Simple queueing models

1 $M/M/1$ queue

This model describes a queue with a single server which serves customers in the order in which they arrive. Customer arrivals constitute a Poisson process of rate λ. Recall that this means that the number of customers arriving during a time period $[t, t+s]$ is a Poisson random variable with mean λs, and it is independent of the past of the arrival process. Equivalently, the inter-arrival times (between the arrival of one customer and the next) are iid Exponential random variables with parameter λ and mean $1/\lambda$. Thus, the arrival process is Markovian, which is what is denoted by the first M in $M/M/1$.

The second M says that the service process is Markovian. Customer service times are exponentially distributed, with parameter μ and mean $1/\mu$, and the service times of different customers are mutually independent. By the memoryless property of the exponential distribution, knowing how much service a customer has already received gives us no information about how much more service it requires. When a customer has been served, it departs from the queue.

The 1 in $M/M/1$ refers to the fact that there is a single server. If there are c servers, we call it an $M/M/c$ queue. Customers are served in their order of arrival. This is called the first-come-first-served (FCFS) or first-in-first-out (FIFO) service discipline. It is not the only possible service discipline. Other examples used in practice include processor sharing (PS) where the server spreads its effort equally among all customers in the queue, and last-come-first-served (LCFS). If needed, we will specify the service discipline in force by writing $M/M/1 - PS$, $M/M/1 - LCFS$ etc., taking FCFS to be the default if nothing is specified. We assume that the queue has an infinite waiting room, so that no arriving customers are turned away. If there is a finite waiting room of size K, we call it an $M/M/1/K$ queue. We assume that the server is busy whenever there is at least one customer in the system. Such a service discipline is called work-conserving. If the service discipline is not work-conserving, we say that the server is on vacation, when there is work in the system but the server isn't working. If the customers are jobs and the servers are machines, 'vacactions' could well correspond to machine breakdowns. We won't be studying non-work conserving queues in this course; however, many of the techniques we learn can be extended to them.

Let X_t denote the number of customers in the queue at time t. What can we say about $P(X_{t+\delta} = j | X_t = i)$ for small δ? Suppose first that $i > 0$. Then, one of two things can happen. We can either have a customer arrival, or the customer currently in service can finish and depart. The probability of the first event is $\lambda\delta + o(\delta)$. The probability of the second is $\mu\delta + o(\delta)$. With the residual probability $1-(\lambda+\mu)\delta+o(\delta)$, the state remains unchanged. Moreover, these probabilities are conditionally independent of the past history of the stochastic process $X_s, s \leq t$, given the current state $X_t = i$. Thus, $X_t, t \geq 0$ is a Markov process with the following transition rates:

$$q_{i,i+1} = \lambda, \ q_{i,i-1} = \mu, \ q_{i,i} = -(\lambda + \mu), \quad \forall i \geq 1.$$

Similar reasoning shows that $q_{0,1} = \lambda$ and $q_{0,0} = -\lambda$. At this point, you might want to draw an arrow diagram to represent this Markov process with states denoting circles, and arrows between them representing transition rates. (The arrow from a state to itself is conventionally omitted as its rate is implicit from those of the outgoing arrows from that state.)

Observe that the Markov process $X_t, t \geq 0$ is a birth-death process since it has state space $\{0, 1, 2, \ldots\}$ and the only allowed transitions from state i are to states $i + 1$ and $i - 1$. Hence, it is reversible provided it has an invariant distribution. Let's suppose for the moment that it has an invariant distribution π. Then π must solve the detailed balance equations

$$\pi_i q_{ij} = \pi_j q_{ji} \quad \forall i, j, \ \text{i.e.,} \ \pi_i \lambda = \pi_{i+1} \mu, \quad i = 0, 1, 2, \ldots, \tag{1}$$

along with the normalisation condition

$$\sum_{i=0}^{\infty} \pi_i = 1. \tag{2}$$

We obtain from (1) that $\pi_i = (\lambda/\mu)\pi_{i-1}$ for all $i \geq 1$. Hence, denoting λ/μ

by ρ, we get $\pi_i = \rho^i \pi_0$. Substituting this in (2), we see that

$$\pi_i = (1-\rho)\rho^i, \ \forall i \geq 0, \quad \text{if } \rho < 1. \tag{3}$$

If $\rho \geq 1$ on the other hand, then there is no invariant distribution. Is there an intuitive explanation of this fact?

Note that $\rho \geq 1$ corresponds to $\lambda \geq \mu$. Now λ is the rate at which customers arrive and μ is the maximum rate at which they can be served. Hence, if $\rho > 1$, then work enters the system faster than it can be served, so it tends to accumulate and the total work (total number of customers) tends to infinity over time. In this case, we see that the queue is unstable. If $\rho < 1$, then all work entering the system can be served and the queue is stable. In this case, there is a steady state, and the invariant distribution describes the steady state or long run behaviour of the system. If $\rho = 1$, then the queue is critically loaded. If both arrival and service processes were deterministic, then the queue would have a steady state behaviour. However, random fluctuations cause work to gradually accumulate in the queue, and the queue length again tends to infinity over time. Nevertheless, the critical case is different from the unstable case in that the queue length grows in proportion to $\sqrt{t}$ rather than t as the time t tends to infinity.

2 Performance measures

From a practical point of view, we are interested in how many customers are typically in the queue, and how long a typical customer needs to wait for service. Let N denote the random number of customers in the queue in steady state, including the customer receiving service (if any). The invariant queue length distribution computed in (3) tells us that

$$P(N = n) = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots. \tag{4}$$

In particular, $P(N = 0) = 1 - \rho$. This is the probability that the queue is empty in steady state. Equivalently, it is the fraction of time that the queue is empty in the long run. This makes intuitive sense because work comes into the queue at rate λ and the server can complete work at rate $\mu > \lambda$. Hence, the fraction of time that the server is busy should be $\lambda/\mu = \rho$, and the fraction of time that the server is idle (because the queue is empty) should be $1 - \rho$.

We now compute the mean number in queue from (4). The most convenient way to do this is using generating functions. We have

$$G_N(z) = E[z^N] = \sum_{n=0}^{\infty}(1-\rho)\rho^n z^n = \frac{1-\rho}{1-\rho z},$$

provided $|z| < 1/\rho$. From this, we obtain

$$E[N] = G'_N(1) = \frac{\rho(1-\rho)}{(1-\rho z)^2} \mid |_{z=1} = \frac{\rho}{1-\rho}. \tag{5}$$

Observe that the mean queue length increases to infinity as ρ increases to 1, i.e., as the load on the queue increases to its critical value. From a practical point of view, this effect can be quite dramatic. The mean queue length at a load of $\rho = 0.9$ is 9, while at a load of $\rho = 0.99$, it is 99. Hence, in a nearly criticality loaded queue, a small increase in the service rate can achieve a huge improvement in performance!

The queue length is a measure of performance seen from the viewpoint of the system operator. The expected waiting time (until the beginning of service), or the expected response time or sojourn time (until the customer departs) might be more relevant from the point of view of the customer. We now compute this quantity.

Consider a typical customer entering the system in steady state. We will show later that the random queue length, N, seen by a typical customer upon arrival (in the case of Poisson arrivals) is the same as the steady state queue length, i.e., it has the invariant queue length distribution given in (4). Now suppose this newly arriving customer sees n customers ahead of him. He needs to wait for all of them to complete service, and for his own service to complete, before he can depart the system. Hence, his response time W is given by

$$W = \tilde{S}_1 + S_2 + S_3 + \ldots + S_n + S_{n+1}. \tag{6}$$

Here, S_{n+1} is his own service time, $S_2, \ldots, S_n$ are the service times of the customers waiting in queue, and $\tilde{S}_1$ is the *residual* service time of the customer currently in service. Clearly, $S_2, \ldots, S_{n+1}$ are iid Exp(μ) random variables, and independent of $\tilde{S}_1$. Moreover, by the memoryless property of the exponential distribution, $\tilde{S}_1$ also has an Exp(μ) distribution. Hence, we have by (6) that

$$E\Big[e^{\theta W} \Bigm| N = n\Big] = \Big(E[e^{\theta S_1}]\Big)^{n+1} = \begin{cases} (\frac{\mu}{\mu-\theta})^{n+1}, & \text{if } \theta < \mu, \\ +\infty, & \text{if } \theta \geq \mu. \end{cases}$$

Taking the expectation of this quantity with respect to the random variable N, whose distribution is given by (4), we get

$$E\Big[e^{\theta W}\Big] = \sum_{n=0}^{\infty} \frac{1-\rho}{\rho}\Big(\frac{\mu\rho}{\mu-\theta}\Big)^{n+1} = \frac{(1-\rho)\lambda}{\rho(\mu-\theta)}\frac{\mu-\theta}{\mu-\lambda-\theta} = \frac{\mu-\lambda}{\mu-\lambda-\theta},$$

if $\theta < \mu-\lambda$, and $E[e^{\theta W}] = \infty$ if $\theta \geq \mu-\lambda$. Comparing this with the moment generating function of an exponential distribution, we see that the response time W has an exponential distribution with parameter $\mu-\lambda$. Hence, the mean response time is

$$E[W] = \frac{1}{\mu-\lambda}. \tag{7}$$

Comparing this with (5), we see that

$$E[N] = \lambda E[W]. \tag{8}$$

This equality is known as *Little's law.* It holds in great generality and not just for the $M/M/1$ queue.

Little's law: Let A_t be any arrival process, i.e., A_t is a counting process counting the number of arrivals up to time t. The service times can have arbitrary distribution, and be dependent on one another or on the arrival times. The service discipline can be arbitrary (e.g., processor sharing or LCFS), and there can be one or more serves. Let N_t denote the queue length (number of customers in the system at time t) and let W_n denote the sojourn time of the n^{th} arriving customer. We assume only that the following limits exist:

$$\lim_{t\to\infty} \frac{A_t}{t} = \lambda, \quad \lim_{t\to\infty} \frac{1}{t}\int_0^t N_s ds = \overline{N}, \quad \lim_{n\to\infty} \sum_{k=0}^{n} W_k = \overline{W}.$$

Then, $\overline{N} = \lambda\overline{W}$. Note that this equality holds for every sample path of the stochastic process for which these limits exist. If the stochastic process is ergodic, then these limits exist (almost surely) and are equal to the corresponding expectations, i.e, $\overline{N} = E[N]$ and $\overline{W} = E[W]$.. We haven't defined ergodicity but, intuitively, it is a notion of the stochastic process being well-behaved in the sense of satisfying the law of large numbers - sample means converge to the true mean.

A proof of this result can be found in many textbooks; see, e.g., Jean Walrand, *An Introduction to Queueing Networks.* The intuition behind it is as follows. Suppose each customer entering the system pays \$1 per unit of time spent in the system. So, customers pay $\$E[W]$ on average. Since they enter the system at rate λ, the rate at which revenue is earned is $\lambda E[W]$. On the other hand, the system earns $\$n$ per unit time when there are n customers in the system. Hence, the average rate at which revenue is earned is $E[N]$. As these are two ways of calculating the same quantity, they must be the same.

3 The PASTA property

In the course of computing the mean waiting time, we assumed that a typical arrival sees the queue in its invariant distribution. At first, this may seem an innocuous assumption, but it is by no means obvious, as the following counterexample shows.

Consider a single-server queue into which customers arrive spaced 2 units apart, deterministically, and suppose also that the arrival time of the first customer is random, uniformly distributed on $[0, 2]$. Suppose that customers need service for iid random times uniformly distributed between 0 and 2. The queue length is not a continuous-time Markov process but it is nevertheless clear how it evolves. If the queue is started empty, then the queue length becomes 1 when the first customer arrives, stays at 1 for a random length of time of unit mean, and uniformly distributed between 0 and 2. Thus, the queue length is guaranteed to return to zero before the next arrival. This cycle repeats itself indefinitely. Now, if we consider any fixed time, then at this time the queue is equally likely to either be empty or to have exactly 1 customer in it. Thus, the steady state queue length distribution assigns probability 0.5 each to states 0 and 1, and probability 0 to all other states. On the other hand, new arrivals always see an empty queue! Thus, the queue length distribution seen by arrivals is not the same as the steady-state queue length distribution.

However, such a situation cannot arise if the arrivals form a homogeneous Poisson process, as in the $M/M/1$ queue or, indeed, in any queue with Poisson arrivals, irrespective of the service time distribution, the service discipline, or the number of servers. This fact is known as the PASTA property (Poisson arrivals see time averages).

More formally, suppose that X_t, $t \geq 0$ is a Markov process on some state space S, having unique invariant distribution π. Suppose that N_t, $t \geq 0$ is a Poisson process satisfying the property that $(N_u - N_t, u > t)$ is independent of X_s, $s \leq t$. If we think of X_t as the queue length process and N_t as the arrival process, then this says that the arrival process is Poisson and that future arrivals are independent of the past of the queue length process. (Clearly, the past of the arrival process won't be independent of the past or current queue length.) We now have:

$$\lim_{\delta \downarrow 0} P(X_t = x | N_{t+\delta} - N_t = 1) = P(X_t = x). \tag{9}$$

In words, this says that knowing that there was an arrival just after time t does not tell us anything about the queue length distribution. In particular, if the chain is in stationarity, $P(X_t = x) = \pi_x$ and (9) says that the state of the Markov process X_t just before an arrival is the same as its invariant distribution. In other words, arrivals see the invariant distribution.

The proof of (9) is easy and follows from Bayes' theorem. Observe that

$$\begin{aligned} P(X_t = x | N_{t+\delta} - N_t = 1) &= \frac{P(X_t = x, N_{t+\delta} - N_t = 1)}{P(N_{t+\delta} - N_t = 1)} \\ &= \frac{P(X_t = x) P(N_{t+\delta} - N_t = 1 | X_t = x)}{P(N_{t+\delta} - N_t = 1)} \\ &= \frac{P(X_t = x) P(N_{t+\delta} - N_t = 1)}{P(N_{t+\delta} - N_t = 1)} \\ &= P(X_t = x). \end{aligned}$$

We have used the fact that future arrivals are independent of the current state in order to obtain the third equality above.

More generally, for any Markovian queueing system, we can compute the queue length distribution seen at arrival or departure epochs. We now illustrate this with an example.

Example: $M/M/1$ queue with balking
This is a queue with Poisson arrivals at rate λ, exponentially distributed job sizes with mean $1/\mu$ and a single server, as for an $M/M/1$ queue. The difference is that not all arriving customers join the queue. A customer who, upon arrival, sees n customers ahead of him, joins the queue with probability $\frac{1}{n+1}$, and departs without service with the residual probability $\frac{n}{n+1}$. Thus, a customer who sees an empty queue is sure to join, but customers become less and less likely to join as the queue gets longer.

The number of customers in the system evolves as a continuous time Markov process, and it can be shown that this process is reversible (being a birth-death chain), and has invariant distribution given by

$$\pi_n = \frac{\rho^n}{n!} e^{-\rho}, \quad n = 0, 1, 2, \ldots$$

We want to compute the number of customers in queue as seen by a typical arrival who chooses to join the queue (excluding the joining customer). We can compute this as follows:

$$\begin{aligned} P(X_A = n) &= P(X_t = n | \text{arrival in } [t, t+dt) \\ &= \frac{P(X_t = n \text{ and arrival in } [t, t+dt))}{P(\text{arrival in } [t, t+dt))} \\ &= \frac{\pi_n \lambda dt \frac{1}{n+1}}{\sum_{k=0}^{\infty} \pi_k \lambda dt \frac{1}{k+1}} \\ &= \frac{\rho^n e^{-\rho}/((n+1)!)}{\sum_{k=0}^{\infty} \rho^k e^{-\rho}/((k+1)!)}. \end{aligned}$$

We have used Bayes' theorem to get the second equality above. Now,

$$\sum_{k=0}^{\infty} \frac{\rho^k}{(k+1)!} = \frac{1}{\rho} \sum_{k=1}^{\infty} \frac{\rho^k}{k!} = \frac{e^{\rho} - 1}{\rho}.$$

Substituting this in the above, we get

$$P(X_A = n) = \frac{\rho^{n+1}}{(n+1)!(e^{\rho} - 1)}.$$

Note that this is different from π_n, demonstrating once again that, if the arrival process isn't Poisson, then a typical arrival need not see the system in its stationary distribution.

We can similarly compute the distribution seen by departures in stationarity. The probability that a departing customer leaves behind n customers in queue is given by

$$\begin{aligned}
P(X_D = n) &= P(X_t = n+1 | \text{departure in } [t, t+dt) \\
&= \frac{P(X_t = n+1 \text{ and departure in } [t, t+dt))}{P(\text{departure in } [t, t+dt))} \\
&= \frac{\pi_{n+1}\mu dt}{\sum_{k=1}^{\infty} \pi_k \mu dt} \\
&= \frac{\rho^{n+1} e^{-\rho}/((n+1)!)}{1 - \pi_0} \\
&= \frac{\rho^{n+1}}{(n+1)!(e^{\rho} - 1)}.
\end{aligned}$$

Note that this is the same as $P(X_A = n)$. This is true of any queue where arrivals and departures happen singly rather than in batches because you can match up arriving and departing customers in such a way that, for every customer who enters a queue with n customers, there is a departure that leaves behind n customers in the queue.

4 Departure process from an $M/M/1$ queue

If an $M/M/1$ queue has arrival rate λ which is strictly smaller than the service rate μ, then the queue is stable and doesn't build up an infinite backlog. Consequently, all arrrivals eventually depart, and the mean departure rate is the same as the arrival rate, λ. But what can we say about the departure process? The answer is given by the following theorem, which is called Burke's theorem.

Theorem 1 *The departure process from an $M/M/1$ queue with arrival rate λ strictly smaller than the service rate μ, is also a Poisson process of rate λ in equilibrium. Moreover, the past of the departure process is independent of the current state.*

Proof. The proof is a pretty straightforward application of reversibility. Note that the queue length process X_t is reversible in equilibrium since it is a birth-death Markov process. Arrivals correspond to upward jumps of X_t and departures to downward jumps. Now, the time reveral of X_t has the same probabilistic properties as X_t itself, but arrivals in forward time correspond to departures in reversed time and vice versa. Hence, the departure process for the forward time process should have the same probability law as the arrival process for its time reversal. But this is a Poisson process because the time reversal is indistinguishable from the original process.

Likewise, the past of the departure process in forward time corresponds to the future of the arrival process in reversed time, which we know to be independent of the current state since this arrival process is Poisson. □

5 $M/M/c$ and $M/M/\infty$ queues

Recall that the first M says that the arrival process is Poisson, the second M that the service times are exponential, and the number refers to the number of servers, which is either a given number c or infinite. While real queues rarely have infinitely many servers, this can often be a good approximation when the number of servers is large! It was first introduced to model telephone exchanges, where the number of available lines may run into the tens of thousands. In systems with a large number of servers, it can often be a matter of judgement whether to model it as an $M/M/c$ or $M/M/\infty$ queue.

Customers are held in a common queue and sent to the first server that becomes available. If one or more are already free when they enter the system, they go to one of these servers at random. As the service time distribution is the same at every server, it doesn't matter how ties are broken.

Let X_t denote the number of customers in the system at time t. If $X_t \leq c$, then X_t servers are busy serving these customers (in parallel), and the remaining $c - X_t$ servers are idle. If $X_t > c$, then all servers are busy, and $X_t - c$ customers are waiting in queue. The possible transitions of this system are that a new customer could arrive, or an existing customer could complete service and leave. The probability of two or more customers leaving simultaneously is zero, as we shall see. The probability that a new customer arrives in the interval $[t, t+\delta]$ is $\lambda\delta + o(\delta)$, independent of the past, since the arrival process is Poisson with rate λ. The probability that an existing customer departs in the interval $[t, t+\delta]$ is the probability that one of the $\min\{X_t, c\}$ customers currently being served completes its service in that interval. This probability is $\mu\delta + o(\delta)$ for each of these customers, independent of the past (because service requirements are exponentially distributed, with parameter

μ), and of one another (since service requirements of different customers are mutually independent). Hence, the probability that exactly one of them departs is

$$1 - (1 - \mu\delta + o(\delta))^{\min\{X_t, c\}} = \min\{X_t, c\}\mu\delta + o(\delta),$$

while the probability that two or more depart is $o(\delta)$. Thus, the departure probability in $[t, t + \delta]$ depends only on the current state X_t, and on no further details of the past. In other words, $X_t, t \geq 0$ is a continuous time Markov process, with transition rates:

$$q_{i,i+1} = \lambda, \quad q_{i,i-1} = \begin{cases} \mu i, & 0 \leq i \leq c, \\ \mu c, & i > c \end{cases} \quad q_{i,i} = -\lambda - \mu \min\{i, c\}.$$

Note that all the statements in the paragraph above also hold if c is replaced by ∞.

Let us now compute the invariant distributions of these queues, starting with the $M/M/\infty$ queue. As the number in system is a birth-death Markov process, it is reversible if it has an invariant distribution, and the invariant distribution π satisfies the detailed balanced equations, which are as follows:

$$\lambda\pi_i = \mu(i + 1)\pi_{i+1}, \quad i \geq 0.$$

Let $\rho = \lambda/\mu$. Then, we can rewrite the above as $\pi_{i+1} = \frac{\rho}{i+1}\pi_i$, which implies that $\pi_i = \frac{\rho^i}{i!}\pi_0$. This can be normalised to be a probability distribution, for any ρ, if we choose $\pi_0 = e^{-\rho}$. Hence, the invariant distribution of the $M/M/\infty$ queue is given by

$$\pi_i = \frac{\rho^i}{i!}e^{-\rho}, \tag{10}$$

which we recognise as the Poisson distribution with parameter ρ. Note that this queue is stable for any combination of arrival rates and (non-zero) service rates. As there are infinitely many servers, the total service rate exceeds the total arrival rate once the number in system becomes sufficiently large.

The mean of a Poisson(ρ) random variable is ρ. (Verify this for yourself.) Hence, the mean number in an $M/M/\infty$ queue is given by $E[N] = \rho$. We can also see this intuitively, as follows. The number of busy servers at any time is the same as the number of customers in the system as there are always enough servers available. Hence, the mean rate at which servers are doing work is the mean number of busy servers, which is also the mean number of customers in the system. On the other hand, the mean rate at which work is brought into the system is the mean arrival rate, λ, times the mean amount of work brought by each customer, $1/\mu$, which is ρ. This should be the same, in equilibrium, as the mean rate at which work is done. Therefore, the mean number of customers in the system should be ρ.

Next, by Little's law, the mean sojourn time of a customer is $E[W] = E[N]/\lambda = 1/\mu$. This is again intuitive. On average, a customer brings in $1/\mu$ amount of work, and doesn't need to wait as there is always a server available. Hence, his sojourn time is the same as his own service requirement.

Next, we compute the invariant distribution of the $M/M/c$ queue. Again, the queue length process is a birth-death Markov process, and so, if the detailed balance equations have a solution which is a probability distribution, then that is the invariant distribution of the Markov process. The detailed balance equations are

$$\lambda \pi_i = \mu \min\{i+1, c\} \pi_{i+1}, \quad i \geq 0,$$

which have the solution

$$\pi_i = \begin{cases} \frac{\rho^i}{i!} \pi_0, & i < c, \\ \frac{\rho^c}{c!} \left(\frac{\rho}{c}\right)^{i-c} \pi_0, & i \geq c, \end{cases} \tag{11}$$

where $\rho = \lambda/\mu$. We want to choose π_0 so as to make $\sum_{i=0}^{\infty} \pi_i = 1$, i.e.,

$$\frac{1}{\pi_0} = \sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \frac{\rho^c}{c!} \sum_{i=c}^{\infty} \left(\frac{\rho}{c}\right)^{i-c} = \sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \frac{\rho^c}{c!} \sum_{j=0}^{\infty} \left(\frac{\rho}{c}\right)^{j}. \tag{12}$$

There is a non-zero solution for π_0 if and only if the above sum is finite, which happens if and only if $\rho < c$. If $\rho \geq c$, the detailed balance equations have no solution which is a probability distribution. This result is intuitive because the maximum rate at which work can be done in the $M/M/c$ queue is c, and if the rate ρ at which work enters the system is greater than this, then the queue is unstable and the waiting times build up to infinity. If $\rho < c$, then the invariant distribution is given by (11) with π_0 being given by the solution of (12). The solution can't be simplified further. The mean queue size and mean waiting time can easily be computed numerically from the invariant distribution, but there are no closed form expressions for them.

6 $M/G/1$ queues: Pollaczek-Khinchin formula

So far, we have only considered queues with Poisson arrivals and exponentially distributed service requirements. That is because it is only in those cases that the queue length process is Markovian, which makes the analysis of such systems tractable. If either the inter-arrival times, or the service times, or both, weren't exponential, then the future evolution of the queue length would depend not only on the current queue length, but on how much time has elapsed since the last arrival or service completion, as this affects the residual time to the next arrival/ service completion. (In the exponential case, there is no such dependence because of the memoryless property of the exponential distribution.) As a result, the analysis of such systems is far from straightforward, and we will not deal with them in this course, with a few exceptions.

The first exception is the $M/G/1$ queue or, more precisely, the $M/G/1-FIFO$ queue, where we make it explicit that the service discipline is first-in-first-out or first-come-first-served. The G here stands for general, which means that we allow arbitrary service time distributions. We will still assume, though, that the service requirements of different customers are iid. As this system is more complicated, we will not aim to derive its invariant distribution, but content ourselves with finding the mean queue length and mean waiting time. The formulas for these quantities are called the Pollaczek-Khinchin (henceforth abbreviated as P-K) formulas.

In the following, we consider an $M/G/1$ queue where arrivals occur according to a Poisson process of rate λ, service times are iid with general distribution, there is a single server operating a FIFO policy, and there is infinite waiting room. We denote the mean service time by $1/\mu$ and define $\rho = \lambda/\mu$ to be consistent with the notation used so far for the $M/M/1$ queue. We shall assume that $\rho < 1$, so that the queue is stable and does not blow up to infinity. We now aim to find an expression for the mean sojourn time in this queue.

Define the following quantities:

S_i : service time required by customer i

R_i : residual service requirement of customer in service when i arrives

N_i : number of customers in system when customer i arrives

W_i : sojourn time of customer i in system

We take R_i to be zero if customer i enters an empty queue. We have

$$W_i = S_i + \Big(R_i + \sum_{j=1}^{N_i - 1} S_{i-j} \Big) \mathbf{1}(N_i \geq 1).$$

If $N_i = 1$, then the sum above is empty, and we take it to be zero. Taking expectations in the above equation, we have

$$\begin{aligned} E[W_i] &= E[S_i] + E\Big(E\Big[R_i \mathbf{1}(N_i \geq 1) \;\Big|\; N_i \Big] \Big) E\Big(E\Big[\sum_{j=1}^{N_i - 1} S_{i-j} \;\Big|\; N_i \Big] \Big) \\ &= E[S_i] + P(N_i \geq 1) E[R_i | N_i \geq 1] + E[S_1] E[(N_i - 1)\mathbf{1}(N_i \geq 1)]. \end{aligned}$$

We have used the linearity of expectation and the independence of the future service times S_{i-j} from N_i to obtain the second equality above. Now, $E[N_i \mathbf{1}(N_i \geq 1)] = E[N_i]$, so we can rewrite the above as

$$E[W] = E[S] + P(N \geq 1) E[R | N \geq 1] + E[S](E[N] - P(N \geq 1)). \tag{13}$$

We have dropped subscripts for notational convenience, keeping in mind that we are referring to quantitites in steady state, and hence that these don't depend on the index i.

Now, N is a random variable denoting the number of customers in system seen by the typical arrival. The arrival process into the $M/G/1$ queue is a Poisson process by assumption and so, by the PASTA property, N has the same distribution as the number of customers in system in stationarity (or in steady state or equilibrium - we use these terms interchangeably). Hence, applying Little's law, we have $E[N] = \lambda E[W]$. Substituting this in (13), and noting that $\lambda E[S] = \lambda / \mu = \rho$, we get

$$(1 - \rho) E[W] = P(N \geq 1) E[R | N \geq 1] + E[S](1 - P(N \geq 1)). \tag{14}$$

It remains to compute $P(N \geq 1)$ and $E[R | N \geq 1]$.

First, we note that customers enter the queue at rate λ, each bringing a random amount of work with them, with mean $1/\mu$. Thus, the average rate at which work comes into the system is $\lambda / \mu = \rho$. This must be the same as the average rate at which work is completed because the system is stable, meaning that no huge backlog of work builds up. Now, the server works at unit rate whenever the queue is non-empty, which means that the long-run fraction of time that the queue is non-empty has to be ρ. In other words,

$P(N \geq 1) = \rho$. (If that argument is not persuasive enough, you can prove this along the same lines as the proof of Little's law. Assume that each customer is charged \$1 per unit time whenever it is at the head of the queue and is being served. Thus, a customer with job size S ends up paying \$ S. But customers enter the system at rate λ, so the rate at which money is paid into the system is $\lambda E[S] = \rho$ dollars per unit time. On the other hand, the server earns \$ 1 per unit time whenever it is working, i.e., whenever the queue is non-empty. As the rate at which the server earns money has to match the rate at which customers are paying it in, the fraction of time the server is busy has to be ρ.) Thus, we can rewrite (14) as

$$(1-\rho)E[W] = \rho E[R|N \geq 1] + (1-\rho)E[S]. \tag{15}$$

Next, we compute $E[R|N \geq 1]$, the mean residual service time of the customer at the head of the queue at the arrival instant of a typical customer, conditional on the queue not being empty at this time. Let us look at a very long interval of the time axis, so the time interval $[0, T]$ for some very large T. The fraction of this interval occupied by the service of customers whose total service requirement is in $(x, x+dx)$ is $\lambda x f_S(x)dx$ where $f_S(x)$ is the density of the service time distribution evaluated at x. (We are making somewhat loose statements involving the infinitesimals dx, but these can be made precise.) To see this, note that approximately λT customers arrive during this time, of which a fraction $f_S(x)dx$ have service requirement in $(x, x+dx)$, and each of them occupies an interval of the time axis of size x when it gets to the head of the queue. The fraction of the interval $[0, T]$ occupied by the service of some customer (i.e., the fraction that the system is non-empty) is then given by $\int_0^\infty \lambda x f_S(x)dx = \lambda E[S]$.

Now, the typical arrival conditioned to enter a non-empty system is equally likely to arrive any time during a busy period. (Though this might seem obvious, it is actually a subtle point and relies on the arrival process being Poisson.) Hence, the probability density for the typical arrival to enter the system during the service of a customer whose total job size is x is given by $xf_S(x)/E[S]$. Moreover, the arrival occurs uniformly during the service period of this customer and hence sees a residual service time of $\frac{x}{2}$, on average. Thus, the mean residual service time seen by a typical arrival who enters a non-empty system is

$$E[R|N \geq 1] = \frac{1}{E[S]} \int_0^\infty \frac{x}{2} x f_S(x)dx = \frac{E[S^2]}{2E[S]}. \tag{16}$$

Substituting this in (15), we get

$$E[W] = E[S] + \frac{\rho}{1-\rho} \frac{E[S^2]}{2E[S]} = E[S] + \frac{\lambda E[S^2]}{2(1-\rho)}. \tag{17}$$

We have used the fact that $\rho = \lambda/\mu = \lambda E[S]$ to obtain the last equality. The formula for the mean sojourn time above consists of two terms, the first of which is the customer's own service time, and the second is the mean waiting time until reaching the head of the queue. Applying Little's law to the above formula, we also get

$$E[N] = \lambda E[W] = \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)}. \tag{18}$$

The formulas (17,18) are referred to as the Pollaczek-Khinchin (PK) formulas for the mean of the queue size and sojourn time. There are also PK formulas for the moment generating functions of these quantities, but we will not be studying them in this course. The goal of this section was to introduce you to some of the methods that can be used to study non-Markovian queues. We will see another such example in the next section. However, but for these two exceptions, we will be restricting ourselves to Markovian queueing models in this course.

7 $M/G/\infty$ queue

This is a model of a queueing system with Poisson arrivals, general iid service times, and infinitely many servers. Thus, an arriving job stays in the system for a period equal to its own service time and then departs. We denote the arrival rate by λ and the mean service time by $1/\mu$. The system is stable for all λ and μ.

Note that the number of customers in the system is not a Markov process since the past of this process contains information about when customers arrived, and hence potentially about how much longer they will remain in the queue. This is easiest to see if all the job sizes are deterministically equal to $1/\mu$. Thus, the methods we learnt for analysing Markov processes are not relevant. Instead, we shall use techniques from the theory of point processes.

Definition A stochastic process on $\mathbb{R}^n$ is called a point process if, with each (measurable) set $A \subseteq \mathbb{R}^n$, it associates a non-negative discrete random variable N_A, and that the collection of random variables $\{N_A, A \subseteq \mathbb{R}\}$ satisfy the following properties:

1. $N_\emptyset = 0$.
2. If $A_1, A_2, A_3, \dots$ are disjoint, and $A = \cup_{j=1}^{\infty} A_j$, then $N_A = \sum_{j=1}^{\infty} N_{A_j}$.

Observe that we already saw an example of a point process on $\mathbb{R}$, namely the Poisson process. If $N.$ satisfies properties 1 and 2 above, it is called a measure. If, in addition, it takes values in the whole numbers, it is called a counting measure. Thus, a point process can be thought of as a random counting measure. More concretely, you can think of it as a collection of randomly placed points, with N_A denoting the random number of points in a set A. Note that the points need not be placed independently.

Definition A point process on $\mathbb{R}^n$ is said to be a Poisson process with intensity function $\lambda(\cdot)$ if the following statements hold:

1. If A and B are disjoint subsets of $\mathbb{R}^n$, then N_A and N_B are independent random variables.
2. For any set A, N_A has the Poisson distribution with mean $\int_A \lambda(\mathbf{x})d\mathbf{x}$. Here $d\mathbf{x} = dx_1 dx_2 \cdots dx_n$ denotes the infinitesimal volume element in $\mathbb{R}^n$.

The Poisson process is said to be homogeneous if the intensity function is a constant, i.e., $\lambda(\mathbf{x}) = \lambda$ for all $\mathbf{x} \in \mathbb{R}^n$. Note that, as for Poisson processes on the real line, the second property above can be rephrased as saying that the probability of having a point in a small neighbourhood around $\mathbf{x}$ is approximately $\lambda(\mathbf{x})$ times the volume of the neighbourhood, and that this is independent of the location of all other points.

Now, consider such a Poisson point process on $\mathbb{R} \times [0, \infty)$ with intensity function $\lambda(x, y) = \lambda f_S(y)$, where λ denotes the arrival rate into the $M/G/\infty$ queue and f_S denotes the density of the service time distribution. We interpret a point at (x, y) as denoting the arrival of a customer at time x with job size y. Note that the intensity has been chosen to correctly model the arrival rate and job size distribution.

Consider a realisation of this Poisson point process. Say we want to know X_0, the number of jobs in the system at time 0. A job that entered the system at time $-t$ will still be in the system at time zero if and only if its job size is smaller than t. Thus, the set of all jobs that are in the system at time zero can be identified with the set of points that lie in the triangular region $A = \{(x, y) : x \leq 0, y > -x\}$. But, by the property of a Poisson process, the number of such points is a Poisson random variable with mean

$$\begin{aligned} \eta &= \int_A \lambda(x, y) dx dy = \int_{y=0}^{\infty} \int_{x=-y}^{0} \lambda f_S(y) dx dy \\ &= \lambda \int_{y=0}^{\infty} y f_S(y) dy = \lambda E[S]. \end{aligned}$$

We have thus shown that the number of jobs in the system at time zero has a Poisson distribution with mean $\rho = \lambda/\mu$. We defined the system on the time interval $(infty, \infty)$, so that by time zero, or by any finite time, the system has reached equilibrium. Hence, the invariant distribution of the number of jobs in the system is Poisson(ρ).

Recall that this is the same as the invariant distribution for an $M/M/\infty$ queue with the same arrival rate and mean service time. Thus, the invariant distribution depends on the service time distribution only through its mean. We call this the *insensitivity property* as the invariant distribution is insensitive to the service time distribution.

8 Some other insensitive queues

Besides the $M/G/\infty$ queue, the $M/G/1 - LIFO$ and $M/G/1 - PS$ queues (with servers employing the last-in-first-out and processor sharing service disciplines respectively) also exhibit the insensitivity property. In their case, the invariant distribution is geometric, just like the $M/M/1$ queue with the same arrival rate and mean service time. We will not prove this fact, but will make use of it, so it is important to remember it.