QUEUEING THEORY

TANVI DESHPANDE

Abstract. In this paper, we discuss two types of queues: $M/M/1$ and $M/M/\infty$ queues. We first introduce stochastic processes and continuous-time Markov chains before finding the steady-state solutions of the two queues.

1. Introduction to Queues

Definition 1.1. A queue consists of a queueing node with one or more "servers". A customer arrives to the queue, may or may not wait a period of time before being served, then spends some time in the queueing node being "served" by a server, after which they leave the queue.

A queue arises when there are limited resources for a service. The actual queue consists of customers or tasks waiting to be completed and a queueing node with server(s) who complete the tasks. Customers enter and leave the queue in a random fashion, according to some stochastic process.

A queue can be used as a mathematical model for a variety of situations, such as waiting for a cashier at a grocery store or handling customer service calls. There are several defining characteristics of a queue.

Definition 1.2 (Characteristics of Queues)**.** A queue possesses the following characteristics: arrival rate, service pattern/rate, queue discipline, queue capacity, and number of servers.

Queues are well-suited to being represented using a continuous-time Markov chain. Before we can actually define queues mathematically, we must give some background on stochastic processes and continuous-time Markov chains.

This paper contains a mixture of results from [1], [4], [5], and [6].

2. Stochastic Processes and the Poisson Process

Definition 2.1. A stochastic process is a collection of random variables indexed by a set, with elements $\{X(\theta)\}$ for $\theta \in \Theta$. Most commonly, Θ, the index set, is time.

For example, in the context of queueing theory, a stochastic process could denote the number of arrivals or departures (the random variable X) over a certain interval of time (the set which these variables are indexed over). It would make sense for the state space of X, then, to be positive integers, since the number of customers in a queue must be a positive integer.

If the index set Θ is a countable set, the process is known as a discrete-time process, and if it is not, such as the case of the real line, it is known as a continuous-time process. Similarly, if the random variable X has a discrete state space, our stochastic process is a discrete space process, and if not, it is a continuous space process. We therefore have four types of stochastic processes based on these classifications; the one that it makes the most sense for us to deal with is a discrete space continuous-time process, because the number of customers must be discrete and time is generally continuous.

The first stochastic process we will define is a Poisson process, which is the most common stochastic process for arrival/departure rate in queueing theory.

Date: October 21, 2020.

Definition 2.2. A Poisson process is a continuous-time stochastic process in which a function $N(t)$ with $t \in [0, \infty)$ represents the number of times a certain event (such as customers or tasks arriving to a queue) occurs at different points in time.

We refer to the stochastic process $\{N(t), t \geq 0\}$, as the *counting process* which counts the number of occurrences of this event up to time t.

Note that in this case, the index set Θ is the subset of the real line $[0, \infty)$, representing time, and the random variable N_t represents the number of arrivals by time t.

Definition 2.3 (Poisson distribution)**.** We first define a Poisson distribution; the Poisson distribution takes one parameter, λ, representing the mean number of occurrences. Then, we have the probability mass function

$$f(k, \lambda) = \mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

This means that the probability of k arrivals occurring at a given time is equal to $f(k, \lambda)$. The expected value and variance of this distribution is λ.

If we are given the rate r rather than the mean number of occurrences, we have $\lambda = rt$ and thus the probability of k occurrences in time interval t is

$$\frac{(rt)^k e^{-rt}}{k!}.$$

Remark 2.4. The following 3 conditions must be met for $N(t)$ to be a Poisson process with a rate parameter of λ:

(1) $N(0) = 0$.

(2) $N(t)$ has independent increments, meaning that the number of arrivals in disjoint intervals are statistically independent. This means that for $r > s > t > u > 0$, $N(r) - N(s)$ and $N(t) - N(u)$ are independent events, or random variables.

(3) The number of occurrences in an interval of length t follows a Poisson distribution, with a mean value of λt.

In addition to these three conditions, a Poisson process has stationary increments; this means that for $t_1 > t_2 > 0$ and $u > 0$, the random variable $N(t_1) - N(t_2)$ has the same distribution as $N(t_1 + u) - N(t_2 + u)$.

Furthermore, if we choose the quantity $\delta = t_1 - t_2$ to be very small, almost a point in time, the probability of there being an occurrence there is the same for any values of t_2 and t_1, essentially stating that the probability of having an event at any point in time is equally likely.

Remark 2.5. The Poisson process is what we refer to as memoryless and orderly.

A memoryless stochastic process is just a process in which the future does not depend on the past's events. The Poisson process being memoryless is a fact which follows from its independent, stationary increments; the independence from the past follows from the independent increments, and the identical distribution of future events is explained by the stationary increments.

An orderly process is one which satisfies the condition

$$\lim_{\Delta t \to 0} \mathbb{P}(N(t + \Delta t) - N(t) > 1 \,|\, N(t + \Delta t) - N(t) \geq 1) = 0.$$

This is important, because it means that there is a negligible probability of two or more arrivals occurring at the same point in time.

Another important distribution in queueing theory is the exponential distribution.

Definition 2.6 (Exponential Distribution)**.** An exponential distribution has probability density function

$$\begin{cases} \delta e^{-\delta x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Note the distinction between *probability mass functions*, such as that of the Poisson distribution, which are for discrete random variables, and *probability density functions*, which are used for continuous random variables.

So, we use an exponential distribution for the length of time between arrivals or events, which is a continuous random variable, and a Poisson distribution for the *number of occurrences* in a certain interval, which is a discrete random variable.

Lemma 2.7. *The expected value of the time between occurrences is* $\frac{1}{\delta}$.

Proof. To find the expected value of a continuous random variable X, we take the integral over all possible values x that X takes multiplied by the probability of x taking that value. Since the exponential distribution takes values in $[0, \infty)$, the bounds of our integral will be 0 and ∞. We evaluate the integral

$$\int_0^\infty x(\delta e^{-\delta x})dx.$$

We can factor out a $\frac{1}{\delta}$ and substitute $y = \delta x$ to get the integral

$$\frac{1}{\delta}\int_0^\infty ye^{-y}dy.$$

Using integration by parts, we can simplify to

$$\frac{1}{\delta}\left[-ye^{-y} - e^{-y}\right]\Big|_0^\infty = \frac{1}{\delta},$$

so the expected value of the time between occurrences is $\frac{1}{\delta}$. □

3. Continuous-Time Markov Chains

3.1. Introduction.

Definition 3.1. A continuous-time Markov chain is one in which the set of possible times is $T = \{t : 0 \leq t < \infty\}$.

This differs from the Markov chains we've seen so far in which T is the nonnegative integers. In this paper, we will deal with continuous-time Markov chains in which the state space is discrete.

Similarly to discrete-time Markov chains, continuous-time Markov chains, or CTMCs, are governed by the following principle:

Given real numbers $t \geq 0$ and $s \geq 0$, a continuous-time Markov chain has the property that

$$\mathbb{P}(X_{t+s} = j \,|\, X_t = i, X_v = k_v, 0 \leq v \leq t) = \mathbb{P}(X_{t+s} = i \,|\, X_t = i),$$

with $i, j \in \Omega$, most commonly for our purposes the positive integers.

We can define continuous-time Markov chains in the language of stochastic processes as well, as discrete-space continuous-time stochastic processes. We have the following two facts, then:

(1) When the chain enters a state i, it stays at that state for an amount of time that is exponentially distributed with parameter δ_i (so, according to the state i) before transitioning again.

(2) When the process leaves a state i, its probability of transitioning to state j is P_{ij}; we have $P_{ii} = 0$ and $\sum_{j\in\Omega} P_{ij} = 1$.

where the second part of the second property is true for all Markov chains.

We also have the related notion of a transition probability function $P_{ij}(t)$, which is the probability that if the Markov chain is at state i at time t_0, then it will be at state j at time $t_0 + t$ for all t_0. We can thus write

$$P_{ij}(t) = \mathbb{P}(X_{t_0+t} = j \,|\, X_{t_0} = i).$$

We denote the matrix of all probabilities $P_{ij}(t)$ as $P(t)$, or the transition function.

3.2. **Steady-State Probabilities.** We define a continuous-time Markov chain to be stable if all of its states are positive recurrent. For irreducible, aperiodic, and stable Markov chains, we define the steady-state or stationary probability of a state j to be

$$\pi_j = \lim_{t\to\infty} P_{ij}(t),$$

and this limit exists because of the three conditions we placed on the chain.

In order to determine the stationary distribution π, we need to define the infinitesimal generator of a CTMC, $Q = P'(0)$. Q is the matrix of infinitesimal rates Q_{ij}, where

$$Q_{ij} = \delta_i P_{ij}$$

for $i \neq j$ and

$$Q_{ii} = -\sum_{i\neq j} Q_{ij}.$$

As defined earlier, δ_i is the rate of leaving state i, so Q_{ij} is the product of the rate of leaving state i and the probability of transitioning to state j, or the rate of transitions from i to j.

Remark 3.2 (Transition-rate matrix)**.** The transition-rate matrix Q with entries q_{ij} of a CTMC also satisfies the following properties:

(1) $0 \leq -q_{ii} < \infty$
(2) $0 \leq -q_{ij}$ for $i \neq j$
(3) $\sum_{j\in\Omega} q_{ij} = 0$ for all $i \in \Omega$.

And, as we defined earlier, since $q_{ii} = -\sum_{i\neq j} q_{ij}$, condition (3) is automatically satisfied.

In order to find the steady-state solution, we solve for π such that $\sum_{i\in\Omega} \pi_i Q_{ij} = 0$ for all j or $\pi Q = 0$.

We can obtain this result through the Chapman-Kolmogorov equations, which say that

$$P(t+h) = P(h)P(t),$$

the continuous-time equivalent to saying that $P(t) = P^t$ for discrete-time chains. Thus we have

$$P(t+h) - P(t) = P(t)(P(h) - I) = P(t)(P(h) - P(0)).$$

We divide by h and take the limit as $h \to 0$:

$$\lim_{x\to 0} \frac{P(t)(P(h) - P(0))}{h},$$

which yields that

$$\frac{dP}{dt} = P(t)P'(0) = P(t)Q.$$

If we set $\frac{dP}{dt} = 0$ for a given transition function P_π, we can say that the probability of transitioning to a certain state does not change over time, and thus, we have achieved stationarity [3]. When we solve for P_π such that $P_\pi Q = 0$, all of the rows of P must be the same in order to be independent of time, as with stationarity for discrete-time Markov chains; if we denote each row as π, we have $\pi Q = 0$, where π represents stationarity or the steady-state solution of the continuous-time Markov chain. Also, we must have $\sum_{i\in\Omega} \pi_i = 1$.

4. Birth-and-Death Processes

Definition 4.1. A birth-and-death process is a type of continuous-time Markov chain in which the change at any occurrence from any state i is either $+1$, or a "birth", or -1, or a "death" (so, to either $i+1$ or $i-1$, akin to a customer entering or leaving a queue).

If the probability of a birth at state i is v_i, the probability of a death at state i is then $1 - v_i$. Because the amount of time between occurrences follows an exponential distribution, the mean amount of time between occurrences is $\frac{1}{\delta_i}$. We can calculate the *birth rate* at state i by saying

$$b_i = \delta_i v_i,$$

then, and similarly, the death rate by saying

$$d_i = \delta_i(1 - v_i),$$

which, after adding the two equations, yields the somewhat intuitive consequence that

$$\delta_i = b_i + d_i,$$

or that the total rate at a certain state is equal to the sum of the birth and death rates, or, more generally, that the total rate of a state of a continuous-time Markov chain is equal to the sums of the rates of the transition to each of the other accessible states.

5. Queueing Theory Notation

With all of the relevant background covered, we are finally ready to actually define and work with queues.

5.1. **Standard Queue Notation.** First, we define the different parameters and notation for queues.

A queue has 5 different parameters: arrival process, service pattern/rate, number of servers, queue capacity, and queue discipline. The arrival process denotes the manner in which customers arrive to the queue, and the service pattern describes the customer service pattern time. The queue capacity is the number of buffers available for customers, including the servers themselves, and lastly, queue discipline represents the manner in which customers in the line waiting are served; some typical disciplines are First In First Out (FIFO)/First Come First Served (FCFS), Last In First Out (LIFO), and random order. If the queue capacity is unlimited and the discipline is FIFO/FCFS, as they are for the purposes of this paper, we generally drop the last two parameters when referring to the queue. So, for example, we denote a FCFS, unlimited capacity queue as $A/S/c$ where A refers to the arrival process, S to the service process and c to the number of servers.

When referring to arrival rates and service patterns, there are a few different conventions to be aware of. M, or Markovian, refers to a Poisson process for arrival or exponential service time distribution; D refers to a deterministic model (which is not stochastic and not covered in this paper); G refers to general, or an arbitrary distribution for both. In this paper, we deal with M/M queues.

Now that we have defined notation and parameters for a queue, we can determine the steady-state solution for a few types queues. The steady-state solution is the state where the probability distribution of the number of customers in the queueing system is independent of time (as opposed to the transient state, where this distribution depends on time).

5.2. **Important Metrics for Queues.** Before moving on to our two types of queues, we'll discuss two important metrics for $G/G/1$ queues and the relations between them. They are the following:

- $\mathbb{E}[Q]$, the expected value of the queue size, including the customer in service.
- $\mathbb{E}[N_q]$, the expected value of the number of customers not in service.
- $\mathbb{E}[D]$, the expected value of the delay of each customer, or the total time from arrival in the queue to completion of service.
- $\mathbb{E}[W_q]$, the expected value of the waiting time in a queue, or the total time form arrival to when a customer's service begins.

Theorem 5.1 (Little's Theorem)**.** *Little's theorem states that for all $G/G/1$ queues in steady-state, $\mathbb{E}[Q] = \lambda\mathbb{E}[D]$, and similarly, that $\mathbb{E}[N_q] = \lambda\mathbb{E}[W_q]$.*

The intuition for the first form of this formula is that a customer who leaves the queueing node after being serves sees on average $\mathbb{E}[Q]$ customers, the expected value of the queue size in steady-state. This is equal to $\lambda\mathbb{E}[D]$, the customers who arrived and were waiting while they were in service. Similar intuition can be applied for the second form of the equation. We will use Little's formula to calculate $\mathbb{E}[Q]$ and $\mathbb{E}[D]$ for the $M/M/1$ queue.

The proof is given in Little's 1960 paper [2].

5.3. $M/M/1$ **Queues.** We'll begin with the $M/M/1$ queue, starting with some general results about queues. For a $G/G/c$ queue with an average service rate of μ and an average customer arrival rate of λ, we can measure the *traffic congestion* by having $\rho = \frac{\lambda}{c\mu}$. It follows that if $\rho > 1$, there is no steady state solution, since the number of people waiting to be served in the queue grows without bound, and if $\rho = 1$, there is no steady-state solution unless the arrival and service rate are deterministic, since the queue is always behind. Therefore, we can only find a steady-state solution if $\rho < 1$.

In the case of an $M/M/1$ queue, we have $c = 1$, so we must have $\rho = \frac{\lambda}{\mu} < 1$ or $\lambda < \mu$.

Theorem 5.2. *The steady-state solution for the $M/M/1$ queue is $\pi_i = \rho^i(1-\rho)$ for $i \geq 0$.*

We will use the techniques described in Section 3 to solve for the steady-state solution (or stationary distribution) π. First, we will define Q using the transition functions $P_{ij}(t)$ and the transition rates δ_i. Next, we will solve for π such that $\pi Q = 0$.

Proof. We will begin by showing a state-transition diagram for the $M/M/1$ queue, which will help us determine Q.

We're given the rates of transition in and out of the queue to be λ and μ respectively. We know then that the rates of transition $q_{i,i+1} = \lambda$ for $i \geq 0$ and $q_{i,i-1} = \mu$ for $i > 0$. There is also a rate of transition for q_{ii}, but all other rates are 0 since an $M/M/1$ queue follows a birth-and-death process. Therefore, since $q_{ii} = -\sum_{i \neq j} q_{ij}$, we have $q_{00} = -\lambda$ and $q_{ii} = -(\lambda + \mu)$ for all $i > 0$. Therefore we have our transition rate matrix Q as follows:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \dots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \dots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \dots \\ 0 & 0 & \mu & -(\lambda+\mu) & \lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

and we must solve for π such that $\pi Q = 0$, or really the infinite-dimensional vector of all 0s. We can obtain the following sequence of equations:

$$-\pi_0\lambda + \pi_1\mu = 0 \tag{5.1}$$

$$\pi_0\lambda - \pi_1(\lambda+\mu) + \pi_2\mu = 0 \tag{5.2}$$

$$\pi_1\lambda - \pi_2(\lambda+\mu) + \pi_3\mu = 0 \tag{5.3}$$

and more generally, following the pattern of columns of Q, we have that $\pi_{i-1}\lambda - \pi_i(\lambda+\mu) + \pi_{i+1}\mu = 0$. Rearranging these equations, we see from equation 5.1 that $\pi_1 = \left(\frac{\lambda}{\mu}\right)\pi_0 = \rho\pi_0$ and using the two-term recurrence relation described, we see that from equation 5.2 that since $\pi_1(\lambda+\mu) = \pi_2\mu + \pi_0\lambda$, we have

$$\begin{aligned} \left(\frac{\lambda(\lambda+\mu)}{\mu}\right)\pi_0 &= \pi_2\mu + \pi_0\lambda \\ \left(\frac{\lambda(\lambda+\mu)}{\mu} - \lambda\right)\pi_0 &= \pi_2\mu \\ \left(\frac{\lambda^2 + \mu\lambda - \mu\lambda}{\mu}\right)\pi_0 &= \pi_2\mu \\ \left(\frac{\lambda}{\mu}\right)^2\pi_0 &= \pi_2 \\ \rho^2\pi_0 &= \pi_2 \end{aligned}$$

and continuing this recursive process, in general, we have

$$\pi_i = \rho^i\pi_0.$$

Furthermore, because π is a probability distribution, we have $\sum_{i=0}^{\infty}\pi_i = 1$, and we can rewrite this as $\sum_{i=0}^{\infty}\rho^i\pi_0$. After we factor out the constant term π_0, we obtain the sum $\pi_0\sum_{i=0}^{\infty}\rho^i$, which is just a geometric

series. Therefore, we have $\pi_0 \left(\frac{1}{1-\rho}\right) = 1$, and solving for π_0, see that $\pi_0 = 1 - \rho$. Plugging this in to our explicit formula in terms of i and π_0, we see that

$$\pi_i = \rho^i(1-\rho).$$

□

We can also calculate $\mathbb{E}[Q]$ and $\mathbb{E}[D]$ using Little's formula.

We can calculate $\mathbb{E}[Q]$ by saying that

$$\mathbb{E}[Q] = \sum_{i=0}^{\infty} i\pi_i = \frac{\rho}{1-\rho},$$

and using Little's theorem, we have $\mathbb{E}[D] = \frac{\lambda\rho}{1-\rho}$.

So, the expected number of customers in the queue when it reaches steady-state is $\frac{\rho}{1-\rho}$ and the average delay faced by a customer who enters the queue in steady-state is $\frac{\lambda\rho}{1-\rho}$

5.4. $M/M/\infty$ **Queues.** Using a similar method, we can find the steady-state solution for an $M/M/\infty$ queue. This queue model, like an $M/M/1$ queue, has a Poisson process for arrival rate and an exponential distribution for service rate; however, it differs from our first queue in that since there are an infinite number of servers, all customers are served immediately. Note that if we define $A = \frac{\lambda}{\mu}$ once again (we use A rather than ρ because we use ρ for single-server queues and A for multiserver queues), the queue will be stable (ie. not grow infinitely) for any nonnegative A, since all customers are served immediately.

The transition rates for the various states also differ slightly from an $M/M/1$ queue; while the probability for births, or $q_{i,i+1}$ for $i \geq 0$, remains to be λ, the probability of a death, or $q_{i,i-1}$ for $i > 0$, is now equal to $i\mu$, because at that time there are i customers being served at the queueing node.

q_{00} is thus once again $-\lambda$, but $q_{ii} = -(\lambda + i\mu)$, because of our definition and because we essentially have a competition between $i+1$ random variables for the next transition, 1 for the next arrival and i for the next departure. Therefore, our transition-rate matrix Q is

$$\begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \dots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \dots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & 0 & \dots \\ 0 & 0 & 3\mu & -(\lambda+3\mu) & \lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

for the $M/M/\infty$ queue, and we once again set up the system of equations from

$$\pi Q = 0.$$

We get the following relations:

$$-\pi_0\lambda + \pi_1\mu = 0 \tag{5.4}$$

$$\pi_0\lambda - \pi_1(\lambda+\mu) + 2\pi_2\mu = 0 \tag{5.5}$$

$$\pi_1\lambda - \pi_2(\lambda+2\mu) + 3\pi_3\mu = 0 \tag{5.6}$$

and in general, $\pi_i(\lambda + i\mu) = \pi_{i-1}\lambda + (n+1)\pi_{n+1}\mu$. Because of the reversibility property of the birth and death process, we have $\pi_0\lambda = \pi_1\mu$, $\pi_1\lambda = 2\pi_2\mu$ and in general, $\pi_i\lambda = (n+1)\pi_{n+1}\mu$.

With this fact we see that $\pi_1 = A\pi_0$, $\pi_2 = \frac{A^2\pi_0}{2}$, and in general that $\pi_n = \frac{A^n\pi_0}{n!}$. Again using the fact that the π_n's sum to 1, we see that

$$1 = \sum_{i=0}^{\infty} \frac{A^n\pi_0}{n!} = \pi_0 \sum_{i=0}^{\infty} \frac{A^n}{n!},$$

which is the expansion of the Maclaurin series for e^x. Therefore we have

$$1 = \pi_0 e^A \implies \pi_0 = e^{-A}$$

and thus $\pi_n = \frac{e^{-A}A^n}{n!}$, meaning that the stationary distribution for an $M/M/\infty$ queue is a Poisson distribution.

References

[1] Hannah Constantin. *Markov Chains and Queueing Theory*. URL: https://www.math.uchicago.edu/~may/VIGRE/VIGRE2011/REUPapers/Constantin.pdf.

[2] John D. C. Little. *A Proof for the Queueing Formula: $L = \lambda W$*. 1960. URL: http://fisherp.scripts.mit.edu/wordpress/wp-content/uploads/2015/11/ContentServer.pdf.

[3] Karl Sigman. *IEOR 6711: Continuous-Time Markov Chains*. 2009. URL: http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-CTMC.pdf.

[4] David K. Smith. *Calculation of Steady-State Probabilities of M/M Queues: Further Approaches*. 2002. URL: http://www.kurims.kyoto-u.ac.jp/EMIS/journals/HOA/ADS/Volume6_1/50.pdf.

[5] Dr. Janos Sztrik. *Basic Queueing Theory*. URL: https://irh.inf.unideb.hu/~jsztrik/education/16/SOR_Main_Angol.pdf.

[6] Moshe Zukerman. *Introduction to Queueing Theory and Stochastic Teletraffic Models*. 2013. eprint: arXiv:1307.2968.