QUEUING MODELS AND MARKOV PROCESSES

Queues form when customer demand for a service cannot be met immediately. They occur because of fluctuations in demand levels so that models of queuing are intrinsically stochastic.

Some definitions

- The number of servers is s
- The mean arrival rate (number of customers per unit of time) is λ
    - this is assumed to follow a Poisson distribution

$$P_n = \frac{\lambda n e^{-\lambda}}{n!}$$

    where P_n is the probability of n arrivals in the time period.

- The mean service rate (number of customers served per unit time per server) is μ
    - this is assumed to follow an exponential distribution

$$P(t) = 1 - e^{-\mu t}$$

    where P(t) is the probability of being served by time period t.

- We assume independence of the two processes λ and μ and require the condition that $s * \mu > \lambda$, otherwise the queue will grow indefinitely.
- The average service time is given by $1 / \mu$.

Costs of Queuing

Waiting Cost (WC)

This depends on the average time spent waiting in line

$$\mathbf{WC = C_w\, L}$$

where **L** is the average number of customers with $\partial \mathbf{L}\,/\partial \mathbf{s} < 0$, and $\mathbf{C_w}$ is the waiting cost per customer per unit of time.

Service Cost (SC)

This is directly proportional to the number of servers s.

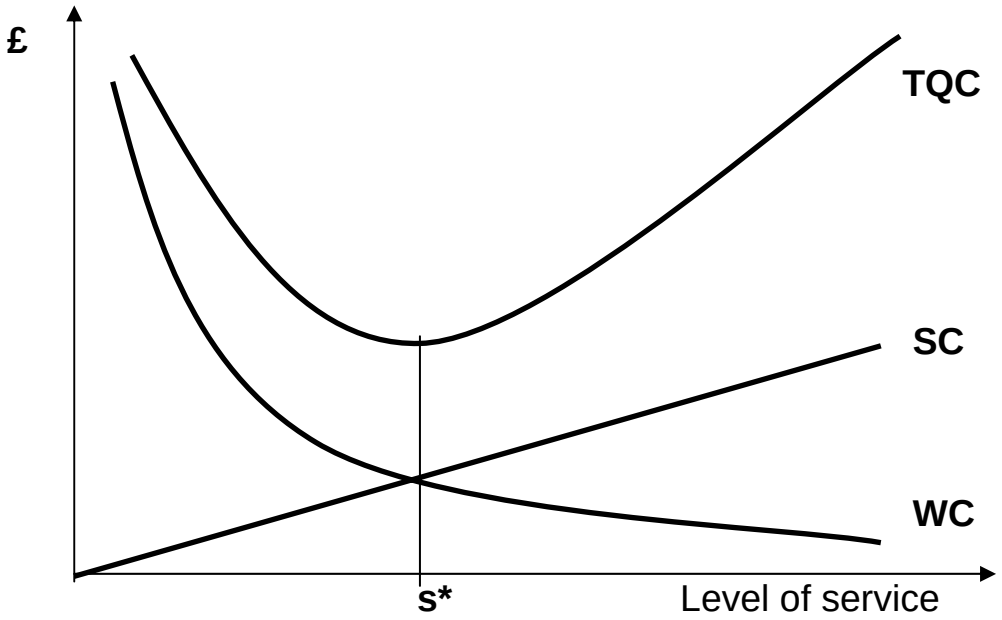
$$\mathbf{SC = C_s\, s}$$

where $\mathbf{C_s}$ is the cost per customer per unit of time. Increasing the number of servers s decreases **WC** but increases **SC**.

Total Cost (TQC)

$$\mathbf{TQC = WC + SC}$$

Total costs are nonlinear in **s** with a minimum at **s***, the optimal level of service.



A Simple Model with One Server

Consider the simplest model with s=1. Then the following results hold:

- the probability of the system being busy: $\rho = \lambda / \mu$.
- the probability of n customers: $P_n = (1-\rho)\,\rho^n$.
- the average number of customers: $L = \lambda / (\mu - \lambda)$.
- the average time spent in the system: $W = L / \lambda$
- the average time spent queuing: $W_q = \lambda / [\mu\,(\mu - \lambda)] = W - 1/\mu$.

A general result in a steady state queuing process is that $L = \lambda W$. This relationship is known as Little's law.

A Model with s Servers

Consider a model with s servers. Then the following results hold:

- the probability of the system being busy: $\rho = \lambda /(s\,\mu)$.
- the probability of 0 customers P_0 is given by the formula

$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \left(\frac{(\lambda/\mu)^s}{(s-1)!}\right)\left(\frac{\mu}{s\mu - \lambda}\right)}$$

- the probability of n customers: P_n is given by

$$P_n = \frac{(\lambda/\mu)^n}{s!\,s^{n-s}} P_0 \quad \text{for} \quad n > s \quad \text{and}$$

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0 \quad \text{for} \quad 0 < n \le s.$$

- the average number of customers: L is given by

$$L = \frac{(\lambda/\mu)^s \lambda\mu}{(s-1)!(s\mu - \lambda)^2} P_0 + \frac{\lambda}{\mu}$$

- the average time spent in the system: $W = L / \lambda$

- the average time spent queuing: $W_q = W - 1/\mu$.

As before, note that Little's law holds and $L = \lambda W$.

MARKOV PROCESSES

Markov processes are used to describe a system moving over time between different states.

Suppose that there are n states: S_1, ..., S_n
and the probability of being in a state at time t: $p_1(t)$, ..., $p_n(t)$

The probabilities are assumed to change over time following a simple stochastic process.

The probabilities $p_1(t)$, ..., $p_n(t)$ can be represented in a row vector

$$\mathbf{p}(t) = [\, p_1(t) \ldots p_n(t) \,].$$

Transitional Probability

The transitional probability $P_{ij}(t)$ is the probability of moving from state i to state j at time t. This is the conditional probability

$$P_{ij}(t) = p_j(t+1) \mid p_i(t)$$

A crucial assumption of the Markov model is that transitional probabilities are independent of time so that

$$P_{ij} = p_j(t+1) \mid p_i(t)$$

The transitional probabilities can be represented in a transition matrix of dimension n x n:

$$\mathbf{P} = \begin{bmatrix} P_{11} & \ldots & P_{1n} \\ & P_{ij} & \\ P_{n1} & \ldots & P_{nn} \end{bmatrix}$$

The elements in each row of the matrix **P** must sum to unity since, whatever state you are in at time t, you must end up in one of the n states in period t+1.

The First Order Markov Process

The simplest Markov process is the first order process defined by the matrix equation

$$\mathbf{p}(t+1) = \mathbf{p}(t)\ \mathbf{P}$$

or, equivalently,

$$p_j(t+1) = \sum_i p_i(t)\ P_{ij}.$$

The first order Markov process is called a *zero memory process* because the state next period depends only on the current state and not on any past states.

Higher Order Markov Processes

More generally, a k^{th} order Markov process can be defined by

$$\mathbf{p}(t+1) = \mathbf{p}(t)\ \mathbf{P}_1 + \mathbf{p}(t-1)\ \mathbf{P}_2 + \ldots + \mathbf{p}(t-k+1)\ \mathbf{P}_k$$

where $\mathbf{P}_1$, $\mathbf{P}_2$, . . . , $\mathbf{P}_k$ are n x n transition matrices. The k^{th} order Markov process depends on the state as far as k-1 periods back.

Predicting Future States

Starting from an initial state p(0) at time t=0 and assuming a first order Markov process, the state at time t=1 is determined by the matrix equation

$$\mathbf{p}(1) = \mathbf{p}(0)\ \mathbf{P}\ .$$

For period t=2 we have

$$\mathbf{p}(2) = \mathbf{p}(1)\ \mathbf{P} = \mathbf{p}(0)\ \mathbf{P}^2$$

where $\mathbf{P}^2 = \mathbf{P} * \mathbf{P}$ is the matrix product of **P** with itself.

In general, by substitution, it can be seen that

$$\mathbf{p}(t) = \mathbf{p}(0)\ \mathbf{P}^t.$$

The Steady State of a Markov Process

Most Markov processes eventually converge to a steady state. This is because in general

Limit $_{t\to\infty}$ $\mathbf{P}^t$ converges to a fixed matrix.

When this is true, after a certain time **p**(t) will not change anymore so that **p**(t) = **p**(t-1) = **p*** and **p*** will satisfy the equation

$$\mathbf{p^*} = \mathbf{p^*}\,\mathbf{P} .$$

The steady state, if it exists, can be computed analytically by solving this system of n equations subject to the adding-up restriction that the elements of **p*** must sum to one, or formally,

$$\sum_i p^*_i = 1.$$

This is a system of n+1 equations in n unknowns but it can be solved by dropping one of the equations.

Special case: Absorbing states

An absorbing state is a state from which, once the state is entered, there is no possibility of exit. An absorbing state is analogous to a black hole in astrophysics. In the transition matrix, an absorbing state will have a row with a one on the diagonal and elsewhere zeroes.

When a transition matrix has absorbing states, then in the limit, everything will end up in these states.

An Example

Suppose the transitional matrix is given by

$$\mathbf{P} = \begin{bmatrix} 0.6 & .0.2 & 0.2 \\ 0.2 & 0.7 & 0.1 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

with initial state

$$\mathbf{p}(0) = [\ 0.324\ 0.441\ 0.235]\ .$$

Then

$$\mathbf{p}(1) = \mathbf{p}(0)\ \mathbf{P}$$

$$= [\ 0.330\ 0.444\ 0.226]\ .$$

and

$$\mathbf{p}(2) = \mathbf{p}(1)\ \mathbf{P} = \mathbf{p}(0)\ \mathbf{P}^2$$

$$= [\ 0.332\ 0.445\ 0.224]\ .$$

Continuing to project forwards

$$\mathbf{p}(11) = \mathbf{p}(0)\ \mathbf{P}^{11}$$

$$= [\ 0.333\ 0.444\ 0.222]$$

and

$$\mathbf{p}(12) = \mathbf{p}(0)\ \mathbf{P}^{12}$$

$$= [\ 0.333\ 0.444\ 0.222]$$

at which point the process has converged to a steady state (to 3 decimal places) so that

$$\mathbf{p}(t > 12) = \mathbf{p}^* = [\ 0.333\ 0.444\ 0.222]\ .$$