Lecture 9 — The $M/G/1$ System

In this lecture we move away from studying purely Markov systems and study the $M/G/1$ queue and the special case of the $M/D/1$ queue. (Note that we could see the $M/M/1$ queue as a special case of the $M/G/1$ queue). The result derived is known as the Pollaczek-Khinchin (P-K) formula. The formula we are working to prove is given by first defining:

$$\overline{X} = \mathrm{E}[X] = \frac{1}{\mu} = \text{Average service time}$$

and

$$\overline{X^2} = \mathrm{E}\left[X^2\right] = \text{Second moment of service time}$$

The P-K formula is then:

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)}$$

where W is the expected customer waiting time in a queue and $\rho = \lambda/\mu = \lambda\overline{X}$ the utilisation as usual.

This lecture we will derive and use the P-K formula and a simple variant.

First let us introduce some notation:

W_i waiting time (in queue) for ith customer.

X_i service time of the ith customer – we assume that these are independent and identically distributed (i.i.d) variables.

N_i number of customers that is found in the queue (not yet being served) when the ith customer arrives.

R_i residual service time found by the ith customer (defined below).

Definition 1. The residual time R_i is the service time remaining to the customer being served when the ith customer arrives at the queue. If no customer is being served ($N_i = 0$) then $R_i = 0$.

A graph will help understand the concept of residual time. Figure 1 shows the residual time in a queue $r(\tau)$ is the residual time remaining at time τ. X_i is the service time of the ith customer (note that the slopes of all the diagonal lines on this graph are, obviously, one). If we take a time t where the system is empty (as shown in the diagram) then define $M(t)$ as the number of customers who have been served and exited the system by time t.

The mean residual time in the interval $[0, t]$ is clearly the average value on the y axis in the interval. This is the area under the curve divided by t which is given by

$$\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{t}\sum_{i=1}^{M(t)} \frac{1}{2}X_i^2.$$

Which we can rewrite as

$$\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)}.$$

Now, assuming the relevant limits exist we have:

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{2}\lim_{t\to\infty}\frac{M(t)}{t}\lim_{t\to\infty}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)}. \tag{1}$$

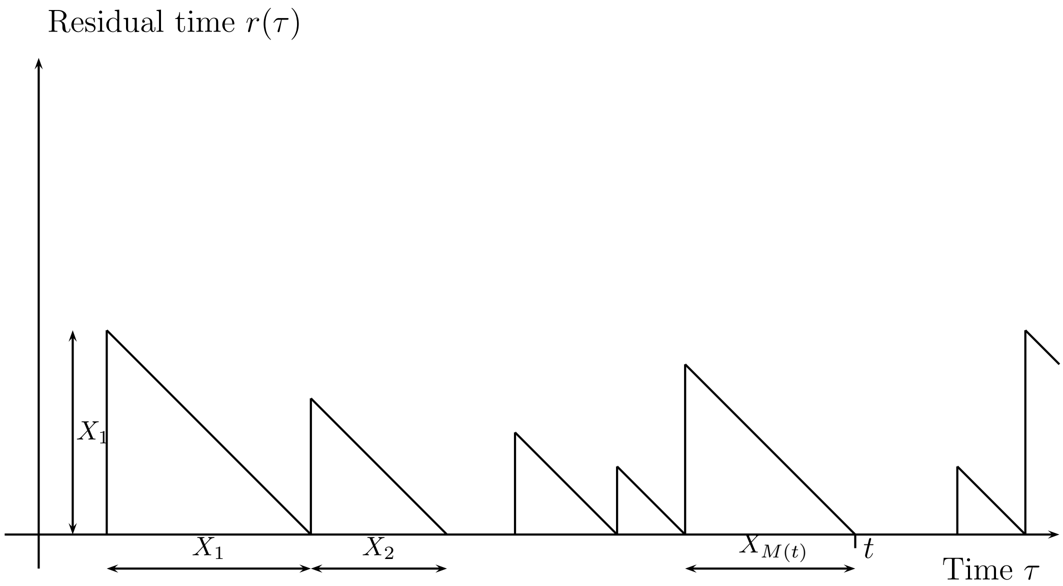

Figure 1: Service Time of Arrivals at an $M/G/1$ queue.

Now, if we assume that the system is ergodic then we can replace these time averages with ensemble averages. In this case define

$$R = \text{Mean residual time} = \lim_{i\to\infty} \mathrm{E}\left[R_i\right],$$

and, if the time average is the state space average, then

$$R = \lim_{t\to\infty} \frac{1}{t}\int_0^t r(\tau)d\tau.$$

Since the system is lossless (no customers ever vanish) then if the number of customers does not rise forever — the number queuing tends to a limit — we can say that the departure rate must equal the arrival rate. That is

$$\lim_{t\to\infty} \frac{M(t)}{t} = \lambda.$$

Therefore equation (1) becomes

$$R = \frac{1}{2}\lambda\overline{X^2}. \tag{2}$$

Now, we know that the waiting time for the $i$ith customer is equal to the residual service time of the customer currently being served plus the total service times of those who are in the queue. This is given by

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j.$$

We note that the X_js are i.i.d by hypothesis. N_i cannot possibly be affected by the X_j values in this sum since those are the service times of customers who are still waiting in this queue.

Therefore, N_i is also independent from the X_j in the above. Therefore we may take expectations as follows

$$\mathrm{E}\left[W_i\right] = \mathrm{E}\left[R_i\right] + \mathrm{E}\left[\sum_{j=i-N_i}^{i-1} \mathrm{E}\left[X_j|N_i\right]\right] = \mathrm{E}\left[R_i\right] + \overline{X}\mathrm{E}\left[N_i\right].$$

Finally, taking the limit as $i \to \infty$ and remembering that $\overline{X} = \frac{1}{\mu}$ then

$$W = R + \frac{1}{\mu}N_Q,$$

where N_Q is the limit as $i \to \infty$ of the expected number found in the queue. By Little's theorem we get

$$N_Q = \lambda W,$$

and therefore

$$W = R + \frac{\lambda}{\mu}W.$$

Rearranging and substuting $\rho = \lambda/\mu$ and our expression for R from equation (2) then

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)},$$

which is the P-K formula we required.

Let us remember the assumptions for this remarkably general formula:

1. The sending process was a Poisson process with parameter λ.
2. The steady state time averages R, W and N_Q exist.
3. The long-term time averages correspond to the state-space averages.
4. The service times X_i are i.i.d. variables.

In our derivation we also assumed that the system was FIFO although this is not, in fact, necessary — it is only necessary that the order of service is independent of the required service time.

Note that the $M/D/1$ queue is the special case of this when all service times are identical. In this case $X_i = \frac{1}{\mu}$ and therefore $\overline{X^2} = \frac{1}{\mu^2}$ and

$$W = \frac{\rho}{2\mu(1-\rho)}.$$

This is the lowest possible value of $\overline{X^2}$ and therefore a lower bound for any $M/G/1$ system. Compare it to the $M/M/1$ system where $\overline{X^2} = 2/\mu^2$ and therefore

$$W = \frac{\rho}{\mu(1-\rho)}.$$

In other words the $M/M/1$ formula has twice the waiting time of the lower bound $M/D/1$ waiting time. We should also note that there is no upper bound on $\overline{X^2}$ therefore it is possible that queues which have a utilisation less than one have an infinite waiting time.

Further $M/G/1$ information

Question

What is the probability that the system is empty when a customer arrives?

Answer

The expected time to serve n customers is $\sum_{i=1}^{n} X_i$. The expected time for n customers to depart is n/λ (since the customers are generated by a Poisson process with rate λ and are also departing at a similar rate as previously stated).

$$\mathbb{P}[\text{Empty}] = \lim_{n\to\infty} \frac{\text{Time taken for } n \text{ customers to depart} - \text{Time serving } n \text{ customers}}{\text{Time taken for } n \text{ customers to depart}},$$

which is

$$\mathbb{P}[\text{Empty}] = \lim_{n\to\infty} \frac{n/\lambda - \sum_{i=1}^{n} X_i}{n/\lambda}.$$

Therefore,

$$\mathbb{P}[\text{Empty}] = 1 - \lambda\overline{X}.$$

Question

What is the average length between busy periods?

Answer

A period between busy periods begins when the last customer exits. It will end when the next customer is generated. Since the generating process is a Poisson and therefore memoryless, the expected time for the next arrival is after a time $1/\lambda$.

Question

What is the average length of a busy period?

Answer

If L is the average length of a busy period then

$$\mathbb{P}[\text{Empty}] = \frac{1/\lambda}{L + 1/\lambda} \tag{3}$$

Substuting from earlier and multiplying top and bottom of RHS by λ

$$1 - \lambda\overline{X} = \frac{1}{\lambda L + 1}.$$

Rearranging gives

$$\lambda L + 1 = \frac{1}{1 - \lambda\overline{X}},$$

and final rearrangement gives

$$L = \frac{\overline{X}}{1 - \lambda\overline{\overline{X}}}$$