MARKOV CHAINS AND QUEUEING THEORY

HANNAH CONSTANTIN

Abstract. In this paper, we introduce queueing processes and find the steady-state solution to the $M/M/1$ queue. A brief background in Markov chains, Poisson processes, and Birth-Death processes is also given.

Contents

1. Introduction to Markov Chains

We will briefly discuss finite (discrete-time) Markov chains, and continuous-time Markov chains, the latter being the most valuable for studies in queuing theory.

1.1. **Finite Markov Chains.**

Definition 1.1. Let T be a set, and $t \in T$ a parameter, in this case signifying time. Let $X(t)$ be a random variable $\forall\ t \in T$. Then the set of random variables $\{X(t), t \in T\}$ is called a stochastic process.

We usually interpret $X(t)$ to be the state of the stochastic process at time t. If T is countable, for example, if we let $t = 0, 1, 2, ...$, then we say that $\{X(t), t \in T\}$ is said to be a *discrete-time* process. If, on the other hand, we let T be an interval of $[0, \infty)$, then the stochastic process is said to be a *continuous-time* process. The set of values of $X(t)$ is called the *state space*, which can also be either discrete (finite or countably infinite), or continuous (a subset of $\mathbb{R}$ or $\mathbb{R}^n$).

Date: August 12, 2011.

Definition 1.2. A stochastic process $\{X(n), n \in \mathbb{N}\}$ is called a Markov chain if, for all times $n \in \mathbb{N}$ and for all states $(i_0, i_1, ... i_n)$

$$P\{X_n = i_n \mid X_0 = i_0, ..., X_{n-1} = i_{n-1}\} = P\{X_n = i_n \mid X_{n-1} = i_{n-1}\} \tag{1.3}$$

In other words, given the present state of the system, we may make predictions about its future state without consulting past states.

Equation (1.3) is called the *Markov property*, and in fact, any stochastic process satisfying the Markov property will be a Markov chain, whether it is a discrete-time (as we defined above), or continuous-time process.

We call the conditional probability

$$P\{X_n = j \mid X_{n-1} = i\}, i, j \in S$$

the *transition probability* from state i to state j, denoted by $p_{ij}(n)$.

A Markov chain is called *time-homogeneous* if $p_{ij}(n)$ does not depend on n. In other words,

$$P\{X_n = j \mid X_{n-1} = i\} = P\{X_{n+m} = j \mid X_{n+m-1} = i\}$$

for $m \in \mathbb{N}$ and $m \geq -(n-1)$. In the future, unless otherwise noted, all Markov chains will be assumed to be time-homogeneous and we will denote the transition probability from state i to state j by p_{ij}.

Given the transition probabilities, we can construct the transition matrix P for the Markov chain. P is an $N \times N$ matrix where the (i, j) entry P_{ij} is p_{ij}. In order for a matrix to be the transition matrix for a Markov chain, it must be a stochastic matrix. In other words, it must satisfy the following two properties:

$$0 \leq P_{ij} \leq 1, \quad 1 \leq i, j \leq N \tag{1.4}$$

$$\sum_{j=1}^{N} P_{ij} = 1, \quad 1 \leq i \leq N. \tag{1.5}$$

Given a transition matrix P, an initial probability distribution ϕ where $\phi(i) = P\{X_0 = i\}, i = 1, ..., N$, we can find the probabilities that the Markov chain will be in a certain state i at a given time n. We define the n-step probabilities p_{ij}^n as the following:

$$p_{ij}^n = P\{X_n = j \mid X_0 = i\} = P\{X_{n+k} = j \mid X_k = i\}.$$

The latter part of the equation follows from time-homogeneity. Then we have

$$P\{X_n = j\} = \sum_{i \in S} \phi(i) p_n(i, j) = \sum_{i \in S} \phi(i) P\{X_n = j \mid X_0 = i\}, \tag{1.6}$$

where S is the state space.

Proposition 1.7. *The n-step transition probability $p_n(i, j)$ is actually the (i, j) entry in the matrix P^n.*

Proof. We will prove this by induction. Let $n = 1$. Then, by the definition of the transition matrix P, p_{ij} is the (i, j) entry and our proposition holds. Now, assume it is true for a given n. Then

$$\begin{aligned} p_{ij}^{n+1} &= P\{X_{n+1} = j \mid X_0 = i\} \\ &= \textstyle\sum_k P\{X_n = k \mid X_0 = i\} P\{X_{n+1} = j \mid X_n = k\} \\ &= \textstyle\sum_k p_{ik}^n p_{kj}. \end{aligned}$$

But since p_{ik}^n is the (i,k) entry of P^n by assumption, the final sum is just the (i,j) entry of $P^nP = P^{n+1}$, as we had wanted. □

The initial probability distribution can be written as a vector: $\vec{\phi}_0 = (\phi_0(1), \ldots, \phi_0(N))$. Then, we can find the distribution at time n, $\phi_n(i) = P\{X_n = i\}$:

$$\vec{\phi}_n = \vec{\phi}_0 P^n.$$

Definition 1.8. Two states i and j of a Markov chain *communicate* iff $\exists\ m, n \geq 0$ such that $p_m(i,j) > 0$ and $p_n(j,i) > 0$.

In other words, i communicates with j (written $i \leftrightarrow j$) if one can eventually reach state j starting from state i, and vice versa. Note that the relation $\leftrightarrow$ between i and j is an equivalence relation. We can use this to separate the Markov chain into what are called *communication classes*. These communication classes are disjoint sets that make up the state space, and every state within a given communication class communicates with every other state in the communication class.

Definition 1.9. A Markov chain is *irreducible* if

$$\forall\ i, j\ \exists\ n = n(i,j) \text{ with } \ p_n(i,j) > 0.$$

Simply put, a Markov chain is irreducible if it has only one communication class.

If a Markov chain is not irreducible, we call it a *reducible* chain. If a communication class is such that with probability 1, the Markov chain will eventually leave the class and never return, then it is called a *transient* class with transient states. Classes that are not transient are *recurrent* classes with recurrent states, and once a Markov chain reaches a recurrent class, it always will return to that class.

Definition 1.10. Let P be the matrix for an irreducible Markov chain. (If the Markov chain is reducible, then we can take P for each of the recurrent classes.) The *period* $d = d(i)$ of a state i is defined to be the greatest common divisor of

$$J_i = \{n \geq 0 : p_n(i,i) > 0\}.$$

If a state has period 1, it is called *aperiodic*. An irreducible chain is either aperiodic or has the same period d for all of its states. We now come to our first theorem:

Theorem 1.11. *Let P be the transition matrix for an irreducible, aperiodic Markov chain. Then there is a unique invariant probability vector $\vec{\pi}$ such that*

$$\vec{\pi} P = \vec{\pi}.$$

Also, if $\vec{\phi}$ is an initial probability vector, then

$$\lim_{n \to \infty} \vec{\phi} P^n = \vec{\pi}.$$

Finally, $\pi(i) > 0\ \forall i$.

Theorem (1.11) is sometimes called the *Ergodic Theorem* for Markov chains. Unfortunately, the proof of the Ergodic Theorem is outside the scope of this paper.

Before we can talk about continuous-time Markov chains, we must define and discuss what is called a *Poisson process.*

1.2. **Poisson Process.**

Consider a stochastic process that counts arrivals, $\{N(t), t \geq 0\}$ where $N(t)$ denotes the total number of arrivals up to time t, and $N(0) = 0$. The stochastic process $N(t)$ is considered a *Poisson process with rate parameter* λ if it satisfies the following three conditions:

i. The probability that an arrival occurs between time t and $t + \Delta t$ is $\lambda \Delta t + o(\Delta t)$, where λ is a constant independent of $N(t)$. We introduce the notation $o(\Delta t)$ to denote a function of Δt satisfying

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

ii. The probability that more than one arrival occurs between t and $t + \Delta t$ is $o(\Delta t)$.

iii. The numbers of arrivals in non-overlapping intervals are statistically independent, so that the process has independent increments.

Let $p_n(t)$ be the probability of n arrivals in a time interval of length t, where $n \in \mathbb{N} \cup 0$. Now,

$$p_n(t + \Delta t) = P\{\ n \text{ arrivals in } t \text{ and none in } \Delta t\ \}$$

$$+P\{\ n - 1 \text{ arrivals in } t \text{ and one in } \Delta t\ \}$$

$$+ \cdots + P\{\ \text{ no arrivals in } t \text{ and } n \text{ in } \Delta t\ \}$$

Using assumptions i, ii, and iii, we have

$$p_n(t + \Delta t) = p_n(t)[1 - \lambda \Delta t - o(\Delta t)] + p_{n-1}(t)[\lambda \Delta t + o(\Delta t)] + o(\Delta t), \tag{1.12}$$

where the last $o(\Delta t)$ represents the terms $P\{\ n - j$ arrivals in t and j in $\Delta t\ \}$ where $2 \leq j \leq n$.
For $n = 0$, we have

$$p_0(t + \Delta t) = p_0(t)[1 - \lambda \Delta t - o(\Delta t)]. \tag{1.13}$$

From (1.12) and (1.13), and combining all the $o(\Delta t)$ terms, we have

$$p_0(t + \Delta t) - p_0(t) = -\lambda \Delta t\, p_0(t) + o(\Delta t) \tag{1.14}$$

and

$$p_n(t + \Delta t) - p_n(t) = -\lambda \Delta t\, p_n(t) + \lambda \Delta t\, p_{n-1}(t) + o(\Delta t). \tag{1.15}$$

From (1.14) and (1.15) we can take the limit as $\Delta t \to 0$ and get the differential-difference equations

$$\lim_{\Delta t \to 0} [\frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + \frac{o(\Delta t)}{\Delta t}], \tag{1.16}$$

$$\lim_{\Delta t \to 0}[\frac{p_n(t+\Delta t) - p_n(t)}{\Delta t} = -\lambda p_n(t) + \lambda p_{n-1}(t)\frac{o(\Delta t)}{\Delta t}], \tag{1.17}$$

which simplify to

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) \tag{1.18}$$

and

$$\frac{dp_n(t)}{dt} = -\lambda p_n(t) + \lambda p_{n-1}(t) \qquad (n \geq 1). \tag{1.19}$$

Now we have an infinite set of linear, first-order ordinary differential equations to solve. Equation (1.18) has the general solution $p_0 = Ce^{-\lambda t}$, where $C = 1$ since $p_0 = 0$. Now, let $n = 1$ in (1.19) so that

$$\frac{dp_1(t)}{dt} = -\lambda p_1(t) + \lambda p_0(t),$$

or

$$\frac{dp_1(t)}{dt} + \lambda p_1(t) = \lambda p_0(t) = \lambda e^{-\lambda t}.$$

Solving this equation gets us

$$p_1(t) = Ce^{-\lambda t} + \lambda t e^{-\lambda t}.$$

Since $p_n(0)$ is 0 $\forall n > 0$ we have $C = 0$, and

$$p_1(t) = \lambda t e^{-\lambda t}.$$

Continuing this process for $n = 2$ and $n = 3$ we get

$$p_2(t) = \frac{(\lambda t)^2}{2} e^{-\lambda t}$$

and

$$p_3(t) = \frac{(\lambda t)^3}{3!} e^{-\lambda t}$$

which leads us to make the following proposition.

Proposition 1.20. *The formula for a Poisson probability distribution with a mean arrival rate of λ is*

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

Proof. We shall prove the proposition by induction. We have already shown it holds true for $n = 1$. Now assume it holds true for some $n = k$. We want to show that $p_{k+1}(t) = \frac{(\lambda t)^{k+1}}{(k+1)!} e^{-\lambda t}$. We know that $\frac{dp_{k+1}(t)}{dt} = -\lambda p_{k+1}(t) + \lambda p_k(t)$ from (1.19). Furthermore, we know that $p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$ by assumption. So we have

$$\frac{dp_{k+1}}{dt} + \lambda p_{k+1}(t) = \lambda p_k(t).$$

Multiplying both sides by $e^{\lambda t}$ we get

$$\frac{d}{dt}(e^{\lambda t} p_{k+1}(t)) = \lambda e^{\lambda t} p_k(t)$$

Integrating leaves us with

$$e^{\lambda t} p_{k+1}(t) = \lambda \int_0^t e^{\lambda s} p_k(s) ds$$

so

$$\begin{aligned} p_{k+1}(t) &= e^{-\lambda t}\lambda\int_0^t e^{\lambda s}p_k(s)ds \\ &= e^{-\lambda t}\lambda\int_0^t e^{\lambda s}\frac{(\lambda s)^k}{k!}e^{-\lambda s} \\ &= e^{-\lambda t}\lambda^{k+1}\frac{t^{k+1}}{(k+1)!} \end{aligned}$$

□

1.3. **Continuous-Time Markov Chains.**

Let $\{X(t),\ t \in T\}$ be a continuous-time Markov chain. This means

$$T = \{t : 0 \leq t < \infty\}.$$

Consider any times $s > t > u \geq 0$ and states i, j; then

$$p_{ij}(u,s) = \sum_r p_{ir}(u,t)p_{rj}(t,s), \tag{1.21}$$

where $p_{ij}(u,s)$ is the probability of moving from state i to state j in the time beginning at u and ending at s, and the summation is over the entire state space of the chain. Letting $u = 0$ and $s = t + \Delta t$ gives

$$p_{ij}(0,t+\Delta t) = \sum_r p_{ir}(0,t)p_{rj}(t,t+\Delta t). \tag{1.22}$$

If we let $p_i(0)$ be the probability that the chain starts in state i at time 0 and $p_j(t)$ be the probability that the chain is in state j at time t regardless of starting state, we can multiply (1.22) by $p_i(0)$ and sum over all states to get

$$\sum_i p_i(0)p_{ij}(0,t+\Delta t) = \sum_r\sum_i p_{ir}(0,t)p_i(0)p_{rj}(t,t+\Delta t),$$

or

$$p_j(t+\Delta t) = \sum_r p_r(t)p_{rj}(t,t+\Delta t). \tag{1.23}$$

For the Poisson process treated earlier,

$$p_{rj}(t,t+\Delta t) = \begin{cases} \lambda\Delta t + o(\Delta t) & \text{if } r = j-1, j \geq 1, \\ 1 - \lambda\Delta t + o(\Delta t) & \text{if } r = j, \\ o(\Delta t) & \text{elsewhere.} \end{cases}$$

Substituting this into (1.23) gets us

$$p_j(t+\Delta t) = [\lambda\Delta t + o(\Delta t)]p_{j-1}(t) + [1 - \lambda\Delta t + o(\Delta t)]p_j(t) + o(\Delta t) \qquad (j \geq 1)$$

which is (1.12). Now if the transition probability functions $p(u,s)$ have continuous functions $q_i(t)$ and $q_{ij}(t)$ associated with them so that

$$P\{\text{ a change of state in } (t,t+\Delta t)\ \} = 1 - p_{ii}(t,t+\Delta t) = q_i(t)\Delta t + o(\Delta t) \tag{1.24}$$

and

$$p_{ij}(t,t+\Delta t) = q_{ij}(t)\Delta t + o(\Delta t), \tag{1.25}$$

then we may take partials of Equation (1.21) and use (1.24) and (1.25) to get the Kolmogorov forward and backwards equations:

$$\frac{\partial}{\partial t}p_{ij}(u,t) = -q_j(t)p_{ij}(u,t) + \sum_{r\neq j} p_{ir}(u,t)q_{rj}(t) \tag{1.26}$$

$$\frac{\partial}{\partial u}p_{ij}(u,t) = q_i(u)p_{ij}(u,t) + \sum_{r\neq i} q_{ir}(u)p_{rj}(u,t) \tag{1.27}$$

1.4. **Birth-Death Processes.**

One example of a continuous-time Markov chain is the birth-death process. The birth-death process is a stochastic process with the property that the net change across an infinitesimal time interval Δt is either -1, 0, or 1, and where the state n signifies the size of the population.

Given the size of the population n, the probability of an increase (in an infinitesimal time period) is

$$P\{\text{ increase } n \to n+1 \text{ in } (t, t+\Delta t)\ \} = \lambda_n \Delta t + o(\Delta t) \qquad (n \geq 0),$$

and the probability of a decrease is

$$P\{\text{ decrease } n \to n-1 \text{ in } (t, t+\Delta t)\ \} = \mu_n \Delta t + o(\Delta t) \qquad (n \geq 1).$$

Therefore the probability the population will stay the same is

$$P\{\ n \to n \text{ in } (t, t+\Delta t)\ \} = 1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t)$$

and

$$q_{ij} = \begin{cases} \lambda_i & \text{if } j = i+1 \\ \mu_i & \text{if } j = i-1, \\ \lambda_i + \mu_i & \text{if } j = i \\ 0 & \text{elsewhere.} \end{cases}$$

Substituting for q_i and q_{ij} we get the matrix Q:

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

Using the Kolmogorov forward equation, we can derive the differential-difference equations for the birth-death process:

$$\frac{dp_j(t)}{dt} = -(\lambda_j + \mu_j)p_j(t) + \lambda_{j-1}p_{j-1}(t) + \mu_{j+1}p_{j+1}(t) \qquad (j \geq 1) \tag{1.28}$$

$$\frac{dp_0(t)}{dt} = -\lambda_0 p_0(t) + \mu_1 p_1(t) \tag{1.29}$$

To find the steady-state solution to a birth-death process, we simply set the differential-difference equations to 0 and we find

$$p_{j+1} = \frac{\lambda_j + \mu_j}{\mu_{j+1}}p_j - \frac{\lambda_{j-1}}{\mu_{j+1}}p_{j-1} \qquad (j \geq 1),$$

$$p_1 = \frac{\lambda_0}{\mu_1}p_0.$$

Plugging in $n = 2$ and $n = 3$ gets us

$$\begin{aligned} p_2 &= \frac{\lambda_1 + \mu_1}{\mu_2} p_1 - \frac{\lambda_0}{\mu_2} p_0 \\ &= \frac{\lambda_1 + \mu_1}{\mu_2} \frac{\lambda_0}{\mu_1} p_0 - \frac{\lambda_0}{\mu_2} p_0 \\ &= \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1}, \end{aligned}$$

and

$$\begin{aligned} p_3 &= \frac{\lambda_2 + \mu_2}{\mu_3} p_2 - \frac{\lambda_1}{\mu_3} p_1 \\ &= \frac{\lambda_2 + \mu_2}{\mu_3} \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0 - \frac{\lambda_1 \lambda_0}{\mu 3 \mu_1} p_0 \\ &= \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1}. \end{aligned}$$

Proposition 1.30. *For the steady-state solution to a birth-death process,*

$$p_n = \frac{\lambda_{n-1} \lambda_{n-2} \dots \lambda_0}{\mu_n \mu_{n-1} \dots \mu_1} p_0 = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \qquad (n \geq 1) \tag{1.31}$$

Proof. We shall prove this by induction. We have already shown the proposition holds for $n = 1, 2, 3$. Now, assume it is true for some $n = k$. Then

$$\begin{aligned} p_{k+1} &= \frac{\lambda_k + \mu_k}{\mu_{k+1}} p_0 \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} - \frac{\lambda_{k-1}}{\mu_{k+1}} p_0 \prod_{i=1}^{k-1} \frac{\lambda_{i-1}}{\mu_i} \\ &= \frac{p_0 \lambda_k}{\mu_{k+1}} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} + \frac{p_0 \mu_k}{\mu_{k+1}} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} - \frac{p_0 \mu_k}{\mu_{k+1}} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} \\ &= p_0 \prod_{i=1}^{k+1} \frac{\lambda_{i-1}}{\mu_i} \end{aligned}$$

□

Since probabilities must sum to 1, we can see that

$$p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1} \tag{1.32}$$

From (1.32), it follows that a necessary and sufficient condition for the existence of a steady-state solution is the convergence of the infinite series

$$1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}.$$

2. Basics of Queueing Processes

A queue is a waiting line; queues are formed whenever the demand for service exceeds the service availability. A queuing system is composed of customers arriving for service, waiting in a queue for the service if necessary, and after being served, leaving the system. The term *customer* is generic and does not imply a human customer necessarily; any unit which needs a form of service is considered a customer.

A queueing system is usually described by five basic characteristics of queueing processes: (1) arrival pattern of customers, (2) service pattern of customers, (3) queue discipline, (4) system capacity, and (5) number of service channels.

In most queueing systems, the arrival pattern will be stochastic. We therefore wish to know the probability distribution describing the interarrival times (times between successive customer arrivals), and also whether customers arrive singly or in groups. If an arrival pattern does not depend on time (in other words, the probability distribution that describes the arrival process is time-independent), it is called a *stationary* arrival pattern. An arrival pattern that is time-dependent is called *nonstationary.*

As with arrival patterns, the most important factor in studying service patterns is the probability distribution that describes the sequence of customer service times. Service may be single or batch (for example, people boarding a train or using an elevator). The service process may depend on the number of customers waiting for service: a server may work more efficiently as the queue increases or become flustered and less productive. This is called *state-dependent* service. Service can also be stationary or nonstationary with respect to time. For example, service may exhibit signs of learning, so that over time, the service becomes more efficient.

Unless arrival and service times are deterministic, the queue length will have no definitive pattern. It follows that a probability distribution for queue lengths will depend on two separate processes: the arrival distribution and the service distribution. These are generally assumed to be mutually independent.

Queue discipline indicates the manner in which customers receive service. The most common discipline is first come, first served (FCFS); however, there are other queue disciplines, such as last come, first served (LCFS), or disciplines involving a set of priorities.

A system may have infinite capacity, or there may be a limit to the amount of customers allowed in the queue. In a finite queueing situation, where there is a maximum service size, customers are forced to balk when the queue size is at its limit. If the system has finite capacity, it would be relevant to know this maximum service size.

A system may have a single server or multiple parallel servers. A customer arriving to find more than one free server may choose at random between them to receive service. If all the servers are busy, the customer joins a queue common to all. The customer at the head of the common queue will receive service at the first server available. In other cases, each parallel server may have a queue; that is, there is no common queue (we see this often in supermarkets). It is generally assumed that parallel channels operate independently of each other.

2.1. **Notation.**

The standard notation to describe a queueing process is $A/B/X/Y/Z$, where A indicates the interarrival time distribution, B indicates the probability distribution describing the service time, X is the number of parallel service channels, Y is the restriction on system capacity, and Z is the queue discipline. X and Y can be any member of $[1, \infty)$, but A and B are described by symbols that represent the probability distributions. For example M is an exponential distribution, D is deterministic, and GD is a general distribution. In the general distribution case, there is no assumption made as to the form of the distribution. Results in these cases are therefore applicable to all probability distributions.

In many cases, only the first three symbols are used. If there is no restriction on system capacity, that is, if $Y = \infty$, the standard is to omit the symbol. Another convention is to omit the queue discipline if it is first come, first served (FCFS). For example, M/D/4 would signify a queueing system with exponential interarrival (Poisson input), deterministic service, four servers, no capacity limit, and first come, first served discipline.

2.2. **System Performance.**

In analyzing queueing systems, we need to find ways to gauge system performance in various arenas. In other words, what makes a queueing system effective? There are three characteristics of systems that are of interest. First, a measure of the typical waiting time of a customer; second, the manner in which customers accumulate; and third, a measure of the idle time of the servers. Since these are most likely stochastic processes, we wish to know their probability distributions, or at least their expected values.

There are two types of customer waiting times: the time a customer spends in the queue itself, and the time a customer spends in the entire system (including service). Depending on the system, one of these may be more valuable than the other. Similarly, there are two forms of customer accumulation: the number of customers in the queue and the total number of customers in the system. Measuring idle service can mean either the percentage of time any particular server is without customers, or the time the entire system is idle.

2.3. **General Relationships and Results.**

Here, we will present some general relationships and results for $G/G/1$ and $G/G/c$ queueing systems. Let the average rate of customers entering the system be λ and the average rate of serving customers be μ. Then, a measure of traffic congestion for a system with c servers is $\rho = \lambda/c\mu$. When $\rho > 1$ ($\lambda > c\mu$), the average number of arrivals into the system exceeds the maximum service rate of the system and the queue size increases without bound as time goes by (unless there is a finite system capacity). It follows that when $\rho > 1$, there is no steady-state solution for the queue size. In fact, the only way to find steady-state results is to have $\rho < 1$. When $\rho = 1$, unless arrivals and service are deterministic, no steady state exists, because randomness prevents the queue from ever becoming empty and the servers are constantly behind, causing the queue to increase without bound. Therefore, if one knows the average arrival rate and the average service rate, the minimum number of parallel servers required to have a steady-state solution to the size of the queue can be found by solving for the smallest c such that $\lambda/c\mu < 1$.

We often want to find the probability distribution for the total number $N(t)$ of customers in the system at time t. $N(t)$ is made up of the number of customers waiting in queue, $N_q(t)$, and those in service, $N_s(t)$. Let $p_n(t) = P\{N(t) = n\}$, and $p_n = P\{N = n\}$ in the steady state. Assuming there are c- server queues in the steady state, we can find the mean number in the system,

$$L = E[N] = \sum_{n=0}^{\infty} np_n,$$

and the expected number in the queue,

$$L_q = E[N_q] = \sum_{n=c+1}^{\infty} (n-c)p_n.$$

Among the most important general results in queueing theory are Little's formulas, which relate the steady state mean system sizes with the steady state average customer waiting times. Let T_q be the time a customer spends waiting in queue prior to service, S be the service time, and $T = T_q = S$ be the total time a customer spends in the system. Then the mean waiting time in the queue is $W_q = E[T_q]$, and the mean waiting time in the system is $W = E[T]$. Little's formulas are

$$L = \lambda W \tag{2.1}$$

and

$$L_q = \lambda W_q. \tag{2.2}$$

Hence it is necessary to find only one of the four expected value measures ($E[N]$, $E[N_q]$, $E[T]$, $E[T_q]$), because of Little's formulas and also the relationship $E[T] = E[T_q] + E[S]$ or equivalently, $W = W_q + 1/\mu$, where μ is the mean service rate. From this we can find the relationship between W and W_q as

$$L - L_q = \lambda(W - W_q) = \lambda(1/\mu) = \lambda/\mu.$$

But $L - L_q = E[N] - E[N_q] = E[N - N_q] = E[S]$ so the expected number of customers in service in the steady state is $r = \lambda/\mu$. In a single-server system, $r = \rho$ and

$$L - L_q = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} p_n = 1 - p_0.$$

Let the probability that any given server is busy in a multiserver system in the steady state be p_b. We know that the expected number present in service at any instant in the steady state is r, hence the expected number present at one server is r/c. Now, by an expected-value argument, we see that

$$r/c = \rho = 0 \cdot (1 - p_b) + 1 \cdot p_b.$$

Therefore, $p_b = \rho$. For $G/G/1$, the probability of the system being idle ($N = 0$) is the same as the probability of a server being idle. Thus, $p_0 = 1 - p_b$ in this case, so $p_0 = 1 - \rho = 1 - r = 1 - \lambda/\mu$. The quantity $\lambda/\mu = r$ is sometimes called the *offered load*, since on average, each customer has $1/\mu$ time units of service and the average arrival rate is λ so λ/μ is the amount of work arriving to the system per unit time. If we divide this by the number of servers c, (yielding ρ), we get the average amount of work coming to each server per unit time.

2.4. The $M/M/1$ Model.

The $M/M/1$ model is a Poisson-input, exponential-service, single-server queue. The density functions for the interarrival and service times are given respectively as

$$a(t) = \lambda e^{-\lambda t},$$
$$b(t) = \mu e^{-\mu t},$$

where $1/\lambda$ is the mean interarrival time and $1/\mu$ is the mean service time. Both the interarrival and service times are exponential, and the arrival and conditional service rates are Poisson, which gets us

$$P\{ \text{ an arrival occurs in an interval of length } \Delta t \ \} = \lambda\Delta t + o(\Delta t)$$
$$P\{ \text{ more than one arrival occurs in } \Delta t \ \} = o(\Delta t)$$
$$P\{ \text{ a service completion in } \Delta t \text{ given the system is not empty } \} = \mu\Delta t + o(\Delta t)$$
$$P\{ \text{ more than one service completion in } \Delta t \text{ given more than one in the system } \} = o(\Delta t)$$

The $M/M/1$ model is a simple birth-death process with $\lambda_n = \lambda$ and $\mu_n = \mu$ for all n. Arrivals are "births" to the system, since if the system is in state n (the state refers to the number of customers in the system), an arrival increases it to state $n+1$. Similarly, departures are "deaths", moving from state n to state $n-1$. Hence, the steady state equations are found to be

$$0 = -(\lambda + \mu)p_n + \mu p_{n+1} + \lambda p_{n-1}, \tag{2.3}$$

$$0 = -\lambda p_0 + \mu p_1 \tag{2.4}$$

or

$$p_{n+1} = \frac{\lambda + \mu}{\mu} p_n - \frac{\lambda}{\mu} p_{n-1} \tag{2.5}$$

$$p_1 = \frac{\lambda}{\mu} p_0 \tag{2.6}$$

We are now going to solve the steady-state difference equations for $\{p_n\}$ using the iterative method. Since the $M/M/1$ system is a birth-death process with constant birth and death rates, we can directly apply (1.31) with $\lambda_n = \lambda$ and $\mu_n = \mu$ for all n. It follows that

$$p_n = p_0 \prod_{i=1}^{n} \frac{\lambda}{\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^n.$$

To find p_0, we remember that probabilities must sum to 1 and it follows that

$$1 = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n p_0.$$

Recall that we defined $\rho = \frac{\lambda}{\mu}$ as the traffic intensity for single-server queues. Rewriting, we find

$$p_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n}$$

.

Now, $\sum_{n=0}^{\infty} \rho^n$ is an infinite geometric series which converges if and only if $\rho < 1$. Thus, the only way for a steady-state solution to exist is also if and only if $\rho < 1$. Since we know the sum of the terms of a convergent geometric series,

$$\sum_{n=0}^{\infty} \rho^n = \frac{1}{1-\rho} \qquad (\rho < 1),$$

we find

$$p_0 = 1 - \rho$$

which confirms the general result for p_0 we derived previously in the $G/G/1$ case. Thus the full steady-state solution for the $M/M/1$ system is the geometric probability function

$$p_n = (1-\rho)\rho^n \qquad (\rho = \frac{\lambda}{\mu} < 1). \tag{2.7}$$

Acknowledgements. I would like to thank my mentor, John Wilmes, for his assistance and helpful comments throughout this process.

References

[1] Donald Gross and Carl M. Harris. *Fundamentals of Queueing Theory.* John Wiley and Sons, Inc. 1998

[2] Gregory F. Lawler. *An Introduction to Stochastic Processes.* Chapman and Hall. 1995.

[3] J. Medhi. *Stochastic Models in Queueing Theory.* Elsevier Science. 1991.