Chapter 5

Queueing theory

All our queues will be Markov processes, and for this we need to have exponential service times. Let Y (≥ 0) denote the duration of a service. It has the exponential distribution with parameter $\mu > 0$. The probability density function is $f(t) = \mu e^{-\mu t}$ for $t \geq 0$ and $f(t) = 0$ for $t < 0$. It is known that

$$\mathbb{E}(Y) = \frac{1}{\mu}, \quad \mathrm{Var}(Y) = \frac{1}{\mu^2}.$$

A useful feature to note is that for any positive number c,

$$\mathbb{P}(Y > c) = \int_c^\infty \mu e^{-\mu t}\, dt = e^{-\mu c}.$$

Applying this gives, for h a small positive number,

$$\mathbb{P}(Y > c + h \mid Y > c) = \frac{\mathbb{P}(Y > c + h)}{\mathbb{P}(Y > c)} = e^{-\mu h} = 1 - \mu h + o(h) \approx 1 - \mu h.$$

This can be interpreted as, given that a service is in progress at time c, then the probability that it is still in progress at time $c + h$ is approximately $1-\mu h$. This does not depend on the value of c, i.e. the process 'has no memory' of when the service started. The probability that the service is completed within time h is approximately μh. So μ can be interpreted as the *completion-of-service rate*. We use this idea to set up differential equations for queues with exponential service times.

5.1 M/M/1

This notation represents a queue with Poisson arrivals, exponential service times and a single server. The queue discipline is first-come first-served. We denote the arrival rate by λ and the service rate by μ per unit time.

Let X_t be the number of the customers in the queueing system at time t and we call X_t the queue size. It should be emphasized that the queue size will always include the customer being served if there is one. So queue size 0 means that there are no customers present, queue size 1 means that there is one customer who is being served but no one is waiting, etc. Let $P_n(t)$ denote the probability that there are n customers in the queueing system at time t. That is

$$P_n(t) = \mathbb{P}(X_t = n), \quad n = 0, 1, 2, \cdots.$$

We shall set up differential equations for $P_n(t)$ and find the stationary distribution of queue size.

First consider the event that the queue size is 0 at time $t+h$, where h is small and positive. Then there might have been queue size 0 at time t with no arrivals during the small interval, which has probability $1-\lambda h+o(h)$, or the queue size might have been 1 at time t with no arrivals during the small interval but the service of the one customer was completed, which has probability $(1-\lambda h+o(h))(\mu h+o(h))=\mu h+o(h)$. Other possibilities involve two or more events of arrivals or departures and have probability $o(h)$. In terms of mathematical equations,

$$\begin{aligned}
\mathbb{P}(X_{t+h}=0) &= \mathbb{P}(X_t=0)\,\mathbb{P}(\text{no arrivals during } [t,t+h] \mid X_t=0)\\
&+ \mathbb{P}(X_t=1)\,\mathbb{P}(\text{no arrivals and 1 departure during } [t,t+h] \mid X_t=1)\\
&+ \mathbb{P}(\text{two or more events of arrivals or departures})\\
&= \mathbb{P}(X_t=0)(1-\lambda h+o(h))+\mathbb{P}(X_t=1)(1-\lambda h+o(h))(\mu h+o(h))+o(h)\\
&= \mathbb{P}(X_t=0)(1-\lambda h)+\mathbb{P}(X_t=1)\mu h+o(h).
\end{aligned}$$

Rearranging gives

$$P_0(t+h)-P_0(t)=-\lambda h P_0(t)+\mu h P_1(t)+o(h).$$

Dividing both sides by h and letting $h\to 0+$ gives

$$P_0'(t)=-\lambda P_0(t)+\mu P_1(t). \tag{5.1}$$

For $n\geq 1$, we have

$$\begin{aligned}
\mathbb{P}(X_{t+h}=n) &= \mathbb{P}(X_t=n-1)\,\mathbb{P}(1 \text{ arrival and no departures during } [t,t+h] \mid X_t=n-1)\\
&+ \mathbb{P}(X_t=n)\,\mathbb{P}(\text{no arrivals and no departures during } [t,t+h] \mid X_t=n)\\
&+ \mathbb{P}(X_t=n+1)\,\mathbb{P}(\text{no arrivals and 1 departure during } [t,t+h] \mid X_t=n+1)\\
&+ \mathbb{P}(\text{two or more events of arrivals or departures})\\
&= \mathbb{P}(X_t=n-1)(\lambda h+o(h))(1-\mu h+o(h))\\
&+ \mathbb{P}(X_t=n)(1-\lambda h+o(h))(1-\mu h+o(h))\\
&+ \mathbb{P}(X_t=n+1)(1-\lambda h+o(h))(\mu h+o(h))\\
&+ o(h)\\
&= \mathbb{P}(X_t=n-1)\lambda h+\mathbb{P}(X_t=n)(1-\lambda h-\mu h)+\mathbb{P}(X_t=n+1)\mu h+o(h).
\end{aligned}$$

Rearranging gives

$$P_n(t+h)-P_n(t)=\lambda h P_{n-1}(t)-(\lambda+\mu)h P_n(t)+\mu h P_{n+1}(t)+o(h).$$

Dividing both sides by h and letting $h\to 0+$ gives

$$P_n'(t)=\lambda P_{n-1}(t)-(\lambda+\mu)P_n(t)+\mu P_{n+1}(t). \tag{5.2}$$

We look for a stationary distribution of queue size. In the stationary situation, $P_n(t)$ will not change with t and so we may write $P_n(t)=P_n$. Noting that $P_n'(t)=0$, we get from (5.1) and (5.2) that

$$\begin{aligned}
-\lambda P_0+\mu P_1 &= 0,\\
\lambda P_{n-1}-(\lambda+\mu)P_n+\mu P_{n+1} &= 0, \quad n\geq 1.
\end{aligned}$$

Rearranging gives

$$\begin{aligned}
-\lambda P_0+\mu P_1 &= 0,\\
-\lambda P_n+\mu P_{n+1} &= -\lambda P_{n-1}+\mu P_n, \quad n\geq 1.
\end{aligned}$$

These imply

$$-\lambda P_n + \mu P_{n+1} = 0,$$

namely

$$P_{n+1} = \frac{\lambda}{\mu} P_n \quad \text{for all } n \geq 0.$$

Therefore

$$P_n = \Big(\frac{\lambda}{\mu}\Big)^n P_0 \quad \text{for all } n \geq 0.$$

The probabilities are in geometric progression. The series will converge if and only if the common ratio λ/μ is less than 1. This means that $\lambda < \mu$, i.e. that the rate at which the customers arrive is less than the rate at which the services are completed. There is a stationary distribution only in this case. If $\lambda \geq \mu$ the queue size will tend to infinity, because the server is unable to cope.

When $\lambda < \mu$, we have

$$1 = \sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} \Big(\frac{\lambda}{\mu}\Big)^n = \frac{P_0}{1-\lambda/\mu} = \frac{\mu P_0}{\mu - \lambda}.$$

This gives $P_0 = (\mu - \lambda)/\lambda$ and the stationary distribution

$$P_n = \frac{\mu - \lambda}{\lambda} \Big(\frac{\lambda}{\mu}\Big)^n, \quad n \geq 0.$$

Let X denote the queue size at the stationary situation. The mean queue size is

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} n P_n = \frac{\mu - \lambda}{\lambda} \sum_{n=1}^{\infty} n \Big(\frac{\lambda}{\mu}\Big)^n.$$

Recalling that $\sum_{n=1}^{n} n x^n = x/(1-x)^2$ for $x \in (0,1)$, we get

$$\mathbb{E}(X) = \frac{\mu - \lambda}{\lambda} \; \frac{\lambda/\mu}{(1-\lambda/\mu)^2} = \frac{\mu}{\mu - \lambda}.$$

Let W denote the number of the customers who are waiting for service at the stationary situation. This is the number of the customers who are really queueing. Clearly, W has the probability distribution

$$\mathbb{P}(W = 0) = P_0 + P_1 \text{ and } \mathbb{P}(W = n) = P_{n+1} \text{ for } n \geq 1.$$

So the mean of the *real queue size* is

$$\mathbb{E}(W) = \sum_{n=0}^{\infty} n \mathbb{P}(W = n) = \sum_{n=1}^{\infty} n P_{n+1} = \frac{\mu - \lambda}{\mu} \sum_{n=1}^{\infty} n \Big(\frac{\lambda}{\mu}\Big)^n = \frac{\mu - \lambda}{\mu} \; \frac{\lambda/\mu}{(1-\lambda/\mu)^2} = \frac{\lambda}{\mu - \lambda}.$$

5.2 M/M/1 with baulking

The same model except that any customer who would arrive when the queue size is a given number K does not actually join the queue, but disappears from the system. This can be used to model

the situation of a waiting room of limited capacity $(K-1)$. The same equations as before, namely (5.1 and (5.2), work for $n \leq K-1$. But, for $n = K$,

$$\begin{aligned}
\mathbb{P}(X_{t+h} = K) &= \mathbb{P}(X_t = K-1)\,\mathbb{P}(\text{1 arrival and no departures during } [t, t+h] \mid X_t = K-1) \\
&+ \mathbb{P}(X_t = K)\,\mathbb{P}(\text{no departures during } [t, t+h] \mid X_t = K) \\
&+ \mathbb{P}(\text{two or more events of arrivals or departures}) \\
&= \mathbb{P}(X_t = K-1)(\lambda h + o(h))(1 - \mu h + o(h)) \\
&+ \mathbb{P}(X_t = K)(1 - \mu h + o(h)) \\
&+ o(h) \\
&= \mathbb{P}(X_t = K-1)\lambda h + \mathbb{P}(X_t = K)(1 - \mu h) + o(h).
\end{aligned}$$

Rearranging gives

$$P_K(t+h) - P_K(t) = \lambda h P_{K-1}(t) - \mu h P_K(t) + o(h).$$

Dividing both sides by h and letting $h \to 0+$ gives

$$P_K'(t) = \lambda P_{K-1}(t) - \mu P_K(t). \tag{5.3}$$

For the stationary distribution P_n $(0 \leq n \leq K)$, we have

$$\begin{aligned}
-\lambda P_0 + \mu P_1 &= 0, \\
\lambda P_{n-1} - (\lambda + \mu) P_n + \mu P_{n+1} &= 0, \quad 1 \leq n \leq K-1, \\
\lambda P_{K-1} - \mu P_K &= 0.
\end{aligned}$$

Rearranging gives

$$\begin{aligned}
-\lambda P_0 + \mu P_1 &= 0, \\
-\lambda P_n + \mu P_{n+1} &= -\lambda P_{n-1} + \mu P_n, \quad 1 \leq n \leq K-1, \\
-\lambda P_{K-1} + \mu P_K &= 0.
\end{aligned}$$

These imply

$$-\lambda P_n + \mu P_{n+1} = 0,$$

namely

$$P_{n+1} = \frac{\lambda}{\mu} P_n \quad \text{for } 0 \leq n \leq K-1.$$

Therefore

$$P_n = \Big(\frac{\lambda}{\mu}\Big)^n P_0 \quad \text{for } 0 \leq n \leq K.$$

So the probabilities form a (finite) geometric progression with common ratio λ/μ. The possible queue sizes are $\{0, 1, \ldots, K\}$. There is no restriction on λ, μ. When $\lambda \neq \mu$, it is easy to show

$$P_n = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}} \Big(\frac{\lambda}{\mu}\Big)^n, \quad 0 \leq n \leq K.$$

Hence the probability that a potential customer baulks (i.e. tries to arrive when the queue size is K and so is lost from the queue) is

$$P_K = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}} \Big(\frac{\lambda}{\mu}\Big)^K.$$

When $\lambda = \mu$, we have the uniform distribution

$$P_n = \frac{1}{1+K}, \quad 0 \leq n \leq K.$$

5.3 M/M/∞

Here there are infinitely many servers. It can be used to model a telephone switchboard where there are plenty of lines available for the calls that may be made. There is no waiting, because each call is connected as soon as it arrives. The queue size is just the number of calls in progress.

In the same way as in M/M/1, we can show

$$P_0'(t) = -\lambda P_0(t) + \mu P_1(t). \tag{5.4}$$

For $n \geq 1$, we observe that when the queue size is n, the probability that one of the calls is finished in a short interval of length h is $n\mu h + o(h)$. Therefore

$$\begin{aligned}
\mathbb{P}(X_{t+h} = n) &= \mathbb{P}(X_t = n-1)\,\mathbb{P}(\text{1 arrival and no departures during } [t, t+h] \mid X_t = n-1) \\
&+ \mathbb{P}(X_t = n)\,\mathbb{P}(\text{no arrivals and no departures during } [t, t+h] \mid X_t = n) \\
&+ \mathbb{P}(X_t = n+1)\,\mathbb{P}(\text{no arrivals and 1 departure during } [t, t+h] \mid X_t = n+1) \\
&+ \mathbb{P}(\text{two or more events of arrivals or departures}) \\
&= \mathbb{P}(X_t = n-1)(\lambda h + o(h))(1 - (n-1)\mu h + o(h)) \\
&+ \mathbb{P}(X_t = n)(1 - \lambda h + o(h))(1 - n\mu h + o(h)) \\
&+ \mathbb{P}(X_t = n+1)(1 - \lambda h + o(h))((n+1)\mu h + o(h)) \\
&+ o(h) \\
&= \mathbb{P}(X_t = n-1)\lambda h + \mathbb{P}(X_t = n)(1 - \lambda h - n\mu h) + \mathbb{P}(X_t = n+1)(n+1)\mu h + o(h).
\end{aligned}$$

Rearranging gives

$$P_n(t+h) - P_n(t) = \lambda h P_{n-1}(t) - (\lambda + n\mu) h P_n(t) + (n+1)\mu h P_{n+1}(t) + o(h).$$

Dividing both sides by h and letting $h \to 0+$ gives

$$P_n'(t) = \lambda P_{n-1}(t) - (\lambda + n\mu) P_n(t) + (n+1)\mu P_{n+1}(t). \tag{5.5}$$

As before we look for a stationary solution P_n, for which the derivatives are 0. We get from (5.4) and (5.5) that

$$\begin{aligned}
-\lambda P_0 + \mu P_1 &= 0, \\
\lambda P_{n-1} - (\lambda + n\mu) P_n + (n+1)\mu P_{n+1} &= 0, \quad n \geq 1.
\end{aligned}$$

Rearranging gives

$$\begin{aligned}
-\lambda P_0 + \mu P_1 &= 0, \\
-\lambda P_n + (n+1)\mu P_{n+1} &= -\lambda P_{n-1} + n\mu P_n, \quad n \geq 1.
\end{aligned}$$

These imply

$$-\lambda P_n + (n+1)\mu P_{n+1} = 0,$$

namely

$$P_{n+1} = \frac{\lambda}{(n+1)\mu} P_n \quad \text{for all } n \geq 0.$$

Therefore

$$P_n = \frac{1}{n!}\Big(\frac{\lambda}{\mu}\Big)^n P_0 \quad \text{for all } n \geq 0.$$

But

$$\sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} \frac{1}{n!}\Big(\frac{\lambda}{\mu}\Big)^n = P_0 \exp(\lambda/\mu).$$

This implies $P_0 = \exp(-\lambda/\mu)$ and

$$P_n = \frac{1}{n!}\Big(\frac{\lambda}{\mu}\Big)^n \exp(-\lambda/\mu), \quad n \geq 0.$$

This is a Poisson distribution with parameter λ/μ.

5.4 M/M/s

Similar queue with s servers. This is a kind of hybrid of the M/M/1 and M/M/∞. So long as the number of customers in the system is no greater than s, there is no waiting. When further customers arrive they wait until a server becomes free. The completion of service rate is $n\mu$ when the queue size is $n \leq s$ and $s\mu$ when the queue size $n \geq s$. The differential equations for $P_n(t)$ are

$$\begin{aligned}
P_0'(t) &= -\lambda P_0(t) + \mu P_1(t), \\
P_n'(t) &= \lambda P_{n-1}(t) - (\lambda + n\mu)P_n(t) + (n+1)\mu P_{n+1}(t), \quad \text{for } 1 \leq n < s, \\
P_n'(t) &= \lambda P_{n-1}(t) - (\lambda + s\mu)P_n(t) + s\mu P_{n+1}(t), \quad \text{for } n \geq s.
\end{aligned}$$

The stationary distribution P_n is given by

$$\begin{aligned}
P_n &= \frac{1}{n!}\Big(\frac{\lambda}{\mu}\Big)^n P_0, \quad \text{for } 0 \leq n \leq s, \\
P_{s+k} &= \Big(\frac{\lambda}{s\mu}\Big)^k P_s, \quad \text{for } k \geq 1.
\end{aligned}$$

The first part of this is like a Poisson distribution and after $n = s$ it is a Geometric distribution with common ratio $\lambda/(s\mu)$. To get a stationary distribution we require $\lambda < s\mu$ for convergence. This says that the arrival rate must be less than the rate at which the s servers can work.

Example. A queue has three servers, arrival rate 2 and completion-of-service rate 1. Find the stationary distribution of queue size, the probability that a customer has to wait, and the mean waiting time.

Here $\lambda = 2,\ \mu = 1,\ s = 3$. For the stationary distribution, we have

$$P_1 = 2P_0, \quad P_2 = \frac{1}{2!}2^2 P_0, \quad P_3 = \frac{1}{3!}2^3 P_0 \quad \text{and} \quad P_{3+k} = \Big(\frac{2}{3}\Big)^k \frac{2^3}{3!} P_0 \quad \text{for } k \geq 1.$$

The sum of the probabilities is

$$P_0\Big[1 + 2 + 2 + \frac{4}{3}\Big(1 + \frac{2}{3} + \Big(\frac{2}{3}\Big)^2 + \cdots\Big)\Big] = P_0\Big(5 + \frac{4}{3}\frac{1}{1 - 2/3}\Big) = 9P_0.$$

So $P_0 = \frac{1}{9}$, and the stationary probabilities are

$$\frac{1}{9},\ \frac{2}{9},\ \frac{2}{9},\ \frac{4}{27},\ \frac{4}{27}\Big(\frac{2}{3}\Big),\ \frac{4}{27}\Big(\frac{2}{3}\Big)^2,\ \cdots.$$

The probability that a customer has to wait is probability that queue size is more than 2 when (s)he arrives, which is

$$1 - \Big(\frac{1}{9} + \frac{2}{9} + \frac{2}{9}\Big) = \frac{4}{9}.$$

To get the mean waiting time, we observe that if the quene size is less than 3 when a customer arrives, (s)he will be served immeadiately so the waiting time is 0; but if the queue size is 3, (s)he has to wait for 1/3 units of time as the mean time between service completions when all three servers are occupied is 1/3; and if the queue size is 4, (s)he has to wait for 2/3, and so on. Therefore

$$\begin{aligned}
\text{the mean waiting time} &= 0 \times (P_0 + P_1 + P_2) + \frac{1}{3}P_3 + \frac{2}{3}P_4 + \cdots \\
&= \frac{4}{27} \times \frac{1}{3}\Big[1 + 2\Big(\frac{2}{3}\Big) + 3\Big(\frac{2}{3}\Big)^2 + \cdots\Big] \\
&= \frac{4}{81} \times \frac{1}{(1-2/3)^2} \\
&= \frac{4}{9}.
\end{aligned}$$

■