

Chapter 4

Continuous-Time Markov Models

4.1 Continuous-Time Markov Chains

In this chapter, we consider a system that is observed continuously, with $X(t)$ being the state at time t , $t \geq 0$. Following the definition of DTMCs, we define *continuous-time Markov chains* (CTMCs) below.

Definition 4.1. (Continuous-Time Markov Chain (CTMC)). A stochastic process $\{X(t), t \geq 0\}$ on state space S is called a CTMC if, for all i and j in S and $t, s \geq 0$,

$$P(X(s+t) = j | X(s) = i, X(u), 0 \leq u \leq s) = P(X(s+t) = j | X(s) = i). \quad (4.1)$$

The CTMC $\{X(t), t \geq 0\}$ is said to be time homogeneous if, for $t, s \geq 0$,

$$P(X(s+t) = j | X(s) = i) = P(X(t) = j | X(0) = i). \quad (4.2)$$

Equation (4.1) implies that the evolution of a CTMC from a fixed time s onward depends on $X(u)$, $0 \leq u \leq s$ (the history of the CTMC up to time s) only via $X(s)$ (the state of the CTMC at time s). Furthermore, for time-homogeneous CTMCs, it is also independent of s ! In this chapter, we shall assume, unless otherwise mentioned, that all CTMCs are time homogeneous and have a finite state space $\{1, 2, \dots, N\}$. For such CTMCs, we define

$$p_{i,j}(t) = P(X(t) = j | X(0) = i), \quad 1 \leq i, j \leq N. \quad (4.3)$$

The N^2 entities $p_{i,j}(t)$ are arranged in matrix form as follows:

$$P(t) = \begin{bmatrix} p_{1,1}(t) & p_{1,2}(t) & p_{1,3}(t) & \cdots & p_{1,N}(t) \\ p_{2,1}(t) & p_{2,2}(t) & p_{2,3}(t) & \cdots & p_{2,N}(t) \\ p_{3,1}(t) & p_{3,2}(t) & p_{3,3}(t) & \cdots & p_{3,N}(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{N,1}(t) & p_{N,2}(t) & p_{N,3}(t) & \cdots & p_{N,N}(t) \end{bmatrix}. \quad (4.4)$$

The matrix above is called the *transition probability matrix* of the CTMC $\{X(t), t \geq 0\}$. Note that $P(t)$ is a matrix of functions of t ; i.e., it has to be specified for each t . The following theorem gives the important properties of the transition probability matrix $P(t)$.

Theorem 4.1. (*Properties of $P(t)$*). A transition probability matrix $P(t) = [p_{i,j}(t)]$ of a time-homogeneous CTMC on state space $S = \{1, 2, \dots, N\}$ satisfies the following:

$$1. \quad p_{i,j}(t) \geq 0, \quad 1 \leq i, j \leq N; t \geq 0, \quad (4.5)$$

$$2. \quad \sum_{j=1}^N p_{i,j}(t) = 1, \quad 1 \leq i \leq N; t \geq 0, \quad (4.6)$$

$$3. \quad p_{i,j}(s+t) = \sum_{k=1}^N p_{i,k}(s)p_{k,j}(t) = \sum_{k=1}^N p_{i,k}(t)p_{k,j}(s), \quad 1 \leq i \leq N; t, s \geq 0. \quad (4.7)$$

Proof. (1) The nonnegativity of $p_{i,j}(t)$ follows because it is a (conditional) probability.

(2) To prove the second assertion, we have

$$\begin{aligned} \sum_{j=1}^N p_{i,j}(t) &= \sum_{j=1}^N \mathbf{P}(X(t) = j | X(0) = i) \\ &= \mathbf{P}(X(t) \in S | X(0) = i). \end{aligned}$$

Since $X(t)$ must take some value in the state space S , regardless of the value of $X(0)$, it follows that the last quantity is 1.

(3) To prove (4.7), we have

$$\begin{aligned} p_{i,j}(s+t) &= \mathbf{P}(X(s+t) = j | X(0) = i) \\ &= \sum_{k=1}^N \mathbf{P}(X(s+t) = j | X(s) = k, X(0) = i) \mathbf{P}(X(s) = k | X(0) = i) \\ &= \sum_{k=1}^N \mathbf{P}(X(s+t) = j | X(s) = k) p_{i,k}(s) \\ &\quad \text{(due to the Markov property at } s) \end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^N \mathbf{P}(X(t) = j | X(0) = k) p_{i,k}(s) \\
&\quad \text{(due to time homogeneity)} \\
&= \sum_{k=1}^N p_{i,k}(s) p_{k,j}(t).
\end{aligned}$$

The other part of (4.7) follows by interchanging s and t . This proves the theorem. $\blacksquare$

Equation (4.7) states the *Chapman–Kolmogorov equations* for CTMCs and can be written in matrix form as

$$P(t + s) = P(s)P(t) = P(t)P(s). \quad (4.8)$$

Thus the transition probability matrices $P(s)$ and $P(t)$ commute! Since AB is generally not the same as BA for square matrices A and B , the property above implies that the transition probability matrices are very special indeed!

Example 4.1. (Two-State CTMC). Suppose the lifetime of a high-altitude satellite is an $\text{Exp}(\mu)$ random variable. Once it fails, it stays failed forever since no repair is possible. Let $X(t)$ be 1 if the satellite is operational at time t and 0 otherwise.

We first show that $\{X(t), t \geq 0\}$ is a CTMC. The state space is $\{0, 1\}$. We directly check (4.1) and (4.2). Consider the case $i = j = 1$. Let the lifetime of the satellite be T . Then $X(s) = 1$ if and only if $T > s$, and if $X(s) = 1$ then $X(u)$ must be 1 for $0 \leq u \leq s$. Hence, using the memoryless property (see Equation (3.5)) of the exponential variable T , we have

$$\begin{aligned}
\mathbf{P}(X(s+t) = 1 | X(s) = 1, X(u) : 0 \leq u \leq s) \\
&= \mathbf{P}(T > s+t | T > s) = e^{-\mu t} \\
&= \mathbf{P}(X(t) = 1 | X(0) = 1).
\end{aligned}$$

Performing similar calculations for all pairs i, j , we establish that $\{X(t), t \geq 0\}$ is a CTMC. Next we compute its transition probability matrix. We have

$$\begin{aligned}
p_{0,0}(t) &= \mathbf{P}(\text{satellite is down at } t | \text{satellite is down at } 0) = 1, \\
p_{1,1}(t) &= \mathbf{P}(\text{satellite is up at } t | \text{satellite is up at } 0) = \mathbf{P}(T > t) = e^{-\mu t}.
\end{aligned}$$

Hence the transition probability matrix is given by

$$P(t) = \begin{bmatrix} 1 & 0 \\ 1 - e^{-\mu t} & e^{-\mu t} \end{bmatrix}. \quad (4.9)$$

We can check by direct calculations that (4.8) are satisfied. $\blacksquare$

In Chapter 2, we described a DTMC by giving its one-step transition probability matrix. How do we describe a CTMC? Giving $P(t)$ for each t is too complicated for most CTMCs. We need a simpler method of describing a CTMC. We develop such a method below based on the insight provided by the example above.

The discussion above implies that the stochastic process $\{X(t), t \geq 0\}$ of Example 4.1 can be a CTMC if and only if the lifetime of the satellite has the memoryless property. We have seen in Theorem 3.1 that the only continuous random variable with the memoryless property is the exponential random variable. This suggests that we should build CTMCs using exponential random variables. This is what we do next.

Let $X(t)$ be the state of a system at time t . Suppose the state space of the stochastic process $\{X(t), t \geq 0\}$ is $\{1, 2, \dots, N\}$. The random evolution of the system occurs as follows. Suppose the system starts in state i . It stays there for an $\text{Exp}(r_i)$ amount of time, called the *sojourn time in state i* . At the end of the sojourn time in state i , the system makes a sudden transition to state $j \neq i$ with probability $p_{i,j}$, independent of how long the system has been in state i . Once in state j , it stays there for an $\text{Exp}(r_j)$ amount of time and then moves to a new state $k \neq j$ with probability $p_{j,k}$, independently of the history of the system so far. And it continues this way forever. A typical sample path of the system is shown in Figure 4.1.

Three remarks are in order at this time. First, the jump probabilities $p_{i,j}$ are not to be confused with the transition probabilities $p_{i,j}(t)$ of (4.3). Think of $p_{i,j}$ as the probability that the system moves to state j when it moves out of state i . Second, $p_{i,i} = 0$. This is because, by definition, the sojourn time in state i is the time the system spends in state i until it moves out of it. Hence, a transition from i to i is not allowed. (Contrast this with the DTMC.) Third, in case state i is absorbing (i.e., the system stays in state i forever once it gets there), we set $r_i = 0$, interpreting an $\text{Exp}(0)$ random variable as taking a value equal to ∞ with probability 1. In such a case, $p_{i,j}$'s have no meaning and can be left undefined.

Theorem 4.2. *The stochastic process $\{X(t), t \geq 0\}$ with parameters $r_i, 1 \leq i \leq N$, and $p_{i,j}, 1 \leq i, j \leq N$, as described above is a CTMC.*

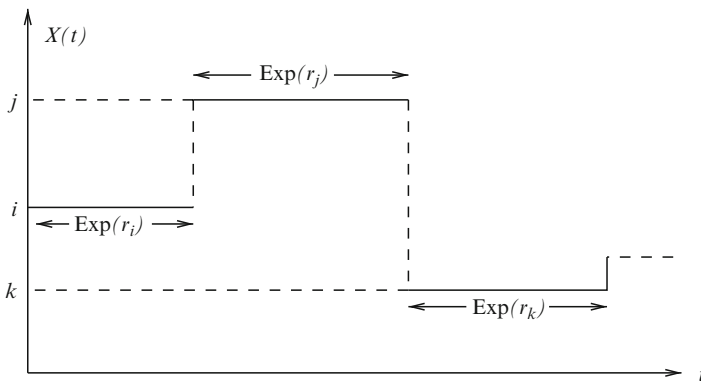


Fig. 4.1 A typical sample path of a CTMC.

Idea of Proof. The rigorous proof of this theorem is tedious, but the idea is simple. Let $s, t \geq 0$ be fixed. Suppose we are given the history of the system up to s , i.e., we are given $X(u), 0 \leq u \leq s$. Also, suppose $X(s) = i$. Then, using the memoryless property of the exponential random variable, we see that the remaining sojourn time of the system in state i from time s onward is again an $\text{Exp}(r_i)$ random variable, independent of the past. At the end of the sojourn time, the system will move to state j with probability $p_{i,j}$. The subsequent sojourn times and the transitions are independent of the history by construction. Hence the evolution of the system from time s onward depends only on i , the state of the system at time s . Hence $\{X(t), t \geq 0\}$ is a CTMC. ■

What is even more important is that all CTMCs on finite state spaces that have nonzero sojourn times in each state can be described this way. We shall not prove this fact here. In all our applications, this condition will be met and hence we can describe a CTMC by giving the parameters $\{r_i, 1 \leq i \leq N\}$ and $\{p_{i,j}, 1 \leq i, j \leq N\}$. This is a much simpler description than giving the transition probability matrix $P(t)$ for all $t \geq 0$.

Analogously to DTMCs, a CTMC can also be represented graphically by means of a directed graph as follows. The directed graph has one node (or vertex) for each state. There is a directed arc from node i to node j if $p_{i,j} > 0$. The quantity $r_{i,j} = r_i p_{i,j}$, called the *transition rate from i to j* , is written next to this arc. Note that there are no self-loops (arcs from i to i) in this representation. This graphical representation is called the *rate diagram* of the CTMC.

Rate diagrams are useful visual representations. We can understand the dynamics of the CTMC by visualizing a particle that moves from node to node in the rate diagram as follows: it stays on node i for an $\text{Exp}(r_i)$ amount of time and then chooses one of the outgoing arcs from node i with probabilities proportional to the rates on the arcs and moves to the node at the other end of the arc. This motion continues forever. The node occupied by the particle at time t is the state of the CTMC at time t .

Note that we can recover the original parameters $\{r_i, 1 \leq i \leq N\}$ and $\{p_{i,j}, 1 \leq i, j \leq N\}$ from the rates $\{r_{i,j}, 1 \leq i, j \leq N\}$ shown in the rate diagram as follows:

$$r_i = \sum_{j=1}^N r_{i,j},$$

$$p_{i,j} = \frac{r_{i,j}}{r_i} \text{ if } r_i \neq 0.$$

It is convenient to put all the rates $r_{i,j}$ in matrix form as follows:

$$R = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & \cdots & r_{1,N} \\ r_{2,1} & r_{2,2} & r_{2,3} & \cdots & r_{2,N} \\ r_{3,1} & r_{3,2} & r_{3,3} & \cdots & r_{3,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{N,1} & r_{N,2} & r_{N,3} & \cdots & r_{N,N} \end{bmatrix}. \quad (4.10)$$

Note that $r_{i,i} = 0$ for all $1 \leq i \leq N$, and hence the diagonal entries in R are always zero. R is called the *rate matrix* of the CTMC. It is closely related to $Q = [q_{i,j}]$, called the *generator matrix* of the CTMC, which is defined as

$$q_{i,j} = \begin{cases} -r_i & \text{if } i = j, \\ r_{i,j} & \text{if } i \neq j. \end{cases} \quad (4.11)$$

Thus the generator matrix Q is the same as the rate matrix R with the diagonal elements replaced by $-r_i$'s. It is common in the literature to describe a CTMC by the Q matrix. However, we shall describe it by the R matrix. Clearly, one can be obtained from the other.

Example 4.2. (Two-State CTMC). Consider the CTMC of Example 4.1. The system stays in state 1 (satellite up) for an $\text{Exp}(\mu)$ amount of time and then moves to state 0. Hence, $r_1 = \mu$ and $p_{1,0} = 1$. Once the satellite fails (state 0), it stays failed forever. Thus state 0 is absorbing. Hence $r_0 = 0$, and $p_{0,1}$ is left undefined. The rate matrix is

$$R = \begin{bmatrix} 0 & 0 \\ \mu & 0 \end{bmatrix}. \quad (4.12)$$

The generator matrix is given by

$$Q = \begin{bmatrix} 0 & 0 \\ \mu & -\mu \end{bmatrix}. \quad (4.13)$$

The rate diagram is shown in Figure 4.2. ■

Example 4.3. (Two-State Machine). Consider a machine that operates for an $\text{Exp}(\mu)$ amount of time and then fails. Once it fails, it gets repaired. The repair time is an $\text{Exp}(\lambda)$ random variable and is independent of the past. The machine is as good as new after the repair is complete. Let $X(t)$ be the state of the machine at time t , 1 if it is up and 0 if it is down. Model this as a CTMC.

The state space of $\{X(t), t \geq 0\}$ is $\{0, 1\}$. The sojourn time in state 0 is the repair time, which is given as an $\text{Exp}(\lambda)$ random variable. Hence $r_0 = \lambda$. After the repair is complete, the machine is up with probability 1; hence $p_{0,1} = 1$. Similarly, $r_1 = \mu$ and $p_{1,0} = 1$. The independence assumptions are satisfied, thus making $\{X(t), t \geq 0\}$ a CTMC. The rate matrix is

$$R = \begin{bmatrix} 0 & \lambda \\ \mu & 0 \end{bmatrix}. \quad (4.14)$$

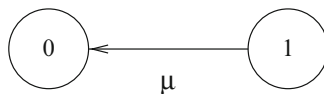
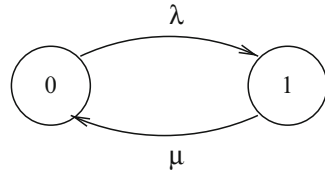


Fig. 4.2 Rate diagram of the satellite system.

Fig. 4.3 Rate diagram of the two-state machine.



The generator matrix is given by

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}. \quad (4.15)$$

The rate diagram is shown in Figure 4.3. Note that describing this CTMC by giving its $P(t)$ matrix would be very difficult since the machine can undergo many failures and repairs up to time t . ■

In the next section, we use the properties of exponential random variables studied in Chapter 3 and develop CTMC models of real-life systems. We give the rate matrix R for the CTMCs and the rate diagrams but omit the generator matrix Q .

4.2 Examples of CTMCs: I

In this section, we consider CTMC models that are built using the exponential random variables and their properties as studied in Chapter 3.

Example 4.4. (Two-Machine Workshop). Consider a workshop with two independent machines, each with its own repair person and each machine behaving as described in Example 4.3. Let $X(t)$ be the number of machines operating at time t . Model $\{X(t), t \geq 0\}$ as a CTMC.

The state space of $\{X(t), t \geq 0\}$ is $\{0, 1, 2\}$. Consider state 0. In this state, both machines are down and hence under repair. Due to the memoryless property, the remaining repair times of the two machines are iid $\text{Exp}(\lambda)$ random variables. The system moves from state 0 to state 1 as soon as one of the two machines finishes service. Hence the sojourn time in state 0 is a minimum of two iid $\text{Exp}(\lambda)$ random variables. Furthermore, by Theorem 3.3, the sojourn time is an $\text{Exp}(2\lambda)$ random variable. Hence $r_0 = 2\lambda$ and $p_{0,1} = 1$. The transition rate from state 0 to state 1 is thus $r_{0,1} = 2\lambda$.

Now consider state 1, where one machine is up and one is down. Let T_1 be the remaining repair time of the down machine, and let T_2 be the remaining life time of the up machine. Again from the memoryless property of the exponential random variables, T_1 is $\text{Exp}(\lambda)$ and T_2 is $\text{Exp}(\mu)$, and they are independent. The sojourn time in state 1 is $\min(T_1, T_2)$, which is an $\text{Exp}(\lambda + \mu)$ random variable. Hence $r_1 = \lambda + \mu$. The system moves to state 2 if $T_1 < T_2$, which has probability $\lambda/(\lambda + \mu)$ and it moves to state 0 if $T_2 < T_1$, which has probability $\mu/(\lambda + \mu)$.

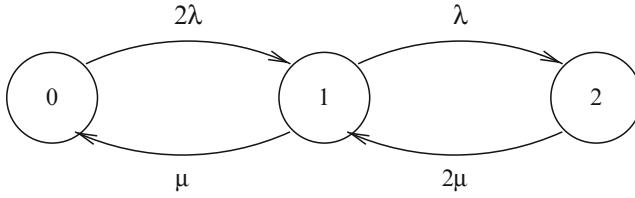


Fig. 4.4 Rate diagram of the two-machine workshop.

Note that the next state is independent of the sojourn time from Theorem 3.4. Hence $p_{1,2} = \lambda/(\lambda + \mu)$ and $p_{1,0} = \mu/(\lambda + \mu)$. The transition rates are $r_{1,2} = \lambda$ and $r_{1,0} = \mu$.

A similar analysis of state 2 yields $r_2 = 2\mu$, $p_{2,1} = 1$, and $r_{2,1} = 2\mu$. Thus $\{X(t), t \geq 0\}$ is a CTMC with rate matrix

$$R = \begin{bmatrix} 0 & 2\lambda & 0 \\ \mu & 0 & \lambda \\ 0 & 2\mu & 0 \end{bmatrix}. \quad (4.16)$$

The rate diagram is as shown in Figure 4.4. ■

The example above produces the following extremely useful insight. Suppose the state of the system at time t is $X(t) = i$. Suppose there are m different events that can trigger a transition in the system state from state i to state j . Suppose the occurrence time of the k th event ($1 \leq k \leq m$) is an $\text{Exp}(\lambda_k)$ random variable. Then, assuming the m events are independent of each other and of the history of the system up to time t , the transition rate from state i to state j is given by

$$r_{i,j} = \lambda_1 + \lambda_2 + \cdots + \lambda_m.$$

If the transition rates can be computed this way for all i and j in the state space, then $\{X(t), t \geq 0\}$ is a CTMC with transition rates $r_{i,j}$. We use this observation in the remaining examples.

Example 4.5. (General Machine Shop). A machine shop consists of N machines and M repair persons. ($M \leq N$.) The machines are identical, and the lifetimes of the machines are independent $\text{Exp}(\mu)$ random variables. When the machines fail, they are serviced in the order of failure by the M repair persons. Each failed machine needs one and only one repair person, and the repair times are independent $\text{Exp}(\lambda)$ random variables. A repaired machine behaves like a new machine. Let $X(t)$ be the number of machines that are functioning at time t . Model $\{X(t), t \geq 0\}$ as a CTMC.

We shall explain with a special case: four machines and two repair persons. The general case is similar. For the special case, the state space of $\{X(t), t \geq 0\}$ is $S = \{0, 1, 2, 3, 4\}$. We obtain the transition rates below. In state 0, all machines are down, and two are under repair. (The other two are waiting for a repair person.) The repair times are iid $\text{Exp}(\lambda)$, and either of the two repair completions will move the system

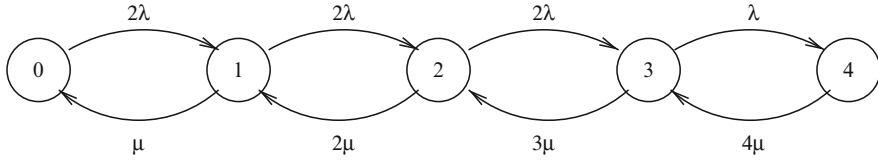


Fig. 4.5 Rate diagram of the general machine shop with four machines and two repair persons.

to state 1. Hence $r_{0,1} = \lambda + \lambda = 2\lambda$. In state 1, one machine is up and three are down, two of which are under repair. Either repair completion will send the system to state 2. Hence $r_{1,2} = 2\lambda$. The failure of the functioning machine, which will happen after an $\text{Exp}(\mu)$ amount of time, will trigger a transition to state 0. Hence $r_{1,0} = \mu$. Similarly, for state 2, we get $r_{2,3} = 2\lambda$ and $r_{2,1} = 2\mu$. In state 3, three machines are functioning and one is down and under repair. (One repair person is idle.) Thus there are three independent failure events that will trigger a transition to state 2, and a repair completion will trigger a transition to state 4. Hence $r_{3,2} = 3\mu$ and $r_{3,4} = \lambda$. Similarly, for state 4, we get $r_{4,3} = 4\mu$.

Thus $\{X(t), t \geq 0\}$ is a CTMC with a rate matrix as follows:

$$R = \begin{bmatrix} 0 & 2\lambda & 0 & 0 & 0 \\ \mu & 0 & 2\lambda & 0 & 0 \\ 0 & 2\mu & 0 & 2\lambda & 0 \\ 0 & 0 & 3\mu & 0 & \lambda \\ 0 & 0 & 0 & 4\mu & 0 \end{bmatrix}. \quad (4.17)$$

The rate diagram is as shown in Figure 4.5. ■

Example 4.6. (Airplane Reliability). A commercial jet airplane has four engines, two on each wing. Each engine lasts for a random amount of time that is an exponential random variable with parameter λ and then fails. If the failure takes place in flight, there can be no repair. The airplane needs at least one engine on each wing to function properly in order to fly safely. Model this system so that we can predict the probability of a trouble-free flight.

Let $X_L(t)$ be the number of functioning engines on the left wing at time t , and let $X_R(t)$ be the number of functioning engines on the right wing at time t . The state of the system at time t is given by $X(t) = (X_L(t), X_R(t))$. Assuming the engine failures are independent of each other, we see that $\{X(t), t \geq 0\}$ is a CTMC on state space

$$S = \{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2), (2, 0), (2, 1), (2, 2)\}.$$

Note that the flight continues in a trouble-free fashion as long as $X_L(t) \geq 1$ and $X_R(t) \geq 1$; i.e., as long as the CTMC is in states $(1, 1), (1, 2), (2, 1), (2, 2)$. The flight crashes as soon as the CTMC visits any state in $\{(0, 0), (0, 1), (0, 2), (1, 0), (2, 0)\}$. We assume that the CTMC continues to evolve even after the airplane crashes!

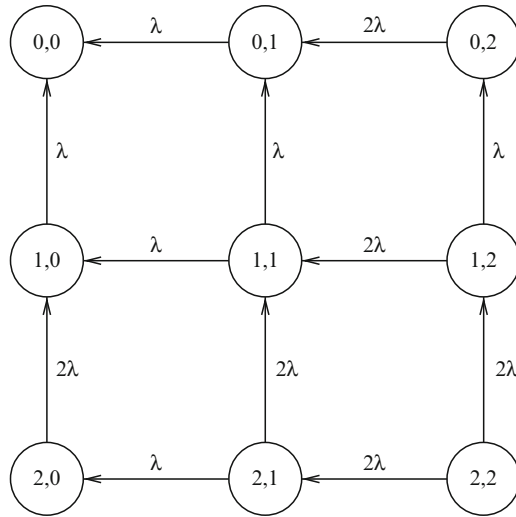


Fig. 4.6 Rate diagram of the airplane reliability system.

Label the states in S 1 through 9 in the order in which they are listed above. Performing the standard triggering event analysis (assuming that no repairs are possible), we get the following rate matrix:

$$R = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2\lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 2\lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2\lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\lambda & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\lambda & 0 & 2\lambda & 0 \end{bmatrix}. \quad (4.18)$$

The rate diagram is as shown in Figure 4.6. ■

4.3 Examples of CTMCs: II

In this section, we consider CTMC models that are built using the Poisson processes and their properties as studied in Chapter 3.

Example 4.7. (Finite-Capacity Single Server Queue). Customers arrive at an automatic teller machine (ATM) according to a $PP(\lambda)$. The space in front of the ATM

can accommodate at most K customers. Thus, if there are K customers waiting at the ATM and a new customer arrives, he or she simply walks away and is lost forever. The customers form a single line and use the ATM in a first-come, first-served fashion. The processing times at the ATM for the customers are iid $\text{Exp}(\mu)$ random variables. Let $X(t)$ be the number of customers at the ATM at time t . Model $\{X(t), t \geq 0\}$ as a CTMC.

The state space of the system is $S = \{0, 1, 2, \dots, K\}$. We shall obtain the transition rates below. In state 0, the system is empty and the only event that can occur is an arrival, which occurs after an $\text{Exp}(\lambda)$ amount of time (due to the $\text{PP}(\lambda)$ arrival process) and triggers a transition to state 1. Hence $r_{0,1} = \lambda$. In state i , $1 \leq i \leq K-1$, there are two possible events: an arrival and a departure. An arrival occurs after an $\text{Exp}(\lambda)$ amount of time and takes the system to state $i+1$, while the departure occurs at the next service completion time (which occurs after an $\text{Exp}(\mu)$ amount of time due to the memoryless property of the service times) and takes the system to state $i-1$. Hence $r_{i,i+1} = \lambda$ and $r_{i,i-1} = \mu$, $1 \leq i \leq K-1$. Finally, in state K , there is no arrival due to capacity limitation. Thus the only event that can occur is a service completion, taking the system to state $K-1$. Hence $r_{K,K-1} = \mu$. Thus $\{X(t), t \geq 0\}$ is a CTMC with rate matrix

$$R = \begin{bmatrix} 0 & \lambda & 0 & \cdots & 0 & 0 \\ \mu & 0 & \lambda & \cdots & 0 & 0 \\ 0 & \mu & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda \\ 0 & 0 & 0 & \cdots & \mu & 0 \end{bmatrix}. \quad (4.19)$$

The rate diagram is as shown in Figure 4.7. We shall see in Chapter 6 that $\{X(t), t \geq 0\}$ is called an $M/M/1/K$ queue in queueing terminology. ■

The CTMCs in Examples 4.5 and 4.7 have a similar structure: both CTMCs move from state i to state $i-1$ or to state $i+1$. There are no other transitions. We describe a general class of such DTMCs below.

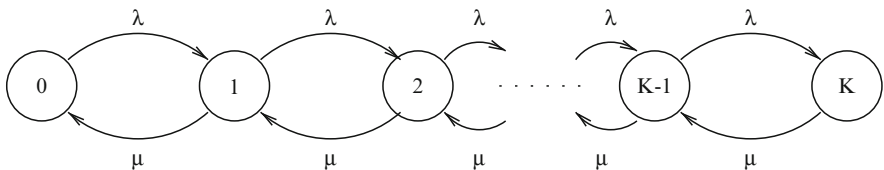


Fig. 4.7 Rate diagram of a single server queue.

Example 4.8. (Finite Birth and Death Process). A CTMC on state space $\{0, 1, 2, \dots, K\}$ with the rate matrix

$$R = \begin{bmatrix} 0 & \lambda_0 & 0 & \cdots & 0 & 0 \\ \mu_1 & 0 & \lambda_1 & \cdots & 0 & 0 \\ 0 & \mu_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda_{K-1} \\ 0 & 0 & 0 & \cdots & \mu_K & 0 \end{bmatrix} \quad (4.20)$$

is called a *finite birth and death process*. The transitions from i to $i + 1$ are called the births, and the λ_i 's are called the birth parameters. The transitions from i to $i - 1$ are called the deaths, and the μ_i 's are called the death parameters. It is convenient to define $\lambda_K = 0$ and $\mu_0 = 0$, signifying that there are no births in state K and no deaths in state 0. A birth and death process spends an $\text{Exp}(\lambda_i + \mu_i)$ amount of time in state i and then jumps to state $i + 1$ with probability $\lambda_i / (\lambda_i + \mu_i)$ or to state $i - 1$ with probability $\mu_i / (\lambda_i + \mu_i)$. ■

Birth and death processes occur quite often in applications and hence form an important class of CTMCs. For example, the CTMC model of the general machine shop in Example 4.5 is a birth and death process on $\{0, 1, \dots, N\}$ with birth parameters

$$\lambda_i = \min(N - i, M)\lambda, \quad 0 \leq i \leq N,$$

and death parameters

$$\mu_i = i\mu, \quad 0 \leq i \leq N.$$

The CTMC model of the single server queue as described in Example 4.7 is a birth and death process on $\{0, 1, \dots, K\}$ with birth parameters

$$\lambda_i = \lambda, \quad 0 \leq i \leq K - 1,$$

and death parameters

$$\mu_i = \mu, \quad 1 \leq i \leq K.$$

We illustrate this with more examples.

Example 4.9. (Telephone Switch). A telephone switch can handle K calls at any one time. Calls arrive according to a Poisson process with rate λ . If the switch is already serving K calls when a new call arrives, then the new call is lost. If a call is accepted, it lasts for an $\text{Exp}(\mu)$ amount of time and then terminates. All call durations are independent of each other. Let $X(t)$ be the number of calls that are being handled by the switch at time t . Model $\{X(t), t \geq 0\}$ as a CTMC.

The state space of $\{X(t), t \geq 0\}$ is $\{0, 1, 2, \dots, K\}$. In state i , $0 \leq i \leq K - 1$, an arrival triggers a transition to state $i + 1$ with rate λ . Hence $r_{i,i+1} = \lambda$. In state K ,

there are no arrivals. In state i , $1 \leq i \leq K$, any of the i calls may complete and trigger a transition to state $i - 1$. The transition rate is $r_{i,i-1} = i\mu$. In state 0, there are no departures. Thus $\{X(t), t \geq 0\}$ is a finite birth and death process with birth parameters

$$\lambda_i = \lambda, \quad 0 \leq i \leq K - 1,$$

and death parameters

$$\mu_i = i\mu, \quad 0 \leq i \leq K.$$

We shall see in Chapter 6 that $\{X(t), t \geq 0\}$ is called an $M/M/K/K$ queue in queueing terminology. ■

Example 4.10. (Call Center). An airline phone-reservation system is called a call center and is staffed by s reservation clerks called agents. An incoming call for reservations is handled by an agent if one is available; otherwise the caller is put on hold. The system can put a maximum of H callers on hold. When an agent becomes available, the callers on hold are served in order of arrival. When all the agents are busy and there are H calls on hold, any additional callers get a busy signal and are permanently lost. Let $X(t)$ be the number of calls in the system, those handled by the agents plus any on hold, at time t . Assume the calls arrive according to a $PP(\lambda)$ and the processing times of the calls are iid $\text{Exp}(\mu)$ random variables. Model $\{X(t), t \geq 0\}$ as a CTMC.

The state space is $\{0, 1, 2, \dots, K\}$, where $K = s + H$. Suppose $X(t) = i$. For $0 \leq i \leq K - 1$, the transition to state $i + 1$ is caused by an arrival, which occurs at rate λ . In state K there are no arrivals. For $1 \leq i \leq K$, the transition to state $i - 1$ is caused by the completion of the processing of any of the $\min(i, s)$ calls that are under service. The completion rate of an individual call is μ ; hence the transition rate from state i to $i - 1$ is $\min(i, s)\mu$. Thus $\{X(t), t \geq 0\}$ is a finite birth and death process with birth parameters

$$\lambda_i = \lambda, \quad 0 \leq i \leq K - 1,$$

and death parameters

$$\mu_i = \min(i, s)\mu, \quad 0 \leq i \leq K.$$

The rate diagram is shown in Figure 4.8. We shall see in Chapter 6 that $\{X(t), t \geq 0\}$ is called an $M/M/s/K$ queue in queueing terminology. ■

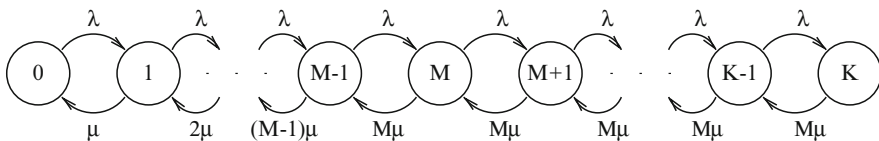


Fig. 4.8 Rate diagram of the call center.

Example 4.11. (Leaky Bucket). Leaky bucket is a traffic control mechanism used in high-speed digital telecommunication networks. It consists of two buffers: a token buffer of size M and a data buffer of size D . Tokens are generated according to a $PP(\mu)$ and are stored in the token buffer. Tokens generated while the token buffer is full are lost. Data packets arrive according to a $PP(\lambda)$. If there is a token in the token buffer when a packet arrives, it immediately removes one token from the token buffer and enters the network. If there are no tokens in the token buffer, it waits in the data buffer if the data buffer is not full. If the data buffer is full, the packet is lost. Let $Y(t)$ be the number of tokens in the token buffer at time t , and let $Z(t)$ be the number of data packets in the data buffer at time t . Model this system as a CTMC.

First note that if there are packets in the data buffer, the token buffer must be empty (otherwise the packets would have already entered the network) and vice versa. Consider a new process $\{X(t), t \geq 0\}$ defined by

$$X(t) = M - Y(t) + Z(t).$$

The state space of $\{X(t), t \geq 0\}$ is $\{0, 1, \dots, K\}$, where $K = M + D$. Also, if $0 \leq X(t) \leq M$, we must have $Z(t) = 0$ and $Y(t) = M - X(t)$, and if $M \leq X(t) \leq K$, we must have $Y(t) = 0$ and $Z(t) = X(t) - M$. Thus we can recover $Y(t)$ and $Z(t)$ from the knowledge of $X(t)$.

Now, suppose $1 \leq X(t) \leq K - 1$. If a token arrives (this happens at rate μ), $X(t)$ decreases by 1, and if a data packet arrives (this happens at rate λ), it increases by 1. If $X(t) = 0$, then $Y(t) = M$, and tokens cannot enter the token buffer, while data packets continue to arrive. If $X(t) = K$, then $Z(t) = D$ and packets cannot enter the data buffer; however, tokens continue to enter. Thus $\{X(t), t \geq 0\}$ is a finite birth and death process with birth parameters

$$\lambda_i = \lambda, \quad 0 \leq i \leq K - 1,$$

and death parameters

$$\mu_i = \mu, \quad 1 \leq i \leq K.$$

Thus $\{X(t), t \geq 0\}$ is the same process as in the single server queue of Example 4.7! ■

Example 4.12. (Inventory Management). A retail store manages the inventory of washing machines as follows. When the number of washing machines in stock decreases to a fixed number k , an order is placed with the manufacturer for m new washing machines. It takes a random amount of time for the order to be delivered to the retailer. If the inventory is at most k when an order is delivered (including the newly delivered order), another order for m items is placed immediately. Suppose the delivery times are iid $\text{Exp}(\lambda)$ and that the demand for the washing machines occurs according to a $PP(\mu)$. Demands that cannot be immediately satisfied are lost. Model this system as a CTMC.

Let $X(t)$ be the number of machines in stock at time t . Note that the maximum number of washing machines in stock is $K = k + m$, which happens if the order is delivered before the next demand occurs. The state space is thus $\{0, 1, \dots, K\}$. In state 0, the demands are lost, and the stock jumps to m when the current outstanding order is delivered (which happens at rate λ). Hence we have $r_{0,m} = \lambda$. In state i , $1 \leq i \leq k$, one order is outstanding. The state changes to $i - 1$ if a demand occurs (which happens at rate μ) and to $i + m$ if the order is delivered. Thus we have $r_{i,i+m} = \lambda$ and $r_{i,i-1} = \mu$. Finally, if $X(t) = i$, $k + 1 \leq i \leq K$, there are no outstanding orders. Hence, the only transition is from i to $i - 1$, and that happens when a demand occurs. Thus, $r_{i,i-1} = \mu$. Thus $\{X(t), t \geq 0\}$ is a CTMC. The rate matrix is shown in (4.21) for the case $k = 3, m = 2$.

$$R = \begin{bmatrix} 0 & 0 & \lambda & 0 & 0 & 0 \\ \mu & 0 & 0 & \lambda & 0 & 0 \\ 0 & \mu & 0 & 0 & \lambda & 0 \\ 0 & 0 & \mu & 0 & 0 & \lambda \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \end{bmatrix}. \quad (4.21)$$

The rate diagram is as shown in Figure 4.9. ■

Example 4.13. (Manufacturing). A simple manufacturing facility consists of a single machine that can be turned on or off. If the machine is on, it produces items according to a Poisson process with rate λ . Demands for the items arrive according to a $PP(\mu)$. The machine is controlled as follows. If the number of items in stock reaches a maximum number K (the storage capacity), the machine is turned off. It is turned on again when the number of items in stock decreases to a prespecified level $k < K$. Model this system as a CTMC.

Let $X(t)$ be the number of items in stock at time t . $\{X(t), t \geq 0\}$ is not a CTMC since we don't know whether the machine is on or off if $k < X(t) < K$. Let $Y(t)$ be the state of the machine at time t , 1 if it is on and 0 if it is off. Then $\{(X(t), Y(t)), t \geq 0\}$ is a CTMC with state space

$$S = \{(i, 1), 0 \leq i < K\} \cup \{(i, 0), k < i \leq K\}.$$

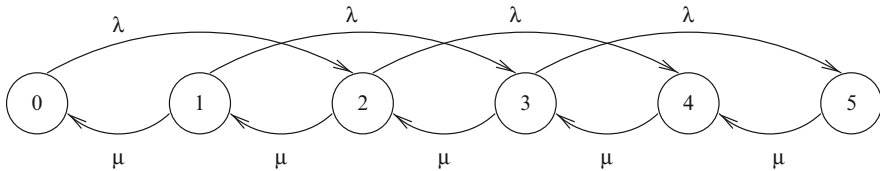


Fig. 4.9 Rate diagram of the inventory management example with $k = 3, r = 2$.