Chapter 24

Markov queues

Queues involve the following essential elements. Customers arrive at random times in random numbers and make random requests of a service. The server then prioritises these requests into queues and responds to each request, the resources and time spent on each request being (to some extent) random. These common features arise in many contexts in life and business, both as a customer and a server. Consider some examples:

- taking an examination paper. The candidate is the server, processing a number of requests (exam questions) within a given time limit.
- e-mail/messaging. You receive a number of messages on your social networking page and decide which to prioritise, which to ignore and and how much time to spend on each response.
- Waiting room. You arrive at an accident and emergency department of a hospital and wonder how long it will be before you are seen.
- Car wash. You operate a car wash and want to maximise your income. How many staff should you employ and how much money can you take per hour ?
- Coffee shop . How many tables and chairs do I need for customers ?
- Supermarket. How many check-out tills do I need to operate ?
- Traffic lights. How should traffic lights be operated at a junction to maximize flow ?
- Call centre. How long are your customers prepared to wait ?
- web browsing. How many file servers are required to keep your web site in operation ?
- streaming audio/video. What quality of video can you watch online ?

This chapter considers some simple continuous-time Markov chains. The specific application is to queuing theory, and in particular to answer the following questions:

Suppose I join a queue of customers, what is the length of this queue (on average) and how does it depend on the arrival rate of new customers ?
How long do I need to wait (on average) in the queue until I'm served ?
How rapidly can I process customers, and what rate of income can I generate ?

24.1 Queues

Let us assume that customers arrive according to a Poisson process. We have already seen that, in theis case (23.32):

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & 0 & -\lambda & \lambda & \cdots \\ \vdots & \vdots & & \ddots & & \ddots \end{pmatrix} \quad . \tag{24.1}$$

Let's modify the Poisson process slightly by considering a system with (Poissonian) arrivals, but also departures. For example, one could think of customers arriving at an airline check-in desk at random times, or patients arriving at an accident and emergency unit of a hospital. For the moment, let us assume that all arrivals will join the queue, if necessary, and wait until the customers/patients who arrived before them are served/treated. The state of the system would be the number of people in this queue. This number would decrease each time a customer was checked in, and increase each time a new customer arrived. So we would have balance between arrivals and departures in the system, but the length of the queue would be a (discrete) stochastic variable since the process of arrival and departure occur at random times. The simplest example of a queuing system is the called the $M/M/1$ queue, which we now consider.

24.2 M/M/1

For the model we suppose that customers arrive as a Poisson process (a Markov *memoryless*process - the first M in $M/M/1$) and once they arrive, they wait to be 'served' with a *service time* described by an exponential distribution, another Markov (*memoryless*) process - denoting the second M in $M/M/1$. Once the customer has been served they leave the system (depart the queue). The 1 in $M/M/1$, denotes the fact that we have 1 service point.

If customers arrive when the server is busy checking in someone, a queue will form. Let us define the state of this system by the number of people in the queue (or simply the 'length' of the queue) at any time. Since we have two competing random processes, arrivals and departures, the state of the system: $X(t) = n \in \{0, 1, 2, 3, \dots\}$ is a discrete random variable.

We can visualise the transitions by using a transition graph, only now the edges (connections) indicate the jump rates. So each arrow to the right denotes and arrival process, each arrow to the left a departure process, as shown below (figure 24.1).

Let us find the jump-rate matrix for this system. We know quite a bit about arrivals, since this is just a Poisson process, and we take this to have a rate λ.

24.2.1 Departures

As regards departures, we can assume that departures are independent of arrivals in the following sense. The server will take the same time with a customer whether the queue has 3 persons or 13 persons waiting! Suppose this service time is an exponential process, with parameter μ. So, let us stop the arrivals for a moment, and think of n persons in the queue. The queue will eventually empty as each person is served in turn.

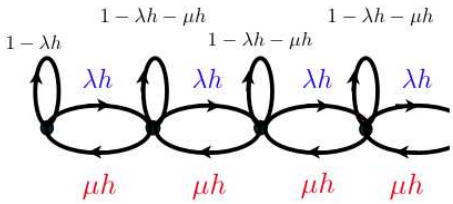

Figure 24.1: Transition graph for an $M/M/1$ queue, with infinite capacity. We have Poisson arrivals with rate λ and exponential departures with rate μ, over a short time h. The state of the system in the length of the queue. The state on the far left is $n = 0$, corresponding to an empty queue. Thus, in this short time, one can stay in the state, n, jump to $n + 1$ by an arrival, or jump back to $n - 1$ with a departure. The system will only have an equilibrium state if the rate of arrivals is less that the rate of departures: $< \mu$.

The probability of the queue reducing from n to $n - 1$ in a time h is simply the probability that the service completes in that time. Let T_s denote the service time, then:

$$P(T_s \leq h) = 1 - e^{-\mu h}$$

and for a very short time $h \to 0$

$$P(T_s \leq h) \approx \mu h + o(h) \qquad . \tag{24.2}$$

That is, in the absence of arrivals:

$$P(X(h) = n - 1 | X(0) = n) = \mu h + o(h) \qquad . \tag{24.3}$$

We see that μ represents the rate of departures, in the same way that λ is the rate of arrivals. I have said that these processes are taken to be independent. So, in this case, over a short time h, the transition graph would have the form shown below (24.1)

An example of what might occur is shown below for $\lambda = 0.3$ and $\mu = 0.4$ in figure (24.2). This illustrates a Monte Carlo simulation of an M/M/1 queue with these parameters. The jumps indicate arrivals and departures at random times. We see the queue length fluctuates at all times, however we can estimate the time average of the length of the queue.

Therefore to include both arrivals and departures is the complete jump-rate matrix, we proceed as follows. Suppose the state of the system is $X(t) = n$, then in a very short time h there can be at most one arrival and/or one departure. Consider the jump rate for the queue to increase on this time. This means having exactly *one* arrival and *no* departures:

$$Q_{n,n+1} = \lim_{h \to 0} \frac{1}{h} \left(P(X(h) = n + 1 | X(0) = n) \right) = \lim_{h \to 0} \frac{1}{h} \left(\lambda h + o(h) \right) \times \left(1 - \mu h + o(h) \right) \qquad . \tag{24.4}$$

That is

$$Q_{n,n+1} = \lambda \qquad . \tag{24.5}$$

The jump rate for the queue to shorten in this time, h, means having exactly *one* departure and *no* arrivals:

$$Q_{n,n-1} = \lim_{h \to 0} \frac{1}{h} \left(P(X(h) = n + 1 | X(0) = n) \right) = \lim_{h \to 0} \frac{1}{h} \left(\mu h + o(h) \right) \times \left(1 - \lambda h + o(h) \right) \tag{24.6}$$

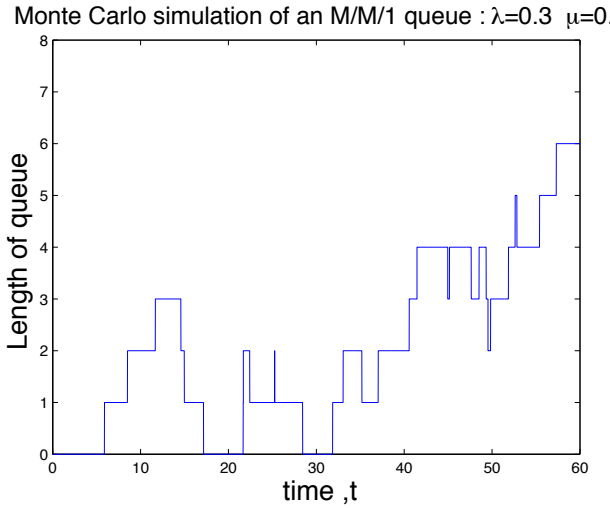

Figure 24.2: Monte Carlo simulation of an M/M/1 queue with arrival rate ($\lambda = 0.3$) and departure rate ($\mu = 0.4$) with infinite capacity. The state of the system is the number of persons in the queue (length of the queue) including the person being served.

that is

$$Q_{n,n-1} = \mu \qquad . \tag{24.7}$$

Finally, for the queue to stay the same length $n \to n$, over this short time h, requires either *no* arrival and *no* departures, or exactly *one* arrival and *one* departure.

$$Q_{n,n} = \lim_{h\to 0} \quad \frac{1}{h}\{P(X(h) = n|X(0) = n) - 1\} \tag{24.8}$$

that is,

$$Q_{nn} = \lim_{h\to 0} \quad \frac{1}{h}\left[(1 - \lambda h + o(h))(1 - \mu h + o(h)) + (\lambda h + o(h))(\mu h + o(h)) - 1\right] \qquad . \tag{24.9}$$

Then neglecting all the tiny terms, that is using the fact that $\lim_{h\to 0} o(h)/h = 0$, we get:

$$Q_{n,n} = -\lambda - \mu \qquad . \tag{24.10}$$

There is a special case when $n = 0$, that is when the queue is empty. In that case no departures are possible and: $Q_{00} = -\lambda$.

That is, explicitly, this is the form of the jump matrix:

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots \\ \mu & -\lambda-\mu & \lambda & 0 & 0 & \cdots & \\ 0 & \mu & -\lambda-\mu & \lambda & 0 & \cdots & \\ 0 & 0 & \mu & -\lambda-\mu & \lambda & \cdots & \\ 0 & 0 & 0 & \mu & -\lambda-\mu & \lambda & \cdots \\ \vdots & \vdots & & \ddots & & \ddots & \end{pmatrix} \tag{24.11}$$

This is the jump rate matrix for an $M/M/1$ queue, with *infinite capacity*. That is, a queue in which we allows the number of persons to be infinite.

Suppose that the queue has a finite capacity. That is, once it reaches a length N, and new arrival occur,

they are not allowed to join the queue. In that case, the only transitions are departures: $N \to N-1$. Then $Q_{N,N-1} = \mu$ and $Q_{NN} = -\mu$. So, for an $M/M/1$ queue with finite capacity, the jump-rate matrix has the form:

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots & 0 \\ \mu & -\lambda-\mu & \lambda & 0 & 0 & \cdots & 0 \\ 0 & \mu & -\lambda-\mu & \lambda & 0 & \cdots & 0 \\ 0 & 0 & \mu & -\lambda-\mu & \lambda & \cdots & 0 \\ \vdots & \vdots & & \ddots & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \mu & -\lambda-\mu & \lambda \\ 0 & \cdots & \cdots & 0 & 0 & \mu & -\mu \end{pmatrix} \tag{24.12}$$

24.3 Equilibrium for the M/M/1 queue

In the queue we have arrivals and departures, so it is natural to ask if these balance out in some way to create an equilibrium state of the system. before discussing this question in detail, we can make some tentative guesses in this direction. Suppose we have an $M/M/1$ queue with infinite capacity. Then if the rate of arrivals (λ) is higher than the rate of departures (μ), that is $\lambda > \mu$, then we will not have equilibrium. That is the queue will, inexorably, lengthen if it we allow people to join without a limit.

However, consider the case in which we might have equilibrium, that is a non-trival $\boldsymbol{\pi}$ such that:

$$\boldsymbol{\pi}\mathsf{Q} = 0 \qquad . \tag{24.13}$$

For the $M/M/1$ queue, with infinite capacity, this means:

$$\begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \cdots & \cdots \end{pmatrix} \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots \\ \mu & -\lambda-\mu & \lambda & 0 & 0 & \cdots & \\ 0 & \mu & -\lambda-\mu & \lambda & 0 & \cdots & \\ 0 & 0 & \mu & -\lambda-\mu & \lambda & \cdots & \\ 0 & 0 & 0 & \mu & -\lambda-\mu & \lambda & \cdots \\ \vdots & \vdots & & \ddots & & \ddots & \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \cdots & \cdots \end{pmatrix} \tag{24.14}$$

Multiplying out the left-hand side gives, for the first column:

$$\pi_0(-\lambda) + \pi_1\mu = 0 \qquad , \qquad \Rightarrow \qquad \pi_1 = (\lambda/\mu)\pi_0 \qquad . \tag{24.15}$$

For the second column:

$$\pi_0(\lambda) + \pi_1(-\lambda-\mu) + \pi_2\mu = 0 \qquad , \qquad \Rightarrow \qquad \pi_2 = (\lambda/\mu)\pi_1 \qquad . \tag{24.16}$$

where we have substituted for, $\pi_0 = (\mu/\lambda)\pi_1$.

We see the following pattern emerging:

$$\boxed{\pi_n = \left(\frac{\lambda}{\mu}\right)\pi_{n-1}} \qquad . \tag{24.17}$$

Then we can use recursion (iteration) to see that:

$$\pi_n = \left(\frac{\lambda}{\mu}\right)\pi_{n-1} = \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)\pi_{n-2} \quad . \tag{24.18}$$

Continuing in this manner until we reach π_0 we have:

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0 \tag{24.19}$$

where π_n in the probability that we have n persons in the queue.

Let us denote the ratio:

$$\boxed{\rho \equiv \frac{\lambda}{\mu}} \tag{24.20}$$

often called the *load factor* or *intensity*. This can be expressed as:

$$\rho = \frac{\text{expected service time}}{\text{expected time between arrivals}} \tag{24.21}$$

Then we must *normalize* the probability distribution, that is, we need to ensure that:

$$\boxed{\sum_{n=0}^{\infty} \pi_n = 1} \quad . \tag{24.22}$$

For this system that means:

$$\sum_{n=0}^{\infty} \rho^n \pi_0 = \pi_0 \sum_{n=0}^{\infty} \rho^n = 1 \tag{24.23}$$

This infinite series is simply the geometric progression (series). The series converges if, and only if, $\rho < 1$. That is we must have $\lambda < \mu$, in other words the rate of arrivals cannot exceed the rate of departures. This is in line with the conjecture we made at the start of this section.

However, when $\rho < 1$, then we can sum the series so that:

$$\sum_{n=0}^{\infty} \rho^n = \frac{1}{1-\rho} \quad .$$

This gives us:

$$\pi_0 \times \frac{1}{1-\rho} = 1 \quad , \quad \Rightarrow \quad \pi_0 = 1 - \rho \quad . \tag{24.24}$$

Then finally we have the probabilty distribution for an $/M/M/1$ queue. which is a *geometric distribution*:

$$\boxed{\pi_n = (1-\rho)\rho^n} \quad , \quad \rho < 1 \quad . \tag{24.25}$$

24.4 Detailed balance

For an equilibrium to exist the transition rates must be in balance. That is the flow 'out' equals the flow 'in'.

$n \qquad n+1$

Recall that *detailed balance* means balance between *each* state, and in particular two neighbouring states that are connected by arrivals and departures. So, in this context we have:

$$\underbrace{\pi_n \mathsf{Q}_{n,n+1}}_{\substack{\text{flow from} \\ n \text{ to } n+1}} = \underbrace{\pi_{n+1} \mathsf{Q}_{n+1,n}}_{\substack{\text{flow from} \\ n+1 \text{ to } n}}$$

This relation must hold for all states in the system.

For this M/M/1 queue - the balance of flows equates to:

$$\boxed{\pi_n \lambda = \pi_{n+1} \mu} \tag{24.26}$$

which is the same relation we obtained from the matrix equations, only this time by a faster route.

In general, the use of *detailed balance* is a much faster method of solution for the equilibrium distribution than solving the linear equations:

$$\boldsymbol{\pi} \mathsf{Q} = 0 \qquad .$$

However, the use of detailed balance requires that the chain is reversible; a property we have not proven and which requires a good deal of extra work.

24.5 Length of the queue and the waiting time

Now that we have the probability distribution, we can answer the questions posed at the beginning of this chapter: how long is the queue on average ?

Since we have a geometric distribution, the expected value for the number of persons in the queue (that is, its length) is, by definition:

$$L = \sum_{n=0}^{\infty} n \pi_n = (1-\rho) \sum_{n=0}^{\infty} n \rho^n \tag{24.27}$$

then the average length of the queue is:

$$\boxed{L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}} \qquad . \tag{24.28}$$

Clearly L increases rapidly as $\lambda \to \mu$, which agrees with intuition that as the rate of arrivals exceeds the rate of departures, the queue will get longer and longer. However, it is interesting to note the analytic form of the divergence.

24.5.1 Cost equation and Little's law

Let's consider more general aspects of queues and the arrival and departure processes. So let $N_A(t)$ be the counting process for arrivals. We suppose that this process is random and homogeneous, but not

necessarily Poissonian. Then over a long time, we can define the time-average rate of arrivals:

$$\lambda_A \equiv \lim_{T\to\infty} \frac{N_A(T)}{T} \quad . \tag{24.29}$$

Note that this 'limit' exists in the sense of convergence according to the law of large numbers, it is not a uniform convergence.

For the moment, suppose no customers leave the system before being served. These arrivals will then either be still in the queue, being served, or left the system (departures) having been served. We can label these stochastic variables as: $N_Q(t), N_S(t)$ and $N_D(t)$, respectively. Thus, assuming customers do not vanish or appear from thin air, we have:

$$N_A(T) = N_Q(t) + N_S(t) + N_D(t) \tag{24.30}$$

Invoking the ergodic theorem, we can then speak of the time-average length of the queue as:

$$L_Q \equiv \lim_{T\to\infty} \frac{N_Q(T)}{T} \quad . \tag{24.31}$$

and the time-average of customers in the system as:

$$L \equiv \lim_{T\to\infty} \frac{N_Q(T) + N_S(T)}{T} \quad . \tag{24.32}$$

Suppose these arrivals are customers who spend a random amount of money at the service point, and our server spend a random amount of time serving them. In fact, we do not need to know anything in detail about the service process to estimate the rate that our system generates money. If the system is in some kind of equilibrium, the inflow of customers is balanced by the outflow, and the customers going in are spending. So we have the *cost equation*:

$$\boxed{\text{average spending rate} = \lambda_A \times \text{average spend per customer}} \quad . \tag{24.33}$$

That is, our earning rate (which is simply the spending rate of customers) does not depend on our speed of service! This appears to defy intuition, but of course we are assuming that customers decide to *wait* as long as it takes to get served. On the other hand this waiting time (and the length of the queue) will depend on the speed of service versus the rate of arrivals. Well we know that *time is money*, and we can use this analogy in a concrete way to relate the waiting time to the length of the queue.

Let W denote the (average) total time a person spends in the system, waiting in the queue and then getting served. Denoting by W_Q the average time spent in the queue after arrival, and $\mathbb{E}(T_s)$ the average time to serve a customer, we have:

$$W = W_Q + \mathbb{E}(T_s) \quad . \tag{24.34}$$

Consider the following bizarre system of charging. Suppose that each customer *pays* continuously as they wait in the system (queueing and serving), let's say a toll of 1 Euro per unit time collected as they wait. There is no payment at the service point. Then the time (on average) each individual spends in the system would be equal to the amount they pay: W. The rate that the system earns, that is the rate of spending of customers would just be L the number of customers in the system, since each of these customers is paying at a rate of 1 Euro per customer, per unit time.

This is called *Little's law*:

$$\boxed{L = \lambda_A W} \quad . \tag{24.35}$$

This has the intuitive meaning that the waiting time, W, is proportional to the number of customers ahead of you, L. Using Little's formula we can estimate the *average waiting time*, W, for an $M/M/1$ queue:

$$\boxed{W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}} \quad . \tag{24.36}$$

24.6 M/M/1 with finite capacity

Consider the case when the queue has a finite length N. In this case, we limit the length of the queue so that any arrival, when the queue is full, are turned away (rejected).

Now consider the equilibrium distribution (the time average) distribution that will arise. Applying the *detailed balance* equations as before:

$$\pi_n \mathsf{Q}_{n,n-1} = \pi_{n-1} \mathsf{Q}_{n-1,n} \quad . \tag{24.37}$$

That is:

$$\pi_n \mu = \pi_{n-1} \lambda \quad . \tag{24.38}$$

As before, using recursion we get:

$$\pi_n = \pi_0 \rho^n \qquad n = 0, 1, 2, \ldots, N \quad . \tag{24.39}$$

However, the normalization condition is a finite geometric series:

$$\sum_{n=0}^{N} \pi_n = 1 \quad . \tag{24.40}$$

This gives:

$$\pi_0 \sum_{n=0}^{N} \rho^n = \pi_0 \times \frac{1 - \rho^{N+1}}{1 - \rho} \quad . \tag{24.41}$$

Thus we can write:

$$\boxed{\pi_n = \frac{(1-\rho)\rho^n}{(1-\rho^{N+1})} \quad , \quad n = 0, 1, 2, \ldots, N} \quad . \tag{24.42}$$

So the probability that the queue is empty, $n = 0$, or full, $n = N$, at any time is:

$$\pi_0 = \frac{(1-\rho)}{(1-\rho^{N+1})} \quad , \quad \pi_N = \frac{(1-\rho)\rho^N}{(1-\rho^{N+1})} \quad . \tag{24.43}$$

The average number of people in the queue (over time) will be:

$$\mathbb{E}(n) = \sum_{n=0}^{N} n \pi_n \tag{24.44}$$

We note that if the arrival rate greatly exceeds the departure rate $\rho \to \infty$, then:

$$\pi_0 \to 0 \quad , \quad \pi_N \to 1 \quad . \tag{24.45}$$

This is as we would expect.

24.7 Lost customers

Suppose the queue is full at some time. Then any customers arriving at that time will be tuned away and 'lost' to the system. Using the ergodic theorem, the fraction of time that the system is full is just: π_N. And since customers arrive at a rate λ that means the rate of customers that are rejected is:

$$\mathsf{Loss} \quad \mathsf{Rate} = \lambda \pi_N \quad . \tag{24.46}$$

EXAMPLE:

Customers arrive at a coffee shop when there is a single service point. The arrival process is Poissonian with a rate $\lambda = 0.5$ per minute.

The service time is a random (exponentially distributed) variable, with average $1/\mu = 2.5$ minutes.

The shop has a maximum capacity of $N = 4$ people, so that customers arriving and finding the shop full, will not join the queue, and go to another place for coffee.

(a) Calculate the probability that the coffee shop is empty.
(b) Calculate the fraction of customers that are lost because the shop is full.
(c) Calculate the rate customers are lost due to the shop being full. (d) If the service time can be improved so tat $\mu = 0.6$ per minute, how does this affect the loss of customers ?

SOLUTION

Let's extract the numbers first:

$$\lambda = 0.5 \quad , \quad \mu = 0.4 \qquad \rho = \frac{0.5}{0.4} = 1.25$$

and, of course $N = 4$.

So the probability that the shop is empty, at any random time, will be:

$$\pi_0 = \frac{(1-\rho)}{(1-\rho^5)} = 0.122 \quad .$$

The probability that the shop is full is:

$$\pi_N = \frac{(1-\rho)\rho^4}{(1-\rho^5)} = 0.297 \quad .$$

So this is the fraction of time that it is full. That is 29.7% of the time it is full when a customer calls. Thus the fraction of lost customers is 0.297. Note that although the arrival rate exceeds the departure rate, the difference is not that great.

The rate at which these customers are lost is:

$$\lambda\pi_4 = 0.5 \times 0.297 = 0.149 \quad \text{per} \quad \text{min} \qquad .$$

Suppose that $\mu = 0.6$, then $\rho = 0.8333$. In that case, the fraction of time that the shop is full is:

$$\pi_4 = \frac{(1-\rho)\rho^4}{(1-\rho^5)} = 0.134 \qquad .$$

So only 13.4% of customers are lost.

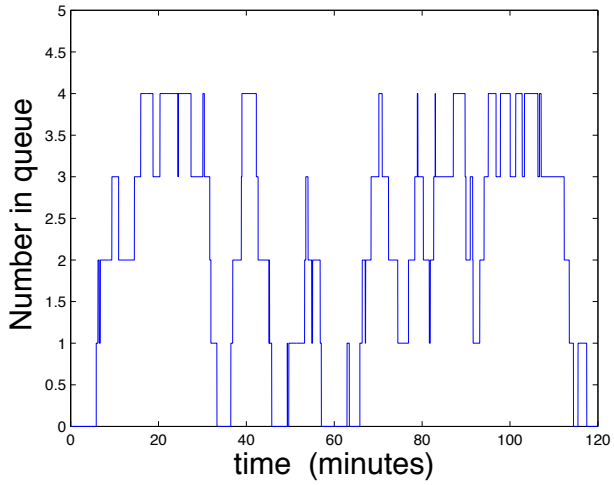

Figure 24.3: Monte Carlo simulation of an M/M/1 queue with finite capacity $N = 4$. For this simulation, the arrival rate exceeds the departure rate, $\lambda = 0.5$ and $\mu = 0.4$. We see many instances in which the queue is full with 4 people, although occasionally it empties completely. The (time) average of the number of people in the queue is the equilibrium distribution

24.8 M/G/1 and the Pollaczek-Khinchin formula

We have seen that, in the $M/M/1$ queue, that so long as (on average) arrivals are less frequent that the departures:

$$\rho \equiv \lambda \mathbb{E}\left(T_s\right) < 1 \qquad , \tag{24.47}$$

where T_s is the service time and λ^{-1} is the average interarrival time, then we can have an equilibrium in which the system is not overloaded.

In fact, this is a general result. Consider a queuing process whereby arrivals are Poissonian as before, but that the service time, T_s, is random with a general distribution $F_{T_s}(\tau)$. This is termed an $M/G/1$ queue, in which the symbol G stands for a *general service process* with 1 server. This is a more realistic scenario for customers arriving at a shop/web site, or patients arriving at a hospital, or passengers arriving at a train station. We are interested in the same kind of questions for such a system, including customer waiting time, and how rapidly can the system process customers.

Consider the queue *system* to be both the customers waiting and those being served. Customers arrive at random, join a queue, wait, get served, and then depart. The process is entirely random, and the length of the queue fluctuates, and on some occasions is empty. We are interested in the time-average behaviour

of this stochastic system.

So at any given time, t, let X_n denote the total number of people in the system left after the n person has left the system. Let Y_n denote the number of new arrivals during the service of the $(n+1)$ th customer. That is Y_n is the number of new arrivals between the nth and $(n+1)$ th customer. Then:

$$X_{n+1} = X_n - 1 + Y_n + \delta_n \tag{24.48}$$

where, the term δ_n indicates whether the queue is empty or not, after the nth departure:

$$\delta_n = \begin{cases} 1 & X_n = 0 \\ 0 & X_n \neq 0 \end{cases} \tag{24.49}$$

We speak of the equilibrium or stationary distribution arising after long times, that is when n is very large. In this case, the system is homogeneous, and according to the ergodic theorem the time-average gives us the equilibrium distribution. Then conditioning on the service time, the expected number of new arrivals is:

$$\mathbb{E}(Y_n) = \mathbb{E}(\mathbb{E}(N(T_s)|T_s)) \quad . \tag{24.50}$$

For a Poisson arrival process, this gives:

$$\mathbb{E}(Y_n) = \int_0^1 \mathbb{E}(N(\tau))\, dF_{T_s}(\tau) = \int_0^1 \lambda\tau dF_{T_s}(\tau) = \lambda\mathbb{E}(T_s) \tag{24.51}$$

and this is simply the load factor already encountered, that is:

$$\boxed{\mathbb{E}(Y_n) = \lambda\mathbb{E}(T_s) = \rho} \quad . \tag{24.52}$$

This result is completely general, given that the arrivals are Poissonian. As usual, if the load factor exceeds 1, that is $\mathbb{E}(Y_n) \geq 1$ an equilibrium cannot exist. Given that, for Poisson arrivals $\mathrm{var}(N(\tau)) = \lambda\tau$, and conditioning in the same way:

$$\mathbb{E}\left(Y_n^2\right) = \int_0^1 \mathbb{E}\left(N(\tau)^2\right) dF_{T_s}(\tau) = \int_0^1 (\lambda\tau + \lambda^2\tau^2) dF_{T_s}(\tau) = \rho + \lambda^2\mathbb{E}\left(T_s^2\right) \quad , \tag{24.53}$$

a result that will be used later.

As a result of the homogeneity of the process, we have:

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(X_n) \tag{24.54}$$

and this would be the average number of customers in the system, that is the 'time average' length of the queue, L. So that, taking expectations of (24.48), we get:

$$\mathbb{E}(\delta_n) = 1 - \mathbb{E}(Y_n) = 1 - \rho \quad . \tag{24.55}$$

So we see that if arrivals are infrequent then $\mathbb{E}(Y_n) \to 0$, and the queue is almost always empty $\mathbb{E}(\delta) = 1$. While if If arrivals are frequent then $\mathbb{E}(Y_n) \to 1$, the queue is almost always full: $\mathbb{E}(\delta) = 0$.

Now, we note that: $\delta_n^2 = \delta_n$ and that $\delta_n X_n = 0$, then we have:

$$X_{n+1}^2 = X_n^2 + 1 + Y_n^2 + \delta_n - 2(X_n + Y_n + \delta_n) + 2X_nY_n + 2Y_n\delta_n \tag{24.56}$$

The number of customers arriving at any time is independent of the length of the queue at that time. Hence Y_n in independent of X_n and indeed δ_n. Taking expectations of this equation, in the limit $n \to \infty$ and using (24.53) gives:

$$\mathbb{E}\left(X_\infty^2\right) = \mathbb{E}\left(X_\infty^2\right) + 1 + \rho + \lambda^2 \mathbb{E}\left(T_s^2\right) + (1-\rho) - 2\mathbb{E}\left(X_\infty\right) - 2\rho - 2(1-\rho) + 2\mathbb{E}\left(X_\infty\right)\rho + 2\rho(1-\rho) \quad (24.57)$$

$$2(1-\rho)\mathbb{E}\left(X_\infty\right) = 2\rho(1-\rho) + \lambda^2 \mathbb{E}\left(T_s^2\right) \quad (24.58)$$

Hence:

$$\mathbb{E}\left(X_\infty\right) = L = \rho + \frac{\lambda^2 \mathbb{E}\left(T_s^2\right)}{2(1-\rho)} \quad . \quad (24.59)$$

a result which is true for any random service process. Let's verify that it holds for the $M/M/1$ case, in which $\mathbb{E}\left(T_s^2\right) = 2/\mu^2$, and:

$$L = \rho + \frac{\rho^2}{(1-\rho)} = \frac{\rho}{(1-\rho)} \quad ,$$

in agreement with the result obtained previously (24.28).

According to Little's law (24.35) the (average) waiting time in the system is $W = L/\lambda$. The waiting time in the queue itself is, on average, $W_Q = W - \mathbb{E}\left(T_s\right)$, that is the time in the system excluding the service time, so that:

$$W_Q = \rho/\lambda + \frac{\lambda \mathbb{E}\left(T_s^2\right)}{2(1-\rho)} - \mathbb{E}\left(T_s\right) \quad . \quad (24.60)$$

This then gives us the formula derived by Pollaczek (1930) and Kinchin (1932):

$$\boxed{W_Q = \frac{\lambda \mathbb{E}\left(T^2\right)}{2(1-\lambda \mathbb{E}\left(T\right))}} \quad . \quad (24.61)$$

EXAMPLE 1

On a saturday morning, cars arrive at *Squeaky Clean* car wash in the manner of a Poisson process at a rate 0.2 per minute. The car wash cycle is a fixed price of £3 and takes exactly 4 minutes.

Calculate the expected number of cars in the system (queue and wash) at any time (on average).
Calculate the average time a customer needs to wait (between arrival and the wash starting).
Calculate the (average) income at the car wash over a 4 hour period.

ANSWER

This is an $M/G/1$ system, although it is greatly simplified by having a fixed service time. So the only stochastic aspect is the arrival process. From the data given we have $\mathbb{E}\left(T_s\right) = 5$ and $\mathbb{E}\left(T_s^2\right) = 25$, with $\rho = 0.2 \times 4 = 0.8$.

Then

$$L = \rho + \frac{\lambda^2 \mathbb{E}\left(T_s^2\right)}{2(1-\rho)} = \quad .$$

EXAMPLE 2

A bar has separate gents and ladies toilets. Male and female customers arrive at the toilets with the same Poisson rate $\lambda = 0.5$ customers per minute. The time spent in the toilet is a continuous random variable with average 0.8 minutes (men) and 1.8 minutes (women), with respective standard deviations 0.2 min and 0.4 min. If the toilet is occupied, customers queue. Calculate the expected time a male/female customer spends in the queue.

ANSWER

This is an $M/G/1$ queue with infinite capacity. Then we can use the Pollaczek-Khinchin formula.

For men:

$\mathbb{E}(T_s) = 0.8$ and $\text{var}(T_s) = (0.2)^2$, therefore $\mathbb{E}\left(T_s^2\right) = 0.68$ min^2. Then,

$$W_Q = \frac{\lambda \mathbb{E}\left(T_s^2\right)}{2(1 - \lambda \mathbb{E}(T))} = 0.28 \text{ min}$$

For women:
$\mathbb{E}(T_s) = 1.8$ and $\text{var}(T_s) = (0.4)^2$, therefore $\mathbb{E}\left(T_s^2\right) = 3.4$ min^2. and hence,

$$W_Q = \frac{\lambda \mathbb{E}\left(T_s^2\right)}{2(1 - \lambda \mathbb{E}(T))} = 8.5 \text{ min}$$

There is a dramatic difference in the values. So, although the mean times differ by roughly a factor of 2, the time spent in the queue is about 30 times longer!