EM 509: Individual Project

Queueing Theory

H. M. P. Bandara (E/14/034)

Introduction

Queueing theory is the mathematical study of waiting lines, or queues. It focuses on a class of models in which customers arrive in some random manner at a service facility. Upon arrival they are made to wait in queue until it is their turn to be served. Once served, they are generally assumed to leave the system.

Classification of Queueing Systems

The simplest form of Kendall's notation consists of 3 letters: a/b/c where

a represents the probability distribution of customer's arrivals
b represents the probability distribution of service time
c is the number of servers
Commonly used letters for *a* or *b* are,

- M to indicate Poisson process (Poisson distribution for arrival and Exponential distribution for service time). M represents Markovian.
- E to indicate Erlang distribution
- D to indicate Deterministic or constant distribution
- G to indicate General probability distribution with known mean and variance

Here the capacity for waiting and servicing in the system is assumed to be infinite in this study.

Some fundamental quantities of interest for queueing models are

L - the average number of customers in the system

L_Q - the average number of customers waiting in queue

W - the average amount of time a customer spends in the system

W_Q - the average amount of time a customer spends waiting in queue

In a general sense, the main interest in any queuing model is the number of customers in the system as a function of time, and in particular, whether the servers can adequately handle the flow of customers.

In this text, only the Discrete-time Queueing chains are focused upon. The distinction between the discrete-time Markov chain (DTMC) and CTMC is that in DTMC there is a "jump" to a new state at times, but in CTMC the "jump" to a new state may occur at any time.

Our main assumptions are as follows:

a. If the queue is empty at a given time, then a random number of new customers arrive at the next time.

b. If the queue is nonempty at a given time, then one customer is served and a random number of new customers arrive at the next time.

c. The number of customers who arrive at each time period forms an independent, identically distributed sequence.

Thus, let X_n denote the number of customers in the system at time $n \in N$, and let U_n denote the number of new customers who arrive at time $n \in N+$. Then $U = (U_1, U_2, \ldots)$ is a sequence of independent random variables, with common probability density function f on N, and

$$X_{n+1} = \begin{cases} U_{n+1}, & X_n = 0 \\ (X_n - 1) + U_{n+1}, & X_n > 0 \end{cases}, \quad n \in \mathbb{N}$$

$X = (X_0, X_1, X_2 \ldots)$ is a discrete-time Markov chain with state space N and transition probability matrix P given by

$P(0,y) = f(y), \qquad y \in N$

$P(x,y) = f(y-x+1), \qquad x \in N+, y \in \{x-1, x, x+1, \ldots\}$

The chain X is the **queueing chain** with arrival distribution defined by f.

The Markov property and the form of the transition matrix follow from the construction of the state process X in term of the IID sequence U. Starting in state 0 (an empty queue), a random number of new customers arrive at the next time unit, governed by the Probability Distribution Function f. Hence the probability of going from state 0 to state y in one step is $f(y)$. Starting in state $x \in N+$, one customer is served and a random number of new customers arrive by the next time unit, again governed by the PDF f. Hence the probability of going from state x to state $y \in \{x-1, x, x+1, \ldots\}$ is $f[y-(x-1)]$. The current value is enough to determine the distribution of the next state. I.e. The state only depends on whether a customer has entered / left the queue from the immediate previous state and not the conditions of the earlier states. Therefore the Markovian property of Queueing models is justified.

Using the birth (arrival)–death (departure) terminology, when the population size is n, let λ_n and μ_n be the infinitesimal transition rates (generators) of birth and death, respectively. When the population is the number of customers in the system, λ_n and μ_n indicate that the arrival and service rates depend on the number in the system. Based on the properties of the Poisson process, i.e., when arrivals are in a Poisson process and service times are exponential, we can make the following probability statements for a transition during $(t, t + \Delta t]$:

birth ($n \geq 0$):

$$P \text{ (one birth)} = \lambda n \Delta t + o(\Delta t),$$
$$P \text{ (no birth)} = 1 - \lambda n \Delta t + o(\Delta t),$$
$$P \text{ (more than one birth)} = o(\Delta t),$$

death ($n > 0$):

$$P \text{ (one death)} = \mu n \Delta t + o(\Delta t),$$
$$P \text{ (no death)} = 1 - \mu n \Delta t + o(\Delta t),$$
$$P \text{ (more than one death)} = o(\Delta t),$$

where $o(\Delta t)$ is such that $\frac{o(Dt)}{Dt} \rightarrow 0$ as $\Delta t \rightarrow 0$.

Let $Q(t)$ be the number of customers in the system at time t.

Define

$P_n(t) = P[Q(t) = n \mid Q(0) = i$

Incorporating the probabilities for transitions during (t, t +Δt], as stated above, we get

$$P_{n,n+1}(\Delta t) = \lambda_n \Delta t + o(\Delta t), \qquad n = 0, 1, 2,\ldots,$$
$$P_{n,n-1}(\Delta t) = \mu_n \Delta t + o(\Delta t), \qquad n = 1, 2, 3,\ldots,$$
$$P_{nn}(\Delta t) = 1 - \lambda_n \Delta t - \mu_n \Delta t + o(\Delta t), \qquad n = 1, 2, 3,\ldots,$$
$$P_{nj}(\Delta t) = o(\Delta t), \qquad j \neq n - 1, n, n + 1.$$

The infinitesimal transition rates above lead to the following generator matrix for the birth-and-death process model of the queueing system:

$$\mathbf{A} = \begin{bmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & & \\ & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}.$$

The generator matrix A leads to the following forward Kolmogorov equations for $P_n(t)$

$$P_0'(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t),$$
$$P_n'(t) = -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t), \quad n = 1, 2, \ldots.$$

Unfortunately, even in simple cases such as $\lambda_n = \lambda$ and $\mu_n = \mu$, that is when the arrivals are Poisson and service times are exponential (M/M/1 queue), deriving $P_n(t)$ explicitly is an arduous process. Furthermore in most of the applications the need for knowing the time-dependent behaviour is not all that critical. The most widely used result, therefore, is the limiting result, determined from by letting t $\rightarrow \infty$.

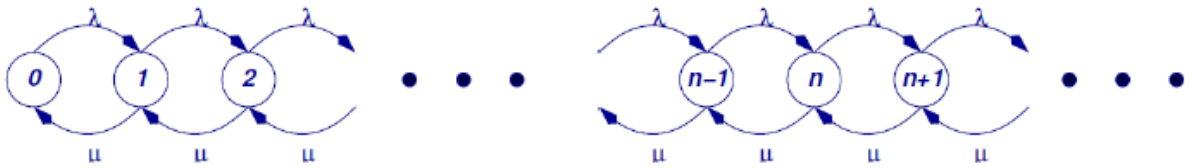

Figure 1: State transition diagram for M/M/1 queue

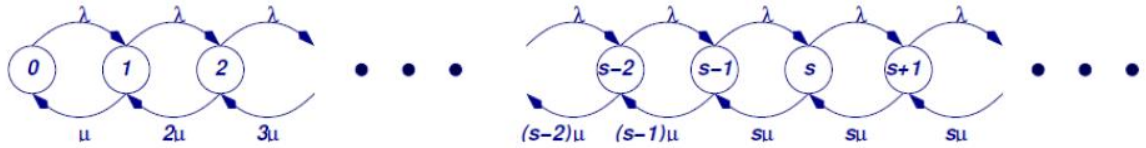

Figure 2: State transition diagram for M/M/s queue

Irreducibility

Consider $X = (X_0, X_1, X_2 \ldots)$ defined above. Let m be the average number of new customers who arrive during a time period.

We can say that the chain X is **irreducible** and **aperiodic.**

Proof:

In a positive state, the chain can move at least one unit to the right and can move one unit to the left at the next step. From state 0, the chain can move two or more units to the right or can stay in 0 at the next step. Thus, **every state leads to every other state** so the chain is irreducible. Since 0 leads back to 0, the chain is aperiodic.

Recurrent and Transient states

Let *m* denote the mean of the arrival distribution, so that

$$m = \sum_{x=0}^{\infty} x\, f(x)$$

Thus *m* is the average number of new customers who arrive during a time period

Let *q* be the probability that the queue eventually empties, starting with a single customer.

The parameter *q* satisfies the equation:

$$q = \sum_{x=0}^{\infty} f(x) q^x$$

Consider the equation $\Phi(q) = q$ where Φ is the **probability generating function** of the distribution that governs the number of new customers that arrive during each period.

q is the smallest solution in (0,1] of the equation $\Phi(t) = t$. Moreover

a. If $m \leq 1$ then $q = 1$ and the chain is recurrent.

b. If $m > 1$ then $0 < q < 1$ and the chain is transient.

Note that the condition in (a) means that on average, one or fewer new customers arrive for each customer served. The condition in (b) means that on average, more than one new customer arrives for each customer served.

Also if $m = 1$ then the queueing chain is null recurrent. Since *m* is the expected number of new customers who arrive during a service period, the results are certainly reasonable.

Stationary distribution

Theorem

The limiting distribution of a positive recurrent irreducible Markov process is also stationary.

A process is said to be stationary if the state distribution is independent of time;

i.e., if

$$P_n(0) = p_n,\ n = 0, 1, 2,...,$$

then

$$P_n(t) = p_n \text{ for all t.}$$

Since we deal with transition distributions conditional on the initial state in stochastic processes, the stationarity means that if we use the stationary distribution as the initial state distribution, from then on all time-dependent distributions will be the same as the one we started with.

In an irreducible, aperiodic, and positive recurrent Markov chain, the limiting probabilities $\{\pi i, i = 0, 1, 2,...\}$ satisfy the equations

$$\Pi j = \sum_{i=0}^{\infty} \Pi j\ Pij$$

for j = 0, 1, 2,...,

$$\sum_{j=0}^{\infty} \Pi j\ = 1$$

The limiting distribution is stationary.

Applications

Real-life applications of queuing theory cover a wide range of applications, such as how to provide faster customer service, improve traffic flow, efficiently ship orders from a warehouse, and design of telecommunications systems, from data networks to call centers. Some of the applications of queueing theory are listed below.

1) Decision making in service facilities

Typically operation of a queuing system typically involves the following decisions.

I. No. of servers at a service facility
II. Efficiency of the servers
III. No. of service facilities

All the decisions above generally lead to the question of appropriate level of service to provide in a queueing system. This level of service generally depends on two considerations

1) the cost incurred by providing the service
2) the amount of waiting for that service

These two considerations create conflicting pressures on the decision maker. The objective of reducing service costs recommends a minimal level of service. On the other hand, long waiting times are undesirable, which recommends a high level of service. Therefore, it is necessary to strive for some type of compromise.

Given that the cost of waiting has been evaluated explicitly, the remainder of the analysis is conceptually straightforward. The objective is to determine the level of service that minimizes the total of the expected cost of service and the expected cost of waiting for that service. This concept is depicted in Fig. 3, where WC denotes waiting cost, SC denotes service cost, and TC denotes total cost. Thus, the mathematical statement of the objective is to

Minimize E(TC) = E(SC) + E(WC).

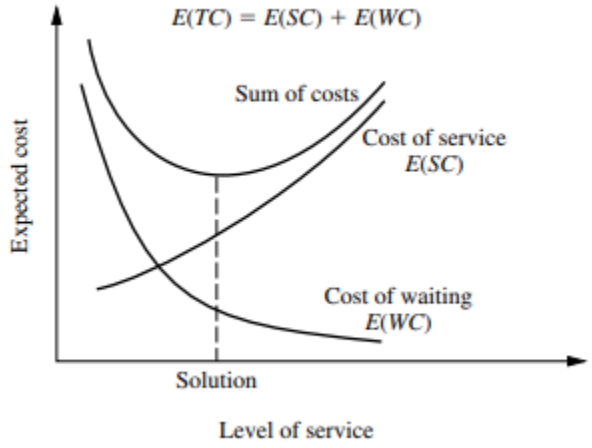

Figure 3: Level of service vs. Expected cost

2) Application of queueing theory in health care

The rising cost of health care can be attributed not only to ageing population and new expensive and advanced treatment modalities but also to inefficiencies in health delivery. Queueing theory application is an attempt to minimise the cost through minimisation of inefficiencies and delays in the system. There are many problems in health care system which can be solved using queueing theory in operational research. Queuing models can be useful in gaining insights on the appropriate degree of specialisation or flexibility to use in organising resources, or on the impact of various priority schemes for determining service order among patients.

References

1. Sheldon M. Ross, Introduction to Probability Models, 10th Edition.
2. U. Narayan Bhat, An Introduction to Queueing Thoery
3. G. f. Newell, Applications of Queueing Theory, 2nd Edition
4. C. Lakshmi & Appa Iyer Sivakumar, Application of queueing theory in health care: A literature review
5. https://people.revoledu.com/kardi/tutorial/Queuing/Kendall-Notation.html
6. https://www.randomservices.org/random/markov/Queuing.html