Proof. We have already established that

$$P_j^{(E)}(t) = \frac{\lambda P_j(t)}{\lambda \sum_{k\in\mathcal{S}} P_k(t)} \quad \text{for all } j \in \mathcal{S}$$

and since $\sum_{k\in S} P_k(t) = 1$, we have that $P_j^{(E)}(t) = P_j(t)$ for all $j \in \mathcal{S}$. Taking the limits as $t \to \infty$ of the above gives $\pi_j^{(E)} = \pi_j$ for all $j \in \mathcal{S}$. □

Note that we have not assumed that the Poisson stream of arrivals necessarily changes the state of the process when it occurs. Very often this is the case, such as for the $M/M/N$ queue, but for the $M/M/N/N$ queue, for example, the Poisson input keeps happening when the system is full even though they are lost.

By the above result, however, we see that the arrival stream to an $M/M/N/N$ queue also has the PASTA property. Consequently,

$$\begin{aligned}\text{the proportion of calls lost} &= \Pr(\text{an arrival is rejected})\\ &= \Pr\{\text{an arrival sees } N \text{ in the system}\} = \pi_N^{(E)}\\ &= \pi_N \quad \text{by PASTA.}\end{aligned}$$

Earlier in this course, we **assumed** that the probability that a call is lost is π_N but, as noted above, this result relies on the PASTA property and is not usually true when the arrival stream is not Poisson.

Example 20. The Birth and Death Process

Let us now consider the birth and death process and how the new arrivals observe the state of the process. Let $\{\pi_j^{(A)}, j \geq 0\}$ be the equilibrium distribution just before arrival points.

Theorem 23. *For an arbitrary birth and death process the equilibrium distribution as seen by those arrivals that change the state of the process can be written as*

$$\pi_j^{(A)} = \pi_0^{(A)} \prod_{\ell=1}^{j} \frac{\lambda_\ell}{\mu_\ell}, \quad \textit{for all } j \geq 0.$$

Note: this is not the same as $\pi_j = \pi_0 \prod_{\ell=1}^{j} \frac{\lambda_{\ell-1}}{\mu_\ell}$.

Proof. Set $\gamma_j = \lambda_j$ for all $j \geq 0$ and then we have by Theorem 21

$$\begin{aligned}\pi_j^{(A)} &= \frac{\lambda_j \pi_j}{\sum_{k\in\mathcal{S}} \lambda_k \pi_k} = \frac{\lambda_j \pi_0}{\sum_{k\in\mathcal{S}} \lambda_k \pi_k} \prod_{\ell=0}^{j-1} \frac{\lambda_\ell}{\mu_{\ell+1}}\\ &= \frac{\lambda_0 \pi_0}{\sum_{k\in\mathcal{S}} \lambda_k \pi_k} \prod_{\ell=1}^{j} \frac{\lambda_\ell}{\mu_\ell} = \pi_0^{(A)} \prod_{\ell=1}^{j} \frac{\lambda_\ell}{\mu_\ell}.\end{aligned}$$

□

Note that if the arrival rate is a constant λ for all j then we have $\pi_j^{(A)} = \pi_j$ (by PASTA).

Example 21. $M/M/N$ queue

We assume that the queue is operating under a first come first served (FCFS) or first in first out (FIFO) discipline. So, the customer will have to wait whenever it arrives to find greater than or equal to N customers already in the system.

(a) **What is the probability that an arriving customer does not have to wait?**

Let W_Q be the random variable for the equilibrium waiting time in the queue (not including the service time). Then, the event $W_Q = 0$ is the event that an arriving customer sees less than N customers in the queue. Hence,

$$\begin{aligned}\Pr(W_Q = 0) &= \sum_{j=0}^{N-1} \pi_j^{(A)} = \sum_{j=0}^{N-1} \pi_j \quad \text{by PASTA} \\ &= \pi_0 \sum_{j=0}^{N-1} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}.\end{aligned}$$

(b) **What is the distribution of the customer's waiting time?**

If the customer arrives to find $N + k$ in the system it has to wait for $k + 1$ services before they are served. During this period, all servers are busy and so the service rate is $N\mu$.

The distribution of the waiting time in the $M/M/N$ queue for a customer that arrives to find $N + k$ in the system is the distribution of time until the $(k + 1)st$ event when the event rate is $N\mu$. The density function for this is the Erlang density function. Thus,

$$\begin{aligned}f_{W_Q}(t) = \frac{\mathrm{d}\Pr(W_Q < t)}{\mathrm{d}t} &= \sum_{k=0}^{\infty} \Pr\begin{pmatrix}\text{arrival finds } N+k \\ \text{in the system}\end{pmatrix} \times \begin{pmatrix}\text{density function} \\ \text{for Erlang}(k+1, N\mu)\end{pmatrix} \\ &= \sum_{k=0}^{\infty} \left(\pi_{N+k}^{(A)}\right) \times \left(N\mu \frac{(N\mu t)^k}{k!} e^{-N\mu t}\right) \\ &= \sum_{k=0}^{\infty} \left(\left(\frac{\lambda}{N\mu}\right)^k \pi_N\right) \times \left(N\mu \frac{(N\mu t)^k}{k!} e^{-N\mu t}\right) \\ &= N\mu e^{-N\mu t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \pi_N \\ &= N\mu e^{(\lambda - N\mu)t} \pi_N.\end{aligned}$$

We can replace π_N in the previous expression by

$$\pi_N = \left(1 - \frac{\lambda}{N\mu}\right) C\left(N, \frac{\lambda}{\mu}\right),$$

where

$$C\left(N, \frac{\lambda}{\mu}\right) = \Pr(\text{all servers are busy}) = \sum_{j=N}^{\infty} \pi_j.$$

Using this definition of $C\left(N, \frac{\lambda}{\mu}\right)$ (known as the Erlang C formula), we can write

$$f_{W_Q}(t) = \frac{\mathrm{d}\Pr(W_Q < t)}{dt} = C\left(N, \frac{\lambda}{\mu}\right)(N\mu - \lambda)e^{-(N\mu-\lambda)t}.$$

Note that $(N\mu - \lambda)e^{-(N\mu-\lambda)t}$ is an exponential density function with parameter $(N\mu - \lambda)$, which is the difference between the maximum service rate and the arrival rate.

The last equation gives the density function for the equilibrium waiting time of a customer that arrives to an $M/M/N$ queue and by integrating this from t to ∞ we obtain the probability that the waiting time is greater than t. That is,

$$\begin{aligned} 1 - F_{W_Q}(t) = \Pr(W_Q > t) &= C\left(N, \frac{\lambda}{\mu}\right)\int_t^\infty (N\mu - \lambda)e^{-(N\mu-\lambda)u}\mathrm{d}u \\ &= C\left(N, \frac{\lambda}{\mu}\right)e^{-(N\mu-\lambda)t}. \end{aligned}$$

(c) **What is the mean waiting time of an arriving customer?**

$$\begin{aligned} \mathbb{E}[W_Q] &= C\left(N, \frac{\lambda}{\mu}\right) \times \begin{pmatrix} \text{Mean of an exponential random} \\ \text{variable with parameter } (N\mu - \lambda) \end{pmatrix} \\ &= \frac{C\left(N, \frac{\lambda}{\mu}\right)}{N\mu - \lambda}. \end{aligned}$$

(d) **What is the conditional waiting time of an arriving customer, given that they have to wait?**

$$\begin{aligned} \Pr(W_Q > t \mid W_Q > 0) &= \frac{\Pr(W_Q > t, W_Q > 0)}{\Pr(W_Q > 0)} \\ &= \frac{\Pr(W_Q > t)}{\Pr(W_Q > 0)} \\ &= \frac{C\left(N, \frac{\lambda}{\mu}\right)e^{-(N\mu-\lambda)t}}{C\left(N, \frac{\lambda}{\mu}\right)} \\ &= e^{-(N\mu-\lambda)t}. \end{aligned}$$

(e) **What is conditional expectation of waiting time, given that they have to wait?**

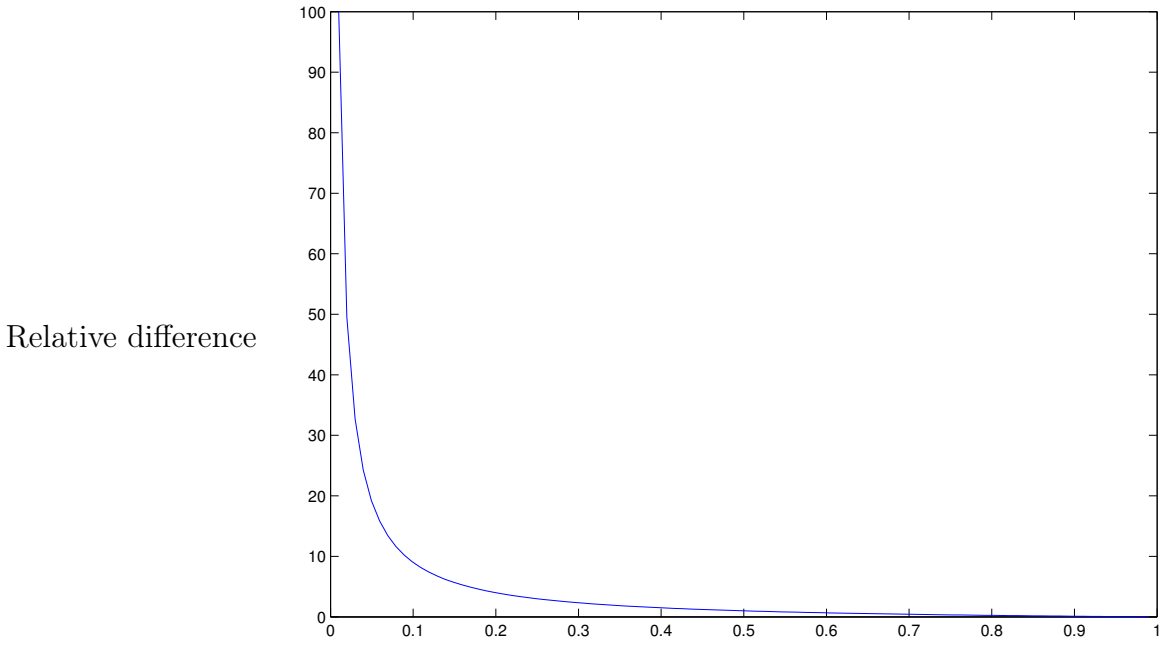
$$\mathbb{E}[W_Q \mid W_Q > 0] = \begin{pmatrix} \text{Mean of an exponential random} \\ \text{variable with parameter } (N\mu - \lambda) \end{pmatrix} = \frac{1}{N\mu - \lambda}.$$

For a single server ($N = 1$), we have

$$\mathbb{E}[W_Q] = \frac{\lambda}{\mu(\mu - \lambda)} \quad \text{and} \quad \mathbb{E}[W_Q \mid W_Q > 0] = \frac{1}{\mu - \lambda}.$$

For a single server ($N = 1$), where $\mu = 1$, plotting the relative difference

$$\frac{\mathbb{E}[W_Q \mid W_Q > 0] - \mathbb{E}[W_Q]}{\mathbb{E}[W_Q]} \quad \text{against } 0 < \lambda < 1 \quad \text{yields}$$



Notice for low offered load that the expected waiting time can be such that the conditional waiting time, given that you have to wait, may be hundreds of times the expected waiting time.