of \( \Pi(z) \) works out to be easier for this expansion. We may write

\[
\Pi(z) = \frac{(1 - \rho)K(z)}{(z - K(z))/(z - 1)} \frac{(1 - \rho)K(z)}{1 - [1 - K(z)]/(1 - z)}
\]

Note that \( \sum_{j=0}^{\infty} z^j(k_{j+1} + k_{j+2} + \ldots) \) can be simplified to write as

\[
\frac{1 - K(z)}{1 - z} = C(z), \quad \text{say.}
\]

[Also see algebraic simplifications leading to (5.3.27)].

For \( |z| \leq 1 \)

\[
|C(z)| = \left| \frac{1 - K(z)}{1 - z} \right| < 1 \quad \text{if } \rho < 1
\]

Now using a geometric series expansion we may write

\[
\Pi(z) = (1 - \rho)K(z) \sum_{j=0}^{\infty} [C(z)]^j
\]

The explicit expression for \( \pi_j \) is obtained by expanding the right hand side of (5.2.22) as a power series in \( z \) and picking the coefficient of \( z^j \) in it.

In a queueing system, the queue length process \( Q(t) \) may be considered with three different time points: (1) when \( t \) is an arrival epoch; (2) when \( t \) is a departure epoch; and (3) when \( t \) is an arbitrary point in time. In general, the distribution of \( Q(t) \) with reference to these three time points may not be the same. However when the arrival process is Poisson, it can be shown that the limiting distributions of \( Q(t) \) in all three cases are the same. The property of the Poisson process that makes this happen is its relationship with the uniform distribution mentioned in Section 3. See Wolff (1982) who coined the acronym PASTA (Poisson Arrivals See Time Averages). For proofs of this property see Cooper (1981) and Gross and Harris (1998).

The probability generating function \( \Pi(z) \) derived in (5.2.19), therefore, also gives the limiting distribution \( \lim_{t \to \infty} Q(t) \). There are several papers in the
literature deriving the transition distribution of $Q(t)$ for finite $t$. Among them are Prabhu and Bhat (1963b) and Bhat (1968) who obtain the transition distribution using recursive methods and renewal theory arguments. The explicit expression for the limiting distribution of $Q(t)$ (and the limiting distribution of $Q_n$ in the imbedded chain case) derived in these papers is given by

$$\pi_0 = 1 - \rho$$

$$\pi_j = (1 - \rho) \int_0^\infty e^{-\lambda t} \sum_{n=0}^\infty \left[ \frac{(\lambda t)^{n+j-1}}{(n+j-1)!} - \frac{(\lambda t)^{n+j}}{(n+j)!} \right] dB_n(t)$$

for $\rho < 1$, where $B_n(t)$ is the $n$-fold convolution of $B(t)$ with itself [Prabhu and Bhat (1963a,b) and Bhat (1968)].

The mean and variance of $\lim_{n \to \infty} Q_n$ can be determined from the probability generating function (5.2.19) through standard techniques. Writing $Q = \lim_{n \to \infty} Q_n$, we have

$$L = E(Q) = \Pi'(1)$$

$$V(Q) = \Pi''(1) + \Pi'(1) - [\Pi'(1)]^2$$

Differentiating $\Pi(z)$ w.r.t. $z$, we get

$$\Pi'(z) = \frac{1}{[z - K(z)]^2} \{[z - K'(z)](1 - \rho)(K'(z) + zK'(z) - K(z)) - [(1 - \rho)(z - 1)K(z)(1 - K'(z))]\}$$

Using L'Hopital's rule twice while taking limits $z \to 1$, we get

$$\Pi'(1) = \frac{2K'(1)K'(1) + K''(1)}{2[1 - K'(1)]}$$

But note from (5.2.9), $K''(1) = \rho$ and

$$K''(1) = \lambda^2 E(S^2)$$

where we have used a generic notation for $S$ for the service time. Substituting from (5.2.26) in (5.2.25) we get, after simplifications,

$$L = E(Q) = \rho + \frac{\lambda^2 E(S^2)}{2(1 - \rho)}$$
CHAPTER 5. IMBEDDED MARKOV CHAIN MODELS

Nothing that $\rho$ is the expected number in service (which is the same as the probability the server is busy in a single server queue), $L_q$, the mean number in the queue is obtained as

$$L_q = \frac{\lambda^2 E(S^2)}{2(1-\rho)} \quad (5.2.28)$$

Extending the differentiation to get $\Pi''(z)$, and taking limits as $z \to 1$ with the multiple use of L’Hoptal’s rule $\Pi''(1)$ can be obtained to get

$$V(Q) = \rho(1-\rho) + \frac{\lambda^2 E(S^2)}{2(1-\rho)} \left[ \frac{3 - 2\rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)}}{3(1-\rho)} \right]$$

Recall that $\sigma^2_S$ is the variance of the service time distribution. Hence $\sigma^2_S = E(S^2) - [E(S)]^2$. Using this expression in (5.2.27) and nothing that $\lambda E(S) = \rho$, we get an alternative form for $E(Q)$.

$$E(Q) = \rho + \frac{\rho^2}{2(1-\rho)} + \frac{\lambda^2 \sigma^2_S}{2(1-\rho)} \quad (5.2.30)$$

which clearly shows that the mean queue length increases with the variance of the service time distribution. For instance, when $\sigma^2_S = 0$, i.e. when the service time is constant (in the queue M/D/1)

$$E(Q) = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{\rho(1 - \frac{\rho}{2})}{1 - \rho} \quad (5.2.31)$$

On the other hand, when the service time distribution is Erlang with mean $\frac{1}{\mu}$ and scale parameter $k$ (i.e. by writing $\lambda = k\mu$ in (2.1.18)), we get $\sigma^2_S = \frac{1}{k^2\mu^2}$ and

$$E(Q) = \rho + \frac{\rho^2}{2(1-\rho)} + \frac{\rho^2}{2k(1-\rho)}$$

$$= \rho + \frac{\rho^2(1+k)}{2k(1-\rho)} \quad (5.2.32)$$

When $k = 1$, we get $E(Q)$ in M/M/1 as

$$E(Q) = \frac{\rho}{1-\rho}.$$
5.2. THE QUEUE \(M/G/1\)

Waiting time

The concept of waiting time has been used earlier in the context of the \(M/M/1\) queue. Since we had the distribution of the queue length explicitly, we were then able to determine the distribution of the waiting time. But in the \(M/G/1\) case the explicit expression for the limiting distribution of the queue length, viz, Eq. (5.2.23), is not easy to handle, even for computations. Consequently we approach this problem indirectly using the probability generating function \(\Pi(z)\) of the queue length.

Assume that the queue discipline is first come, first served. Let \(T\) be the total time spent by the customer in the system in waiting and service which we may call system time or time in system, and \(T_q\) the actual waiting time, both as \(t \to \infty\). Let \(E(T) = W\) and \(E(T_q) = W_q\). Also let \(F(\cdot)\) be the distribution function of \(T\) with a Laplace-Stieltjes transform

\[
\Phi(\theta) = \int_0^\infty e^{-\theta t} dF(t) \quad \mathcal{R}(\theta) > 0.
\]

Consider a customer departing from the system. It has spent a total time of \(T\), in waiting and service. Suppose the departing customer leaves \(n\) customers behind; clearly these customers have arrived during its time in system \(T\). Then we have

\[
P(Q = n) = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF(t) \quad n \geq 0 \quad (5.2.33)
\]

Using generating functions,

\[
\Pi(z) = \sum_{n=0}^{\infty} P(Q = n) z^n = \sum_{n=0}^{\infty} z^n \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF(t)
\]

\[
= \int_0^\infty e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda z)^n}{n!} dF(t)
\]

\[
= \Phi(\lambda - \lambda z) \quad (5.2.34)
\]

Comparing (5.2.19) with (5.2.33) we have

\[
\frac{(1 - \rho)(z - 1)K(z)}{z - K(z)} = \Phi(\lambda - \lambda z) \quad (5.2.35)
\]
Recall that
\[ K(z) = \psi(\lambda - \lambda z) \]

Substituting in (5.2.35),
\[ \Phi(\lambda - \lambda z) = \frac{(1 - \rho)(z - 1)\psi(\lambda - \lambda z)}{z - \psi(\lambda - \lambda z)} \]

Writing \( \lambda - \lambda z = \theta \), we get \( z = 1 - \frac{\theta}{\lambda} \);
\[ \Phi(\theta) = \frac{(1 - \rho)\lambda \psi(\theta)}{\psi(\theta) - (\lambda - \theta)/\lambda} \]
\[ = \frac{(1 - \rho)\theta \psi(\theta)}{\theta - \lambda[1 - \psi(\theta)]} \]  \hspace{1cm} (5.2.36)

Since the system time \( T \) is the sum of the actual waiting time \( T_q \) and service time \( S \), defining the Laplace-Stieltjes transform of the distribution of \( T_q \) as \( \Phi_q(\theta) \), we have
\[ \Phi(\theta) = \Phi_q(\theta)\psi(\theta) \] \hspace{1cm} (5.2.37)

Comparing (5.2.36) and (5.2.37) we write
\[ \Phi_q(\theta) = \frac{(1 - \rho)\theta}{\theta - \lambda[1 - \psi(\theta)]} \] \hspace{1cm} (5.2.38)

which can be expressed as
\[ \Phi_q(\theta) = \frac{1 - \rho}{1 - \frac{\lambda}{\theta}[1 - \psi(\theta)]} \]
\[ = (1 - \rho) \sum_{n=0}^{\infty} \left[ \frac{\lambda}{\theta} [1 - \psi(\theta)] \right]^n \] \hspace{1cm} (5.2.39)

In using the geometric series for (5.2.39), we can show that \( |\frac{\lambda}{\theta}[1 - \psi(\theta)]| < 1 \) for \( \rho < 1 \).

In Chapter 3 we have introduced a renewal process as a sequence of independent and identically distributed random variables. Suppose \( t_{n+1} - t_n = Z_n \) is the \( n \)th member of such a sequence. Let \( t \) be a time point such that \( t_n < t \leq t_{n+1} \). Then \( t_{n+1} - t = R(t) \) is known as the forward recurrence time (also known as
5.2. THE QUEUE M/G/1

excess life in the terminology of Reliability Theory). If \( B(\cdot) \) is the distribution function of \( Z_n \), it is possible to show, as \( t \to \infty \), \( r_t(s) \), the density function of \( R(t) \), can be given as

\[
\lim_{t \to \infty} r_t(x) = \frac{1}{E[Z_n]}[1 - B(x)]
\]

(5.2.40)

Using this concept, we can invert (5.2.39) to give the distribution function of \( T_q \) as

\[
F_q(t) = (1 - \rho) \sum_{n=0}^{\infty} \rho^n R^{(n)}(t)
\]

(5.2.41)

where \( R^{(n)}(t) \) is the \( n \)-fold convolution of the remaining service distribution \( R(t) \) (forward recurrence time) with itself.

As stated in Chapter 4, Little’s law \( (L = \lambda W) \) applies broadly to queueing systems with only some restrictions on structure and discipline. Hence using the law on (5.2.27) and (5.2.28) we get

\[
W = \frac{1}{\mu} + \frac{\lambda E(S^2)}{2(1 - \rho)}
\]

(5.2.42)

\[
W_q = \frac{\lambda E(S^2)}{2(1 - \rho)}
\]

(5.2.43)

These means can also be determined from the transforms \( \Phi(\theta) \) and \( \Phi_q(\theta) \). For, we have

\[
W = E(T) = \Phi'(0)
\]

\[
\sigma_T^2 = V(T) = \Phi''(0) - [\Phi'(0)]^2
\]

and similar expressions for \( W_q \) and \( \sigma_{T_q}^2 \). The following result, derived in this manner, might be useful in some applications.

\[
\sigma_{T_q}^2 = V(T_q) = \frac{\lambda E(S^3)}{3(1 - \rho)} + \frac{\lambda^2 E(S^2)^2}{4(1 - \rho)^2}
\]

(5.2.44)

The busy period

In the context of an imbedded Markov chain, the length of the busy period is measured in terms of the number of transitions of the chain without visiting
the state 0. Let $B_i$ be the number of transitions of the Markov chain before it enters state 0 for the first time, having initially started from state $i$. Let

$$g_i^{(n)} = P[B_i = n] \quad n = 1, 2, \ldots$$  \hfill (5.2.45)

A key property of $B_i$ is that it can be thought of as the sum of $i$ random variables each with the distribution of $B_1$. This is equivalent to saying that the transition $i \to 0$ can be considered to be occurring in $i$ segments, $i \to i-1$, $i-1 \to i-2, \ldots, 1 \to 0$. This is justified by the fact that the downward transition can occur only one step at a time. Since all these transitions are structurally similar to each other we can consider $B_i$ as the sum of $i$ random variables each with the distribution of $B_1$. Consequently $g_i^{(n)}$ is the $i$-fold convolution of $g_1^{(n)}$ with itself. Thus for the probability generating function of $g_i^{(n)}$

$$G_i(z) = \sum_{n=1}^{\infty} g_i^{(n)} z^n = [G_1(z)]^i$$ \hfill (5.2.46)

Noting that the busy period cannot end before the $i$th transition of the Markov chain, we have

$$g_i^{(i)} = k_0^{(i)}$$

$$g_i^{(n)} = \sum_{r=1}^{n-1} k_r^{(i)} g_r^{(n-i)}$$ \hfill (5.2.47)

where $k_r^{(i)}$ is the $i$-fold convolution of the probability $k_r$ that $r$ customers arrive during a service period. [See (5.2.1)].

For $i = 1$

$$g_1^{(1)} = k_0$$

and

$$g_1^{(n)} = \sum_{r=1}^{n-1} k_r g_r^{(n-1)} \quad n \geq 1$$

Multiplying by appropriate powers of $z$ to both sides of these equations we convert them into a single one in generating functions

$$g_1^{(1)} z = k_0 z$$
\[
\sum_{n=2}^{\infty} g_1^{(n)} z^n = \sum_{n=2}^{\infty} z^n \sum_{r=1}^{n-1} g_r^{(n-1)}
\]

\[G_1(z) \equiv z k_0 + z \sum_{r=1}^{\infty} k_r \sum_{n=r+1}^{\infty} z^{n-1} g_r^{(n-1)}\]

\[= z \left[ k_0 + \sum_{r=1}^{\infty} k_r [G_1(z)]^r \right] = z K[G_1(z)] \tag{5.2.48}\]

From the definition of \(K(z)\) earlier, we have

\[K(z) = \psi(\lambda - \lambda z)\]

where \(\psi(\theta)\) is Laplace-Stieltjes transform of the service time distribution. Thus the p.g.f. \(G_1(z) \equiv G(z)\) is such that it satisfies the functional equation

\[\omega = z \psi(\lambda - \lambda \omega) \tag{5.2.49}\]

It is possible to show that \(G(z)\) is the least positive root (\(\leq 1\)) of the functional equation when \(\rho < 1\) and determine explicit expressions for specific distributions. There are other ways of deriving the busy period distribution in explicit forms [see Prabhu and Bhat (1963)]. Nevertheless we can easily obtain the mean length of the busy period by implicit differentiation of (5.2.49). We have

\[G(z) = z \psi(\lambda - \lambda G(z))\]

On differentiation,

\[G'(z) = \psi(\lambda - \lambda G(z)) + z \psi'(\lambda - \lambda G(z))[-\lambda G'(z)]\]

As \(z \to 1\),

\[G'(1) = \psi(0) + \psi'(0)[-\lambda G'(1)]\]

Rearranging terms

\[G'(1)[1 + \lambda \psi'(0)] = 1\]

\[G'(1) = \frac{1}{1 + \lambda \psi'(0)}\]
Referring back to the definitions given earlier
\[ E[B_1] = G'(1) = \frac{1}{1 - \rho}, \] (5.2.50)

Since we are counting the number of transitions, to get the exact mean length of a busy period we multiply by the mean length of time taken for each transition, viz, the service period. Hence
\[ \text{Mean length of the busy period} = \frac{E(S)}{(1 - \rho)} \] (5.2.51)

The queue M/G/1/K

Consider the M/G/1 queue described earlier, with the restriction that the capacity for the number of customers in the system is \( K \). Since the state space for the imbedded Markov chain is the number in the system soon after departure, \( K \) will not be included in the state space. \( S = \{0, 1, 2, \ldots, K - 1\} \). Thus corresponding to (5.2.2) we have the relation
\[ Q_{n+1} = \begin{cases} 
\min (Q_n + X_{n+1} - 1, K - 1) & \text{if } Q_n > 0 \\
\min (X_{n+1}, K - 1) & \text{if } Q_n = 0 
\end{cases} \] (5.2.52)

Using the probability distribution \( \{k_j, 0, 1, 2, \ldots\} \) defined in (5.2.1), we get the transition probability matrix

\[
P = \begin{bmatrix}
0 & 1 & 2 & \cdots & K - 1 \\
0 & k_0 & k_1 & k_2 & \cdots & 1 - \sum_{j=0}^{k-2} k_j \\
1 & k_0 & k_1 & k_2 & \cdots & 1 - \sum_{j=0}^{k-2} k_j \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
K - 1 & k_0 & k_1 & \cdots & 1 - \sum_{j=0}^{k-3} k_j 
\end{bmatrix}
\] (5.2.53)

Let \( \pi = (\pi_0, \pi_1, \ldots, \pi_{k-1}) \) be the limiting distribution for the state of the Markov chain. These probabilities are determined by solving the equations
\[ \pi_j = \sum_i \pi_i P_{ij} \quad j = 0, 1, 2, \ldots, K - 1 \]
5.2. THE QUEUE M/G/1

\[ \sum_{j=0}^{K-1} \pi_j = 1 \]  \hspace{1cm} (5.2.54)

The first \( K - 1 \) equations are identical to those for M/G/1 with no capacity restriction. Therefore we can use the computational method outlined in (5.2.15) for the solution of (5.2.54). We may note here that one of the \( k \) simultaneous equations in (5.2.54) is redundant because of the Markov chain structure of the coefficients. In its place we use the normalizing condition \( \sum_{j=0}^{K-1} \pi_j = 1 \) for the solution. We may also note from the computational solution technique that the finite case solution is obtained by using the same \( \nu_i \)'s as in the infinite case for \( i = 0, 1, 2, \ldots, K - 1 \) and determining

\[
\pi_0 = \left[ \sum_{i=0}^{K-1} \nu_i \right]^{-1} \\
\pi_i = \pi_0 \nu_i \]  \hspace{1cm} (5.2.55)

The discussion of the waiting time distribution is a bit complicated in M/G/1/K, since for the Markov chain the state space is only \( \{0, 1, 2, \ldots, K - 1\} \) while our arrival may find \( K \) customers in the system (before a departure). Thus from the viewpoint of an arrival, we need the limiting distribution at an arbitrary point in time. For further details, the readers are referred to Gross and Harris (1998), p. 231-232.

The busy period analysis given earlier for the queue M/G/1 cannot be easily modified for the finite capacity case. Computationally, the best approach seems to be to consider the busy period as a first passage problem in the irreducible Markov chain from state 1 to state 0. This can be done by converting state 0 into an absorbing state and using the concept of the Fundamental Matrix in the determination of the expected number of transitions required for the first passage transition. For details, the readers are referred to Bhat and Miller (2002) Ch. 2. A further discussion of this method is also given in Chapter 8.

Example 5.2.1
Consider a computer network node in which requests for data arrive in a Poisson process at the rate of 0.5 per unit time. Assume that the data retrieval (service) takes a constant amount of one unit of time.

We can model this system as an M/D/1 queue and use the techniques developed in this section for its analysis. We have

\[ k_j = e^{-0.5 \frac{(0.5)^j}{j!}} \quad j = 0, 1, 2, \ldots \]

which on an evaluation gives

\[ k_0 = 0.607; \quad k_1 = 0.303; \quad k_2 = 0.076; \quad k_3 = 0.012; \quad k_4 = 0.002 \]

The Laplace-Stieltjes transform of the service time distribution

\[ \psi(\theta) = e^{-(0.5)\theta} \]

and the probability generating function of \( k_j, j = 0, 1, 2 \ldots \)

\[ K(z) = e^{-0.5(1-z)} \]

These give, the p.g.f. \( \Pi(z) \) of the limiting distribution as

\[ \Pi(z) = \frac{(1 - 0.5)(z - 1)e^{-0.5(1-z)}}{z - e^{-0.5(1-z)}} \]

Instead of using \( \Pi(z) \) to obtain the distribution explicitly, we use the computational method described earlier. We have

\[ \nu_1 = \frac{1 - k_0}{k_0} = 0.647 \]
\[ \nu_2 = \frac{1 - k_1}{k_0} \nu_1 - \frac{k_1}{k_0} = 0.244 \]
\[ \nu_3 = 0.074 \]
\[ \nu_4 = 0.022 \]
\[ \nu_5 = 0.006 \]
\[ \nu_6 = 0.002 \]
\[ \nu_7 = 0.001 \]
5.2. THE QUEUE M/G/1

\[ \sum_{i=0}^{7} \nu_i = 1.996 \]

\[ \pi_0 = \left( \sum_{i=0}^{7} \nu_i \right)^{-1} = 0.501 \]

\[ \pi_i = \nu_i \pi_0 \quad i = 1, 2, \ldots, 7 \]

Thus we get

\[ \pi_0 = 0.501; \quad \pi_1 = 0.324; \quad \pi_2 = 0.122; \quad \pi_3 = 0.037; \]

\[ \pi_4 = 0.011; \quad \pi_5 = 0.003; \quad \pi_6 = 0.001; \quad \pi_7 = 0.000 \]

The mean number of customers in the system as \( t \to \infty \) can be determined either from the formula (5.2.31) or from the distribution \( \pi \) determined above.

We get

\[ L = E(Q) = 0.75 \]

Using Little’s law \( L = \lambda W \), the mean system time can be obtained as

\[ W = 0.75 / 0.5 = 1.5 \text{ time units} \]

Also for the length of a busy period \( B \), we have

\[ E(B_1) = \frac{1}{1 - 0.5} = 2 \text{ time units} \]

\( \text{Ans.} \)

Example 5.2.2

In an automobile garage with a single mechanic, from the records kept by the owner, the distribution of the number of vehicles arriving during the service time of a vehicle is obtained as follows:

\[ P \text{ (no new arrivals)} = 0.5 \]

\[ P \text{ (one new arrival)} = 0.3 \]

\[ P \text{ (two new arrivals)} = 0.2 \]
If we assume that the arrival of vehicles for service follow a Poisson distribution, we can model this system as an M/G/1 queue, even when we do not have a distribution form for the service times. With this assumption, we get

\[ k_0 = 0.5; \quad k_1 = 0.3; \quad k_2 = 0.2 \]

with \( E (\# \text{ of arrivals during one service period}) = 0.7 = \text{traffic intensity } \rho \). The computational method for the determination of the limiting distribution is the most appropriate since no distribution form is available for the service time.

Using equations (5.2.15) we get

\[ \nu_0 = 1; \quad \nu_1 = 1; \quad \nu_2 = 0.8; \quad \nu_3 = 0.32; \]
\[ \nu_4 = 0.128; \quad \nu_5 = 0.051; \quad \nu_6 = 0.021 \quad \nu_7 = 0.008; \]
\[ \nu_8 = 0.003; \quad \nu_9 = 0.001; \quad \nu_{10} = 0.001 \]

Hence \( \sum_{i=0}^{10} \nu_i = 3.333 \). Since \( \pi_0 = \left( \sum_{i=0}^{10} \nu_i \right)^{-1} \) and \( \pi_i = \nu_i \pi_0 \) we get

\[ \pi_0 = 0.300; \quad \pi_1 = 0.300; \quad \pi_2 = 0.240; \quad \pi_3 = 0.096; \]
\[ \pi_4 = 0.038; \quad \pi_5 = 0.015; \quad \pi_6 = 0.006; \quad \pi_7 = 0.002; \]
\[ \pi_8 = 0.001; \quad \pi_9 = 0.000. \]

The mean of this distribution is obtained as

\[ L = E(Q) = 1.353. \]

Using Little’s law, for the mean system time we get

\[ W = 1.353/0.7 = 1.933 \text{ service time units.} \]

Note that we use the mean service time as the unit time for the purpose of determining the mean waiting time.

\[ \text{Ans.} \]
5.3 THE QUEUE G/M/1

Let customers arrive at time points \( t_0 = 0, t_1, t_2, \ldots \) and get served by a single server. Let \( Z_n = t_{n+1} - t_n, n = 1, 2, 3, \ldots \), be independent and identically distributed random variables with distribution function \( A(\cdot) \) with mean \( a \). Also let the service time distribution be exponential with mean \( 1/\mu \). It should be noted that this system has been traditionally represented by the symbol GI/M/1 (GI-General Independent). We use the symbolic representation G/M/1 for symmetry with the system M/G/1. Also I in GI does not really add any additional information.

Let \( Q(t) \) be the number of customers in the system at time \( t \) and define \( Q(t_n - 0) = Q_n, n = 1, 2, \ldots \). Thus \( Q_n \) is the number in the system just before the \( n \)th arrival. Define \( X_n \) as the number of potential service completions during the inter-arrival period \( Z_n \). (Note that we use the word ‘potential’ to indicate that there may not be \( X_n \) actual service completions, if the number of customers in the system soon after \( t_n \) is less than that number.) Let \( b_j, j = 0, 1, 2, \ldots \) be the distribution of \( X_n \). We have

\[
b_j = P(X_n = j) = \int_0^{\infty} e^{-\mu t} \left( \frac{\mu t}{j} \right)^j dA(t) \quad (5.3.1)
\]

Now consider the relationship between \( Q_n \) and \( Q_{n+1} \). We have

\[
Q_{n+1} = \begin{cases} 
Q_n + 1 - X_{n+1} & \text{if } Q_n + 1 - X_{n+1} > 0 \\
0 & \text{if } Q_n + 1 - X_{n+1} \leq 0
\end{cases} \quad (5.3.2)
\]

We should note that since \( X_{n+1} \) is defined as the potential number of departures, \( Q_n + 1 - X_{n+1} \) can be \(< 0\). Clearly \( Q_{n+1} \) does not depend on any random variable with an earlier index parameter than \( n \); hence \( \{Q_n, n = 0, 1, 2, \ldots \} \) is a Markov chain imbedded in the queue length process. From (5.3.2) we get the transition probability

\[
P_{ij} = P(Q_{n+1} = j|Q_n = i)
\]
CHAPTER 5. IMBEDDED MARKOV CHAIN MODELS

\[ P(X_{n+1} = i - j + 1) \text{ if } j > 0 \]
\[ P(X_{n+1} \geq i + 1) \text{ if } j = 0 \]

giving

\[ P_{ij} = b_{i-j+1} \quad j > 0 \]
\[ P_{i0} = \sum_{r=i+1}^{\infty} b_r. \quad (5.3.3) \]

The transition probability matrix takes the form

\[
\begin{pmatrix}
0 & 1 & 2 & 3 & \ldots \\
0 & \sum_1^\infty b_r & b_0 \\
1 & \sum_2^\infty b_r & b_1 & b_0 \\
2 & \sum_3^\infty b_r & b_2 & b_1 & b_0 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\
\end{pmatrix}
\]

\[ (5.3.4) \]

For the Markov chain to be irreducible \( b_0 > 0 \) and \( b_0 + b_1 < 1 \). (These two conditions can be justified in much the same way as for the queue M/G/1.) We can easily determine that the Markov chain is aperiodic. Let

\[ \phi(\theta) = \int_0^\infty e^{-\theta t} dA(t) \quad Re(\theta) > 0 \]

be the Laplace-Stieltjes transform of \( A(\cdot) \). Using \( \phi(\theta) \), the probability generating function of \( \{b_j\} \) is obtained as

\[ \beta(z) = \sum_{j=0}^{\infty} b_j z^j \quad |z| \leq 1 \]
\[ = \int_0^\infty e^{-(\mu - \mu z)t} dA(t) \]
\[ = \phi(\mu - \mu z). \]

Following the definitions given in (5.2.8), we get (using generic symbols \( X \) and \( Z \) for \( X_n \) and \( Z_n \)).

\[ E(Z) = B'(1) = -\mu \phi'(0) = a\mu \quad (5.3.5) \]
5.3. THE QUEUE G/M/1

We define the traffic intensity $\rho = (\text{arrival rate})/(\text{service rate})$. From (5.3.5) we get

$$\rho = \frac{1}{a\mu} \quad (5.3.6)$$

It can be shown that the Markov chain is positive recurrent when $\rho < 1$, null recurrent when $\rho = 1$ and transient when $\rho > 1$. (See discussion under M/G/1 for the implications of these properties. Also a proof is provided later in (5.3.31) and the remarks following that equation.)

The $n$-step transition probabilities $P^{(n)}_{ij} (i, j = 0, 1, 2, \ldots)$ of the Markov chain $\{Q_n\}$ are obtained as elements of the $n$th power of $P$. The observations made under M/G/1 regarding the behavior $P$ holds in the G/M/1 case as well. For analytical expressions for $P^{(n)}_{ij}$ the readers may refer to the same references, Takács (1962), Prabhu and Bhat (1963), and Prabhu (1965). In practice however, if the state space can be restricted to a manageable size depending on the computer power, successive multiplication of $P$ to get its power $P^n$ is likely to turn out to be the best course of action.

Limiting distribution

Let $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ be the limiting probabilities defined as $\pi_j = \lim_{n \rightarrow \infty} P^{(n)}_{ij}$. Based on Theorem 5.2.1, this limiting distribution exists when the Markov chain is irreducible aperiodic and positive recurrent, i.e., when $\rho < 1$. Theorem 5.2.2 provides the method to determine the limiting distribution. Thus from (5.2.12) we have the equations

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, \quad j = 0, 1, 2, \ldots$$

$$\sum_{0}^{\infty} \pi_j = 1$$
Using $P_{ij}$’s from (5.3.4) we get

$$
\pi_0 = \sum_{i=0}^{\infty} \pi_i (\sum_{r=i+1}^{\infty} b_r)
$$

$$
\pi_1 = \pi_0 b_0 + \pi_1 b_1 + \pi_2 b_2 + \ldots
$$

$$
\pi_2 = \pi_1 b_0 + \pi_2 b_1 + \pi_3 b_2 + \ldots
$$

$$
\vdots
$$

$$
\pi_j = \sum_{r=0}^{\infty} \pi_{j+r-1} b_r \quad (j \geq 1)
$$

(5.3.7)

The best computational method for the determination of the limiting distribution seems to be the direct matrix multiplication to get $P^n$ for increasing values of $n$ until the rows can be considered to be reasonably identical. The computational technique suggested for M/G/1 [(see Eq. (5.2.15)] does not work because of the lower triangular structure of $P$. As we will see later in the discussion of the finite queue G/M/1/K, unless we strat with a large enough $K$, restricting the state space to a finite value alters the last row of the matrix on which the technique has to be anchored.

In this case, however, (5.3.7) can be easily solved by the use of finite difference methods. This procedure is mathematically simple as well as elegant.

Define the finite difference operator $D$ as

$$
D \pi_i = \pi_{i+1}
$$

(5.3.8)

Using (5.3.8), equation (5.3.7) can be written as

$$
\pi_{j-1}(D - b_0 - Db_1 - D^2 b_2 - D^3 b_3 - \ldots) = 0
$$

(5.3.9)

Appealing to finite difference methods, a nontrivial solution to the equation (5.3.9) is obtained by solving its characteristic equation

$$
D - b_0 - Db_1 - D^2 b_2 - \ldots = 0
$$

$$
D = \sum_{j=0}^{\infty} b_j D^j
$$

$$
D = \beta(D).
$$

(5.3.10)
5.3. THE QUEUE G/M/1

Hence the solution to (5.3.10) should satisfy the functional equation

\[ z = \beta(z) \quad (5.3.11) \]

In (5.3.10) and (5.3.11) we have used the fact that \( \beta(z) \) is the probability generating function of \( \{b_j, j = 0, 1, 2, \ldots\} \).

To obtain roots of (5.3.11), consider two equations \( y = z \) and \( y = \beta(z) \). The intersections of these two equations gives the required roots.

We also have the following properties.

\( \beta(0) = b_0 > 0; \beta(1) = \sum_0^\infty b_j = 1; \beta'(1) = \rho^{-1} \)

\( \beta''(z) = 2b_2 + 6b_3 z + \ldots > 0 \) for \( z > 0 \).

Hence \( \beta'(z) \) is monotone increasing and therefore \( \beta(z) \) is convex.

Of the two equations, \( y = z \) is a straight line passing through 0 and since \( \beta(0) = b_0 > 0, \beta(1) = 1, \) and \( \beta(z) \) is convex, equation \( y = z \) and \( y = \beta(z) \) intersect in almost two points; one of them is at \( z = 1 \). Let \( \zeta_s \) be the second root. Whether \( \zeta_s \) lies to the left or to the right of 1 is dependent on the value of the traffic intensity \( \rho \).

**Case 1**: \( \rho < 1 \).

When \( \rho < 1, \beta'(1) > 1 \); then \( y = \beta(z) \) intersects \( y = z \) approaching from below at \( z = 1 \). But \( b_0 > 0 \). Hence \( \zeta_s < 1 \).

**Case 2**: \( \rho > 1 \)

. When \( \rho > 1, \beta'(1) < 1 \). Then \( y = \beta(z) \) intersects \( y = z \) approaching from above at \( z = 1 \). Hence \( \zeta_s > 1 \).

**Case 3**: \( \rho = 1 \)

. In this case \( \beta'(1) = 1 \) and \( y = z \) is a tangent to \( y = \beta(z) \) at \( z = 1 \). This means \( \zeta_s \) and 1 coincide.
Let $\zeta$ be the least positive root. We have $\zeta < 1$ if $\rho < 1$ and $\zeta = 1$ is $\rho \geq 1$. This root is used in the solution of the finite different equation (5.3.9). (Note that our solution is in terms of probabilities which are $\leq 1$, the root we use must be $\leq 1$ as well.)

Going back to the difference equation (5.3.9), we can say that

$$\pi_j = c\zeta^j \quad (j > 0) \quad (5.3.12)$$

is a solution. Since $\zeta < 1$, $\sum_0^\infty \pi_j = 1$, we get

$$\sum_j \pi_j = c \sum_0^\infty \zeta^j = \frac{c}{1-\zeta} = 1$$

giving

$$c = 1 - \zeta$$

Substituting this back into (5.3.12), we get

$$\pi_j = (1 - \zeta)\zeta^j \quad j = 0, 1, 2, \ldots \quad (5.3.13)$$

as the limiting distribution of the state of the system in the queue G/M/1.

Note that $\zeta$ is the root of the equation

$$z = \phi(\mu - \mu z) \quad (5.3.14)$$

In most cases the root $\zeta$ of (5.3.14) has to be determined using numerical techniques.

With the geometric structure for the limiting distribution (5.3.13), the mean and variance of the number in the system, say $Q$, are easily obtained. We have (the superscript $A$ denotes arrival point restriction)

$$L^A = E(Q^A) = \frac{\zeta}{1-\zeta}; \quad L_q^A = \frac{\zeta^2}{1-\zeta}$$

$$V(Q^A) = \frac{\zeta}{(1-\zeta)^2} \quad (5.3.15)$$

It is important to note that the imbedded Markov chain analysis gives the properties of the number in the system at arrival epochs. (For convenience we
have used the number before the arrivals). As pointed out under the discussion on the M/G/1 queue the limiting distributions of the number of customers in the system at arrival epochs, departure epochs, and at arbitrary points in time are the same only when the arrivals occur as a Poisson process. Otherwise we have to make appropriate adjustments to the distribution derived above. In this context, results derived in Prabhu (1965) and Bhat (1968) are worth mentioning. Writing \( p_j = \lim_{t \to \infty} P[Q(t) = j] \), where \( Q(t) \) is the number at an arbitrary time \( t \), these authors arrive at the following explicit expression for the limiting distribution \( \{ p_j, j = 0, 1, 2 \ldots \} \), when \( \rho < 1 \).

\[
p_0 = 1 - \rho \\
p_j = \rho \lambda (1 - \zeta) \sum_{n=j}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{n-j}}{(n-j)!} \int_{\tau=0}^{t} [(t-\tau) + \frac{j-1}{n} \tau][1 - A(t-\tau)] dA_n(\tau) \quad j \geq 1 \tag{5.3.16}
\]

Alternatively, \( p_j \) can also be expressed as

\[
p_j = \rho (1 - \zeta) \zeta^{j-1} \quad j \geq 1 \tag{5.3.17}
\]

From (5.3.17) the following results follow on mean queue length.

\[
L = \frac{\rho}{1 - \zeta}; \quad L_q = \frac{\rho \zeta}{1 - \zeta} \tag{5.3.18}
\]

As an example, consider the queue M/M/1. Let \( A(t) = 1 - e^{-\lambda t} \ (t \geq 0) \). Then we have

\[
\phi(\theta) = \frac{\lambda}{\lambda + \theta} \\
\phi(\mu - \mu z) = \frac{\lambda}{\lambda + \mu - \mu z}.
\]

Now, the functional equation takes the form

\[
z = \frac{\lambda}{\lambda + \mu - \mu z} \\
-\mu z^2 + (\lambda + \mu)z - \lambda = 0.
\]
This quadratic equation has two roots 1 and $\frac{1}{\mu} = \rho$. Substituting $\rho$ in place of $\zeta$ in (5.3.16) - (5.3.18) we have the limiting distribution and mean values for the queue M/M/1, which match with the results derived in Chapter 4.

**Waiting time**

To determine the distribution of the waiting time of a customer we need the distribution of the number of customers in the system at the time of its arrival. The limiting distribution derived in (5.3.13) is in fact an arrival point distribution in G/M/1. Furthermore its structure is the same as the geometric distribution we had for M/M/1, with $\zeta$ taking the place of $\rho$ of the M/M/1 result. The service times of customers in the system are exponential with rate $\mu$, also as in M/M/1. Hence the waiting time results for G/M/1 have the same forms as those for M/M/1 with $\zeta$ replacing $\rho$. Without going to the detail of their derivation, we can write,

$$F_q(t) = P(T_q \leq t) = 1 - \zeta e^{-\mu(1-\zeta)t}$$

$$W_q = E[T_q] = \frac{\zeta}{\mu(1 - \zeta)}$$

$$V[T_q] = \frac{\zeta(2 - \zeta)}{\mu^2(1 - \zeta)^2}. \quad (5.3.19)$$

The time $T$ spent by the customer in the system is obtained by adding service time to $T_q$. We get

$$W = E(T) = E(T_q) + \frac{1}{\mu} = \frac{1}{\mu(1 - \zeta)}$$

$$V(T) = V[T_q + S] = \frac{1}{\mu^2(1 - \zeta)^2}. \quad (5.3.20)$$

**Busy cycle**

A busy cycle of a G/M/1 queue, when modeled as an imbedded Markov chain is the number of transitions the process takes to go from state 0 to state 0 for the first time. This interval is also known as the *recurrence time* of state 0. The
5.3. THE QUEUE G/M/1

busy cycle includes the busy period when the server is continuously busy, and the idle period, when there is no customer in the system. Let $R$ denote the number of transitions in a busy cycle. (Note that we are using a generic symbol $R$, with the assumption that all such busy cycles have the same distribution.) Let $h_j^{(n)}$ be the probability that the number of customers, just before the $n$th arrival in a busy cycle is $j$. Working backward from $n$, considering the arrival time of the first of those $j$ customers we can write, for $j \geq 1$,

$$
\begin{align*}
  h_j^{(j)} &= b_0^{(j)} \\
  h_j^{(n)} &= \sum_r h_r^{(n-j)} b_r^{(j)} & n \geq j
\end{align*}
$$

(5.3.21)

where $b_r^{(j)}$ is the $j$-fold convolution of $b_r$ with itself. Looking back to relations (5.2.45) we see that (5.3.21) is structurally similar to (5.2.45) with $h_j^{(n)}$ replacing $g_i^{(n)}$ and $b_r^{(i)}$ replacing $h_r^{(i)}$. Define

$$
H_j(z) = \sum_{n=j}^{\infty} h_j^{(n)} z^n \quad |z| \leq 1
$$

(5.3.22)

Using arguments similar to those used in determining $G(z)$, we can show that

$$
H_j(z) = [\eta(z)]^j, \quad j \geq 1
$$

(5.3.23)

where $\eta(z)$ is unique root in the unit circle of the equation

$$
\omega = z\beta(\omega)
$$

(5.3.24)

The distribution of $R$ is given by $h_0^{(n)}$. Considering the transitions during the $n$th transition interval, we have

$$
\begin{align*}
  h_0^{(n)} &= \sum_{r=1}^{n-1} h_r^{(n-1)} \left( \sum_{k=r+1}^{\infty} b_k \right) \\
  h_0^{(1)} &= \sum_{k=1}^{\infty} b_k
\end{align*}
$$

(5.3.25)

Taking generating functions

$$
H_0(z) = \sum_{n=1}^{\infty} h_0^{(n)} z^n
$$
\[ \begin{align*}
\sum_{k=1}^{\infty} b_k z + \sum_{n=2}^{\infty} \sum_{r=1}^{n-1} h_r^{(n-1)} \left( \sum_{k=r+1}^{\infty} b_k \right) & \quad (5.3.26)
\end{align*} \]

The right hand side of (5.3.26) can be simplified as follows. For ease of notation write \( \sum_{r+1}^{\infty} b_k = \beta_r \). The RHS of (5.3.26) simplifies to

\[ \begin{align*}
\beta_0 z + z \sum_{n=2}^{\infty} z^{n-1} \sum_{r=1}^{\infty} \beta_r h_r^{(n-1)} & \\
= z \left[ \beta_0 + \sum_{r=1}^{\infty} \beta_r \sum_{n=r+1}^{\infty} h_r^{(n-1)} z^{n-1} \right] & \\
= z \left[ \sum_{r=0}^{\infty} \beta_r \eta(z)^r \right]
\end{align*} \]

where we have used (5.3.22) and (5.3.23). But

\[ \sum_{r=0}^{\infty} \beta_r z^r = \frac{1 - \beta(z)}{1 - z} \]

since

\[ \sum_{r=0}^{\infty} z^r \sum_{j=r+1}^{\infty} b_j = \sum_{j=1}^{\infty} b_j \sum_{r=0}^{j-1} z^r \]
\[ = \sum_{j=1}^{\infty} b_j \left( \sum_{r=0}^{\infty} z^r - \sum_{r=j}^{\infty} z^r \right) \]
\[ = \sum_{j=1}^{\infty} b_j \left[ \frac{1}{1 - z} - z^j \sum_{r=j}^{\infty} z^{r-j} \right] \]
\[ = \frac{1 - \beta(z)}{1 - z}. \]

Thus we get

\[ H_0(z) = \frac{z - z\beta(\eta(z))}{1 - \eta(z)} \]

But \( \eta(z) \) is such that

\[ \eta(z) = \frac{z}{\beta(\eta(z))} \]

Hence

\[ H_0(z) = \frac{z - \eta(z)}{1 - \eta(z)} \quad (5.3.27) \]
Letting \( z \to 1 \) in \( H_0(z) \), we can show that \( R \) is a proper random variable (i.e. \( P(R < \infty) \)) when \( \rho \leq 1 \). The expected length of the busy cycle (recurrence time of state 0) is obtained as \( H_0'(z) \). We have

\[
H_0'(z) = \frac{[1 - \eta(z)][1 - \eta'(z)] + [z - \eta(z)]\eta'(z)}{[1 - \eta(z)]^2}
= \frac{1 - \eta'(z)}{1 - \eta(z)} + \frac{[z - \eta(z)]\eta'(z)}{[1 - \eta(z)]^2}
\]

To simplify (5.3.28) further, we need values for \( \eta(1) \) and \( \eta'(1) \). Referring back to the functional equation (5.3.24), we find that for \( z = 1 \) it is the solution of the functional equation (5.3.11) which we have found to be the least positive root \( \zeta \) (< 1 or = 1). Hence

\[
\eta(1) = \zeta \text{ if } \rho < 1 \text{ and } = 1, \text{ if } \rho \geq 1.
\]  

Consider \( \eta(z) = z\beta[\eta(z)] \). We get

\[
\eta'(z) = z\beta'[\eta(z)]\eta'(z) + \beta[\eta(z)]
\eta'(z)[1 - z\beta'[\eta(z)]] = \beta(\eta(z))
\]

Letting \( z \to 1 \), and using (5.3.29)

\[
\eta'(1)[\beta'(\zeta)] = \beta(\zeta)
\eta'(1) = \frac{\beta(\zeta)}{1 - \beta'(\zeta)}
\]

Substituting from (5.3.29) and (5.3.30) in (5.3.28)

\[
\lim_{z \to 1} H_0'(z) = \left[1 - \frac{\beta(\zeta)}{1 - \beta'(\zeta)}\right] \frac{1}{1 - \zeta} + \left[(1 - \zeta) \frac{\beta(\zeta)}{1 - \beta'(\zeta)}\right] \times \frac{1}{(1 - \zeta)^2}
= \frac{1}{1 - \zeta} < \infty \text{ if } \rho < 1.
\]

Similarly, we can also show that \( H_0'_{\lim z \to 1}(z) = \infty \) when \( \rho = 1 \).

We may note that these results establish the classification properties of positive recurrence, null recurrence and transience of the imbedded Markov chain.
The mean length of the busy cycle is obtained as the product (expected number of transitions) \( \times \) (mean interarrival time).

\[
E[busycycle] = \frac{E(Z)}{1 - \zeta}
\]  

(5.3.32)

Since the busy period terminates during the last transition of the Markov chain and the transition interval (interarrival time) has a general distribution, the determination of the mean busy period is too complicated to be covered at this stage.

**The queue G/M/s**

The imbedded Markov chain analysis of the queue G/M/1 can be easily extended to the multi-server queue G/M/s. Since the Markov chain is defined at arrival points, the structure of the process is similar to that of G/M/1, except for the transition probabilities. Retaining the same notations, for the relationship between \( Q_n \) and \( Q_{n+1} \), we get

\[
Q_{n+1} = \begin{cases} 
Q_n + 1 - X_{n+1} & \text{if } Q_n + 1 - X_{n+1} > 0 \\
0 & \text{if } Q_n + 1 - X_{n+1} \leq 0,
\end{cases}
\]

where \( X_{n+1} \) is the total number of potential customers who can be served by \( s \) servers during an interarrival time with distribution \( A(\cdot) \).

To determine transition probabilities \( P_{ij} \) \((i, j = 0, 1, 2 \ldots)\) we have to consider three cases for the initial value \( i \), and the final value \( j \): \( i + 1 \geq j \geq s \); \( i + 1 \leq s \) and \( j \leq s \); \( i + 1 > s \) but \( j < s \). Note that when \( Q_n = i \), the transition starts with \( i + 1 \), and \( j \) is always \( \leq i + 1 \). Since the service times are exponential with density \( \mu e^{-\mu x} \) \((x > 0)\), the probability that a server will complete service during \((0, t]\) is \(1 - e^{-\mu t}\) and the probability that the service will continue beyond \( t \) is \( e^{-\mu t}\). Incorporating these concepts along with the assumptions that the servers work independently of each other we get the following expressions for \( P_{ij} \).
5.3. THE QUEUE G/M/1

Case (1): \( i + 1 \geq j \geq s \)

\[
P_{ij} = \int_0^\infty e^{-s\mu t} \frac{(i + 1 - j)!}{(i + 1)!} dA(t) \tag{5.3.33}
\]

This represents \( i + 1 - j \) service completions during an interarrival period, when all \( s \) servers are busy.

Case (2): \( i + 1 \leq s \) and \( j \leq s \)

\[
P_{ij} = \left( \frac{i + 1}{i + 1 - j} \right) \int_0^\infty (1 - e^{-\mu t})^{i+1-j} e^{-j\mu t} dA(t) \tag{5.3.34}
\]

This expression takes into account the event \( i + 1 - j \) out of \( i + 1 \) customers completing service during \( (0, t] \) while customers are still being served. Because of the independence of servers among one another, each service can be considered a Bernoulli trial and the outcome has a binomial distribution with success probability \( 1 - e^{-\mu t} \).

Case (3): \( i + 1 > s \) but \( j < s \)

\[
P_{ij} = \int_0^t \int_{\tau=0}^t e^{-s\mu \tau} \frac{(s\mu \tau)^{i-s}}{(i-s)!} s\mu \left( \frac{s}{s-j} \right) [1 - e^{-\mu(t-\tau)}]^{s-j} e^{-j\mu(t-\tau)} d\tau dA(t) \tag{5.3.35}
\]

\[
(5.3.36)
\]

Initially \( i + 1 - s \) service completions occur with rate \( s\mu \), then \( s - j \) out of the remaining \( s \) complete their service independently of each other.

The transition probability matrix of the imbedded chain has a structure similar to the one displayed in (5.3.4). Because of the structure of \( P_{ij} \) values under cases (2) and (3), the finite difference solution given earlier for the limiting distribution need major modifications. Interested readers are referred to Gross and Harris (1998, pp. 256-258) or Takács (1962). Taking into consideration the complexities of these procedures, the computational method developed for G/M/1/K could turn out to be advantageous in this case, if it is possible to work with a finite limit for the number of customers in the system.
5.3.1 The queue G/M/1/K

Consider the G/M/1 queue described earlier with the restrictions that the system can accommodate only $K$ customers at a time. Since the imbedded chain is defined just before an arrival epoch, the number of customers in the system soon after the arrival epoch is $K$, whether it is $K$ or $K-1$ before that time point. If it is $K$ before, the arriving customer does not get admitted to the system. Thus in place of (5.3.2) we have the relation

$$Q_{n+1} = \begin{cases} \min (Q_n + 1 - X_{n+1}, K) & \text{if } Q_n + 1 - X_{n+1} > 0 \\ 0 & \text{if } Q_n + 1 - X_{n+1} \leq 0 \end{cases} \quad (5.3.37)$$

Using probabilities $b_j$, $j = 0, 1, 2 \ldots$ defined in (5.3.2) the transition probability matrix $P$ can be displayed as follows:

$$P = \begin{bmatrix} 0 & 1 & 2 & \ldots & K-1 & K \\ \sum_1^\infty b_r & b_0 & 0 \\ \sum_2^\infty b_r & b_1 & b_0 \\ \vdots & \vdots & \vdots \\ \sum_K^\infty b_r & b_{K-1} & b_{K-2} & \ldots & b_1 & b_0 \\ \sum_K^\infty b_r & b_{K-1} & b_{K-2} & \ldots & b_1 & b_0 \end{bmatrix} \quad (5.3.38)$$

Note that the last two rows of the matrix $P$ are identical because the Markov chain effectively starts off with $K$ customers from either of the states $K-1$ and $K$.

Let $\pi = (\pi_0, \pi_1, \ldots, \pi_K)$ be the limiting distributions of the imbedded chain. Writing out $\pi_j = \sum_{i=0}^K \pi_i P_{ij}$, we have

$$\begin{align*}
\pi_0 &= \sum_{i=0}^K \pi_i \left( \sum_{r=i+1}^\infty b_r \right) \\
\pi_i &= \pi_0 b_0 + \pi_1 b_1 + \ldots + \pi_{K-1} b_{K-1} + \pi_K b_{K-1} \\
\pi_2 &= \pi_1 b_0 + \pi_2 b_1 + \ldots + \pi_{K-1} b_{K-1} + \pi_K b_{K-2}
\end{align*}$$
5.3. THE QUEUE G/M/1

\[ \pi_{K-1} = \pi_{K-2}b_0 + \pi_{K-1}b_1 + \pi_Kb_1 \]
\[ \pi_K = \pi_{K-1}b_0 + \pi_Kb_0 \]  

(5.3.39)

If the value of \( K \) is not too large solving these simultaneous equations in \( \pi_j, j = 0, 1, 2, \ldots, K \), along with the normalizing condition \( \sum_0^K \pi_j = 1 \) directly, could be computationally practical. Or for that matter getting \( P^n \) for increasing values of \( n \) until the row elements are close to being identical will also give the limiting distribution under these circumstances. An alternative procedure is to develop a computational recursion as done in the case of the M/G/1 queue [see Eq. (5.2.15)].

To do so we start with the last equation of (5.3.38) and define

\[ \nu_i = \frac{\pi_i}{\pi_{i-1}}, \quad i = 1, 2, \ldots, K. \]

We have

\[ \pi_i = \nu_i\pi_{i-1} \]
\[ = \nu_i\nu_{i-1}\pi_{i-2} \]
\[ = \nu_i\nu_{i-1}\ldots\nu_1\pi_0 \]  

(5.3.40)

From the last equation in (5.3.38) we get

\[ \nu_K = b_0 + \nu_Kb_0 \]
\[ \nu_K = \frac{b_0}{1 - b_0} \]

From the next to the last equation in (5.3.38) we get

\[ \nu_{K-1} = b_0 + \nu_{K-1}b_1 + \nu_K\nu_{K-1}b_1 \]
\[ \nu_{K-1} = \frac{b_0}{1 - b_1 - \nu_Kb_1} \]

and so on.
CHAPTER 5. IMBEDDED MARKOV CHAIN MODELS

Since $\sum_0^K \pi_j = 1$, from (5.3.39) we get

$$(1 + \nu_1 + \nu_1\nu_2 + \ldots + \nu_1\nu_2\ldots\nu_K)\pi_0 = 1$$

and hence

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^K \Pi_{r=1}^i \nu_r} \tag{5.3.41}$$

Substituting these back in (5.3.39) we get $\pi_j, j = 0, 1, 2, \ldots K$.

We may note that in developing the recursion we have defined $\nu_i$'s as ratio of consecutive $\pi_i$'s, unlike in the case of (5.2.15). We do so for the reason that $\pi_j$'s decrease in value as $j$ increases and dividing by a very small $\pi_j$ is likely to result in large computational errors. Looking at the structure of the limiting distribution of G/M/1, the ratio of consecutive terms of $\pi$, are likely to be close to the constant $\zeta$.

**Example 5.3.1**

In a service center job arrivals occur in a deterministic process, one job per one unit of time. Service is provided by a single server with an exponential service time distribution with rate 1.5 jobs per unit time.

In order to determine the limiting distribution, using a D/M/1 model, we note that

$$\phi(\theta) = \int_0^\infty e^{-\theta t}dA(t)$$

$$= e^{-\theta}$$

With an exponential service time distribution we have $\mu = 1.5$. Hence

$$\beta(z) = \phi(\mu - \mu z)$$

$$= e^{-1.5(1-z)}$$

The limiting distribution is expressed in terms of $\zeta$ which is the unique root in the unit circle, of the functional equation

$$z = e^{-1.5(1-z)}$$
We can easily solve this equation by successive substitution starting with \( z = 0.4 \). We get

\[
\begin{array}{cc}
z & \beta(z) \\
0.400 & 0.407 \\
0.407 & 0.411 \\
0.411 & 0.413 \\
0.413 & 0.415 \\
0.415 & 0.416 \\
0.416 & 0.416 \\
\end{array}
\]

We use \( \zeta = 0.416 \) in the limiting distribution \( \pi_0 = (\pi_0, \pi_1, \pi_2, \ldots) \) given by (5.3.13). We get

\[
\begin{align*}
\pi_0 &= 0.584; & \pi_1 &= 0.243; & \pi_2 &= 0.101; & \pi_3 &= 0.042; \\
\pi_4 &= 0.017; & \pi_5 &= 0.007; & \pi_6 &= 0.003; & \pi_7 &= 0.001.
\end{align*}
\]

\textit{Ans.}

\section*{Example 5.3.2}

Consider the service center Example 5.3.1 above with a capacity restriction of \( K \) customers in the system.

In this case we use the computational recursion developed in (5.3.39), for two values of \( K = 4 \) and 7.

The distribution of the potential number of customers served during an interarrival period is Poisson with mean 1.5. We have

\[
\begin{align*}
b_0 &= 0.223; & b_1 &= 0.335; & b_2 &= 0.251; & b_3 &= 0.125 \\
b_4 &= 0.047; & b_5 &= 0.015; & b_6 &= 0.003; & b_7 &= 0.001.
\end{align*}
\]

\( K = 4 \)

Using \( \nu_i = \pi_i / \pi_{i-1} \) in (5.3.38) with \( K = 4 \), we get

\[
\nu_4 = b_0 [1 - b_0]^{-1}
\]
\[\begin{align*}
\nu_3 &= b_0 [1 - b_1 - \nu_4 b_1] \\
\nu_2 &= b_0 [1 - b_1 - \nu_3 b_2 - \nu_4 \nu_3 b_2] \\
\nu_1 &= b_1 [1 - b_1 - \nu_2 b_2 - \nu_3 \nu_2 b_3 - \nu_4 \nu_3 \nu_2 b_3]^{-1}
\end{align*}\]

Substituting appropriate values of \(b_j = 0, 1, 2, 3\), we get

\[\nu_1 = 0.413; \quad \nu_2 = 0.398; \quad \nu_3 = 0.392; \quad \nu_4 = 0.287\]

But we have

\[\begin{align*}
\pi_4 &= \nu_4 \nu_3 \nu_2 \nu_1 \pi_0 \\
\pi_3 &= \nu_3 \nu_2 \nu_1 \pi_0 \\
\pi_2 &= \nu_2 \nu_1 \pi_0 \\
\pi_1 &= \nu_1 \pi_0
\end{align*}\]

Using \(\sum_0^4 \pi_j = 1\), we get

\[\pi_0 = [1 + \nu_1 + \nu_1 \nu_2 + \nu_1 \nu_2 \nu_3 + \nu_1 \nu_2 \nu_3 \nu_4]^{-1} = 0.602.\]

Thus we have the limiting distribution

\[\pi_0 = 0.602; \quad \pi_1 = 0.249; \quad \pi_2 = 0.099; \quad \pi_3 = 0.039; \quad \pi_4 = 0.001\]

\(\text{Ans.}\)

\(K = 7\)

Looking at the structure of \(\nu_i\)’s, it is clear that \(\nu_4, \ldots, \nu_1\) determined above, in fact, yield \(\nu_7, \ldots, \nu_4\), when \(K = 7\). Extending the equations to determine the remaining \(\nu\)’s, viz, \(\nu_3, \nu_2, \) and \(\nu_1\), we get the following set of values.

\[\nu_7 = 0.287; \quad \nu_6 = 0.392; \quad \nu_5 = 0.398; \quad \nu_4 = 0.413; \quad \nu_3 = 0.415; \quad \nu_2 = 0.416; \quad \nu_1 = 0.417\]

Converting these back to \(\pi\)’s, we get
\[ \pi_0 = 0.585; \quad \pi_1 = 0.244; \quad \pi_2 = 0.101 ; \]
\[ \pi_3 = 0.042; \quad \pi_4 = 0.017; \quad \pi_5 = 0.007 ; \]
\[ \pi_6 = 0.003; \quad \pi_7 = 0.001. \]

A comparison of these values with those obtained under Example 5.3.1 shows that when \( K = 7 \) the effect of the capacity limit is negligible for the long run distribution of the process.

References


REFERENCES


