Chapter 3

An Introduction to
Stochastic Processes

3.1 Stochastic Process

In this chapter we introduce basic concepts used in analyzing queueing systems. Analysis techniques are developed later in conjunction with the discussion of specific systems.

Uncertainties in model characteristics lead us to random variables as the basic building block for the queueing model. However, a random variable represents an event in a random phenomenon. In queueing systems, and all systems that operate over time (or space or any other parameter), the model must be able to represent the system over time. That means we need a sequence or a family of random variables to represent a phenomenon over time. Let $T$ be the range of time of interest. Time can be continuous or discrete. We denote the time $t \in T$, when it is continuous and $n \in T$, when it is discrete. Then the family of random variables $\{X(t), t \in T\}$ or the sequence of random variables $\{X_n, n \in T\}$ is known as a stochastic process. (A sample value of a random variable can be
looked upon as a snapshot, whereas, a sample path of a stochastic process can be considered a video.) The space in which $X(t)$ or $X_n$ assume values is known as the state space and $T$ is known as the parameter space. Another way of saying is that a stochastic process is a family or a sequence of random variables indexed by a parameter.

### 3.2 The Stationary Process

The properties of a random variable are determined from its distribution. For instance $P(X(t) \leq x)$ for a specified $t$ gives the distribution of the stochastic process at time $t$. To get the properties of the process for all $t \in T$, one has to get the distribution of $X(t)$ for all $t$, which in its generality is intractable. Instead, if we consider a discrete set of points $t_1, t_2, \ldots, t_n \in T$ and find

$$F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = P(X(t_1) \leq x_1, X(t_2) \leq x_2, \ldots, X(t_n) \leq x_n)$$

we have a fairly general representation of the process. However, generating information from such an expression is a formidable task as well. Accordingly we seek properties of the process when its dependence structure is known.

Let, for $s < t$,

$$
\mu(t) = E[X(t)]; \quad \mu_2(t) = E[(X(t))^2] \\
\mu_{11}(s, t) = E[X(s)X(t)]; \quad \sigma^2(t) = \mu_s(t) - [\mu(t)]^2
$$

the covariance function

$$
\gamma(s, t) = \mu_{11}(s, t) - \mu(s)\mu(t)
$$

and the correlation function $\rho(s, t) = \gamma(s, t) / [\sigma(s)\sigma(t)]$. When the mean and variance do not depend on the time parameter $t$, we have

$$
\mu(t) = \mu; \quad \sigma^2(t) = \sigma^2; \quad \gamma(s, t) = \gamma(0, t - s).
$$

The stochastic process is now called covariance stationary (or weak, or wide sense stationary). But suppose, $F(x_1, x_2, \ldots, x_n; t_1 + h, t_2 + h, \ldots, t_n + h) =$
3.3. POINT, REGENERATIVE, AND RENEWAL PROCESSES

F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n). Then the process \{X(t) + t\epsilon T\} is said to be strictly stationary. This implies that a translation in the origin of the time parameter does not affect the distribution properties of the process. It should be noted that a strictly stationary process is also covariance stationary; but a covariance stationary process need not be strictly stationary. A wide array of time series models involving stationary and non-stationary processes (in which even their mean value functions such as \mu(t) and \sigma^2(t) depend on t) are used to analyze random phenomena in the real world; e.g. stock prices, national income, sales figures, population figures, industrial output, etc. The analyses procedure involves the identification of the appropriate model using either the covariance and correlation functions or a harmonic analysis of the periodic functions that make up the time series. Some examples of the time series models used are, autoregressive processes, moving average processes, and their various combinations. For an introduction to stationary processes and time series analysis the readers are referred to Bhat and Miller (2002), Ch. 11 and references cited in it.

Even though the underlying processes of queueing systems are time series, their structure allows us to use an entirely different set of procedures in their analysis. Queueing processes are the product of arrivals and service. Even when we define continuous state processes such as waiting times, arrival and departure points are embedded in them. The next three sections describe commonly occurring processes used in the analysis of queueing systems.

3.3 Point, Regenerative, and Renewal Processes

Point process

Consider randomly located discrete set of points in the parameter space T. These points may represent events such as arrivals in a queueing system or accidents on a stretch of road. Let N(t), t\epsilon T be the number of points in (0, t].
Then the counting process \( N(t) \) is known as a point process (see, Lewis (1972)).

There are processes in which the points may be of different types. For instance, the arrival of two types of customers. Then the process is identified as a marked point process.

**Regenerative process**

Consider a stochastic process \( \{X(t), t \in T\} \) and a discrete set of points \( t_1 < t_2 < \ldots < t_n \in T \). Suppose the distribution properties of the process from \( t_i \) onwards is the same for all \( i = 1, 2, \ldots, n \). Then we can consider the process regenerating itself at these points.

**Renewal Process**

Consider a discrete set of points \( (t_0, t_1, t_2, \ldots) \) at which a specified event occurs and let \( t_i - t_{i-1} = Z_i \) (\( i = 1, 2, \ldots \)), be independent and identically distributed (i.i.d) random variables. The process of the sequence of random variables \( (Z_1, Z_2, \ldots) \) is known as a renewal process. Let \( N(t) \) be the process representing the number of events occurring in \( (0, t] \). It is known as the renewal counting process. The periods \( Z_i \) (\( i = 1, 2, \ldots \)) are renewal periods. Since the renewal periods one i.i.d., it is clearly seen that the renewal process is also a regenerative process.

In the context of queueing systems, when the inter-arrival times are i.i.d., the arrivals form a renewal process. But, since a departure cannot take place when there are no customers in the system, the departure process is not renewal even when service times are i.i.d. random variables. They form a renewal process only during the period customers are continuously busy. (Periods when customers are continuously busy are known as busy periods. They are followed by idle periods during which the server is idle). When the queue discipline dictates that the server does not stay idle when there are customers in the system, the starting points of busy periods form another set of renewal points, with the sequence of
busy period-idle period pairs forming the renewal periods. The renewal process framework is useful in analyzing some advanced classes of queueing systems.

3.4 Markov Process

Some of the simple models widely used in queueing theory are based on Markov processes. Suppose, a stochastic process \( \{X(t), t \in T\} \) is such, that

\[
P[X(t) \leq x | X(t_1) = x_1, x(t_2) = x_2, \ldots, X(t_n) = x_n] = P[X(t) \leq x | X(t_n) = x_n] (t_1 < t_2 \ldots < t_n < t)
\]

\[
= F(x_n, x; t_n, t) \tag{3.4.1}
\]

Then \( \{X(t)\} \) is a Markov process. When \( T \) and the state space are discrete the parallel definition is given as

\[
P(X_n = j | x_{n_1} = i_1, x_{n_2} = i_2, \ldots, X_{n_k} = i_k) = P(X_n = j | X_{n_k} = i_k)
\]

\[
= P^{(n_k, n)}_{i_k, j} \tag{3.4.2}
\]

Now the process \( \{X_n, n = 1, 2, \ldots\} \) is called the Markov chain.

The dependence structure exhibited here is a one-step dependence, in which the state of the process is dependent only on the last parameter point at which full information of the process is available. As can be seen in the following chapters, the property of Markov dependence simplifies the analysis while retaining essential characteristics of the systems.

Since the time parameter in a Markov process has a specific range we use transition distributions or probabilities of the process in its analysis. These are conditional statements, conditioned on the process value at the initial values of \( t \). An unconditional distribution or the probability (in the discrete case) can be obtained by the usual method of removing the condition.

The fundamental property of the Markov process is given by the Chapman-Kolmogorov relation. For the transition probabilities of Markov processes we
use the following notations depending on the nature of state and parameter spaces.

(i) Discrete state, discrete parameter:

\[ P_{i,j}^{m,n} = P(X_n = j | X_m = i), \quad m < n, \tag{3.4.3} \]

(ii) Discrete state, continuous parameter:

\[ P_{i,j}(s,t) = P[X(t) = j | X(s) = i], \quad s < t, \tag{3.4.4} \]

(iii) Continuous state, discrete parameter:

\[ F(x_m, x; m,n) = P(X_n \leq x | X_m = x_m), \quad m < n, \tag{3.4.5} \]

(iv) Continuous state, continuous parameter

\[ F(x_n, x; t_n, t) = P[X(t) \leq x | X(t_n) = x_n], \quad t_n < t \tag{3.4.6} \]

Corresponding to these four cases, the Chapman-Kolmogorov relation can be given as follows

(i)

\[ P_{i,j}^{(m,n)} = \sum_{k \in S} P_{i,k}^{(m,r)} P_{k,j}^{(r,n)} \quad (m < r < n) \tag{3.4.7} \]

(ii)

\[ P_{i,j}(s,t) = \sum_{k \in S} P_{i,k}(s,u) P_{k,j}(u,t) \quad (s < u < t) \tag{3.4.8} \]

(iii)

\[ F(x_m, x; m,n) = \int_{y \in S} dy F(x_m, y; m, r) \cdot F(y, x; r, n) \quad (m < r < n) \tag{3.4.9} \]

(iv)

\[ F(x_n, x; s, t) = \int_{y \in S} dy F(x_n, y; s, u) \cdot F(y, x; u, t) \quad (s < u < t) \tag{3.4.10} \]
3.4. MARKOV PROCESS

These equations can be easily established by considering the transitions of the process in two time periods \((m, r)\) and \((r, n)\) when the time parameter is discrete and \((s, u)\) and \((u, t)\) when the time parameter is continuous, and using the basic definition of the Markov process. For instance, when both the state and parameter spaces are discrete, the probability of the transition from the initial state \(i\) to a state \(k\) \((k \in S)\) in time period \((m, r)\) is \(P_{ik}^{(m, r)}\) and from state \(k\) to state \(j\) in time period \((r, n)\) is \(P_{kj}^{(r, n)}\). Eq. (3.4.7) now follows by multiplying these two probabilities and summing over all values of \(k \in S\). Similar arguments establish (3.4.8) - (3.4.10).

The stochastic processes underlying the queueing systems considered in this book primarily belong to two classes: discrete state and parameter spaces (case (i) above) and discrete state and continuous parameter space (case (ii) above). Here we provide the conceptual framework for the method by which Eqs. (3.4.7) and (3.4.8) can be used in their analysis.

Case (i): Discrete state and parameter space

Let \(\{X_n, n = 0, 1, 2 \ldots\}\) be a time homogeneous Markov chain. By time homogeneous we mean that the transition probabilities \(P_{ij}^{(m,n)}\) and \(P_{ij}^{(m+k,n+k)}\) are the same. Without loss of generality we use \(m = 0\) and write

\[
P_{ij}^{(n)} = P(X_n = j | X_0 = i) \quad (3.4.11)
\]

For convenience write \(P_{ij}^{(1)} = P_{ij}\) as the one-step transition probability. In matrix notation we have

\[
P^{(n)} = \begin{bmatrix}
P_{00}^{(n)} & P_{01}^{(n)} & P_{02}^{(n)} & \cdots \\
P_{10}^{(n)} & P_{11}^{(n)} & P_{12}^{(n)} & \cdots \\
P_{20}^{(n)} & P_{21}^{(n)} & P_{22}^{(n)} & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix} \quad (3.4.12)
\]

When \(n = 1\), the matrix \(P = P^{(1)}\) is known as the transition probability matrix.
Note that $0 \leq P_{ij}^{(n)} \leq 1$ and the row sums of $P^{(n)}$ (i.e., $\sum_{j \in S} P_{ij}^{(n)}$) are equal to 1. With these notational simplifications, (3.4.7) can be written as

$$P_{ij}^{(n)} = \sum_{k \in S} P_{ik}^{(r)} P_{kj}^{(n-r)} \quad 0 < r < n$$

or

$$P^{(n)} = P^{(r)} P^{(n-r)}$$

By iterating on the value of $r = 1, 2, \ldots, n$, it is easy to show

$$P^{(n)} = P^n$$  \hspace{1cm} (3.4.13)

showing the $n$-step transition probabilities are given by the elements of the $n$th power of the one-step transition probability matrix.

**Case (ii): Discrete state and continuous parameter space**

As in Case (i) consider time-homogeneous Markov process in which transition probabilities $P_{ij}(s, t)$ and $P_{ij}(s + u, t + u)$ are the same. Without loss of generality, use $s = 0$ and write

$$P_{ij}(t) = P[X(t) = j | X(0) = i]$$  \hspace{1cm} (3.4.14)

In matrix notation the probabilities of transition among states $i, j \in S$ can be given as elements of the matrix

$$P(t) = ||P_{ij}(t)||$$

Because of the continuous nature of the time parameter, we cannot get a result similar to (3.4.13). Instead of the product representation, here we derive differential equations from which $P_{ij}(t)$ can be determined. To start with note the following properties that are either obvious or assumed.

(i) $P_{ij}(t) \geq 0$;

(ii) $\sum_{j \in S} P_{ij}(t) = 1$;
3.4. MARKOV PROCESS

(iii) \( P_{ij}(s + t) = \sum_{k \in S} P_{ik}(s)P_{kj}(t) \);

(iv) \( P_{ij}(t) \) is continuous;

(v) \( \lim_{t \to 0} P_{ij}(t) = 1 \) if \( i = j \), and \( = 0 \), otherwise.

Note that properties (i) (ii) are obvious from the transition structure and (iii) is a restatement of Chapman-Kolmogorov relation. The properties (iv) and (v) are necessary (hence assumed) for deriving the differential equations.

Using Taylor series expansion, we may write

\[
P_{ij}(t + \Delta t) = P_{ij}(t) + \Delta t P'_{ij}(t) + \frac{\Delta t^2}{2} P''_{ij}(t) + \ldots
\]

Setting \( t = 0 \)

\[
P_{ij}(\Delta t) = P_{ij}(0) + \Delta t P'_{ij}(0) + \frac{\Delta t^2}{2} P''_{ij}(0) + \ldots
\]

Rewriting these equations, we have

\[
\frac{P_{ij}(\Delta t)}{\Delta t} = P'_{ij}(0) + \frac{o(\Delta t)}{\Delta t} \quad i \neq j \quad (3.4.15)
\]

\[
\frac{P_{ij}(\Delta t) - 1}{\Delta t} = P'_{ii}(0) + \frac{o(\Delta t)}{\Delta t} \quad (3.4.16)
\]

where \( o(\Delta t) \) is defined such that \( o(\Delta t)/\Delta t \to 0 \) as \( \Delta t \to 0 \). Let \( \Delta t \to 0 \) in Eqs. (3.4.15) and (3.4.16) and denote \( P'_{ij}(0) = \lambda_{ij} \) \((j \neq i)\) and \( P'_{ii}(0) = -\lambda_{ii} \).

Therefore we have

\[
\sum_{j \neq i} \lambda_{ij} - \lambda_{ii} = 0 \quad (3.4.17)
\]

Noting that \( \lambda_{ij} \) are infinitesimal transition rates, which are derivatives at zero as in (3.4.15) and (3.4.16), it is easy to see, that (3.4.17) is the direct consequence of the property \( \sum_{j \in S} P_{ij}(t) = 1 \). These transition rates are also known as generators, displayed in a matrix as

\[
A = \begin{bmatrix}
-\lambda_{00} & \lambda_{01} & \lambda_{12} & \ldots \\
\lambda_{10} & -\lambda_{11} & \lambda_{12} & \ldots \\
\lambda_{20} & \lambda_{21} & -\lambda_{22} & \ldots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix} \quad (3.4.18)
\]
In continuous time Markov processes with discrete states the generator matrix $A$ plays the part of the transition probability matrix $P$ (matrix (3.4.12) with $n = 1$) in the analysis of the process.

In order to derive the differential equations for the determination of $P_{ij}(t)$ we proceed as follows. From Chapman-Kolmogorov relation property (iii) above we have

$$P_{ij}(t + s) = \sum_{k \in S} P_{ik}(t)P_{kj}(s)$$

Set $s = \Delta t$; then

$$P_{ij}(t + \Delta t) = \sum_{k \in S} P_{ik}(t)P_{kj}(\Delta t)$$

Subtracting $P_{ij}(t)$ from both sides of the equation and dividing by $\Delta t$,

$$\frac{P_{ij}(t + \Delta t) - P_{ij}(t)}{\Delta t} = \sum_{k \neq j} \frac{P_{ik}(t)P_{kj}(\Delta t)}{\Delta t} + P_{ij}(t)\frac{P_{jj}(t) - 1}{\Delta t}$$

Let $\Delta t \to 0$; we get

$$P'_{ij}(t) = -\lambda_{jj}P_{ij}(t) + \sum_{k \neq j} \lambda_{kj}P_{ik}(t) \tag{3.4.19}$$

In deriving (3.4.19) we have used the definition of $\lambda_{ij}$ given in (3.4.15) and (3.4.16). Eqs. (3.4.19) for $i, j \in S$ are known as forward Kolmogorov equations.

In matrix notation we can write them as

$$P'(t) = P(t)A. \tag{3.4.20}$$

The transition probability $P_{ij}(t)$ can be determined by solving these differential equations along with the boundary condition $P(0) = I$.

Backward Kolmogorov equations can be obtained in a similar manner, by starting with the relation

$$P_{ij}(\Delta t + t) = \sum_{k \in S} P_{ik}(\Delta t)P_{kj}(t)$$
The corresponding matrix equation can be given as

\[ P'(t) = AP(t) \]

Formally, the solution for both sets of equations can be given as

\[ P(t) = e^{At} = I + \sum_{n=1}^{\infty} \frac{A^n t^n}{n!} \]  \hspace{1cm} (3.4.21)

If a stochastic process in a queueing system can be represented as a Markov process with discrete states and continuous parameter space, the analysis starts with deriving the appropriate \( A \) (generator) matrix and the corresponding forward Kolmogorov equations. This procedure will be elaborated when we take up specific queueing systems.

### 3.4.1 The Poisson process

In Chapter 2 we have introduced events whose inter occurrence times are exponential. We have also listed the following properties:

1. Events occurring in nonoverlapping intervals of time are independent of each other.

2. There is a constant \( \lambda \) such that the probabilities of occurrence of events in a small interval of length \( \Delta t \) are given as follows:
   
   a. \( P\{\text{Number of events occurring in } (t, t + \Delta t] = 0\} = 1 - \lambda \Delta t + o(\Delta t) \)
   
   b. \( P\{\text{Number of events occurring in } (t, t + \Delta t] = 1\} = \lambda \Delta t + o(\Delta t) \)
   
   c. \( P\{\text{Number of events occurring in } (t, t + \Delta t] > 1\} = o(\Delta t) \)

where \( o(\Delta t) \) is such that \( o(\Delta t)/\Delta t \to 0 \) as \( \Delta t \to 0 \).

Using the notations and equations developed for Markov processes, in this context, we have

\[ P'_{ij}(0) = \lambda; \quad P'_{ii}(0) = -\lambda \]
resulting in a generator matrix

$$A = \begin{bmatrix}
-\lambda & \lambda & 0 & 0 & 0 & \ldots \\
0 & -\lambda & \lambda & 0 & 0 & \ldots \\
0 & 0 & -\lambda & \lambda & 0 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix} \quad (3.4.22)$$

The Poisson process is a counting process whose initial value is 0, i.e. $X(0) = 0$. Writing $P_{0n}(t) = P_n(t)$, for convenience, and noting that $\mathbf{P}(t) = (P_0(t), P_1(t), \ldots)$ and $\mathbf{P}'(t) = (P'_0(t), P'_1(t), \ldots)$ the individual equations in (3.4.19) can be written out as

$$P'_0(t) = -\lambda P_0(t)$$
$$P'_n(t) = -\lambda P_n(t) + \lambda P_{n-1}(t) \quad n > 0$$

with $P_0(0) = 1$ and $P_n(0) = 0$ for $n > 0$. Solving these differential equations we get

$$P_n(t) = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \quad n = 0, 1, 2, \ldots \quad (3.4.23)$$

As we see in later chapters, solutions to equations in (3.4.19) are not easily determined. In the case of the Poisson process because of the bi-diagonal structure of $A$ and the constant element $\lambda$ the differential equations could be solved using standard methods. When such simplifications are not available, we have to use Laplace transforms and probability generating functions in their solutions. (see Bailey (1964) and Bhat and Miller (2002).)

### 3.4.2 Sojourn time in a state

In the modeling of queueing systems, it helps to understand what the elements of matrix $A$ of (3.4.18) stand for. As indicated earlier, by definition (see (3.4.15) and (3.4.16)), $\lambda_{ij}$, $j \neq i$, is the instantaneous rate for the transition $i \rightarrow j$. From (3.4.17) we also know that $\sum_{j \neq i} \lambda_{ij} = \lambda_{ii}$. That means $\lambda_{ii}$ is the sum of all the instantaneous transition rates out of state $i$. 
This allows us to interpret $1/\lambda_{ii}$ as the mean length of time the process stays in state $i$ during a visit. The length of time the process stays in a state during a visit is known as the *sojourn time* in that state. We can show that the sojourn time of the process in state $i$ has an exponential distribution with mean $1/\lambda_{ii}$.

Let $Z_i$ be the sojourn time of the Markov process in state $i$ (the time the process resides in a state before moving out of it) and let

$$G_i(x) = P(Z_i > x).$$

Using the properties of transition in a Markov process for an increment $\Delta x$, we write

$$G_i(x + \Delta x) = G_i(x)[P_{ii}(\Delta x)] + o(\Delta x)$$

$$\frac{G_i(x + \Delta x) - G_i(x)}{\Delta x} = G_i(x) \frac{[P_{ii}(\Delta x) - 1]}{\Delta x} + \frac{o(\Delta x)}{\Delta x}.$$ 

Let $\Delta x \to 0$. Then

$$G_i'(x) = -\lambda_{ii}G_i(x)$$

where we have used the definition of $\lambda_{ii}$ given in (3.4.16). Now

$$\frac{d\ln G_i(x)}{dx} = \frac{G_i'(x)}{G_i(x)} = -\lambda_{ii}$$

from (3.4.25)

giving

$$\ln G_i(x) = -\lambda_{ii}x$$

$$G_i(x) = e^{-\lambda_{ii}x}$$

Hence

$$P(Z_i \leq x) = 1 - e^{-\lambda_{ii}x}$$

which is an exponential distribution with mean $1/\lambda_{ii}$.

For a discussion of special forms of Markov processes used in stochastic modeling the readers are referred to Bhat and Miller (2002) and advanced books cited in that text.
3.5 Classification of States

In order to describe a stochastic process we need to specify the state space and the parameter space. The parameter space is easily categorized as being discrete or continuous. The state space, however, in addition to begin discrete or continuous, may include states or groups of states with special properties.

The states of a discrete state stochastic process fall into groups depending on how they interact with each other. The basic property defining this interaction is communication. If state \(i\) can be reached from state \(j\) in a finite number of steps, then \(i\) is said to be accessible from \(j\). If \(i\) and \(j\) are accessible to each other, they are said to communicate. Now it is not hard to visualize all communicating states forming a single group, known as an equivalence class. If a Markov chain has all its states belonging to a single equivalence class, it is said to be irreducible.

For instance consider the number of customers, \(Q_n\), in a queueing system, at discrete time points \(t_n, n = 0, 1, 2, \ldots\). Assume that \(t_n\) are such that \(\{Q_n, n = 0, 1, 2 \ldots\}\) can be modeled as a Markov chain. When no restrictions are imposed on the transitions of \(\{Q_n\}\) it is easy to note that all states of the Markov chain communicate with each other and hence form a single equivalence class. Alternatively we may think of a finite queueing system which ceases to operate when \(Q_n\) hits the value, say \(M\).

**Example 3.5.1**

\(M\) machines are in operation in a service facility. The facility ceases its operation when all machines become inoperative. Let the number of failed machines be the state of the process. Now the state \(M\) of the Markov chain is accessible from all other states \([0, 1, 2, \ldots, M - 1]\); but other states are not accessible from \(M\). Then we have two equivalence classes: \([M]\), and \([0, 1, 2, \ldots, M - 1]\). Since the process stops in \(M\), it is known as an absorbing state.
Example 3.5.2

Suppose now the system is modified such that the facility is not shut down when all $M$ machines are inoperative. One or more of them are repaired to bring the facility back into operation. Now all states $[0, 1, 2, \ldots, M]$ belong to the same class. Comparing the states belonging to the equivalence class $[0, 1, 2, \ldots, M]$ of Example 3.5.1 and the equivalence class $[0, 1, 2, \ldots, M]$ of Example 3.5.2 we can make the following observation. The Markov chain starting from anyone of the states in the class $[0, 1, \ldots, M - 1]$ from Example 3.5.1 will not remain in any of these states when $n \to \infty$, because at some stage, it is bound to get absorbed in $M$. On the other hand the Markov chain of Example 3.5.2 will remain in the class even when $n \to \infty$. This behavior of the Markov chain allows us to classify the states, and the equivalence classes themselves into being recurrent or transient.

(1) Starting from state $i$, if the Markov chain is certain to return to $i$, the state is said to be recurrent. Since all states in the equivalence class communicate with each other, the class itself is recurrent. A further classification is made based on the value of recurrence time, which is the mean time the process takes to return to the same state. If the recurrence time is finite, the state (and the class to which it belongs) is known as positive recurrent. If it is infinite the state and the class are known as null recurrent. Note that an absorbing state is recurrent.

(2) Starting from state $i$, if the Markov chain’s return to that state is not certain it is said to be transient. Since all states in the equivalence class communicate with each other, then the class itself is transient.

The classification of states of a stochastic process such as queue length (number of customers in the system) plays a major role in understanding its behavior. We give below some of the properties that can be deduced from the nature of the states of the process.
1. If there are transient states in the state space of the process, in the long run \((n \to \infty)\), the process will not be found in those states. Thus, if there are transient as well as recurrent states in the state space, the process will always end up in the recurrent states.

2. A process starting out in a recurrent state \(i\) will always remain in the recurrent equivalence class to which state \(i\) belongs.

3. Because of Properties 1 and 2 above, only processes with irreducible Markov chain models need to be considered to understand the long run behavior of the system. Later we shall see that we can establish conditions under which limiting distributions exist for such processes.

4. When the state space includes both transient and recurrent states, one of the characteristics of interest is the transition from the transient states to a state in the recurrent class. For instance the distribution properties of the busy period in a queueing system can be determined by considering 0 as an absorbing state for the queue length process, while all other states are transient.

For an elaboration on the classification of states and their usefulness in stochastic modeling readers are referred to Bhat and Miller (2002).

### 3.6 Phase-type Distributions

In Chapter 2 we postponed the description of a phase-type (PH-distribution) because it required results from Markov processes. For illustration we use the generalized Erlang distribution, one of the simpler PH-distributions, given by the Laplace transform \(\psi(\theta)\) of Eq. (2.1.22).

\[
\psi(\theta) = \prod_{i=1}^{k} \left( \frac{\lambda_i}{\theta + \lambda_i} \right) 
\]  

(3.6.1)
3.6. PHASE-TYPE DISTRIBUTIONS

Generalized Erlang can be generated as the distribution of the total time a process takes to traverse \( k \) phases, with phase \( i \) lasting a duration that has an exponential distribution with mean \( 1/\lambda_i \). Recalling the properties of the Markov process, we can identify it as a Markov process with states \{1, 2, \ldots, k, k + 1\} of which state \( k + 1 \) is absorbing. The generator matrix of the process can be given as

\[
A = \begin{pmatrix}
  -\lambda_1 & \lambda_1 & 0 & \cdots & 0 \\
  -\lambda_2 & \lambda_2 & 0 & \cdots & 0 \\
  -\lambda_3 & \lambda_3 & 0 & \cdots & 0 \\
    &    &    &    &    \\
  0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}
\]

(3.6.2)

where \( T (m \times m) \) and \( T^0 (m \times 1) \) are submatrices.

Let \( P_i(t) \) be the probability that the process is in state \( i \) at time \( t \). We should note that, because \( k + 1 \) is an absorbing state ultimately the process will come to reside in that state. Therefore \( \sum_{i=1}^{k} P_i(t) \) is the probability that the process is in one of the transient states \{1, 2, \ldots, k\} at time \( t \). Let \( Y_k \) be the time the process takes to traverse all the \( k \) states. Then

\[
P(Y_k > t) = \sum_{i=1}^{k} P_i(t)
\]

Hence

\[
P(Y_k \leq t) = 1 - \sum_{i=1}^{k} P_i(t) = P_{k+1}(t)
\]

(3.6.3)

For the Markov process with generator matrix \( A \) we can write down the forward Kolmogorov equations as

\[
P'_i(t) = -\lambda_1 P_i(t)
\]
\[ P'_i(t) = -\lambda_i P_i(t) + \lambda_{i-1} P_i(t) \quad 1 < i \leq k \]
\[ P'_{k+1}(t) = \lambda_k P_k(t) \]  \hspace{1cm} (3.6.4)

with \( P_1(0) = 1 \) and \( P_i(0) = 0 \) for \( i > 0 \). Define the Laplace transform
\[ \phi_i(\theta) = \int_0^\infty e^{-\theta t} P_i(t) dt \]

Taking transforms of equations (3.6.4) we get
\[ -1 + \phi_1(\theta) = -\lambda_1 \phi_1(\theta) \]
\[ \theta \phi_i(\theta) = -\lambda_i \phi_i(\theta) + \lambda_{i-1} \phi_{i-1}(\theta) \quad 1 < i \leq k \]

Solving these equations recursively we get
\[ \phi_k(\theta) = \left( \frac{1}{\theta + \lambda_k} \right) \Pi_{i=1}^{k-1} \left( \frac{\lambda_i}{\theta + \lambda_i} \right) \]  \hspace{1cm} (3.6.5)

If \( f_k(y) \) is the probability density of \( Y_k \), it is easy to note that \( P'_{k+1}(t) \) is in fact \( f_k(y) \). Thus, from the last equation in (3.6.4) and Eq. (3.6.5) we have, the Laplace transform of \( Y_k \) as,
\[ \int_0^\infty e^{-\theta y} f_k(y) dy = \Pi_{i=1}^{k} \left( \frac{\lambda_i}{\theta + \lambda_i} \right) \]
which is the same as (3.6.1).

Referring back to the general form of the solution to the forward Kolmorov equations given by (3.4.20), the distribution function of \( Y_k \) given by (3.6.3) can be given as
\[ F(t) = 1 - \alpha \exp(Tt) e' \text{ for } t \geq 0 \]  \hspace{1cm} (3.6.6)
where \( \alpha = (1,0,0,\ldots,0) \) and \( e' = (1,1,\ldots,1) \).

Generalizing this structure, Neuts (1981) has defined the PH-distribution as the time until absorption in a finite Markov process of the type with generator
\[
A = \begin{bmatrix}
T & T^0 \\
0 & 0
\end{bmatrix}
\]
where the $m \times m$ matrix $T$ satisfies $T_{ii} < 0$ for $1 \leq i \leq k$, and $T_{ij} \geq 0$ for $i \neq j$. Also $Te + T^0 = 0$. The initial probability vector $\alpha, \alpha_{k+1}$ is such that $\alpha e + \alpha_{k+1} = 1$. States $\{1, 2, \ldots, k\}$ are transient and $k + 1$ is absorbing. A large number of PH-distributions can be generated by using different structures for $T$. For the properties of PH-distribution and its use in queueing theory the readers are referred to Neuts (1981, 1989).

References


