Mathematical Foundations of Computer Networking

Srinivasan Keshav

ADDISON-WESLEY PROFESSIONAL COMPUTING SERIES

Mathematical Foundations of Computer Networking

6

Stochastic Processes and Queueing Theory

6.1 Overview

Queues arise naturally when entities demanding service, or **customers**, interact asynchronously with entities providing service, or **servers**. **Service requests** may arrive at a server when it is either unavailable or busy serving other requests. In such cases, service demands must be either queued or dropped. A server that queues demands instead of dropping them can smooth over fluctuations in the rate of service and in the rate of request arrivals, leading to the formation of a queueing system. The study of the probabilistic behavior of such systems is the subject of queueing theory.

Here are some examples of queueing systems.

1. The arrival and service of packets at the output queue of a switch. Packets may arrive when the output link is busy, in which case the packets, which are implicit service requests, must be buffered (i.e., queued).
2. The arrival and service of HTTP requests at a Web server. If the Web server is busy serving a request, incoming requests are queued.
3. The arrival of telephone calls to a switch-control processor in the telephone network. The processor may be unable to service the call because the switch is busy. In this case, the call is queued, awaiting the release of network resources.

Given a queueing system, we would like to compute certain quantities of interest, such as

- The **queueing delay**: the expected waiting time for a service request
- The **backlog**: the mean number of service requests that are awaiting service
- The **utilization factor**: the expected fraction of time that the server is busy
- The **drop rate**: the expected fraction of service requests that must be dropped because there is no more space left in the queue
- The mean length of a busy or idle period

Queueing theory allows us to compute these quantities—both for a single queue and for interconnected networks of queues—as long as the incoming traffic and the servers obey certain simplifying conditions. Unfortunately, measurements show that traffic in real networks does *not* obey these conditions. Moreover, we cannot mathematically analyze most networks that are subjected to realistic traffic workloads. Nevertheless, it is worth studying queueing theory for two important reasons. First, it gives us fundamental insights into the behavior of queueing systems. These insights apply even to systems that are mathematically intractable. Second, the solutions from queueing theory—even from unrealistic traffic models—are a reasonable first approximation to reality. Therefore, as long as we keep in mind that results from queueing theory are only approximate and are meant primarily to give an insight into a real system, we can derive considerable benefit from the mathematical theory of queues.

6.1.1 A General Queueing System

We now introduce some standard notation. A **queue** is formed when **customers** present **service requests**, or **jobs**, to one or more **servers**. Customers arrive at a rate of λ customers/second at times t, and the time between arrivals is described by the **interarrival time distribution** $A(t) = P(\text{time between arrivals} \leq t)$. We denote the service time by x, which has a **service-time distribution** $B(x) = P(\text{service time} \leq x)$ with a mean service rate of μ customers/second. Customers are assumed to wait in the queue with a **mean waiting time** T. Note that in this chapter, we will study primarily a single queue in isolation.

6.1.2 Little's Theorem

Little's theorem is a fundamental result that holds true for *all* arrival and service processes. The theorem states that the mean number of customers in a queueing system is the product of their mean waiting time and their mean arrival rate.

Proof of Little's Theorem: Suppose that customers arrive to an empty queue[1] at a mean rate of λ customers/second. An average of λt customer arrive in t seconds. Let T denote the mean waiting time in seconds of a newly arriving customer. The total time spent waiting in the queue across all customers during the time interval is therefore λTt customer-seconds.

Let N denote the mean number of customers in the queue. In 1 second, these N customers accumulate N customer-seconds of total waiting time. Thus, in t seconds, they accumulate a total of Nt customer-seconds of waiting time. This must equal λTt, which implies that $N=\lambda T$.

Note that this argument is independent of the length of the time interval t. Moreover, it does not depend on the order of service of customers, the number of servers, or the way in which customers arrive. Thus, it is a powerful and general law applicable to all queueing systems.

EXAMPLE 6.1: LITTLE'S THEOREM

Suppose that you receive e-mail at the average rate of one message every 5 minutes. If you read all your incoming mail instantaneously once an hour, what is the average time that a message remains unread?

Solution:

The mean message arrival rate is 12 messages/hour. Because you read e-mail once an hour, the mean number of unread messages is 6 (the expected number of messages received in half an hour). By Little's theorem, this is the product of the mean arrival rate and the mean waiting time, which immediately tells us that the mean time for a message to be unread is $\frac{6}{12}$ hours = 30 minutes.

EXAMPLE 6.2: LITTLE'S THEOREM

Suppose that 10,800 HTTP requests arrive at a Web server over the course of the busiest hour of the day. If the mean waiting time for service should be under 6 seconds, what should be the largest allowed queue length?

Solution:

The arrival rate λ = 10,800/3600 = 3 requests/second. We want $T \leq 6$. Now, $N = \lambda T$, so $T = N/\lambda$. This means that $N/\lambda \leq 6$ or that $N \leq 6 * 3 = 18$. So, if the mean

1. The same reasoning applies when customers arrive to a nonempty queue, but arrivals to an empty queue simplifies the analysis.

queueing delay is to be no larger than 6 seconds, the mean queue length should not exceed 18. In practice, the Web server could return a server-busy response when the queue exceeds 18 requests. This is conservative because then 18 is the *maximum* queue length rather than its mean.

6.2 Stochastic Processes

The foundation of queueing theory is the mathematical study of a **stochastic process**. Such a process is used to model the arrival and service processes in a queue. We will both intuitively and mathematically define a stochastic process and then study some standard stochastic processes.

EXAMPLE 6.3: DETERMINISTIC AND STOCHASTIC PROCESSES

Consider a staircase with ten steps numbered 1 through 10 and a person standing on the first step, which is numbered 1. Suppose that a clock next to the staircase ticks once a second starting at time 1. Finally, assume that it takes zero time to climb each step.

If the person were to climb one step at each clock tick, we can predict exactly where the person would be at each time step. At time 0, the person is on step 1 and would stay there until just before time 1. When the clock ticks and time increments to 1, the person would be at step 2 and would stay there until just before time 2. At time 2, the person would be on step 3, and so on. We therefore call the act of climbing the staircase in this fashion a **deterministic** process.

In contrast, suppose that the person climbs one step or goes down one step or stays on the same step with some (potentially zero) probability. With this change, we lose predictability. That is, we no longer know exactly where the person will be at any moment in time; we can only attach probabilities to the set of places where the person *could* be at that time. The process is no longer deterministic but instead **stochastic**.

We capture this by means of a random variable X_i that denotes the step the person is on immediately after the ith clock tick. The random variable is associated with the probability distribution $\pi(i)$ over the positions where the person could be immediately after that time. For instance, at time 0, the person is at step 1 with probability 1, so the distribution of X_0 over the domain {1, 2,...,10} is given by the discrete probability distribution $\pi(0)$ = {1.0, 0,..., 0}. Suppose that the probability that the person goes up is p, that the person goes down is q, and that the person stays on the same step is $1 - p - q$, except at

step 1, where the probability of going up is p, and the probability of staying on the same step is $1 - p$. Then, immediately after time 1 (after the first clock tick), $\pi(1) = \{1 - p, p, 0,...,0\}$. Similarly, $\pi(2) = \{(1 - p)^2 + pq, p(1 - p - q) + (1 - p)p, p^2, 0,...,0\}$, and so on. To compute $\pi(i)$ given $\pi(i - 1)$, we determine the different ways that we can reach each particular step, summing the probability over all possible ways to reach that step.

Note that we distinguish between the distribution of the random variables at each time step and the actual trajectory taken by a person. For a given trajectory, at each time instant, the person is, of course, only on one step of the staircase. The trajectories are created by sampling from the distributions $\pi(i)$. A trajectory is also therefore called a **sample path**.

This example suggests the following definition of a **stochastic process**: a family of random variables X_i that are indexed by the time index i. The value of the random variable in a particular trajectory is also called the **state** of the stochastic process at that point in time. Without loss of generality, we can think of the states as being chosen from the integers from 1 to N. Thus, we can imagine the process as "moving" from the state corresponding to the value taken by random variable X_i in a given trajectory to the state corresponding to the value taken by random variable X_{i+1} at time i+1, just like the person moves from one stair to another. As we have shown, given the probabilities of moving from one step to another, we can, in principle, compute the distribution of each X_i: the distribution of the stochastic process over the state space at that time.

Time is discrete in this example. In other situations, a system is better modeled when time is continuous. In this case, the family of random variables corresponding to the stochastic process consists of the variables $X(t_1)$, $X(t_2)$,..., where the t_i represent the times at which the state transitions occur. Given the probability of moving from step to step, we can compute $\pi(t_{i+1})$, the distribution of $X(t_{i+1})$, from $\pi(t_i)$ the distribution of $X(t_i)$.

In the example, the person's movements were limited to moving up or down one step on each clock tick. We could, instead, allow the person to go from a given step to any other step—not just the steps above and below—with some probability. Indeed, this distribution of probabilities could differ at different steps and even differ at each clock tick! And, finally, the person could be on a ramp, so that the amount of movement could be a real positive or negative quantity rather than an integer. These variations are all within the scope of definition of a stochastic process, but the analysis of the corresponding processes is progressively more difficult. We will first describe some standard types of stochastic processes and then focus on the simplest ones.

6.2.1 Discrete and Continuous Stochastic Processes

A stochastic process can be classified as a discrete or a continuous process in two ways. (1) The values assumed by the random variables, or the **state space**, can be discrete or continuous, and (2) the index variable, or time, also can be discrete or continuous.

A **discrete-space process** is one in which the random variables X_i take on discrete values. Without loss of generality, we can think of the state in a discrete-space process as being indexed by an integer in the range 1,2,...N.

EXAMPLE 6.4: DISCRETE-SPACE PROCESS

Continuing with Example 6.3, we see that the set of possible states is the set of stairs, which forms a discrete set.

A **continuous-space process** is one in which the random variables take on values from a finite or infinite continuous interval or a set of such intervals.

EXAMPLE 6.5: CONTINUOUS-STATE PROCESS

Continuing with Example 6.3, consider a person walking up and down a ramp rather than a stair. This would allow movements by real amounts. Therefore, the random variable corresponding to the state of the process can take on real values, and the corresponding stochastic process would be a continuous-space process.

In a **discrete-time** process, the indices of the random variables are integers. We can think of the stochastic process in a particular trajectory as moving from one state to another at these points in time.

EXAMPLE 6.6: DISCRETE-TIME PROCESS

Continuing with Example 6.3, this corresponds to a person moving from one step to another exactly at each clock tick. Such a stochastic process is also called a **stochastic sequence**.

In a **continuous-time** process, the times when the process can move to a new state are chosen from a real interval.

EXAMPLE 6.7: CONTINUOUS-TIME PROCESS

Continuing with Example 6.3, this corresponds to a person moving from stair to stair at will, independent of the clock.

Stochastic processes corresponding to all four combinations of {discrete space, continuous space} and {discrete time, continuous time} are well known.

6.2.2 Markov Processes

An important aspect of a stochastic process is how the probability of transitioning from one state to another is influenced by past history. Continuing with our staircase example—a discrete-time and discrete-space stochastic process—consider a person who is allowed to go from any stair to any other stair. Moreover, we will ask the person to obey the following rules: If he or she arrives at stair 5 from stair 6, move to stair 3. If, however, he or she arrives at stair 5 from stair 3, move to stair 9. In all other cases, move to stair 1. Suppose that at some point in time, we see that the person is on stair 5. What happens next?

The answer is: We don't know. It depends on where the person was in the previous time step. Stated more precisely, the distribution of the random variable X_{n+1} (i.e., $\pi(n+1)$) when $X_n = 5$ depends on the value of X_{n-1}. Generalizing from this example, we can define more complex stochastic processes for which $\pi(n+1)$ depends not only on X_n but also on X_{n-1}, X_{n-2},..., X_1. Such systems are inherently complex, and there is little we can say about them.

In an attempt to curb this complexity, consider the following rule:

$$\pi(n+1) \text{ depends only on the value of } X_n$$

This rule simplifies the situation: If the person is on step 5 at time n, we know π_{n+1} *independent of the prior history.* As we will see, this allows us to easily compute many quantities of interest about the process. Moreover, many naturally occurring stochastic processes obey this rule. Owing to these two facts, stochastic processes that obey this rule are given a special name: **Markov processes**, in honor of A. N. Markov, who first studied them in 1907.

Formally, for the case of discrete-time stochastic processes, we state the **Markov property** as

$$\begin{aligned} P(X_{n+1} = j | X_n = i_n, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, X_{n-3} = i_{n-3}, \ldots, X_1 = i_1) \\ = P(X_{n+1} = j | X_n = i_n) \end{aligned} \quad \textbf{(EQ 6.1)}$$

The conditional probability $P(X_{n+1} = j | X_n = i_n)$ is called the **transition probability** to go from state i_n to state j.

EXAMPLE 6.8: MARKOV PROCESS

Consider a discrete-time discrete-space Markov process whose state space is {1,2,3}. Let $P(X_3 = 1 \mid X_2 = 1) = 0.2$; $P(X_3 = 2 \mid X_2 = 1) = 0.4$; $P(X_3 = 3 \mid X_2 = 1) = 0.4$. Suppose that we know that $P(X=1) = 1$. Then, the Markov property allows us to compute $\pi(3)$ no matter which path was taken to reach the state 1 at time 2.

Note that for a discrete-time stochastic process at time step n, we already know the sequence of prior states. At time n, we are usually interested in computing $\pi(n+1)$ given this past history. The Markov property allows us to forget everything about history except the value of the current random variable, which encapsulates all past history. This is similar in spirit to the memorylessness property of the exponential distribution (see Section 1.6.3).

A similar property holds true for continuous-time stochastic processes. For simplicity, we will first study discrete-time processes that obey the Markov property, also known as **discrete-time Markov chains**, before considering continuous-time Markov processes.

6.2.3 Homogeneity, State-Transition Diagrams, and the Chapman-Kolmogorov Equations

A stochastic process may satisfy the Markov property even if its transition probabilities vary over time. A process with time-dependent transition probabilities is called a **nonhomogeneous Markov process**. Of course, this greatly complicates the analysis, but we can simplify it by decreeing that the transition probabilities should be time-independent. In our example, this means that when the person is on a particular step—say, step 4—the probability of going to any other step is always the same, no matter *when* the person got to step 4. Such a process is called a **homogeneous Markov process**. For a homogeneous Markov process, we define the *time-independent* transition probability between state i and state j as $p_{ij} = P(X_n = j \mid X_{n-1} = i)$ for any n.

EXAMPLE 6.9: HOMOGENEOUS MARKOV PROCESS

Consider the discrete-time discrete-space stochastic process in Example 6.8. If this process were homogeneous, we need not consider exactly one point in

time, such as time 2. Instead, if $P(X_{n+1}= 1 \mid X_n = 1) = 0.2$; $P(X_{n+1}=2 \mid X_n = 1) = 0.4$; $P(X_{n+1} = 3 \mid X_n = 1) = 0.4$, we can compute the distribution X_{n+1} given X_n for all values of n.

The state-transition probabilities for a homogeneous Markov chain with N states have two equivalent representations. The first is in the form of an $N \times N$ transition matrix $\mathbf{A}$ whose elements are the probabilities $p_{ij.}$ This representation has the attractive property that the probability of going from any state i to state j in two steps is given by the elements of $\mathbf{A^2}$. The second is as a graph (see Example 6.10). In this representation, vertices represent states, and the annotation on an edge from vertex i to vertex j is p_{ij}. This visually represents a Markov chain. Note that a non-homogeneous Markov chain requires such a state-transition diagram for each time step.

EXAMPLE 6.10: REPRESENTING A HOMOGENEOUS MARKOV PROCESS

Continuing with Example 6.9: We have already seen that $p_{11} = 0.2$, $p_{12} = 0.4$, $p_{13} = 0.4$. Suppose that $p_{21} = 1.0$, $p_{22} = 0$, $p_{23} = 0$, and $p_{31} = 0.5$, $p_{32} = 0.25$, $p_{33} = 0.25$. Then, we can represent it in two ways as follows and as shown in Figure 6.1.

$$A = \begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 1.0 & 0 & 0 \\ 0.5 & 0.25 & 0.25 \end{bmatrix}$$

where $\boldsymbol{A}$ is a right stochastic matrix (see Section 3.6).

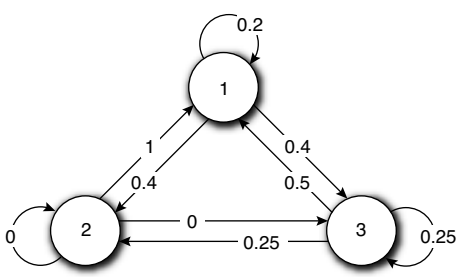


Figure 6.1 State-transition diagram for Example 6.10

Given the set of transition probabilities p_{ij} for a homogeneous Markov chain, we can define the m-step transition probability from state i to state j denoted $p_{ij}^{(m)}$ by

$$p_{ij}^{(m)} = P(X_{n+m}=j \mid X_n = i) = \sum_k p_{ij}^{(m-1)} p_{kj} \qquad m = 2,3,... \qquad \textbf{(EQ 6.2)}$$

where the sum of products form comes from summing across independent events (that of going to some intermediate state in $m - 1$ steps), and each term is a product because it is the combination of two independent events (going from state i to state k and from state k to state m). These relations exist only because of the Markovian nature of the chain. They are important enough that they are given their own name: the **Chapman-Kolmogorov** equations, which can also be stated as:

$$p_{ij}^{(m)} = P(X_{n+m}=j \mid X_n = i) = \sum_k p_{ik} p_{kj}^{(m-1)} \qquad m = 2,3,... \qquad \textbf{(EQ 6.3)}$$

Comparing the two, we see that the first formulation traces the trajectory of the process as it goes from state i to state k in $m - 1$ steps and from state k to state j in one step. The second formulation traces the trajectory of the process as it goes from state i to state k in one step and from state k to state j in $m - 1$ steps. Clearly, these are equivalent recursions.

6.2.4 Irreducibility

If every state of a stochastic process can be reached from every other state after a finite number of steps, the process is called **irreducible**; otherwise, it is **reducible**. Moreover, if states of a stochastic process can be separated into subsets that are mutually unreachable, we call each such set a **separable subchain**.

EXAMPLE 6.11: A REDUCIBLE MARKOV CHAIN

Consider the Markov chain in Figure 6.2. Here, the transition probabilities p_{ij} for i even and j odd or i odd and j even are 0. Therefore, if the initial state of the process is an even-numbered state, the trajectory of process is confined to even-numbered states. Alternatively, a process that starts from an odd-numbered state will forever stay in odd-numbered states. The even-numbered states are unreachable from the odd-numbered steps' states, and the chain, therefore, is reducible. Indeed, we could separate the even-numbered and odd-numbered states into separate chains that would equivalently describe the process. We can generalize this idea to construct stochastic processes that can be decomposed into as many subchains as we wish.

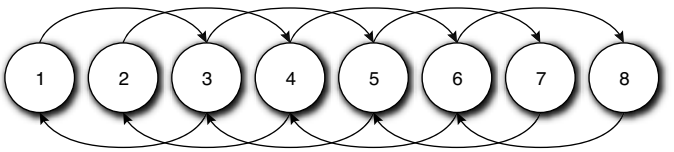


Figure 6.2 A reducible Markov chain

6.2.5 Recurrence

For every state j of a stochastic process, one of two conditions must hold: After entering state j, either the probability of reentering state j after a finite number of steps is 1, or there is some nonzero probability that the state is not reentered after a finite number of steps. In Example 6.3, this is equivalent to saying that after stepping on a stair—say, stair 6—it either is certain that the person will return to stair 6, or there is a nonzero probability that the person will not return to stair 6. If return to a state is certain, we call the state **recurrent**; otherwise, we call it **transient**.

Let f_j^n denote the probability that the *first* return to state j is after n steps. State j is recurrent if $\sum_{n=1}^{\infty} f_j^n = 1$ and transient otherwise.

Although a state is recurrent, its expected recurrence period, defined as $\sum_{n=1}^{\infty} n f_j^n$, may be infinite. This sum may diverge if f_j^n is sufficiently large for large values of n. In such cases, the mean recurrence period is infinite, and the state is called **recurrent null**. Otherwise, it is called **recurrent non-null**.

6.2.6 Periodicity

Given a recurrent state j, suppose that the only way to return to that state is to take r, $2r$, $3r$... steps, with $r \geq 2$. We then call the state j **periodic**, with a period r. Periodic states arise when the Markov chain has a cycle. A trivial way to check whether a state is periodic is to see whether it has a self-loop, that is $p_{jj} > 0$. If so, the state can be reentered with any desired number of steps, which makes $r = 1$ and the state **aperiodic**. For an irreducible Markov chain, if *any* state has a self-loop, all states are aperiodic.

EXAMPLE 6.12: PERIODIC AND APERIODIC MARKOV CHAINS

The Markov chain in Figure 6.2 is periodic with period 2, and the chain in Figure 6.1 is aperiodic.

6.2.7 Ergodicity

The sequence of states visited by a stochastic process is called its **trajectory**. For example, a valid trajectory for the chain in Figure 6.1 is $1 \to 2 \to 1 \to 3 \to \ldots$. Given a trajectory, we can compute a statistic on it, such as the fraction of the trajectory spent in a particular state. This statistic is the limiting ratio of the number of occurrences of that state to the length of the trajectory. Because trajectories can be made as long as we desire, if the limiting statistic exists, it can be approximated as closely as we wish. We call such a limit a **time average**.

Now, consider a set of instances of the same stochastic process. At each time step, each instance changes state according to the same transition probabilities. Nevertheless, the trajectory associated with any pair of processes in the ensemble may differ owing to their stochastic nature. Suppose that we want to compute a statistic on the ensemble of trajectories at any time step. For instance, we may wish to compute the fraction of trajectories in state 1 at time step 10. If the limiting statistic exists, we can approximate this statistic as closely as we wish by making the ensemble sufficiently large. We call such a limit an **ensemble average**.

An interesting question is whether a statistic computed as a time average is the same as a statistic computed as an ensemble average. As the next example shows, this need not necessarily be the case!

EXAMPLE 6.13: TIME AND SPACE AVERAGES

Consider the stochastic process shown in Figure 6.2. Suppose that the initial state is 1 and that the statistic we wish to compute is the fraction of time spent in an odd-numbered state. Clearly, no matter how long the length of the trajectory, this time average will be 1.0. Now, consider an ensemble of instances of this process where the initial state is chosen equally probably as 1 or 2. For this ensemble, the limiting value of the statistic at any time step will be 0.5 and is the ensemble average. For this stochastic process, the time average differs from the ensemble average.

Intuitively, a stochastic process is **ergodic** if every statistic computed as a time average over a sufficiently long single trajectory can also be computed as an ensemble average over a sufficiently large number of trajectories. For this to be true, the sequence of states visited by the stochastic process over time should look statistically identical to the set of states occupied by an ensemble of processes at a single time step. We now consider the conditions under which a stochastic process is ergodic.[2]

2. Many mutually incompatible definitions of ergodicity exist in the literature. The definition presented here was chosen because it has a simple intuitive basis.

To begin with, we define a *state j* to be ergodic if it is recurrent non-null and aperiodic. Continuing with Example 6.3, it is a stair that the person will return to (recurrent), with a mean recurrence period that is finite (non-null), and such that the returning times do not have a least common divisor larger than 1 (aperiodic). If all the states in a Markov chain are ergodic, other than for a finite set of transient states, the chain itself is ergodic.[3] It can be shown that a finite aperiodic irreducible Markov chain is always ergodic. (In other words, all states of a finite irreducible Markov chain are recurrent non-null.)

EXAMPLE 6.14: ERGODIC MARKOV CHAIN

The Markov chain in Figure 6.1 is finite, aperiodic, and irreducible. Therefore, it is also ergodic.

A chain that is ergodic is insensitive to its initial state $\pi(0)$: Independent of its initial state, $\pi(n)$, the distribution of X_n (for reasonably large values of n) is the same. Nonergodic chains are (1) recurrent null (so that they may take a long time to return to some state), or (2) reducible (so that some parts of the chain do not communicate with others), or (3) periodic (so that quantities of interest also share the same period).

6.2.8 A Fundamental Theorem

We now have enough terminology to state (without proof) a fundamental theorem of queueing theory.

Theorem 6.11: The states of an irreducible Markov chain are one of the following: all transient, all recurrent null, or recurrent non-null. If any state is periodic, all states are periodic with the same period *r*.

Intuitively, this categorizes all Markov chains into a few types. The first are those where the process goes from state to state but never returns to any state. In this case, all states are transient. In the second and third types of chain, the process returns to at least one of the states. But the chain is irreducible, and so we can go from that state to all other states. Therefore, if the process can return to any one state, it can by definition return to all other states, which makes all states recurrent. In the second type, the transition probabilities are such that the expected recurrence period is infinite, so that all states are recurrent null. In the third type,

3. We allow a finite number of transient states in the chain because over sufficiently long trajectories or for a sufficiently large ensemble, their contribution to any statistic is negligible.

the expected recurrence period for all states is finite. For this type, we have two subtypes: the periodic recurrent non-null chains, whose states all share the same period, and the aperiodic recurrent non-null (ergodic) chains, for which no such period can be defined.

6.2.9 Stationary (Equilibrium) Probability of a Markov Chain

Recall that for a homogeneous Markov chain, the state-transition probabilities are time-independent. For a homogeneous chain, we expect that the probability of *being* in any particular state to also be time-independent. (If the probability of going from one state to another does not depend on time, the probability of being in any state shouldn't either.)

Of course, the probability of being in a particular state may depend on the initial state, especially for nonergodic chains, which are sensitive to their initial conditions. We define the **stationary probability distribution** of a Markov chain as follows: Suppose that we start with the initial distribution $\pi(0) = \pi^*$. Then, π^* is also the stationary distribution of the chain, if for all n, $\pi(n) = \pi^*$. Intuitively, if we start with the probability of being in each state j as defined by the stationary distribution, the transitions from each state according to the transition probabilities do not change the probability of being in each state.

EXAMPLE 6.15: STATIONARY DISTRIBUTION

Compute the stationary distribution of the Markov chain in Figure 6.3.

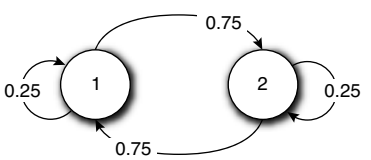


Figure 6.3 A simple Markov chain

Solution:

By symmetry, we guess that the stationary probability of being in state 1 is 0.5 and of being in state 2 is 0.5; that is, $\pi(0) = \pi^* = [0.5\ 0.5]$. After one time step, the probability of being in state 1 is 0.25*0.5 + 0.75*0.5, where the first term is the probability of remaining in state 1, and the second term is the probability of going from state 2 to state 1; we sum these probabilities because these are independent events. As expected, this sums to 0.5, so that the probability of being in state 1 in time step 2 is also 0.5. Symmetrically, if the probability of

being in state 2 at time 1 is 0.5, the probability of being in state 2 at time 2 is also 0.5. Therefore, the stationary probability distribution of this chain is indeed $\pi^* = [0.5\ 0.5]$.

6.2.10 A Second Fundamental Theorem

We now state a second fundamental theorem that allows us to compute stationary probabilities for any Markov chain.

Theorem 6.2: In an irreducible and aperiodic homogeneous Markov chain, the limiting probability distribution

$$\pi^* = \lim_{n \to \infty} \pi(n) \qquad \text{(EQ 6.4)}$$

always exists and is independent of the initial state probability distribution $\pi(0)$. Moreover, if all the states are ergodic (being recurrent non-null, in addition to being aperiodic), then π_j^*, the stationary probability of being in state j, is nonzero and can be uniquely determined by solving the following set of equations:

$$\begin{aligned} \sum_j \pi_j^* &= 1 \\ \pi_j^* &= \sum_i \pi_i^* p_{ij} \end{aligned} \qquad \text{(EQ 6.5)}$$

This theorem provides us with a simple set of equations to determine the stationary probability that the Markov chain is in any particular state. We need only verify that the set of states is finite, memoryless (i.e., satisfies the Markov property), irreducible (all states can be reached from each other), and aperiodic (e.g., because of at least one self-loop). These properties can be verified through simple inspection. Then, we can solve the preceding system of equations to obtain the stationary probability of being in each state.

EXAMPLE 6.16: STATIONARY PROBABILITY OF A MARKOV CHAIN

Compute the stationary probability for the Markov chain in Figure 6.1.

Solution:

Note that this chain is ergodic, so we can use Theorem 6.2 to obtain the following equations:

$$\begin{aligned}
\pi_1^* &= 0.2\pi_1^* + 1\pi_2^* + 0.5\pi_3^* \\
\pi_2^* &= 0.4\pi_1^* + 0\pi_2^* + 0.25\pi_3^* \\
\pi_3^* &= 0.4\pi_1^* + 0\pi_2^* + 0.25\pi_3^* \\
1 &= \pi_1^* + \pi_2^* + \pi_3^*
\end{aligned}$$

We solve this system of equations by using, for example, Gaussian elimination (see Section 3.4.2) to obtain $\pi_1^* = 15/31$; $\pi_2^* = 8/31$; $\pi_3^* = 8/31$, which is the stationary probability distribution of the chain.

6.2.11 Mean Residence Time in a State

Besides knowing the stationary probability of being in a particular state of a Markov chain, we would also like to know how long the process spends in each state. This duration can be computed by first obtaining the probability P(system stays in state j for m additional steps given that it just entered state j). The probability that the system stays in the same state after one time step is clearly p_{jj}. Moreover, after one time step, being Markovian, the process has no memory that it was in that state earlier. Therefore, the probability of staying in the state for m steps is given by $p_{jj}^m(1 - p_{jj})$, which is a geometrically distributed random variable with parameter $(1 - p_{jj})$ (see Section 1.5.3). This allows us to compute the mean of the distribution: the expected residence time in state j, as $1/(1 - p_{jj})$.

EXAMPLE 6.17: RESIDENCE TIME

Compute the residence times in each state of the Markov chain shown in Figure 6.1.

Solution:

$p_{11} = 0.2$, so E(residence time in state 1) = $1/(1 - 0.2) = 1/0.8 = 1.25$.

$p_{22} = 0$, so E(residence time in state 1) = $1/(1 - 0) = 1$.

$p_{33} = 0.25$, so E(residence time in state 1) = $1/(1 - 0.25) = 1/0.75 = 1.33$.

6.3 Continuous-Time Markov Chains

Our discussion so far has focused on discrete-time Markov chains, where state transitions happen every clock tick. We now turn our attention to continuous-time chains, where state transitions can happen independent of clock ticks. Most of the intuitions developed for discrete-time chains carry through to continuous-time

chains, with a few modifications. The main point of difference is that we need to consider the time instants t_1, t_2,... when state transitions occur, rather than assuming that a state transition occurs at every clock tick. We briefly state the main results for a continuous-time stochastic process and then focus on a specific type of continuous-time process: the birth-death process.

6.3.1 Markov Property for Continuous-Time Stochastic Processes

We first state the Markov property for continuous-time stochastic processes. The stochastic process $X(t)$ forms a continuous-time Markov chain if for all integers n and for any sequence of times $t_1, t_2,...,t_{n+1}$ such that $t_1 < t_2 <...< t_{n+1}$:

$$P(X(t_{n+1}) = j \mid X(t_1) = i_1, X(t_2) = i_2,...,X(t_n) = i_n) = \mathrm{P}(X(t_{n+1} = j) \mid X(t_n) = i_n) \quad \textbf{(EQ 6.6)}$$

Intuitively, this means that the future ($X(t_{n+1})$) depends on the past only through the current state i_n.

The definitions of homogeneity, irreducibility, recurrence, periodicity, and ergodicity introduced for discrete-time Markov chains in Section 6.2.3 continue to hold for continuous-time chains with essentially no change, so we will not restate them here.

6.3.2 Residence Time in a Continuous-Time Markov Chain

Analogous to the geometric distribution of residence times in a discrete-time chain, residence times are exponentially distributed for a continuous-time Markov chain for essentially the same reasons. If we denote the residence time in state j by R_j, the exponential distribution gives us the memorylessness property:

$$P(R_j > s + t | R_j > s) = P(R_j > t) \quad \textbf{(EQ 6.7)}$$

6.3.3 Stationary Probability Distribution for a Continuous-Time Markov Chain

The definition of the stationary probability of a continuous-time Markov chain closely follows that of a discrete-time Markov chain. Therefore, we omit the intermediate details and present the set of equations necessary to compute the stationary probability of a continuous-time homogeneous Markov chain.

Define the transition probability of going from state i to state j by

$$p_{ij}(t) = P(X(s + t) = j | X(s) = i) \quad \textbf{(EQ 6.8)}$$

Intuitively, this means that if the process is at state i at any time s, the probability that it will get to state j after a time interval t is given by $p_{ij}(t)$. This is independent of the value of s because the process is homogeneous.

Define the quantity q_{ij}, which denotes the **rate** at which the process departs from state i to state j (where j and i differ) when in state i:

$$q_{ij} = \lim_{\Delta t \to 0} p_{ij}(\Delta t)/\Delta t \quad \text{(EQ 6.9)}$$

That is, the probability that the process transitions from i to j during any interval of length Δt time units, conditional on its already being at state i, is $q_{ij}\Delta t$. We also define the negative quantity q_{ii} by

$$q_{ii} = -\sum_{j \neq i} q_{ij} \quad \text{(EQ 6.10)}$$

Then, $-q_{ii}$ is the rate at which the process does *not* stay in state i (i.e., departs to some other state) during an interval of length Δt time units. Because $\sum_j p_{ij}(t) = 1$ (at any time t, the chain transitions to *some* state, including the current state), we see that

$$\sum_j q_{ij}(t) = 0 \quad \text{(EQ 6.11)}$$

With these quantities in hand, we can define the time evolution of the probability of being in state j at time t, defined as $\pi_j(t)$, by

$$\frac{d\pi_j(t)}{dt} = q_{jj}\pi_j(t) + \sum_{k \neq j} q_{kj}\pi_k(t) \quad \text{(EQ 6.12)}$$

For ergodic continuous-time Markov chains, as $t \to \infty$, these probabilities converge to the stationary probability distributions π_j^*, which are implicitly defined by

$$\begin{aligned} q_{jj}\pi_j^* + \sum_{k \neq j} q_{kj}\pi_k^* &= 0 \\ \sum_j \pi_j^* &= 1 \end{aligned} \quad \text{(EQ 6.13)}$$

Note that this is Equation 6.12 with the rate of change of the probability set to 0—which is what one would expect for a stationary probability—and with the time-dependent probabilities replaced by their time-independent limiting values.

This ends our brief summary of continuous-time Markov processes. Instead of studying general continuous-time processes, we instead focus on a smaller but very important subclass: continuous-time birth-death processes.

6.4 Birth-Death Processes

This section discusses a special class of continuous-time homogenous Markov chains having the property that state transitions are permitted from state j only to states $j - 1$ and $j + 1$ (if these states exist). This class is well suited to describe such processes as the arrival and departure of customers from a queue (the subject of queueing theory, after all!), where the state index corresponds to the number of customers awaiting service. More precisely, if the number of customers in the system is j, the Markov chain is considered to be in state j. Customer arrivals cause the number of customers in the system to increase by one, which moves the process to state $j + 1$, and this happens at a rate $q_{j,j+1}$. Similarly, customer departures (due to service) cause the process to move from state j to state $j - 1$, and this happens at the rate $q_{j,j-1}$. In keeping with standard terminology, we denote:

$$\begin{aligned} \lambda_j &= q_{j,j+1} \\ \mu_j &= q_{j,j-1} \end{aligned} \qquad \textbf{(EQ 6.14)}$$

as the **birth** and **death rates**, respectively. Note that these rates can be state-dependent but cannot be time-dependent, owing to homogeneity. Also, by definition, at state i, the transition rates q_{ij} are 0 for all j other than i, $i - 1$, and $i + 1$. Given this fact, Equation 6.11, and Equation 6.14, we find that

$$q_{jj} = -(\lambda_j + \mu_j) \qquad \textbf{(EQ 6.15)}$$

For a birth-death process, being in state j has the intuitive meaning that the population size is j; that is, there are j customers in the queueing system. Note that when $j = 1$, we have one customer in the system—the customer that is receiving service—and *none* are in the queue. Generalizing, in state j, we have one customer receiving service and $j - 1$ in the queue awaiting service.

6.4.1 Time-Evolution of a Birth-Death Process

Because a birth-death process is a continuous-time Markov chain, the time evolution of $\pi_j(t)$ is given by Equation 6.12. We substitute Equation 6.14 and Equation 6.15 to find

$$\begin{aligned} \frac{d\pi_j(t)}{dt} &= -(\lambda_j + \mu_j)\pi_j(t) + \lambda_{j-1}\pi_{j-1}(t) + \mu_{j+1}\pi_{j+1}(t) && \text{j} \geq 1 \\ \frac{d\pi_0(t)}{dt} &= -\lambda_0\pi_0(t) + \mu_1\pi_1(t) && \text{j} = 0 \end{aligned} \qquad \textbf{(EQ 6.16)}$$

This describes the time evolution of a birth-death system. In practice, solving these equations is complex and does not give too many insights into the structure of the system. These are better obtained from the stationary probability distribution, which we study next.

6.4.2 Stationary Probability Distribution of a Birth-Death Process

Because a birth-death process is an ergodic continuous-time Markov chain, its stationary probability distribution is given by Equation 6.13. Denoting the stationary probability of being in state j by $\pi_j^* P_j$ and substituting Equation 6.14 and Equation 6.15 into Equation 6.13, we obtain the following equations:

$$\begin{aligned} 0 &= -(\lambda_j + \mu_j)\pi_j^* + \lambda_{j-1}\pi_{j-1}^* + \mu_{j+1}\pi_{j+1}^* && \text{j} \geq 1 \\ 0 &= -\lambda_0\pi_0^* + \mu_1\pi_1^* && \text{j} = 0 \\ & \sum_j \pi_j^* = 1 \end{aligned} \qquad \text{(EQ 6.17)}$$

In matrix form, we can write the first two equations as

$$\boldsymbol{PQ} = \mathbf{0} \qquad \text{(EQ 6.18)}$$

where

$$\boldsymbol{P} = \begin{bmatrix} \pi_0^* & \pi_1^* & \pi_2^* & \ldots \end{bmatrix}$$

$$\boldsymbol{Q} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & \ldots & \ldots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & \ldots & \ldots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix} \qquad \text{(EQ 6.19)}$$

The two matrices are infinite-dimensional if the population size is unbounded. Moreover, by defining $\boldsymbol{P}(t)$ as

$$\boldsymbol{P}(t) = \begin{bmatrix} \pi_0(t) & \pi_1(t) & \pi_2(t) & \ldots \end{bmatrix}$$

we can rewrite Equation 6.16 as

$$\text{d}\boldsymbol{P}(t)/\text{dt} = \boldsymbol{P}(t)\boldsymbol{Q} \qquad \text{(EQ 6.20)}$$

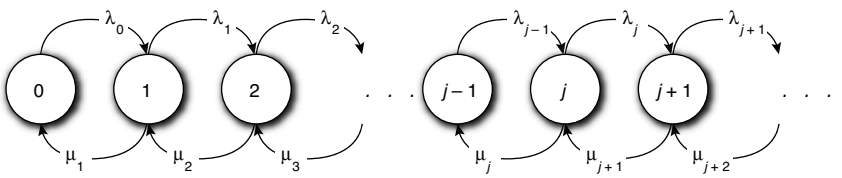


Figure 6.4 State-transition-rate diagram for a birth-death process

6.4.3 Finding the Transition-Rate Matrix

The $\boldsymbol{Q}$ matrix defined by Equation 6.19 is also called the **transition-rate matrix**. It is important because it allows us to derive both the time-dependent evolution of the system (i.e., $P_j(t)$), through Equation 6.20, and the long-term probability of being in state j, through Equation 6.18. Thus, in practice, the first step in studying a birth-death process is to write down its $\boldsymbol{Q}$ matrix.

Consider the representation of a generic birth-death process in Figure 6.4. Here, we represent each state j by a circle, and we label the arc from state j to state k with the transition rate q_{jk}. From this figure, we can determine $\boldsymbol{Q}$ for a birth-death process as follows. Note that the diagonal elements of $\boldsymbol{Q}$ (i.e., q_{jj}) are the negative of the quantities *leaving* state j. Focusing on the jth column, the $q_{j-1,j}$th elements, immediately above the diagonal, such as element q_{01}, are the rates entering state j from state $j - 1$, i.e., λ_{j-1} and the $q_{j+1,j}$th elements, such as element q_{32}, immediately below the diagonal, are the rates entering state j from state $j + 1$. All other elements are 0. In each row, the quantities sum to zero, due to Equation 6.11.

Thus, given the state-transition-rate diagram, it is possible to quickly construct $\boldsymbol{Q}$ and use it to obtain the time-dependent and time-independent (stationary) probabilities of being in each state. We now use this approach to study some standard birth-death systems. Note that this inspection approach also applies to all continuous-time Markov chains, where we can determine the elements of the $\boldsymbol{Q}$ matrix by inspecting the corresponding state-transition-rate diagram, then solving for $\boldsymbol{P}$ and $\boldsymbol{P}(t)$ by using Equation 6.18 and Equation 6.20, respectively.

EXAMPLE 6.18: TRANSITION-RATE MATRIX FOR A BIRTH-DEATH PROCESS

Consider the state-rate-transition diagram in Figure 6.5. What are the $\boldsymbol{P}$ and $\boldsymbol{Q}$ matrices for this system? What are the equations for its time evolution and the long-term probability of being in each state?

The $\boldsymbol{P}$ matrix is $\left[\pi_0^* \; \pi_1^* \; \pi_2^* \; \pi_3^*\right]$ and $\boldsymbol{P}(t) = \left[\pi_0(t) \; \pi_1(t) \; \pi_2(t) \; \pi_3(t)\right]$. By inspection, we can write the $\boldsymbol{Q}$ matrix as

$$\mathbf{Q} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 5 & -10 & 5 & 0 \\ 0 & 8 & -12 & 4 \\ 0 & 0 & 10 & -10 \end{bmatrix}$$

Therefore, the time evolution of state probabilities is given by

$$\frac{d}{dt}\left(\begin{bmatrix} \pi_0(t) & \pi_1(t) & \pi_2(t) & \pi_3(t) \end{bmatrix}\right) = \begin{bmatrix} \pi_0(t) & \pi_1(t) & \pi_2(t) & \pi_3(t) \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 & 0 \\ 5 & -10 & 5 & 0 \\ 0 & 8 & -12 & 4 \\ 0 & 0 & 10 & -10 \end{bmatrix}$$

$$\begin{bmatrix} \dot{\pi}_0(t) & \dot{\pi}_1(t) & \dot{\pi}_2(t) & \dot{\pi}_3(t) \end{bmatrix} = \begin{bmatrix} \pi_0(t) & \pi_1(t) & \pi_2(t) & \pi_3(t) \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 & 0 \\ 5 & -10 & 5 & 0 \\ 0 & 8 & -12 & 4 \\ 0 & 0 & 10 & -10 \end{bmatrix}$$

which is a system of differential equations that can be solved to give the evolution of the time-varying probabilities $\pi_i(t)$.

The long-term probability of being in each state is given by

$$\begin{bmatrix} \pi_0^* & \pi_1^* & \pi_2^* & \pi_3^* \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 & 0 \\ 5 & -10 & 5 & 0 \\ 0 & 8 & -12 & 4 \\ 0 & 0 & 10 & -10 \end{bmatrix} = 0$$

which is a system of linear equations in four variables that can be solved to obtain the stationary probability distribution of the chain.

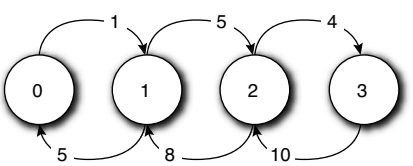


Figure 6.5 A simple birth-death process

6.4.4 A Pure-Birth (Poisson) Process

Consider a system in which $\lambda_j = \lambda$ for all j (the departure rate from all states is the same) and $\mu_j = 0$ for all j (the death rate is 0). This represents a **Poisson** process whose population grows without bound and whose rate of growth is λ independent of the population size, that is, we expect the population to grow by 1 every $1/\lambda$ seconds independent of the current population size.

We study two properties of this process: the probability of being in state j at time t, which corresponds to having j arrivals in time t, and the distribution of interarrival times, that is, the expected time between going from any state to the adjacent state.

We can derive the probability of being in any state directly from Equation 6.16. Substituting the values for λ and μ in this equation, we get

$$\begin{aligned} \frac{d\pi_j(t)}{dt} &= -\lambda\pi_j(t) + \lambda\pi_{j-1}(t) \qquad j \geq 1 \\ \frac{d\pi_0(t)}{dt} &= -\lambda\pi_0(t) \end{aligned} \tag{EQ 6.21}$$

The second equation is a trivial differential equation whose solution is given by

$$\pi_0(t) = e^{-\lambda t} \tag{EQ 6.22}$$

We substitute this into the first equation to get

$$\frac{d\pi_1(t)}{dt} = -\lambda\pi_1(t) + \lambda e^{-\lambda t} \tag{EQ 6.23}$$

whose solution is

$$\pi_1(t) = \lambda e^{-\lambda t} \tag{EQ 6.24}$$

By repeatedly substituting this into the first equation, we obtain

$$\pi_j(t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t} \tag{EQ 6.25}$$

which is the density function for the Poisson distribution (see Section 1.5.4) with parameter λt. Thus, for a Poisson process with parameter λ, the probability of j arrivals in time t, which is also the probability of being in state j at time t, is given by a Poisson distribution with parameter λt. Because the mean of the Poisson distribution is also its parameter, the expected number of arrivals in time t is λt. This is intuitively pleasing: The arrival rate is λ, so in time t we should see, on average, λt arrivals.

EXAMPLE 6.19: POISSON PROCESS

Consider packets arriving to a link as a Poisson process at a mean rate of five packets/second. What is the probability that the link receives two packets after 2 seconds and ten packets after 2 seconds?

Solution:

We have $\lambda = 5$ and $t = 2$, so the Poisson parameter is 10. The probability of having two packets arrive to the link after 2 seconds is $\pi_2(2) = (10^2/2!)\, e^{-10} = 50 * e^{-10} = 2.26 * 10^{-3}$. This is a rather unlikely event.

The probability of having ten packets arrive to the link after 2 seconds is $\pi_{10}(2) = (10^{10}/10!)\, e^{-10} = 0.125$. Note that the expected number of packet arrivals after 2 seconds is ten, yet the probability that the expected number of packets is actually achieved is only 1 in 8!

We now derive the interarrival time distribution for a Poisson process. Let a denote the continuous random variable that represents the time between any two arrivals: We seek the distribution for a. Consider the cumulative density function of a, given by the probability $P(a \le t) = 1 - P(a > t)$. But $P(a > t)$ is just P(0 customer arrivals in time $(0, t)$) $= 1 - \pi_0(t) = 1 - e^{-\lambda t}$, $t \ge 0$. The density function is given by differentiating this expression to get

$$a(t) = \lambda e^{-\lambda t} \qquad \textbf{(EQ 6.26)}$$

We recognize this as an exponential distribution (see Section 1.6.3). This gives us the following important result:

> The interarrival times for a Poisson process
> are drawn from an exponential distribution.

We note that the exponential distribution is memoryless. Thus, for a Poisson process, not only is the rate of transitioning to the next state (the birth rate) independent of the current population size, but also the *time* at which this transition occurs does not depend on how long the process has been at the current population size.

6.4.5 Stationary Probability Distribution for a Birth-Death Process

We now return to computing the stationary probability distribution for a general birth-death process, using Equation 6.17. From the second equation, we immediately obtain

$$\pi_1^* = \frac{\lambda_0}{\mu_1}\pi_0^* \quad \text{(EQ 6.27)}$$

Substituting this into the first equation, we find that

$$\pi_2^* = \frac{\lambda_0\lambda_1}{\mu_1\mu_2}\pi_0^* \quad \text{(EQ 6.28)}$$

Repeating this substitution, we find that P_j is given by

$$\pi_j^* = \frac{\lambda_0\lambda_1\ldots\lambda_{j-1}}{\mu_1\mu_2\ldots\mu_j}\pi_0^* = \pi_0^*\prod_{i=0}^{j-1}\frac{\lambda_i}{\mu_{i+1}} \quad \text{(EQ 6.29)}$$

We therefore obtain the long-term probabilities of being in any state j as a function of the probability of being in state 0 and the system parameters. Knowing that these probabilities sum to 1, we can determine

$$\pi_0^* = \frac{1}{1+\sum_{j=1}^{\infty}\prod_{i=0}^{j-1}\frac{\lambda_i}{\mu_{i+1}}} \quad \text{(EQ 6.30)}$$

This can be substituted back into Equation 6.29 to obtain the long-term probability of being in any state j. Of course, we need to ensure that the series in the denominator of Equation 6.30 actually converges! Otherwise, π_0^* is undefined, and so are all the other π_i^*. It turns out that the condition for convergence, as well as for the chain to be ergodic, is the existence of a value j_0 such that for all values of $j > j_0$, $\lambda_j < \mu_j$. We interpret this to mean that after the population reaches some threshold j_0, the rate of departures must exceed the rate of arrivals. This makes intuitive sense: Otherwise, the population size will grow (in expectation) without bound, and the probability of any particular population size will be 0.

EXAMPLE 6.20: GENERAL EQUILIBRIUM SOLUTION

Find the equilibrium probabilities of being in each state for the birth-death process shown in Figure 6.5.

Solution:

From Equation 6.30, we get

$$\pi_0^* = 1/[1 + 1/5 + (1 * 5)/(5 * 8) + (1 * 5 * 4)/5 * 8 * 10)] = 0.73$$

This can be substituted into Equation 6.29 to obtain

$$\pi_1^* = 1/5\ \pi_0^* = 0.2 * 0.73 = 0.146.$$

$$\pi_1^* = 1/8\ \pi_0^* = 0.125 * 0.73 = 0.09.$$

$$\pi_2^* = 1/20\ \pi_0^* = 0.05 * 0.73 = 0.0365.$$

As a check, 0.73 + 0.146 + 0.09 + 0.0365 = 1.0025, which is within the rounding error.

6.5 The M/M/1 Queue

The M/M/1 queue is the simplest nontrivial queueing system. Here, the a/b/c notation, also called **Kendall notation**, denotes the following.

- The "a" portion in the notation, the arrival process, is Markovian; it is a Poisson process with exponentially distributed interarrival times.
- The "b" portion in the notation, the departure process, is Markovian; it is a Poisson process with exponentially distributed interdeparture times.
- The "c" portion in the notation, the system is a single server.

Extended forms of the notation describe the size of the buffers available—we assume an infinite number—as well as the service discipline—we assume first come, first served—and other queueing parameters. However, the three-parameter version of the notation is the one that is commonly used.

The M/M/1 queueing system is a birth-death Markov process with a state-independent arrival rate λ and a state-independent departure rate μ. These rates are therefore also independent of the population size. This is counterintuitive, in that the rate of departure from a small population is the same as the rate of departure from a large population.

We study the long-term behavior of the M/M/1 queue by removing state dependence (the subscript j) in the transition rates in the analysis of Section 6.4.5. From Equation 6.29, we find that

$$\pi_j^* = \pi_0^* \prod_{i=0}^{j-1} \frac{\lambda}{\mu} = \pi_0^* \left(\frac{\lambda}{\mu}\right)^j \qquad j \geq 0 \qquad \textbf{(EQ 6.31)}$$

To obtain π_0^*, we use Equation 6.30 to obtain

$$\pi_0^* = \frac{1}{\left(1 + \sum_{j=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^j\right)} \quad \text{(EQ 6.32)}$$

When $\lambda < \mu$, the infinite sum in the denominator converges, and the denominator reduces to $\left(\frac{1}{1-\frac{\lambda}{\mu}}\right)$, so that

$$\pi_0^* = 1 - \frac{\lambda}{\mu} \quad \text{(EQ 6.33)}$$

The ratio λ/μ represents the intensity of the arrival rate as a fraction of the service rate and can be viewed as the utilization of the system. The value is important enough that it deserves its own symbol, ρ, which allows us to write Equation 6.33 as

$$\pi_0^* = 1 - \rho \quad \text{(EQ 6.34)}$$

This equation has the intuitive meaning that the probability that the system is idle (i.e., π_0^*) is (1 – utilization). It can be shown that this relationship is true for *all* queueing systems whose population size is unbounded.

We now use Equation 6.31 to obtain the stationary probability of being in any state j as

$$\pi_j^* = \rho^j(1 - \rho) \quad \text{(EQ 6.35)}$$

Note that this is a geometric distribution.

EXAMPLE 6.21: M/M/1 QUEUE

Consider a link to which packets arrive as a Poisson process at a rate of 300 packets/sec such that the time taken to service a packet is exponentially distributed. Suppose that the mean packet length is 500 bytes and that the link capacity is 1.5 Mbps. What is the probability that the link's queue has one, two, and ten packets, respectively?

Solution:

The packet length is 500 bytes = 4,000 bits, so the link service rate of 1,500,000 bits/sec = 375 packets/sec. Therefore, the utilization is 300/375 = 0.8. When the link queue has one packet, it is in state $j = 2$, because one packet is being served at that time. Thus, we need $\pi_2^* = 0.8^2 * 0.2 = 0.128$. For the queue having two packets, we compute $\pi_3^* = 0.8^3 * 0.2 = 0.1$. For ten packets in the queue, we compute $\pi_{11}^* = 0.8^{11} * 0.2 = 0.0067$, a fairly small quantity. Thus, even when

the utilization is high (80%), the queue size is quite small, rarely exceeding ten packets.

Note that the long-term probability that the population size is j depends only on the utilization of the system. As the utilization increases, the probability of reaching larger population sizes increases. To see this analytically, consider the mean number of customers in the system, which is also the mean population size, denoted $\bar{N}$ defined by

$$\bar{N} = \sum_{j=0}^{\infty} j\pi_j^* \qquad \textbf{(EQ 6.36)}$$

It can be shown that when this sum converges (i.e., when $\lambda < \mu$):

$$\text{Mean number of customers in the system} = \bar{N} = \frac{\rho}{(1-\rho)} \qquad \textbf{(EQ 6.37)}$$

EXAMPLE 6.22: MEAN NUMBER OF CUSTOMERS IN THE QUEUE

Compute the mean number of packets in the system of Example 6.19.

Solution:

The utilization is 0.8, so the mean number of packets in the system is $0.8/(1 - 0.8) = 0.8/0.2 = 4$. Of these, we expect three to be in the queue, and one to be in service.

It is obvious from Equation 6.37 that as $\rho \to 1$, $\bar{N} \to \infty$. That is, as the arrival rate approaches the service rate, the expected number of customers in the system grows without bound. This is somewhat unexpected: After all, the arrival rate is smaller than the service rate. Why, then, should the number of customers grow? The reason is that we are dealing with stochastic processes. Even though the arrival rate, on average, is lower than the service rate, there will be time periods when the short-term arrival rate exceeds the service rate. For instance, even if the mean arrival rate is one customer per second, there will be short intervals during which two or even three customers may arrive in 1 second. During this time, the queue builds up and is drained when the service rate exceeds the arrival rate. In fact, there is an interesting asymmetry in the system: When the short-term arrival rate exceeds the short-term service rate, the queue builds up, but when the service rate exceeds the arrival rate, if the queue is empty, the system does not build up "service credits." The server is merely idle. Thus, the system tends to build up queues that are drained only over time. This is reflected in the fact that as the utilization of the system increases, the mean number of customers in the system increases sharply.

It is remarkable that this fundamental insight into the behavior of a real queueing system can be derived with only elementary queueing theory. Moreover, this insight carries over to all other queueing systems: As the utilization approaches 1, the system becomes **congested**. The behavior of the mean queue length, which also corresponds to the waiting time, through Little's theorem, is shown in Figure 6.6.

It is clear that the queue length asymptotes to infinity as the utilization approaches 1. In networking terms, this means that as the arrival rate approaches a link's capacity, the queue at the immediately preceding router or switch will grow without bound, causing packet loss. This analysis allows us to derive a practical guideline: We should provision enough service capacity so that the system utilization never exceeds a threshold of around 70%. Alternatively, if this threshold is exceeded, either service requests should be dropped, or new service capacity should be made available so that the utilization decreases.

Another related quantity of interest for this queue is the mean waiting time in the queue. From Little's theorem, the mean number of customers in the system is the product of their mean waiting time and their mean arrival rate, so $\frac{\rho}{(1-\rho)}c$ = mean waiting time *λ, which means that

$$\text{Mean waiting time} = \frac{\frac{\rho}{\lambda}}{(1-\rho)} = \frac{\frac{1}{\mu}}{(1-\rho)} \qquad \textbf{(EQ 6.38)}$$

This quantity also grows without bound as the utilization approaches 1.

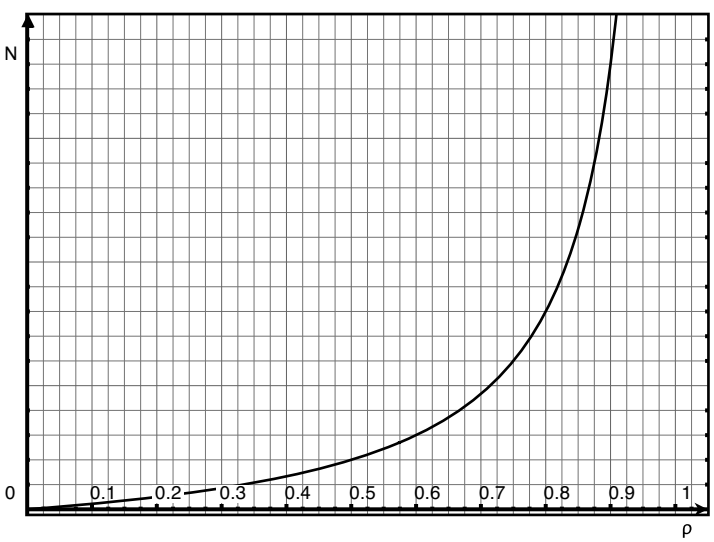


Figure 6.6 Mean queue length as a function of utilization

EXAMPLE 6.23: M/M/1 QUEUE WAITING TIME

What is the mean waiting time for a packet in the queue described in Example 6.20?

Solution:

For this queue, $\mu = 375$ and $\rho = 0.8$. So, the mean waiting time is $(1/375)/(1 - 0.8) = 5/375$ seconds = 13.3 ms.

6.6 Two Variations on the M/M/1 Queue

We now briefly consider two variations on the M/M/1 queue, essentially to give insight into how one proceeds with the analysis of a queueing system.

6.6.1 The M/M/∞ Queue: A Responsive Server

Suppose that a provider of service capacity brings on a new server to serve every arriving customer. This is like a private bank where new agents are brought on to provide individual attention to each customer when she or he arrives. This system can be modeled as a queue with an infinite number of servers, though, at any time, the number of servers is finite.

We can model and analyze this queue by using the same techniques as with an M/M/1 queue. We start with the state-transition-rate diagram shown in Figure 6.7. Note that μ_j, the rate of departure from the jth queue, is $j\mu$, which models the fact that when there are j customers, there are j servers. From the diagram, we can directly invoke Equation 6.29 to write down π_j^*, the stationary probability of being in state j, as

$$\pi_j^* = \pi_0^* \prod_{i=0}^{j-1} \frac{\lambda}{(i+1)\mu} = \pi_0^* \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} \qquad \text{(EQ 6.39)}$$

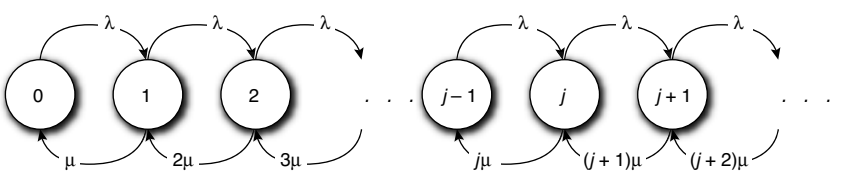


Figure 6.7 State-transition-rate diagram for an M/M/∞ queue

We solve for π_0^* by using Equation 6.30, as

$$\pi_0^* = \frac{1}{\left[1 + \sum_{j=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}\right]} \quad \text{(EQ 6.40)}$$

Recalling the standard expansion $e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}$, we see that

$$\pi_0^* = e^{-\frac{\lambda}{\mu}} \quad \text{(EQ 6.41)}$$

and

$$\pi_j^* = \pi_0^* \prod_{i=0}^{j-1} \frac{\lambda}{(i+1)\mu} = e^{-\frac{\lambda}{\mu}} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} \quad \text{(EQ 6.42)}$$

Equation 6.42 shows that the stationary probability of being in state *j* is given by the Poisson distribution with parameter λ/μ. Thus, with "infinite" servers, the number of customers in the queue follows the Poisson distribution. This allows us to compute the expected number of customers in the queue as the mean of the Poisson, which is its parameter: λ/μ. All other parameters of interest for this queueing system can be derived from Equation 6.42.

EXAMPLE 6.24: RESPONSIVE SERVER

Suppose that customers arrive at a private bank, modeled as a responsive server, as a Poisson process at the rate of ten customers/hour. Suppose that a customer's needs can be met on average in 20 minutes and that the service-time distribution is exponentially distributed. What is the probability that there are five customers in the bank at any point in time?

Solution:

We have $\lambda = 10$ and $\mu = 3$ (i.e., three customers can be served an hour, on average, by a server). Thus, $\pi_0^* = e^{-10/3} = 0.036$. We need to find $\pi_5^* = 0.036 * (10/3)^5 * 1/5! = 0.123$. Thus, there is a nearly one in eight chance that there will be five customers in the bank at any given time.

6.6.2 M/M/1/K: Bounded Buffers

Suppose that the queueing system has only $K - 1$ buffers. In this case, the population size, including the customer in service, cannot grow beyond K, and arrivals when the system is in state K are lost, similar to packet loss when arriving at a full queue. To model this, we can simply ignore arrivals to state K, which means that we will never enter states $K + 1$, $K + 2$,.... This results in a state-transition-rate diagram shown in Figure 6.8.

The state transition rates are therefore

$$\lambda_j = \begin{cases} \lambda & j < K \\ 0 & j \geq K \end{cases} \quad \text{(EQ 6.43)}$$

and

$$\mu_j = \mu \qquad j=1,2,...,K$$

We can therefore use Equation 6.29 to write π_j^* as

$$\pi_j^* = \begin{cases} \pi_0^* \left(\frac{\lambda}{\mu}\right)^j & j \leq K \\ 0 & j > K \end{cases} \quad \text{(EQ 6.44)}$$

We use Equation 6.30 to obtain

$$\pi_0^* = \frac{1}{\left[1 + \sum_{j=1}^{K} \left(\frac{\lambda}{\mu}\right)^j\right]} \quad \text{(EQ 6.45)}$$

Given the standard result $\sum_{k=0}^{n-1} r^k = \frac{1-r^n}{1-r}$, we can simplify this to

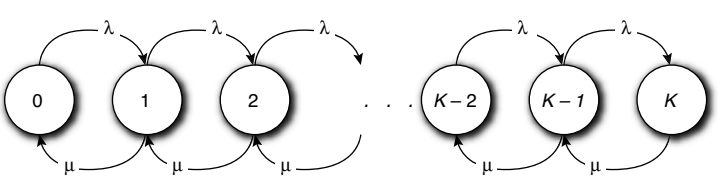


Figure 6.8 State-transition-rate diagram for an M/M/1/K queue

$$\pi_0^* = \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{K+1}} \quad \textbf{(EQ 6.46)}$$

So, we can now write Equation 6.44 as

$$\pi_j^* = \begin{cases} \dfrac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{K+1}} \left(\frac{\lambda}{\mu}\right)^j = \dfrac{1-\rho}{1-\rho^{K+1}}\rho^j & j \le K \\ 0 & j > K \end{cases} \quad \textbf{(EQ 6.47)}$$

As before, given these probabilities, we can compute all quantities of interest about the queueing system, such as the distribution of the queue length, the mean number of customers in the queue, and the mean waiting time. In particular, the intuitive meaning of π_K^* is the probability that the system is "full" when it has a buffer of size $K - 1$. So, π_K^* can be interpreted as the **blocking probability** of an M/M/1/K queue. We can then choose K as a sizing parameter to make π_K^* as small as desired.

Note that in this system, $\pi_0^* \neq 1 - \rho$, because the system size is bounded. Therefore, the number of customers served in a chosen time period may be lower than what the utilization indicates, because customer arrivals when the queue is full are lost. Moreover, the system is stable by definition, independent of the utilization, because excess arrivals are automatically dropped.

EXAMPLE 6.25: M/M/1/K QUEUE

Consider the system of Equation 6.21 but with the restriction that the queue has only four buffers. What is the probability that three of these are in use? How many buffers should we provision to ensure that the blocking probability is no more than 10^{-6}?

Solution:

We have $K = 5$, and $\frac{\lambda}{\mu} = 0.8$. From Equation 6.46, $\pi_0^* = (1 - 0.8)/(1 - 0.8^6) = 0.27$. If three buffers are in use, the system is in state $j = 4$. From Equation 6.44, we get $\pi_4^* = 0.27(0.8)^4 = 0.11$.

To size the buffer, we have to choose K such that $\pi_K^* < 10^{-6}$. We solve for K^* by using the inequality $10^{-6} > ((0.2)(0.8)^K)/(1 - 0.8^{K+1})$, to obtain $K^* = 55$. Thus, we need 54 buffers to satisfy this blocking probability.

6.7 Other Queueing Systems

Advanced texts of queueing theory describe queueing systems that go beyond the Markovian and exponential framework. Those queueing systems become much more difficult to analyze, so we will merely state two important results.

6.7.1 M/D/1: Deterministic Service Times

Consider a queueing system in which arrivals are from a Poisson process, but service times are deterministic. That is, as long as the queue is nonempty, the interdeparture time is deterministic rather than exponentially distributed. Representing the interdeparture time (a constant) by μ and the utilization by $\rho = \lambda/\mu$, it can be shown that the system is stable (i.e., the queue length is finite) as long as $\lambda < \mu$. Moreover, the long-term probability that the number of customers in the system is j (i.e., P_j) is given by

$$\pi_j^* = \begin{cases} 1-\rho & j = 0 \\ (1-\rho)(e^{\rho}-1) & j = 1 \\ (1-\rho)\left(\sum_{i=0}^{j} \frac{(-1)^{j-i}(i\rho)^{j-i-1}(i\rho+j-i)e^{i\rho}}{(j-1)!}\right) & j > 1 \end{cases} \quad \text{(EQ 6.48)}$$

This allows us to derive the mean number of customers in the system as

$$\text{Mean customers in the system} = \rho + \frac{\rho^2}{2(1-\rho)} \quad \text{(EQ 6.49)}$$

and the mean response time as

$$\text{Mean response time} = \frac{1}{\mu} + \frac{\rho}{2\mu(1-\rho)} \quad \text{(EQ 6.50)}$$

Other quantities of interest regarding the M/D/1 queue can be found in standard texts on queueing theory.

6.7.2 G/G/1

Once the arrival and service processes become non-Poisson, the analysis of even a single queue becomes challenging. For such systems, few results are available other than Little's theorem, and if the queue size is unbounded, $\pi_0^* = 1 - \rho$. A detailed study of such queues is beyond the scope of this text.

6.7.3 Networks of Queues

So far, we have studied the behavior only of a single queue. This is like studying a network with a single router: not very interesting! What happens when we link the output of a queue to the input of another queue, as we do in any computer network? Intuitively, we are making the departure process of one queue the arrival process for the second queue. Moreover, we may have more than one departure process mix to form the arrival process. Can this be analyzed?

We represent this composite system, also called a *tandem of queues* as shown in Figure 6.9. Here, each queue is shown by a buffer (with customers or jobs in it) and a server (represented by a circle). Jobs served by the servers on the left enter the queue of the server on the right. Each queue and associated server is also called a *node* (drawing on the obvious graph analogy).

If all the queues on the left are M/M/1 queues, recall that their departure processes are Poisson. Moreover, it can be shown that the mixture of Poisson processes is also a Poisson process whose parameter is the sum of the individual processes. Therefore, the input to the queue on the right is a Poisson process that can be analyzed as an M/M/1 queue. This leads to the fundamental insight that a tandem of M/M/1 queues is analytically tractable. Because the departure process of an M/M/m queue (i.e., a queue with m servers) is also Poisson, this result holds true for tandems of M/M/m queues.

We can make things a bit more complicated: We can allow customers to enter *any* queue (node) as a Poisson process, and we can also allow customers that leave a node to exit the system altogether with some probability or join any other node in the system with some probability. Note that this can potentially lead to cycles, where customers go through some set of nodes more than once. Nevertheless, in 1963, the American mathematician J. R. Jackson was able to show that these networks behave as if each M/M/m queue was being fed by a single Poisson stream. Such networks are also called **Jacksonian** networks in his honor. For a Jacksonian network, we have a strong result: Let $\pi^*_{k_1 k_2 \ldots k_n}$ denote the long-term probability

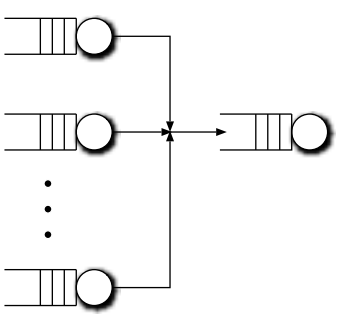


Figure 6.9 A network of queues

that there are k_1 customers at the first node, k_2 customers at the second node, and so on. Then:

$$\pi^*_{k_1 k_2 \ldots k_n} = \pi^*_{k_1} \pi^*_{k_2} \ldots \pi^*_{k_n} \quad \textbf{(EQ 6.51)}$$

That is, the joint probability of having a certain number of customers in each queue is the product of the individual probabilities. We interpret this to mean that each queue in the system acts as if it were independent of the others. This **product form** of the probability distribution greatly simplifies analysis.

Despite the elegance and power of Jacksonian network analysis, it can be rarely applied to study practical computer networks, because customers (packets) in real networks rarely arrive as a Poisson process. Thus, the output process is also non-Poisson, which makes subsequent analysis complex. In recent years, the development of network calculus and stochastic network calculus has allowed significant inroads into the study of the performance of non-Jacksonian networks.

6.8 Further Reading

The definitive introductory text on queueing theory is the two-volume text by L. Kleinrock, *Queueing Systems*, Wiley Interscience, 1975. A modern and thorough treatment of Markov chains can be found in P. Bremaud, *Markov Chains*, Springer, 1999. Further details on network calculus can be found in J.-Y. Le Boudec and P. Thiran, *Network Calculus*, Springer, 2001.

6.9 Exercises

1. Little's theorem

Patients arriving at the Grand River Hospital Emergency Room have a mean waiting time of 3 hours. It has been found that, averaged over the period of a day, patients arrive at the rate of one every 5 minutes.

a. How many patients are awaiting treatment on average at any given point in time?

b. What should the size of the waiting room be so that it can always accommodate arrivals?

2. A stochastic process

Consider that in Example 6.3, a person is on an infinite staircase on stair 10 at time 0 and potentially moves once every clock tick. Suppose that the person

moves from stair i to stair $i + 1$ with probability 0.2, and from stair i to stair $i - 1$ with probability 0.2 (the probability of staying on stair i is 0.6). Compute the probability that the person is on each stair at time 1 (after the first move), time 2, and time 3.

3. Discrete- and continuous-state and time processes

Come up with your own examples for all four combinations of discrete state/discrete time/continuous state/continuous time processes.

4. Markov process

Is the process in Exercise 2 a Markov process? Why or why not?

5. Homogeneity

Is the process in Exercise 2 homogeneous? Why or why not?

6. Representation

a. Represent the process in Exercise 2 by using a transition matrix and a state transition diagram.

b. Do the rows in this matrix have to sum to 1? Do the columns in this matrix have to sum to 1? Why or why not?

c. Now, assume that the staircase has only four steps. Make appropriate assumptions (what are these?) to represent this finite process as a transition matrix and a state-transition diagram.

7. Reducibility

Is the chain in Exercise 2 reducible? Why or why not?

8. Recurrence

Is state 1 in the chain in Exercise 6(c) recurrent? Compute $f_1^{\,1}$, $f_1^{\,2}$, and $f_1^{\,3}$.

9. Periodicity

Is the chain in Exercise 2 periodic? If not, give an example of a chain with period N for arbitrary $N > 1$.

10. Ergodicity

Is any state in the chain of Exercise 6(c) nonergodic? Why or why not?

11. Stationary probability

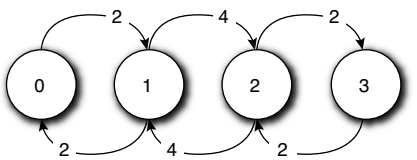
Compute the stationary probability distribution of the chain in Exercise 6(c).

12. Residence times

Compute the residence time in each state of the Markov chain in Exercise 6(c).

13. Stationary probability of a birth-death process

Consider the following state-rate-transition diagram.



a. Compare this with the state-transition probability diagram in Exercise 6(c). What features are the same, and what differ?

b. Write down the $\boldsymbol{Q}$ matrix for this system.

c. Use the $\boldsymbol{Q}$ matrix to compute the stationary-probability distribution of this chain.

14. Poisson process

Prove that the interdeparture time of a pure-death process is exponentially distributed.

15. Stationary probabilities of a birth-death process

Use Equation 6.30 to compute the stationary probability of the birth-death process in Exercise 13.

16. M/M/1 queue

Is the birth-death process in Exercise 13 M/M/1? Why or why not?

17. M/M/1 queue

Consider a link to which packets arrive as a Poisson process at a rate of 450 packets/sec such that the time taken to service a packet is exponentially distributed. Suppose that the mean packet length is 250 bytes and that the link capacity is 1 Mbps.

a. What is the probability that the link's queue has one, two, and ten packets respectively?

b. What is the mean number of packets in the system? What is the mean number in the queue?

c. What is the mean waiting time?

18. Responsive (M/M/∞) server

Compute the ratio of π_j^* for a responsive server to the same value for an M/M/1 queue. How does this ratio behave as a function of j?

19. M/M/1/K server

Assume that the queueing system in Exercise 17 has ten buffers. Compute an upper bound on the probability of packet loss.

20. M/D/1 queue

Compute the mean number of customers in an M/D/1 system that has a utilization of 0.9.

a. How does this compare with a similarly loaded M/M/1 system?

b. Compute the ratio of the mean number of customers as a function of ρ.

c. Use this to compare the behavior of an M/D/1 queue with that of an M/M/1 queue under heavy load.