11

Queues

Summary. A queue may be specified by its arrival process and its queueing and service disciplines. Queues with exponentially distributed interarrival and service times are the easiest to study, since such processes are Markov chains. Imbedded Markov chains allow an analysis when either the interarrival or the service times are exponentially distributed. The general case may be studied using the theory of random walks via Lindley's equation. Open and closed networks of Markovian queues may be studied via their stationary distributions.

11.1 Single-server queues

As summarized in Section 8.4, with each queue we can associate two sequences $\{X_n : n \geq 1\}$ and $\{S_n : n \geq 1\}$ of independent positive random variables, the X_n being interarrival times with common distribution function F_X and the S_n being service times with common distribution function F_S. We assume that customers arrive in the manner of a renewal process with interarrival times $\{X_n\}$, the nth customer arriving at time $T_n = X_1 + X_2 + \cdots + X_n$. Each arriving customer joins the line of customers who are waiting for the attention of the *single* server. When the nth customer reaches the head of this line he is served for a period of length S_n, after which he leaves the system. Let $Q(t)$ be the number of waiting customers at time t (including any customer whose service is in progress at t); clearly $Q(0) = 0$. Thus $Q = \{Q(t) : t \geq 0\}$ is a random process whose finite-dimensional distributions (fdds) are specified by the distribution functions F_X and F_S. We seek information about Q. For example, we may ask:

(a) When is Q a Markov chain, or when does Q contain an imbedded Markov chain?

(b) When is Q asymptotically stationary, in the sense that the distribution of $Q(t)$ settles down as $t \to \infty$?

(c) When does the queue length grow beyond all bounds, in that the server is not able to cope with the high rate of arrivals?

The answers to these and other similar questions take the form of applying conditions to the distribution functions F_X and F_S; the style of the analysis may depend on the types of these distributions functions. With this in mind, it is convenient to use a notation for the queueing system which incorporates information about F_X and F_S. The most common notation scheme describes each system by a triple A/B/s, where A describes F_X, B describes F_S, and s is the number of servers. Typically, A and B may each be one of the following:

$$\begin{aligned}
\mathrm{D}(d) &\equiv \text{almost surely concentrated at the value } d \text{ (D for `deterministic')},\\
\mathrm{M}(\lambda) &\equiv \text{exponential, parameter } \lambda \text{ (M for `Markovian')},\\
\Gamma(\lambda, k) &\equiv \text{gamma, parameters } \lambda \text{ and } k,\\
\mathrm{G} &\equiv \text{some general distribution, fixed but unspecified.}
\end{aligned}$$

(1) Example. M(λ)/M(μ)/1. Interarrival times are exponential with parameter λ and service times are exponential with parameter μ. Thus customers arrive in the manner of a Poisson process with intensity λ. The process $Q = \{Q(t)\}$ is a continuous-time Markov chain with state space $\{0, 1, 2, \dots\}$; this follows from the lack-of-memory property of the exponential distribution. Furthermore, such systems are the *only* systems whose queue lengths are homogeneous Markov chains. Why is this? ●

(2) Example. M(λ)/D(1)/1. Customers arrive in the manner of a Poisson process, and each requires a service time of constant length 1. The process Q is not a Markov chain, but we shall see later that there exists an imbedded discrete-time Markov chain $\{Q_n : n \geq 0\}$ whose properties provide information about Q. ●

(3) Example. G/G/1. In this case we have no special information about F_X or F_S. Some authors denote this system by GI/G/1, reserving the title G/G/1 to denote a more complicated system in which the interarrival times may not be independent. ●

The notation M(λ) is sometimes abbreviated to M alone. Thus Example (1) becomes M/M/1; this slightly unfortunate abbreviation does *not* imply that F_X and F_S are the same. A similar remark holds for systems described as G/G/1.

Broadly speaking, there are two types of statement to be made about the queue Q:

(a) 'time-dependent' statements, which contain information about the queue for finite values of t;

(b) 'limiting' results, which discuss the asymptotic properties of the queue as $t \to \infty$. These include conditions for the queue length to grow beyond all bounds.

Statements of the former type are most easily made about M(λ)/M(μ)/1, since this is the only Markovian system; such conclusions are more elusive for more general systems, and we shall generally content ourselves with the asymptotic properties of such queues.

In the subsequent sections we explore the systems M/M/1, M/G/1, G/M/1, and G/G/1, in that order. For the reader's convenience, we present these cases roughly in order of increasing difficulty. This is not really satisfactory, since we are progressing from the specific to the general, so we should like to stress that queues with Markovian characteristics are very special systems and that their properties do not always indicate features of more general systems.

Here is a final piece of notation.

(4) Definition. The **traffic intensity** ρ of a queue is defined as $\rho = \mathbb{E}(S)/\mathbb{E}(X)$, the ratio of the mean of a typical service time to the mean of a typical interarrival time.

We assume throughout that neither $\mathbb{E}(S)$ nor $\mathbb{E}(X)$ takes the value zero or infinity.

We shall see that queues behave in qualitatively different manners depending on whether $\rho < 1$ or $\rho > 1$. In the latter case, service times exceed interarrival times, on average, and the queue length grows beyond all bounds with probability 1; in the former case, the queue attains an equilibrium as $t \to \infty$. It is a noteworthy conclusion that the threshold between instability and stability depends on the mean values of F_X and F_S alone.

11.2 M/M/1

The queue M(λ)/M(μ)/1 is very special in that Q is a continuous-time Markov chain. Furthermore, reference to (6.11.1) reminds us that Q is a birth–death process with birth and death rates given by

$$\lambda_n = \lambda \quad \text{for all } n, \qquad \mu_n = \begin{cases} \mu & \text{if } n \geq 1, \\ 0 & \text{if } n = 0. \end{cases}$$

The probabilities $p_n(t) = \mathbb{P}(Q(t) = n)$ satisfy the Kolmogorov forward equations in the usual way:

$$\text{(1)} \qquad \frac{dp_n}{dt} = \lambda p_{n-1}(t) - (\lambda + \mu)p_n(t) + \mu p_{n+1}(t) \quad \text{for} \quad n \geq 1,$$

$$\text{(2)} \qquad \frac{dp_0}{dt} = -\lambda p_0(t) + \mu p_1(t),$$

subject to the boundary conditions $p_n(0) = \delta_{0n}$, the Kronecker delta. It is slightly tricky to solve these equations, but routine methods provide the answer after some manipulation. There are at least two possible routes: either use generating functions or use Laplace transforms with respect to t. We proceed in the latter way here, and define the Laplace transform† of p_n by

$$\widehat{p}_n(\theta) = \int_0^\infty e^{-\theta t} p_n(t)\, dt.$$

(3) **Theorem.** *We have that* $\widehat{p}_n(\theta) = \theta^{-1}[1 - \alpha(\theta)]\alpha(\theta)^n$ *where*

$$\text{(4)} \qquad \alpha(\theta) = \frac{(\lambda + \mu + \theta) - \sqrt{(\lambda + \mu + \theta)^2 - 4\lambda\mu}}{2\mu}.$$

The actual probabilities $p_n(t)$ can be deduced in terms of Bessel functions. It turns out that $p_n(t) = K_n(t) - K_{n+1}(t)$ where

$$K_n(t) = \int_0^t (\lambda/\mu)^{\frac{1}{2}n} n s^{-1} e^{-s(\lambda+\mu)} I_n\big(2s\sqrt{\lambda\mu}\big)\, ds$$

and $I_n(x)$ is a modified Bessel function (see Feller 1971, p. 482), defined to be the coefficient of z^n in the power series expansion of $\exp[\frac{1}{2}x(z + z^{-1})]$. See Exercise (5) for another representation of $p_n(t)$.

†Do not confuse $\widehat{p}_n$ with the Laplace–Stieltjes transform $p_n^*(\theta) = \int_0^\infty e^{-\theta t}\, dp_n(t)$.

Proof. Transform (1) and (2) to obtain

$$\mu\widehat{p}_{n+1} - (\lambda + \mu + \theta)\widehat{p}_n + \lambda\widehat{p}_{n-1} = 0 \quad \text{for} \quad n \geq 1, \tag{5}$$

$$\mu\widehat{p}_1 - (\lambda + \theta)\widehat{p}_0 = -1, \tag{6}$$

where we have used the fact (see equation (14) of Appendix I) that

$$\int_0^\infty e^{-\theta t}\frac{dp_n}{dt}\,dt = \theta\widehat{p}_n - \delta_{0n}, \qquad \text{for all } n.$$

Equation (5) is an ordinary difference equation, and standard techniques (see Appendix I) show that it has a unique solution which is bounded as $\theta \to \infty$ and which is given by

$$\widehat{p}_n(\theta) = \widehat{p}_0(\theta)\alpha(\theta)^n \tag{7}$$

where α is given by (4). Substitute (7) into (6) to deduce that $\widehat{p}_0(\theta) = [1 - \alpha(\theta)]/\theta$ and the proof is complete. Alternatively, $\widehat{p}_0(\theta)$ may be calculated from the fact that $\sum_n p_n(t) = 1$, implying that $\sum_n \widehat{p}_n(\theta) = \theta^{-1}$. ■

The asymptotic behaviour of $Q(t)$ as $t \to \infty$ is deducible from (3), but more direct methods yield the answer more quickly. Remember that Q is a Markov chain.

(8) Theorem. *Let $\rho = \lambda/\mu$ be the traffic intensity.*

(a) *If $\rho < 1$, then $\mathbb{P}(Q(t) = n) \to (1-\rho)\rho^n = \pi_n$ for $n \geq 0$, where π is the unique stationary distribution.*

(b) *If $\rho \geq 1$, there is no stationary distribution, and $\mathbb{P}(Q(t) = n) \to 0$ for all n.*

The result is very natural. It asserts that the queue settles down into equilibrium if and only if interarrival times exceed service times on average. We shall see later that if $\rho > 1$ then $\mathbb{P}(Q(t) \to \infty \text{ as } t \to \infty) = 1$, whilst if $\rho = 1$ then the queue length experiences wild oscillations with no reasonable bound on their magnitudes.

Proof. The process Q is an irreducible chain. Let us try to find a stationary distribution, as done in greater generality at (6.11.2) for birth–death chains. Let $t \to \infty$ in (1) and (2) to find that the mass function π is a stationary distribution if and only if

$$\begin{aligned} \pi_{n+1} - (1+\rho)\pi_n + \rho\pi_{n-1} &= 0 \quad \text{for} \quad n \geq 1, \\ \pi_1 - \rho\pi_0 &= 0. \end{aligned} \tag{9}$$

(The operation of taking limits is justifiable by (6.9.20) and the uniformity of Q.) The general solution to (9) is

$$\pi_n = \begin{cases} A + B\rho^n & \text{if } \rho \neq 1, \\ A + Bn & \text{if } \rho = 1, \end{cases}$$

where A and B are arbitrary constants. Thus the only bounded solution to (9) with bounded sum is

$$\pi_n = \begin{cases} B\rho^n & \text{if } \rho < 1, \\ 0 & \text{if } \rho \geq 1. \end{cases}$$

Hence, if $\rho < 1$, $\pi_n = (1-\rho)\rho^n$ is a stationary distribution, whilst if $\rho \geq 1$ then there exists no stationary distribution. By Theorem (6.9.21), the proof is complete. ■

There is an alternative derivation of the asymptotic behaviour (8) of Q, which has other consequences also. Let U_n be the epoch of time at which the nth change in Q occurs. That is to say

$$U_0 = 0, \quad U_{n+1} = \inf\{t > U_n : Q(t) \neq Q(U_n+)\}.$$

Now let $Q_n = Q(U_n+)$ be the number of waiting customers immediately after the nth change in Q. Clearly $\{Q_n : n \geq 0\}$ is a random walk on the non-negative integers, with

$$Q_{n+1} = \begin{cases} Q_n + 1 & \text{with probability } \dfrac{\lambda}{\lambda+\mu} = \dfrac{\rho}{1+\rho}, \\ Q_n - 1 & \text{with probability } \dfrac{\mu}{\lambda+\mu} = \dfrac{1}{1+\rho}, \end{cases}$$

whenever $Q_n \geq 1$ (see paragraph A after (6.11.12) for a similar result for another birth–death process). When $Q_n = 0$ we have that

$$\mathbb{P}(Q_{n+1} = 1 \mid Q_n = 0) = 1,$$

so that the walk leaves 0 immediately after arriving there; it is only in this regard that the walk differs from the random walk (6.4.15) with a retaining barrier. Look for stationary distributions of the walk in the usual way to find (*exercise*) that there exists such a distribution if and only if $\rho < 1$, and it is given by

(10)
$$\pi_0 = \tfrac{1}{2}(1-\rho), \quad \pi_n = \tfrac{1}{2}(1-\rho^2)\rho^{n-1} \quad \text{for} \quad n \geq 1.$$

Follow the argument of Example (6.4.15) to find that

$$\{Q_n\} \text{ is } \begin{cases} \text{non-null persistent} & \text{if } \rho < 1, \\ \text{null persistent} & \text{if } \rho = 1, \\ \text{transient} & \text{if } \rho > 1. \end{cases}$$

Equation (10) differs from the result of (8) because the walk $\{Q_n\}$ and the process Q behave differently at the state 0. It is possible to deduce (8) from (10) by taking account of the times which elapse between the jumps of the walk (see Exercise (11.2.6) for details). It is clear now that $Q_n \to \infty$ almost surely as $n \to \infty$ if $\rho > 1$, whilst $\{Q_n\}$ experiences large fluctuations in the symmetric case $\rho = 1$.

Exercises for Section 11.2

1. Consider a random walk on the non-negative integers with a reflecting barrier at 0, and which moves rightwards or leftwards with respective probabilities $\rho/(1+\rho)$ and $1/(1+\rho)$; when at 0, the particle moves to 1 at the next step. Show that the walk has a stationary distribution if and only if $\rho < 1$, and in this case the unique such distribution $\boldsymbol{\pi}$ is given by $\pi_0 = \frac{1}{2}(1-\rho)$, $\pi_n = \frac{1}{2}(1-\rho^2)\rho^{n-1}$ for $n \geq 1$.

2. Suppose now that the random walker of Exercise (1) delays its steps in the following way. When at the point n, it waits a random length of time having the exponential distribution with parameter θ_n before moving to its next position; different 'holding times' are independent of each other and of further information concerning the steps of the walk. Show that, subject to reasonable assumptions on the θ_n, the ensuing continuous-time process settles into an equilibrium distribution $\boldsymbol{\nu}$ given by $\nu_n = C\pi_n/\theta_n$ for some appropriate constant C.

By applying this result to the case when $\theta_0 = \lambda, \theta_n = \lambda + \mu$ for $n \geq 1$, deduce that the equilibrium distribution of the M(λ)/M(μ)/1 queue is $\nu_n = (1-\rho)\rho^n$, $n \geq 0$, where $\rho = \lambda/\mu < 1$.

3. Waiting time. Consider a M(λ)/M(μ)/1 queue with $\rho = \lambda/\mu$ satisfying $\rho < 1$, and suppose that the number $Q(0)$ of people in the queue at time 0 has the stationary distribution $\pi_n = (1-\rho)\rho^n$, $n \geq 0$. Let W be the time spent by a typical new arrival before he begins his service. Show that the distribution of W is given by $\mathbb{P}(W \leq x) = 1 - \rho e^{-x(\mu-\lambda)}$ for $x \geq 0$, and note that $\mathbb{P}(W = 0) = 1 - \rho$.

4. A box contains i red balls and j lemon balls, and they are drawn at random without replacement. Each time a red (respectively lemon) ball is drawn, a particle doing a walk on $\{0, 1, 2, \dots\}$ moves one step to the right (respectively left); the origin is a retaining barrier, so that leftwards steps from the origin are suppressed. Let $\pi(n; i, j)$ be the probability that the particle ends at position n, having started at the origin. Write down a set of difference equations for the $\pi(n; i, j)$, and deduce that

$$\pi(n; i, j) = A(n; i, j) - A(n+1; i, j) \quad \text{for } i \leq j + n$$

where $A(n; i, j) = \binom{i}{n} \big/ \binom{j+n}{n}$.

5. Let Q be a M(λ)/M(μ)/1 queue with $Q(0) = 0$. Show that $p_n(t) = \mathbb{P}(Q(t) = n)$ satisfies

$$p_n(t) = \sum_{i,j \geq 0} \pi(n; i, j) \left(\frac{(\lambda t)^i e^{-\lambda t}}{i!} \right) \left(\frac{(\mu t)^j e^{-\mu t}}{j!} \right)$$

where the $\pi(n; i, j)$ are given in the previous exercise.

6. Let $Q(t)$ be the length of an M(λ)/M(μ)/1 queue at time t, and let $Z = \{Z_n\}$ be the jump chain of Q. Explain how the stationary distribution of Q may be derived from that of Z, and vice versa.

7. Tandem queues. Two queues have one server each, and all service times are independent and exponentially distributed, with parameter μ_i for queue i. Customers arrive at the first queue at the instants of a Poisson process of rate λ ($< \min\{\mu_1, \mu_2\}$), and on completing service immediately enter the second queue. The queues are in equilibrium. Show that:

(a) the output of the first queue is a Poisson process with intensity λ, and that the departures before time t are independent of the length of the queue at time t,

(b) the waiting times of a given customer in the two queues are not independent.

11.3 M/G/1

M/M/1 is the only queue which is a Markov chain; the analysis of other queueing systems requires greater ingenuity. If either interarrival times or service times are exponentially distributed then the general theory of Markov chains still provides a method for studying the queue. The reason is that, for each of these two cases, we may find a discrete-time Markov chain which is imbedded in the continuous-time process Q. We consider M/G/1 in this section, which is divided into three parts dealing with equilibrium theory, the 'waiting time' of a typical customer, and the length of a typical 'busy period' during which the server is continuously occupied.

(A) Asymptotic queue length. Consider M(λ)/G/1. Customers arrive in the manner of a Poisson process with intensity λ. Let D_n be the time of departure of the nth customer from the system, and let $Q(D_n)$ be the number of customers which he leaves behind him in the system on his departure (really, we should write $Q(D_n+)$ instead of $Q(D_n)$ to make clear that the departing customer is not included). Then $Q(D) = \{Q(D_n) : n \geq 1\}$ is a sequence of random variables. What can we say about a typical increment $Q(D_{n+1}) - Q(D_n)$? If $Q(D_n) > 0$, the $(n+1)$th customer begins his service time immediately at time D_n; during this service time of length S_{n+1}, a random number, U_n say, of customers arrive and join the waiting line. Therefore the $(n+1)$th customer leaves $U_n + Q(D_n) - 1$ customers behind him as he departs. That is,

$$Q(D_{n+1}) = U_n + Q(D_n) - 1 \quad \text{if} \quad Q(D_n) > 0. \tag{1}$$

If $Q(D_n) = 0$, the server must wait for the $(n+1)$th arrival before she† sets to work again. When this service is complete, the $(n+1)$th customer leaves exactly U_n customers behind him where U_n is the number of arrivals during his service time, as before. That is,

$$Q(D_{n+1}) = U_n \quad \text{if} \quad Q(D_n) = 0. \tag{2}$$

Combine (1) and (2) to obtain

$$Q(D_{n+1}) = U_n + Q(D_n) - h(Q(D_n)) \tag{3}$$

where h is defined by

$$h(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x \le 0. \end{cases}$$

Equation (3) holds for any queue. However, in the case of $M(\lambda)/G/1$ the random variable U_n depends *only* on the length of time S_{n+1}, and is independent of $Q(D_n)$, because of the special properties of the Poisson process of arrivals. We conclude from (3) that $Q(D)$ is a Markov chain.

(4) **Theorem.** *The sequence $Q(D)$ is a Markov chain with transition matrix*

$$\mathbf{P}_D = \begin{pmatrix} \delta_0 & \delta_1 & \delta_2 & \cdots \\ \delta_0 & \delta_1 & \delta_2 & \cdots \\ 0 & \delta_0 & \delta_1 & \cdots \\ 0 & 0 & \delta_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where

$$\delta_j = \mathbb{E}\left(\frac{(\lambda S)^j}{j!} e^{-\lambda S}\right)$$

and S is a typical service time.

The quantity δ_j is simply the probability that exactly j customers join the queue during a typical service time.

Proof. We need to show that $\mathbf{P}_D$ is the correct transition matrix. In the notation of Chapter 6,

$$p_{0j} = \mathbb{P}\big(Q(D_{n+1}) = j \mid Q(D_n) = 0\big) = \mathbb{E}\big(\mathbb{P}(U_n = j \mid S)\big)$$

†Recall the convention of Section 8.4 that customers are male and servers female.

where $S = S_{n+1}$ is the service time of the $(n+1)$th customer. Thus

$$p_{0j} = \mathbb{E}\left(\frac{(\lambda S)^j}{j!} e^{-\lambda S}\right) = \delta_j$$

as required, since, conditional on S, U_n has the Poisson distribution with parameter λS. Likewise, if $i \geq 1$ then

$$p_{ij} = \mathbb{E}\big(\mathbb{P}(U_n = j - i + 1 \mid S)\big) = \begin{cases} \delta_{j-i+1} & \text{if } j - i + 1 \geq 0, \\ 0 & \text{if } j - i + 1 < 0. \end{cases} \qquad \blacksquare$$

This result enables us to observe the behaviour of the process $Q = \{Q(t)\}$ by evaluating it at the time epochs $D_1, D_2, \dots$ and using the theory of Markov chains. It is important to note that this course of action provides reliable information about the asymptotic behaviour of Q only because $D_n \to \infty$ almost surely as $n \to \infty$. The asymptotic behaviour of $Q(D)$ is described by the next theorem.

(5) Theorem. *Let $\rho = \lambda\mathbb{E}(S)$ be the traffic intensity.*

(a) *If $\rho < 1$, then $Q(D)$ is ergodic with a unique stationary distribution $\boldsymbol{\pi}$, having generating function*

$$G(s) = \sum_j \pi_j s^j = (1-\rho)(s-1)\frac{M_S(\lambda(s-1))}{s - M_S(\lambda(s-1))},$$

where M_S is the moment generating function of a typical service time.

(b) *If $\rho > 1$, then $Q(D)$ is transient.*

(c) *If $\rho = 1$, then $Q(D)$ is null persistent.*

Here are some consequences of this theorem.

(6) Busy period. A *busy period* is a period of time during which the server is continuously occupied. The length B of a typical busy period behaves similarly to the time B' between successive visits of the chain $Q(D)$ to the state 0. Thus

$$\begin{array}{llll} \text{if} \quad \rho < 1 & \text{then} \quad \mathbb{E}(B) < \infty, & & \\ \text{if} \quad \rho = 1 & \text{then} \quad \mathbb{E}(B) = \infty, & \mathbb{P}(B = \infty) = 0, \\ \text{if} \quad \rho > 1 & \text{then} \quad \mathbb{P}(B = \infty) > 0. & \end{array}$$

See the forthcoming Theorems (17) and (18) for more details about B.

(7) Stationarity of Q. It is an immediate consequence of (5) and Theorem (6.4.17) that $Q(D)$ is asymptotically stationary whenever $\rho < 1$. In this case it can be shown that Q is asymptotically stationary also, in that $\mathbb{P}(Q(t) = n) \to \pi_n$ as $t \to \infty$. Roughly speaking, this is because $Q(t)$ forgets more and more about its origins as t becomes larger.

Proof of (5). The sequence $Q(D)$ is irreducible and aperiodic. We proceed by applying Theorems (6.4.3), (6.4.10), and (6.4.13).

(a) Look for a root of the equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}_D$. Any such $\boldsymbol{\pi}$ satisfies

$$\textbf{(8)} \qquad \pi_j = \pi_0\delta_j + \sum_{i=1}^{j+1} \pi_i\delta_{j-i+1}, \quad \text{for} \quad j \geq 0.$$

First, note that if π_0 (≥ 0) is given, then (8) has a unique solution $\boldsymbol{\pi}$. Furthermore, this solution has non-negative entries. To see this, add equations (8) for $j = 0, 1, \dots, n$ and solve for π_{n+1} to obtain

$$\textbf{(9)} \qquad \pi_{n+1}\delta_0 = \pi_0\epsilon_n + \sum_{i=1}^{n} \pi_i\epsilon_{n-i+1} \quad \text{for} \quad n \geq 0$$

where

$$\epsilon_n = 1 - \delta_0 - \delta_1 - \cdots - \delta_n > 0 \quad \text{because} \quad \sum_j \delta_j = 1.$$

From (9), $\pi_{n+1} \geq 0$ whenever $\pi_i \geq 0$ for all $i \leq n$, and so

$$\textbf{(10)} \qquad \pi_n \geq 0 \quad \text{for all } n$$

if $\pi_0 \geq 0$, by induction. Return to (8) to see that the generating functions

$$G(s) = \sum_j \pi_j s^j, \quad \Delta(s) = \sum_j \delta_j s^j,$$

satisfy

$$G(s) = \pi_0\Delta(s) + \frac{1}{s}[G(s) - \pi_0]\Delta(s)$$

and therefore

$$\textbf{(11)} \qquad G(s) = \frac{\pi_0(s-1)\Delta(s)}{s - \Delta(s)}.$$

The vector $\boldsymbol{\pi}$ is a stationary distribution if and only if $\pi_0 > 0$ and $\lim_{s\uparrow 1} G(s) = 1$. Apply L'Hôpital's rule to (11) to discover that

$$\pi_0 = 1 - \Delta'(1) > 0$$

is a necessary and sufficient condition for this to occur, and thus there exists a stationary distribution if and only if

$$\textbf{(12)} \qquad \Delta'(1) < 1.$$

However,

$$\Delta(s) = \sum_j s^j \mathbb{E}\left(\frac{(\lambda S)^j}{j!} e^{-\lambda S}\right) = \mathbb{E}\left(e^{-\lambda S} \sum_j \frac{(\lambda s S)^j}{j!}\right)$$
$$= \mathbb{E}(e^{\lambda S(s-1)}) = M_S\big(\lambda(s-1)\big)$$

where M_S is the moment generating function of S. Thus

$$\Delta'(1) = \lambda M_S'(0) = \lambda\mathbb{E}(S) = \rho \tag{13}$$

and condition (12) becomes $\rho < 1$. Thus $Q(D)$ is non-null persistent if and only if $\rho < 1$. In this case, $G(s)$ takes the form given in (5a).

(b) Recall from Theorem (6.4.10) that $Q(D)$ is transient if and only if there is a bounded non-zero solution $\{y_j : j \geq 1\}$ to the equation

$$y_1 = \sum_{i=1}^{\infty} \delta_i y_i, \tag{14}$$

$$y_j = \sum_{i=0}^{\infty} \delta_i y_{j+i-1} \quad \text{for} \quad j \geq 2. \tag{15}$$

If $\rho > 1$ then $\Delta(s)$ satisfies

$$0 < \Delta(0) < 1, \quad \Delta(1) = 1, \quad \Delta'(1) > 1,$$

from (13). Draw a picture (or see Figure 5.1) to see that there exists a number $b \in (0, 1)$ such that $\Delta(b) = b$. By inspection, $y_j = 1 - b^j$ solves (14) and (15), and (b) is shown.

(c) $Q(D)$ is transient if $\rho > 1$ and non-null persistent if and only if $\rho < 1$. We need only show that $Q(D)$ is persistent if $\rho = 1$. But it is not difficult to see that $\{y_j : j \neq 0\}$ solves equation (6.4.14), when y_j is given by $y_j = j$ for $j \geq 1$, and the result follows. ■

(B) Waiting time. When $\rho < 1$ the imbedded queue length settles down into an equilibrium distribution π. Suppose that a customer joins the queue after some large time has elapsed. He will wait a period W of time before his service begins; W is called his *waiting time* (this definition is at odds with that used by some authors who include the customer's service time in W). The distribution of W should not vary much with the time of the customer's arrival since the system is 'nearly' in equilibrium.

(16) Theorem. *The waiting time W has moment generating function*

$$M_W(s) = \frac{(1-\rho)s}{\lambda + s - \lambda M_S(s)}$$

when the imbedded queue is in equilibrium.

Proof. The condition that the imbedded queue be in equilibrium amounts to the supposition that the length $Q(D)$ of the queue on the departure of a customer is distributed according to the stationary distribution π. Suppose that a customer waits for a period of length W and then is served for a period of length S. On departure he leaves behind him all those customers who have arrived during the period, length $W + S$, during which he was in the system. The number Q of such customers is Poisson distributed with parameter $\lambda(W + S)$, and so

$$\begin{aligned}
\mathbb{E}(s^Q) &= \mathbb{E}\big(\mathbb{E}(s^Q \mid W, S)\big) \\
&= \mathbb{E}(e^{\lambda(W+S)(s-1)}) \\
&= \mathbb{E}(e^{\lambda W(s-1)})\mathbb{E}(e^{\lambda S(s-1)}) \quad \text{by independence} \\
&= M_W\big(\lambda(s-1)\big)M_S\big(\lambda(s-1)\big).
\end{aligned}$$

However, Q has distribution π given by (5a) and the result follows. ■

(C) Busy period: a branching process. Finally, put yourself in the server's shoes. She may not be as interested in the waiting times of her customers as she is in the frequency of her tea breaks. Recall from (6) that a *busy period* is a period of time during which she is continuously occupied, and let B be the length of a typical busy period. That is, if the first customer arrives at time T_1 then

$$B = \inf\{t > 0 : Q(t + T_1) = 0\};$$

The quantity B is well defined whether or not $Q(D)$ is ergodic, though it may equal $+\infty$.

(17) Theorem. *The moment generating function M_B of B satisfies the functional equation*

$$M_B(s) = M_S\big(s - \lambda + \lambda M_B(s)\big).$$

It can be shown that this functional equation has a unique solution which is the moment generating function of a (possibly infinite) random variable (see Feller 1971, pp. 441, 473). The server may wish to calculate the probability

$$\mathbb{P}(B < \infty) = \lim_{x\to\infty} \mathbb{P}(B \le x)$$

that she is eventually free. It is no surprise to find the following, in agreement with (6).

(18) Theorem. *We have that*

$$\mathbb{P}(B < \infty)\begin{cases} = 1 & \text{if } \rho \le 1, \\ < 1 & \text{if } \rho > 1. \end{cases}$$

This may remind you of a similar result for the extinction probability of a branching process. This is no coincidence; we prove (17) and (18) by methods first encountered in the study of branching processes.

Proof of (17) and (18). Here is an imbedded branching process. Call customer C_2 an 'offspring' of customer C_1 if C_2 joins the queue while C_1 is being served. Since customers arrive in the manner of a Poisson process, the numbers of offspring of different customers are independent random variables. Therefore, the family tree of the 'offspring process' is that of a branching process. The mean number of offspring of a given customer is given in the notation of the proof of (5) as $\mu = \Delta'(1)$, whence $\mu = \rho$ by (13). The offspring process is ultimately extinct if and only if the queue is empty at some time later than the first arrival. That is,

$$\mathbb{P}(B < \infty) = \mathbb{P}(Z_n = 0 \text{ for some } n),$$

where Z_n is the number of customers in the nth generation of the process. We have by Theorem (5.4.5) that $\eta = \mathbb{P}(Z_n = 0 \text{ for some } n)$ satisfies

$$\eta = 1 \quad \text{if and only if} \quad \mu \le 1.$$

Therefore $\mathbb{P}(B < \infty) = 1$ if and only if $\rho \le 1$, as required for (18).

Each individual in this branching process has a service time; B is the sum of these service times. Thus

$$B = S + \sum_{j=1}^{Z} B_j \tag{19}$$

where S is the service time of the first customer, Z is the number of offspring of this customer, and B_j is the sum of the service times of the jth such offspring together with all his descendants in the offspring process (this is similar to the argument of Problem (5.12.11)). The two terms on the right side of (19) are *not* independent of each other; after all, if S is large then Z is likely to be large as well. However, condition on S to obtain

$$M_B(s) = \mathbb{E}\left(\mathbb{E}\left\{\exp\left[s\left(S + \sum_{j=1}^{Z} B_j\right)\right] \,\middle|\, S\right\}\right)$$

and remember that, conditional on Z, the random variables $B_1, B_2, \dots, B_Z$ are independent with the same distribution as B to obtain

$$M_B(s) = \mathbb{E}\big(e^{sS} G_{\mathrm{Po}(\lambda S)}\{M_B(s)\}\big)$$

where $G_{\mathrm{Po}(\mu)}$ is the probability generating function of the Poisson distribution with parameter μ. Therefore

$$M_B(s) = \mathbb{E}(e^{S(s-\lambda+\lambda M_B(s))})$$

as required. ■

Exercises for Section 11.3

1. Consider M(λ)/D(d)/1 where $\rho = \lambda d < 1$. Show that the mean queue length at moments of departure in equilibrium is $\frac{1}{2}\rho(2-\rho)/(1-\rho)$.

2. Consider M(λ)/M(μ)/1, and show that the moment generating function of a typical busy period is given by

$$M_B(s) = \frac{(\lambda + \mu - s) - \sqrt{(\lambda + \mu - s)^2 - 4\lambda\mu}}{2\lambda}$$

for all sufficiently small but positive values of s.

3. Show that, for a M/G/1 queue, the sequence of times at which the server passes from being busy to being free constitutes a renewal process.

11.4 G/M/1

The system G/M(μ)/1 contains an imbedded discrete-time Markov chain also, and this chain provides information about the properties of $Q(t)$ for large t. This section is divided into two parts, dealing with the asymptotic behaviour of $Q(t)$ and the waiting time distribution.

(A) Asymptotic queue length. This time, consider the epoch of time at which the nth customer *joins* the queue, and let $Q(A_n)$ be the number of individuals who are ahead of him in the system at the moment of his arrival. The quantity $Q(A_n)$ includes any customer whose service is in progress; more specifically, $Q(A_n) = Q(T_n-)$ where T_n is the instant of the nth arrival. The argument of the last section shows that

$$(1) \qquad Q(A_{n+1}) = Q(A_n) + 1 - V_n$$

where V_n is the number of departures from the system during the interval $[T_n, T_{n+1})$ between the nth and $(n+1)$th arrival. This time, V_n depends on $Q(A_n)$ since not more than $Q(A_n)+1$ individuals may depart during this interval. However, service times are exponentially distributed, and so, conditional upon $Q(A_n)$ and $X_{n+1} = T_{n+1} - T_n$, the random variable V_n has a truncated Poisson distribution

$$(2)\qquad \mathbb{P}(V_n = d \mid Q(A_n) = q,\ X_{n+1} = x) = \begin{cases} \dfrac{(\mu x)^d}{d!} e^{-\mu x} & \text{if } d \le q, \\ \displaystyle\sum_{m>q} \frac{(\mu x)^m}{m!} e^{-\mu x} & \text{if } d = q+1. \end{cases}$$

Anyway, given $Q(A_n)$, the random variable V_n is independent of the sequence $Q(A_1)$, $Q(A_2)$, $\dots$, $Q(A_{n-1})$, and so $Q(A) = \{Q(A_n) : n \ge 1\}$ is a Markov chain.

(3) Theorem. *The sequence $Q(A)$ is a Markov chain with transition matrix*

$$\mathbf{P}_A = \begin{pmatrix} 1-\alpha_0 & \alpha_0 & 0 & 0 & \dots \\ 1-\alpha_0-\alpha_1 & \alpha_1 & \alpha_0 & 0 & \dots \\ 1-\alpha_0-\alpha_1-\alpha_2 & \alpha_2 & \alpha_1 & \alpha_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where

$$\alpha_j = \mathbb{E}\left(\frac{(\mu X)^j}{j!} e^{-\mu X}\right)$$

and X is a typical interarrival time.

The quantity α_j is simply the probability that exactly j events of a Poisson process occur during a typical interarrival time.

Proof. This proceeds as for Theorem (11.3.4). ■

(4) Theorem. *Let $\rho = \{\mu\mathbb{E}(X)\}^{-1}$ be the traffic intensity.*

(a) *If $\rho < 1$, then $Q(A)$ is ergodic with a unique stationary distribution π given by*

$$\pi_j = (1-\eta)\eta^j \quad \textit{for} \quad j \ge 0$$

where η is the smallest positive root of $\eta = M_X(\mu(\eta-1))$ and M_X is the moment generating function of X.

(b) *If $\rho > 1$, then $Q(A)$ is transient.*

(c) *If $\rho = 1$, then $Q(A)$ is null persistent.*

If $\rho < 1$ then $Q(A)$ is asymptotically stationary. Unlike the case of M/G/1, however, the stationary distribution π given by (4a) need *not* be the limiting distribution of Q itself; to see an example of this, just consider D(1)/M/1.

Proof. Let Q_d be an M(μ)/G/1 queue whose service times have the same distribution as the interarrival times of Q (the queue Q_d is called the *dual* of Q, but more about that later). The traffic intensity ρ_d of Q_d satisfies

$$(5)\qquad \rho\rho_\mathrm{d} = 1.$$

From the results of Section 11.3, Q_{d} has an imbedded Markov chain $Q_{\mathrm{d}}(D)$, obtained from the values of Q_{d} at the epochs of time at which customers depart. We shall see that $Q(A)$ is non-null persistent (respectively transient) if and only if the imbedded chain $Q_{\mathrm{d}}(D)$ of Q_{d} is transient (respectively non-null persistent) and the results will follow immediately from Theorem (11.3.5) and its proof.

(a) Look for non-negative solutions $\boldsymbol{\pi}$ to the equation

$$\text{(6)} \qquad \boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}_A$$

which have sum $\boldsymbol{\pi}\mathbf{1}' = 1$. Expand (6), set

$$\text{(7)} \qquad y_j = \pi_0 + \pi_1 + \cdots + \pi_{j-1} \quad \text{for} \quad j \geq 1,$$

and remember that $\sum_j \alpha_j = 1$ to obtain

$$\text{(8)} \qquad y_1 = \sum_{i=1}^{\infty} \alpha_i y_i,$$

$$\text{(9)} \qquad y_j = \sum_{i=0}^{\infty} \alpha_i y_{j+i-1} \quad \text{for} \quad j \geq 2.$$

These are the same equations as (11.3.14) and (11.3.15) for Q_{d}. As in the proof of Theorem (11.3.5), it is easy to check that

$$\text{(10)} \qquad y_j = 1 - \eta^j$$

solves (8) and (9) whenever

$$A(s) = \sum_{j=0}^{\infty} \alpha_j s^j$$

satisfies $A'(1) > 1$, where η is the unique root in the interval $(0, 1)$ of the equation $A(s) = s$. However, write A in terms of M_X, as before, to find that $A(s) = M_X(\mu(s-1))$, giving

$$A'(1) = \rho_{\mathrm{d}} = \frac{1}{\rho}.$$

Combine (7) and (10) to find the stationary distribution for the case $\rho < 1$. If $\rho \geq 1$ then $\rho_{\mathrm{d}} \leq 1$ by (5), and so (8) and (9) have no bounded non-zero solution since otherwise $Q_{\mathrm{d}}(D)$ would be transient, contradicting Theorem (11.3.5). Thus $Q(A)$ is non-null persistent if and only if $\rho < 1$.

(b) To prove transience, we seek bounded non-zero solutions $\{y_j : j \geq 1\}$ to the equations

$$\text{(11)} \qquad y_j = \sum_{i=1}^{j+1} y_i \alpha_{j-i+1} \quad \text{for} \quad j \geq 1.$$

Suppose that $\{y_j\}$ satisfies (11), and that $y_1 \geq 0$. Define $\boldsymbol{\pi} = \{\pi_j : j \geq 0\}$ as follows:

$$\pi_0 = y_1\alpha_0, \quad \pi_1 = y_1(1-\alpha_0), \quad \pi_j = y_j - y_{j-1} \quad \text{for} \quad j \geq 2.$$

It is an easy exercise to show that π satisfies equation (11.3.8) with the δ_j replaced by the α_j throughout. But (11.3.8) possesses a non-zero solution with bounded sum if and only if $\rho_{\text{d}} < 1$, which is to say that $Q(A)$ is transient if and only if $\rho = 1/\rho_{\text{d}} > 1$.

(c) $Q(A)$ is transient if and only if $\rho > 1$, and is non-null persistent if and only if $\rho < 1$. If $\rho = 1$ then $Q(A)$ has no choice but null persistence. ■

(B) Waiting time. An arriving customer waits for just as long as the server needs to complete the service period in which she is currently engaged and to serve the other waiting customers. That is, the nth customer waits for a length W_n of time:

$$W_n = Z_1^* + Z_2 + Z_3 + \cdots + Z_{Q(A_n)} \quad \text{if} \quad Q(A_n) > 0$$

where Z_1^* is the *excess* (or *residual*) *service time* of the customer at the head of the queue, and $Z_2, Z_3, \dots, Z_{Q(A_n)}$ are the service times of the others. Given $Q(A_n)$, the Z_i are independent, but Z_1^* does not in general have the same distribution as $Z_2, Z_3, \dots$. In the case of G/M(μ)/1, however, the lack-of-memory property helps us around this difficulty.

(12) Theorem. *The waiting time W of an arriving customer has distribution*

$$\mathbb{P}(W \le x) = \begin{cases} 0 & \textit{if } x < 0, \\ 1 - \eta e^{-\mu(1-\eta)x} & \textit{if } x \ge 0, \end{cases}$$

where η is given in (4a), *when the imbedded queue is in equilibrium.*

Note that W has an atom of size $1 - \eta$ at the origin.

Proof. By the lack-of-memory property, W_n is the sum of $Q(A_n)$ independent exponential variables. Use the equilibrium distribution of $Q(A)$ to find that

$$M_W(s) = (1 - \eta) + \eta \frac{\mu(1 - \eta)}{\mu(1 - \eta) - s}$$

which we recognize as the moment generating function of a random variable which either equals zero (with probability $1 - \eta$) or is exponentially distributed with parameter $\mu(1 - \eta)$ (with probability η). ■

Finally, here is a word of caution. There is another quantity called *virtual* waiting time, which must not be confused with *actual* waiting time. The latter is the actual time spent by a customer after his arrival; the former is the time which a customer *would* spend if he were to arrive at some particular instant. The equilibrium distributions of these waiting times may differ whenever the stationary distribution of Q differs from the stationary distribution of the imbedded Markov chain $Q(A)$.

Exercises for Section 11.4

1. Consider G/M(μ)/1, and let $\alpha_j = \mathbb{E}((\mu X)^j e^{-\mu X}/j!)$ where X is a typical interarrival time. Suppose the traffic intensity ρ is less than 1. Show that the equilibrium distribution π of the imbedded chain at moments of arrivals satisfies

$$\pi_n = \sum_{i=0}^{\infty} \alpha_i \pi_{n+i-1} \qquad \text{for } n \ge 1.$$

Look for a solution of the form $\pi_n = \theta^n$ for some θ, and deduce that the unique stationary distribution is given by $\pi_j = (1-\eta)\eta^j$ for $j \geq 0$, where η is the smallest positive root of the equation $s = M_X(\mu(s-1))$.

2. Consider a G/M(μ)/1 queue in equilibrium. Let η be the smallest positive root of the equation $x = M_X(\mu(x-1))$ where M_X is the moment generating function of an interarrival time. Show that the mean number of customers ahead of a new arrival is $\eta(1-\eta)^{-1}$, and the mean waiting time is $\eta\{\mu(1-\eta)\}^{-1}$.

3. Consider D(1)/M(μ)/1 where $\mu > 1$. Show that the continuous-time queue length $Q(t)$ does not converge in distribution as $t \to \infty$, even though the imbedded chain at the times of arrivals is ergodic.

11.5 G/G/1

If neither interarrival times nor service times are exponentially distributed then the methods of the last three sections fail. This apparent setback leads us to the remarkable discovery that queueing problems are intimately related to random walk problems. This section is divided into two parts, one dealing with the equilibrium theory of G/G/1 and the other dealing with the imbedded random walk.

(A) **Asymptotic waiting time.** Let W_n be the waiting time of the nth customer. There is a useful relationship between W_n and W_{n+1} in terms of the service time S_n of the nth customer and the length X_{n+1} of time between the nth and the $(n+1)$th arrivals.

(1) **Theorem. Lindley's equation.** *We have that*

$$W_{n+1} = \max\{0, W_n + S_n - X_{n+1}\}.$$

Proof. The nth customer is in the system for a length $W_n + S_n$ of time. If $X_{n+1} > W_n + S_n$ then the queue is empty at the $(n+1)$th arrival, and so $W_{n+1} = 0$. If $X_{n+1} \leq W_n + S_n$ then the $(n+1)$th customer arrives while the nth is still present, but only waits for a period of length $W_n + S_n - X_{n+1}$ before the previous customer leaves. ■

We shall see that Lindley's equation implies that the distribution functions

$$F_n(x) = \mathbb{P}(W_n \leq x)$$

of the W_n converge as $n \to \infty$ to some limit function $F(x)$. Of course, F need not be a proper distribution function; indeed, it is intuitively clear that the queue settles down into equilibrium if and only if F is a distribution function which is not defective.

(2) **Theorem.** *Let $F_n(x) = \mathbb{P}(W_n \leq x)$. Then*

$$F_{n+1}(x) = \begin{cases} 0 & \text{if } x < 0, \\ \displaystyle\int_{-\infty}^{x} F_n(x-y)\, dG(y) & \text{if } x \geq 0, \end{cases}$$

where G is the distribution function of $U_n = S_n - X_{n+1}$. Thus the limit $F(x) = \lim_{n\to\infty} F_n(x)$ exists.

Note that $\{U_n : n \geq 1\}$ is a collection of independent identically distributed random variables.

Proof. If $x \geq 0$ then

$$\begin{aligned}\mathbb{P}(W_{n+1} \leq x) &= \int_{-\infty}^{\infty} \mathbb{P}(W_n + U_n \leq x \mid U_n = y)\, dG(y)\\ &= \int_{-\infty}^{x} \mathbb{P}(W_n \leq x - y)\, dG(y) \quad \text{by independence,}\end{aligned}$$

and the first part is proved. We claim that

$$F_{n+1}(x) \leq F_n(x) \quad \text{for all } x \text{ and } n. \tag{3}$$

If (3) holds then the second result follows immediately; we prove (3) by induction. Trivially, $F_2(x) \leq F_1(x)$ because $F_1(x) = 1$ for all $x \geq 0$. Suppose that (3) holds for $n = k - 1$, say. Then, for $x \geq 0$,

$$F_{k+1}(x) - F_k(x) = \int_{-\infty}^{x} \big[F_k(x-y) - F_{k-1}(x-y)\big]\, dG(y) \leq 0$$

by the induction hypothesis. The proof is complete. ■

It follows that the distribution functions of $\{W_n\}$ converge as $n \to \infty$. It is clear, by monotone convergence, that the limit $F(x)$ satisfies the Wiener–Hopf equation

$$F(x) = \int_{-\infty}^{x} F(x-y)\, dG(y) \quad \text{for} \quad x \geq 0;$$

this is not easily solved for F in terms of G. However, it is not too difficult to find a criterion for F to be a proper distribution function.

(4) Theorem. *Let $\rho = \mathbb{E}(S)/\mathbb{E}(X)$ be the traffic intensity*

(a) *If $\rho < 1$, then F is a non-defective distribution function.*
(b) *If $\rho > 1$, then $F(x) = 0$ for all x.*
(c) *If $\rho = 1$ and* $\mathrm{var}(U) > 0$, *then $F(x) = 0$ for all x.*

An explicit formula for the moment generating function of F when $\rho < 1$ is given in Theorem (14) below. Theorem (4) classifies the stability of G/G/1 in terms of the sign of $1 - \rho$; note that this information is obtainable from the distribution function G since

$$\rho < 1 \quad \Leftrightarrow \quad \mathbb{E}(S) < \mathbb{E}(X) \quad \Leftrightarrow \quad \mathbb{E}(U) = \int_{-\infty}^{\infty} u\, dG(u) < 0 \tag{5}$$

where U is a typical member of the U_i. We call the process *stable* when $\rho < 1$.

The crucial step in the proof of (4) is important in its own right. Use Lindley's equation (1) to see that:

$$\begin{aligned} W_1 &= 0,\\ W_2 &= \max\{0, W_1 + U_1\} = \max\{0, U_1\},\\ W_3 &= \max\{0, W_2 + U_2\} = \max\{0, U_2, U_2 + U_1\},\end{aligned}$$

and in general

$$(6) \qquad W_{n+1} = \max\{0, U_n, U_n + U_{n-1}, \dots, U_n + U_{n-1} + \cdots + U_1\}$$

which expresses W_{n+1} in terms of the partial sums of a sequence of independent identically distributed variables. It is difficult to derive asymptotic properties of W_{n+1} directly from (6) since every non-zero term changes its value as n increases from the value k, say, to the value $k+1$. The following theorem is the crucial observation.

(7) Theorem. *The random variable W_{n+1} has the same distribution as*

$$W'_{n+1} = \max\{0, U_1, U_1 + U_2, \dots, U_1 + U_2 + \cdots + U_n\}.$$

Proof. The vectors $(U_1, U_2, \dots, U_n)$ and $(U_n, U_{n-1}, \dots, U_1)$ are sequences with the same joint distribution. Replace each U_i in (6) by U_{n+1-i}. ■

That is to say, W_{n+1} and W'_{n+1} are *different* random variables but they have the *same* distribution. Thus

$$F(x) = \lim_{n\to\infty} \mathbb{P}(W_n \le x) = \lim_{n\to\infty} \mathbb{P}(W'_n \le x).$$

Furthermore,

$$(8) \qquad W'_n \le W'_{n+1} \quad \text{for all } n \ge 1,$$

a monotonicity property which is not shared by $\{W_n\}$. This property provides another method for deriving the existence of F in (2).

Proof of (4). From (8), the limit $W' = \lim_{n\to\infty} W'_n$ exists almost surely (and, in fact, pointwise) but may be $+\infty$. Furthermore,

$$(9) \qquad W' = \max\{0, \Sigma_1, \Sigma_2, \dots\}$$

where

$$\Sigma_n = \sum_{j=1}^{n} U_j$$

and $F(x) = \mathbb{P}(W' \le x)$. Thus

$$F(x) = \mathbb{P}(\Sigma_n \le x \text{ for all } n) \quad \text{if} \quad x \ge 0,$$

and the proof proceeds by using properties of the sequence $\{\Sigma_n\}$ of partial sums, such as the strong law (7.5.1):

$$(10) \qquad \frac{1}{n}\Sigma_n \xrightarrow{\text{a.s.}} \mathbb{E}(U) \quad \text{as} \quad n \to \infty.$$

Suppose first that $\mathbb{E}(U) < 0$. Then

$$\mathbb{P}(\Sigma_n > 0 \text{ for infinitely many } n) = \mathbb{P}\left(\frac{1}{n}\Sigma_n - \mathbb{E}(U) > |\mathbb{E}(U)| \text{ i.o.}\right) = 0$$

by (10). Thus, from (9), W' is almost surely the maximum of only finitely many terms, and so $\mathbb{P}(W' < \infty) = 1$, implying that F is a non-defective distribution function.

Next suppose that $\mathbb{E}(U) > 0$. Pick any $x > 0$ and choose N such that

$$N \geq \frac{2x}{\mathbb{E}(U)}.$$

For $n \geq N$,

$$\begin{aligned}\mathbb{P}(\Sigma_n \geq x) &= \mathbb{P}\left(\frac{1}{n}\Sigma_n - \mathbb{E}(U) \geq \frac{x}{n} - \mathbb{E}(U)\right) \\ &\geq \mathbb{P}\left(\frac{1}{n}\Sigma_n - \mathbb{E}(U) \geq -\tfrac{1}{2}\mathbb{E}(U)\right).\end{aligned}$$

Let $n \to \infty$ and use the weak law to find that

$$\mathbb{P}(W' \geq x) \geq \mathbb{P}(\Sigma_n \geq x) \to 1 \quad \text{for all } x.$$

Therefore W' almost surely exceeds any finite number, and so $\mathbb{P}(W' < \infty) = 0$ as required.

In the case when $\mathbb{E}(U) = 0$ these crude arguments do not work and we need a more precise measure of the fluctuations of Σ_n; one way of doing this is by way of the law of the iterated logarithm (7.6.1). If $\text{var}(U) > 0$ and $\mathbb{E}(U_1^2) < \infty$, then $\{\Sigma_n\}$ enjoys fluctuations of order $O(\sqrt{n \log\log n})$ in both positive and negative directions with probability 1, and so

$$\mathbb{P}(\Sigma_n \geq x \text{ for some } n) = 1 \quad \text{for all } x.$$

There are other arguments which yield the same result. ■

(B) Imbedded random walk. The sequence $\Sigma = \{\Sigma_n : n \geq 0\}$ given by

$$\Sigma_0 = 0, \quad \Sigma_n = \sum_{j=1}^{n} U_j \quad \text{for} \quad n \geq 1, \tag{11}$$

describes the path of a particle which performs a random walk on $\mathbb{R}$, jumping by an amount U_n at the nth step. This simple observation leads to a wealth of conclusions about queueing systems. For example, we have just seen that the waiting time W_n of the nth customer has the same distribution as the maximum W_n' of the first n positions of the walking particle. If $\mathbb{E}(U) < 0$ then the waiting time distributions converge as $n \to \infty$, which is to say that the maximum displacement $W' = \lim W_n'$ is almost surely finite. Other properties also can be expressed in terms of this random walk, and the techniques of reflection and reversal which we discussed in Section 3.10 are useful here.

The limiting waiting time distribution is the same as the distribution of the maximum

$$W' = \max\{0, \Sigma_1, \Sigma_2, \dots\},$$

and so it is appropriate to study the so-called 'ladder points' of Σ. Define an increasing sequence $L(0), L(1), \dots$ of random variables by

$$L(0) = 0, \quad L(n+1) = \min\big\{m > L(n) : \Sigma_m > \Sigma_{L(n)}\big\};$$

that is, $L(n+1)$ is the earliest epoch m of time at which Σ_m exceeds the walk's previous maximum $\Sigma_{L(n)}$. The $L(n)$ are called *ladder points*; *negative ladder points* of Σ are defined similarly as the epochs at which Σ attains new minimum values. The result of (4) amounts to the assertion that

$$\mathbb{P}(\text{there exist infinitely many ladder points}) = \begin{cases} 0 & \text{if } \mathbb{E}(U) < 0, \\ 1 & \text{if } \mathbb{E}(U) > 0. \end{cases}$$

The total number of ladder points is given by the next lemma.

(12) Lemma. *Let $\eta = \mathbb{P}(\Sigma_n > 0$ for some $n \geq 1)$ be the probability that at least one ladder point exists. The total number Λ of ladder points has mass function*

$$\mathbb{P}(\Lambda = l) = (1-\eta)\eta^l \quad \textit{for} \quad l \geq 0.$$

Proof. The process Σ is a discrete-time Markov chain. Thus

$$\mathbb{P}(\Lambda \geq l+1 \mid \Lambda \geq l) = \eta$$

since the path of the walk after the lth ladder point is a copy of Σ itself. ■

Thus the queue is stable if $\eta < 1$, in which case the maximum W' of Σ is related to the height of a typical ladder point. Let

$$Y_j = \Sigma_{L(j)} - \Sigma_{L(j-1)}$$

be the difference in the displacements of the walk at the $(j-1)$th and jth ladder points. Conditional on the value of Λ, $\{Y_j : 1 \leq j \leq \Lambda\}$ is a collection of independent identically distributed variables, by the Markov property. Furthermore,

$$\text{(13)} \qquad W' = \Sigma_{L(\Lambda)} = \sum_{j=1}^{\Lambda} Y_j;$$

this leads to the next lemma, relating the waiting time distribution to the distribution of a typical Y_j.

(14) Lemma. *If the traffic intensity ρ satisfies $\rho < 1$, the equilibrium waiting time distribution has moment generating function*

$$M_W(s) = \frac{1-\eta}{1-\eta M_Y(s)}$$

where M_Y is the moment generating function of Y.

Proof. We have that $\rho < 1$ if and only if $\eta < 1$. Use (13) and (5.1.25) to find that

$$M_W(s) = G_\Lambda(M_Y(s)).$$

Now use the result of Lemma (12). ■

Lemma (14) describes the waiting time distribution in terms of the distribution of Y. Analytical properties of Y are a little tricky to obtain, and we restrict ourselves here to an elegant description of Y which provides a curious link between pairs of 'dual' queueing systems.

The server of the queue enjoys busy periods during which she works continuously; in between busy periods she has *idle periods* during which she drinks tea. Let I be the length of her first idle period.

(15) Lemma. *Let* $L = \min\{m > 0;\ \Sigma_m < 0\}$ *be the first negative ladder point of* Σ. *Then* $I = -\Sigma_L$.

That is, I equals the absolute value of the depth of the first negative ladder point. It is of course possible that Σ has *no* negative ladder points.

Proof. Call a customer *lucky* if he finds the queue empty as he arrives (customers who arrive at exactly the same time as the previous customer departs are deemed to be unlucky). We claim that the $(L+1)$th customer is the first lucky customer after the very first arrival. If this holds then (15) follows immediately since I is the elapsed time between the Lth departure and the $(L+1)$th arrival:

$$I = \sum_{j=1}^{L} X_{j+1} - \sum_{j=1}^{L} S_j = -\Sigma_L.$$

To verify the claim remember that

$$\textbf{(16)} \qquad W_n = \max\{0, V_n\} \quad \text{where} \quad V_n = \max\{U_{n-1}, U_{n-1}+U_{n-2}, \dots, \Sigma_{n-1}\}$$

and note that the nth customer is lucky if and only if $V_n < 0$. Now

$$V_n \geq \Sigma_{n-1} \geq 0 \quad \text{for} \quad 2 \leq n \leq L,$$

and it remains to show that $V_{L+1} < 0$. To see this, note that

$$U_L + U_{L-1} + \cdots + U_{L-k} = \Sigma_L - \Sigma_{L-k-1} \leq \Sigma_L < 0$$

whenever $0 \leq k < L$. Now use (16) to obtain the result. ■

Now we are ready to extract a remarkable identity which relates 'dual pairs' of queueing systems.

(17) Definition. If Q is a queueing process with interarrival time distribution F_X and service time distribution F_S, its **dual process** Q_{d} is a queueing process with interarrival time distribution F_S and service time distribution F_X.

For example, the dual of M(λ)/G/1 is G/M(λ)/1, and vice versa; we made use of this fact in the proof of (11.4.4). The traffic densities ρ and ρ_{d} of Q and Q_{d} satisfy $\rho\rho_{\mathrm{d}} = 1$; the processes Q and Q_{d} cannot both be stable except in pathological instances when all their interarrival and service times almost surely take the same constant value.

(18) Theorem. *Let* Σ *and* Σ_{d} *be the random walks associated with the queue* Q *and its dual* Q_{d}. *Then* $-\Sigma$ *and* Σ_{d} *are identically distributed random walks.*

Proof. Let Q have interarrival times $\{X_n\}$ and service times $\{S_n\}$; Σ has jumps of size $U_n = S_n - X_{n+1}$ $(n \geq 1)$. The reflected walk $-\Sigma$, which is obtained by reflecting Σ in the x-axis, has jumps of size $-U_n = X_{n+1} - S_n$ $(n \geq 1)$ (see Section 3.10 for more details of the reflection principle). Write $\{S'_n\}$ and $\{X'_n\}$ for the interarrival and service times of Q_d; Σ_d has jumps of size $U'_n = X'_n - S'_{n+1}$ $(n \geq 1)$, which have the same distribution as the jumps of $-\Sigma$. ■

This leads to a corollary.

(19) Theorem. *The height Y of the first ladder point of Σ has the same distribution as the length I_d of a typical idle period in the dual queue.*

Proof. From (15), $-I_\mathrm{d}$ is the height of the first ladder point of $-\Sigma_\mathrm{d}$, which by (18) is distributed as the height Y of the first ladder point of Σ. ■

Here is an example of an application of these facts.

(20) Theorem. *Let Q be a stable queueing process with dual process Q_d. Let W be a typical equilibrium waiting time of Q and I_d a typical idle period of Q_d. Their moment generating functions are related by*

$$M_W(s) = \frac{1-\eta}{1-\eta M_{I_\mathrm{d}}(s)}$$

where $\eta = \mathbb{P}(W > 0)$.

Proof. Use (14) and (19). ■

An application of this result is given in Exercise (2). Another application is a second derivation of the equilibrium waiting time distribution (11.4.12) of G/M/1; just remark that the dual of G/M/1 is M/G/1, and that idle periods of M/G/1 are exponentially distributed (though, of course, the server does not have many such periods if the queue is unstable).

Exercises for Section 11.5

1. Show that, for a G/G/1 queue, the starting times of the busy periods of the server constitute a renewal process.

2. Consider a G/M(μ)/1 queue in equilibrium, together with the dual (unstable) M(μ)/G/1 queue. Show that the idle periods of the latter queue are exponentially distributed. Use the theory of duality of queues to deduce for the former queue that: (a) the waiting-time distribution is a mixture of an exponential distribution and an atom at zero, and (b) the equilibrium queue length is geometric.

3. Consider G/M(μ)/1, and let G be the distribution function of $S - X$ where S and X are typical (independent) service and interarrival times. Show that the *Wiener–Hopf equation*

$$F(x) = \int_{-\infty}^{x} F(x-y)\,dG(y), \qquad x \geq 0,$$

for the limiting waiting-time distribution F is satisfied by $F(x) = 1 - \eta e^{-\mu(1-\eta)x}$, $x \geq 0$. Here, η is the smallest positive root of the equation $x = M_X(\mu(x-1))$, where M_X is the moment generating function of X.

11.6 Heavy traffic

A queue settles into equilibrium if its traffic intensity ρ is less than 1; it is unstable if $\rho > 1$. It is our shared personal experience that many queues (such as in doctors' waiting rooms and at airport check-in desks) have a tendency to become unstable. The reason is simple: employers do not like to see their employees idle, and so they provide only just as many servers as are necessary to cope with the arriving customers. That is, they design the queueing system so that ρ is only slightly smaller than 1; the ensuing queue is long but stable, and the server experiences 'heavy traffic'. As $\rho \uparrow 1$ the equilibrium queue length Q_ρ becomes longer and longer, and it is interesting to ask for the rate at which Q_ρ approaches infinity. Often it turns out that a suitably scaled form of Q_ρ is asymptotically exponentially distributed. We describe this here for the M/D/1 system, leaving it to the readers to amuse themselves by finding corresponding results for other queues. In this special case, $Q_\rho \simeq Z/(1-\rho)$ as $\rho \uparrow 1$ where Z is an exponential variable.

(1) Theorem. *Let $\rho = \lambda d$ be the traffic intensity of the $M(\lambda)/D(d)/1$ queue, and let Q_ρ be a random variable with the equilibrium queue length distribution. Then $(1-\rho)Q_\rho$ converges in distribution as $\rho \uparrow 1$ to the exponential distribution with parameter 2.*

Proof. Use (11.3.5) to see that Q_ρ has moment generating function

$$M_\rho(s) = \frac{(1-\rho)(e^s - 1)}{\exp[s - \rho(e^s - 1)] - 1} \quad \text{if} \quad \rho < 1. \tag{2}$$

The moment generating function of $(1-\rho)Q_\rho$ is $M_\rho((1-\rho)s)$, and we make the appropriate substitution in equation (2). Now let $\rho \uparrow 1$ and use L'Hôpital's rule to deduce that $M_\rho((1-\rho)s) \to 2/(2-s)$. ∎

Exercise for Section 11.6

1. Consider the M(λ)/M(μ)/1 queue with $\rho = \lambda/\mu < 1$. Let Q_ρ be a random variable with the equilibrium queue distribution, and show that $(1-\rho)Q_\rho$ converges in distribution as $\rho \uparrow 1$, the limit distribution being exponential with parameter 1.

11.7 Networks of queues

A customer departing from one queue may well be required to enter another. Under certain circumstances, this customer may even, at a later stage, be required to re-enter the original queue. Networks of queues provide natural models for many situations in real life, ranging from the manufacture of components in complex industrial processes to an application for a visa for travel to another country.

We make concrete our notion of a queueing network as follows. There is a finite set S of 'stations' labelled $s_1, s_2, \dots, s_c$. At time t, station i contains $Q_i(t)$ individuals, so that the state of the system may be represented as the vector $\mathbf{Q}(t) = (Q_1(t), Q_2(t), \dots, Q_c(t))$. We assume for simplicity that the process $\mathbf{Q}$ is Markovian, taking values in the set $\mathcal{N} = \{\mathbf{n} = (n_1, n_2, \dots, n_c) : n_i = 0, 1, 2, \dots \text{ for } 1 \le i \le c\}$ of sequences of non-negative integers. The migration of customers between stations will be described in a rather general way in order to enable a breadth of applications.

We begin with a little notation. We write $\mathbf{e}_k = (0, 0, \dots, 0, 1, 0, \dots, 0)$ for the row vector of length c with 1 in the kth position and 0 elsewhere.

Assume that $\mathbf{Q}(t) = \mathbf{n}$. Three types of event may occur in the short time interval $(t, t+h)$, at rates given by the following. For $i \neq j$,

$$\mathbf{Q}(t+h) = \begin{cases} \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j & \text{with probability } \lambda_{ij}\phi_i(n_i)h + \mathrm{o}(h), \\ \mathbf{n} + \mathbf{e}_j & \text{with probability } \nu_j h + \mathrm{o}(h), \\ \mathbf{n} - \mathbf{e}_i & \text{with probability } \mu_i\phi_i(n_i)h + \mathrm{o}(h), \end{cases}$$

where λ_{ij}, ν_j, μ_i are constants and the ϕ_i are functions such that $\phi_i(0) = 0$ and $\phi_i(n) > 0$ for $n \geq 1$. We assume for later use that $\lambda_{ii} = 0$ for all i. Thus a single customer moves from station i to station j at rate $\lambda_{ij}\phi_i(n_i)$, a new arrival occurs at station j at rate ν_j, and a single customer departs the system from station i at rate $\mu_i\phi_i(n_i)$. Note that departures from a station may occur at a rate which depends on the number of customers at that station.

Queueing networks defined in this rather general manner are sometimes termed 'migration processes' or 'Jackson networks'. Here are some concrete instances. First, the network is termed a 'closed migration process' if $\nu_j = \mu_j = 0$ for all j, since in this case no arrival from or departure to the outside world is permitted. If some ν_j or μ_j is strictly positive, the network is termed an 'open migration process'. Closed migration processes are special inasmuch as they are restricted to a subset of $\mathcal{N}$ containing vectors $\mathbf{n}$ having constant sum.

(3) Example. Suppose that each station i has r servers, and that each customer at that station requires a service time having the exponential distribution with parameter γ_i. On departing station i, a customer proceeds to station j ($\neq i$) with probability p_{ij}, or departs the system entirely with probability $q_i = 1 - \sum_{j:j\neq i} p_{ij}$. Assuming the usual independence, the corresponding migration process has parameters given by $\phi_i(n) = \min\{n, r\}$, $\lambda_{ij} = \gamma_i p_{ij}$, $\mu_i = \gamma_i q_i$, $\nu_j = 0$. ●

(4) Example. Suppose that customers are invisible to one another, in the sense that each proceeds around the network at rates which do not depend on the positions of other customers. In this case, we have $\phi_i(n) = n$. ●

(A) Closed migration processes. We shall explore the equilibrium behaviour of closed processes, and we assume therefore that $\nu_j = \mu_j = 0$ for all j. The number of customers is constant, and we denote this number by N.

Consider first the case $N = 1$, and suppose for convenience that $\phi_j(1) = 1$ for all j. When at station i, the single customer moves to station j at rate λ_{ij}. The customer's position is a continuous-time Markov chain with generator $\mathbf{H} = (h_{ij})$ given by

(5) $$h_{ij} = \begin{cases} \lambda_{ij} & \text{if } i \neq j, \\ -\sum_k \lambda_{ik} & \text{if } i = j. \end{cases}$$

It has an equilibrium distribution $\boldsymbol{\alpha} = (\alpha_i : i \in S)$ satisfying $\boldsymbol{\alpha}\mathbf{H} = \mathbf{0}$, which is to say that

(6) $$\sum_j \alpha_j \lambda_{ji} = \alpha_i \sum_j \lambda_{ij} \quad \text{for } i \in S,$$

and we assume henceforth that $\boldsymbol{\alpha}$ satisfies these equations. We have as usual that $\alpha_i > 0$ for all i whenever this chain is irreducible, and we suppose henceforth that this is the case. It is the case that $\sum_i \alpha_i = 1$, but this will be irrelevant in the following.

The equilibrium distribution in the case of a general closed migration process is given in the following theorem. We write $\mathcal{N}_N$ for the set of all vectors in $\mathcal{N}$ having sum N. Any empty product is to be interpreted as 1.

(7) Theorem. *The equilibrium distribution of an irreducible closed migration process with N customers is*

$$\text{(8)} \qquad \pi(\mathbf{n}) = B_N \prod_{i=1}^{c} \left\{ \frac{\alpha_i^{n_i}}{\prod_{r=1}^{n_i} \phi_i(r)} \right\}, \quad \mathbf{n} \in \mathcal{N}_N,$$

where B_N is the appropriate normalizing constant.

Note the product form of the equilibrium distribution in (8). This does not imply the independence of queue lengths in equilibrium since they are constrained to have constant sum N and must therefore in general be dependent.

Proof. The process has at most one equilibrium distribution, and any such distribution $\boldsymbol{\gamma} = (\gamma(\mathbf{n}) : \mathbf{n} \in \mathcal{N}_N)$ satisfies the equations

$$\text{(9)} \qquad \sum_{i,j} \gamma(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j)\lambda_{ji}\phi_j(n_j + 1) = \gamma(\mathbf{n}) \sum_{i,j} \lambda_{ij}\phi_i(n_i), \quad \mathbf{n} \in \mathcal{N}_N.$$

This is a complicated system of equations. If we may solve the equation 'for each i', then we obtain a solution to (9) by summing over i. Removing from (9) the summation over i, we obtain the 'partial balance equations'

$$\text{(10)} \qquad \sum_{j} \gamma(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j)\lambda_{ji}\phi_j(n_j + 1) = \gamma(\mathbf{n}) \sum_{j} \lambda_{ij}\phi_i(n_i), \quad \mathbf{n} \in \mathcal{N}_N^i,\ i \in S,$$

where $\mathcal{N}_n^i$ is the subset of $\mathcal{N}$ containing all vectors $\mathbf{n}$ with $n_i \geq 1$. It suffices to check that (8) satisfies (10). With $\boldsymbol{\pi}$ given by (8), we have that

$$\pi(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) = \pi(\mathbf{n}) \frac{\alpha_j \phi_i(n_i)}{\alpha_i \phi_j(n_j + 1)}$$

whence $\boldsymbol{\pi}$ satisfies (10) if and only if

$$\sum_j \frac{\alpha_j \phi_i(n_i)}{\alpha_i \phi_j(n_j + 1)} \cdot \lambda_{ji}\phi_j(n_j + 1) = \sum_j \lambda_{ij}\phi_i(n_i), \quad \mathbf{n} \in \mathcal{N}_N^i,\ i \in S.$$

The latter equation is simply (6), and the proof is complete. ■

(11) Example. A company has exactly K incoming telephone lines and an ample number of operators. Calls arrive in the manner of a Poisson process with rate ν. Each call occupies an operator for a time which is exponentially distributed with parameter λ, and then lasts a further period of time having the exponential distribution with parameter μ. At the end of this time, the call ceases and the line becomes available for another incoming call. Arriving calls are lost if all K channels are already in use.

Although the system of calls is a type of *open* queueing system, one may instead consider the *lines* as customers in a network having three stations. That is, at any time t the state vector of the system may be taken to be $\mathbf{n} = (n_1, n_2, n_3)$ where n_1 is the number of free lines, n_2 is the number of calls being actively serviced by an operator, and n_3 is the number of calls still in operation but no longer utilizing an operator. This leads to a closed migration process with transition rates given by

$$\begin{aligned} \lambda_{12} &= \nu, & \phi_1(n) &= I_{\{n\geq 1\}}, \\ \lambda_{23} &= \lambda, & \phi_2(n) &= n, \\ \lambda_{31} &= \mu, & \phi_3(n) &= n, \end{aligned}$$

where $I_{\{n\geq 1\}}$ is the indicator function that $n \geq 1$. It is an easy exercise to show from (6) that the relative sizes of $\alpha_1, \alpha_2, \alpha_3$ satisfy $\alpha_1 : \alpha_2 : \alpha_3 = \nu^{-1} : \lambda^{-1} : \mu^{-1}$, whence the equilibrium distribution is given by

$$\text{(12)} \qquad \pi(n_1, n_2, n_3) = B \cdot \frac{1}{\nu^{n_1}} \cdot \frac{1}{\lambda^{n_2} n_2!} \cdot \frac{1}{\mu^{n_3} n_3!}, \qquad n_1 + n_2 + n_3 = K,$$

for an appropriate constant B. ●

(B) Open migration processes. We turn now to the general situation in which customers may enter or leave the system. As in the case of a closed migration process, it is valuable to consider first an auxiliary process containing exactly one customer. We attach another station, labelled ∞, to those already in existence, and we consider the following closed migration process on the augmented set $S \cup \{\infty\}$ of stations. There is a unique customer who, when at station i, moves to station j $(\neq i)$ at rate:

$$\begin{cases} \lambda_{ij} & \text{if } 1 \leq i, j \leq c, \\ \mu_i & \text{if } j = \infty, \\ \nu_j & \text{if } i = \infty. \end{cases}$$

We assume henceforth that this auxiliary process is irreducible. Let $\mathbf{J}$ be its generator. The chain has a unique stationary distribution $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_c, \beta_\infty)$ which satisfies $\boldsymbol{\beta}\mathbf{J} = \mathbf{0}$. In particular,

$$\beta_\infty \nu_i + \sum_{j\in S} \beta_j \lambda_{ji} = \beta_i \Big(\mu_i + \sum_{j \in S} \lambda_{ij} \Big) \quad \text{for } i \in S.$$

Note that $\beta_i > 0$ for $i \in S \cup \{\infty\}$. We set $\alpha_i = \beta_i / \beta_\infty$, to obtain a vector $\boldsymbol{\alpha} = (\alpha_i : i \in S)$ with strictly positive entries such that

$$\text{(13)} \qquad \nu_i + \sum_j \alpha_j \lambda_{ji} = \alpha_i \Big(\mu_i + \sum_j \lambda_{ij} \Big) \quad \text{for } i \in S.$$

We shall make use of this vector $\boldsymbol{\alpha}$ in very much the same way as we used equation (6) for closed migration processes. We let

$$D_i = \sum_{n=0}^{\infty} \frac{\alpha_i^n}{\prod_{r=1}^n \phi_i(r)}.$$

(14) Theorem. *Assume that the above auxiliary process is irreducible, and that $D_i < \infty$ for all $i \in S$. The open migration process has equilibrium distribution*

$$\text{(15)} \qquad \pi(\mathbf{n}) = \prod_{i=1}^{c} \pi_i(n_i), \qquad \mathbf{n} \in \mathcal{N},$$

where

$$\pi_i(n_i) = D_i^{-1} \frac{\alpha_i^{n_i}}{\prod_{r=1}^{n_i} \phi_i(r)}.$$

Proof. The distribution $\boldsymbol{\gamma} = (\gamma(\mathbf{n}) : \mathbf{n} \in \mathcal{N})$ is an equilibrium distribution if $\boldsymbol{\gamma}\mathbf{G} = \mathbf{0}$ where $\mathbf{G} = (g(\mathbf{n}, \mathbf{n}') : \mathbf{n}, \mathbf{n}' \in \mathcal{N})$ is the generator of the Markov chain $\mathbf{Q}$, which is to say that

(16)

$$\sum_i \gamma(\mathbf{n}-\mathbf{e}_i)g(\mathbf{n}-\mathbf{e}_i, \mathbf{n}) + \sum_{i,j} \gamma(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)g(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j, \mathbf{n}) + \sum_j \gamma(\mathbf{n}+\mathbf{e}_j)g(\mathbf{n}+\mathbf{e}_j, \mathbf{n})$$
$$= \gamma(\mathbf{n}) \left(\sum_i g(\mathbf{n}, \mathbf{n}-\mathbf{e}_i) + \sum_{i,j} g(\mathbf{n}, \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) + \sum_j g(\mathbf{n}, \mathbf{n}+\mathbf{e}_j) \right).$$

These are solved by $\boldsymbol{\gamma}$ if it satisfies the 'partial balance equations'

$$\text{(17)} \quad \gamma(\mathbf{n}-\mathbf{e}_i)g(\mathbf{n}-\mathbf{e}_i, \mathbf{n}) + \sum_j \gamma(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)g(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j, \mathbf{n})$$
$$= \gamma(\mathbf{n}) \left(g(\mathbf{n}, \mathbf{n}-\mathbf{e}_i) + \sum_j g(\mathbf{n}, \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) \right), \quad i \in S,$$

together with the equation

$$\text{(18)} \qquad \sum_j \gamma(\mathbf{n}+\mathbf{e}_j)g(\mathbf{n}+\mathbf{e}_j, \mathbf{n}) = \gamma(\mathbf{n}) \sum_j g(\mathbf{n}, \mathbf{n}+\mathbf{e}_j).$$

We now substitute (15) into (17) and divide through by $\pi(\mathbf{n})$ to find that $\boldsymbol{\pi}$ satisfies (17) by reason of (13). Substituting $\boldsymbol{\pi}$ into (18), we obtain the equation

$$\sum_j \alpha_j \mu_j = \sum_j \nu_j,$$

whose validity follows by summing (13) over i. ■

Equation (15) has the important and striking consequence that, in equilibrium, the queue lengths at the different stations are independent random variables. Note that this equilibrium statement does not apply to the queueing process itself: the queue processes $Q_i(\cdot)$, $i \in S$, are highly dependent on one another.

It is a remarkable fact that the reversal in time of an open migration process yields another open migration process.

(19) Theorem. *Let* $\mathbf{Q} = (\mathbf{Q}(t) : -\infty < t < \infty)$ *be an open migration process and assume that, for all* t, $\mathbf{Q}(t)$ *has distribution* $\boldsymbol{\pi}$ *given by Theorem* (14). *Then* $\mathbf{Q}'(t) = \mathbf{Q}(-t)$ *is an open migration network with parameters*

$$\lambda'_{ij} = \frac{\alpha_j \lambda_{ji}}{\alpha_i}, \quad \nu'_j = \alpha_j \mu_j, \quad \mu'_i = \frac{\nu_i}{\alpha_i}, \quad \phi'_i(\cdot) = \phi_i(\cdot),$$

where $\boldsymbol{\alpha}$ *satisfies* (13).

Proof. It is a straighforward exercise in conditional probabilities (see Problem (6.15.16)) to show that $\mathbf{Q}'$ is a Markov chain with generator $\mathbf{G}' = (g'(\mathbf{m}, \mathbf{n}) : \mathbf{m}, \mathbf{n} \in \mathcal{N})$ given by

$$\begin{aligned} g'(\mathbf{n}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) &= \frac{\pi(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) g(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j, \mathbf{n})}{\pi(\mathbf{n})} = \lambda'_{ij}\phi_i(n_i), \\ g'(\mathbf{n}, \mathbf{n} - \mathbf{e}_i) &= \frac{\pi(\mathbf{n} - \mathbf{e}_i) g(\mathbf{n} - \mathbf{e}_i, \mathbf{n})}{\pi(\mathbf{n})} = \mu'_i \phi_i(n_i), \\ g'(\mathbf{n}, \mathbf{n} + \mathbf{e}_j) &= \frac{\pi(\mathbf{n} + \mathbf{e}_j) g(\mathbf{n} + \mathbf{e}_j, \mathbf{n})}{\pi(\mathbf{n})} = \nu'_j, \end{aligned}$$

as required. ■

Here is one noteworthy consequence of this useful theorem. Let $\mathbf{Q}$ be an open migration network in equilibrium, and consider the processes of departures from the network from the various stations. That is, let $D_i(t)$ be the number of customers who depart the system from station i during the time interval $[0, t]$. These departures correspond to arrivals at station i in the reversed process $\mathbf{Q}'$. However, such arrival processes are independent Poisson processes, with respective parameters $\nu'_i = \alpha_i \mu_i$. It follows that the departure processes are independent Poisson processes with these parameters.

Exercises for Section 11.7

1. Consider an open migration process with c stations, in which individuals arrive at station j at rate ν_j, individuals move from i to j at rate $\lambda_{ij}\phi_i(n_i)$, and individuals depart from i at rate $\mu_i \phi_i(n_i)$, where n_i denotes the number of individuals currently at station i. Show when $\phi_i(n_i) = n_i$ for all i that the system behaves as though the customers move independently through the network. Identify the explicit form of the stationary distribution, subject to an assumption of irreducibility, and explain a connection with the Bartlett theorem of Problem (8.7.6).

2. Let Q be an M(λ)/M(μ)/s queue where $\lambda < s\mu$, and assume Q is in equilibrium. Show that the process of departures is a Poisson process with intensity λ, and that departures up to time t are independent of the value of $Q(t)$.

3. Customers arrive in the manner of a Poisson process with intensity λ in a shop having two servers. The service times of these servers are independent and exponentially distributed with respective parameters μ_1 and μ_2. Arriving customers form a single queue, and the person at the head of the queue moves to the first free server. When both servers are free, the next arrival is allocated a server chosen according to one of the following rules:

(a) each server is equally likely to be chosen,

(b) the server who has been free longer is chosen.

Assume that $\lambda < \mu_1 + \mu_2$, and the process is in equilibrium. Show in each case that the process of departures from the shop is a Poisson process, and that departures prior to time t are independent of the number of people in the shop at time t.

4. **Difficult customers.** Consider an M(λ)/M(μ)/1 queue modified so that on completion of service the customer leaves with probability δ, or rejoins the queue with probability $1-\delta$. Find the distribution of the total time a customer spends being served. Hence show that equilibrium is possible if $\lambda < \delta\mu$, and find the stationary distribution. Show that, in equilibrium, the departure process is Poisson, but if the rejoining customer goes to the end of the queue, the composite arrival process is not Poisson.

5. Consider an open migration process in equilibrium. If there is no path by which an individual at station k can reach station j, show that the stream of individuals moving directly from station j to station k forms a Poisson process.

11.8 Problems

1. **Finite waiting room.** Consider M(λ)/M(μ)/k with the constraint that arriving customers who see N customers in the line ahead of them leave and never return. Find the stationary distribution of queue length for the cases $k = 1$ and $k = 2$.

2. **Baulking.** Consider M(λ)/M(μ)/1 with the constraint that if an arriving customer sees n customers in the line ahead of him, he joins the queue with probability $p(n)$ and otherwise leaves in disgust.

(a) Find the stationary distribution of queue length if $p(n) = (n+1)^{-1}$.

(b) Find the stationary distribution π of queue length if $p(n) = 2^{-n}$, and show that the probability that an arriving customer joins the queue (in equilibrium) is $\mu(1-\pi_0)/\lambda$.

3. **Series.** In a Moscow supermarket customers queue at the cash desk to pay for the goods they want; then they proceed to a second line where they wait for the goods in question. If customers arrive in the shop like a Poisson process with parameter λ and all service times are independent and exponentially distributed, parameter μ_1 at the first desk and μ_2 at the second, find the stationary distributions of queue lengths, when they exist, and show that, at any given time, the two queue lengths are independent in equilibrium.

4. **Batch (or bulk) service.** Consider M/G/1, with the modification that the server may serve up to m customers simultaneously. If the queue length is less than m at the beginning of a service period then she serves everybody waiting at that time. Find a formula which is satisfied by the probability generating function of the stationary distribution of queue length at the times of departures, and evaluate this generating function explicitly in the case when $m = 2$ and service times are exponentially distributed.

5. Consider M(λ)/M(μ)/1 where $\lambda < \mu$. Find the moment generating function of the length B of a typical busy period, and show that $\mathbb{E}(B) = (\mu-\lambda)^{-1}$ and $\text{var}(B) = (\lambda+\mu)/(\mu-\lambda)^3$. Show that the density function of B is

$$f_B(x) = \frac{\sqrt{\mu/\lambda}}{x} e^{-(\lambda+\mu)x} I_1\left(2x\sqrt{\lambda\mu}\right) \qquad \text{for } x > 0$$

where I_1 is a modified Bessel function.

6. Consider M(λ)/G/1 in equilibrium. Obtain an expression for the mean queue length at departure times. Show that the mean waiting time in equilibrium of an arriving customer is $\frac{1}{2}\lambda\mathbb{E}(S^2)/(1-\rho)$ where S is a typical service time and $\rho = \lambda\mathbb{E}(S)$.

Amongst all possible service-time distributions with given mean, find the one for which the mean waiting time is a minimum.

7. Let W_t be the time which a customer would have to wait in a M(λ)/G/1 queue if he were to arrive at time t. Show that the distribution function $F(x;t) = \mathbb{P}(W_t \le x)$ satisfies

$$\frac{\partial F}{\partial t} = \frac{\partial F}{\partial x} - \lambda F + \lambda\mathbb{P}(W_t + S \le x)$$

where S is a typical service time, independent of W_t.

Suppose that $F(x, t) \to H(x)$ for all x as $t \to \infty$, where H is a distribution function satisfying $0 = h - \lambda H + \lambda \mathbb{P}(U + S \le x)$ for $x > 0$, where U is independent of S with distribution function H, and h is the density function of H on $(0, \infty)$. Show that the moment generating function M_U of U satisfies

$$M_U(\theta) = \frac{(1-\rho)\theta}{\lambda + \theta - \lambda M_S(\theta)}$$

where ρ is the traffic intensity. You may assume that $\mathbb{P}(S = 0) = 0$.

8. Consider a G/G/1 queue in which the service times are constantly equal to 2, whilst the interarrival times take either of the values 1 and 4 with equal probability $\frac{1}{2}$. Find the limiting waiting time distribution.

9. Consider an extremely idealized model of a telephone exchange having infinitely many channels available. Calls arrive in the manner of a Poisson process with intensity λ, and each requires one channel for a length of time having the exponential distribution with parameter μ, independently of the arrival process and of the duration of other calls. Let $Q(t)$ be the number of calls being handled at time t, and suppose that $Q(0) = I$.

Determine the probability generating function of $Q(t)$, and deduce $\mathbb{E}(Q(t))$, $\mathbb{P}(Q(t) = 0)$, and the limiting distribution of $Q(t)$ as $t \to \infty$.

Assuming the queue is in equilibrium, find the proportion of time that no channels are occupied, and the mean length of an idle period. Deduce that the mean length of a busy period is $(e^{\lambda/\mu} - 1)/\lambda$.

10. Customers arrive in a shop in the manner of a Poisson process with intensity λ, where $0 < \lambda < 1$. They are served one by one in the order of their arrival, and each requires a service time of unit length. Let $Q(t)$ be the number in the queue at time t. By comparing $Q(t)$ with $Q(t+1)$, determine the limiting distribution of $Q(t)$ as $t \to \infty$ (you may assume that the quantities in question converge). Hence show that the mean queue length in equilibrium is $\lambda(1 - \frac{1}{2}\lambda)/(1-\lambda)$.

Let W be the waiting time of a newly arrived customer when the queue is in equilibrium. Deduce from the results above that $\mathbb{E}(W) = \frac{1}{2}\lambda/(1-\lambda)$.

11. Consider M(λ)/D(1)/1, and suppose that the queue is empty at time 0. Let T be the earliest time at which a customer departs leaving the queue empty. Show that the moment generating function M_T of T satisfies

$$\log\left(1 - \frac{s}{\lambda}\right) + \log M_T(s) = (s - \lambda)\big(1 - M_T(s)\big),$$

and deduce the mean value of T, distinguishing between the cases $\lambda < 1$ and $\lambda \ge 1$.

12. Suppose $\lambda < \mu$, and consider a M(λ)/M(μ)/1 queue Q in equilibrium.

(a) Show that Q is a reversible Markov chain.

(b) Deduce the equilibrium distributions of queue length and waiting time.

(c) Show that the times of departures of customers form a Poisson process, and that $Q(t)$ is independent of the times of departures prior to t.

(d) Consider a sequence of K single-server queues such that customers arrive at the first in the manner of a Poisson process, and (for each j) on completing service in the jth queue each customer moves to the $(j+1)$th. Service times in the jth queue are exponentially distributed with parameter μ_j, with as much independence as usual. Determine the (joint) equilibrium distribution of the queue lengths, when $\lambda < \mu_j$ for all j.

13. Consider the queue M(λ)/M(μ)/k, where $k \ge 1$. Show that a stationary distribution $\boldsymbol{\pi}$ exists if and only if $\lambda < k\mu$, and calculate it in this case.

Suppose that the cost of operating this system in equilibrium is

$$Ak + B\sum_{n=k}^{\infty}(n - k + 1)\pi_n,$$

the positive constants A and B representing respectively the costs of employing a server and of the dissatisfaction of delayed customers.

Show that, for fixed μ, there is a unique value λ^* in the interval $(0, \mu)$ such that it is cheaper to have $k = 1$ than $k = 2$ if and only if $\lambda < \lambda^*$.

14. Customers arrive in a shop in the manner of a Poisson process with intensity λ. They form a single queue. There are two servers, labelled 1 and 2, server i requiring an exponentially distributed time with parameter μ_i to serve any given customer. The customer at the head of the queue is served by the first idle server; when both are idle, an arriving customer is equally likely to choose either.
(a) Show that the queue length settles into equilibrium if and only if $\lambda < \mu_1 + \mu_2$.
(b) Show that, when in equilibrium, the queue length is a time-reversible Markov chain.
(c) Deduce the equilibrium distribution of queue length.
(d) Generalize your conclusions to queues with many servers.

15. Consider the D(1)/M(μ)/1 queue where $\mu > 1$, and let Q_n be the number of people in the queue just before the nth arrival. Let Q_μ be a random variable having as distribution the stationary distribution of the Markov chain $\{Q_n\}$. Show that $(1 - \mu^{-1})Q_\mu$ converges in distribution as $\mu \downarrow 1$, the limit distribution being exponential with parameter 2.

16. Taxis arrive at a stand in the manner of a Poisson process with intensity τ, and passengers arrive in the manner of an (independent) Poisson process with intensity π. If there are no waiting passengers, the taxis wait until passengers arrive, and then move off with the passengers, one to each taxi. If there is no taxi, passengers wait until they arrive. Suppose that initially there are neither taxis nor passengers at the stand. Show that the probability that n passengers are waiting at time t is $(\pi/\tau)^{\frac{1}{2}n} e^{-(\pi+\tau)t} I_n(2t\sqrt{\pi\tau})$, where $I_n(x)$ is the modified Bessel function, i.e., the coefficient of z^n in the power series expansion of $\exp\{\frac{1}{2}x(z + z^{-1})\}$.

17. Machines arrive for repair as a Poisson process with intensity λ. Each repair involves two stages, the ith machine to arrive being under repair for a time $X_i + Y_i$, where the pairs (X_i, Y_i), $i = 1, 2, \dots$, are independent with a common joint distribution. Let $U(t)$ and $V(t)$ be the numbers of machines in the X-stage and Y-stage of repair at time t. Show that $U(t)$ and $V(t)$ are independent Poisson random variables.

18. Ruin. An insurance company pays independent and identically distributed claims $\{K_n : n \geq 1\}$ at the instants of a Poisson process with intensity λ, where $\lambda\mathbb{E}(K_1) < 1$. Premiums are received at constant rate 1. Show that the maximum deficit M the company will ever accumulate has moment generating function

$$\mathbb{E}(e^{\theta M}) = \frac{(1-\rho)\theta}{\lambda + \theta - \lambda\mathbb{E}(e^{\theta K})}.$$

19. (a) **Erlang's loss formula.** Consider M(λ)/M(μ)/s with baulking, in which a customer departs immediately if, on arrival, he sees all the servers occupied ahead of him. Show that, in equilibrium, the probability that all servers are occupied is

$$\pi_s = \frac{\rho^s/s!}{\sum_{j=0}^{s} \rho^j/j!}, \qquad \text{where } \rho = \lambda/\mu.$$

(b) Consider an M(λ)/M(μ)/∞ queue with channels (servers) numbered $1, 2, \dots$. On arrival, a customer will choose the lowest numbered channel that is free, and be served by that channel. Show in the notation of part (a) that the fraction p_c of time that channel c is busy is $p_c = \rho(\pi_{c-1} - \pi_c)$ for $c \geq 2$, and $p_1 = \pi_1$.