MA423 - Fundamentals of Operations Research
Lecture 7: Queueing Theory

Katerina Papadaki[1]

Academic year 2017/18

[1]London School of Economics and Political Science. Houghton Street London WC2A 2AE (k.p.papadaki@lse.ac.uk)

Contents

Chapter 1

Queueing theory - continued

Appendices to these Lecture Notes A number of appendices are included at the end of these lecture notes. Appendix 2.1 goes into some details about the Poisson process, and how it relates to the Poisson and the exponential distribution. Appendices 2.2 and 2.2 give expressions for finite and infinite sums of geometric or arithmetic-geometric progressions, which will be used several times during this two lectures.

1.1 Revision of last lecture

Remember that $p_n(t)$ is the probability that there are n customers in the system at time t and $p_n = \lim_{t\to\infty} p_n(t)$ if this limit exists. So if the system has a steady state then the limit exists and p_n is the probability that the system in the long run has n customers in it. The p_n's are called steady state probabilities.

In the last lecture we modelled a queueing system as a birth and death process and we used the balance equations:

State	Balance equation
$n = 0$	$p_0\lambda_0 = p_1\mu_1$.
$n = 1$	$p_0\lambda_0 + p_2\mu_2 = p_1\lambda_1 + p_1\mu_1$
$n = 2$	$p_1\lambda_1 + p_3\mu_3 = p_2\lambda_n + p_2\mu_n$
$\vdots$	
n	$p_{n-1}\lambda_{n-1} + p_{n+1}\mu_{n+1} = p_n\lambda_n + p_n\mu_n$
$\vdots$	

Table 1.1: Balance equations for the steady state of the birth-death process.

to derive the steady state probabilities:

In the steady state, for $n = 1, 2, 3, \ldots$, the probability that n customers are in the system is

$$p_n = C_n p_0, \tag{1.1}$$

where we define

$$C_n := \frac{\lambda_0 \lambda_1 \lambda_2 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \cdots \mu_n}.$$

In the steady state, the probability that no customers are in the system is

$$p_0 = \left(1 + C_1 + C_2 + C_3 + \cdots\right)^{-1}. \tag{1.2}$$

To find the balance equations we used the "mean rate in = mean rate out" principle. In the following section we will prove the balance equations.

1.2 Proof of balance equations

First we prove the balance equations for $n = 0$ (see Table 1.1). Consider a very small time internal $[t, t + dt]$, where dt is small enough that only one event arrival or departure can occur. To be in state $n = 0$ at time $t + dt$ then system in time t can only be in states 0 or 1 since only one event can occur. Specifically, it was in state 0 at time t and state 0 in time $t + dt$ if no events occurred (here only an arrival can occur); it was in state 1 at time t and state 0 in time $t + dt$ if one departure occurred. Given that the arrival process is Poisson and we are in state 0, the probability of an arrival in $[t, t + dt]$ is $\lambda_0 dt$ and the probability of no arrival is $1 - \lambda_0 dt$. Given that the service process is Poisson and we are in state 1, the probability of a departure in $[t, t + dt]$ (given $n \geq 1$) is $\mu_1 dt$. These are shown in Figure 1.1.

Thus, the probability $p_0(t + dt)$ of being in state 0 at time $t + dt$ is equal to the probability of being in state 0 at time t, $p_0(t)$, and not getting an arrival with probability $1 - \lambda_0 dt$ plus the probability of being in state 1 at time t, $p_1(t)$, and getting a departure with probability $\mu_1 dt$. This gives us:

$$p_0(t + dt) = p_0(t)(1 - \lambda_0 dt) + p_1(t)\mu_1 dt.$$

Rewriting we get:

$$\frac{p_0(t + dt) - p_0(t)}{dt} = p_1(t)\mu_1 - p_0(t)\lambda_0.$$

Taking the limit $dt \to 0$ we get:

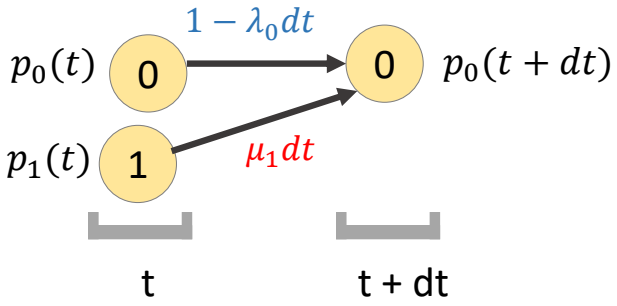


Figure 1.1: Diagram of balance equations for $n = 0$.

$$p'_0(t) = p_1(t)\mu_1 - p_0(t)\lambda_0, \tag{1.3}$$

where $p'_0(t)$ is the derivative of $p_0(t)$ with respect to t. Now assuming that there exists a steady state, then we have:

$$\lim_{t\to\infty} p'_0(t) = 0$$
$$\lim_{t\to\infty} p_0(t) = p_0$$
$$\lim_{t\to\infty} p_1(t) = p_1,$$

where the first limit comes from the fact that the probability $p_0(t)$ does not change in steady state and the last two limits come from the definitions of steady state probabilities. If we take the limit of $t \to \infty$ on (1.3) we get the balance equation for $n = 0$:

$$p_0\lambda_0 = p_1\mu_1 \tag{1.4}$$

Now we prove the balance equations for $n \geq 1$. Consider a very small time internal $[t, t+dt]$ as before. To be in state n at time $t + dt$ the system in time t can only be in states $n - 1$ or n or $n + 1$ since only one event can occur in $[t, t + dt]$. Specifically, the system goes from $n - 1$ to n with an arrival with probability $\lambda_{n-1}dt$; it goes from n to n when no arrival and no departure occur with probability $(1-\lambda_n dt)(1-\mu_n dt)$ (where we can multiply these because we assume that the arrival and service processes are independent); it goes from $n+1$ to n with a departure with probability $\mu_{n+1}dt$. These are shown in Figure 1.2.

Thus, the probability $p_n(t+dt)$ of being in state n at time $t+dt$ is equal to the probability of being in state $n-1$ at time t, $p_{n-1}(t)$, and getting an arrival with probability $\lambda_{n-1}dt$ plus the probability of being in state n at time t, $p_n(t)$, and getting no event with probability $(1-\lambda_n dt)(1-\mu_n dt)$ plus the probability of being in state $n+1$ at time t, $p_{n+1}(t)$, and getting

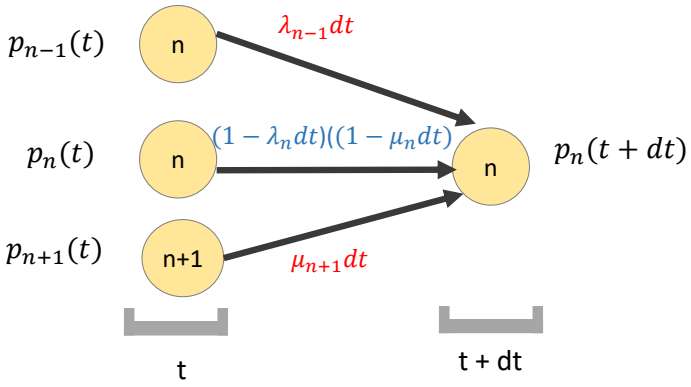


Figure 1.2: Diagram of balance equations for $n \geq 1$.

a departure with probability $\mu_{n+1}dt$. This gives us:

$$p_n(t+dt) = p_{n-1}(t)\lambda_{n-1}dt + p_n(t)(1-\lambda_n dt)(1-\mu_n dt) + p_{n+1}(t)\mu_{n+1}dt.$$

Rewriting we get:

$$\frac{p_n(t+dt) - p_n(t)}{dt} = p_{n-1}(t)\lambda_{n-1} + p_{n+1}(t)\mu_{n+1} - p_n(t)(\lambda_n + \mu_n) + p_n(t)\lambda_n\mu_n dt$$

Taking the limit $dt \to 0$ we get:

$$p'_n(t) = p_{n-1}(t)\lambda_{n-1} + p_{n+1}(t)\mu_{n+1} - p_n(t)(\lambda_n + \mu_n). \tag{1.5}$$

Now assuming that there exists a steady state, if we take the limit of $t \to \infty$ on (1.5) we get the balance equations for any $n \geq 1$:

$$p_{n-1}\lambda_{n-1} + p_{n+1}\mu_{n+1} = p_n(\lambda_n + \mu_n). \tag{1.6}$$

1.3 Little's law

Before we proceed, we present Little's law – named after John D.C. Little who proved a version of it in 1961 – which holds in great generality for any arrival-departure process. We consider a system (which could be, for example, a queuing system, or just the queue in the queueing system) where customers arrive over time, starting at time 0, spend a certain amount of time in the system, and then leave the system. We denote by

- $N(t)$ = the number of arrivals in the system in the time interval $[0, t]$,

- $L(t)$ = the number of customers in the system at time t,
- W_n = the amount of time that the nth customer to arrive spends in the system.

Note that here we make no assumption at all on the probability distributions of the inter-arrival times or of the time spent in the system. Little's law gives a general relationship between the following three quantities:

- The arrival rate $\lambda = \lim_{t \to +\infty} \frac{N(t)}{t}$ (that is, the average number of arrivals over time);
- The average time spent in the system $W = \lim_{n \to +\infty} \frac{1}{n} \sum_{j=1}^{n} W_j$;
- The average number in the system $L = \lim_{t \to +\infty} \frac{1}{t} \int_0^t L(\tau)\, d\tau$.

Note that $\int_0^t L(\tau)\, d\tau$ gives the area under $L(\tau)$ in the interval $[0, t]$ and thus this divided by t will give the height of the rectangle that has the same area as $L(\tau)$, and this height is basically the average L in $[0, t]$.

Little's law. Assuming that λ and W exist and are finite, we have $L = \lambda W$ (which means that L is also finite).

Remarks:

- Note that the above law is interesting only in the case where an infinite number of customers arrive in the system. If a finite number of customers arrive then we would have $\lambda = 0$, $L = 0$ and then the law holds trivially.
- Imagine a queueing system where the arrival rate exceeds the service rate, in which case the sojourn times in the system go to infinity because the servers cannot keep up with the customers joining the system. In this case, W is infinite and the law does not apply.

1.4 Arrival rates independent on the state

In this section we assume that there is only one server, and that the arrival rates λ_n are independent of the state n of the system This means that

$$\lambda_n \equiv \lambda, \quad n = 0, 1, 2, 3, \ldots$$

Note, in particular, that the above assumption implies that the queuing system has infinite capacity. Indeed, if the capacity was capped, then the arrival rate would drop to 0 if the system was filled at capacity, as no more customers are allowed in.

A further assumption that we make is that all servers are "identical", in the sense that we assume that each server is able to serve at the same rate μ per unit time.

1.4.1 Single server

Furthermore, since there is only one server, the service rates μ_n are also independent on n, therefore we assume

$$\mu_n \equiv \mu, \quad n = 1, 2, 3, \dots$$

It will be convenient to define the *traffic intensity* to be

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

The traffic intensity ρ is the ratio of the mean number of arrivals and mean number of serviced customers per unit time. Note that we need to assume that $\rho < 1$, otherwise the arrival rate is greater than the serving rate and the system will never reach a steady state because the queue will increase indefinitely.

Note that, in this case, the coefficients C_n defined in the previous section become

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n.$$

Using equation (1.2) from the previous section, we obtain

$$p_0 = (1 + \rho + \rho^2 + \rho^3 + \cdots)^{-1} = \left(\frac{1}{1-\rho}\right)^{-1},$$

where we used the infinite sum of a geometric progression (see equation (2.1) in Appendix 2.2).

In the steady state, the probability that there are no customers in the system is

$$p_0 = 1 - \rho. \tag{1.7}$$

By equation (1.1), we also obtain

In the steady state, for $n = 1, 2, 3, \dots$, the probability that n customers are in the system is

$$p_n = \rho^n(1 - \rho). \tag{1.8}$$

Key indicators of the queue - single server

1. The **probability that the server is idle** is $p_0 = 1 - \rho = 1 - \frac{\lambda}{\mu}$ so that:
2. The **proportion of time the server is busy** is $1 - p_0 = \rho$, and:

3. If we denote by X the random variable representing the number of customers being served at any given moment (so X equals either 0 or 1), then the **expected number of customers being served** is

$$S = \mathbb{E}[X] = 0 \cdot p_0 + 1 \cdot (1 - p_0) = \rho.$$

4. If we denote by L_S the **expected number of customer in the system**, then

$$L_S = \sum_{n=0}^{\infty}(np_n) = 0p_0 + 1p_1 + 2p_2 + \cdots = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = \frac{\rho}{1-\rho}.$$

where the second to last inequality follows from (1.8) and the last inequality follows from the fact that the infinite sum of the arithmetic-geometric progression $n\rho^n$ equals $\rho/(1-\rho)^2$ (see equation (2.2) in Appendix 2.3).

Alternatively, substituting $\rho = \lambda/\mu$, we can express L_S as

$$L_S = \frac{\lambda}{\mu - \lambda}.$$

5. If we denote by L_q the **expected number of customers in the queue**, then

$$L_q = L_S - S = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}.$$

6. Let W_q denote the **expected queuing time**. By Little's law (considering the queue itself as a system, where people depart when they leave the queue to be served)

$$W_q = \frac{L_q}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)},$$

where the last equation follows from substituting $\rho = \lambda/\mu$.

7. Let W_S be defined to be the **expected time in the system**. Applying Little's law to the whole queuing system, we get

$$W_S = \frac{L_S}{\lambda} = \frac{1}{\mu - \lambda}.$$

An alternative derivation is from the expression for W_q previously obtained

$$W_S = \mathbb{E}[\text{queuing time} + \text{service time}] = W_q + \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu} = \frac{1}{\mu-\lambda}.$$

8. **The conditional expected queuing time**, denoted by w_q is expected time a customer has to queue given that there is already a queue when s/he enters the system. Note that

$W_q = 0 \cdot p_0 + w_q(1 - p_0)$. Hence

$$w_q = \frac{W_q}{1 - p_0} = \frac{1}{\mu - \lambda},$$

where the equality follows from $W_q = \frac{\lambda}{\mu(\mu-\lambda)}$ and $p_0 = 1 - \frac{\lambda}{\mu}$.

Example 1.1. Fred Smith runs a small newsagents shop. At around mid-day, when the local school has its lunch hour, his only customers are school children. All the goods are on or behind the counter, so a child is either queuing or being served. On average a child enters the shop every 40 seconds during this period and can be served in 30 seconds.

Arrival rate is $\lambda = 90$ customers per hour, and service rate is $\mu = 120$ customers per hour, so the traffic intensity is $\rho = 0.75$.

1. The probability that the Fred is idle is $p_0 = 1 - \rho = 0.25$.
2. The proportion of time that Fred is busy is 0.75.
3. The expected number of children being served is $\rho = 0.75$.
4. The expected number of children in the shop is $L_S = \frac{0.75}{1-0.75} = 3$.
5. The expected queue length is $L_q = 3 - 0.75 = 2.25$.
6. A typical child's expected time spent in the queue is $W_q = \frac{L_q}{\lambda} = \frac{2{,}25}{90} = \frac{1}{40}$ hours, that is, 90 seconds.
7. A typical child's expected time spent in the shop is $W_S = W_q + \frac{1}{\mu} = \frac{1}{40} + \frac{1}{120} = \frac{1}{30}$ hours, that is, 2 minutes.
8. A typical child's expected time spent queuing if someone is already in the shop when s/he arrives is $w_q = \frac{1}{\mu-\lambda} = \frac{1}{30}$ hours, that is 2 minutes.

1.4.2 Multiple servers

Suppose we have two servers working at the same time with rate μ each. What is the total service rate of both servers together? Let W_1, W_2 be the service times of servers 1 and 2 respectively. We assume that the service processes are Poisson with rate parameters μ_1 and μ_2 and they are independent between servers. This means that random variables W_1, W_2 are independent exponential with rate parameters μ_1 and μ_2. We will use the following property of exponential random variables from the previous lecture:

> The minimum of k independent exponentially distributed random variables with parameters $\mu_1, \ldots, \mu_k$ is an exponential random variable with parameter $\mu_1 + \mu_2 + \cdots + \mu_k$.

This means that if the two servers are working at the same time the random variable $\min\{W_1, W_2\}$, which is the time until the next departure is exponentially distributed with rate parameter $\mu_1 + \mu_2$.

Now suppose that we have multiple servers and each of them has the same service rate μ. Then the service rate μ_n of the whole system depends on the number n of customers in the system. Indeed, when $n \leq s$ only n servers are busy, so that the service rate is $n\mu$, whereas all s servers are busy when $n > s$, in which case the service rate is $s\mu$. Note that here we are using the assumption that service times are exponentially distributed. Hence

$$\mu_n = \begin{cases} n\mu, & n = 1, \ldots, s \\ s\mu, & n = s+1, s+2, \ldots \end{cases}$$

To compute the steady state distribution p_n, $n = 0, 1, 2, \ldots$, we need to compute the coefficients C_n appearing in equation (1.1).

When $n \leq s$,

$$C_n = \frac{\lambda_0 \lambda_1 \lambda_2 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \cdots \mu_n} = \frac{\lambda^n}{(1\mu)(2\mu)\cdots(n\mu)} = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n.$$

For $n > s$, we have

$$C_n = \frac{\lambda_0 \lambda_1 \lambda_2 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \cdots \mu_n} = \frac{\lambda^n}{(1\mu)(2\mu)\cdots(s\mu)\underbrace{(s\mu)\cdots(s\mu)}_{n-s \text{ times}}} = \frac{1}{s!}\left(\frac{1}{s^{n-s}}\right)\left(\frac{\lambda}{\mu}\right)^n$$

We define the *traffic intensity* to be

$$\rho = \frac{\lambda}{s\mu},$$

so that the system is stable if $\rho < 1$, and the queue increases indefinitely if $\rho \geq 1$.

Then we can write C_n in terms of ρ, thus obtaining the following

In the steady state, for $n = 1, 2, 3, \ldots$, the probability that n customers are in the system is

$$p_n = C_n p_0,$$

where

$$C_n = \begin{cases} \dfrac{(s\rho)^n}{n!}, & n = 1, \ldots, s \\ \dfrac{(s\rho)^s}{s!}\rho^{n-s}, & n = s+1, s+2, \ldots \end{cases} \tag{1.9}$$

From equation (1.2) we can calculate the value of p_0. In order to do this, we need to compute

$$\sum_{n=1}^{\infty} C_n = \sum_{n=1}^{s} \frac{(s\rho)^n}{n!} + \frac{(s\rho)^s}{s!} \sum_{n=s+1}^{\infty} \rho^{n-s}.$$

Little can be done about the first sum, although it is simply the sum of s terms. For the second sum, we have

$$\sum_{n=s+1}^{\infty} \rho^{n-s} = \rho \sum_{n=s+1}^{\infty} \rho^{n-(s+1)} = \rho \sum_{i=0}^{\infty} \rho^i = \frac{\rho}{1-\rho},$$

since the last is the sum of a geometric progression.

Therefore, we have shown the following.

In the steady state, the probability that there are no customers in the system is

$$p_0 = \left(1 + \sum_{n=1}^{s} \frac{(s\rho)^n}{n!} + \frac{(s\rho)^s}{s!}\frac{\rho}{1-\rho}\right)^{-1}. \tag{1.10}$$

To find the other steady state probabilities we use: $p_n = C_n p_0$.

Key indicators of the queue - multiple servers

1. **Expected proportion of idle time per server.** When there are $n < s$ customers in the system, only n of the s servers will be busy, so the average proportion of idle servers is $1 - \frac{n}{s}$. This is also the proportion of time a server is idle. For example, when $n = 2$, $s = 3$, there is 1 idle server out of 3 and thus the proportion of idle servers is $\frac{1}{3}$. Futher, the proportion of time a server is idle is $\frac{1}{3}$ (when $n = 2$).

   Thus the expected proportion of idle time per server is

   $$\sum_{n=0}^{s-1} \left(1 - \frac{n}{s}\right) p_n + \sum_{n=s}^{\infty} \left(1 - \frac{s}{s}\right) p_n = \sum_{n=0}^{s-1} \left(1 - \frac{n}{s}\right) p_n$$

   This is the sum of a s terms, so it can be computed for any given problem.

2. L_q = **Expected number in the queue.** If $n \leq s$, then the queue is empty, else there are $n - s$ people in the queue. Therefore

   $$\begin{aligned} L_q &= 1p_{s+1} + 2p_{s+2} + 3p_{s+3} + \cdots \\ &= p_0(C_{s+1} + 2C_{s+2} + 3C_{s+3} + \cdots) \\ &= p_0 \frac{(s\rho)^s}{s!}(\rho + 2\rho^2 + 3\rho^3 + \cdots). \end{aligned}$$

   Since $n\rho^n$ $(n = 1, 2, 3 \ldots)$ is an arithmetic-geometric series, $\rho + 2\rho^2 + 3\rho^3 + \cdots = \rho/(1-\rho)^2$, therefore

   $$L_q = p_0 \frac{(s\rho)^s}{s!}\frac{\rho}{(1-\rho)^2}, \tag{1.11}$$

   where p_0 is given by (1.10).

3. W_q = **Expected waiting time in the queue.** By Little's law,

$$W_q = \frac{L_q}{\lambda},$$

where L_q is given by (1.11).

4. W_S = **Expected time in the system**. As before, this is

$$W_S = \mathbb{E}[\text{queuing time + service time}] = W_q + \frac{1}{\mu} = \frac{L_q}{\lambda} + \frac{1}{\mu}.$$

5. L_S = **Expected number of customers in the system.** By Little's law, this is

$$L_S = \lambda W_S = L_q + \frac{\lambda}{\mu}.$$

6. S = **Expected number of customers being served.** This is the expected number of people in the queueing system minus the expected number of people in the queue, that is

$$S = L_S - L_q = \frac{\lambda}{\mu}.$$

Example 1.2. Fred Smith has hired an assistant to help him in his shop when children are waiting to buy. As before Fred can serve a customer in 30 seconds, and his new assistant is equally fast on average. As before, children enter the shop every 40 seconds, on average.

In this example, $s = 2$, arrival rate is $\lambda = 90$ customers per hour, as before, and service rates are $\mu_1 = 120$ customers per hour if there is only one customer in the system, and $\mu_n = 240$ customers an hour if there are $n \geq 2$ customers in the system. The traffic intensity is $\rho = \frac{90}{240} = \frac{3}{8} = 0.375$.

According to (1.10),

$$p_0 = \left(1 + (0.75 + \frac{(0.75)^2}{2}) + (\frac{(0.75)^2}{2} \cdot \frac{0.375}{1 - 0.375})\right)^{-1} = (2.2)^{-1} = \frac{5}{11}.$$

1. The expected proportion of idle time per server is

$$1 \cdot p_0 + (1 - \frac{1}{2}) \cdot p_1 + 0 \cdot (1 - p_1 - p_2) = p_0 + \frac{1}{2}p_1 = p_0(1.375) = 0.625$$

hours per server

2. The expected number in the queue is

$$L_q = \frac{5}{11}\frac{(0.75)^2}{2}\frac{0.375}{(1 - 0.375)^2} \approx 0.123$$

people.

3. The expected waiting time is
$$W_q = \frac{L_q}{90} = 0.0013$$
hours, that is, about 4.9 seconds.

4. The expected time in the system is $W_S = W_q + \frac{1}{\mu} = 4.9 + 30 = 34.9$ seconds.

5. The expected number of customers in the system is $0.123 + 0.75 \approx 0.872$.

6. The expected number of customers being served is $S = 90/120 = 0.75$.

1.5 Arrival and/or service rates dependent on the state

In the following, we give examples of systems where the arrival and service rates depend on the number of customers in the system. In all these examples, we will assume that the queuing system has finite capacity and thus the number of possible states is finite. We consider two types of situations that guarantee this: first the case where we have a finite capacity waiting room (first two examples); second the case where we have a finite population (last example).

1.5.1 Single server - finite waiting room

Example 1.3. Example 1.1 is extended as follows (here we assume that Fred is the only server). Fred Smith is becoming concerned about the volume of shoplifting going on in his shop, and decides to limit the number queuing or being served to 3. Because the shop now becomes more attractive to these children when Fred is busy serving a child, the arrival rate then increases to 120 per hour when someone is already in the shop, and stays at 90 when the shop is empty. For what fraction of his time is Fred now idle? How many customers will he expect to lose because of this strategy? How long is the queue, on average?

Here there is a "finite waiting room" situation because at most 3 children are allowed in the system (and no others are allowed to be in the shop browsing). The value of n, the number in the system, is thus one of 0, 1, 2 or 3 only.

We have that $\lambda_0 = 90$, whereas $\lambda_n = 120$ children/hour, for $n = 1, 2$, but $\lambda_n = 0$ for $n = 3$. For every state $n = 1, 2, 3$, the service rate is $\mu_n = \mu = 120$ as in Example 1.1.

We must calculate p_0, p_1, p_2 and p_3, and we know that p_n will be zero for $n \geq 4$.

By Equation 1.1, we have that

$$\begin{aligned} p_1 &= \frac{\lambda_0}{\mu_1} p_0 = \frac{90}{120} p_0 = \frac{3}{4} p_0; \\ p_2 &= \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0 = \frac{90 \cdot 120}{120 \cdot 120} p_0; \\ p_3 &= \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} p_0 = \frac{90 \cdot 120 \cdot 120}{120 \cdot 120 \cdot 120} = \frac{3}{4} p_0. \end{aligned}$$

Since we know that the sum of the probability must be 1, we have

$$p_0 + p_1 + p_2 + p_3 = 1,$$

therefore

$$p_0(1 + \frac{3}{4} + \frac{3}{4} + \frac{3}{4}) = 1,$$

which gives

$$p_0 = \frac{1}{1 + \frac{9}{4}} = \frac{4}{13}, \quad p_1 = p_2 = p_3 = \frac{3}{4} \cdot \frac{4}{13} = \frac{3}{13}$$

Note that the computation of p_0 has not required us to sum an infinite series. This is always the case when there is a finite waiting room.

Fred is now idle for just over 30 percent of his time, whereas previously he was idle for only 25 percent of his time. In one hour he expects to serve $(1-4/13)*120 = 83.08$ children. His increase in idle time is approximately equal to $(\frac{4}{13} - 0.25) = \frac{3}{52}$ of an hour each hour (approximately 3.46 minutes per hour), during which time he would expect to serve $\frac{3}{52} \cdot 120 \approx 6.92$ children. So his "limited waiting room" strategy costs him 6.92 customers per hour on average. He should ask himself how this compares with the money he loses from shoplifting.

In the last part we need to compute the expected number in the queue, L_q. The queue length will be 1 or 2, with probabilities p_2 and p_3 respectively. Hence

$$L_q = 1 \cdot p_2 + 2 \cdot p_3 = \frac{3}{13} + 2 \cdot \frac{3}{13} = \frac{9}{13} \approx 0.6923$$

children.

1.5.2 Multiple servers - finite waiting room

Example 1.4. Example 1.3 is modified as follows. Smith has hired an assistant to help him in his shop when children are waiting to buy. As before Fred can serve a customer in 30 seconds, and his new assistant is equally fast on average. Whenever the shop contains at most one child a child will usually enter after 15 seconds on average, which is faster than before because now children expect to be served promptly if there are two servers, but this rate drops to one every 30 seconds if there are two children in the shop and to only one a minute if there are already 3 in the shop when a new child arrives at the shop. Fred will allow this extra (fourth) child to enter the shop in comparison to example 1.3, because he feels with two people he's got a better chance of spotting a shoplifter, but this is his limit: no more than four children are allowed in the shop. For what proportion of the day does Fred expect to be idle? How many customers will now be served per hour? If he makes an average of 15p gross profit per child and he pays his assistant 9 for his hour's work each school day, is Fred going to find this more profitable than when he limited the number of children to just three at most and didn'tt have an assistant?

The number of customers in the shop, n, is now between 0 and 4.

When $n = 1$, $\mu_1 = 120$ children per hour are served by either Fred of his assistant, but when $n = 2, 3, 4$ $\mu_n = 240$ per hour. The values of the arrival rates are $\lambda_0 = \lambda_1 = 240$ per hour, $\lambda_2 = 120$ per hour, $\lambda_3 = 60$ per hour, and $\lambda_n = 0$ for $n = 4, 5, 6, \ldots$.

Substituting into the equations (1.1):

$$\begin{aligned} p_1 &= \tfrac{240}{120} p_0 = 2p_0 \\ p_2 &= \tfrac{240 \cdot 240}{120 \cdot 240} p_0 = 2p_0 \\ p_3 &= \tfrac{240 \cdot 240 \cdot 120}{120 \cdot 240^2} p_0 = p_0 \\ p_4 &= \tfrac{240 \cdot 240 \cdot 120 \cdot 60}{120 \cdot 240^3} = \tfrac{1}{4} p_0. \end{aligned}$$

Thus $p_0(1 + 2 + 2 + 1 + \frac{1}{4}) = 1$.

Hence $p_0 = 4/25$, $p_1 = 8/25$, $p_2 = 8/25$, $p_3 = 4/25$, and $p_3 = 1/25$.

Fred is idle whenever the number in the system is $n = 0$, and on average half the time when $n = 1$ (because when there is only one child in the shop Fred and the assistant will be equally likely to be serving the child), so that the expected proportion of time Fred is idle is $p_0 + \frac{1}{2}p_1 = 8/25$ of the hour. Hence the number of children Fred expects to serve during the hour is $(1 - 8/25) \cdot 120 \approx 81.6$ children and, since the assistant is expected to serve the same number on average, approximately 163.2 children will be served during the lunch hour.

This is a large increase on the corresponding figure, 83.08, in Example 1.3, so he expects an extra 80.12 children to be served, contributing on average $80.12 \cdot 0.15 = 12.02$ additional gross profit. Since the wages of the assistant are only 9, Fred can expect on average that this (third) strategy will generate an extra 3.02 profit per school-day lunch-time.

1.5.3 Single server - finite population

Example 1.5. A 24 hour copy shop has N copy machines. Each machine fails at a rate of λ times a week. When a machine fails it is placed in a queue to be served by a mechanic. The mechanic repairs at a rate of μ machines per week. What fraction of his time is the mechanic idle? Suppose $N = 4$, $\lambda = 3$, and $\mu = 6$. What is the expected number of machines that are in the shop and working? What fraction of machines are broken at any time?

Here we have an example of a "finite population" since the arrivals come from the finite set of working machines at the copy shop. The machine failing rate λ per week be the rate that each machine fails. This means that if there is one machine in the copy shop working and $N - 1$ in the mechanics workshop, the arrival rate to the mechanics workshop is λ. If there are two machines in the copy shop and $N - 2$ at the mechanic then the arrival rate is 2λ. Thus, if there are n in the queueing system (i.e. at the mechanics workshop) then there are $N - n$ at the copy shop and the arrival rate is then $\lambda_n = (N - n)\lambda$ for $n = 0, 1, 2, N - 1$; if $n = N$ then $\lambda_N = 0$, since there are no machines in the copy shop to fail.

The service rates are $\mu_n = \mu$ for $n = 1, 2, 3, 4$.

By equation (1.1):

$$
\begin{aligned}
p_1 &= \frac{\lambda_0}{\mu_1} p_0 = \frac{N\lambda}{\mu} p_0 = N \left(\frac{\lambda}{\mu}\right) p_0; \\
p_2 &= \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0 = N(N-1) \left(\frac{\lambda}{\mu}\right)^2 p_0; \\
p_3 &= \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} p_0 = N(N-1)(N-2) \left(\frac{\lambda}{\mu}\right)^3 p_0; \\
&\dots \\
p_N &= N! \left(\frac{\lambda}{\mu}\right)^N p_0;
\end{aligned}
$$

Thus the fraction of the time that the mechanic is idle is p_0 which is equal to:

$$
p_0 = \left(1 + N \left(\frac{\lambda}{\mu}\right) + N(N-1) \left(\frac{\lambda}{\mu}\right)^2 + \dots N! \left(\frac{\lambda}{\mu}\right)^N\right)^{-1}
$$

Substituting $N = 4$, $\lambda = 3$, and $\mu = 6$, we get:

$$
p_0 = \left(1 + 2 + 3 + 3 + \frac{3}{2}\right)^{-1} = \frac{2}{21}
$$

This gives:

$$
\begin{aligned}
p_0 &= \frac{2}{21} \\
p_1 &= \frac{4}{21} \\
p_2 &= \frac{6}{21} \\
p_3 &= \frac{6}{21} \\
p_4 &= \frac{3}{21}
\end{aligned}
$$

First we want to compute the expected number of machines in the queueing sytem:

$$
L_S = 0p_0 + 1p_1 + 2p_2 + 3p_3 + 4p_4 = \frac{4 + 12 + 18 + 12}{21} = \frac{46}{21}
$$

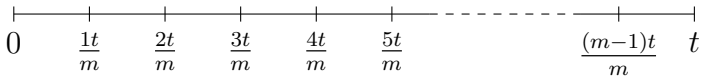
Then the expected number of machines working in the shop is $N - L_S = 4 - \frac{46}{21} = \frac{38}{21}$. The expected number of broken machines are $\frac{46}{21}$ out of 4 and thus the fraction of broken machines is $\frac{46/21}{4} = \frac{23}{46}$.

Chapter 2

Appendices

2.1 Relation between Poisson process and exponential distribution

To compute $\mathbb{P}[X(t) = n]$, sub-divide the time interval [0,t) into a very large number m of sub-intervals of width t/m



If m is large enough, then by assumption b) the probability of exactly one arrival occurring in any given time interval is $\lambda t/m$, and the probability of 0 arrivals is $1 - \lambda t/m$.

Thus "n arrivals in $[0, t)$" means: any n of these m sub-intervals each have exactly one arrival, and the other $m - n$ have no arrivals. By assumption a), the probabilities of arrival in each of the m intervals are independent, thus the probability of exactly n arrivals is given by the binomial distribution

$$f_m(n) = \binom{m}{n} \left(\frac{\lambda t}{m}\right)^n \left(1 - \frac{\lambda t}{m}\right)^{m-n}, \quad n = 0, 1, 2, 3, \ldots$$

It follows from our discussion (recalling that $\binom{m}{n} = \frac{m!}{n!(m-n)!}$) that

$$\begin{aligned}
\mathbb{P}[X(t) = n] &= \lim_{m\to\infty} f_m(n) \\
&= \frac{(\lambda t)^n}{n!} \cdot \lim_{m\to\infty} \frac{m!}{(m-n)!}\frac{1}{m^n} \cdot \lim_{m\to\infty} \left(1 - \frac{\lambda t}{m}\right)^m \cdot \lim_{m\to\infty} \left(1 - \frac{\lambda t}{m}\right)^{-n} \\
&= \frac{(\lambda t)^n}{n!} e^{-\lambda t}
\end{aligned}$$

where the last equality comes from the fact that the first limit and third limit on the second row equal 1, and by the well known fact that the second limit equals $e^{-\lambda t}$.

To compute $a(t)$, consider its cumulative distribution function $A(t)$. $A(t)$ is the probability that the time between two consecutive arrivals is no more than t, so $1 - A(t)$ is the probability

that the time between two arrivals is at least t. Because of memorylessness , $1 - A(t)$ is the probability that there are no arrivals in the time interval $[0, t)$, which by the previous discussion gives

$$1 - A(t) = \mathbb{P}[X(t) = 0] = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}.$$

Recalling that, by definition of cumulative distribution function, $A(t) := \int_0^t a(\tau) d\tau$, we have

$$\int_0^t a(\tau) d\tau = 1 - e^{-\lambda t}.$$

Taking the derivative on both sides we obtain

$$a(t) = \lambda e^{-\lambda t}.$$

The mean inter-arrival time is

$$\begin{aligned} \mathbb{E}[\text{inter-arrival time}] &= \int_0^\infty t a(t)\, dt = \int_0^\infty \lambda t e^{-\lambda t}\, dt = \frac{1}{\lambda} \int_0^\infty x e^{-x}\, dx \\ &= \frac{1}{\lambda} \left(-(1+x)e^{-x} \right) \Big|_0^\infty = \frac{1}{\lambda}. \end{aligned}$$

2.2 Geometric progression

The *geometric progression* is a sequence of the form $1, a, a^2, a^3, \ldots\ a^n, \ldots$, where $a > 0$. The sum of the first n members of the sequence is

$$S_n = 1 + a + a^2 + a^3 + \cdots + a^n.$$

Multiplying by $1 - a$ on both sides of the above equality, we obtain

$$(1-a)S_n = (1 + a + a^2 + a^3 \cdots + a^n) - (a + a^2 + a^3 + \cdots a^n + a^{n+1}) = 1 - a^{n+1},$$

implying

$$S_n = \frac{1 - a^{n+1}}{1 - a}.$$

If $a \geq 1$, then the infinite sum of the geometric progression is $+\infty$, whereas if $a < 1$ the sum is finite, and it is

$$1 + a + a^2 + a^3 + \cdots = \lim_{n \to +\infty} S_n = \lim_{n \to +\infty} \frac{1 - a^{n+1}}{1 - a} = \frac{1}{1 - a}. \tag{2.1}$$

2.3 Arithmetic-geometric progression

The *arithmetic geometric progression* is a sequence of the form $a, 2a^2, 3a^3, \dots\ na^n, \dots$, where $a > 0$. The sum of the first n members of the sequence is

$$S_n = a + 2a^2 + 3a^3 + \cdots + na^n.$$

Multiplying by $1 - a$ on both sides of the above equality, we obtain

$$(1-a)S_n = (a+2a^2+3a^3\cdots+na^n)-(a^2+2a^3+3a^4+\cdots(n-1)a^n+na^{n+1}) = a+a^2+a^3+\cdots+a^n-na^{n+1}.$$

implying

$$S_n = \frac{a}{1-a}(1 + a + \cdots + a^{n-1}) - \frac{na^{n+1}}{1-a}.$$

Note that the term between parenthesis is the sum of the first $n-1$ elements of the geometric progression, therefore it equals $(1 - a^n)/(1 - a)$, and we obtain

$$S_n = \frac{a - a^{n+1}}{(1-a)^2} - \frac{na^{n+1}}{1-a} = \frac{a}{(1-a)^2} + \frac{a^{n+1}\left(n(a-1)-1\right)}{(1-a)^2}$$

If $a \geq 1$, then the infinite sum of the geometric progression is $+\infty$, whereas if $a < 1$ the sum is finite, and it is

$$a + 2a^2 + 3a^3 + \cdots = \lim_{n\to+\infty} S_n = \frac{a}{(1-a)^2}. \tag{2.2}$$