Lecture 7

In this lecture an example of a very simple continuous time Markov chain is examined. The theory of birth-death processes is covered and finally the M/M/1 queue is solved.

A very simple continuous time Markov chain

An extremely simple continuous time Markov chain is the chain with two states 0 and 1. Let λ_0 be the flow rate from zero to one and λ_1 be the flow rate from one to zero. The transition rate matrix is given by

$$\mathbf{Q} = \begin{bmatrix} -\lambda_0 & \lambda_0 \\ \lambda_1 & -\lambda_1 \end{bmatrix}.$$

The probability vector is given by

$$\frac{d\mathbf{f}(t)}{dt} = \mathbf{f}(t)\mathbf{Q}.$$

This gives the two balance equations

$$\begin{aligned} \frac{df_0(t)}{dt} &= -\lambda_0 f_0(t) + \lambda_1 f_1(t) \\ \frac{df_1(t)}{dt} &= -\lambda_1 f_1(t) + \lambda_0 f_0(t). \end{aligned}$$

Now, since there are only states, for any t then $f_0(t) + f_1(t) = 1$. Hence the first equation becomes

$$\frac{df_0(t)}{dt} = \lambda_1 - (\lambda_0 + \lambda_1) f_0(t). \tag{1}$$

A solution is suspected of the form

$$f_0(t) = K + Ce^{-(\lambda_0+\lambda_1)t}.$$

Differentiating and substituting into (1) gives

$$-C(\lambda_0 + \lambda_1)e^{-(\lambda_0+\lambda_1)t} = \lambda_1 - (\lambda_0 + \lambda_1)K - C(\lambda_0 + \lambda_1)e^{-(\lambda_0+\lambda_1)t}.$$

Therefore $K = \lambda_1/(\lambda_0 + \lambda_1)$. To determine the boundary condition then assume $f_0(0) = F_0$. Substituting gives

$$F_0 = \frac{\lambda_1}{\lambda_0 + \lambda_1} + C.$$

Therefore the final equation is

$$f_0(t) = \frac{\lambda_1}{\lambda_0 + \lambda_1} + (F_0 - \frac{\lambda_1}{\lambda_0 + \lambda_1})e^{-(\lambda_0+\lambda_1)t}.$$

The first part of this (the constant) is the equilibrium solution and the second part is a damped term which shows how the probabilities approach the equilibrium.

Note that, in general solving a set of differential equations is not as easy as this. However, solving the balance equations is more practical. In this case the balance equations become

$$\begin{aligned} -\lambda_0\pi_0 + \lambda_1\pi_1 &= 0 \\ \lambda_0\pi_0 - \lambda_1\pi_1 &= 0. \end{aligned}$$

Note that the two equations are obviously linearly dependent (the second is -1 times the first. It will always be the case that one equation produced like this is a linear combination of the others and, as with discrete time Markov chains, it is necessary to use the fact that the π_i sum to zero. This gives

$$\pi_0 + \pi_1 = 1.$$

Solving the simultaneous equations gives,

$$\begin{aligned} -\lambda_0\pi_0 + \lambda_1 - \lambda_1\pi_0 &= 0 \\ \pi_0 &= \frac{\lambda_1}{\lambda_0 + \lambda_1} \\ \pi_1 &= \frac{\lambda_0}{\lambda_0 + \lambda_1} \end{aligned}$$

Note that the equation for π_0 is in agreement with the equation obtained for $f_0(t)$. In general, solving the simultaneous equations to find the π_i will be much easier than solving a set of differential equations to find $f_i(t)$ although, obviously, the latter provides transient information as well as the final equilibrium position.

Birth-Death Processes

It is now time to see how continuous time Markov chains can be used in queuing and, finally, to get some answers for the elusive M/M/1 queue which was the original aim of introducing Markov chains. First it is necessary to introduce one more new concept, the birth-death process.

A birth-death process is a process where the population k may increase or decrease according to certain rules. Specifically, when the population is k then the population may increase to $k+1$ in the manner of a Poisson process with rate λ_k and may decrease to $k-1$ in the manner of a Poisson process with rate μ_k. The parameter λ_k is known as the birth rate for population k and the parameter μ_k is known as the death rate for population k. It is usually assumed that $\mu_0 = 0$ (if there is no population then nobody can die and that k begins at 0 (or some positive integer. Formally, let $B(t, \delta t)$ be the number of births in the interval $[t, t + \delta t$ and $D(t, \delta t)$ be the number of deaths in the same interval. Let $X(t)$ be the population at time t. The birth-death process is the defined by the following constraints.

$$\begin{aligned} &\mathbb{P}\left[B(t,\delta t) = 1 | X(t) = k\right] = \lambda_k \delta t + o(\delta t) \\ &\mathbb{P}\left[B(t,\delta t) = 0 | X(t) = k\right] = 1 - \lambda_k \delta t + o(\delta t) \\ &\mathbb{P}\left[B(t,\delta t) > 1 | X(t) = k\right] = o(\delta t) \\ &\mathbb{P}\left[D(t,\delta t) = 1 | X(t) = k\right] = \mu_k \delta t + o(\delta t) \\ &\mathbb{P}\left[D(t,\delta t) = 0 | X(t) = k\right] = 1 - \mu_k \delta t + o(\delta t) \\ &\mathbb{P}\left[D(t,\delta t) > 1 | X(t) = k\right] = o(\delta t) \end{aligned}$$

The usual manipulations with differential difference equations follow. Let $P_k(t) = \mathbb{P}\left[X(t) = k\right]$. Now

$$\begin{aligned} P_k(t+\delta t) &= (1 - \lambda_k \delta t - \mu_k \delta t) P_k(t) + \lambda_{k-1}\delta t P_{k-1}(t) + \mu_{k+1}\delta t P_{k+1}(t) + o(\delta t) \\ \frac{P_k(t+\delta t) - P_k(t)}{\delta t} &= -(\lambda_k + \mu_k) P_k(t) + \lambda_{k-1} P_{k-1}(t) + \mu_{k+1} P_{k+1}(t) + \frac{o(\delta t)}{\delta t} \\ \frac{dP_k(t)}{dt} &= -(\lambda_k + \mu_k) P_k(t) + \lambda_{k-1} P_{k-1}(t) + \mu_{k+1} P_{k+1}(t). \end{aligned}$$

This obviously suggests a formulation as a continuous time Markov chain. The transition rate matrix is given by

$$\mathbf{Q} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \dots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Note that the λ_k and μ_k are all constrained to be 0 or larger. If some $\lambda_K = 0$ then K is a maximum population for the system (assuming the population begins below K). If $\lambda_K = 0$ and the population starts in a state below K then no λ_i where $i > K$ need be defined since these states will never be reached. This can be thought of as a birth-death process with a finite population. If some $\mu_K = 0$ then the population will never fall below K once it reaches K.

A number of interesting (well, to queuing theory people anyway) results are available from correct choice of these parameters. If we consider the population k as the number of people in the queue, a birth as an arrival in the queue and a death as a departure from the queue then this is a practical (if somewhat dramatic) way to model queuing processes.

The simplest queue, the M/M/1 queue is simply the birth death process with the birth rate and death rate constant $\lambda_i = \lambda$ for all $i \in \mathbb{Z}^+$ and $\mu_i = \mu$ for all $i \in \mathbb{N}$ with $\mu_0 = 0$ as usual. The rate λ represents the arrival rate at the queue and the rate μ represents the service rate.

Aside: The "pure birth" process

First, however, consider the "pure birth" process where $\lambda_i = \lambda$ for all $i \in \mathbb{Z}^+$ and $\mu_i = 0$ for all $i \in \mathbb{Z}^+$. Using

$$\frac{d\mathbf{f}(t)}{dt} = \mathbf{f}(t)\mathbf{Q},$$

then

$$\begin{aligned} \frac{df_i(t)}{dt} &= \lambda f_{i-1}(t) - \lambda f_i(t) \qquad i > 0 \\ \frac{df_0(t)}{dt} &= -\lambda f_0(t). \end{aligned}$$

Now, assuming that the system is initially empty ($f_0(0) = 1$ and $f_i(0) = 0$ for $i > 0$) then the process begins to look familiar. From the equations above then it can be shown that

$$\begin{aligned} f_0(t) &= e^{-\lambda t} \\ f_1(t) &= (-\lambda t)e^{-\lambda t} \\ f_2(t) &= \frac{(-\lambda t)^2 e^{-\lambda t}}{2}, \end{aligned}$$

and so on. If $X(t)$ is the state of the chain at time t and $X(0) = 0$ then the familiar equation

$$\mathbb{P}\left[X(t) = n\right] = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

with $n \in \mathbb{Z}^+$ is recovered.

Solving the general birth death process

Usually the transient behaviour of the system is of less interest than the equilibrium positions. Assuming that the system is ergodic then the equation

$$\boldsymbol{\pi}\mathbf{Q} = 0,$$

can be used to discover that

$$\begin{aligned} \lambda_{i-1}\pi_{i-1} + \mu_{i+1}\pi_{i+1} &= \lambda_i\pi_i + \mu_i\pi_i \qquad i > 0 \\ \mu_1\pi_1 &= \lambda_0\pi_0. \end{aligned}$$

Taking the second of these equations gives

$$\pi_1 = \frac{\pi_0\lambda_0}{\mu_1}.$$

Taking $i = 1$ in the first of them gives

$$\lambda_0\pi_0 + \mu_2\pi_2 = \lambda_1\pi_1 + \mu_1\pi_1,$$

Substituting the equation for π_1 and rearranging gives

$$\pi_2 = \frac{\lambda_1\lambda_0\pi_0}{\mu_2\mu_1}.$$

A similar process gives

$$\pi_3 = \frac{\lambda_2\lambda_1\lambda_0\pi_0}{\mu_3\mu_2\mu_1}.$$

This leads to the suspicion that the general form is

$$\pi_k = \pi_0 \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i}. \tag{2}$$

Proving this is left as an exercise for the student (quite literally, it is in Worksheet two for this course).

Equation (2) can obtain an expression for any π_k in terms of π_0 but it is still necessary to find an expression for π_0. This can be done using the fact that the sum of all the π_k must be 1. Therefore

$$\begin{aligned} \pi_0 &= 1 - \sum_{k=1}^{\infty} \pi_k \\ &= 1 - \sum_{k=1}^{\infty} \pi_0 \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i}. \end{aligned}$$

This can be rearranged to give

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i}}. \tag{3}$$

These two equations (2) and (3) can be used to find the equilibrium probabilities for any birth-death process providing the product can be evaluated. To do this it is valuable to introduce the concept of utilisation.

Utilisation

Utilisation is defined as the proportion of the queues capacity which is being used. However, this definition is somewhat vague. The utilisation, ρ is given by the equation

$$\rho = \frac{\lambda}{\mu},$$

where λ is the mean arrival rate for the system and μ is the maximum service rate. A clear definition of μ is hard to obtain. In cases where there is some N such that $\mu_i = C$ for all $i > N$ it is clear that $\mu = C$ (this is the case for the M/M/1 model). In cases where the birth death process has some maximum population K and $\mu_i \geq \mu_{i-1}$ for $0 < i \leq K$ then $\mu = \mu_K$. In cases where μ_i increases and does not reach a limit then μ is infinite. Fortunately, in most of the cases dealt with in this course the value to use will be clear. In the $M/M/1$ system $\rho = \lambda/\mu$.

The point where $\rho = 1$ is a criticla point for a queuing system as this is the point where all the traffic that could possibly be handled is arriving. As will be seen, this is a point where the system breaks down. If $\rho \geq 1$ then an infinite birth-death process is not ergodic.

Finally, the M/M/1 model solved

Getting a full solution to the M/M/1 model (including transients) is difficult but the equilibrium solution is easy to find. For the M/M/1 process the arrival process is a constant Poisson process with a rate λ and the server process is a Poisson process with rate μ. Substituting into the birth death equations gives

$$\pi_k = \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} \pi_0 = \rho^k \pi_0$$

for $k \in \mathbb{N}$ with $\rho = \lambda/\mu$ as the utilisation. Using the second equation for birth death processes gives

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \rho^k} = \frac{1}{1 + \rho/(1-\rho)} = 1 - \rho.$$

This can be though of as the probability that the queue is empty and as can be seen this falls linearly from 1 (when the system utilisation is zero, that is no arrivals occur and the system is always empty) to 1 (when the system is at full utilisation and the queue never falls to zero).

The obvious final step in the solution is to calculate the expected queue length when the system is in equilibrium. Call this $\mathrm{E}[q]$, it is given by

$$\mathrm{E}[q] = \sum_{i=0}^{\infty} i\pi_i = \sum_{i=0}^{\infty} i(1-\rho)\rho^i.$$

A nice trick can be used here. First rewrite as

$$\begin{aligned}
\sum_{i=0}^{\infty} i(1-\rho)\rho^i &= (1-\rho)\rho \sum_{i=0}^{\infty} i\rho^{i-1} \\
&= (1-\rho)\rho \sum_{i=0}^{\infty} \frac{d\rho^i}{di} \\
&= (1-\rho)\rho \frac{d}{di} \sum_{i=0}^{\infty} \rho^i \\
&= (1-\rho)\rho \frac{d}{di} \frac{1}{1-\rho} \\
&= \frac{\rho}{1-\rho}.
\end{aligned}$$

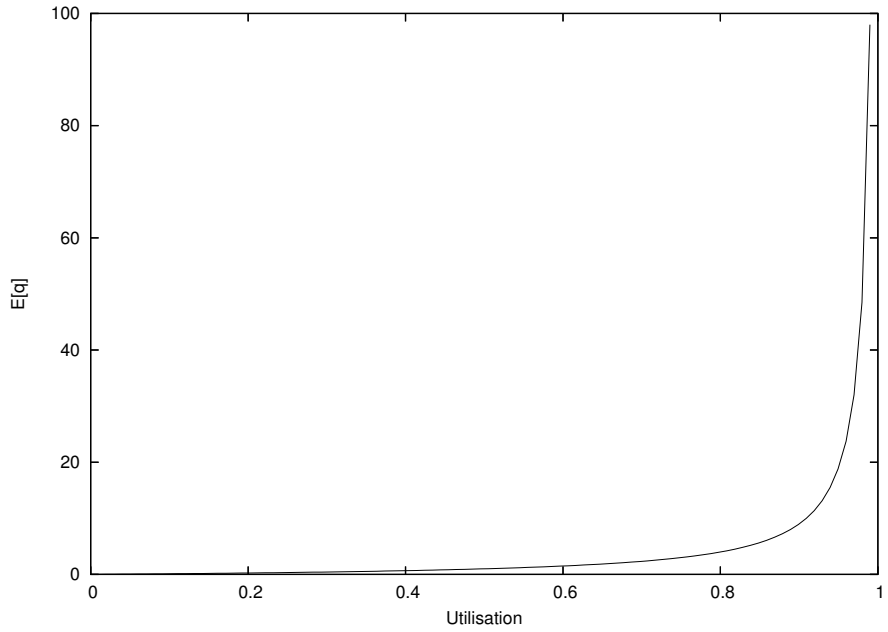

Figure 1: Utilisation versus expected queue size for an M/M/1 queue.

The expected queue length begins at 0 and heads to infinity as ρ approaches one. (Remember that this solution is only valid for an ergodic chain with $\rho < 1$.)

Little's Theorem will give the average delay as

$$T = \frac{\rho}{\lambda(1-\rho)} = \frac{1/\mu}{1-\rho} = \frac{1}{\mu - \lambda}.$$