Lecture 24: Observed distributions in CTMCs – Little's Law and Pollaczek-Khinchin mean value formulae

Concepts checklist

At the end of this lecture, you should be able to:

- *state* and *apply* Little's Law; and,
- *state* and *apply* the Pollaczek-Khinchin mean value formulae.

Little's Law

For the $M/M/N$ queue, there is a relation between the average number of waiting customers in the queue and the mean waiting time. The average number of waiting customers (not in service) $\mathbb{E}[L_Q]$ is given by

$$\mathbb{E}[L_Q] = \lambda \mathbb{E}[W_Q]. \tag{24}$$

Equality (24), which is an example of Little's Law, holds in systems much more complex than the $M/M/N$ queue. To see this, let Γ be any section of a queueing system. Assume that the queueing system is stationary (in equilibrium) and let

- $\overline{L}(\Gamma)$ be the average number of customers in Γ,
- $\overline{W}(\Gamma)$ be the average time a customer spends in Γ and
- $\overline{\lambda}(\Gamma)$ be the average number of customers entering Γ per unit time.

Then Little's Law gives a relationship between $\overline{L}(\Gamma)$, $\overline{W}(\Gamma)$ and $\lambda(\Gamma)$.

Theorem 24 (Little's Law.). *In equilibrium,*

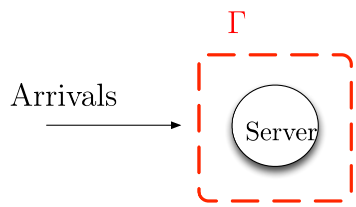
$$\overline{L}(\Gamma) = \overline{\lambda}(\Gamma)\overline{W}(\Gamma). \tag{25}$$

Example 22. $M/G/1/1$ queue

Consider a single-server queue with general service times with mean $1/\mu$, and assume that blocked calls are lost.

Goal: determine the probability that the server is busy.

Consider Γ to be the server:



The input rate $\overline{\lambda}(\Gamma)$ into Γ is $\lambda(1-\Pr(\text{server busy}))$.
The average time $\overline{W}(\Gamma)$ in Γ is given by $\overline{W}=1/\mu$.
The average number $\overline{L}(\Gamma)$ of customers in service is given by

$$\overline{L}=1\times\Pr(\text{server busy})+0\times(1-\Pr(\text{ server busy})).$$

By Little's Law, we have

$$\Pr(\text{server busy})=\frac{\lambda(1-\Pr(\text{server busy}))}{\mu},$$

which implies

$$\Pr\left(\text{server busy}\right)=\frac{a}{1+a},\quad\text{where}\quad a=\frac{\lambda}{\mu}.$$

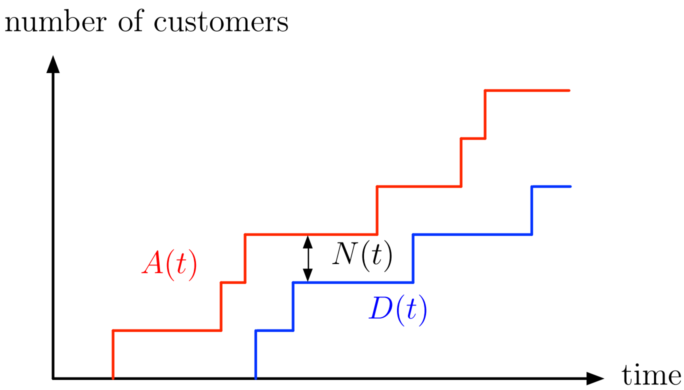
Comparing this with the exponential service time case we see that

$$\Pr\left(\text{server busy}\right)=B(1,a)=\frac{a}{1+a}.$$

A more general result can be shown to apply:

The equilibrium distribution for the $M/G/N/N$ system, including the probability that a call is blocked can be shown to depend on the service time distribution only through its mean. That is, the exponential distribution with mean $1/\mu$ gives the correct equilibrium distribution for an $M/G/N/N$ queue with a general service time distribution with mean $1/\mu$.

Proof. Let $A(t)$ be the number of arrivals to Γ in $(0,t)$, $D(t)$ be the number of departures from Γ in $(0,t)$, and $N(t)=A(t)-D(t)$ be the number in Γ at time t.



On the time interval $[0,t)$, let $\overline{\lambda_t}(\Gamma)$ be the mean arrival rate in $[0,t)$, then

$$\overline{\lambda_t}(\Gamma)=A(t)/t.$$

Let $S(t)$ be the total area between curves $A(t)$ and $D(t)$ up to time t, then

$$S(t)=\int_0^t N(u)du=\text{ cumulative time spent in }\Gamma\text{ up to time }t.$$

Then,

$$\overline{W}_t(\Gamma)=\text{ mean time each customer spends in }\Gamma=\frac{S(t)}{A(t)}$$

$$\text{and}\quad \overline{L}_t(\Gamma)=\text{ average number of customers in }\Gamma\text{ during }[0,t)=\frac{S(t)}{t}.$$

Hence,

$$\overline{L}_t(\Gamma) = \overline{\lambda}_t(\Gamma)\overline{W}_t(\Gamma).$$

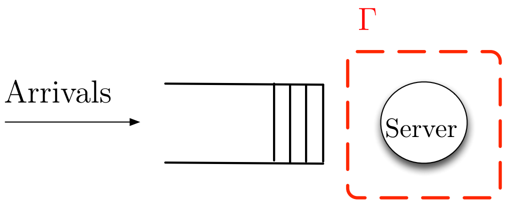
Letting $t \to \infty$, we have $\overline{L}(\Gamma) = \overline{\lambda}(\Gamma)\overline{W}(\Gamma)$ as required. $\square$

Note: We have made no assumption about Poisson arrivals, correlations between customers behaviour, the number of servers, exponential service times etc. The queueing system is completely arbitrary and yet the expression is valid.

Example 23. $M/G/1$ queue

Consider a single-server queue with general service times and an infinite capacity.

Goal: to determine the probability that the server is busy.



Here, we have $\overline{W} = 1/\mu$ and the input rate into $\Gamma = \lambda$. Using Little's law,

$$\overline{L} = \Pr(\text{server busy}) = \lambda\overline{W} = a.$$

Therefore, the probability π_0 that the $M/G/1$ system is empty, is $\pi_0 = 1 - a$.

Pollaczek-Khinchin mean value formulae for an arbitrary $M/G/1$ queue

Let q_n be the number of customers left behind by the departure of the nth customer of an $M/G/1$ queue, and let v_n be the number of customers that arrive during the service of the nth customer.

Then we have

$$q_{n+1} = \begin{cases} q_n - 1 + v_{n+1} & \text{if } q_n > 0 \\ v_{n+1} & \text{if } q_n = 0 \end{cases},$$

which can be combined to get

$$q_{n+1} = q_n - \mathbb{1}_{\{q_n>0\}} + v_{n+1}, \quad \text{where } \mathbb{1}_{\{q_n>0\}} = \begin{cases} 1 & \text{if } q_n > 0 \\ 0 & \text{if } q_n = 0 \end{cases}.$$

Now, take the expectations of v_n and q_n. Assume that we can take limits as $n \to \infty$ and that $\mathbb{E}[v_n] \to \mathbb{E}[v]$ and $\mathbb{E}[q_n] \to \mathbb{E}[q]$. Then, from before we get

$$\mathbb{E}[q] = \mathbb{E}[q] - \mathbb{E}[\mathbb{1}_{\{q>0\}}] + \mathbb{E}[v],$$
$$\Rightarrow \mathbb{E}[\mathbb{1}_{\{q>0\}}] = \mathbb{E}[v].$$

On the other hand, as

$$\mathbb{E}[\mathbb{1}_{\{q>0\}}] = 0.\Pr\{q = 0\} + 1.\Pr\{q > 0\} = \Pr\{q > 0\},$$

so we have $\Pr(q > 0) = \mathbb{E}[v]$. In words, the probability that the queue is non-empty at a departure instant is equal to the expected number of arrivals in a service.

Note: if $\mathbb{E}[v] > 1$ there will, on average, be more than one arrival in each service, the queue will be unstable, and the above assumptions about the legality of taking limits will not hold.

We get more information from $q_{n+1} = q_n - \mathbb{1}_{\{q_n>0\}} + v_{n+1}$ by squaring it:

$$q_{n+1}^2 = q_n^2 + \mathbb{1}_{\{q_n>0\}}^2 + v_{n+1}^2 - 2q_n\mathbb{1}_{\{q_n>0\}} + 2q_nv_{n+1} - 2\mathbb{1}_{\{q_n>0\}}v_{n+1}.$$

Note that $\mathbb{1}_{\{q_n>0\}}^2 = \mathbb{1}_{\{q_n>0\}}$ and $q_n\mathbb{1}_{\{q_n>0\}} = q_n$. Then, by the independence of the number of arrivals and the queue length we have

$$\mathbb{E}[q_nv_{n+1}] = \mathbb{E}[q_n]\mathbb{E}[v_{n+1}] \quad \text{and} \quad \mathbb{E}[\mathbb{1}_{\{q_n>0\}}v_{n+1}] = \mathbb{E}[\mathbb{1}_{\{q_n>0\}}]\mathbb{E}[v_{n+1}].$$

We then take expectations, and use the above result that $\mathbb{E}[\mathbb{1}_{\{q>0\}}] = \mathbb{E}[v]$ to show that

$$\begin{aligned} \mathbb{E}[q^2] &= \mathbb{E}[q^2] + \mathbb{E}[v] + \mathbb{E}[v^2] - 2\mathbb{E}[q] + 2\mathbb{E}[q]\mathbb{E}[v] - 2\mathbb{E}[v]^2 \\ \Rightarrow \quad \mathbb{E}[q] &= \mathbb{E}[v] + \frac{\mathbb{E}[v^2] - \mathbb{E}[v]}{2(1-\mathbb{E}[v])}. \end{aligned}$$

It can be argued, using a graph of arrivals and departures, that the departure points leaving n customers in the queue can be matched one-to-one with the arrival points which find n customers in the queue and therefore that the distribution as left behind by departures is the same as the distribution seen by arrivals.

We can therefore appeal to PASTA to conclude that

$$\mathbb{E}[q] = \mathbb{E}[v] + \frac{\mathbb{E}[v^2] - \mathbb{E}[v]}{2(1-\mathbb{E}[v])} \tag{26}$$

is also the mean queue length at any arbitrary time.

Using Little's Law, we can further get an expression for the mean time spent by a customer in the queue as

$$\mathbb{E}[w] = \frac{1}{\lambda}\left[\mathbb{E}[v] + \frac{\mathbb{E}[v^2] - \mathbb{E}[v]}{2(1-\mathbb{E}[v])}\right]. \tag{27}$$

Now, using $\Pr(v = k) = \displaystyle\int_0^\infty \frac{(\lambda t)^k}{k!}e^{-\lambda t}\mathrm{d}B(t)$ — k arrivals during a service time — we can show that

$$\mathbb{E}(v^2) - \mathbb{E}(v) = \lambda^2\mathbb{E}[Y^2],$$

where $\mathbb{E}[Y^2]$ is the second (non-central) moment of the service time distribution $B(t)$.

Equations (26) and (27) can now be written in the equivalent forms

$$\mathbb{E}[q] = a + \frac{\lambda^2\mathbb{E}[Y^2]}{2(1-a)}, \tag{28}$$

$$\text{and} \quad \mathbb{E}[w] = \frac{1}{\lambda}\left[a + \frac{\lambda^2\mathbb{E}[Y^2]}{2(1-a)}\right]. \tag{29}$$

Equations (28) and (29) are known as the *Pollaczek-Khinchin mean value formulae* for the queue length and waiting time respectively. They give expressions for the mean queue length and the mean waiting time for an arbitrary $M/G/1$ queue.