4 $M/G/1$ queue

In the $M/G/1$ queue customers arrive according to a Poisson process with rate λ and they are treated in order of arrival. The service times are independent and identically distributed with distribution function $F_B(\cdot)$ and density $f_B(\cdot)$. For stability we have to require that the occupation rate

$$\rho = \lambda E(B) \tag{1}$$

is less than one. In this chapter we will derive the limiting or equilibrium distribution of the number of customers in the system and the distributions of the sojourn time and the waiting time. It is further shown how the means of these quantities can be obtained by using the mean value approach.

4.1 Which limiting distribution?

The state of the $M/G/1$ queue can be described by the pair (n, x) where n denotes the number of customers in the system and x the service time already received by the customer in service. We thus need a two-dimensional state description. The first dimension is still discrete, but the other one is continuous and this essentially complicates the analysis. However, if we look at the system just after departures, then the state description can be simplified to n only, because $x = 0$ for the new customer (if any) in service. Denote by L_k^d the number of customers left behind by the kth departing customer. In the next section we will determine the limiting distribution

$$d_n = \lim_{k\to\infty} P(L_k^d = n).$$

The probability d_n can be interpreted as the fraction of customers that leaves behind n customers. But in fact we are more interested in the limiting distribution p_n defined as

$$p_n = \lim_{t\to\infty} P(L(t) = n),$$

where $L(t)$ is the number of customers in the system at time t. The probability p_n can be interpreted as the fraction of time there are n customers in the system. From this distribution we can compute, e.g., the mean number of customers in the system. Another perhaps even more important distribution is the limiting distribution of the number of customers in the system seen by an arriving customer, i.e.,

$$a_n = \lim_{k\to\infty} P(L_k^a = n),$$

where L_k^a is the number of customers in the system just before the kth arriving customer. From this distribution we can compute, e.g., the distribution of the sojourn time. What is the relation between these three distributions? It appears that they all are the same.

Of course, from the PASTA property we already know that $a_n = p_n$ for all n. We will now explain why also $a_n = d_n$ for all n. Taking the state of the system as the number of customers therein, the changes in state are of a nearest-neighbour type: if the system is in state n, then an arrival leads to a transition from n to $n+1$ and a departure from n to $n-1$. Hence, in equilibrium, the number of transitions per unit time from state n to $n+1$ will be the same as the number of transitions per unit time from $n+1$ to n. The former transitions correspond to arrivals finding n customers in the system, the frequency of which is equal to the total number of arrivals per unit time, λ, multiplied with the fraction of customers finding n customers in the system, a_n. The latter transitions correspond to departures leaving behind n customers. The frequency of these transitions is equal to the total number of departures per unit time, λ, multiplied with the fraction of customers leaving behind n customers, d_n. Equating both frequencies yields $a_n = d_n$. Note that this equality is valid for any system where customers arrive and leave one by one. Thus it also holds for, e.g., the G/G/c queue.

Summarizing, for the $M/G/1$ queue, arrivals, departures and outside observers all see the same distribution of number of customers in the system, i.e., for all n,

$$a_n = d_n = p_n.$$

4.2 Departure distribution

In this section we will determine the distribution of the number of customers left behind by a departing customer when the system is in equilibrium.

Denote by L_k^d the number of customers left behind by the kth departing customer. We first derive an equation relating the random variable L_{k+1}^d to L_k^d. The number of customers left behind by the $k+1$th customer is clearly equal to the number of customers present when the kth customer departed minus one (since the $k+1$th customer departs himself) plus the number of customers that arrives during his service time. This last number is denoted by the random variable A_{k+1}. Thus we have

$$L_{k+1}^d = L_k^d - 1 + A_{k+1},$$

which is valid if $L_k^d > 0$. In the special case $L_k^d = 0$, it is readily seen that

$$L_{k+1}^d = A_{k+1}.$$

From the two equations above it is immediately clear that the sequence $\{L_k^d\}_{k=0}^{\infty}$ forms a Markov chain. This Markov chain is usually called the *imbedded Markov chain*, since we look at imbedded points on the time axis, i.e., at departure instants.

We now specify the transition probabilities

$$p_{i,j} = P(L_{k+1}^d = j | L_k^d = i).$$

Clearly $p_{i,j} = 0$ for all $j < i - 1$ and $p_{i,j}$ for $j \geq i - 1$ gives the probability that exactly $j - i + 1$ customers arrived during the service time of the $k+1$th customer. This holds for $i > 0$. In state 0 the kth customer leaves behind an empty system and then $p_{0,j}$ gives the probability that during the service time of the $k+1$th customer exactly j customers arrived. Hence the matrix P of transition probabilities takes the form

$$P = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\ 0 & 0 & \alpha_0 & \alpha_1 & \cdots \\ 0 & 0 & 0 & \alpha_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where α_n denotes the probability that during a service time exactly n customers arrive. To calculate α_n we note that given the duration of the service time, t say, the number of customers that arrive during this service time is Poisson distributed with parameter λt. Hence, we have

$$\alpha_n = \int_{t=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_B(t) dt. \tag{2}$$

The transition probability diagram is shown in figure 1.

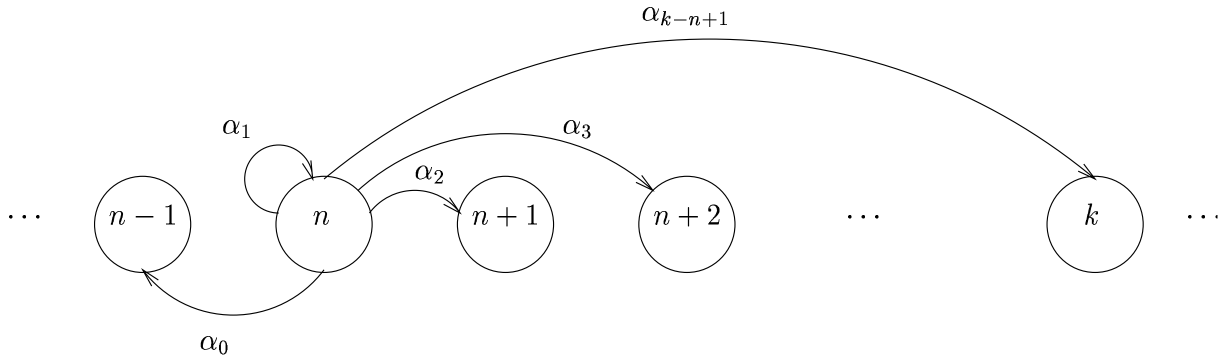

Figure 1: Transition probability diagram for the $M/G/1$ imbedded Markov chain

This completes the specification of the imbedded Markov chain. We now wish to determine its limiting distribution. Denote the limiting distribution of L_k^d by $\{d_n\}_{n=0}^{\infty}$ and the limiting random variable by L^d. So

$$d_n = P(L^d = n) = \lim_{k \to \infty} P(L_k^d = n).$$

The limiting probabilities d_n, which we know are equal to p_n, satisfy the equilibrium equations

$$\begin{aligned} d_n &= d_{n+1}\alpha_0 + d_n\alpha_1 + \cdots + d_1\alpha_n + d_0\alpha_n \\ &= \sum_{k=0}^{n} d_{n+1-k}\alpha_k + d_0\alpha_n, \qquad n = 0, 1, \ldots \end{aligned} \tag{3}$$

To solve the equilibrium equations we will use the generating function approach. Let us introduce the probability generating functions

$$P_{L^d}(z) = \sum_{n=0}^{\infty} d_n z^n, \qquad P_A(z) = \sum_{n=0}^{\infty} \alpha_n z^n,$$

which are defined for all $z \leq 1$. Multiplying (3) by z^n and summing over all n leads to

$$\begin{aligned} P_{L^d}(z) &= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} d_{n+1-k}\alpha_k + d_0\alpha_n \right) z^n \\ &= z^{-1} \sum_{n=0}^{\infty} \sum_{k=0}^{n} d_{n+1-k} z^{n+1-k} \alpha_k z^k + \sum_{n=0}^{\infty} d_0 \alpha_n z^n \\ &= z^{-1} \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} d_{n+1-k} z^{n+1-k} \alpha_k z^k + d_0 P_A(z) \\ &= z^{-1} \sum_{k=0}^{\infty} \alpha_k z^k \sum_{n=k}^{\infty} d_{n+1-k} z^{n+1-k} + d_0 P_A(z) \\ &= z^{-1} P_A(z)(P_{L^d}(z) - d_0) + d_0 P_A(z). \end{aligned}$$

Hence we find

$$P_{L^d}(z) = \frac{d_0 P_A(z)(1 - z^{-1})}{1 - z^{-1} P_A(z)} \, .$$

To determine the probability d_0 we note that d_0 is equal to p_0, which is the fraction of time the system is empty. Hence $d_0 = p_0 = 1 - \rho$ (alternatively, d_0 follows from the requirement $P_{L^d}(1) = 1$). So, by multiplying numerator and denominator by $-z$ we obtain

$$P_{L^d}(z) = \frac{(1-\rho) P_A(z)(1-z)}{P_A(z) - z} \, . \tag{4}$$

By using (2), the generating function $P_A(z)$ can be rewritten as

$$\begin{aligned} P_A(z) &= \sum_{n=0}^{\infty} \int_{t=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_B(t) dt z^n \\ &= \int_{t=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\lambda t z)^n}{n!} e^{-\lambda t} f_B(t) dt \\ &= \int_{t=0}^{\infty} \sum_{n=0}^{\infty} e^{-(\lambda - \lambda z)t} f_B(t) dt \\ &= \widetilde{B}(\lambda - \lambda z) \end{aligned} \tag{5}$$

Substitution of (5) into (4) finally yields

$$P_{L^d}(z) = \frac{(1-\rho)\widetilde{B}(\lambda - \lambda z)(1-z)}{\widetilde{B}(\lambda - \lambda z) - z} \, . \tag{6}$$

This formula is one form of the *Pollaczek-Khinchin formula*. In the following sections we will derive similar formulas for the sojourn time and waiting time. By differentiating formula (6) we can determine the moments of the queue length. To find its distribution, however, we have to invert formula (6), which usually is very difficult. In the special case that $\widetilde{B}(s)$ is a quotient of polynomials in s, i.e., *a rational function*, then in principle the right-hand side of (6) can be decomposed into partial fractions, the inverse transform of which can be easily determined. The service time has a rational transform for, e.g., mixtures of Erlang distributions or Hyperexponential distributions. The inversion of (6) is demonstrated below for exponential and Erlang-2 service times.

Example 4.1 *($M/M/1$)*
Suppose the service time is exponentially distributed with mean $1/\mu$. Then

$$\widetilde{B}(s) = \frac{\mu}{\mu + s}\,.$$

Thus

$$P_{L^d}(z) = \frac{(1-\rho)\frac{\mu}{\mu+\lambda-\lambda z}(1-z)}{\frac{\mu}{\mu+\lambda-\lambda z} - z} = \frac{(1-\rho)\mu(1-z)}{\mu - z(\mu+\lambda-\lambda z)} = \frac{(1-\rho)\mu(1-z)}{(\mu-\lambda z)(1-z)} = \frac{1-\rho}{1-\rho z}\,.$$

Hence

$$d_n = p_n = (1-\rho)\rho^n, \qquad n = 0, 1, 2, \ldots$$

Example 4.2 *($M/E_2/1$)*
Suppose the service time is Erlang-2 distributed with mean $2/\mu$. Then

$$\widetilde{B}(s) = \left(\frac{\mu}{\mu + s}\right)^2,$$

so

$$\begin{aligned} P_{L^d}(z) &= \frac{(1-\rho)\left(\frac{\mu}{\mu+\lambda-\lambda z}\right)^2(1-z)}{\left(\frac{\mu}{\mu+\lambda-\lambda z}\right)^2 - z} \\ &= \frac{(1-\rho)\mu^2(1-z)}{\mu^2 - z(\mu+\lambda-\lambda z)^2} \\ &= \frac{(1-\rho)(1-z)}{1 - z(1+\rho(1-z)/2)^2} \\ &= \frac{1-\rho}{1-\rho z - \rho^2 z(1-z)/4}\,. \end{aligned}$$

For $\rho = 1/3$ we then find

$$\begin{aligned} P_{L^d}(z) &= \frac{2/3}{1 - z/3 - z(1-z)/36} = \frac{24}{36 - 13z + z^2} \\ &= \frac{24}{(4-z)(9-z)} = \frac{24/5}{4-z} - \frac{24/5}{9-z} = \frac{6/5}{1-z/4} - \frac{8/15}{1-z/9}\,. \end{aligned}$$

Hence,

$$d_n = p_n = \frac{6}{5}\left(\frac{1}{4}\right)^n - \frac{8}{15}\left(\frac{1}{9}\right)^n, \qquad n = 0, 1, 2, \ldots$$

Example 4.3 *($M/H_2/1$)*
Suppose that $\lambda = 1$ and that the service time is hyperexponentially distributed with parameters $p_1 = 1 - p_2 = 1/4$ and $\mu_1 = 1$, $\mu_2 = 2$. So the mean service time is equal to $1/4 \cdot 1 + 3/4 \cdot 1/2 = 5/8$. The Laplace-Stieltjes transform of the service time is given by

$$\widetilde{B}(s) = \frac{1}{4} \cdot \frac{1}{1+s} + \frac{3}{4} \cdot \frac{2}{2+s} = \frac{1}{4} \cdot \frac{8+7s}{(1+s)(2+s)}\,.$$

Thus we have

$$\begin{aligned}
P_{L^d}(z) &= \frac{\frac{3}{8}\frac{1}{4}\frac{15-7z}{(2-z)(3-z)}(1-z)}{\frac{1}{4}\frac{15-7z}{(2-z)(3-z)} - z} \\
&= \frac{3}{8} \cdot \frac{(15-7z)(1-z)}{(15-7z) - 4z(2-z)(3-z)} \\
&= \frac{3}{8} \cdot \frac{15-7z}{(3-2z)(5-2z)} \\
&= \frac{3}{8} \cdot \frac{9/4}{3-2z} + \frac{3}{8} \cdot \frac{5/4}{5-2z/5} \\
&= \frac{9/32}{1-2z/3} + \frac{3/32}{1-2z/5}\,.
\end{aligned}$$

So

$$d_n = p_n = \frac{9}{32}\left(\frac{2}{3}\right)^n + \frac{3}{32}\left(\frac{2}{5}\right)^n, \qquad n = 0, 1, 2, \ldots$$

4.3 Distribution of the sojourn time

We now turn to the calculation of how long a customer spends in the system. We will show that there is a nice relationship between the transforms of the time spent in the system and the departure distribution.

Let us consider a customer arriving at the system in equilibrium. Denote the total time spent in the system for this customer by the random variable S with distribution function $F_S(\cdot)$ and density $f_S(\cdot)$. The distribution of the number of customers left behind upon departure of our customer is equal to $\{d_n\}_{n=0}^{\infty}$ (since the system is in equilibrium). In considering a first-come first-served system it is clear that all customers left behind are precisely those who arrived during his stay in the system. Thus we have (cf. (2))

$$d_n = \int_{t=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_S(t) dt.$$

Hence, we find similarly to (5) that

$$P_{L^d}(z) = \widetilde{S}(\lambda - \lambda z).$$

Substitution of this relation into (6) yields

$$\widetilde{S}(\lambda - \lambda z) = \frac{(1-\rho)\widetilde{B}(\lambda - \lambda z)(1-z)}{\widetilde{B}(\lambda - \lambda z) - z}\,.$$

Making the change of variable $s = \lambda - \lambda z$ we finally arrive at

$$\widetilde{S}(s) = \frac{(1-\rho)\widetilde{B}(s)s}{\lambda \widetilde{B}(s) + s - \lambda}\,. \tag{7}$$

This formula is also known as the *Pollaczek-Khinchin formula.*

Example 4.4 *(M/M/1)*
For exponential service times with mean $1/\mu$ we have

$$\widetilde{B}(s) = \frac{\mu}{\mu + s}\,.$$

Thus

$$\widetilde{S}(s) = \frac{(1-\rho)\frac{\mu}{\mu+s}s}{\lambda\frac{\mu}{\mu+s} + s - \lambda} = \frac{(1-\rho)\mu s}{\lambda\mu + (s-\lambda)(\mu+s)} = \frac{(1-\rho)\mu s}{(\mu-\lambda)s + s^2} = \frac{\mu(1-\rho)}{\mu(1-\rho)+s}\,.$$

Hence, S is exponentially distributed with parameter $\mu(1-\rho)$, i.e.,

$$F_S(t) = P(S \le t) = 1 - e^{-\mu(1-\rho)t}, \qquad t \ge 0.$$

Example 4.5 *($M/E_2/1$)*
Suppose that $\lambda = 1$ and that the service time is Erlang-2 distributed with mean 1/3, so

$$\widetilde{B}(s) = \left(\frac{6}{6+s}\right)^2.$$

Then it follows that (verify)

$$F_S(t) = \frac{8}{5}(1 - e^{-3t}) - \frac{3}{5}(1 - e^{-8t}), \qquad t \ge 0.$$

Example 4.6 *($M/H_2/1$)*
Consider example 4.3 again. From (7) we obtain (verify)

$$F_S(t) = \frac{27}{32}(1 - e^{-t/2}) + \frac{5}{32}(1 - e^{-3t/2}), \qquad t \ge 0.$$

4.4 Distribution of the waiting time

We have that S, the time spent in the system by a customer, is the sum of W (his waiting time) and B (his service time), where W and B are independent. Since the transform of the sum of two independent random variables is the product of the transforms of these two random variables, it holds that

$$\widetilde{S}(s) = \widetilde{W}(s) \cdot \widetilde{B}(s). \tag{8}$$

Together with (7) it follows that

$$\widetilde{W}(s) = \frac{(1-\rho)s}{\lambda \widetilde{B}(s) + s - \lambda}\,, \tag{9}$$

which is the third form of the Pollaczek-Khinchin formula.

Example 4.7 *(M/M/1)*
For exponential service times with mean $1/\mu$ we have

$$\widetilde{B}(s) = \frac{\mu}{\mu + s}\,.$$

Then from (9) it follows that (verify)

$$\widetilde{W}(s) = (1-\rho) + \rho \cdot \frac{\mu(1-\rho)}{\mu(1-\rho) + s}\,.$$

The inverse transform yields

$$F_W(t) = P(W \leq t) = (1-\rho) + \rho(1 - e^{-\mu(1-\rho)t}), \qquad t \geq 0.$$

Hence, with probability $(1-\rho)$ the waiting time is zero (i.e., the system is empty on arrival) and, given that the waiting time is positive (i.e., the system is not empty on arrival), the waiting time is exponentially distributed with parameter $\mu(1-\rho)$.

4.5 Mean value approach

The mean waiting time can of course be calculated from the Laplace-Stieltjes transform (9) by differentiating and substituting $s = 0$. In this section we show that the mean waiting time can also be determined directly (i.e., without transforms) with the *mean value approach*.

A new arriving customer first has to wait for the *residual service time* of the customer in service (if there is one) and then continues to wait for the servicing of all customers who were already waiting in the queue on arrival. By PASTA we know that with probability ρ the server is busy on arrival. Let the random variable R denote the residual service time and let L^q denote the number of customers waiting in the queue. Hence,

$$E(W) = E(L^q)E(B) + \rho E(R),$$

and by Little's law (applied to the queue),

$$E(L^q) = \lambda E(W).$$

So we find

$$E(W) = \frac{\rho E(R)}{1-\rho}\,. \tag{10}$$

Formula (10) is commonly referred to as the Pollaczek-Khinchin mean value formula. It remains to calculate the mean residual service time. In the following section we will show that

$$E(R) = \frac{E(B^2)}{2E(B)}\,, \tag{11}$$

which may also be written in the form

$$E(R) = \frac{E(B^2)}{2E(B)} = \frac{\sigma_B^2 + E(B)^2}{2E(B)} = \frac{1}{2}(c_B^2+1)E(B). \tag{12}$$

An important observation is that, clearly, the mean waiting time only depends upon the first two moments of service time (and not upon its distribution). So in practice it is sufficient to know the mean and standard deviation of the service time in order to estimate the mean waiting time.

4.6 Residual service time

Suppose that our customer arrives when the server is busy and denote the total service time of the customer in service by X. Further let $f_X(\cdot)$ denote the density of X. The basic observation to find $f_X(\cdot)$ is that it is more likely that our customer arrives in a long service time than in a short one. So the probability that X is of length x should be proportional to the length x as well as the frequency of such service times, which is $f_B(x)dx$. Thus we may write

$$P(x \le X \le x + dx) = f_X(x)dx = Cxf_B(x)dx,$$

where C is a constant to normalize this density. So

$$C^{-1} = \int_{x=0}^{\infty} xf_B(x)dx = E(B).$$

Hence

$$f_X(x) = \frac{xf_B(x)}{E(B)}\,.$$

Given that our customer arrives in a service time of length x, the arrival instant will be a random point within this service time, i.e., it will be uniformly distributed within the service time interval $(0,x)$. So

$$P(t \le R \le t + dt | X = x) = \frac{dt}{x}, \qquad t \le x.$$

Of course, this conditional probability is zero when $t > x$. Thus we have

$$P(t \leq R \leq t + dt) = f_R(t)dt = \int_{x=t}^{\infty} \frac{dt}{x} f_X(x)dx = \int_{x=t}^{\infty} \frac{f_B(x)}{E(B)} dxdt = \frac{1 - F_B(t)}{E(B)} dt.$$

This gives the final result

$$f_R(t) = \frac{1 - F_B(t)}{E(B)} \, ,$$

from which we immediately obtain, by partial integration,

$$E(R) = \int_{t=0}^{\infty} t f_R(t)dt = \frac{1}{E(B)} \int_{t=0}^{\infty} t(1 - F_B(t))dt = \frac{1}{E(B)} \int_{t=0}^{\infty} \frac{1}{2} t^2 f_B(t)dt = \frac{E(B^2)}{2E(B)} \, .$$

This computation can be repeated to obtain all moments of R, yielding

$$E(R^n) = \frac{E(B^{n+1})}{(n+1)E(B)} \, .$$

Example 4.8 *(Erlang service times)*
For an Erlang-r service time with mean r/μ we have

$$E(B) = \frac{r}{\mu} \, , \qquad \sigma^2(B) = \frac{r}{\mu^2} \, ,$$

so

$$E(B^2) = \sigma^2(B) + (E(B))^2 = \frac{r(1+r)}{\mu^2} \, .$$

Hence

$$E(R) = \frac{1+r}{2\mu}$$