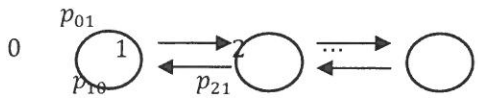
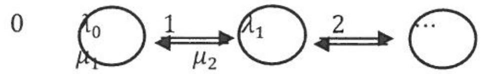
MATH 4581: Queues



In a continuous time queue, which is a continuous time Markov chain, a transition can occur any time, and it is meaningless to talk about "next time step". We will think of a queue as a continuous time **birth-death process**, which is a process where the only transitions are between adjacent states, i.e. the state (number of people in the queue) changes by $+1$ or -1. Let λ_n and μ_n be the birth rate and death rate for state n. Then, in place of the state-transition diagrams for discrete time Markov chains, we introduce **state-transition-rate** diagrams for continuous time birth-death process.



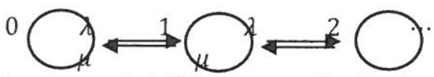
The first question we will answer is of the steady-state solution (analogous to the fixed-vector in discrete-time ergodic Markov chains), namely, in the long run, what is the probability that the system is in state i? Let p_i be the steady-state probability the system is in state i. When the system is in steady state, the "net flow of probability" in and out of a state is zero.

$$\mu_1 p_1 = \lambda_0 p_0$$
$$\mu_{i+1}p_{i+1} + \lambda_{i-1}p_{i-1} = \lambda_i p_i + \mu_i p_i \quad \Rightarrow \quad \mu_{i+1}p_{i+1} = \lambda_i p_i + \mu_i p_i - \lambda_{i-1}p_{i-1}$$

These equations, together with normalization $\sum p_i = 1$, give the steady-state probabilities p_i.

We are ready to look at some examples of queues. A queue has 4 components and is denoted by $B/D/m/n$, where B is the type of arrival (birth), D is the type of departure (death), m is the number of servers, and n is the maximum capacity of the system (omitted if infinite). We will focus mainly on cases where arrival and departure are assumed to be Poisson processes, denoted by M (memoryless?)

Example 1a In the basic queue, **M/M/1**, there is one server and infinite capacity, and the arrival and departures are Poisson processes with rates λ and μ respectively for all states. We start with a state-transition-rate diagram.



The steady-state probabilities p_i are calculated as follows.

$$\mu p_1 = \lambda p_0 \quad \Rightarrow \quad p_1 = \rho p_0, \qquad \rho = \frac{\lambda}{\mu}$$
$$\mu p_2 = \lambda p_1 + \mu p_1 - \lambda p_0 = \lambda p_1 \quad \Rightarrow \quad p_2 = \rho p_1 = \rho^2 p_0$$
$$\mu p_i = \lambda p_{i-1} \quad \Rightarrow \quad p_i = \rho p_{i-1} = \rho^i p_0$$

Normalization gives

$$\sum_{i=0}^{\infty} p_i = 1 \quad \Rightarrow \quad \sum_{i=0}^{\infty} \rho^i p_0 = 1 \quad \Rightarrow \quad p_0 = \frac{1}{\sum_i \rho^i} = 1 - \rho \qquad \text{if } \rho < 1$$

If $\rho \geq 1$, i.e. $\lambda \geq \mu$, there is no steady state; the number of people will keep increasing in the system.

When we have the steady-state probabilities, we can calculate the expected number of people in the system and in the queue (waiting line). Let N and N_q be the number of people in the system and the queue respectively, then

$$E(N) = \sum_k kP(N = k) = \sum_k kp_k$$

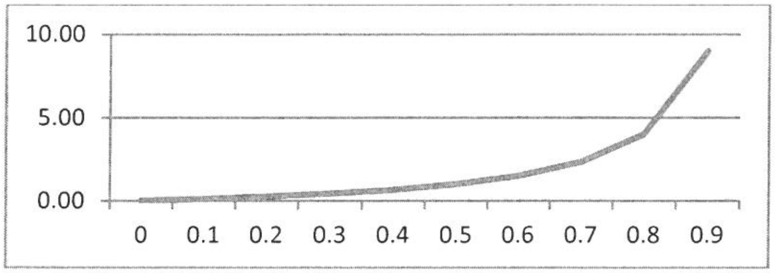
Note that $N_q = N - c$ if $N > c$ (c is the number of servers), and $N_q = 0$ if $N \leq c$, so

$$P(N_q = 0) = p_0 + p_1 + \cdots + p_{c-1}, \; P(N_q = k) = p_{k+c} \text{ for } k \geq 0$$
$$E(N_q) = \sum_k kP(N_q = k) = \sum_k kp_{k+c}$$

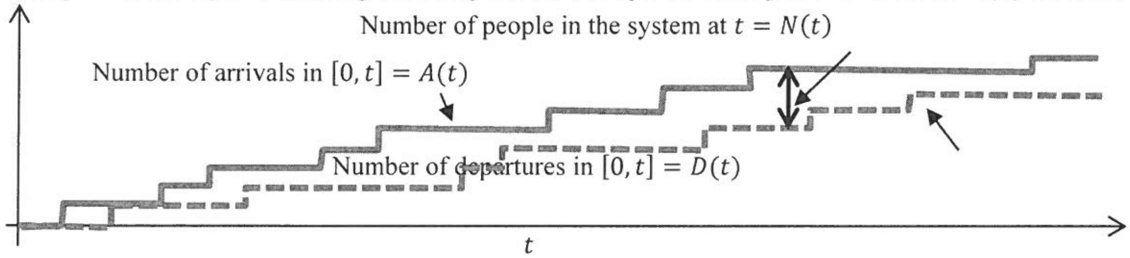
Example 1b In M/M/1,

$$E(N) = \sum_{k=0}^{\infty} kp_k = p_0 \sum_{i=0}^{\infty} k\rho^k = (1-\rho)\rho \sum_{i=1}^{\infty} k\rho^{k-1} = (1-\rho)\rho\left(\frac{d}{d\rho}\sum_{i=0}^{\infty}\rho^k\right) = (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

$$E(N_q) = \sum_{k=1}^{\infty} kp_{k+1} = p_0 \sum_{k=1}^{\infty} k\rho^{k+1} = (1-\rho)\rho^2 \sum_{k=1}^{\infty} k\rho^{k-1} = (1-\rho)\rho^2\left(\frac{d}{d\rho}\sum_{k=0}^{\infty}\rho^i\right) = (1-\rho)\rho^2\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{\rho^2}{1-\rho}$$



As shown in the above plot of $E(N)$ versus ρ, $E(N)$ increases rapidly when ρ approaches 1. Let T and W be the time a customer spends in the system (waiting and being served) and spends waiting in line. What are $E(T)$ and $E(W)$?



The area between the graphs for $A(t)$ and $D(t)$ is the accumulated customer time. As $t \to \infty$, the area is $E(N)t$. On the other hand, the cumulated customer time is $A(t)E(t)$. (Why?) So, $E(N)t = A(t)E(t)$, i.e. $E(N) = (A(t)/t)E(T)$, or

$$E(T) = \frac{E(N)}{\lambda}$$

This is known as **Little's Formula**. It is not surprising that there is an analogous version of Little's Theorem for W.

$$E(W) = \frac{E(N_q)}{\lambda}$$

Another way to calculate $E(W)$ is to recognize that Since $T = W + service\ time$, $E(T) = E(W) + 1/\mu$, and

$$E(W) = E(T) - \frac{1}{\mu}$$

Example 1c In M/M/1,

$$E(T) = \frac{E(N)}{\lambda} = \frac{1}{\lambda}\left(\frac{\lambda}{\mu - \lambda}\right) = \frac{1}{\mu - \lambda}$$

$$E(W) = E(T) - \frac{1}{\mu} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\mu - \mu + \lambda}{\mu(\mu - \lambda)} = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$E(W) = \frac{1}{\lambda}\left(\frac{\rho^2}{1 - \rho}\right) = \frac{1}{\lambda}\left(\frac{\lambda^2}{\mu^2(1 - \lambda/\mu)}\right) = \frac{\lambda}{\mu(\mu - \lambda)}$$

As we have seen, there are 4 quantities of interest, $E(N)$, $E(N_q)$, $E(T)$, and $E(W)$. Usually we first calculate $E(N)$ directly from the stationary probabilities p_i's. There are two ways to complete the calculation of the other quantities. We can calculate $E(N_q)$ directly with the pdf of N_q (derived from p_i's), then use Little's theorems for $E(T)$ and $E(W)$. Alternatively, we may first calculate $E(T)$ using Little's theorem, then $E(W) = E(T) - 1/\mu$, and $E(N_q) = \lambda E(W)$.
Finally, we are interested in the pdf's of T and of W. We will show how to obtain the pdf's for M/M/1.

Example 1d In M/M/1, suppose an arriving customer finds there are $N = m$ people in the system, then the time he/she spends in the system is the sum of the times m customers spend plus his/her own, i.e. it is the sum of $m + 1$ independent exponential distributions. We have seen in earlier that the sum of exponential distributions is the Erlang distribution (recall T_k in a Poisson process), so

$$f_T(t|N = m) = \frac{\mu^{m+1}t^m}{m!}e^{-\mu t}, \quad t \geq 0$$

To obtain the pdf for T, we sum the conditional pdf's over all values of m, weighted by $P(N = m)$.

$$f_T(t) = \sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!}\mu e^{-\mu t}P(N = m) = \sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!}\mu e^{-\mu t}(1 - \rho)\rho^m = \mu e^{-\mu t}(1 - \rho)\sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!}\rho^m$$

$$= \mu e^{-\mu t}\left(1 - \frac{\lambda}{\mu}\right)\sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!}\left(\frac{\lambda}{\mu}\right)^m = (\mu - \lambda)e^{-\mu t}\sum_{m=0}^{\infty} \frac{t^m}{m!}\lambda^m = (\mu - \lambda)e^{-\mu t}e^{\lambda t} = (\mu - \lambda)e^{-(\mu - \lambda)t}$$

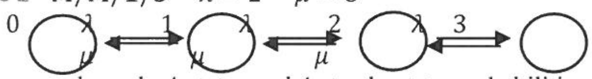
It should not surprise you that if you use $f_T(t)$ to find $E(T)$, you get $1/(\mu - \lambda)$, what we got from Little's theorem.
Next we find the pdf for W. Note that while the probability that $T = 0$ is zero, there is a non-zero probability that $W = 0$: $P(W = 0) = P(N_q = 0) = p_0$. When there are $N_q = m$ people in the waiting line, then the time he/she spends waiting in line is the sum of $m + 1$ independent exponential distributions.
To obtain the pdf for W for $W > 0$, we sum over $m \geq 0$.

$$f_W(t) = \sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!} \mu e^{-\mu t} P(N_q = m) = \sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!} \mu e^{-\mu t} p_{m+1} = \sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!} \mu e^{-\mu t} (1-\rho)\rho^{m+1}$$

$$= \mu e^{-\mu t}(1-\rho)\rho \sum_{m=0}^{\infty} \frac{(\mu \rho t)^m}{m!} = \mu e^{-\mu t}(1-\rho)\rho e^{\mu\rho t} = \rho(\mu - \mu\rho)e^{-\mu t}e^{\lambda t} = \rho(\mu-\lambda)e^{-(\mu-\lambda)t}$$

In the next examples, we discuss a system with **finite capacity**, and see what differences the finite capacity introduces.

Example 2 **M/M/1/3** $\lambda = 2$ $\mu = 3$



First, there can be only 4 states and 4 steady state probabilities, p_0, p_1, p_2, and p_3. The balance equations are the same as **M/M/1**, but there are only 3, resulting in

$$p_1 = \frac{\lambda}{\mu} p_0 = \frac{2}{3} p_0 \qquad p_2 = \frac{2}{3} p_1 = \left(\frac{2}{3}\right)^2 p_0 \qquad p_3 = \frac{2}{3} p_2 = \left(\frac{2}{3}\right)^3 p_0$$

To find p_0, we apply the normalization condition $p_0 + p_1 + p_2 + p_3 = 1$.

$$p_0 \left(1 + \frac{2}{3} + \left(\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^3\right) = 1$$

$$p_0 = \frac{27}{65} \qquad p_1 = \frac{2}{3} p_0 = \frac{18}{65} \qquad p_2 = \frac{2}{3} p_1 = \frac{12}{65} \qquad p_3 = \frac{2}{3} p_2 = \frac{8}{65}$$

$$E(N) = \sum_{k=0}^{3} k p_k = \frac{18}{65} + (2)\frac{12}{65} + (3)\frac{8}{65} = \frac{66}{65}$$

The probability distribution and expected value of N_q can be calculated.

$$P(N_q = 0) = p_0 + p_1 = \frac{45}{65} \qquad P(N_q = 1) = p_2 = \frac{12}{65} \qquad P(N_q = 2) = p_3 = \frac{8}{65}$$

$$E(N_q) = (1)\frac{12}{65} + (2)\frac{8}{65} = \frac{28}{65}$$

Before we compute the expected values and pdf's of T and W, note that because of the finite capacity, some arriving customers will be turned away, and this happen when the system is at maximum capacity, i.e. with probability $p_3 = 8/65$. We may then ask, **<u>given</u> that a customer enters the system**, how much time does he/she spends in the system? We may apply Little's Theorem to find $E(T)$, but have to first calculate the **effective arrival rate**. Because arriving customers are turned away with probability p_3, the effective or average arrival is $\lambda_a = \lambda(1 - p_3) = 2(57/65) = 114/65$.

$$E(T) = \frac{E(N)}{\lambda_a} = \frac{66}{65} / \frac{114}{65} = \frac{11}{19} = .579$$

$$E(W) = \frac{E(N_q)}{\lambda_a} = \frac{28}{65}\frac{65}{114} = \frac{14}{57} = .246 \ \left(Another\ way{:}\ E(W) = E(T) - E(service) = \frac{11}{19} - \frac{1}{\mu} = \frac{14}{57}\right)$$

Suppose an arriving customer finds there are $N = m$ people in the system ($m \le 2$), then the time he/she spends in the system is the sum of $m + 1$ independent exponential distributions, which is an Erlang distribution. To obtain the pdf for T, we sum the conditional pdf's over all values of m, weighted by $P(N = m)$.

$$f_T(t) = \frac{\sum_{m=0}^{2} \frac{(\mu t)^m}{m!} \mu e^{-\mu t} P(N = m)}{P(customer\ enters\ the\ system)} = \frac{65}{57}\left(\left(\frac{27}{65}\right)3e^{-3t} + \left(\frac{18}{65}\right)3^2 t e^{-3t} + \left(\frac{12}{65}\right)\frac{3^3 t^2 e^{-3t}}{2}\right)$$

$$= \frac{27e^{-3t}}{19}(1 + 2t + 4t^2)$$

The probability distribution for W is $P(W = 0) = p_0/(1 - p_3) = 9/19$, for $W > 0$,

$$f_W(t) = \frac{\sum_{m=1}^{2} \frac{(\mu t)^{m-1}}{(m-1)!} \mu e^{-\mu t} P(N = m)}{P(customer\ enters\ the\ system)} = \frac{65}{57}\left(\left(\frac{18}{65}\right)3e^{-3t} + \left(\frac{12}{65}\right)3^2 t e^{-3t}\right) = \frac{18e^{-3t}}{19}(1 + 2t)$$

What if we have **multiple servers**? Let's first consider M/M/c, which has infinite capacity and c servers. Suppose the number of customers does not exceed c, then all customers in the system are being served. If there is one person in the system, the rate of going down to 0 people is the service rate μ. If there are two people, the system to go down by one with the rate 2μ, because when we have two independent Poisson processes with rates μ_a and μ_b, the merged process is a Poisson process with rate $\mu_a + \mu_b$. Another way to get this is to recall that the minimum of two exponential variables with rate μ is an exponential variable with rate 2μ. Similarly, if there are 3 people, the system goes down to 2 with the rate of 3μ. In general, when k people are being served, we have k independent Poisson processes in parallel and time of the first person finishes being served is minimum of k exponential variables each with rate μ, and the minimum time is an exponential with rate $k\mu$. Suppose the number of customers exceeds c, then c customers are served (the others are waiting in line), and the rate of going down by one is $c\mu$ The rate transition diagram is as follows.

$\mu \quad \leftarrow 2\mu \quad \leftarrow 3\mu \quad \leftarrow c\mu \quad c\mu \quad \rightleftarrows \quad \leftarrow$

$$p_1 = \frac{\lambda}{\mu} p_0 = \rho p_0 \qquad \rho = \frac{\lambda}{\mu}$$

$$p_2 = \frac{\lambda}{2\mu} p_1 = \frac{\rho^2}{2} p_0, \quad p_3 = \frac{\lambda}{3\mu} p_2 = \frac{\rho^3}{3!} p_0, \quad \cdots, \quad p_k = \frac{\rho^k}{k!} p_0, \quad \cdots, \quad p_c = \frac{\rho^c}{c!} p_0$$

$$p_{c+k} = \frac{\rho^{c+k}}{c!\, c^k} p_0 = p_c \left(\frac{\rho}{c}\right)^k = p_c \varrho^k, \quad \varrho = \frac{\rho}{c}$$

$$\sum_{j=0}^{\infty} p_j = \sum_{j=0}^{c-1} p_j + \sum_{j=c}^{\infty} p_j = p_0 \sum_{j=0}^{c-1} \frac{\rho^k}{k!} + \frac{\rho^c}{c!} p_0 \sum_{k=0}^{\infty} \varrho^k = p_0 \left(\sum_{j=0}^{c-1} \frac{\rho^k}{k!} + \frac{\rho^c}{c!\,(1-\varrho)} \right)$$

$$\sum_{j=0}^{\infty} p_j = 1 \;\Rightarrow\; p_0 = \left(\sum_{j=0}^{c-1} \frac{\rho^k}{k!} + \frac{\rho^c}{c!\,(1-\varrho)} \right)^{-1}$$

The distribution for N is given by p_k's: $P(N = k) = p_k$. We can proceed to calculate $E(N) = \sum k p_k$. However, that can be a little complicated. On the hand, the distribution of N_q is simpler: $P(N_q = 0) = p_0 + \cdots + p_c$, $P(N_q = k) = p_{c+k} = p_c \varrho^k$,

$$E(N_q) = \sum_{k=1}^{\infty} k P(N_q = k) = p_c \varrho \sum_{k=1}^{\infty} k \varrho^{k-1} = p_c \varrho \left(\frac{d}{d\varrho} \sum_{k=0}^{\infty} \varrho^k \right) = p_c \varrho \frac{d}{d\varrho} \left(\frac{1}{1-\varrho} \right) = \frac{p_c \varrho}{(1-\varrho)^2}$$

We can now calculate

$$E(W) = \frac{E(N_q)}{\lambda}$$

$$E(T) = E(W) + \frac{1}{\mu}$$

$$E(N) = \lambda E(T)$$

Finally let's calculate the pdf's for T and W. Again, W is easier. First of all, $P(W = 0) = p_0 + \cdots + p_{c-1}$. For $W > 0$, if there are c people in the system being served by c servers, you need to wait for one of them to be done, and we know that this is a Poisson process with rate $c\mu$ (see discussion on top of page). If there are k people in the waiting line, you need to wait for $k + 1$ people to be done, each time with rate $c\mu$, you have an Erlang distribution,

$$f_W(t | N_q = k) = \frac{(c\mu)^{k+1} t^k}{k!} e^{-c\mu t}, \quad t > 0$$

$$f_W(t) = c\mu e^{-c\mu t} \sum_{k=0}^{\infty} \frac{(c\mu t)^k}{k!} P(N_q = k) = c\mu e^{-c\mu t} \sum_{k=0}^{\infty} \frac{(c\mu t)^k}{k!} p_c \varrho^k = p_c c\mu e^{-c\mu t} \sum_{m=0}^{\infty} \frac{(\mu \rho t)^k}{k!}$$
$$= p_c c\mu e^{-c\mu t} e^{\mu \rho t} = p_c c\mu e^{-\mu c(1-\varrho)t}$$

The pdf for T turns out to be quite complicated. We have $T = W + S$, where $S \sim Exponential(\mu)$ is the service time and is independent of W. We need to consider 2 cases, $W = 0$ and $W > 0$, because if $W = 0$, T is simply S,

$$f_T(t | W = 0) = \mu e^{-\mu t}$$

And if $W > 0$, T is the sum of two exponentials. We use convolution integral to find $f_T(t | W > 0)$, but first note that

$$P(W > 0) = P(N \geq c) = \frac{p_c}{1 - \varrho}$$

$$f_W(t | W > 0) = \frac{f_W(t)}{P(N \geq c)} = (1 - \varrho) c\mu e^{-\mu c(1-\varrho)t}$$

$$f_T(t | W > 0) = \int_{-\infty}^{\infty} f_W(y | W > 0) f_S(t - y | W > 0) dy = \int_0^t (1 - \varrho) c\mu e^{-\mu c(1-\varrho)y} \mu e^{-\mu(t-y)}\, dy$$
$$= (1 - \varrho) c\mu^2 \int_0^t e^{-\mu(c+1-\rho)y} e^{-\mu t}\, dy = -\frac{(1-\varrho) c\mu^2}{\mu(c + 1 - c\varrho)} e^{-\mu(c-1-\rho)y} e^{-\mu t} \Big|_0^t$$
$$= \frac{(1-\varrho) c\mu}{c - 1 - \rho} \left(e^{-\mu t} - e^{-\mu(c-\rho)t} \right)$$

$$f_T(t) = P(W = 0) f_T(t | W = 0) + P(W > 0) f_T(t | W > 0) = \frac{1 - \varrho - p_c}{1 - \varrho} \mu e^{-\mu t} + \frac{p_c c\mu}{c - 1 - \rho} \left(e^{-\mu t} - e^{-\mu(c-\rho)t} \right)$$

The final case is that of multiple servers **and** finite capacity, where we put together what we have learned so far. There are exercises dealing with this in homework. In homework you will also see queues with non-constant arrival rates.

Chapter 10 Brownian Motion

A. Continuous State Processes

Recall that a stochastic process is *a collection or a sequence of random variables indexed by time*, $\{Y(t)\}$. Time may be discrete or continuous, so may the random variables. So far, we have seen processes where time is discrete (Bernoulli