

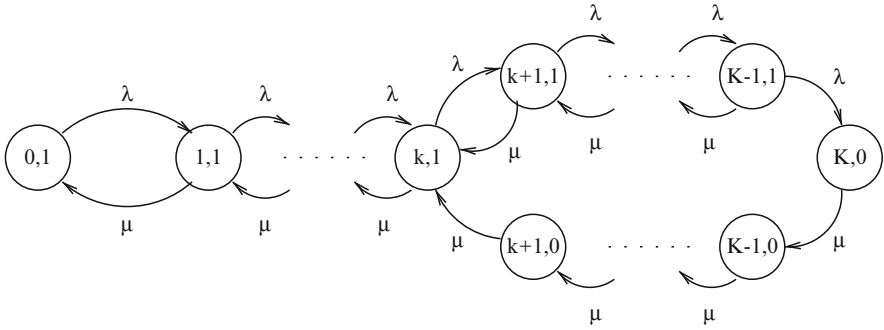


Fig. 4.10 Rate diagram of the manufacturing system.

Note that the machine is always on if the number of items is k or less. Hence we do not need the states $(i, 0), 0 \leq i \leq k$. The usual triggering event analysis yields the following transition rates:

$$\begin{aligned} r_{(i,1),(i+1,1)} &= \lambda, \quad 0 \leq i < K-1, \\ r_{(K-1,1),(K,0)} &= \lambda, \\ r_{(i,1),(i-1,1)} &= \mu, \quad 1 \leq i \leq K-1, \\ r_{(i,0),(i-1,0)} &= \mu, \quad k+1 < i \leq K, \\ r_{(k+1,0),(k,1)} &= \mu. \end{aligned}$$

The rate diagram is shown in Figure 4.10. ■

Armed with these examples, we proceed with the analysis of CTMCs in the next section.

4.4 Transient Analysis: Uniformization

Let $\{X(t), t \geq 0\}$ be a CTMC on state space $S = \{1, 2, \dots, N\}$, and let $R = [r_{i,j}]$ be its rate matrix. In this section, we develop methods of computing the transient distribution of $\{X(t), t \geq 0\}$; that is, we compute the pmf of $X(t)$ for a given t . We assume that the probability distribution of the initial state, $X(0)$, is known. Then we have

$$\mathbf{P}(X(t) = j) = \sum_{i=1}^N \mathbf{P}(X(t) = j | X(0) = i) \mathbf{P}(X(0) = i), \quad 1 \leq j \leq N. \quad (4.22)$$

Thus we need to compute $\mathbf{P}(X(t) = j | X(0) = i) = p_{i,j}(t)$ in order to obtain the pmf of $X(t)$. Hence we concentrate on computing the transition matrix $P(t) = [p_{i,j}(t)]$ in this section. We first introduce the notation needed before stating the main result in Theorem 4.3.

We have seen before that the CTMC spends an $\text{Exp}(r_i)$ amount of time in state i ($r_i = \sum_{j=1}^N r_{i,j}$) and, if $r_i > 0$, jumps to state j with probability $p_{i,j} = r_{i,j}/r_i$. Now let r be any finite number satisfying

$$r \geq \max_{1 \leq i \leq N} \{r_i\}. \quad (4.23)$$

Define a matrix $\hat{P} = [\hat{p}_{i,j}]$ as

$$\hat{p}_{i,j} = \begin{cases} 1 - \frac{r_i}{r} & \text{if } i = j, \\ \frac{r_{i,j}}{r} & \text{if } i \neq j. \end{cases} \quad (4.24)$$

Example 4.14. Suppose R is given as

$$R = \begin{bmatrix} 0 & 2 & 3 & 0 \\ 4 & 0 & 2 & 0 \\ 0 & 2 & 0 & 2 \\ 1 & 0 & 3 & 0 \end{bmatrix}. \quad (4.25)$$

Then we have $r_1 = 5, r_2 = 6, r_3 = 4$, and $r_4 = 4$. Hence, from (4.23), we can choose $r = \max\{5, 6, 4, 4\} = 6$. Then, from (4.24), we get

$$\hat{P} = \begin{bmatrix} \frac{1}{6} & \frac{2}{6} & \frac{3}{6} & 0 \\ \frac{4}{6} & 0 & \frac{2}{6} & 0 \\ 0 & \frac{2}{6} & \frac{2}{6} & \frac{2}{6} \\ \frac{1}{6} & 0 & \frac{3}{6} & \frac{2}{6} \end{bmatrix}. \quad (4.26)$$

Note that $\hat{P}$ is stochastic. ■

Using the notation above, the next theorem gives a method of computing $P(t)$.

Theorem 4.3. (The Matrix $P(t)$). *The transition probability matrix $P(t) = [p_{i,j}(t)]$ is given by*

$$P(t) = \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} \hat{P}^k. \quad (4.27)$$

We present the proof of the theorem above in Appendix 4A at the end of this chapter. Equation (4.27) provides a very stable method of computing $P(t)$. It is known as the *uniformization method* since it involves observing the CTMC at rate r that is a uniform upper bound on the transition rate r_i out of state $i \in S$. In numerical computations, we approximate $P(t)$ by using the first M terms of the infinite series in (4.27). In order to make the method numerically implementable,

we need to describe how to choose M to guarantee a specified accuracy. A good rule of thumb is to choose

$$M \approx \max\{rt + 5 * \sqrt{rt}, 20\}.$$

A more precise algorithm is developed in Appendix 4B. Since the uniformization algorithm terminates more quickly (i.e., it uses less terms of the series) if the value of rt is small, it makes sense to use the smallest value of r allowed by (4.23), namely

$$r = \max_{1 \leq i \leq N} \{r_i\}.$$

This is what we do in the next three examples. However, in some cases it is more convenient to use an r that is larger than the one above. This is illustrated in Example 4.18.

Example 4.15. Let $\{X(t), t \geq 0\}$ be a CTMC with state space $\{1, 2, 3, 4\}$ and the rate matrix given in Example 4.14. Using the values of r and $\hat{P}$ from that example, we get

$$P(t) = \sum_{k=0}^{\infty} e^{-6t} \frac{(6t)^k}{k!} \begin{bmatrix} \frac{1}{6} & \frac{2}{6} & \frac{3}{6} & 0 \\ \frac{4}{6} & 0 & \frac{2}{6} & 0 \\ 0 & \frac{2}{6} & \frac{2}{6} & \frac{2}{6} \\ \frac{1}{6} & 0 & \frac{3}{6} & \frac{2}{6} \end{bmatrix}^k. \quad (4.28)$$

We compute $P(t)$ by using the first M terms of the infinite series in (4.28). We use $\epsilon = .00001$ in the algorithm in Appendix 4B and display the value of $P(t)$ and the value of M :

$$P(.5) = \begin{bmatrix} .2506 & .2170 & .3867 & .1458 \\ .2531 & .2384 & .3744 & .1341 \\ .1691 & .1936 & .4203 & .2170 \\ .1580 & .1574 & .3983 & .2862 \end{bmatrix}, \quad M = 13,$$

$$P(1) = \begin{bmatrix} .2062 & .2039 & .3987 & .1912 \\ .2083 & .2053 & .3979 & .1885 \\ .1968 & .1984 & .4010 & .2039 \\ .1920 & .1940 & .4015 & .2125 \end{bmatrix}, \quad M = 19,$$

$$P(5) = \begin{bmatrix} .2000 & .2000 & .4000 & .2000 \\ .2000 & .2000 & .4000 & .2000 \\ .2000 & .2000 & .4000 & .2000 \\ .2000 & .2000 & .4000 & .2000 \end{bmatrix}, \quad M = 56.$$

It can be numerically verified that $P(1) = P(.5)P(.5)$, thus verifying the Chapman–Kolmogorov equations (4.8). Some rows do not sum to 1 due to rounding errors. ■

Example 4.16. (General Machine Shop). Consider the machine shop of Example 4.5. Suppose the shop has four machines and two repair persons. Suppose the life-times of the machines are exponential random variables with mean 3 days, while the repair times are exponential random variables with mean 2 hours. The shop operates 24 hours a day. Suppose all the machines are operating at 8:00 a.m. Monday. What is the expected number of working machines at 5:00 p.m. the same Monday?

Using the notation of Example 4.5, we see that the repair rate is $\lambda = .5$ per hour and the failure rate is $\mu = 1/72$ per hour. (Note that we must use consistent units.) Let $X(t)$ be the number of working machines at time t . Then, from Example 4.5, we see that $\{X(t), t \geq 0\}$ is a CTMC on $\{0, 1, 2, 3, 4\}$ with the rate matrix

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{72} & 0 & 1 & 0 & 0 \\ 0 & \frac{2}{72} & 0 & 1 & 0 \\ 0 & 0 & \frac{3}{72} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{4}{72} & 0 \end{bmatrix}. \quad (4.29)$$

We can take 8:00 a.m. as time 0 (hours). Thus we are given $X(0) = 4$. We are asked to compute the expected number of working machines at 5:00 p.m.; i.e., after 9 hours. Thus we need $E(X(9))$. From (4.23), we get $r = \frac{74}{72}$, and from (4.24), we get

$$\hat{P} = \begin{bmatrix} \frac{2}{74} & \frac{72}{74} & 0 & 0 & 0 \\ \frac{1}{74} & \frac{1}{74} & \frac{72}{74} & 0 & 0 \\ 0 & \frac{2}{74} & 0 & \frac{72}{74} & 0 \\ 0 & 0 & \frac{3}{74} & \frac{35}{74} & \frac{36}{74} \\ 0 & 0 & 0 & \frac{4}{74} & \frac{70}{74} \end{bmatrix}. \quad (4.30)$$

Using an error tolerance of $\epsilon = .00001$ and $t = 9$ hours in the uniformization algorithm of Appendix 4B, we get

$$P(9) = \begin{bmatrix} .0002 & .0020 & .0136 & .1547 & .8295 \\ .0000 & .0006 & .0080 & .1306 & .8607 \\ .0000 & .0002 & .0057 & .1154 & .8787 \\ .0000 & .0002 & .0048 & .1070 & .8880 \\ .0000 & .0001 & .0041 & .0987 & .8971 \end{bmatrix}. \quad (4.31)$$

The algorithm used the first 26 terms of the infinite series to compute the matrix above. Now we are given that $X(0) = 4$. Hence the pmf of $X(9)$ is given by the last

row of the matrix above. Hence the expected number of working machines at 5:00 p.m. is given by

$$E(X(9)) = 0 * .0000 + 1 * .0001 + 2 * .0041 + 3 * .0987 + 4 * .8971 = 3.8928.$$

Note that the probability that all machines are working at 5:00 p.m. is .8971; however, this is not the probability that there were no failures from 8:00 a.m. to 5:00 p.m. There may have been one or more failures and the corresponding repairs. ■

Example 4.17. (Airplane Reliability). Consider the four-engine jet airplane of Example 4.6. Suppose each engine can give trouble-free service for 200 hours on average before it needs to be shut down due to any problem. Assume that all four engines are operational at the beginning of a 6-hour flight. What is the probability that this flight lands safely?

We assume that the only trouble on the flight is from the engine failures. Let $\{X(t), t \geq 0\}$ be the CTMC as in Example 4.6. The rate matrix is as given in (4.18) with $\lambda = 1/200 = .005$ per hour. The flight lands safely if the state of the CTMC stays in the set $\{5, 6, 8, 9\}$ for the duration $[0, 6]$. Since this CTMC cannot return to this set of states once it leaves it, the required probability is given by the probability that $X(6)$ is in $\{5, 6, 8, 9\}$ given that $X(0) = 9$, the initial state. Using the uniformization algorithm of Appendix 4B, we get as the probability of a safe landing

$$p_{9,5}(6) + p_{9,6}(6) + p_{9,8}(6) + p_{9,9}(6) = .0033 + .0540 + .0540 + .8869 = .9983.$$

Thus the probability of a crash is .0017. This is far too large for any commercial airline! ■

Example 4.18. (Two-State CTMC). Consider the two-state CTMC of Example 4.3 with state space $\{0, 1\}$ and rate matrix

$$R = \begin{bmatrix} 0 & \lambda \\ \mu & 0 \end{bmatrix}. \quad (4.32)$$

Here $r_0 = \lambda$ and $r_1 = \mu$. Hence (4.23) would suggest $r = \max\{\lambda, \mu\}$. However, it is more convenient to choose

$$r = \lambda + \mu.$$

Then the $\hat{P}$ matrix of (4.24) is given by

$$\hat{P} = \begin{bmatrix} \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \\ \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \end{bmatrix}. \quad (4.33)$$

It can be readily verified that

$$\hat{p}^k = \begin{bmatrix} \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \\ \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \end{bmatrix}, \quad k \geq 1. \quad (4.34)$$

Substituting in (4.27), we get

$$\begin{aligned} P(t) &= \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} \hat{p}^k \\ &= e^{-rt} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \sum_{k=1}^{\infty} e^{-rt} \frac{(rt)^k}{k!} \begin{bmatrix} \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \\ \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \end{bmatrix} \\ &= e^{-rt} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \\ \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \end{bmatrix} \sum_{k=1}^{\infty} e^{-rt} \frac{(rt)^k}{k!} \\ &= e^{-rt} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \\ \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \end{bmatrix} (1 - e^{-rt}) \\ &= \begin{bmatrix} \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \\ \frac{\mu}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \end{bmatrix} + \begin{bmatrix} \frac{\lambda}{\lambda + \mu} & -\frac{\lambda}{\mu + \lambda} \\ -\frac{\lambda}{\lambda + \mu} & \frac{\lambda}{\lambda + \mu} \end{bmatrix} e^{-rt}. \quad (4.35) \end{aligned}$$

From this we can read

$$p_{1,1}(t) = \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t}.$$

Thus the probability that a machine is up at time t given that it was up at time 0 is given by the expression above. Note that the machine could have gone through many failures and repairs up to time t . The probability above includes the sum total of all such possibilities. ■

4.5 Occupancy Times

In this section, we shall do the short-term analysis of a CTMC $\{X(t), t \geq 0\}$; i.e., we study the behavior of the CTMC over a finite interval $[0, T]$. We concentrate on the *occupancy time* of a given state; i.e., the expected length of time the system spends in that state during a given interval of time. Here we develop a method for computing such quantities.

Let $X(t)$ be the state of a system at time t , and assume that $\{X(t), t \geq 0\}$ is a CTMC with state space $S = \{1, 2, \dots, N\}$ and rate matrix R . Let $m_{i,j}(T)$ be the expected amount of time the CTMC spends in state j during the interval $[0, T]$, starting in state i . The quantity $m_{i,j}(T)$ is called the occupancy time of state j until time T starting from state i . The following theorem shows a way of computing $m_{i,j}(T)$.

Theorem 4.4. (The Occupancy Times). *Let $P(t) = [p_{i,j}(t)]$ be the transition probability matrix of $\{X(t), t \geq 0\}$. Then*

$$m_{i,j}(T) = \int_0^T p_{i,j}(t) dt, \quad 1 \leq i, j \leq N. \quad (4.36)$$

Proof. Let i and j be fixed, and define

$$Y_j(t) = \begin{cases} 1 & \text{if } X(t) = j, \\ 0 & \text{otherwise.} \end{cases}$$

The total amount of time spent in state j by the CTMC during $[0, T]$ is then given by

$$\int_0^T Y_j(t) dt.$$

Hence we get

$$\begin{aligned} m_{i,j}(T) &= \mathbf{E} \left(\int_0^T Y_j(t) dt \mid X(0) = i \right) \\ &= \int_0^T \mathbf{E}(Y_j(t) \mid X(0) = i) dt \\ &\quad \text{(the integral and } \mathbf{E} \text{ can be interchanged here)} \\ &= \int_0^T \mathbf{P}(Y_j(t) = 1 \mid X(0) = i) dt \\ &= \int_0^T \mathbf{P}(X(t) = j \mid X(0) = i) dt \\ &= \int_0^T p_{i,j}(t) dt \end{aligned}$$

as desired. ■

Example 4.19. (Two-State Machine). Consider the two-state machine of Example 4.3. Suppose the expected time until failure of the machine is 10 days, while the expected repair time is 1 day. Suppose the machine is working at the beginning of January. Compute the expected total uptime of the machine in the month of January.

Let $\{X(t), t \geq 0\}$ be as in Example 4.3. Using the data given here in time units of days, we get $\mu = .1 \text{ days}^{-1}$ and $\lambda = 1 \text{ days}^{-1}$. Since there are 31 days in January, the expected uptime in January is given by $m_{1,1}(31)$. From Example 4.18, we have

$$\begin{aligned} p_{1,1}(t) &= \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t} \\ &= \frac{10}{11} + \frac{1}{11} e^{-1.1t}. \end{aligned} \quad (4.37)$$

Substituting in (4.36), we get

$$\begin{aligned} m_{1,1}(31) &= \int_0^{31} p_{1,1}(t) dt \\ &= \int_0^{31} \left(\frac{10}{11} + \frac{1}{11} e^{-1.1t} \right) dt \\ &= \frac{10}{11} \cdot 31 + \frac{1}{11} \cdot \frac{1}{1.1} (1 - e^{-34.1}) \\ &= 28.26 \text{ days.} \end{aligned}$$

Thus the expected uptime during January is 28.26 days. Hence the expected downtime must be $31 - 28.26 = 2.74$ days. The reader should compute $m_{1,0}(31)$ and verify this independently. ■

Note that Theorem 4.4 can be used directly if $p_{i,j}(t)$ is known as a function of t in closed form, as was the case in the example above. However, $p_{i,j}(t)$ is rarely known in closed form and mostly has to be computed using the infinite series representation of (4.27). How do we compute $m_{i,j}(T)$ in such cases?

First we compute the uniformization constant r from (4.23) and the $\hat{P}$ matrix from (4.24). Next we construct an $N \times N$ occupancy matrix,

$$M(t) = [m_{i,j}(t)]. \quad (4.38)$$

The next theorem gives an infinite series expression for $M(T)$, analogous to Theorem 4.3.

Theorem 4.5. (The Matrix $M(T)$). *Let Y be a $P(rT)$ random variable. Then*

$$M(T) = \frac{1}{r} \sum_{k=0}^{\infty} P(Y > k) \hat{P}^k, \quad T \geq 0. \quad (4.39)$$

Proof. From (4.27), we get

$$p_{i,j}(t) = \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} [\hat{P}^k]_{i,j}.$$

We also have

$$\begin{aligned} \int_0^T e^{-rt} \frac{(rt)^k}{k!} dt &= \frac{1}{r} \left(1 - \sum_{l=0}^k e^{-rT} \frac{(rT)^l}{l!} \right) \\ &= \frac{1}{r} \mathbf{P}(Y > k). \end{aligned}$$

Substituting in (4.36), we get

$$\begin{aligned} m_{i,j}(T) &= \int_0^T p_{i,j}(t) dt \\ &= \int_0^T \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} [\hat{P}^k]_{i,j} dt \\ &= \sum_{k=0}^{\infty} \left(\int_0^T e^{-rt} \frac{(rt)^k}{k!} dt \right) [\hat{P}^k]_{i,j} \\ &= \sum_{k=0}^{\infty} \frac{1}{r} \mathbf{P}(Y > k) [\hat{P}^k]_{i,j}. \end{aligned}$$

Putting the equation above in matrix form yields (4.39). ■

The theorem above provides a very stable method of computing the matrix $M(T)$. An algorithm to compute $M(T)$ based on the theorem above is developed in Appendix 4C. We illustrate the theorem above with several examples below.

Example 4.20. (Single Server Queue). Consider the ATM queue of Example 4.7. Suppose the customers arrive at a rate of 10 per hour and take on average 4 minutes at the machine to complete their transactions. Suppose that there is space for at most five customers in front of the ATM and that it is idle at 8:00 a.m. What is the expected amount of time the machine is idle during the next hour?

The arrival process is Poisson with rate $\lambda = 10$ per hour. The processing times are iid exponential with mean 4/60 hours; i.e., with parameter $\mu = 15$ per hour. The capacity is $K = 5$. Let $X(t)$ be the number of customers in the queue at time t . With these parameters, $\{X(t), t \geq 0\}$ is a CTMC with state space $\{0, 1, 2, 3, 4, 5\}$ and rate matrix given by

$$R = \begin{bmatrix} 0 & 10 & 0 & 0 & 0 & 0 \\ 15 & 0 & 10 & 0 & 0 & 0 \\ 0 & 15 & 0 & 10 & 0 & 0 \\ 0 & 0 & 15 & 0 & 10 & 0 \\ 0 & 0 & 0 & 15 & 0 & 10 \\ 0 & 0 & 0 & 0 & 15 & 0 \end{bmatrix}. \quad (4.40)$$

We are given $X(0) = 0$. The expected idle time is thus given by $m_{0,0}(1)$. Using the uniformization algorithm of Appendix 4C with $\epsilon = .00001$, we get

$$M(1) = \begin{bmatrix} .4451 & .2548 & .1442 & .0814 & .0465 & .0280 \\ .3821 & .2793 & .1604 & .0920 & .0535 & .0326 \\ .3246 & .2407 & .2010 & .1187 & .0710 & .0441 \\ .2746 & .2070 & .1780 & .1695 & .1046 & .0663 \\ .2356 & .1806 & .1598 & .1569 & .1624 & .1047 \\ .2124 & .1648 & .1489 & .1492 & .1571 & .1677 \end{bmatrix}.$$

The algorithm uses the first 44 terms of the infinite series to compute the result above. Thus $m_{0,0}(1) = .4451$ hours; i.e., if the ATM is idle at the beginning, it is idle on average $.4451 * 60 = 26.71$ minutes of the first hour. Similarly, the queue is filled to capacity for $m_{0,5} = .0280$ hours = 1.68 minutes of the first hour. ■

Example 4.21. (Manufacturing). Consider the manufacturing operation described in Example 4.13. Suppose the system operates 24 hours a day, the demands occur at the rate of five per hour, and the average time to manufacture one item is 10 minutes. The machine is turned on whenever the stock of manufactured items is reduced to two, and it stays on until the stock rises to four, at which point it is turned off. It stays off until the stock is reduced to two, and so on. Suppose the stock is at four (and the machine is off) at the beginning. Compute the expected amount of time during which the machine is on during the next 24 hours.

Let $X(t)$ be the number of items in stock at time t , and 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

$$S = \{1 = (0, 1), 2 = (1, 1), 3 = (2, 1), 4 = (3, 1), 5 = (4, 0), 6 = (3, 0)\}.$$

The demand rate is $\mu = 5$ per hour and the production rate is $\lambda = 6$ per hour. Following the analysis of Example 4.13, we get as the rate matrix of the CTMC

$$R = \begin{bmatrix} 0 & 6 & 0 & 0 & 0 & 0 \\ 5 & 0 & 6 & 0 & 0 & 0 \\ 0 & 5 & 0 & 6 & 0 & 0 \\ 0 & 0 & 5 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 \\ 0 & 0 & 5 & 0 & 0 & 0 \end{bmatrix}. \quad (4.41)$$

We are given that the initial state of the system is $(X(0), Y(0)) = (4, 0)$; i.e., the initial state is 5. The machine is on whenever the system is in state 1, 2, 3, or 4. Hence the total expected on time during 24 hours is given by $m_{5,1}(24) + m_{5,2}(24) + m_{5,3}(24) + m_{5,4}(24)$. We use the uniformization algorithm of Appendix 4C with $T = 24$ and $\epsilon = .00001$. The algorithm uses the first 331 terms of the infinite series to yield

$$M(24) = \begin{bmatrix} 4.0004 & 4.6322 & 5.4284 & 2.9496 & 3.5097 & 3.4798 \\ 3.8602 & 4.6639 & 5.4664 & 2.9703 & 3.5345 & 3.5047 \\ 3.7697 & 4.5553 & 5.5361 & 3.0084 & 3.5802 & 3.5503 \\ 3.7207 & 4.4965 & 5.4656 & 3.0608 & 3.6431 & 3.6132 \\ 3.7063 & 4.4793 & 5.4448 & 2.9586 & 3.7204 & 3.6906 \\ 3.7380 & 4.5173 & 5.4905 & 2.9835 & 3.5503 & 3.7204 \end{bmatrix}.$$

Hence the total expected on time is given by

$$3.7063 + 4.4793 + 5.4448 + 2.9586 = 16.5890 \text{ hours.}$$

If the system had no stock in the beginning, the total expected uptime would be

$$4.0004 + 4.6322 + 5.4284 + 2.9496 = 17.0106 \text{ hours. } \blacksquare$$

4.6 Limiting Behavior

In the study of the limiting behavior of DTMCs in Section 2.5, we studied three quantities: the limiting distribution, the stationary distribution, and the occupancy distribution. In this section, we study similar quantities for the CTMCs. Let $\{X(t), t \geq 0\}$ be a CTMC with state space $S = \{1, 2, \dots, N\}$ and rate matrix R . In Section 4.4, we saw how to compute $P(X(t) = j)$ for $1 \leq j \leq N$. In this section, we study

$$p_j = \lim_{t \rightarrow \infty} P(X(t) = j). \quad (4.42)$$

If the limit above exists, we call it the *limiting probability* of being in state j . If it exists for all $1 \leq j \leq N$, we call

$$p = [p_1, p_2, \dots, p_N]$$

the *limiting distribution* of the CTMC. If the limiting distribution exists, we ask if it is unique and how to compute it if it is unique.

Here we recapitulate some important facts from Appendix 4A. The reader does not need to follow the appendix if he or she is willing to accept these facts as true. Recall the definition of the uniformization constant from (4.23). We opt to choose strict inequality here as follows:

$$r > \max_{1 \leq j \leq N} \{r_j\}.$$

Construct $\hat{P}$ from R by using (4.24). We have shown in Theorem 4.13 that $\hat{P}$ is a stochastic matrix. Let $\{\hat{X}_n, n \geq 0\}$ be a DTMC on state space $\{1, 2, \dots, N\}$ with

transition probability matrix $\hat{P}$. Let $N(t)$ be a $P(r)$, and assume that it is independent of $\{\hat{X}_n, n \geq 0\}$. Then we saw in Appendix 4A that

$$P(X(t) = j) = P(\hat{X}_{N(t)} = j).$$

Hence the limiting distribution of $X(t)$, if it exists, is identical to the limiting distribution of $\hat{X}_n$. We use this fact below.

Suppose $\{\hat{X}_n, n \geq 0\}$ is irreducible; i.e., it can go from any state i to any state j in a finite number of steps. However, $\hat{X}_n = i$ for some n if and only if $X(t) = i$ for some t . This implies that the CTMC $\{X(t), t \geq 0\}$ can go from any state i to any state j in a finite amount of time. This justifies the following definition.

Definition 4.2. (Irreducible CTMC). The CTMC $\{X(t), t \geq 0\}$ is said to be irreducible if the corresponding DTMC $\{\hat{X}_n, n \geq 0\}$ is irreducible.

Using this definition, we get the next theorem.

Theorem 4.6. (Limiting Distribution). *An irreducible CTMC $\{X(t), t \geq 0\}$ with rate matrix $R = [r_{i,j}]$ has a unique limiting distribution $p = [p_1, p_2, \dots, p_N]$. It is given by the solution to*

$$p_j r_j = \sum_{i=1}^N p_i r_{i,j}, \quad 1 \leq j \leq N, \quad (4.43)$$

$$\sum_{i=1}^N p_i = 1. \quad (4.44)$$

Proof. The CTMC $\{X(t), t \geq 0\}$ is irreducible; hence the DTMC $\{\hat{X}_n, n \geq 0\}$ is irreducible. The choice of the uniformization constant r implies that $\hat{p}_{j,j} = 1 - r_j/r > 0$ for all j . Thus the DTMC $\{\hat{X}_n, n \geq 0\}$ is aperiodic and irreducible. Hence, from Theorem 2.10, it has a limiting distribution

$$\lim_{n \rightarrow \infty} P(\hat{X}_n = j) = \pi_j, \quad 1 \leq j \leq N.$$

The limiting distribution satisfies (2.35) and (2.36) as follows:

$$\pi_j = \sum_{i=1}^N \pi_i \hat{p}_{i,j}$$

and

$$\sum_{j=1}^N \pi_j = 1.$$

Now, $N(t)$ goes to ∞ as t goes to ∞ . Hence

$$\begin{aligned}
 p_j &= \lim_{t \rightarrow \infty} P(X(t) = j) \\
 &= \lim_{t \rightarrow \infty} P(\hat{X}_{N(t)} = j) \\
 &= \lim_{n \rightarrow \infty} P(\hat{X}_n = j) \\
 &= \pi_j.
 \end{aligned} \tag{4.45}$$

Hence, the limiting distribution p exists and is unique, and is equal to the limiting distribution of the DTMC! Substituting $p_j = \pi_j$ in the balance equations satisfied by π , we get

$$p_j = \sum_{i=1}^N p_i \hat{p}_{i,j} \tag{4.46}$$

and

$$\sum_{j=1}^N p_j = 1.$$

The last equation produces (4.44). Substituting in (4.46) for $\hat{p}_{i,j}$ from (4.24), we get

$$\begin{aligned}
 p_j &= \sum_{i=1}^N p_i \hat{p}_{i,j} \\
 &= p_j \hat{p}_{j,j} + \sum_{i \neq j} p_i \hat{p}_{i,j} \\
 &= p_j \left(1 - \frac{r_j}{r}\right) + \sum_{i \neq j} p_i \frac{r_{i,j}}{r}.
 \end{aligned}$$

Canceling p_j from both sides and rearranging, we get

$$p_j \frac{r_j}{r} = \sum_{i \neq j} p_i \frac{r_{i,j}}{r}.$$

Canceling the common factor r and recalling that $r_{i,i} = 0$, we get (4.43). ■

We can think of $p_j r_j$ as the rate at which the CTMC leaves state j and $p_i r_{i,j}$ as the rate at which the CTMC enters state j from state i . Thus, (4.43) says that the limiting probabilities are such that the rate of entering state j from all other states is the same as the rate of leaving state j . Hence it is called the *balance equation* or *rate-in–rate-out* equation.

The balance equations can easily be written down from the rate diagram as follows. Imagine that there is a flow from node i to node j at rate $p_i r_{i,j}$ on every arc (i, j) in the rate diagram. Then the balance equations simply say that the total flow into a node from all other nodes in the rate diagram equals the total flow out of that node to all other nodes.

In Chapter 2, we saw that the limiting distribution of a DTMC is also its stationary distribution. Is the same true for CTMCs as well? The answer is given by the next theorem.

Theorem 4.7. (Stationary Distribution). *Let $\{X(t), t \geq 0\}$ be an irreducible CTMC with limiting distribution p . Then it is also the stationary distribution of the CTMC; i.e.,*

$$\begin{aligned} P(X(0) = j) &= p_j \text{ for } 1 \leq j \leq N \Rightarrow \\ P(X(t) = j) &= p_j \text{ for } 1 \leq j \leq N, t \geq 0. \end{aligned}$$

Proof. From the proof of Theorem 4.6, $p_j = \pi_j$ for all $1 \leq j \leq N$. Suppose $P(X(0) = j) = P(\hat{X}_0 = j) = p_j$ for $1 \leq j \leq N$. Then, from Corollary 2.3,

$$P(\hat{X}_k = j) = \pi_j = p_j, \quad 1 \leq j \leq N, k \geq 0,$$

and

$$\begin{aligned} P(X(t) = j) &= \sum_{i=1}^N P(X(t) = j | X(0) = i) P(X(0) = i) \\ &= \sum_{i=1}^N P(X(0) = i) \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} [\hat{P}^k]_{i,j} \\ &= \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} \sum_{i=1}^N P(X(0) = i) P(\hat{X}_k = j | \hat{X}_0 = i) \\ &= \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} \sum_{i=1}^N P(\hat{X}_k = j | \hat{X}_0 = i) P(\hat{X}_0 = i) \\ &= \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} P(\hat{X}_k = j) \\ &= \sum_{k=0}^{\infty} e^{-rt} \frac{(rt)^k}{k!} p_j \\ &= p_j \end{aligned}$$

as desired. ■

The next question is about the occupancy distribution. In DTMCs, the limiting distribution is also the occupancy distribution. Does this hold true for the CTMCs? The answer is yes if we define the occupancy of state i to be the expected long-run fraction of the time the CTMC spends in state i . The main result is given in the next theorem.

Theorem 4.8. (Occupancy Distribution). *Let $m_{i,j}(T)$ be the expected total time spent in state j up to time T by an irreducible CTMC starting in state i . Then*

$$\lim_{T \rightarrow \infty} \frac{m_{i,j}(T)}{T} = p_j,$$

where p_j is the limiting probability that the CTMC is in state j .

Proof. From Theorem 4.5, we see that

$$M(T) = \frac{1}{r} \sum_{k=0}^{\infty} \mathbf{P}(Y > k) \hat{P}^k, \quad T \geq 0,$$

where $M(T) = [m_{i,j}(T)]$ and Y is a $P(rT)$ random variable. We have

$$\mathbf{P}(Y > k) = \sum_{m=k+1}^{\infty} a_m,$$

where

$$a_m = e^{-rT} \frac{(rT)^m}{m!}.$$

Substituting, we get

$$\begin{aligned} M(T) &= \frac{1}{r} \sum_{k=0}^{\infty} \sum_{m=k+1}^{\infty} a_m \hat{P}^k \\ &= \frac{1}{r} \sum_{m=1}^{\infty} a_m \sum_{k=0}^{m-1} \hat{P}^k \\ &= \frac{1}{r} \sum_{m=1}^{\infty} e^{-rT} \frac{(rT)^{m-1}}{(m-1)!} \frac{rT}{m} \sum_{k=0}^{m-1} \hat{P}^k. \end{aligned}$$

Hence

$$\begin{aligned} \frac{M(T)}{T} &= \sum_{m=1}^{\infty} e^{-rT} \frac{(rT)^{m-1}}{(m-1)!} \frac{1}{m} \sum_{k=0}^{m-1} \hat{P}^k \\ &= \sum_{m=0}^{\infty} e^{-rT} \frac{(rT)^m}{m!} \frac{1}{m+1} \sum_{k=0}^m \hat{P}^k \\ &= \mathbf{E} \left(\frac{1}{N(T)+1} \sum_{k=0}^{N(T)} \hat{P}^k \right), \end{aligned}$$

where $N(T)$ is the number of events in a PP(r) up to time T . From Theorems 4.7 and 4.8, we see that

$$\lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{k=0}^n [\hat{P}^k]_{i,j} = \pi_j = p_j.$$

Also, as $T \rightarrow \infty$, $N(T) \rightarrow \infty$. Using these two observations, we get

$$\begin{aligned} \lim_{T \rightarrow \infty} \frac{m_{i,j}(T)}{T} &= \lim_{T \rightarrow \infty} \mathbb{E} \left(\frac{1}{N(T)+1} \sum_{k=0}^{N(T)} [\hat{P}^k]_{i,j} \right) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{k=0}^n [\hat{P}^k]_{i,j} \\ &= p_j \end{aligned}$$

as desired. ■

Thus, an irreducible CTMC has a unique limiting distribution, which is also its stationary distribution and its occupancy distribution. It can be computed by solving the balance equation (4.43) along with the normalizing equation (4.44). We illustrate the concepts above by means of several examples below.

Example 4.22. (Two-State Machine). Consider the two-state machine as described in Example 4.3. Compute the limiting distribution of the state of the machine.

The two-state CTMC of Example 4.3 is irreducible. Hence it has a unique limiting distribution $[p_0, p_1]$. The two balance equations are

$$\lambda p_0 = \mu p_1,$$

$$\mu p_1 = \lambda p_0.$$

Note that they are the same! The normalizing equation is

$$p_0 + p_1 = 1.$$

The solution is given by

$$p_0 = \frac{\mu}{\lambda + \mu}, \quad p_1 = \frac{\lambda}{\lambda + \mu}.$$

Note that this is also the stationary distribution and the occupancy distribution of the machine. (Check this using the $P(t)$ matrix from Example 4.18.)

Suppose the expected time until failure of the machine is 10 days, while the expected repair time is 1 day. Thus $\lambda = 1$ and $\mu = .1$. Hence

$$p_0 = \frac{1}{11}, \quad p_1 = \frac{10}{11}.$$

Thus, in the long run, the machine spends 10 out of 11 days in working condition. This makes intuitive sense in view of the fact that the machine stays down for 1 day, followed by being up for 10 days on average! ■

Example 4.23. (Finite Birth and Death Process). Let $\{X(t), t \geq 0\}$ be a birth and death process as described in Example 4.8. Assume that all the birth rates $\{\lambda_i, 0 \leq i < K\}$ and death rates $\{\mu_i, 1 \leq i \leq K\}$ are positive. Then the CTMC is irreducible and hence has a unique limiting distribution. We compute it below.

The balance equations are as follows:

$$\begin{aligned}\lambda_0 p_0 &= \mu_1 p_1, \\ (\lambda_1 + \mu_1) p_1 &= \lambda_0 p_0 + \mu_2 p_2, \\ (\lambda_2 + \mu_2) p_2 &= \lambda_1 p_1 + \mu_3 p_3, \\ &\vdots \\ (\lambda_{K-1} + \mu_{K-1}) p_{K-1} &= \lambda_{K-2} p_{K-2} + \mu_K p_K, \\ \mu_K p_K &= \lambda_{K-1} p_{K-1}.\end{aligned}$$

From the first equation above, we get

$$p_1 = \frac{\lambda_0}{\mu_1} p_0.$$

Adding the first two balance equations, we get

$$\lambda_1 p_1 = \mu_2 p_2,$$

which yields

$$p_2 = \frac{\lambda_1}{\mu_2} p_1 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0.$$

Proceeding this way (adding the first three balance equations, etc.), we get, in general,

$$p_i = \rho_i p_0, \quad 0 \leq i \leq K, \quad (4.47)$$

where $\rho_0 = 1$ and

$$\rho_i = \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i}, \quad 1 \leq i \leq K. \quad (4.48)$$

The only unknown left is p_0 , which can be computed by using the normalization equation as follows:

$$p_0 + p_1 + \cdots + p_{K-1} + p_K = (\rho_0 + \rho_1 + \cdots + \rho_{K-1} + \rho_K) p_0 = 1.$$

This yields

$$p_0 = \frac{1}{\sum_{j=0}^K \rho_j}. \quad (4.49)$$

Combining this with (4.47), we get

$$p_i = \frac{\rho_i}{\sum_{j=0}^K \rho_j}, \quad 0 \leq i \leq K. \quad (4.50)$$

This also gives the stationary distribution and the occupancy distribution of the birth and death processes. ■

Example 4.24. (The General Machine Shop). Consider the machine shop of Example 4.16, with four machines and two repair persons. Compute the long-run probability that all the machines are up. Also compute the long-run fraction of the time that both repair persons are busy.

Let $X(t)$ be the number of working machines at time t . Then $\{X(t), t \geq 0\}$ is a birth and death process on $\{0, 1, 2, 3, 4\}$ with birth rates $\lambda_0 = 1, \lambda_1 = 1, \lambda_2 = 1, \lambda_3 = \frac{1}{2}$ and death rates $\mu_1 = \frac{1}{72}, \mu_2 = \frac{2}{72}, \mu_3 = \frac{3}{72}, \mu_4 = \frac{4}{72}$. Hence we get

$$\rho_0 = 1, \rho_1 = 72, \rho_2 = 2592, \rho_3 = 62,208, \rho_4 = 559,872.$$

Thus

$$\sum_{i=0}^4 \rho_i = 624,745.$$

Hence, using (4.50), we get

$$\begin{aligned} p_0 &= \frac{1}{624,745} = 1.6001 \times 10^{-6}, \\ p_1 &= \frac{72}{624,745} = 1.1525 \times 10^{-4}, \\ p_2 &= \frac{2592}{624,745} = 4.149 \times 10^{-3}, \\ p_3 &= \frac{62,208}{624,745} = 9.957 \times 10^{-2}, \\ p_4 &= \frac{559,872}{624,745} = 8.962 \times 10^{-1}. \end{aligned}$$

Thus, in the long run, all the machines are functioning 89.6% of the time. Both repair persons are busy whenever the system is in state 0, 1, or 2. Hence the long-run fraction of the time that both repair persons are busy is given by $p_0 + p_1 + p_2 = .00427$. ■

Example 4.25. (Finite-Capacity Single Server Queue). Consider the finite-capacity single server queue of Example 4.7. Let $X(t)$ be the number of customers in the queue at time t . Then we have seen that $\{X(t), t \geq 0\}$ is a birth and death process on state space $\{0, 1, \dots, K\}$ with birth rates

$$\lambda_i = \lambda, \quad 0 \leq i < K,$$

and death rates

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

Now define

$$\rho = \frac{\lambda}{\mu}.$$

Using the results of Example 4.23, we get

$$\rho_i = \rho^i, \quad 0 \leq i \leq K,$$

and hence, assuming $\rho \neq 1$,

$$\sum_{i=0}^K \rho_i = \sum_{i=0}^K \rho^i = \frac{1 - \rho^{K+1}}{1 - \rho}.$$

Substituting in (4.50), we get

$$p_i = \frac{1 - \rho}{1 - \rho^{K+1}} \rho^i, \quad 0 \leq i \leq K. \quad (4.51)$$

This gives the limiting distribution of the number of customers in the single server queue. We leave it to the reader to verify that, if $\rho = 1$, the limiting distribution is given by

$$p_i = \frac{1}{K+1}, \quad 0 \leq i \leq K. \quad \blacksquare \quad (4.52)$$

Example 4.26. (Leaky Bucket). Consider the leaky bucket of Example 4.11. Suppose the token buffer size is $M = 10$ and the data buffer size is $D = 14$. The tokens are generated at the rate $\mu = 1$ per millisecond, and data packets arrive at the rate $\lambda = 1$ per millisecond. What fraction of the time is the data buffer empty? What fraction of the time is it full?

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 . In Example 4.11, we saw that the process $\{X(t), t \geq 0\}$ defined by

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

behaves like a single server queue with capacity $K = M + D = 24$, arrival rate $\lambda = 1$, and service rate $\mu = 1$. Hence $\rho = \lambda/\mu = 1$. Thus, from (4.52), we get

$$p_i = \frac{1}{25}, \quad 0 \leq i \leq 24.$$

Now, from the analysis in Example 4.11, the data buffer is empty whenever the $\{X(t), t \geq 0\}$ process is in states 0 through $M = 10$. Hence the long-run fraction of the time the data buffer is empty is given by

$$\sum_{i=0}^{10} p_i = \frac{11}{25} = .44.$$