MATH 4792/5792 Probabilistic Modeling

Yaning Liu

2024-08-01

Table of contents

Preface

This is a book based on the classnotes of MATH 4792/5792

1 Stochastic Processes

A **stochastic process** $\{X(t), t \in T\}$ is a collection of random variables. That is, for each $t \in T$, $X(t)$ is a random variable. The index t is often interpreted as time and, as a result, we refer to $X(t)$ as the **state** of the process at time t. For example, $X(t)$ might equal the total number of customers that have entered a supermarket by time t; or the number of customers in the supermarket at time t; or the total amount of sales that have been recorded in the market by time t; etc.

The set T is called the **index** set of the process. When T is a countable set the stochastic process is said to be a **discrete-time** process. If T is an interval of the real line, the stochastic process is said to be a **continuous-time** process. For instance, $\{X_n, n = 0, 1, \dots\}$ is a discrete-time stochastic process indexed by the nonnegative integers; while $\{X(t), t \geq 0\}$ is a continuous-time stochastic process indexed by the nonnegative real numbers.

The **state space** of a stochastic process is defined as the set of all possible values that the random variables $X(t)$ can assume.

Thus, a stochastic process is a family of random variables that describes the evolution through time of some (physical) process.

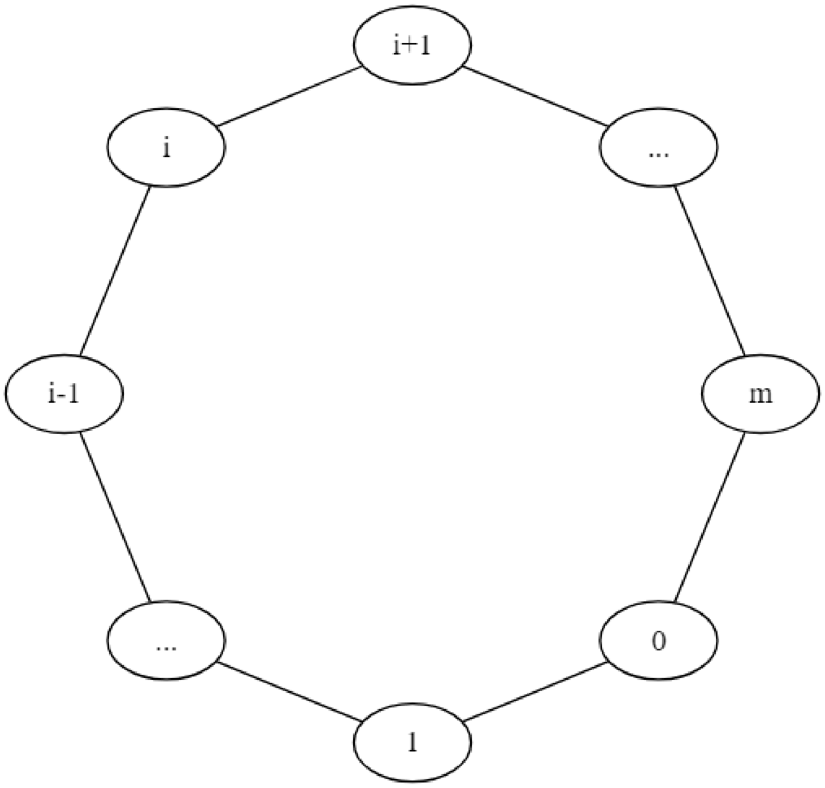
Example 1.1. Consider a particle that moves along a set of $m + 1$ nodes, labeled $0, 1, \dots, m$, that are arranged around a circle. At each step the particle is equally likely to move one position in either the clockwise or counterclockwise direction. That is, if X_n is the position of the particle after its nth step then

$$P\{X_{n+1} = i + 1 | X_n = i\} = P\{X_{n+1} = i - 1 | X_n = i\} = \frac{1}{2}$$

where $i + 1 \equiv 0$ when $i = m$, and $i - 1 \equiv m$ when $i = 0$. Suppose now that the particle starts at 0 and continues to move around according to the preceding rules until all the nodes $1, 2, \dots, m$ have been visited. What is the probability that node $i, i = 1, \dots, m$, is the last one visited?

Solution 1.1. Surprisingly enough, the probability that node i is the last node visited can be determined without any computations. To do so, consider the first time that the particle is at one of the two neighbors of node i, that is, the first time that the particle is at one of the nodes $i - 1$ or $i + 1$ (with $m + 1 \equiv 0$). Suppose it is at node $i - 1$ (the argument in the alternative situation is identical). Since neither node i nor $i+1$ has yet been visited, it follows that i will be the last node visited if and only if $i+1$ is visited before i. This is so because in order to visit $i+1$ before i the particle will have to visit all the nodes on the counterclockwise path from $i-1$ to $i+1$ before it visits i. But the probability that a particle at node $i-1$ will visit $i+1$ before i is just the probability that a particle will progress $m-1$ steps in a specified direction before progressing one step in the other direction. That is, it is equal to the probability that a gambler who starts with one unit, and wins one when a fair coin turns up heads and loses one when it turns up tails, will have his fortune go up by $m-1$ before he goes broke. Hence, because the preceding implies that the probability that node i is the last node visited is the same for all i, and because these probabilities must sum to 1, we obtain

$$P\{i \text{ is the last node visited}\} = 1/m, i = 1, \dots, m$$



Remark 1.1. The argument used in Example 25.1 also shows that a gambler who is equally likely to either win or lose one unit on each gamble will be down n before being up 1 with probability $1/(n+1)$; or equivalently

$$P\{\text{gambler is up 1 before being down } n\} = \frac{n}{n+1},\ i = 1, \dots, m$$

Suppose now we want the probability that the gambler is up 2 before being down n. Upon conditioning on whether he reaches up 1 before down n, we obtain that

$$\begin{aligned}
&P\{\text{ gambler is up 2 before being down } n\} \\
&\quad = P\{\text{ up 2 before down } n \mid \text{ up 1 before down } n\}\frac{n}{n+1} \\
&\quad = P\{\text{ up 1 before down } n+1\}\frac{n}{n+1} \\
&\quad = \frac{n+1}{n+2}\frac{n}{n+1} = \frac{n}{n+2}
\end{aligned}$$

Repeating this argument yields that

$$P\{\text{gambler is up } k \text{ before being down } n\} = \frac{n}{n+k}$$

Below, we use Python to simulate the process and numerically determine the probability that a node is the last one visited, and compare the numerical results with the theoretical ones.

```python
import numpy as np

def last_visited_node(m):
    '''
    Simulate the node that is last visited when all the m nodes numbered
    from 1 to m have been visited
    input:
    m: int, the m+1 nodes are numbered from 0 to m
    output:
    n: int, between 1 and m, the number of the node that is last visited
    '''
    current_state = 0
    unvisited_nodes = np.arange(1, m+1)

    while len(unvisited_nodes) != 0:
        rn = np.random.random()
        if rn > 0.5:
            if current_state < m:
                current_state += 1
            else:
                current_state = 0
            if current_state in unvisited_nodes:
                indices_to_delete = np.where(unvisited_nodes == current_state)[0]
                unvisited_nodes = np.delete(unvisited_nodes, indices_to_delete)
        else:
            if current_state != 0:
                current_state -= 1
            else:
                current_state = m
            if current_state in unvisited_nodes:
                indices_to_delete = np.where(unvisited_nodes == current_state)[0]
                unvisited_nodes = np.delete(unvisited_nodes, indices_to_delete)
    n = current_state
    return n

def prob_nodes(m, sample_size):
    """
    Compute the probability that each of the m nodes (numbered 1 to m) is visited
    lastly when all the m+1 nodes have been visited

    input:
    m: int, the m+1 nodes are numbered from 0 to m
    sample_size: int, sample size for the simulation, i.e.,
    how many experiments are performed

    output:
    probs: numpy array of shape (m,), the collection of m probabilities,
    each representing the chance that the corresponding node is visited
    lastly when all the m nodes have been visited
    """
    probs = np.zeros(m)
    for i in range(sample_size):
        node = last_visited_node(m)  # node can be an integer between 1 and m
        probs[node-1] += 1
    probs /= probs.sum()

    return probs
```

Consider a case where there are 11 nodes ($m = 10$), and we simulate the process 10000 times and compute the probability that each node is last visited . The result is in agreement with the theoretical one ($1/m$).

```
print(prob_nodes(10, 10000))
```

```
[0.099  0.0979 0.104  0.0975 0.107  0.0981 0.1005 0.1035 0.0959 0.0966]
```

2 Introduction to Markov Chains

2.1 Introduction

Consider a stochastic process $\{X_n, n = 0, 1, 2, \dots\}$ that takes on a finite or countable number of possible values. Unless otherwise mentioned, this set of possible values of the process will be denoted by the set of nonnegative integers $\{0, 1, 2, \dots\}$. If $X_n = i$, then the process is said to be in state i at time n. We suppose that whenever the process is in state i, there is a fixed probability P_{ij} that it will next be in state j. That is, we suppose that

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\} = P_{ij} \tag{2.1}$$

for all states $i_0, i_1, \dots, i_{n-1}, i, j$ and all $n \geq 0$. Such a stochastic process is known as a **Markov chain**. Equation 2.1 may be interpreted as stating that, for a Markov chain, the conditional distribution of any future state X_{n+1} given the past states $X_0, X_1, \dots, X_{n-1}$ and the present state X_n, is independent of the past states and depends only on the present state.

The value P_{ij} represents the probability that the process will, when in state i, next make a transition into state j. Since probabilities are nonnegative and since the process must make a transition into some state, we have that

$$P_{ij} \geq 0, \quad i, j \geq 0; \quad \sum_{j=0}^{\infty} P_{ij} = 1, \quad i = 0, 1, \dots$$

Let P denote the matrix of one-step transition probabilities P_{ij} , so that

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ P_{i0} & P_{i1} & P_{i2} & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Example 2.1. (Forecasting the Weather) Suppose that the chance of rain tomorrow depends on previous weather conditions only through whether or not it is raining today and not on past weather conditions. Suppose also that if it rains today, then it will rain tomorrow with probability α; and if it does not rain today, then it will rain tomorrow with probability β.

If we say that the process is in state 0 when it rains and state 1 when it does not rain, then the preceding is a two-state Markov chain whose transition probabilities are given by

$$P = \begin{bmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{bmatrix}$$

Example 2.2. (A Random Walk Model) A Markov chain whose state space is given by the integers $i = 0, \pm 1, \pm 2, ...$ is said to be a random walk if, for some number $0 < p < 1$,

$$P_{i,i+1} = p = 1 - P_{i,i-1}, \quad i = 0, \pm 1, ...$$

The preceding Markov chain is called a **random walk** for we may think of it as being a model for an individual walking on a straight line who at each point of time either takes one step to the right with probability p or one step to the left with probability $1-p$.

Example 2.3. (A Gambling Model) Consider a gambler who, at each play of the game, either wins \$1 with probability p or loses \$1 with probability $1-p$. If we suppose that our gambler quits playing either when he goes broke or he attains a fortune of $\$N$, then the gambler's fortune is a Markov chain having transition probabilities

$$\begin{aligned} P_{i,i+1} &= p = 1 - P_{i,i-1}, \quad i = 1, 2, \dots, N-1 \\ P_{00} &= P_{NN} = 1 \end{aligned}$$

States 0 and N are called **absorbing states** since once entered they are never left. Note that the preceding is a finite state random walk with absorbing barriers (states 0 and N).

Example 2.4. In most of Europe and Asia annual automobile insurance premiums are determined by use of a Bonus Malus (Latin for Good-Bad) system. Each policyholder is given a positive integer valued state and the annual premium is a function of this state (along, of course, with the type of car being insured and the level of insurance). A policyholder's state changes from year to year in response to the number of claims made by that policyholder. Because lower numbered states correspond to lower annual premiums, a policyholder's state will usually decrease if he or she had no claims in the preceding year, and will generally increase if he or she had at least one claim. (Thus, no claims is good and typically results in a decreased premium, while claims are bad and typically result in a higher premium.)

For a given Bonus Malus system, let $s_i(k)$ denote the next state of a policyholder who was in state i in the previous year and who made a total of k claims in that year. If we suppose that the number of yearly claims made by a particular policyholder is a Poisson random variable with parameter λ, then the successive states of this policyholder will constitute a Markov chain with transition probabilities

$$P_{i,j} = \sum_{k: s_i(k) = j} e^{-\lambda} \frac{\lambda^k}{k!}, \quad j \geq 0$$

Whereas there are usually many states (20 or so is not atypical), the following table specifies a hypothetical Bonus Malus system having four states.

		Next	state	if	
State	Annual Premium	0 claims	1 claim	2 claims	≥ 3 claims
1	200	1	2	3	4
2	250	1	3	4	4
3	400	2	4	4	4
4	600	3	4	4	4

Thus, for instance, the table indicates that $s_2(0) = 1$; $s_2(1) = 3$; $s_2(k) = 4$, $k \geq 2$. Consider a policyholder whose annual number of claims is a Poisson random variable with parameter λ. If a_k is the probability that such a policyholder makes k claims in a year, then

$$a_k = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \geq 0$$

For the Bonus Malus system specified in the preceding table, the transition probability matrix of the successive states of this policyholder is

$$P = \begin{bmatrix} a_0 & a_1 & a_2 & 1 - a_0 - a_1 - a_2 \\ a_0 & 0 & a_1 & 1 - a_0 - a_1 \\ 0 & a_0 & 0 & 1 - a_0 \\ 0 & 0 & a_0 & 1 - a_0 \end{bmatrix}$$

2.2 Chapman–Kolmogorov Equations

We have already defined the one-step transition probabilities P_{ij} . We now define the n-step transition probabilities P_{ij}^n to be the probability that a process in state i will be in state j after n additional transitions. That is,

$$P_{ij}^n = P\{X_{n+k} = j | X_k = i\}, \quad n \geq 0, i, j \geq 0$$

Of course $P_{ij}^1 = P_{ij}$. The **Chapman–Kolmogorov equations** provide a method for computing these n-step transition probabilities. These equations are

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m \quad \text{for all } n, m \geq 0, \text{ all } i, j \tag{2.2}$$

and are most easily understood by noting that $P_{ik}^n P_{kj}^m$ represents the probability that starting in i the process will go to state j in $n + m$ transitions through a path which takes it into state

k at the nth transition. Hence, summing over all intermediate states k yields the probability that the process will be in state j after $n+m$ transitions. Formally, we have

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P\{X_{n+m} = j | X_n = k, X_0 = i\} P\{X_n = k | X_0 = i\} = \sum_{k=0}^{\infty} P_{ik}^{n} P_{kj}^{m}$$

If we let $P^{(n)}$ denote the matrix of n-step transition probabilities P_{ij}^{n} , then Equation 2.2 asserts that

$$P^{(n+m)} = P^{(n)} \cdot P^{(m)}$$

where the dot represents matrix multiplication. Hence, in particular,

$$P^{(2)} = P^{(1+1)} = P \cdot P = P^2$$

and by induction

$$P^{(n)} = P^{(n-1+1)} = P^{n-1} \cdot P = P^n$$

That is, the n-step transition matrix may be obtained by multiplying the matrix P by itself n times.

Example 2.5. Consider Example 25.1 in which the weather is considered as a two-state Markov chain. If $\alpha = 0.7$ and $\beta = 0.4$, then calculate the probability that it will rain four days from today given that it is raining today.

Solution 2.1. The one-step transition probability matrix is given by

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

Hence,

$$P^{(2)} = P^2 = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \cdot \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix}$$

$$P^{(4)} = (P^2)^2 = \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix} \cdot \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix} = \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix}$$

and the desired probability P_{00}^{4} equals 0.5749.

So far, all of the probabilities we have considered are conditional probabilities. For instance, P_{ij}^{n} is the probability that the state at time n is j given that the initial state at time 0 is i. If the unconditional distribution of the state at time n is desired, it is necessary to specify the probability distribution of the initial state. Let us denote this by

$$\alpha_i \equiv P\{X_0 = i\}, \quad i \geq 0, \sum_{i=0}^{\infty} \alpha_i = 1$$

All unconditional probabilities may be computed by conditioning on the initial state. That is,

$$P\{X_n = j\} = \sum_{i=0}^{\infty} P\{X_n = j | X_0 = i\} P\{X_0 = i\} = \sum_{i=0}^{\infty} P_{ij}^n \alpha_i$$

For instance, if $\alpha_0 = 0.4$, $\alpha_1 = 0.6$, in Example 25.4, then the (unconditional) probability that it will rain four days after we begin keeping weather records is

$$P\{X_4 = 0\} = 0.4P_{00}^4 + 0.6P_{10}^4 = (0.4)(0.5749) + (0.6)(0.5668) = 0.5700$$

Suppose now that you want to determine the probability that a Markov chain enters any of a specified set of states $\mathcal{A}$ by time n. One way to accomplish this is to reset the transition probabilities out of states in $\mathcal{A}$ to

$$P\{X_{m+1} = j | X_m = i\} = \begin{cases} 1, & \text{if } i \in \mathcal{A}, j = i \\ 0, & \text{if } i \in \mathcal{A}, j \neq i \end{cases}$$

That is, transform all states in $\mathcal{A}$ into absorbing states which once entered can never be left. Because the original and transformed Markov chain follows identical probabilities until a state in $\mathcal{A}$ is entered, it follows that the probability the original Markov chain enters a state in $\mathcal{A}$ by time n is equal to the probability that the transformed Markov chain is in one of the states of $\mathcal{A}$ at time n.

Example 2.6. A pensioner receives 2 (thousand dollars) at the beginning of each month. The amount of money he needs to spend during a month is independent of the amount he has and is equal to i with probability $P_i, i = 1, 2, 3, 4, \sum_{i=1}^4 P_i = 1$. If the pensioner has more than 3 at the end of a month, he gives the amount greater than 3 to his son. If, after receiving his payment at the beginning of a month, the pensioner has a capital of 5, what is the probability that his capital is ever 1 or less at any time within the following four months?

Solution 2.2. To find the desired probability, we consider a Markov chain with the state equal to the amount the pensioner has at the end of a month. Because we are interested in whether this amount ever falls as low as 1, we will let 1 mean that the pensioner's end-of-month fortune has ever been less than or equal to 1. Because the pensioner will give any end-of-month amount greater than 3 to his son, we need only consider the Markov chain with states $1, 2, 3$ and transition probability matrix $Q = [Q_{i,j}]$ given by

$$\begin{bmatrix} 1 & 0 & 0 \\ P_3 + P_4 & P_2 & P_1 \\ P_4 & P_3 & P_1 + P_2 \end{bmatrix}$$

To understand the preceding, consider $Q_{2,1}$, the probability that a month that ends with the pensioner having the amount 2 will be followed by a month that ends with the pensioner having less than or equal to 1. Because the pensioner will begin the new month with the amount $2 + 2 = 4$, his ending capital will be less than or equal to 1 if his expenses are either 3 or 4. Thus, $Q_{2,1} = P_3 + P_4$. The other transition probabilities are similarly explained.

Suppose now that $P_i = 1/4, i = 1, 2, 3, 4$. The transition probability matrix is

$$\begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$$

Squaring this matrix and then squaring the result gives the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{222}{256} & \frac{13}{256} & \frac{21}{256} \\ \frac{201}{256} & \frac{21}{256} & \frac{34}{256} \end{bmatrix}$$

Because the pensioner's initial end of month capital was 3, the desired answer is $Q^4_{3,1} = \frac{201}{256}$.

Let $\{X_n, n \geq 0\}$ be a Markov chain with transition probabilities $P_{i,j}$. If we let $Q_{i,j}$ denote the transition probabilities that transform all states in $\mathcal{A}$ into absorbing states, then

$$Q_{i,j} = \begin{cases} 1, & \text{if } i \in \mathcal{A}, j = i \\ 0, & \text{if } i \in \mathcal{A}, j \neq i \\ P_{i,j}, & \text{otherwise} \end{cases}$$

For $i, j \notin \mathcal{A}$, the n stage transition probability $Q^n_{i,j}$ represents the probability that the original chain, starting in state i, will be in state j at time n without ever having entered any of the states in $\mathcal{A}$. For instance, in Example 13.5, starting with 5 at the beginning of January, the probability that the pensioner's capital is 4 at the beginning of May without ever having been less than or equal to 1 in that time is $Q^4_{3,2} = 21/256$.

We can also compute the conditional probability of X_n given that the chain starts in state i and has not entered any state in $\mathcal{A}$ by time n, as follows. For $i, j \notin \mathcal{A}$,

$$\begin{aligned} P\{X_n = j | X_0 = i, X_k \notin \mathcal{A}, k = 1, \ldots, n\} &= \frac{P\{X_n = j, X_k \notin \mathcal{A}, k = 1, \ldots, n | X_0 = i\}}{P\{X_k \notin \mathcal{A}, k = 1, \ldots, n | X_0 = i\}} \\ &= \frac{Q^n_{i,j}}{\sum_{r \notin \mathcal{A}} Q^n_{i,r}} \end{aligned}$$

Now we use Python to numerically compute the probability analytically calculated in Example 13.5.

```python
import numpy as np

def prob_1_or_less(P, sample_size):
    """
    Compute the probability that the capital is ever 1 or less at any time
    within the following four months

    input:
    P: a 1d numpy array of shape (4), the probability of spending i,
    i=1,2,3,4
    sample_size: int, sample size for the simulation, i.e., how many
    experiments are performed

    output:
    prob, float, the probability that the capital is ever 1 or less
    at any time within the following four months
    """
    count = 0  # count the number of times when the capital falls at or below 1
    P_cumsum = np.cumsum(P)  # cumulative sum of probabilities to simulate expenditures
    for i in range(sample_size):
        # initial capital. Use 3 instead of 5, since we add 2 at the beginning of each month
        capital = 3.0
        for j in range(4):  # simulate 4 months
            capital += 2.0
            expenditure = np.where((np.random.random()<P_cumsum) == 1)[0][0] + 1
            capital -= expenditure
            if capital <= 1.0:  # Check if capital is <= 1, if so, get out of the inner loop
                count += 1
                break
            if capital > 3:
                capital = 3.0
    prob = count / sample_size
    return prob
```

```python
P = np.array([0.25, 0.25, 0.25, 0.25])
sample_size = 100000
prob = prob_1_or_less(P, sample_size)
print('Computed probability is ', prob, 'and the theoretical probability is ', 201/256)
```

```
Computed probability is  0.78766 and the theoretical probability is  0.78515625
```

The numerical result verifies our theoretical result.

3 Classification of States for Markov Chains

3.1 Classification of States

State j is said to be accessible from state i if $P_{ij}^n > 0$ for some $n \geq 0$. Note that this implies that state j is accessible from state i if and only if, starting in i, it is possible that the process will ever enter state j . This is true since if j is not accessible from i, then

$$\begin{aligned} P\{\text{ever enter } j|\text{start in } i\} &= P\left\{\cup_{n=0}^{\infty}\{X_n = j\}|X_0 = i\right\} \\ &\leq \sum_{n=0}^{\infty} P\{X_n = j|X_0 = i\} = \sum_{n=0}^{\infty} P_{ij}^n = 0 \end{aligned}$$

Two states i and j that are accessible to each other are said to **communicate**, and we write $i \leftrightarrow j$.

Note that any state communicates with itself since, by definition,

$$P_{ii}^0 = P\{X_0 = i|X_0 = i\} = 1$$

The relation of communication satisfies the following three properties:

1. State i communicates with state i, all $i \geq 0$.
2. If state i communicates with state j, then state j communicates with state i.
3. If state i communicates with state j, and state j communicates with state k, then state i communicates with state k.

Properties 1 and 2 follow immediately from the definition of communication. To prove 3, suppose that i communicates with j , and j communicates with k. Thus, there exist integers n and m such that $P_{ij}^n > 0$, $P_{jk}^m > 0$. Now by the Chapman–Kolmogorov equations, we have that

$$P_{ik}^{n+m} = \sum_{r=0}^{\infty} P_{ir}^n P_{rk}^m \geq P_{ij}^n P_{jk}^m > 0$$

Hence, state k is accessible from state i. Similarly, we can show that state i is accessible from state k. Hence, states i and k communicate.

Two states that communicate are said to be in the same **class**. It is an easy consequence of 1, 2, and 3 that any two classes of states are either identical or disjoint. In other words, the concept of communication divides the state space up into a number of separate classes. The Markov chain is said to be **irreducible** if there is only one class, that is, if all states communicate with each other.

Example 3.1. Consider the Markov chain consisting of the three states 0, 1, 2 and having transition probability matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

It is easy to verify that this Markov chain is irreducible. For example, it is possible to go from state 0 to state 2 since

$$0 \to 1 \to 2$$

That is, one way of getting from state 0 to state 2 is to go from state 0 to state 1 (with probability $\frac{1}{2}$) and then go from state 1 to state 2 (with probability $\frac{1}{4}$).

Example 3.2. Consider a Markov chain consisting of the four states 0, 1, 2, 3 and having transition probability matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The classes of this Markov chain are $\{0, 1\}$, $\{2\}$, and $\{3\}$. Note that while state 0 (or 1) is accessible from state 2, the reverse is not true. Since state 3 is an absorbing state, that is, $P_{33} = 1$, no other state is accessible from it.

For any state i we let f_i denote the probability that, starting in state i, the process will ever reenter state i. State i is said to be **recurrent** if $f_i = 1$ and **transient** if $f_i < 1$.

Suppose that the process starts in state i and i is recurrent. Hence, with probability 1, the process will eventually reenter state i. However, by the definition of a Markov chain, it follows that the process will be starting over again when it reenters state i and, therefore, state i will eventually be visited again. Continual repetition of this argument leads to the conclusion that if state i is recurrent then, starting in state i, the process will reenter state i again and again and again—in fact, infinitely often.

On the other hand, suppose that state i is transient. Hence, each time the process enters state i there will be a positive probability, namely, $1 - f_i$, that it will never again enter that state. Therefore, starting in state i, the probability that the process will be in state i for exactly n time periods equals $f_i^{n-1}(1 - f_i)$, $n \geq 1$. In other words, if state i is transient then, starting in state i, the number of time periods that the process will be in state i has a geometric distribution with finite mean $1/(1 - f_i)$.

From the preceding two paragraphs, it follows that state i is recurrent if and only if, starting in state i, the expected number of time periods that the process is in state i is infinite. But, letting

$$I_n = \begin{cases} 1, & \text{if } X_n = i \\ 0, & \text{if } X_n \neq i \end{cases}$$

we have that $\sum_{n=0}^{\infty} I_n$ represents the number of periods that the process is in state i. Also,

$$\begin{aligned} E\left[\sum_{n=0}^{\infty} I_n | X_0 = i\right] &= \sum_{n=0}^{\infty} E[I_n | X_0 = i] \\ &= \sum_{n=0}^{\infty} P\{X_n = i | X_0 = i\} = \sum_{n=0}^{\infty} P_{ii}^n \end{aligned}$$

We have thus proven the following.

Proposition 3.1. *State i is*

$$\textit{recurrent if } \sum_{n=1}^{\infty} P_{ii}^n = \infty,$$

$$\textit{transient if } \sum_{n=1}^{\infty} P_{ii}^n < \infty,$$

The argument leading to the preceding proposition is doubly important because it also shows that a transient state will only be visited a finite number of times (hence the name transient). This leads to the conclusion that in a finite-state Markov chain not all states can be transient. To see this, suppose the states are $0, 1, \dots, M$ and suppose that they are all transient. Then after a finite amount of time (say, after time T_0) state 0 will never be visited, and after a time (say, T_1) state 1 will never be visited, and after a time (say, T_2) state 2 will never be visited, and so on. Thus, after a finite time $T = \max\{T_0, T_1, \dots, T_M\}$ no states will be visited. But as the process must be in some state after time T we arrive at a contradiction, which shows that at least one of the states must be recurrent.

Another use of Proposition 24.1 is that it enables us to show that recurrence is a class property.

Corollary 3.1. *If state i is recurrent, and state i communicates with state j, then state j is recurrent.*

Proof. To prove this we first note that, since state i communicates with state j, there exist integers k and m such that $P_{ij}^k > 0$, $P_{ji}^m > 0$. Now, for any integer n

$$P_{jj}^{m+n+k} \geq P_{ji}^m P_{ii}^n P_{ij}^k$$

This follows since the left side of the preceding is the probability of going from j to j in $m+n+k$ steps, while the right side is the probability of going from j to j in $m+n+k$ steps via a path that goes from j to i in m steps, then from i to i in an additional n steps, then from i to j in an additional k steps.

From the preceding we obtain, by summing over n, that

$$\sum_{n=1}^{\infty} P_{jj}^{m+n+k} \geq P_{ji}^{m} P_{ij}^{k} \sum_{n=1}^{\infty} P_{ii}^{n} = \infty$$

since $P_{ji}^{m} P_{ij}^{k} > 0$ and $\sum_{n=1}^{\infty} P_{ii}^{n}$ is infinite since state i is recurrent. Thus, by Proposition 24.1 it follows that state j is also recurrent. □

Remark 3.1.

1. Corollary 23.1 also implies that transience is a class property. For if state i is transient and communicates with state j , then state j must also be transient. For if j were recurrent then, by Corollary 23.1, i would also be recurrent and hence could not be transient.
2. Corollary 23.1 along with our previous result that not all states in a finite Markov chain can be transient leads to the conclusion that all states of a finite irreducible Markov chain are recurrent.

Example 3.3. Let the Markov chain consisting of the states $0, 1, 2, 3$ have the transition probability matrix

$$P = \begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Determine which states are transient and which are recurrent.

Solution 3.1. It is a simple matter to check that all states communicate and, hence, since this is a finite chain, all states must be recurrent.

Example 3.4. Consider the Markov chain having states $0, 1, 2, 3, 4$ and

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{2} \end{bmatrix}$$

Determine the recurrent state.

Solution 3.2. This chain consists of the three classes $\{0, 1\}$, $\{2, 3\}$, and $\{4\}$. The first two classes are recurrent and the third transient.

Example 3.5. (A Random Walk) Consider a Markov chain whose state space consists of the integers $i = 0, \pm 1, \pm 2, ...$, and have transition probabilities given by

$$P_{i,i+1} = p = 1 - P_{i,i-1}, \quad i = 0, \pm 1, \pm 2, ...$$

where $0 < p < 1$. In other words, on each transition the process either moves one step to the right (with probability p) or one step to the left (with probability $1 - p$). One colorful interpretation of this process is that it represents the wanderings of a drunken man as he walks along a straight line. Another is that it represents the winnings of a gambler who on each play of the game either wins or loses one dollar.

Since all states clearly communicate, it follows from Corollary 23.1 that they are either all transient or all recurrent. So let us consider state 0 and attempt to determine if $\sum_{n=1}^{\infty} P_{00}^n$ is finite or infinite.

Since it is impossible to be even (using the gambling model interpretation) after an odd number of plays we must, of course, have that

$$P_{00}^{2n-1} = 0, \quad n = 1, 2, ...$$

On the other hand, we would be even after $2n$ trials if and only if we won n of these and lost n of these. Because each play of the game results in a win with probability p and a loss with probability $1 - p$, the desired probability is thus the binomial probability

$$P_{00}^{2n} = \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{n!n!} (p(1-p))^n, \quad n = 1, 2, 3, ...$$

By using an approximation, due to Stirling, which asserts that

$$n! \sim n^{n+1/2} e^{-n} \sqrt{2\pi} \tag{3.1}$$

where we say that $a_n \sim b_n$ when $\lim_{n \to \infty} a_n / b_n = 1$, we obtain

$$P_{00}^{2n} \sim \frac{(4p(1-p))^n}{\sqrt{\pi n}}$$

Now it is easy to verify, for positive a_n, b_n, that if $a_n \sim b_n$, then $\sum_n a_n < \infty$ if and only if $\sum_n b_n < \infty$. Hence, $\sum_{n=1}^{\infty} P_{00}^n$ will converge if and only if

$$\sum_{n=1}^{\infty} \frac{(4p(1-p))^n}{\sqrt{\pi n}}$$

does. However, $4p(1-p) \leq 1$ with equality holding if and only if $p = \frac{1}{2}$. Hence, $\sum_{n=1}^{\infty} P_{00}^{n} = \infty$ if and only if $p = \frac{1}{2}$. Thus, the chain is recurrent when $p = \frac{1}{2}$ and transient if $p \neq \frac{1}{2}$.

When $p = \frac{1}{2}$, the preceding process is called a **symmetric random walk**. We could also look at symmetric random walks in more than one dimension. For instance, in the two-dimensional symmetric random walk the process would, at each transition, either take one step to the left, right, up, or down, each having probability $\frac{1}{4}$. That is, the state is the pair of integers (i, j) and the transition probabilities are given by

$$P_{(i,j),(i+1,j)} = P_{(i,j),(i-1,j)} = P_{(i,j),(i,j+1)} = P_{(i,j),(i,j-1)} = \frac{1}{4}$$

By using the same method as in the one-dimensional case, we now show that this Markov chain is also recurrent.

Since the preceding chain is irreducible, it follows that all states will be recurrent if state $0 = (0, 0)$ is recurrent. So considerP_{00}^{2n}. Now after $2n$ steps, the chain will be back in its original location if for some i, $0 \leq i \leq n$, the $2n$ steps consist of i steps to the left, i to the right, $n-i$ up, and $n-i$ down. Since each step will be either of these four types with probability $\frac{1}{4}$, it follows that the desired probability is a multinomial probability. That is,

$$\begin{aligned} P_{00}^{2n} &= \sum_{i=0}^{n} \frac{(2n)!}{i!i!(n-i)!(n-i)!} \left(\frac{1}{4}\right)^{2n} = \sum_{i=0}^{n} \frac{(2n)!}{n!n!} \frac{n!}{(n-i)!i!} \frac{n!}{(n-i)!i!} \left(\frac{1}{4}\right)^{2n} \\ &= \left(\frac{1}{4}\right)^{2n} \binom{2n}{n} \sum_{i=0}^{n} \binom{n}{i} \binom{n}{n-i} = \left(\frac{1}{4}\right)^{2n} \binom{2n}{n} \binom{2n}{n} \end{aligned}$$

where the last equality uses the combinatorial identity

$$\binom{2n}{n} = \sum_{i=0}^{n} \binom{n}{i} \binom{n}{n-i}$$

which follows upon noting that both sides represent the number of subgroups of size n one can select from a set of n white and n black objects. Now,

$$\binom{2n}{n} = \frac{(2n)!}{n!n!} \sim \frac{(2n)^{2n+1/2} e^{-2n} \sqrt{2\pi}}{n^{2n+1} e^{-2n} (2\pi)} = \frac{4^n}{\sqrt{\pi n}}$$

where we used Stirling's approximation in the second step. Hence from Equation 3.2 we see that

$$P_{00}^{2n} \sim \frac{1}{\pi n} \tag{3.2}$$

which shows that $\sum_n P_{00}^{2n} = \infty$, and thus all states are recurrent.

Interestingly enough, whereas the symmetric random walks in one and two dimensions are both recurrent, all higher-dimensional symmetric random walks turn out to be transient. (For instance, the three-dimensional symmetric random walk is at each transition equally likely to move in any of six ways—either to the left, right, up, down, in, or out.)

The following Python code verifies Equation 3.2.

```python
import numpy as np

def p_00_2n(n, sample_size):
    """
    The function computes P^{2n}_{00} for a 2D symmetric random walk,
    the probability that state 0 is visited again in 2n steps.
    The probability is known to be about 1/(pi n) when n is large.

    input:
    n: int, 2n is the number of steps
    sample_size: int, sample size, the number of experiments performed

    output:
    prob: P^{2n}_{00}, the probability that state 0 is visited again in 2n steps.
    """
    count = 0
    for i in range(sample_size):
        current_state = np.array([0,0])
        for j in range(2*n):
            rn = np.random.random()  # simulate a random number
            # Decide where to move
            if rn <= 0.25:
                current_state[0] -= 1
            elif rn <= 0.5:
                current_state[0] += 1
            elif rn <= 0.75:
                current_state[1] -= 1
            else:
                current_state[1] += 1
        if np.array_equal(current_state, np.array([0,0])):
            count += 1
    prob = count / sample_size
    return prob
```

```python
n = 1000
sample_size = 100000
print('Computed probability is: ', p_00_2n(n, sample_size),
      'and analytical probability is: ', 1/(np.pi*n))
```

```
Computed probability is:  0.00024 and analytical probability is:  0.0003183098861837907
```

Example 3.6. (On the Ultimate Instability of the Aloha Protocol) Consider a communications facility in which the numbers of messages arriving during each of the time periods $n = 1, 2, ...$ are independent and identically distributed random variables. Let $a_i = P\{i \text{ arrivals}\}$, and suppose that $a_0 + a_1 < 1$. Each arriving message will transmit at the end of the period in which it arrives. If exactly one message is transmitted, then the transmission is successful and the message leaves the system. However, if at any time two or more messages simultaneously transmit, then a collision is deemed to occur and these messages remain in the system. Once a message is involved in a collision it will, independently of all else, transmit at the end of each additional period with probability p—the so-called Aloha protocol (because it was first instituted at the University of Hawaii). We will show that such a system is asymptotically unstable in the sense that the number of successful transmissions will, with probability 1, be finite.

To begin let X_n denote the number of messages in the facility at the beginning of the nth period, and note that $\{X_n, n \geq 0\}$ is a Markov chain. Now for $k \geq 0$ define the indicator variables I_k by

$$I_k = \begin{cases} 1, & \text{if the first time that the chain departs state k it directly goes to state k } -1 \\ 0, & \text{otherwise} \end{cases}$$

and let it be 0 if the system is never in state k, $k \geq 0$. (For instance, if the successive states are $0, 1, 3, 4, ...$, then $I_3 = 0$ since when the chain first departs state 3 it goes to state 4; whereas, if they are $0, 3, 3, 2, ...$, then $I_3 = 1$ since this time it goes to state 2.) Now,

$$E\left[\sum_{k=0}^{\infty} I_k\right] = \sum_{k=0}^{\infty} E[I_k] = \sum_{k=0}^{\infty} P\{I_k = 1\} \leq \sum_{k=0}^{\infty} P\{I_k = 1 | k \text{ is ever visited}\} \tag{3.3}$$

Now, $P\{I_k = 1 | k \text{ is ever visited}\}$ is the probability that when state k is departed the next state is $k - 1$. That is, it is the conditional probability that a transition from k is to $k - 1$ given that it is not back into k, and so

$$P\{I_k = 1 | k \text{ is ever visited}\} = \frac{P_{k,k-1}}{1 - P_{kk}}$$

Because

$$\begin{aligned} P_{k,k-1} &= a_0 kp(1-p)^{k-1}, \\ P_{k,k} &= a_0[1 - kp(1-p)^{k-1}] + a_1(1-p)^k \end{aligned}$$

which is seen by noting that if there are k messages present on the beginning of a day, then (a) there will be $k - 1$ at the beginning of the next day if there are no new messages that day and exactly one of the k messages transmits; and (b) there will be k at the beginning of the next day if either

1. there are no new messages and it is not the case that exactly one of the existing k messages transmits (otherwise the transmission would be successful), or
2. there is exactly one new message (which automatically transmits) and none of the other k messages transmits.

Substitution of the preceding into Equation 3.3 yields

$$E\left[\sum_{k=0}^{\infty} I_k\right] \leq \sum_{k=0}^{\infty} \frac{a_0 kp(1-p)^{k-1}}{1-a_0[1-kp(1-p)^{k-1}]-a_1(1-p)^k} < \infty$$

where the convergence follows by noting that when k is large the denominator of the expression in the preceding sum converges to $1-a_0$ and so the convergence or divergence of the sum is determined by whether or not the sum of the terms in the numerator converge and $\sum_{k=0}^{\infty} k(1-p)^{k-1} < \infty$.

Hence, $E\left[\sum_{k=0}^{\infty} I_k\right] < \infty$, which implies that $\sum_{k=0}^{\infty} I_k < \infty$ with probability 1 (for if there was a positive probability that $\sum_{k=0}^{\infty} I_k$ could be ∞, then its mean would be ∞). Hence, with probability 1, there will be only a finite number of states that are initially departed via a successful transmission; or equivalently, there will be some finite integer N such that whenever there are N or more messages in the system, there will never again be a successful transmission. From this (and the fact that such higher states will eventually be reached—why?) it follows that, with probability 1, there will only be a finite number of successful transmissions.

4 Limiting Probabilities for Markov Chains

4.1 Limiting Probabilities

In Example 4 in Lecture 2, where we considered a two-state Markov chain with one-step transition probability given by

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix},$$

we calculated $P^{(4)}$, which is

$$P^{(4)} = \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix}.$$

From this, it follows that $P^{(8)} = P^{(4)} \cdot P^{(4)}$ is given (to three significant places) by

$$P^{(8)} = \begin{bmatrix} 0.572 & 0.428 \\ 0.570 & 0.430 \end{bmatrix}$$

Note that the matrix $P^{(8)}$ is almost identical to the matrix $P^{(4)}$, and secondly, that each of the rows of $P^{(8)}$ has almost identical entries. In fact it seems that P_{ij}^n is converging to some value (as $n \to \infty$) which is the same for all i. In other words, there seems to exist a limiting probability that the process will be in state j after a large number of transitions, and this value is independent of the initial state.

To make the preceding heuristics more precise, two additional properties of the states of a Markov chain need to be considered. State i is said to have **period** d if $P_{ii}^n = 0$ whenever n is not divisible by d, and d is the largest integer with this property. For instance, starting in i, it may be possible for the process to enter state i only at the times $2, 4, 6, 8, \cdots$, in which case state i has period 2. A state with period 1 is said to be **aperiodic**. It can be shown that periodicity is a class property. That is, if state i has period d, and states i and j communicate, then state j also has period d.

If state i is recurrent, then it is said to be **positive recurrent** if, starting in i, the expected time until the process returns to state i is finite. It can be shown that positive recurrence is a class property. While there exist recurrent states that are not positive recurrent (such states are called **null recurrent**), it can be shown that in a finite-state Markov chain all recurrent states are positive recurrent. Positive recurrent, aperiodic states are called **ergodic**.

We are now ready for the following important theorem which we state without proof.

Theorem 4.1. *For an irreducible ergodic Markov chain* $\lim_{n\to\infty} P_{ij}^n$ *exists and is independent of* i*. Furthermore, letting*

$$\pi_j = \lim_{n\to\infty} P_{ij}^n, \quad j \geq 0$$

then π_j *is the unique nonnegative solution of*

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, \quad j \geq 0,$$

$$\sum_{j=0}^{\infty} \pi_j = 1 \tag{4.1}$$

Remark 4.1.

(i) Given that $\pi_j = \lim_{n\to\infty} P_{ij}^n$ exists and is independent of the initial state i, it is not difficult to (heuristically) see that the π's must satisfy Equation 4.1. Let us derive an expression for $P\{X_{n+1} = j\}$ by conditioning on the state at time n. That is,

$$P\{X_{n+1} = j\} = \sum_{i=0}^{\infty} P\{X_{n+1} = j | X_n = i\} P\{X_n = i\} = \sum_{i=0}^{\infty} P_{ij} P\{X_n = i\}$$

Letting $n \to \infty$, and assuming that we can bring the limit inside the summation, leads to

$$\pi_j = \sum_{i=0}^{\infty} P_{ij} \pi_i$$

(ii) It can be shown that π_j , the limiting probability that the process will be in state j at time n, also equals the long-run proportion of time that the process will be in state j .

(iii) If the Markov chain is irreducible, then there will be a solution to

$$\pi_j = \sum_i \pi_i P_{ij}, \quad j \geq 0,$$

$$\sum_j \pi_j = 1$$

if and only if the Markov chain is positive recurrent. If a solution exists then it will be unique, and π_j will equal the long run proportion of time that the Markov chain is in state j . If the chain is aperiodic, then π_j is also the limiting probability that the chain is in state j.

Example 4.1. Consider Example 1 in Lecture 2, in which we assume that if it rains today, then it will rain tomorrow with probability α; and if it does not rain today, then it will rain tomorrow with probability β. If we say that the state is 0 when it rains and 1 when it does not rain, then by Equation 4.1 the limiting probabilities π_0 and π_1 are given by

$$\begin{aligned} \pi_0 &= \alpha\pi_0 + \beta\pi_1 \\ \pi_1 &= (1-\alpha)\pi_0 + (1-\beta)\pi_1 \\ \pi_0 + \pi_1 &= 1 \end{aligned}$$

which yields that

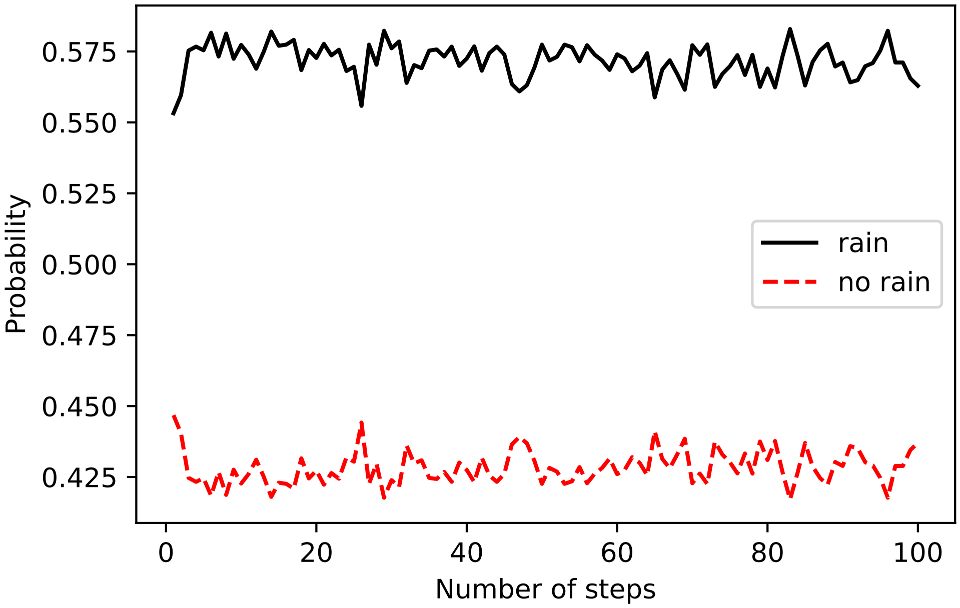
$$\pi_0 = \frac{\beta}{1+\beta-\alpha}, \quad \pi_1 = \frac{1-\alpha}{1+\beta-\alpha}$$

For example if $\alpha = 0.7$ and $\beta = 0.4$, then the limiting probability of rain is $\pi_0 = \frac{4}{7} = 0.571$.

```python
import numpy as np
import matplotlib.pyplot as plt

def simulate_rain(P, n):
    """
    Simulate the rain problem

    input:
    P: numpy 2d array of shape (2,2), the transition probability matrix
    n: int, number of steps to simulate

    output:
    state: int, 0 or 1, representing the final state.
    """
    state = np.random.randint(2)  # randomly initialize a state
    for i in range(n):
        if state == 0:
            if np.random.random() < P[0,0]:
                state = 0
            else:
                state = 1
        else:
            if np.random.random() < P[1,0]:
                state = 0
            else:
                state = 1
    return state
```

```
alpha = 0.7
beta = 0.4
sample_size = 10000  # For each fixed number of steps, use 10000 simulations to determine th
n_max = 100  # Consider 1, 2, 3, up until 100 steps
P = np.array([[alpha, 1-alpha], [beta, 1-beta]])
p = np.zeros((n_max,2))  # Store all the probabilities corresponding to all numbers of steps
for n in range(1, n_max+1):
    for i in range(sample_size):
        if simulate_rain(P, n) == 0:
            p[n-1, 0] += 1
        else:
            p[n-1, 1] += 1
    p[n-1,:] /= p[n-1,:].sum()
```

```
fig = plt.figure()
plt.plot(np.arange(1, n_max+1), p[:, 0], 'k-', label='rain')
plt.plot(np.arange(1, n_max+1), p[:, 1], 'r--', label='no rain')
plt.xlabel('Number of steps')
plt.ylabel('Probability')
plt.legend()
plt.show(fig)
```



Example 4.2. Consider a three-state Markov chain having a transition probability matrix

$$P = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

In the long run, what proportion of time is the process in each of the three states?

Solution 4.1. The limiting probabilities $\pi_i, i = 0, 1, 2$, are obtained by solving the set of equations in Equation 4.1. In this case these equations are

$$\begin{aligned} \pi_0 &= 0.5\pi_0 + 0.3\pi_1 + 0.2\pi_2 \\ \pi_1 &= 0.4\pi_0 + 0.4\pi_1 + 0.4\pi_2 \\ \pi_2 &= 0.1\pi_0 + 0.3\pi_1 + 0.5\pi_2 \\ \pi_0 + \pi_1 + \pi_2 &= 1 \end{aligned}$$

Solving yields

$$\pi_0 = \frac{21}{62}, \quad \pi_1 = \frac{23}{62}, \quad \pi_2 = \frac{18}{62}$$

Example 4.3. (The Hardy–Weinberg Law and a Markov Chain in Genetics) Consider a large population of individuals, each of whom possesses a particular pair of genes, of which each individual gene is classified as being of type A or type a. Assume that the proportions of individuals whose gene pairs are AA, aa, or Aa are, respectively, p_0, q_0, and r_0 ($p_0 + q_0 + r_0 = 1$). When two individuals mate, each contributes one of his or her genes, chosen at random, to the resultant offspring. Assuming that the mating occurs at random, in that each individual is equally likely to mate with any other individual, we are interested in determining the proportions of individuals in the next generation whose genes are AA, aa, or Aa. Calling these proportions p, q, and r, they are easily obtained by focusing attention on an individual of the next generation and then determining the probabilities for the gene pair of that individual.

To begin, note that randomly choosing a parent and then randomly choosing one of its genes is equivalent to just randomly choosing a gene from the total gene population. By conditioning on the gene pair of the parent, we see that a randomly chosen gene will be type A with probability

$$P\{A\} = P\{A|AA\}p_0 + P\{A|aa\}q_0 + P\{A|Aa\}r_0 = p_0 + r_0/2$$

Similarly, it will be type a with probability

$$P\{a\} = q_0 + r_0/2$$

Thus, under random mating a randomly chosen member of the next generation will be type AA with probability p, where

$$p = P\{A\}P\{A\} = (p_0 + r_0/2)^2$$

Similarly, the randomly chosen member will be type aa with probability

$$q = P\{a\}P\{a\} = (q_0 + r_0/2)^2$$

and will be type Aa with probability

$$r = 2P\{A\}P\{a\} = 2(p_0 + r_0/2)(q_0 + r_0/2)$$

Since each member of the next generation will independently be of each of the three gene types with probabilities p, q, r, it follows that the percentages of the members of the next generation that are of type AA, aa, or Aa are respectively p, q, and r.

If we now consider the total gene pool of this next generation, then $p + r/2$, the fraction of its genes that are A, will be unchanged from the previous generation. This follows either by arguing that the total gene pool has not changed from generation to generation or by the following simple algebra:

$$\begin{aligned} p + r/2 &= (p_0 + r_0/2)^2 + (p_0 + r_0/2)(q_0 + r_0/2) \\ &= (p_0 + r_0/2)[p_0 + r_0/2 + q_0 + r_0/2] \\ &= p_0 + r_0/2 \quad \text{since } p_0 + r_0 + q_0 = 1 \\ &= P\{A\} \end{aligned} \tag{4.2}$$

Thus, the fractions of the gene pool that are A and a are the same as in the initial generation. From this it follows that, under random mating, in all successive generations after the initial one the percentages of the population having gene pairs AA, aa, and Aa will remain fixed at the values p, q, and r. This is known as the *Hardy–Weinberg law.*

Suppose now that the gene pair population has stabilized in the percentages p, q, r, and let us follow the genetic history of a single individual and her descendants. (For simplicity, assume that each individual has exactly one offspring.) So, for a given individual, let X_n denote the genetic state of her descendant in the nth generation. The transition probability matrix of this Markov chain, namely,

$$\begin{array}{c} \\ AA \\ aa \\ Aa \end{array} \begin{array}{c} \begin{array}{ccc} AA & aa & Aa \end{array} \\ \left[\begin{array}{ccc} p + \frac{r}{2} & 0 & q + \frac{r}{2} \\ 0 & q + \frac{r}{2} & p + \frac{r}{2} \\ \frac{p}{2} + \frac{r}{4} & \frac{q}{2} + \frac{r}{4} & \frac{p}{2} + \frac{q}{2} + \frac{r}{2} \end{array}\right] \end{array}$$

is easily verified by conditioning on the state of the randomly chosen mate. It is quite intuitive (why?) that the limiting probabilities for this Markov chain (which also equal the fractions of the individual's descendants that are in each of the three genetic states) should just be p, q, and r. To verify this we must show that they satisfy Equation 4.1. Because one of the equations in Equation 4.1 is redundant, it suffices to show that

$$\begin{aligned} p &= p\left(p + \frac{r}{2}\right) + r\left(\frac{p}{2} + \frac{r}{4}\right) = \left(p + \frac{r}{2}\right)^2 \\ q &= q\left(q + \frac{r}{2}\right) + r\left(\frac{q}{2} + \frac{r}{4}\right) = \left(q + \frac{r}{2}\right)^2 \\ p + q + r &= 1 \end{aligned}$$

But this follows from Equation 4.2, and thus the result is established.

The following result is quite useful.

Proposition 4.1. *Let $\{X_n, n \geq 1\}$ be an irreducible Markov chain with stationary probabilities $\pi_j, j \geq 0$, and let r be a bounded function on the state space. Then, with probability 1,*

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} r(X_n)}{N} = \sum_{j=0}^{\infty} r(j)\pi_j$$

Proof. If we let $a_j(N)$ be the amount of time the Markov chain spends in state j during time periods $1, \dots, N$, then

$$\sum_{n=1}^{N} r(X_n) = \sum_{j=0}^{\infty} a_j(N) r(j)$$

Since $a_j(N)/N \to \pi_j$ the result follows from the preceding upon dividing by N and then letting $N \to \infty$. □

If we suppose that we earn a reward $r(j)$ whenever the chain is in state j, then Proposition 24.1 states that our average reward per unit time is $\sum_j r(j)\pi_j$.

Example 4.4. For the four state Bonus Malus automobile insurance system specified in Lecture 2, find the average annual premium paid by a policyholder whose yearly number of claims is a Poisson random variable with mean 1/2.

Solution 4.2. With $a_k = e^{-1/2}\frac{(1/2)^k}{k!}$, we have

$$a_0 = 0.6065, \quad a_1 = 0.3033, \quad a_2 = 0.0758$$

Therefore, the Markov chain of successive states has the following transition probability matrix.

$$\begin{bmatrix} 0.6065 & 0.3033 & 0.0758 & 0.0144 \\ 0.6065 & 0.0000 & 0.3033 & 0.0902 \\ 0.0000 & 0.6065 & 0.0000 & 0.3935 \\ 0.0000 & 0.0000 & 0.6065 & 0.3935 \end{bmatrix}$$

The stationary probabilities are given as the solution of

$$\begin{aligned} \pi_1 &= 0.6065\pi_1 + 0.6065\pi_2 \\ \pi_2 &= 0.3033\pi_1 + 0.6065\pi_3 \\ \pi_3 &= 0.0758\pi_1 + 0.3033\pi_2 + 0.6065\pi_4 \\ \pi_1 + \pi_2 + \pi_3 + \pi_4 &= 1 \end{aligned}$$

Rewriting the first three of these equations gives:

$$\begin{aligned}
\pi_2 &= \frac{1-0.6065}{0.6065}\pi_1, \\
\pi_3 &= \frac{\pi_2 - 0.3033\pi_1}{0.6065} \\
\pi_4 &= \frac{\pi_3 - 0.0758\pi_1 - 0.3033\pi_2}{0.6065}
\end{aligned}$$

or

$$\begin{aligned}
\pi_2 &= 0.6488\pi_1, \\
\pi_3 &= 0.5697\pi_1 \\
\pi_4 &= 0.4900\pi_1
\end{aligned}$$

Using that $\sum_{i=1}^{4} \pi = 1$ gives the solution (rounded to four decimal places)

$$\pi_1 = 0.3692, \quad \pi_2 = 0.2395, \quad \pi_3 = 0.2103, \quad \pi_4 = 0.1809$$

Therefore, the average annual premium paid is

$$200\pi_1 + 250\pi_2 + 400\pi_3 + 600\pi_4 = 326.375$$

5 Some Markov Chains Applications

5.1 The Gambler's Ruin Problem

Consider a gambler who at each play of the game has probability p of winning one unit and probability $q = 1 - p$ of losing one unit. Assuming that successive plays of the game are independent, what is the probability that, starting with i units, the gambler's fortune will reach N before reaching 0?

If we let X_n denote the player's fortune at time n, then the process $\{X_n, n = 0, 1, 2, \dots\}$ is a Markov chain with transition probabilities

$$\begin{aligned} P_{00} &= P_{NN} = 1, \\ P_{i,i+1} &= p = 1 - P_{i,i-1}, \quad i = 1, 2, \dots, N-1 \end{aligned}$$

This Markov chain has three classes, namely, $\{0\}$, $\{1, 2, \dots, N-1\}$, and $\{N\}$; the first and third class being recurrent and the second transient. Since each transient state is visited only finitely often, it follows that, after some finite amount of time, the gambler will either attain his goal of N or go broke.

Let $P_i, i = 0, 1, \dots, N$, denote the probability that, starting with i, the gambler's fortune will eventually reach N. By conditioning on the outcome of the initial play of the game we obtain

$$P_i = pP_{i+1} + qP_{i-1}, \quad i = 1, 2, \dots, N-1$$

or equivalently, since $p + q = 1$,

$$pP_i + qP_i = pP_{i+1} + qP_{i-1}$$

or

$$P_{i+1} - P_i = \frac{q}{p}\left(P_i - P_{i-1}\right), \quad i = 1, 2, \dots, N-1$$

Hence, since $P_0 = 0$, we obtain from the preceding line that

$$
\begin{aligned}
P_2 - P_1 &= \frac{q}{p}(P_1 - P_0) = \frac{q}{p} P_1, \\
P_3 - P_2 &= \frac{q}{p}(P_2 - P_1) = \left(\frac{q}{p}\right)^2 P_1, \\
&\vdots \\
P_i - P_{i-1} &= \frac{q}{p}(P_{i-1} - P_{i-2}) = \left(\frac{q}{p}\right)^{i-1} P_1, \\
&\vdots \\
P_N - P_{N-1} &= \left(\frac{q}{p}\right)(P_{N-1} - P_{N-2}) = \left(\frac{q}{p}\right)^{N-1} P_1
\end{aligned}
$$

Adding the first $i-1$ of these equations yields

$$P_i - P_1 = P_1 \left[\left(\frac{q}{p}\right) + \left(\frac{q}{p}\right)^2 + \cdots + \left(\frac{q}{p}\right)^{i-1} \right]$$

or

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)} P_1, & \text{if } \frac{q}{p} \neq 1 \\ i P_1, & \text{if } \frac{q}{p} = 1 \end{cases}$$

Now, using the fact that $P_N = 1$, we obtain that

$$P_1 = \begin{cases} \frac{1-(q/p)}{1-(q/p)^N}, & \text{if } p \neq \frac{1}{2} \\ \frac{1}{N}, & \text{if } p = \frac{1}{2} \end{cases}$$

and hence

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)^N}, & \text{if } p \neq \frac{1}{2} \\ \frac{i}{N}, & \text{if } p = \frac{1}{2} \end{cases} \tag{5.1}$$

Note that, as $N \to \infty$,

$$P_i \to \begin{cases} 1 - \left(\frac{q}{p}\right)^i, & \text{if } p > \frac{1}{2} \\ 0, & \text{if } p \leq \frac{1}{2} \end{cases}$$

Thus, if $p > \frac{1}{2}$, there is a positive probability that the gambler's fortune will increase indefinitely; while if $p \leq \frac{1}{2}$, the gambler will, with probability 1, go broke against an infinitely rich adversary.

```python
import numpy as np

def compute_Pi(p, N, i, sample_size):
    """
    Compute Pi, i=0,1,...,N, the probability that, starting with i,
    the gambler’s fortune will eventually reach N

    input:
    p: float, the probability p_{i,i+1}
    N: int, the last state
    i: int between 0 and N (inclusive), the starting state
    sample_size: int, the sample size, i.e,
    how many experiments are performed

    output:
    Pi: float, the probability Pi
    """
    count = 0
    for j in range(sample_size):
        state = i
        while state != 0 and state != N:
            if np.random.random() < p:
                state += 1
            else:
                 state -= 1
        if state == N:
            count += 1
    Pi = count / sample_size
    return Pi
```

We consider a case where $p = 0.5$, so $P_i = \frac{i}{N}$.

```
p = 0.5
N = 10
i = 3
sample_size = 10000
print('The simulated probability is: ', compute_Pi(p, N, i, sample_size),
      ' and the theoretical probability is: ', i/N)
```

```
The simulated probability is:  0.3133  and the theoretical probability is:  0.3
```

Now consider a case where $p = 0.3 \neq 0.5$. In this case, $P_i = \frac{1-(q/p)^i}{1-(q/p)^N}$.

```
p = 0.3
N = 10
i = 8
sample_size = 10000
print('The simulated probability is: ', compute_Pi(p, N, i, sample_size),
      ' and the theoretical probability is: ', (1-((1-p)/p)**i)/(1-((1-p)/p)**N))
```

```
The simulated probability is:  0.187  and the theoretical probability is:  0.183502787729599
```

Example 5.1. Suppose Max and Patty decide to flip pennies; the one coming closest to the wall wins. Patty, being the better player, has a probability 0.6 of winning on each flip. (a) If Patty starts with five pennies and Max with ten, what is the probability that Patty will wipe Max out? (b) What if Patty starts with 10 and Max with 20?

Solution 5.1.

(a) The desired probability is obtained from Equation 5.1 by letting $i = 5$, $N = 15$, and $p = 0.6$. Hence, the desired probability is

$$\frac{1-\left(\frac{2}{3}\right)^5}{1-\left(\frac{2}{3}\right)^{15}} \approx 0.87$$

(b) The desired probability is

$$\frac{1-\left(\frac{2}{3}\right)^{10}}{1-\left(\frac{2}{3}\right)^{30}} \approx 0.98$$

5.2 A Model for Algorithmic Efficiency

The following optimization problem is called a linear program:

$$\begin{aligned} &\text{minimize } cx, \\ &\text{subject to} Ax = b, x \geq 0 \end{aligned}$$

where A is an $m \times n$ matrix of fixed constants; $c = (c_1, \dots, c_n)$ and $b = (b_1, \dots, b_m)$ are vectors of fixed constants; and $x = (x_1, \dots, x_n)$ is the n-vector of nonnegative values that is to be chosen to minimize $cx = \sum_{i=1}^{n} c_i x_i$. Supposing that $n > m$, it can be shown that the optimal x can always be chosen to have at least $n - m$ components equal to 0—that is, it can always be taken to be one of the so-called extreme points of the feasibility region.

The simplex algorithm solves this linear program by moving from an extreme point of the feasibility region to a better (in terms of the objective function cx) extreme point (via the pivot operation) until the optimal is reached. Because there can be as many as $N = \binom{n}{m}$ such extreme points, it would seem that this method might take many iterations, but, surprisingly to some, this does not appear to be the case in practice.

To obtain a feel for whether or not the preceding statement is surprising, let us consider a simple probabilistic (Markov chain) model as to how the algorithm moves along the extreme points. Specifically, we will suppose that if at any time the algorithm is at the jth best extreme point then after the next pivot the resulting extreme point is equally likely to be any of the $j-1$ best. Under this assumption, we show that the time to get from the Nth best to the best extreme point has approximately, for large N, a normal distribution with mean and variance equal to the logarithm (base e) of N.

Consider a Markov chain for which $P_{11} = 1$ and

$$P_{ij} = \frac{1}{i-1}, \quad j = 1, \dots, i-1, \, i > 1$$

and let T_i denote the number of transitions needed to go from state i to state 1. A recursive formula for $E[T_i]$ can be obtained by conditioning on the initial transition:

$$E[T_i] = 1 + \frac{1}{i-1} \sum_{j=1}^{i-1} E[T_j]$$

Starting with E[T_1] = 0, we successively see that

$$\begin{aligned} E[T_2] &= 1, \\ E[T_3] &= 1 + \frac{1}{2}, \\ E[T_4] &= 1 + \frac{1}{3}(1 + 1 + \frac{1}{2}) = 1 + \frac{1}{2} + \frac{1}{3} \end{aligned}$$

and it is not difficult to guess and then prove inductively that

$$E[T_i] = \sum_{j=1}^{i-1} 1/j$$

However, to obtain a more complete description of T_N, we will use the representation

$$T_N = \sum_{j=1}^{N-1} I_j$$

where

$$I_j = \begin{cases} 1, & \text{if the process ever enters } j \\ 0, & \text{otherwise} \end{cases}$$

The importance of the preceding representation stems from the following:

Proposition 5.1. *$I_1, \dots, I_{N-1}$ are independent and*

$$P\{I_j = 1\} = 1/j, \quad 1 \le j \le N-1$$

Proof. Given $I_{j+1}, \dots, I_N$, let $n = \min\{i : i > j, I_i = 1\}$ denote the lowest numbered state, greater than j , that is entered. Thus we know that the process enters state n and the next state entered is one of the states $1, 2, \dots, j$. Hence, as the next state from state n is equally likely to be any of the lower number states $1, 2, \dots, n-1$ we see that

$$P\{I_j = 1 | I_{j+1}, \dots, I_N\} = \frac{1/(n-1)}{j/(n-1)} = 1/j$$

Hence, $P\{I_j = 1\} = 1/j$, and independence follows since the preceding conditional probability does not depend on $I_{j+1}, \dots, I_N$. □

Corollary 5.1.

(i) $E[T_N] = \sum_{j=1}^{N-1} 1/j$.

(ii) $Var(T_N) = \sum_{j=1}^{N-1} (1/j)(1-1/j)$.

(iii) For N large, T_N has approximately a normal distribution with mean $\log N$ and variance $\log N$.

Proof. Parts (i) and (ii) follow from Proposition 24.1 and the representation $T_N = \sum_{j=1}^{N-1} I_j$. Part (iii) follows from the central limit theorem since

$$\int_1^N \frac{dx}{x} < \sum_{j=1}^{N-1} 1/j < 1 + \int_1^{N-1} \frac{dx}{x}$$

or

$$\log N < \sum_{j=1}^{N-1} 1/j < 1 + \log(N-1)$$

and so

$$\log N \approx \sum_{j=1}^{N-1} 1/j$$

□

Returning to the simplex algorithm, if we assume that n, m, and $n-m$ are all large, we have by Stirling's approximation that

$$N = \binom{n}{m} \sim \frac{n^{n+1/2}}{(n-m)^{n-m+1/2} m^{m+1/2} \sqrt{2}\pi}$$

and so, letting $c = n/m$,

$$\begin{aligned}\log N \sim \left(mc + \frac{1}{2}\right) \log{(mc)} - \left(m(c-1) + \frac{1}{2}\right) \log{(m(c-1))} \\ - \left(m + \frac{1}{2}\right) \log m - \frac{1}{2} \log{(2\pi)}\end{aligned}$$

or

$$\log N \sim m\left[c \log \frac{c}{c-1} + \log{(c-1)}\right]$$

Now, as $\lim_{x \to \infty} x \log{[x/(x-1)]} = 1$, it follows that, when c is large,

$$\log N \sim m[1 + \log{(c-1)}]$$

Thus, for instance, if $n = 8000$, $m = 1000$, then the number of necessary transitions is approximately normally distributed with mean and variance equal to $1000(1 + \log 7) \approx 3000$. Hence, the number of necessary transitions would be roughly between

$$3000 \pm 2\sqrt{3000} \text{ or roughly } 3000 \pm 110$$

95 percent of the time.

6 Time Reversible Markov Chains

Consider a stationary ergodic Markov chain (that is, an ergodic Markov chain that has been in operation for a long time) having transition probabilities P_{ij} and stationary probabilities π_i, and suppose that starting at some time we trace the sequence of states going backward in time. That is, starting at time n, consider the sequence of states $X_n, X_{n-1}, X_{n-2}, ...$ It turns out that this sequence of states is itself a Markov chain with transition probabilities Q_{ij} defined by

$$\begin{aligned}
Q_{ij} &= P\{X_m = j | X_{m+1} = i\} \\
&= \frac{P\{X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\}} \\
&= \frac{P\{X_m = j\} P\{X_{m+1} = i | X_m = j\}}{P\{X_{m+1} = i\}} \\
&= \frac{\pi_j P_{ji}}{\pi_i}
\end{aligned}$$

To prove that the reversed process is indeed a Markov chain, we must verify that

$$P\{X_m = j | X_{m+1} = i, X_{m+2}, X_{m+3}, ...\} = P\{X_m = j | X_{m+1} = i\}$$

To see that this is so, suppose that the present time is $m+1$. Now, since $X_0, X_1, X_2, ...$ is a Markov chain, it follows that the conditional distribution of the future $X_{m+2}, X_{m+3}, ...$ given the present state X_{m+1} is independent of the past state X_m. However, independence is a symmetric relationship (that is, if A is independent of B, then B is independent of A), and so this means that given X_{m+1}, X_m is independent of $X_{m+2}, X_{m+3}, ...$ But this is exactly what we had to verify.

Thus, the reversed process is also a Markov chain with transition probabilities given by

$$Q_{ij} = \frac{\pi_j P_{ji}}{\pi_i}$$

If $Q_{ij} = P_{ij}$ for all i, j, then the Markov chain is said to be **time reversible**. The condition for time reversibility, namely, $Q_{ij} = P_{ij}$, can also be expressed as

$$\pi_i \, P_{ij} = \pi_j \, P_{ji} \quad \text{for all } i, \ j \tag{6.1}$$

The condition in Equation 6.1 can be stated that, for all states i and j, the rate at which the process goes from i to j (namely, $\pi_i P_{ij}$) is equal to the rate at which it goes from j to i (namely,

$\pi_j P_{ji}$). It is worth noting that this is an obvious necessary condition for time reversibility since a transition from i to j going backward in time is equivalent to a transition from j to i going forward in time; that is, if $X_m = i$ and $X_{m-1} = j$, then a transition from i to j is observed if we are looking backward, and one from j to i if we are looking forward in time. Thus, the rate at which the forward process makes a transition from j to i is always equal to the rate at which the reverse process makes a transition from i to j; if time reversible, this must equal the rate at which the forward process makes a transition from i to j.

If we can find nonnegative numbers, summing to one, that satisfy Equation 6.1, then it follows that the Markov chain is time reversible and the numbers represent the limiting probabilities. This is so since if

$$x_i P_{ij} = x_j P_{ji} \quad \text{for all } i, j, \quad \sum_i x_i = 1 \tag{6.2}$$

then summing over i yields

$$\sum_i x_i \, P_{ij} = x_j \sum_i P_{ji} = x_j, \quad \sum_i x_i = 1$$

and, because the limiting probabilities π_i are the unique solution of the preceding, it follows that $x_i = \pi_i$ for all i.

Example 6.1. Consider a random walk with states $0, 1, \dots, M$ and transition probabilities

$$\begin{aligned} &P_{i,i+1} = \alpha_i = 1 - P_{i,i-1}, \quad i = 1, \dots, M-1, \\ &P_{0,1} = \alpha_0 = 1 - P_{0,0}, \\ &P_{M,M} = \alpha_M = 1 - P_{M,M-1} \end{aligned}$$

Without the need for any computations, it is possible to argue that this Markov chain, which can only make transitions from a state to one of its two nearest neighbors, is time reversible. This follows by noting that the number of transitions from i to $i+1$ must at all times be within 1 of the number from $i+1$ to i. This is so because between any two transitions from i to $i+1$ there must be one from $i+1$ to i (and conversely) since the only way to reenter i from a higher state is via state $i+1$. Hence, it follows that the rate of transitions from i to $i+1$ equals the rate from $i+1$ to i, and so the process is time reversible.

We can easily obtain the limiting probabilities by equating for each state $i = 0, 1, \dots, M-1$ the rate at which the process goes from i to $i+1$ with the rate at which it goes from $i+1$ to i. This yields

$$\begin{aligned} \pi_0 \alpha_0 &= \pi_1 (1 - \alpha_1), \\ \pi_1 \alpha_1 &= \pi_2 (1 - \alpha_2), \\ &\vdots \\ \pi_i \alpha_i &= \pi_{i+1} (1 - \alpha_{i+1}), \quad i = 0, 1, \dots, M-1 \end{aligned}$$

Solving in terms of π_0 yields

$$\begin{aligned}\pi_1 &= \frac{\alpha_0}{1-\alpha_1}\pi_0, \\ \pi_2 &= \frac{\alpha_1}{1-\alpha_2}\pi_1 = \frac{\alpha_1\alpha_0}{(1-\alpha_2)(1-\alpha_1)}\pi_0\end{aligned}$$

and, in general,

$$\pi_i = \frac{\alpha_{i-1}\cdots\alpha_0}{(1-\alpha_i)\cdots(1-\alpha_1)}\pi_0, \quad i = 1, 2, \ldots, M$$

Since $\sum_0^M \pi_i = 1$, we obtain

$$\pi_0\left[1 + \sum_{j=1}^{M} \frac{\alpha_{j-1}\cdots\alpha_0}{(1-\alpha_j)\cdots(1-\alpha_1)}\right] = 1$$

or

$$\pi_0 = \left[1 + \sum_{j=1}^{M} \frac{\alpha_{j-1}\cdots\alpha_0}{(1-\alpha_j)\cdots(1-\alpha_1)}\right]^{-1} \tag{6.3}$$

and

$$\pi_i = \frac{\alpha_{i-1}\cdots\alpha_0}{(1-\alpha_i)\cdots(1-\alpha_1)}\pi_0, \quad i = 1, \ldots, M \tag{6.4}$$

For instance, if $\alpha_i \equiv \alpha$, then

$$\begin{aligned}\pi_0 &= \left[1 + \sum_{j=1}^{M} \left(\frac{\alpha}{1-\alpha}\right)^j\right]^{-1} \\ &= \frac{1-\beta}{1-\beta^{M+1}}\end{aligned}$$

and, in general,

$$\pi_i = \frac{\beta^i(1-\beta)}{1-\beta^{M+1}}, \quad i = 0, 1, \ldots, M$$

where

$$\beta = \frac{\alpha}{1-\alpha}$$

Example 6.2. Consider an arbitrary connected graph having a number w_{ij} associated with arc (i, j) for each arc. One instance of such a graph is given by the following figure. Now consider a particle moving from node to node in this manner: If at any time the particle resides at node i, then it will next move to node j with probability P_{ij} where

$$P_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$

and where w_{ij} is 0 if (i, j) is not an arc. For instance, for the graph below, $P_{12} = 3/(3+1+2) = \frac{1}{2}$.

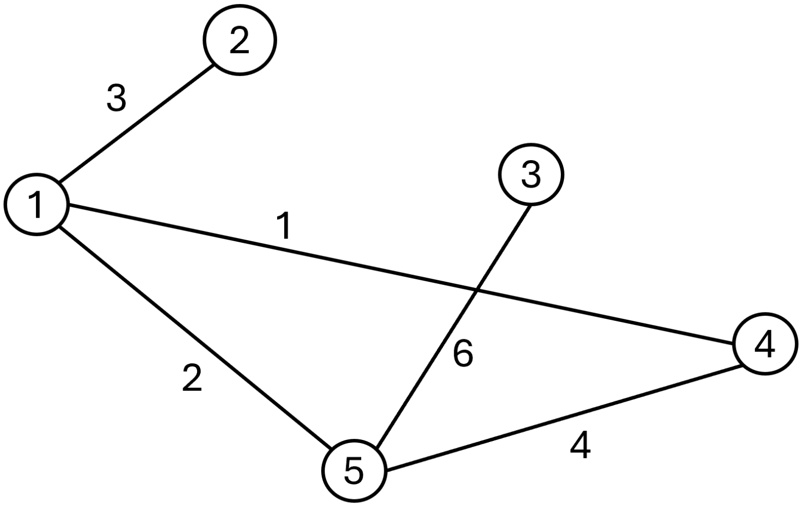


Figure 6.1: A connected graph with arc weights.

The time reversibility equations

$$\pi_i P_{ij} = \pi_j P_{ji}$$

reduce to

$$\pi_i \frac{w_{ij}}{\sum_j w_{ij}} = \pi_j \frac{w_{ji}}{\sum_i w_{ji}}$$

or, equivalently, since $w_{ij} = w_{ji}$

$$\frac{\pi_i}{\sum_j w_{ij}} = \frac{\pi_j}{\sum_i w_{ji}}$$

which is equivalent to

$$\frac{\pi_i}{\sum_j w_{ij}} = c$$

or

$$\pi_i = c \sum_j w_{ij}$$

or, since $1 = \sum_i \pi_i$

$$\pi_i = \frac{\sum_j w_{ij}}{\sum_i \sum_j w_{ij}}$$

Because the π_i's given by this equation satisfy the time reversibility equations, it follows that the process is time reversible with these limiting probabilities. For the graph above we have that

$$\pi_1 = \frac{6}{32}, \quad \pi_2 = \frac{3}{32}, \quad \pi_3 = \frac{6}{32}, \quad \pi_4 = \frac{5}{32}, \quad \pi_5 = \frac{12}{32}$$

.

If we try to solve Equation 6.2 for an arbitrary Markov chain with states $0, 1, \dots, M$, it will usually turn out that no solution exists. For example, from Equation 6.2,

$$x_i P_{ij} = x_j P_{ji},$$
$$x_k P_{kj} = x_j P_{jk}$$

implying (if $P_{ij} P_{jk} > 0$) that

$$\frac{x_i}{x_k} = \frac{P_{ji} P_{kj}}{P_{ij} P_{jk}}$$

which in general need not equal P_{ki}/P_{ik}. Thus, we see that a necessary condition for time reversibility is that

$$P_{ik} P_{kj} P_{ji} = P_{ij} P_{jk} P_{ki} \quad \text{for all } i, j, k \tag{6.5}$$

which is equivalent to the statement that, starting in state i, the path $i \to k \to j \to i$ has the same probability as the reversed path $i \to j \to k \to i$. To understand the necessity of this, note that time reversibility implies that the rate at which a sequence of transitions from i to k to j to i occurs must equal the rate of ones from i to j to k to i (why?), and so we must have

$$\pi_i P_{ik} P_{kj} P_{ji} = \pi_i P_{ij} P_{jk} P_{ki}$$

implying Equation 6.5 when $\pi_i > 0$.

In fact, we can show the following.

Theorem 6.1. *An ergodic Markov chain for which $P_{ij} = 0$ whenever $P_{ji} = 0$ is time reversible if and only if starting in state i, any path back to i has the same probability as the reversed path. That is, if*

$$P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,i} = P_{i,i_k} P_{i_k,i_{k-1}} \cdots P_{i_1,i} \tag{6.6}$$

for all states $i, i_1, \dots, i_k$.

Proof. We have already proven necessity. To prove sufficiency, fix states i and j and rewrite Equation 6.6 as

$$P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,j} P_{ji} = P_{ij} P_{j,i_k} \cdots P_{i_1,i}$$

Summing the preceding over all states $i_1, \dots, i_k$ yields

$$P_{ij}^{k+1} P_{ji} = P_{ij} P_{ji}^{k+1}$$

Letting $k \to \infty$ yields

$$\pi_j P_{ji} = P_{ij} \pi_i$$

which proves the theorem. □

The concept of the reversed chain is useful even when the process is not time reversible. To illustrate this, we start with the following proposition whose proof is left as an exercise.

Proposition 6.1. *Consider an irreducible Markov chain with transition probabilities P_{ij}. If we can find positive numbers $\pi_i, i \geq 0$, summing to one, and a transition probability matrix $Q = [Q_{ij}]$ such that*

$$\pi_i P_{ij} = \pi_j Q_{ji} \tag{6.7}$$

then the Q_{ij} are the transition probabilities of the reversed chain and the π_i are the stationary probabilities both for the original and reversed chain.

The importance of the preceding proposition is that, by thinking backward, we can sometimes guess at the nature of the reversed chain and then use the set of Equation 6.7 to obtain both the stationary probabilities and the Q_{ij}.

Example 6.3. A single bulb is necessary to light a given room. When the bulb in use fails, it is replaced by a new one at the beginning of the next day. Let X_n equal i if the bulb in use at the beginning of day n is in its ith day of use (that is, if its present age is i). For instance, if a bulb fails on day $n-1$, then a new bulb will be put in use at the beginning of day n and so $X_n = 1$. If we suppose that each bulb, independently, fails on its ith day of use with probability $p_i, i \geqslant 1$, then it is easy to see that $\{X_n, n \geqslant 1\}$ is a Markov chain whose transition probabilities are as follows:

$$\begin{aligned} P_{i,1} &= P\{\text{ bulb, on its } i\text{th day of use, fails}\} \\ &= P\{\text{life of bulb} = i | \text{life of bulb} \geqslant i\} \\ &= \frac{P\{L = i\}}{P\{L \geqslant i\}} \end{aligned}$$

where L, a random variable representing the lifetime of a bulb, is such that $P\{L = i\} = p_i$. Also,

$$P_{i,i+1} = 1 - P_{i,1}$$

Suppose now that this chain has been in operation for a long (in theory, an infinite) time and consider the sequence of states going backward in time. Since, in the forward direction, the state is always increasing by 1 until it reaches the age at which the item fails, it is easy to see that the reverse chain will always decrease by 1 until it reaches 1 and then it will jump to a random value representing the lifetime of the (in real time) previous bulb. Thus, it seems that the reverse chain should have transition probabilities given by

$$\begin{aligned} Q_{i,i-1} &= 1, \quad i > 1 \\ Q_{1,i} &= p_i, \quad i \geqslant 1 \end{aligned}$$

To check this, and at the same time determine the stationary probabilities, we must see if we can find, with the $Q_{i,j}$ as previously given, positive numbers $\{\pi_i\}$ such that

$$\pi_i P_{i,j} = \pi_j Q_{j,i}$$

To begin, let $j = 1$ and consider the resulting equations:

$$\pi_i \, P_{i,1} = \pi_1 \, Q_{1,i}$$

This is equivalent to

$$\pi_i \frac{P\{L = i\}}{P\{L \geqslant i\}} = \pi_1 P\{L = i\}$$

or

$$\pi_i = \pi_1 P\{L \geqslant i\}$$

Summing over all i yields

$$1 = \sum_{i=1}^{\infty} \pi_i = \pi_1 \sum_{i=1}^{\infty} P\{L \geqslant i\} = \pi_1 E[L]$$

and so, for the preceding Q_{ij} to represent the reverse transition probabilities, it is necessary for the stationary probabilities to be

$$\pi_i = \frac{P\{L \geqslant i\}}{E[L]}, \quad i \geqslant 1$$

To finish the proof that the reverse transition probabilities and stationary probabilities are as given, all that remains is to show that they satisfy

$$\pi_i P_{i,i+1} = \pi_{i+1} Q_{i+1,i}$$

which is equivalent to

$$\frac{P\{L \geqslant i\}}{E[L]} \left(1 - \frac{P\{L = i\}}{P\{L \geqslant i\}}\right) = \frac{P\{L \geqslant i+1\}}{E[L]}$$

and which is true since $P\{L \geqslant i\} - P\{L = i\} = P\{L \geqslant i+1\}$.

7 Hidden Markov Chains

7.1 Hidden Markov Chains

Let $X_n, n = 1, 2, ...$ be a Markov chain with transition probabilities $P_{i,j}$ and initial state probabilities $p_i = P\{X_1 = i\}$, $i \geq 0$. Suppose that there is a finite set $\mathcal{S}$ of signals, and that a signal from $\mathcal{S}$ is emitted each time the Markov chain enters a state. Further, suppose that when the Markov chain enters state j then, independently of previous Markov chain states and signals, the signal emitted is s with probability $p(s|j), \sum_{s \in \mathcal{S}} p(s|j) = 1$. That is, if S_n represents the n-th signal emitted, then

$$\begin{aligned} P\{S_1 = s | X_1 = j\} &= p(s|j) \\ P\{S_n = s | X_1, S_1, ..., X_{n-1}, S_{n-1}, X_n = j\} &= p(s|j) \end{aligned}$$

A model of the preceding type in which the sequence of signals $S_1, S_2, ...$ is observed, while the sequence of underlying Markov chain states $X_1, X_2, ...$ is unobserved, is called a *hidden Markov chain* model.

Example 7.1. Consider a production process that in each period is either in a good state (state 1) or in a poor state (state 2). If the process is in state 1 during a period then, independent of the past, with probability 0.9 it will be in state 1 during the next period and with probability 0.1 it will be in state 2. Once in state 2, it remains in that state forever. Suppose that a single item is produced each period and that each item produced when the process is in state 1 is of acceptable quality with probability 0.99, while each item produced when the process is in state 2 is of acceptable quality with probability 0.96.

If the status, either acceptable or unacceptable, of each successive item is observed, while the process states are unobservable, then the preceding is a hidden Markov chain model. The signal is the status of the item produced, and has value either a or u, depending on whether the item is acceptable or unacceptable. The signal probabilities are

$$\begin{aligned} p(u|1) = 0.01, &\quad p(a|1) = 0.99, \\ p(u|2) = 0.04, &\quad p(a|2) = 0.96, \end{aligned}$$

while the transition probabilities of the underlying Markov chain are

$$P_{1,1} = 0.9 = 1 - P_{1,2}, \quad P_{2,2} = 1$$

Although $\{S_n, n \geq 1\}$ is not a Markov chain, it should be noted that, conditional on the current state X_n, the sequence $S_n, X_{n+1}, S_{n+1}, \ldots$ of future signals and states is independent of the sequence $X_1, S_1, \ldots, X_{n-1}, S_{n-1}$ of past states and signals.

Let $S^n = (S_1, \ldots, S_n)$ be the random vector of the first n signals. For a fixed sequence of signals $s_1, \ldots, s_n$, let $s_k = (s_1, \ldots, s_k)$, $k \leq n$. To begin, let us determine the conditional probability of the Markov chain state at time n given that $S^n = s_n$. To obtain this probability, let

$$F_n(j) = P\{S^n = s_n, X_n = j\}$$

and note that

$$P\{X_n = j | S^n = s_n\} = \frac{P\{S^n = s_n, X_n = j\}}{P\{S^n = s_n\}} = \frac{F_n(j)}{\sum_i F_n(i)}$$

Now,

$$\begin{aligned} F_n(j) &= P\{S^{n-1} = s_{n-1}, S_n = s_n, X_n = j\} \\ &= \sum_i P\{S^{n-1} = s_{n-1}, X_{n-1} = i, X_n = j, S_n = s_n,\} \\ &= \sum_i F_{n-1}(i) P\{X_n = j, S_n = s_n | S^{n-1} = s_{n-1}, X_{n-1} = i\} \\ &= \sum_i F_{n-1}(i) P\{X_n = j, S_n = s_n | X_{n-1} = i\} \\ &= \sum_i F_{n-1}(i) P_{i,j} p(s_n|j) \\ &= p(s_n|j) \sum_i F_{n-1}(i) P_{i,j} \end{aligned} \tag{7.1}$$

where the preceding used that

$$\begin{aligned} P\{X_n = j, S_n = s_n | X_{n-1} = i\} &= P\{S_n = s_n | X_n = j, X_{n-1} = i\} P\{X_n = j | X_{n-1} = i\} \\ &= P\{S_n = s_n | X_n = j\} P_{i,j} = p(s_n|j) P_{i,j} \end{aligned}$$

Starting with

$$F_1(i) = P\{X_1 = i, S_1 = s_1\} = p_i p(s_1|i)$$

we can use Equation Equation 7.1 to recursively determine the functions $F_2(i), F_3(i), \ldots$ up to $F_n(i)$.

Example 7.2. Suppose in Example 25.1 that $P\{X_1 = 1\} = 0.8$. Given that the successive conditions of the first 3 items produced are a, u, a,

(i) what is the probability that the process was in its good state when the third item was produced;
(ii) what is the probability that X_4 is 1;
(iii) what is the probability that the next item produced is acceptable?

Solution 7.1. With $\mathbf{s}_3 = (a, u, a)$, we have

$$\begin{aligned}
&F_1(1) = (0.8)(0.99) = 0.792,\\
&F_1(2) = (0.2)(0.96) = 0.192\\
&F_2(1) = 0.01[0.792(0.9) + 0.192(0)] = 0.007128,\\
&F_2(2) = 0.04[0.792(0.1) + (0.192)(1)] = 0.010848\\
&F_3(1) = 0.99[(0.007128)(0.9)] \approx 0.006351,\\
&F_3(2) = 0.96[(0.007128)(0.1) + 0.010848] \approx 0.011098
\end{aligned}$$

Therefore, the answer to part (i) is

$$P\{X_3 = 1 \mid \mathbf{s}_3\} \approx \frac{0.006351}{0.006351 + 0.011098} \approx 0.364$$

To compute $P\{X_4 = 1 \mid \mathbf{s}_3\}$, condition on X_3 to obtain

$$\begin{aligned}
P\{X_4 = 1 \mid \mathbf{s}_3\} =& P\{X_4 = 1 \mid X_3 = 1, \mathbf{s}_3\} P\{X_3 = 1 \mid \mathbf{s}_3\}\\
&+ P\{X_4 = 1 \mid X_3 = 2, \mathbf{s}_3\} P\{X_3 = 2 \mid \mathbf{s}_3\}\\
=& P\{X_4 = 1 \mid X_3 = 1, \mathbf{s}_3\}(0.364) + P\{X_4 = 1 \mid X_3 = 2, \mathbf{s}_3\}(0.636)\\
=& 0.364P_{1,1} + 0.636P_{2,1}\\
=& 0.3276
\end{aligned}$$

To compute $P\{S_4 = a \mid \mathbf{s}_3\}$, condition on X_4

$$\begin{aligned}
P\{S_4 = a \mid \mathbf{s}_3\} =& P\{S_4 = a \mid X_4 = 1,\ s_3\} P\{X_4 = 1 \mid s_3\}\\
&+ P\{S_4 = a \mid X_4 = 2, \mathbf{s}_3\} P\{X_4 = 2 \mid \mathbf{s}_3\}\\
=& P\{S_4 = a \mid X_4 = 1\}(0.3276) + P\{S_4 = a \mid X_4 = 2\}(1 - 0.3276)\\
=& (0.99)(0.3276) + (0.96)(0.6724) = 0.9698
\end{aligned}$$

We now use Python to simulate the conditional probabilities and compare the simulated results with what we obtained above.

```python
import numpy as np

def HMM_cond_prob(P, p_init, p_signal, s_obs, sample_size):
    """
    Compute the conditional probabilities p(X_n=j | S^n=s_n), the probability that X_n=j give

    input:
    P: 2d array of shape (#ofstates, #ofstates), transition probability
    p_init: 1d array of shape (#ofstates), the initial probability of states
    p_signal: 2d array of shape (#ofstates, #ofsignals), the probability of signal given sta
              the row sums should be equal to 1
    s_obs: 1d array of shape (#ofsignalobservations), the collection of observed signals, co
    sample_size: int, the sample size, i.e., how many experiments to perform

    output:
    prob, 1d array of shape (#ofstates)
    """
    n_states = P.shape[0]  # the number of states
    n = s_obs.size  # the time n
    count = np.zeros(n_states, dtype='int')  # count how many times each state j happens at
    for i in range(sample_size):
        # keep simulating the process until we get exactly the signal observations s_obs
        # This will count as one successful simulation
        signals = np.zeros(n, dtype='int')
        while True:
            state = np.where(np.random.random() < np.cumsum(p_init))[0][0] + 1  # initial sta
            signals[0] = np.where(np.random.random() < np.cumsum(p_signal[state-1,:]))[0][0]
            for j in range(n-1):
                state = np.where(np.random.random() < np.cumsum(P[state-1,:]))[0][0] + 1  # 
                signals[j+1] = np.where(np.random.random() < np.cumsum(p_signal[state-1,:]))
            if np.array_equal(signals, s_obs):
                count[state-1] += 1  # one case where X_n=state is found
                break
    prob = count / count.sum()

    return prob
```

```
# State 1: good state, State 2: poor state
# Signal 1: unacceptable, Signal 2: acceptable
P = np.array([[0.9, 0.1], [0.0, 1.0]])
p_init = np.array([0.8, 0.2])
p_signal = np.array([[0.01, 0.99], [0.04, 0.96]])
s_obs = np.array([2, 1, 2], dtype='int')
sample_size = 10000

prob = HMM_cond_prob(P, p_init, p_signal, s_obs, sample_size)

print('Computed p(X3=1|s3) = ', prob[0], 'Computed p(X3=2|s3) = ', prob[1] )
```

The probabilities obtained in Part (i) have been verified numerically. Similarly, one can slightly modify the HMM_cond_prob function to compute the probabilities in Parts (ii) and (iii).

To compute $P\left\{\mathbf{S}^n=\mathbf{s}_n\right\}$, use the identity $P\left\{\mathbf{S}^n=\mathbf{s}_n\right\}=\sum_i F_n(i)$ along with the recursion Equation 7.1. If there are N states of the Markov chain, this requires computing nN quantities $F_n(i)$, with each computation requiring a summation over N terms. This can be compared with a computation of $P\left\{\mathbf{S}^n=\mathbf{s}_n\right\}$ based on conditioning on the first n states of the Markov chain to obtain

$$\begin{aligned}
P\left\{\mathbf{S}^n=\mathbf{s}_n\right\} &= \sum_{i_1,\dots,i_n} P\left\{\mathbf{S}^n=\mathbf{s}_n \mid X_1=i_1,\dots,X_n=i_n\right\} P\left\{X_1=i_1,\dots,X_n=i_n\right\} \\
&= \sum_{i_1,\dots,i_n} p\left(s_1 \mid i_1\right)\cdots p\left(s_n \mid i_n\right) p_{i_1} P_{i_1,i_2} P_{i_2,i_3}\cdots P_{i_{n-1},i_n}
\end{aligned}$$

The use of the preceding identity to compute $P\left\{\mathbf{S}^n=\mathbf{s}_n\right\}$ would thus require a summation over N^n terms, with each term being a product of $2n$ values, indicating that it is not competitive with the previous approach.

The computation of $P\left\{\mathbf{S}^n=\mathrm{s}_n\right\}$ by recursively determining the functions $F_k(i)$ is known as the *forward approach*. There also is a *backward approach*, which is based on the quantities $B_k(i)$, defined by

$$B_k(i)=P\left\{S_{k+1}=s_{k+1},\dots,S_n=s_n \mid X_k=i\right\}$$

A recursive formula for $B_k(i)$ can be obtained by conditioning on X_{k+1}.

$$\begin{aligned}
B_k(i) &= \sum_j P\left\{S_{k+1}=s_{k+1},\dots,S_n=s_n \mid X_k=i, X_{k+1}=j\right\} P\left\{X_{k+1}=j \mid X_k=i\right\} \\
&= \sum_j P\left\{S_{k+1}=s_{k+1},\dots,S_n=s_n \mid X_{k+1}=j\right\} P_{i,j} \\
&= \sum_j P\left\{S_{k+1}=s_{k+1} \mid X_{k+1}=j\right\} \\
&\quad \times P\left\{S_{k+2}=s_{k+2},\dots,S_n=s_n \mid S_{k+1}=s_{k+1}, X_{k+1}=j\right\} P_{i,j} \\
&= \sum_j p\left(s_{k+1} \mid j\right) P\left\{S_{k+2}=s_{k+2},\dots,S_n=s_n \mid X_{k+1}=j\right\} P_{i,j} \\
&= \sum_j p\left(s_{k+1} \mid j\right) B_{k+1}(j) P_{i,j}
\end{aligned} \tag{7.2}$$

Starting with

$$\begin{aligned}
B_{n-1}(i) &= P\left\{S_n=s_n \mid X_{n-1}=i\right\} \\
&= \sum_j P_{i,j} p\left(s_n \mid j\right)
\end{aligned}$$

we would then use Equation Equation 7.2 to determine the function $B_{n-2}(i)$, then $B_{n-3}(i)$, and so on, down to $B_1(i)$. This would then yield $P\left\{\mathbf{S}^n=\mathbf{s}_n\right\}$ via

$$\begin{aligned}
P\left\{\mathbf{S}^n=\mathrm{s}_n\right\} &= \sum_i P\left\{S_1=s_1,\dots,S_n=s_n \mid X_1=i\right\} p_i \\
&= \sum_i P\left\{S_1=s_1 \mid X_1=i\right\} P\left\{S_2=s_2,\dots,S_n=s_n \mid S_1=s_1, X_1=i\right\} p_i
\end{aligned}$$

$$
\begin{aligned}
&= \sum_i p\left(s_1 \mid i\right) P\left\{S_2 = s_2, \ldots, S_n = s_n \mid X_1 = i\right\} p_i \\
&= \sum_i p\left(s_1 \mid i\right) B_1(i) p_i
\end{aligned}
$$

Another approach to obtaining $P\left\{\mathbf{S}^n = \mathbf{s}_n\right\}$ is to combine both the forward and backward approaches. Suppose that for some k we have computed both functions $F_k(j)$ and $B_k(j)$. Because

$$
\begin{aligned}
P\left\{\mathbf{S}^n = \mathbf{s}_n, X_k = j\right\} =& P\left\{\mathbf{S}^k = \mathbf{s}_k, X_k = j\right\} \\
& \times P\left\{S_{k+1} = s_{k+1}, \ldots, S_n = s_n \mid \mathbf{S}^k = \mathbf{s}_k, X_k = j\right\} \\
=& P\left\{\mathbf{S}^k = \mathbf{s}_k, X_k = j\right\} P\left\{S_{k+1} = s_{k+1}, \ldots, S_n = s_n \mid X_k = j\right\} \\
=& F_k(j) B_k(j)
\end{aligned}
$$

we see that

$$
P\left\{\mathbf{S}^n = \mathbf{s}_n\right\} = \sum_j F_k(j) B_k(j)
$$

The beauty of using the preceding identity to determine $P\left\{\mathbf{S}^n = \mathbf{s}_n\right\}$ is that we may simultaneously compute the sequence of forward functions, starting with F_1, as well as the sequence of backward functions, starting at B_{n-1}. The parallel computations can then be stopped once we have computed both F_k and B_k for some k.

8 The Exponential Distribution

8.1 Definition

A continuous random variable X is said to have an exponential distribution with parameter $\lambda, \lambda > 0$, if its probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

or, equivalently, if its cdf is given by

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \begin{cases} 1 - e^{-\lambda x}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

The mean of the exponential distribution, $E[X]$, is given by

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x)\,dx \\ &= \int_{0}^{\infty} \lambda x e^{-\lambda x}\,dx \end{aligned}$$

Integrating by parts ($u = x, dv = \lambda e^{-\lambda x} dx$) yields

$$E[X] = -x e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}\,dx = \frac{1}{\lambda}$$

The moment generating function $\phi(t)$ of the exponential distribution is given by

$$\begin{aligned} \phi(t) &= E[e^{tX}] \\ &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x}\,dx \\ &= \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda \end{aligned} \tag{8.1}$$

All the moments of X can now be obtained by differentiating Equation 8.1. For example,

$$\begin{aligned}E[X^2] &= \frac{d^2}{dt^2}\phi(t)\bigg|_{t=0} \\ &= \frac{2\lambda}{(\lambda - t)^3}\bigg|_{t=0} \\ &= \frac{2}{\lambda^2}\end{aligned}$$

Consequently,

$$\begin{aligned}\text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2}\end{aligned}$$

Example 8.1. (Exponential Random Variables and Expected Discounted Returns) Suppose that you are receiving rewards at randomly changing rates continuously throughout time. Let $R(x)$ denote the random rate at which you are receiving rewards at time x. For a value $\alpha \geqslant 0$, called the discount rate, the quantity

$$R = \int_0^\infty e^{-\alpha x} R(x)\, dx$$

represents the total discounted reward. (In certain applications, α is a continuously compounded interest rate, and R is the present value of the infinite flow of rewards.) Whereas

$$E[R] = E\left[\int_0^\infty e^{-\alpha x} R(x)\, dx\right] = \int_0^\infty e^{-\alpha x} E[R(x)]\, dx$$

is the expected total discounted reward, we will show that it is also equal to the expected total reward earned up to an exponentially distributed random time with rate α.

Let T be an exponential random variable with rate α, that is independent of all the random variables $R(x)$. We want to argue that

$$\int_0^\infty e^{-\alpha x} E[R(x)]\, dx = E\left[\int_0^T R(x)\, dx\right]$$

To show this, define for each $x \geqslant 0$, a random variable $I(x)$ by

$$I(x) = \begin{cases} 1, & \text{if } x \leqslant T \\ 0, & \text{if } x > T \end{cases}$$

and note that

$$\int_0^T R(x)\, dx = \int_0^\infty R(x) I(x)\, dx$$

Thus,

$$\begin{aligned} E\left[\int_0^T R(x)\,dx\right] &= E\left[\int_0^\infty R(x)I(x)\,dx\right] \\ &= \int_0^\infty E[R(x)I(x)]dx \\ &= \int_0^\infty E[R(x)]E[I(x)]dx \qquad \text{by independence} \\ &= \int_0^\infty E[R(x)]P\{T \geqslant x\}\,dx \\ &= \int_0^\infty e^{-\alpha x}E[R(x)]\,dx \end{aligned}$$

Therefore, the expected total discounted reward is equal to the expected total (undiscounted) reward earned by a random time that is exponentially distributed with a rate equal to the discount factor.

8.2 Properties of the Exponential Distribution

A random variable X is said to be without memory, or *memoryless*, if

$$P\{X > s + t \mid X > t\} = P\{X > s\} \quad \text{for all } s, t \geqslant 0 \tag{8.2}$$

If we think of X as being the lifetime of some instrument, then Equation 8.2 states that the probability that the instrument lives for at least $s + t$ hours given that it has survived t hours is the same as the initial probability that it lives for at least s hours. In other words, if the instrument is alive at time t,then the distribution of the remaining amount of time that it survives is the same as the original lifetime distribution; that is, the instrument does not remember that it has already been in use for a time t. The condition in Equation 8.2 is equivalent to

$$\frac{P\{X > s + t,\ X > t\}}{P\{X > t\}} = P\{X > s\}$$

or

$$P\{X > s + t\} = P\{X > s\}P\{X > t\} \tag{8.3}$$

Since Equation 8.3 is satisfied when X is exponentially distributed (for $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda I}$), it follows that exponentially distributed random variables are memoryless

Example 8.2. Suppose that the amount of time one spends in a bank is exponentially distributed with mean ten minutes, that is, $\lambda = \frac{1}{10}$. What is the probability that a customer will spend more than fifteen minutes in the bank? What is the probability that a customer will spend more than fifteen minutes in the bank given that she is still in the bank after ten minutes?

Solution 8.1. If X represents the amount of time that the customer spends in the bank, then the first probability is just

$$P\{X > 15\} = e^{-15\lambda} = e^{-3/2} \approx 0.220$$

The second question asks for the probability that a customer who has spent ten minutes in the bank will have to spend at least five more minutes. However, since the exponential distribution does not “remember” that the customer has already spent ten minutes in the bank, this must equal the probability that an entering customer spends at least five minutes in the bank. That is, the desired probability is just

$$P\{X > 5\} = e^{-5\lambda} = e^{-1/2} \approx 0.604$$

Python simulations can be used to estimate these probabilities as shown below.

```python
import numpy as np
from scipy.stats import expon

# What is the probability that a customer will spend
# more than fifteen minutes in the bank?
sample_size = 100000
times = expon.rvs(scale = 10.0, size = sample_size)
prob = np.sum(times > 15) / sample_size
print('the probability that a customer will spend more than '
      'fifteen minutes in the bank is: \n', prob,
      ' and the theorectical probability is: ', np.exp(-3/2))
```

```python
# What is the probability that a customer will spend more than fifteen minutes
# in the bank given that she is still in the bank after ten minutes?
sample_size = 10000
count = 0
times_gt_10 = np.zeros(sample_size)  # store all the cases where time is greater than 10
while count < sample_size:
    time = expon.rvs(scale = 10.0)
    if time >= 10:
        times_gt_10[count] = time
        count += 1

prob = np.sum(times_gt_10 > 15) / sample_size
```

```
print('the probability that a customer will spend more than fifteen minutes \n '
      'in the bank given that she is still in the bank after ten minutes is: \n', prob,
      ' and the theorectical probability is: ', np.exp(-1/2))
```

Example 8.3. Consider a post office that is run by two clerks. Suppose that when Mr. Smith enters the system he discovers that Mr. Jones is being served by one of the clerks and Mr. Brown by the other. Suppose also that Mr. Smith is told that his service will begin as soon as either Jones or Brown leaves. If the amount of time that a clerk spends with a customer is exponentially distributed with mean $1/\lambda$, what is the probability that, of the three customers, Mr. Smith is the last to leave the post office?

The answer is obtained by this reasoning: Consider the time at which Mr. Smith first finds a free clerk. At this point either Mr. Jones or Mr. Brown would have just left and the other one would still be in service. However, by the lack of memory of the exponential, it follows that the amount of time that this other man (either Jones or Brown) would still have to spend in the post office is exponentially distributed with mean $1/\lambda$. That is, it is the same as if he were just starting his service at this point. Hence, by symmetry, the probability that he finishes before Smith must equal $\frac{1}{2}$. .

Now we numerically simulate the processes and compute the probability.

```
sample_size = 10000
lam = 0.1  # arbitrarily picked
count = 0
for i in range(sample_size):
    [t_Jones, t_Brown] = expon.rvs(scale = 1/lam, size=2)  # Generate the service time for Jo
    if expon.rvs(scale = 1/lam) > max(t_Jones, t_Brown) - min(t_Jones, t_Brown):  # Mr. Smit
        count += 1
prob = count / sample_size
print('The probability that Mr. Smith is the last to leave is: ', prob,
      '\n and the theoretical probability is: ', 1/2)
```

Example 8.4. The dollar amount of damage involved in an automobile accident is an exponential random variable with mean 1000. Of this, the insurance company only pays that amount exceeding (the deductible amount of) 400. Find the expected value and the standard deviation of the amount the insurance company pays per accident.

Solution 8.2. If X is the dollar amount of damage resulting from an accident, then the amount paid by the insurance company is $(X - 400)^+$, (where a^+ is defined to equal a if $a > 0$ and to equal 0 if $a \leqslant 0$). Whereas we could certainly determine the expected value and variance of $(X-400)^+$ from first principles, it is easier to condition on whether X exceeds 400. So,let

$$I = \begin{cases} 1, & \text{if} X > 400 \\ 0, & \text{if} X \leqslant 400 \end{cases}$$

Let $Y = (X-400)^+$ be the amount paid. By the lack of memory property of the exponential, it follows that if a damage amount exceeds 400, then the amount by which it exceeds it is exponential with mean 1000. Therefore,

$$\begin{aligned} E[Y|I=1] &= 1000 \\ E[Y|I=0] &= 0 \\ \operatorname{Var}(Y|I=1) &= (1000)^2 \\ \operatorname{Var}(Y|I=0) &= 0 \end{aligned}$$

which can be conveniently written as

$$E[Y|I] = 10^3 I, \quad \operatorname{Var}(Y|I) = 10^6 I$$

Because I is a Bernoulli random variable that is equal to 1 with probability $e^{-0.4}$, we obtain

$$E[Y] = E\Big[E[Y|I]\Big] = 10^3 E[I] = 10^3 e^{-0.4} \approx 670.32$$

and, by the conditional variance formula

$$\begin{aligned} \operatorname{Var}(Y) &= E\Big[\operatorname{Var}(Y|I)\Big] + \operatorname{Var}\left(E[Y|I]\right) \\ &= 10^6 e^{-0.4} + 10^6 e^{-0.4}(1-e^{-0.4}) \end{aligned}$$

where the final equality used that the variance of a Bernoulli random variable with parameter p is $p(1-p)$. Consequently,

$$\sqrt{\operatorname{Var}(Y)} \approx 944.09$$

Numerical results are as follows:

```
n_accidents = 1000000
deductible = 400
lam = 1/1000
payment = expon.rvs(scale=1/lam, size=n_accidents)
payment[payment <= deductible] = 0.0  # if deductible is >= payment for an accident, payment
payment[payment > deductible] -= deductible  # otherwise, payment = payment - deductible
mean = np.mean(payment)
std = np.std(payment)
```

```
print('The expected value of the amount the insurance companypays per accident is: ', mean,
      '\n and the theoretical probability is: ', 10**3*np.exp(-0.4))
print('The standard deviation of the amount the insurance companypays per accident is: ', st
      '\n and the theoretical probability is: ', np.sqrt(10**6*np.exp(-0.4)+10**6*np.exp(-0.4
```

It turns out that not only is the exponential distribution "memoryless" but it is the unique distribution possessing this property. To see this, suppose that X is memoryless and let $\bar{F}(x) = P\{X > x\}$. Then by Equation 8.3 it follows that

$$\bar{F}(s+t) = \bar{F}(s)\bar{F}(t)$$

That is, $\bar{F}(x)$ satisfies the functional equation

$$g(s+t) = g(s)g(t)$$

However, it turns out that the only right continuous solution of this functional equation is

$$g(x) = e^{-\lambda x *}$$

and since a distribution function is always right continuous we must have

$$\bar{F}(x) = e^{-\lambda x}$$

or

$$F(x) = P\{X \leqslant x\} = 1 - e^{-\lambda x}$$

which shows that X is exponentially distributed.

Example 8.5. Let $X_1, \dots, X_n$ be independent exponential random variables with respective rates $\lambda_1, \dots, \lambda_n$, where $\lambda_i \neq \lambda_j$ when $i \neq j$. Let T be independent of these random variables and suppose that

$$\sum_{j=1}^{n} P_j = 1 \quad \text{where} P_j = P\{T = j\}$$

The random variable X_T is said to be a *hyperexponential* random variable. To see how such a random variable might originate, imagine that a bin contains n different types of batteries, with a type j battery lasting for an exponential distributed time with rate $\lambda_j, j = 1, \dots, n$. Suppose further that P_j is the proportion of batteries in the bin that are type j for each $j = 1, \dots, n$. If a battery is randomly chosen, in the sense that it is equally likely to be any of the batteries in the bin, then the lifetime of the battery selected will have the hyperexponential distribution specified in the preceding.

To obtain the distribution function F of $X = X_T$, condition on T. This yields

$$\begin{aligned}1 - F(t) &= P\{X > t\}\\ &= \sum_{i=1}^{n} P\{X > t|T = i\}P\{T = i\}\\ &= \sum_{i=1}^{n} P_i e^{-\lambda_i t}\end{aligned}$$

Differentiation of the preceding yields f, the density function of X.

$$f(t) = \sum_{i=1}^{n} \lambda_i P_i e^{-\lambda_i t}$$

Consequently, the failure rate function of a hyperexponential random variable is

$$r(t) = \frac{\sum_{j=1}^{n} P_j \lambda_j e^{-\lambda_j t}}{\sum_{i=1}^{n} P_i e^{-\lambda_i t}}$$

By noting that

$$\begin{aligned}P\{T = j|X > t\} &= \frac{P\{X > t|T = j\}P\{T = j\}}{P\{X > t\}}\\ &= \frac{P_j e^{-\lambda_j t}}{\sum_{i=1}^{n} P_i e^{-\lambda_i t}}\end{aligned}$$

we see that the failure rate function $r(t)$ can also be written as

$$r(t) = \sum_{j=1}^{n} \lambda_j P\{T = j|X > t\}$$

If $\lambda_1 < \lambda_i$, for all $i > 1$,then

$$\begin{aligned}P\{T = 1|X > t\} = &\ \frac{P_1 e^{-\lambda_1 t}}{P_1 e^{-\lambda_1 t} + \sum_{i=2}^{n} P_i e^{-\lambda_i t}}\\ = &\ \frac{P_1}{P_1 + \sum_{i=2}^{n} P_i e^{-(\lambda_i - \lambda_1)t}}\\ &\to 1 \quad \text{as}\, t \to \infty\end{aligned}$$

Similarly, $P\left\{T = i|X > t\right\} \to 0$ when $i \neq 1$, thus showing that

$$\lim_{t \to \infty} r(t) = \min_i \lambda_i$$

That is, as a randomly chosen battery ages its failure rate converges to the failure rate of the exponential type having the smallest failure rate, which is intuitive since the longer the battery lasts, the more likely it is a battery type with the smallest failure rate.

9 Further Properties of the Exponential Distribution

Let $X_1, \dots, X_n$ be independent and identically distributed exponential random variables having mean $1/\lambda$. $X_1 + \cdots + X_n$ has a gamma distribution with parameters n and λ. Let us now give a second verification of this result by using mathematical induction. Because there is nothing to prove when $n = 1$, let us start by assuming that $X_1 + \cdots + X_{n-1}$ has density given by

$$f_{X_1+\cdots+X_{n-1}}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-2}}{(n-2)!}$$

Hence,

$$\begin{aligned} f_{X_1+\cdots+X_{n-1}+X_n}(t) &= \int_0^\infty f_{X_n}(t-s) f_{X_1+\cdots+X_{n-1}}(s)\, ds \\ &= \int_0^t \lambda e^{-\lambda(t-s)} \lambda e^{-\lambda s} \frac{(\lambda s)^{n-2}}{(n-2)!}\, ds \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \end{aligned}$$

which proves the result.

Another useful calculation is to determine the probability that one exponential random variable is smaller than another. That is, suppose that X_1 and X_2 are independent exponential random variables with respective means $1/\lambda_1$ and $1/\lambda_2$; what is $P\{X_1 < X_2\}$? This probability is easily calculated by conditioning on X_1 :

$$\begin{aligned} P\{X_1 < X_2\} &= \int_0^\infty P\{X_1 < X_2 | X_1 = x\} \lambda_1 e^{-\lambda_1 x}\, dx \\ &= \int_0^\infty P\{x < X_2\} \lambda_1 e^{-\lambda_1 x}\, dx \\ &= \int_0^\infty e^{-\lambda_2 x} \lambda_1 e^{-\lambda_1 x}\, dx \qquad (9.1) \\ &= \int_0^\infty \lambda_1 e^{-(\lambda_1+\lambda_2)x}\, dx \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2} \end{aligned}$$

Suppose that $X_1, X_2, \dots, X_n$ are independent exponential random variables, with X_i having rate $\mu_i, i = 1, \dots, n$. It turns out that the smallest of the X_i is exponential with a rate equal to the sum of the μ_i. This is shown as follows:

$$\begin{aligned} P\{\text{minimum}(X_1, \dots, X_n) > x\} &= P\{X_i > x \text{ for each } i = 1, \dots, n\} \\ &= \prod_{i=1}^{n} P\{X_i > x\} \quad \text{(by independence)} \\ &= \prod_{i=1}^{n} e^{-\mu_i x} \\ &= \exp\left\{ -\left(\sum_{i=1}^{n} \mu_i \right) x \right\} \end{aligned} \tag{9.2}$$

Example 9.1. (Analyzing Greedy Algorithms for the Assignment Problem) A group of n people is to be assigned to a set of n jobs, with one person assigned to each job. For a given set of n^2 values $C_{ij}, i, j = 1, \dots, n$, a cost C_{ij} is incurred when person i is assigned to job j. The classical assignment problem is to determine the set of assignments that minimizes the sum of the n costs incurred.

Rather than trying to determine the optimal assignment, let us consider two heuristic algorithms for solving this problem. The first heuristic is as follows. Assign person 1 to the job that results in the least cost. That is, person 1 is assigned to job j_1 where $C(1, j_1) = \text{minimum}_j C(1, j)$. Now eliminate that job from consideration and assign person 2 to the job that results in the least cost. That is, person 2 is assigned to job j_2 where $C(2, j_2) = \text{minimum}_{j \neq j_1} C(2, j)$. This procedure is then continued until all n persons are assigned. Since this procedure always selects the best job for the person under consideration, we will call it Greedy Algorithm A.

The second algorithm, which we call Greedy Algorithm B, is a more "global" version of the first greedy algorithm. It considers all n^2 cost values and chooses the pair i_1, j_1 for which $C(i, j)$ is minimal. It then assigns person i_1 to job j_1. It then eliminates all cost values involving either person i_1 or job j_1 [so that $(n-1)^2$ values remain] and continues in the same fashion. That is, at each stage it chooses the person and job that have the smallest cost among all the unassigned people and jobs.

Under the assumption that the C_{ij} constitute a set of n^2 independent exponential random variables each having mean 1, which of the two algorithms results in a smaller expected total cost?

Solution 9.1. Suppose first that Greedy Algorithm A is employed. Let C_i denote the cost associated with person $i, i = 1, \dots, n$. Now C_1 is the minimum of n independent exponentials each having rate 1; so by Equation 9.2 it will be exponential with rate n. Similarly, C_2 is the minimum of $n-1$ independent exponentials with rate l, and so is exponential with rate $n-1$.

Indeed, by the same reasoning C_i will be exponential with rate $n - i + 1, i = 1, \dots, n$. Thus, the expected total cost under Greedy Algorithm A is

$$\begin{aligned} E_A[\text{total cost}] &= E[C_1 + \cdots + C_n] \\ &= \sum_{i=1}^{n} 1/i \end{aligned}$$

Let us now analyze Greedy Algorithm B. Let C_i be the cost of the i th personjob pair assigned by this algorithm. Since C_1 is the minimum of all the n^2 values C_{ij}, it follows from Equation 9.2 that C_1 is exponential with rate n^2. Now, it follows from the lack of memory property of the exponential that the amounts by which the other C_{ij} exceed C_1 will be independent exponentials with rates 1. As a result, C_2 is equal to C_1 plus the minimum of $(n-1)^2$ independent exponentials with rate l. Similarly, C_3 is equal to C_2 plus the minimum of $(n-2)^2$ independent exponentials with rate 1, and so on. Therefore, we see that

$$\begin{aligned} &E[C_1] = 1/n^2, \\ &E[C_2] = E[C_1] + 1/(n-1)^2, \\ &E[C_3] = E[C_2] + 1/(n-2)^2, \\ &E[C_j] = E[C_{j-1}] + 1/(n-j+1)^2, \\ &E[C_n] = E[C_{n-1}] + 1 \end{aligned}$$

Therefore,

$$\begin{aligned} &E[C_1] = 1/n^2, \\ &E[C_2] = 1/n^2 + 1/(n-1)^2, \\ &E[C_3] = 1/n^2 + 1/(n-1)^2 + 1/(n-2)^2, \\ &E[C_n] = 1/n^2 + 1/(n-1)^2 + 1/(n-2)^2 + \cdots + 1 \end{aligned}$$

Adding up all the $E[C_i]$ yields that

$$\begin{aligned} E_B[\text{total cost}] &= n/n^2 + (n-1)/(n-1)^2 + (n-2)/(n-2)^2 + \cdots + 1 \\ &= \sum_{i=1}^{n} \frac{1}{i} \end{aligned}$$

The expected cost is thus the same for both greedy algorithms.

Next, we numerically verify the results.

```python
import numpy as np
from scipy.stats import expon

def cost_algo_A(C):
    """
    Given the cost matrix, compute the cost using greedy algorithm A

    input:
    C: 2d square numpy array of shape(#of people, #of people), the costs
       C will be changed in place inside the cod

    ouptut:
    min_cost: float, the minimum cost from greedy algorithm A
    """
    n_people = C.shape[0]
    n_job = n_people

    min_cost = 0.0
    for i in range(n_people):
        min_cost += np.min(C[i,:])
        C[:,np.argmin(C[i,:])] = np.inf  # make sure assigned jobs are not assigned again
    return min_cost
```

```
n = 50  # fifty people and jobs
lam = 1.0
sample_size = 10000
EcostA = 0.0  # expectation of cost computed from algo A
for i in range(sample_size):
    C = expon.rvs(scale=1/lam, size=(n,n))
    EcostA += cost_algo_A(C)
EcostA /= sample_size
print('For algorithm A, the computed expectation of cost is: ', EcostA,
      '\n while the theoretical value is: ', np.sum(1/np.arange(1, n+1)))
```

```
def cost_algo_B(C):
    """
    Given the cost matrix, compute the cost using greedy algorithm B

    input:
    C: 2d square numpy array of shape(#of people, #of people), the costs
       C will be changed in place inside the cod

    ouptut:
    min_cost: float, the minimum cost from greedy algorithm B
    """
    n_people = C.shape[0]
    n_job = n_people
    min_cost = 0.0
    for i in range(n_people):
        min_cost += np.min(C)
        ind = np.unravel_index(np.argmin(C), C.shape)
        C[ind[0], :] = np.inf  # eliminate the person
        C[:, ind[1]] = np.inf  # eliminate the job
    return min_cost
```

```
n = 50  # fifty people and jobs
lam = 1.0
sample_size = 10000
EcostB = 0.0  # expectation of cost computed from algo A
for i in range(sample_size):
    C = expon.rvs(scale=1/lam, size=(n,n))
    EcostB += cost_algo_B(C)
EcostB /= sample_size
print('For algorithm B, the computed expectation of cost is: ', EcostB,
      '\n while the theoretical value is: ', np.sum(1/np.arange(1, n+1)))
```

Let $X_1, \dots, X_n$ be independent exponential random variables, with respective rates $\lambda_1, \dots, \lambda_n$. A useful result, generalizing Equation 9.1, is that X_i is the smallest of these with probability $\lambda_i / \sum_j \lambda_j$. This is shown as follows:

$$\begin{aligned} P\left\{X_i = \min_j X_j\right\} &= P\left\{X_i < \min_{j \neq i} X_j\right\} \\ &= \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \end{aligned}$$

where the final equality uses Equation 9.1 along with the fact that $\min_{j \neq i} X_j$ is exponential with rate $\sum_{j \neq i} \lambda_j$.

Another important fact is that $\min_i X_i$ and the rank ordering of the X_i are independent. To see why this is true, consider the conditional probability that $X_{i_1} < X_{i_2} < \dots < X_{i_n}$ given that the minimal value is greater than t. Because $\min_i X_i > t$ means that all the X_i are greater than t, it follows from the lack of memory property of exponential random variables that their remaining lives beyond t remain independent exponential random variables with their original rates. Consequently,

$$\begin{aligned} P\Big\{X_{i_1} < \dots < X_{i_n} \Big| \min_i X_i > t\Big\} &= P\Big\{X_{i_1} - t < \dots < X_{i_n} - t \Big| \min_i X_i > t\Big\} \\ &= P\{X_{i_1} < \dots < X_{i_n}\} \end{aligned}$$

which proves the result.

Example 9.2. Suppose you arrive at a post office having two clerks at a moment when both are busy but there is no one else waiting in line. You will enter service when either clerk becomes free. If service times for clerk i are exponential with rate $\lambda_i, i = 1, 2$, find $E[T]$, where T is the amount of time that you spend in the post office.

Solution 9.2. Let R_i denote the remaining service time of the customer with clerk $i, i = 1, 2$, and note, by the lack of memory property of exponentials, that R_1 and R_2 are independent exponential random variables with respective rates λ_1 and λ_2. Conditioning on which of R_1 or R_2 is the smallest yields

$$\begin{aligned} E[T] &= E[T|R_1 < R_2]P\{R_1 < R_2\} + E[T|R_2 \leqslant R_1]P\{R_2 \leqslant R_1\} \\ &= E[T|R_1 < R_2]\frac{\lambda_1}{\lambda_1 + \lambda_2} + E[T|R_2 \leqslant R_1]\frac{\lambda_2}{\lambda_1 + \lambda_2} \end{aligned}$$

Now, with S denoting your service time

$$\begin{aligned} E[T|R_1 < R_2] &= E[R_1 + S|R_1 < R_2] \\ &= E[R_1|R_1 < R_2] + E[S|R_1 < R_2] \\ &= E[R_1|R_1 < R_2] + \frac{1}{\lambda_1} \\ &= \frac{1}{\lambda_1 + \lambda_2} + \frac{1}{\lambda_1} \end{aligned}$$

The final equation used that conditional on $R_1 < R_2$ the random variable R_1 is the minimum of R_1 and R_2 and is thus exponential with rate $\lambda_1 + \lambda_2$; and also that conditional on $R_1 < R_2$ you are served by server 1.

As we can show in a similar fashion that

$$E[T|R_2 \leqslant R_1] = \frac{1}{\lambda_1 + \lambda_2} + \frac{1}{\lambda_2}$$

we obtain the result

$$E[T] = \frac{3}{\lambda_1 + \lambda_2}$$

Another way to obtain $E[T]$ is to write T as a sum, take expectations, and then condition where needed. This approach yields

$$\begin{aligned} E[T] &= E[\min(R_1, R_2) + S] \\ &= E[\min(R_1, R_2)] + E[S] \\ &= \frac{1}{\lambda_1 + \lambda_2} + E[S] \end{aligned}$$

To compute $E[S]$, we condition on which of R_1 and R_2 is smallest.

$$\begin{aligned}
E[S] &= E[S|R_1 < R_2]\frac{\lambda_1}{\lambda_1+\lambda_2} + E[S|R_2 \leqslant R_1]\frac{\lambda_2}{\lambda_1+\lambda_2} \\
&= \frac{2}{\lambda_1+\lambda_2}
\end{aligned}$$

10 The Poisson Process

10.1 The Poisson Process

10.1.1 Counting Processes

A stochastic process $\{N(t), t \geq 0\}$ is said to be a counting process if $N(t)$ represents the total number of "events" that occur by time t. Some examples of counting processes are the following:

(a) If we let $N(t)$ equal the number of persons who enter a particular store at or prior to time t, then $\{N(t), t \geq 0\}$ is a counting process in which an event corresponds to a person entering the store. Note that if we had let $N(t)$ equal the number of persons in the store at time t, then $\{N(t), t \geq 0\}$ would not be a counting process (why not?).

(b) If we say that an event occurs whenever a child is born, then $\{N(t), t \geq 0\}$ is a counting process when $N(t)$ equals the total number of people who were born by time t. [Does $N(t)$ include persons who have died by time t? Explain why it must.]

(c) If $N(t)$ equals the number of goals that a given soccer player scores by time t, then $\{N(t), t \geq 0\}$ is a counting process. An event of this process will occur whenever the soccer player scores a goal.

From its definition we see that for a counting process $N(t)$ must satisfy:

(i) $N(t) \geq 0$.

(ii) $N(t)$ is integer valued.

(iii) If $s < t$, then $N(s) \leq N(t)$

(iv) For $s < t$, $N(t) - N(s)$ equals the number of events that occur in the interval $(s, t]$.

A counting process is said to possess *independent increments* if the numbers of events that occur in disjoint time intervals are independent. For example, this means that the number of events that occur by time 10 [that is, $N(10)$] must be independent of the number of events that occur between times 10 and 15 [that is, $N(15) - N(10)$].

The assumption of independent increments might be reasonable for example (a), but it probably would be unreasonable for example (b). The reason for this is that if in example (b) $N(t)$

is very large, then it is probable that there are many people alive at time t; this would lead us to believe that the number of new births between time t and time $t + s$ would also tend to be large [that is, it does not seem reasonable that $N(t)$ is independent of $N(t + s) - N(t)$, and so $\{N(t), t \geq 0\}$ would not have independent increments in example (b)]. The assumption of independent increments in example (c) would be justified if we believed that the soccer player's chances of scoring a goal today do not depend on "how he's been going." It would not be justified if we believed in "hot streaks" or "slumps."

A counting process is said to possess *stationary increments* if the distribution of the number of events that occur in any interval of time depends only on the length of the time interval. In other words, the process has stationary increments if the number of events in the interval $(s, s + t)$ has the same distribution for all s.

The assumption of stationary increments would only be reasonable in example (a) if there were no times of day at which people were more likely to enter the store. Thus, for instance, if there was a rush hour (say, between 12 P.M. and 1 P.M.) each day, then the stationarity assumption would not be justified. If we believed that the earth's population is basically constant (a belief not held at present by most scientists), then the assumption of stationary increments might be reasonable in example (b). Stationary increments do not seem to be a reasonable assumption in example (c) since, for one thing, most people would agree that the soccer player would probably score more goals while in the age bracket 25–30 than he would while in the age bracket 35–40. It may, however, be reasonable over a smaller time horizon, such as one year.

10.1.2 Definition of the Poisson Process

One of the most important counting processes is the Poisson process which is defined as follows:

Definition 10.1. The counting process $\{N(t), t \geq 0\}$ is said to be a *Poisson process having rate* λ, $\lambda > 0$, if

(i) $N(0) = 0$

(ii) The process has independent increments.

(iii) The number of events in any interval of length t is Poisson distributed with mean λt. That is, for all $s, t \geq 0$

$$P\{N(t + s) - N(s) = n\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad n = 0, 1, ...$$

Note that it follows from condition (iii) that a Poisson process has stationary increments and also that

$$E[N(t)] = \lambda t$$

which explains why λ is called the rate of the process.

To determine if an arbitrary counting process is actually a Poisson process, we must show that conditions (i), (ii), and (iii) are satisfied. Condition (i), which simply states that the counting of events begins at time $t = 0$, and condition (ii) can usually be directly verified from our knowledge of the process. However, it is not at all clear how we would determine that condition (iii) is satisfied, and for this reason an equivalent definition of a Poisson process would be useful.

As a prelude to giving a second definition of a Poisson process we shall define the concept of a function $f(\cdot)$ being $o(h)$.

Definition 10.2. The function $f(\cdot)$ is said to be $o(h)$ if

$$\lim_{h\to 0} \frac{f(h)}{h} = 0$$

Example 10.1.

(i) The function $f(x) = x^2$ is $o(h)$ since

$$\lim_{h\to 0} \frac{f(h)}{h} = \lim_{h\to 0} \frac{h^2}{h} = \lim_{h\to 0} h = 0$$

(ii) The function $f(x) = x$ is not $o(h)$ since

$$\lim_{h\to 0} \frac{f(h)}{h} = \lim_{h\to 0} \frac{h}{h} = \lim_{h\to 0} 1 = 1 \neq 0$$

(iii) If $f(\cdot)$ is $o(h)$ and $g(\cdot)$ is $o(h)$, then so is $f(\cdot) + g(\cdot)$. This follows since

$$\lim_{h\to 0} \frac{f(h) + g(h)}{h} = \lim_{h\to 0} \frac{f(h)}{h} + \lim_{h\to 0} \frac{g(h)}{h} = 0 + 0 = 0$$

(iv) If $f(\cdot)$ is $o(h)$, then so is $g(\cdot) = cf(\cdot)$. This follows since

$$\lim_{h\to 0} \frac{cf(h)}{h} = c \lim_{h\to 0} \frac{f(h)}{h} = c \cdot 0 = 0$$

(v) From (iii) and (iv) it follows that any finite linear combination of functions, each of which is $o(h)$, is $o(h)$.

In order for the function $f(\cdot)$ to be $o(h)$ it is necessary that $f(h)/h$ go to zero as h goes to zero. But if h goes to zero, the only way for $f(h)/h$ to go to zero is for $f(h)$ to go to zero faster than h does. That is, for h small, $f(h)$ must be small compared with h.

We are now in a position to give an alternate definition of a Poisson process.

Definition 10.3. The counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process having rate $\lambda, \lambda > 0$, if

(i) $N(0) = 0$.

(ii) The process has stationary and independent increments.

(iii) $P\{N(h) = 1\} = \lambda h + o(h)$.

(iv) $P\{N(h) \geq 2\} = o(h)$.

Theorem 10.1. *Definition 25.1 and Definition 10.3 are equivalent.*

Proof. We show that Definition 10.3 implies Definition 25.1, and leave it to you to prove the reverse. To start, fix $u \geq 0$ and let

$$g(t) = E[\exp\{-uN(t)\}]$$

We derive a differential equation for $g(t)$ as follows:

$$\begin{aligned} g(t+h) &= E[\exp\{-uN(t+h)\}] \\ &= E[\exp\{-uN(t)\}\exp\{-u(N(t+h)-N(t))\}] \\ &= E[\exp\{-uN(t)\}]E[\exp\{-u(N(t+h)-N(t))\}] \quad \text{by independent increments} \\ &= g(t)E[\exp\{-uN(h)\}] \quad \text{by stationary increments} \end{aligned} \tag{10.1}$$

Now, assumptions (iii) and (iv) imply that

$$P\{N(h) = 0\} = 1 - \lambda h + o(h)$$

Hence, conditioning on whether $N(h) = 0$ or $N(h) = 1$ or $N(h) \geq 2$ yields

$$\begin{aligned} E[\exp\{-uN(h)\}] &= 1 - \lambda h + o(h) + e^{-u}(\lambda h + o(h)) + o(h) \\ &= 1 - \lambda h + e^{-u}\lambda h + o(h) \end{aligned} \tag{10.2}$$

Therefore, from Equation 10.1 and Equation 10.2 we obtain that

$$g(t+h) = g(t)(1 - \lambda h + e^{-u}\lambda h) + o(h)$$

implying that

$$\frac{g(t+h) - g(t)}{h} = g(t)\lambda(e^{-u} - 1) + \frac{o(h)}{h}$$

Letting $h \to 0$ gives

$$g'(t) = g(t)\lambda(e^{-u} - 1)$$

or, equivalently,

$$\frac{g'(t)}{g(t)} = \lambda(e^{-u} - 1)$$

Integrating, and using $g(0) = 1$, shows that

$$\log g(t) = \lambda t(e^{-u} - 1)$$

or

$$g(t) = \exp\{\lambda t(e^{-u} - 1)\}$$

or

$$g(t) = \exp\{\lambda t(e^{-u} - 1)\}$$

That is, the Laplace transform of $N(t)$ evaluated at u is $e^{\lambda t(e^{-u}-1)}$. Since that is also the Laplace transform of a Poisson random variable with mean λt, the result follows from the fact that the distribution of a nonnegative random variable is uniquely determined by its Laplace transform. □

10.1.3 Interarrival and Waiting Time Distributions

Consider a Poisson process, and let us denote the time of the first event by T_1. Further, for $n > 1$, let T_n denote the elapsed time between the $(n-1)$st and the nth event. The sequence $\{T_n, n = 1, 2, ...\}$ is called the sequence of interarrival times. For instance, if $T_1 = 5$ and $T_2 = 10$, then the first event of the Poisson process would have occurred at time 5 and the second at time 15.

We shall now determine the distribution of the T_n. To do so, we first note that the event $\{T_1 > t\}$ takes place if and only if no events of the Poisson process occur in the interval $[0, t]$ and thus,

$$P\{T_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}$$

Hence, T_1 has an exponential distribution with mean $1/\lambda$.

(Recall: A continuous random variable X is said to have an exponential distribution with parameter λ, $\lambda > 0$, if its probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

or, equivalently, if its cdf is given by

$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

)

Now,

$$P\{T_2 > t\} = E[P\{T_2 > t|T_1]\}]$$

However,

$$\begin{aligned} P\{T_2 > t | T_1 = s\} &= P\{0 \text{ events in } (s, s+t] | T_1 = s\} \\ &= P\{0 \text{ events in } (s, s+t]\} \\ &= e^{-\lambda t} \end{aligned} \tag{10.3}$$

where the last two equations followed from independent and stationary increments. Therefore, from Equation 10.3 we conclude that T_2 is also an exponential random variable with mean $1/\lambda$ and, furthermore, that T_2 is independent of T_1. Repeating the same argument yields the following.

Proposition 10.1. *$T_n, n = 1, 2, ...,$ are independent identically distributed exponential random variables having mean $1/\lambda$.*

Remark 10.1. The proposition should not surprise us. The assumption of stationary and independent increments is basically equivalent to asserting that, at any point in time, the process probabilistically restarts itself. That is, the process from any point on is independent of all that has previously occurred (by independent increments), and also has the same distribution as the original process (by stationary increments). In other words, the process has no memory, and hence exponential interarrival times are to be expected.

Another quantity of interest is S_n, the arrival time of the nth event, also called the *waiting time* until the nth event. It is easily seen that

$$S_n = \sum_{i=1}^{n} T_i, \quad n \geq 1$$

and hence from Proposition 24.1 and a property related to the exponential distribution it follows that S_n has a gamma distribution with parameters n and λ. That is, the probability density of S_n is given by

$$f_{S_n}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t \geq 0 \tag{10.4}$$

Equation 10.4 may also be derived by noting that the nth event will occur prior to or at time t if and only if the number of events occurring by time t is at least n. That is,

$$N(t) \geq n \Leftrightarrow S_n \leq t$$

Hence,

$$F_{S_n}(t) = P\{S_n \leq t\} = P\{N(t) \geq n\} = \sum_{j=n}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!}$$

which, upon differentiation, yields

$$\begin{aligned} f_{S_n}(t) &= -\sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} + \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} + \sum_{j=n+1}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=n}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \end{aligned}$$

Example 10.2. Suppose that people immigrate into a territory at a Poisson rate $\lambda = 1$ per day.

(a) What is the expected time until the tenth immigrant arrives?

(b) What is the probability that the elapsed time between the tenth and the eleventh arrival exceeds two days?

Solution 10.1.

(a) $E[S_{10}] = 10/\lambda = 10$ days.

(b) $P\{T_{11} > 2\} = e^{-2\lambda} = e^{-2} \approx 0.133$.

Now we numerically derive the results using Python.

```python
import numpy as np
from scipy.stats import expon

def generate_Poisson_process(lam, n):
    """
    Simulate a Poisson process with Poisson rate lam until the nth event happens.

    input:
    lam: float, the Poisson rate
    n: int, the process ends when the nth event happens.

    output:
    T: a 1d array that contains all the interarrival times
    """

    T = expon.rvs(scale = 1/lam, size=n)
    return T
```

```
# Part (a)
sample_size = 10000  # sample size for calculating probabilities
n = 10  # 10th event happens
lam = 1.0
S10 = np.zeros(sample_size)  # S10 stores all the simulated S_{10}
for i in range(sample_size):
    S10[i] = np.sum(generate_Poisson_process(lam, n))
print('By simulation, the expected time until the tenth immigrant arrives is: ', S10.mean(),
      ' days, and the theoretical result is 10 days')
```

```
# Part (b)
sample_size = 10000  # sample size for calculating probabilities
n = 11  # 11th event happens
lam = 1.0
T11 = np.zeros(sample_size)  # T11 stores all the simulated T_{11}
for i in range(sample_size):
    T11[i] = generate_Poisson_process(lam, n)[-1]
P_T11_gt_2 = np.sum(T11>2) / sample_size
print('The simulated probability that the elapsed time between the 10th and 11th arrival exc
      P_T11_gt_2, ' days, and the theoretical result is ', np.exp(-2))
```

Proposition 24.1 also gives us another way of defining a Poisson process. Suppose we start with a sequence $\{T_n, n \geq 1\}$ of independent identically distributed exponential random variables each having mean $1/\lambda$. Now let us define a counting process by saying that the nth event of this process occurs at time

$$S_n \equiv T_1 + T_2 + \cdots + T_n$$

The resultant counting process $\{N(t), t \geq 0\}$ will be Poisson with rate λ.

Remark 10.2. Another way of obtaining the density function of S_n is to note that because S_n is the time of the nth event,

$$\begin{aligned} P\{t < S_n < t+h\} &= P\{N(t) = n-1, \text{ one event in } (t, t+h)\} + o(h) \\ &= P\{N(t) = n-1\}, P\{\text{one event in } (t, t+h)\} + o(h) \\ &= e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}[\lambda h + o(h)] + o(h) \\ &= \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}h + o(h) \end{aligned}$$

where the first equality uses the fact that the probability of 2 or more events in $(t, t+h)$ is $o(h)$. If we now divide both sides of the preceding equation by h and then let $h \to 0$, we obtain

$$f_{S_n}(t) = \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}$$

11 Further Properties of Poisson Processes

11.1 Further Properties of Poisson Processes

Consider a Poisson process $\{N(t), t \geq 0\}$ having rate λ, and suppose that each time an event occurs it is classified as either a type I or a type II event. Suppose further that each event is classified as a type I event with probability p or a type II event with probability $1-p$, independently of all other events. For example, suppose that customers arrive at a store in accordance with a Poisson process having rate λ; and suppose that each arrival is male with probability $\frac{1}{2}$ and female with probability $\frac{1}{2}$. Then a type I event would correspond to a male arrival and a type II event to a female arrival.

Let $N_1(t)$ and $N_2(t)$ denote respectively the number of type I and type II events occurring in $[0,t]$. Note that $N(t) = N_1(t) + N_2(t)$.

Proposition 11.1. *$\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are both Poisson processes having respective rates λp and $\lambda(1-p)$. Furthermore, the two processes are independent.*

Proof. It is easy to verify that $\{N_1(t), t \geq 0\}$ is a Poisson process with rate λp by verifying that it satisfies Definition 3 in Lecture 7.

- $N_1(0) = 0$ follows from the fact that $N(0) = 0$.
- It is easy to see that $\{N_1(t), t \geq 0\}$ inherits the stationary and independent increment properties of the process $\{N(t), t \geq 0\}$. This is true because the distribution of the number of type I events in an interval can be obtained by conditioning on the number of events in that interval, and the distribution of this latter quantity depends only on the length of the interval and is independent of what has occurred in any nonoverlapping interval.
- $$\begin{aligned} P\{N_1(h) = 1\} &= P\{N_1(h) = 1 | N(h) = 1\} P\{N(h) = 1\} \\ &\quad + P\{N_1(h) = 1 | N(h) \geq 2\} P\{N(h) \geq 2\} \\ &= p(\lambda h + o(h) + o(h)) \\ &= \lambda p h + o(h) \end{aligned}$$
- $P\{N_1(h) \geq 2\} \leq P\{N(h) \geq 2\} = o(h)$

Thus we see that $\{N_1(t), t \geq 0\}$ is a Poisson process with rate λp and, by a similar argument, that $\{N_2(t), t \geq 0\}$ is a Poisson process with rate $\lambda(1-p)$. Because the probability of a type I event in the interval from t to $t+h$ is independent of all that occurs in intervals that do not overlap $(t, t+h)$, it is independent of knowledge of when type II events occur, showing that the two Poisson processes are independent. □

Example 11.1. If immigrants to area A arrive at a Poisson rate of ten per week, and if each immigrant is of English descent with probability $\frac{1}{12}$, then what is the probability that no people of English descent will emigrate to area A during the month of February?

Solution 11.1. By the previous proposition it follows that the number of Englishmen emigrating to area A during the month of February is Poisson distributed with mean $4(10)(\frac{1}{12}) = \frac{10}{3}$. Hence the desired probability is $e^{-10/3}$

```
import numpy as np
from scipy.stats import poisson

def generate_immigrating_process(lam, n, p):
    """
    Simulate a immigrating Poisson process with Poisson rate lam for n periods. The chance o

    input:
    lam: float, the Poisson rate
    n: int, the number of periods to simulate.
    p: float, probability for an immigrant to be English

    output:
    E: int, the number of English immigrants during n periods.
    """

    E = 0
    for i in range(n):  # simulate n periods
        n_of_immigrants =  poisson.rvs(mu=lam)
        E += np.sum(np.random.random(n_of_immigrants) < p )

    return E
```

```
lam = 10
n = 4  # 4 periods in February
p = 1/12
sample_size = 100000
```

```
E = np.zeros(sample_size, dtype = 'int')
for i in range(sample_size):
    E[i] = generate_immigrating_process(lam, n, p)
prob = np.sum(E == 0) / sample_size
print('By simulation, no people of English descent will emigrate to area A during the month
      prob, ' \n and the theoretical probability is: ', np.exp(-10/3))
```

Example 11.2. Suppose nonnegative offers to buy an item that you want to sell arrive according to a Poisson process with rate λ. Assume that each offer is the value of a continuous random variable having density function $f(x)$. Once the offer is presented to you, you must either accept it or reject it and wait for the next offer. We suppose that you incur costs at a rate c per unit time until the item is sold, and that your objective is to maximize your expected total return, where the total return is equal to the amount received minus the total cost incurred. Suppose you employ the policy of accepting the first offer that is greater than some specified value y. (Such a type of policy, which we call a y-policy, can be shown to be optimal.) What is the best value of y?

Solution 11.2. Let us compute the expected total return when you use the y-policy, and then choose the value of y that maximizes this quantity. Let X denote the value of a random offer, and let $\bar{F}(x) = P\{X > x\} = \int_x^\infty f(u)du$ be its tail distribution function. Because each offer will be greater than y with probability $\bar{F}(y)$, it follows that such offers occur according to a Poisson process with rate $\lambda\bar{F}(y)$. Hence, the time until an offer is accepted is an exponential random variable with rate $\lambda\bar{F}(y)$. Letting $R(y)$ denote the total return from the policy that accepts the first offer that is greater than y, we have

$$\begin{aligned}
E[R(y)] &= E[\text{accepted offer}] - cE[\text{time to accept}] \\
&= E[X|X > y] - \frac{c}{\lambda\bar{F}(y)} \\
&= \int_0^\infty x f_{X|X>y}(x)\, dx - \frac{c}{\lambda\bar{F}(y)} \\
&= \int_y^\infty x \frac{f(x)}{\bar{F}(y)}\, dx - \frac{c}{\lambda\bar{F}(y)} \\
&= \frac{\int_y^\infty x f(x)\, dx - c/\lambda}{\bar{F}(y)}
\end{aligned}$$

Differentiation yields that

$$\frac{d}{dy}E[R(y)] = 0 \Leftrightarrow -\bar{F}(y)yf(y) + \left(\int_y^\infty x f(x)\, dx - \frac{c}{\lambda}\right) f(y) = 0$$

Therefore, the optimal value of y satisfies

$$y\bar{F}(y) = \int_y^\infty x f(x)\, dx - \frac{c}{\lambda}$$

or

$$y \int_y^\infty f(x)\, dx = \int_y^\infty x f(x)\, dx - \frac{c}{\lambda}$$

or

$$\int_y^\infty (x - y) f(x)\, dx = \frac{c}{\lambda} \tag{11.1}$$

We now argue that the left-hand side of the preceding is a nonincreasing function of y. To do so, note that, with a^+ defined to equal a if $a > 0$ or to equal 0 otherwise, we have

$$\int_y^\infty (x - y) f(x)\, dx = E[(X - y)^+]$$

Because $(X - y)^+$ is a nonincreasing function of y, so is its expectation, thus showing that the left hand side of Equation 11.1 is a nonincreasing function of y. Consequently, if $E[X] < c/\lambda$—in which case there is no solution of Equation 11.1—then it is optimal to accept any offer; otherwise, the optimal value y is the unique solution of Equation 11.1.

It follows from Proposition 24.1 that if each of a Poisson number of individuals is independently classified into one of two possible groups with respective probabilities p and $1 - p$, then the number of individuals in each of the two groups will be independent Poisson random variables. Because this result easily generalizes to the case where the classification is into any one of r possible groups, we have the following application to a model of employees moving about in an organization.

Example 11.3. Consider a system in which individuals at any time are classified as being in one of r possible states, and assume that an individual changes states in accordance with a Markov chain having transition probabilities $P_{ij}, i, j = 1, \dots, r$. That is, if an individual is in state i during a time period then, independently of its previous states, it will be in state j during the next time period with probability P_{ij}. The individuals are assumed to move through the system independently of each other. Suppose that the numbers of people initially in states $1, 2, \dots, r$ are independent Poisson random variables with respective means $\lambda_1, \lambda_2, \dots, \lambda_r$. We are interested in determining the joint distribution of the numbers of individuals in states $1, 2, \dots, r$ at some time n.

Solution 11.3. For fixed i, let $N_j(i), j = 1, \dots, r$ denote the number of those individuals, initially in state i, that are in state j at time n. Now each of the (Poisson distributed) number of people initially in state i will, independently of each other, be in state j at time n with probability `$P^n_{ij}` , where P_{ij}^n is the n-stage transition probability for the Markov chain having transition probabilities P_{ij} . Hence, the $N_j(i), j = 1, \dots, r$ will be independent Poisson random variables with respective means $\lambda_i P^n_{ij}, j = 1, \dots, r$. Because the sum of independent Poisson random variables is itself a Poisson random variable, it follows that the number of individuals in state j at time n—namely $\sum_{i=1}^{r} N_j(i)$—will be independent Poisson random variables with respective means $\sum_i \lambda_i P^n_{ij}$, for $j = 1, \dots, r$.

(The Coupon Collecting Problem) There are m different types of coupons. Each time a person collects a coupon it is, independently of ones previously obtained, a type j coupon with probability $p_j, \sum_{j=1}^{m} p_j = 1$. Let N denote the number of coupons one needs to collect in order to have a complete collection of at least one of each type. Find $E[N]$.

Solution 11.4. If we let N_j denote the number one must collect to obtain a type j coupon, then we can express N as

$$N = \max_{1 \le j \le m} N_j$$

However, even though each N_j is geometric with parameter p_j, the foregoing representation of N is not that useful, because the random variables N_j are not independent.

We can, however, transform the problem into one of determining the expected value of the maximum of independent random variables. To do so, suppose that coupons are collected at times chosen according to a Poisson process with rate $\lambda = 1$. Say that an event of this Poisson process is of type j, $1 \le j \le m$, if the coupon obtained at that time is a type j coupon. If we now let $N_j(t)$ denote the number of type j coupons collected by time t , then it follows from Proposition 24.1 that $\{N_j(t), t \ge 0\}$, $j = 1, \dots, m$ are independent Poisson processes with respective rates $\lambda p_j = p_j$. Let X_j denote the time of the first event of the jth process, and let

$$X = \max_{1 \leqslant j \leqslant m} X_j$$

denote the time at which a complete collection is amassed. Since the X_j are independent exponential random variables with respective rates p_j, it follows that

$$\begin{aligned} P\{X < t\} &= P\{\max X_j < t\} \\ &= P\{X_j < t, \text{ for } j = 1, \dots, m\} \\ &= \prod_{j=1}^{m} (1 - e^{-p_j t}) \end{aligned}$$

Therefore,

$$\begin{aligned} E[X] &= \int_0^\infty P\{X > t\}\, dt \\ &= \int_0^\infty \left\{ 1 - \prod_{j=1}^{m} (1 - e^{-p_j t}) \right\} dt \end{aligned} \tag{11.2}$$

It remains to relate $E[X]$, the expected time until one has a complete set, to $E[N]$, the expected number of coupons it takes. This can be done by letting T_i denote the ith interarrival time of the Poisson process that counts the number of coupons obtained. Then it is easy to see that

$$X = \sum_{i=1}^{N} T_i$$

Since the T_i are independent exponentials with rate 1, and N is independent of the T_i, we see that

$$E[X|N] = NE[T_i] = N$$

Therefore,

$$E[X] = E[N]$$

and so $E[N]$ is as given in Equation 11.2.

Let us now compute the expected number of types that appear only once in the complete collection. Letting I_i equal l if there is only a single type i coupon in the final set, and letting it equal 0 otherwise, we thus want

$$\begin{aligned} E\left[\sum_{i=1}^{m} I_i\right] &= \sum_{i=1}^{m} E[I_i] \\ &= \sum_{i=1}^{m} P\{I_i = 1\} \end{aligned}$$

Now there will be a single type i coupon in the final set if a coupon of each type has appeared before the second coupon of type i is obtained. Thus, letting S_i denote the time at which the second type i coupon is obtained, we have

$$P\{I_i = 1\} = P\{X_j < S_i, \text{ for all } j \neq i\}$$

Using that S_i has a gamma distribution with parameters (2, pi), this yields

$$\begin{aligned} P\{I_i = 1\} &= \int_0^\infty P\{X_j < S_i \text{ for all } j \neq i | S_i = x\} p_i e^{-p_i x}\, p_i x \, dx \\ &= \int_0^\infty P\{X_j < x, \text{ for all } j \neq i\} p_i^2 x\, e^{-p_i x} \, dx \\ &= \int_0^\infty \prod_{j \neq i} (1 - e^{-p_j x})\, p_i^2 x e^{-p_i x} \, dx \end{aligned}$$

Therefore, we have the result

$$\begin{aligned} E\left[\sum_{i=1}^{m} I_i\right] &= \int_0^\infty \sum_{i=1}^{m} \prod_{j \neq i} (1 - e^{-p_j x}) p_i^2 x e^{-p_i x} \, dx \\ &= \int_0^\infty x \prod_{j=1}^{m} (1 - e^{-p_j x}) \sum_{i=1}^{m} p_i^2 \frac{e^{-p_i x}}{1 - e^{-p_i x}} \, dx \quad \blacksquare \end{aligned}$$

12 Conditional Distribution of the Arrival Times

12.1 Conditional Distribution of the Arrival Times

Suppose we are told that exactly one event of a Poisson process has taken place by time t, and we are asked to determine the distribution of the time at which the event occurred. Now, since a Poisson process possesses stationary and independent increments it seems reasonable that each interval in $[0, t]$ of equal length should have the same probability of containing the event. In other words, the time of the event should be uniformly distributed over $[0, t]$. This is easily checked since, for $s \leq t$,

$$\begin{aligned}
P\{T_1 < s|N(t) = 1\} &= \frac{P\{T_1 < s, N(t) = 1\}}{P\{N(t) = 1\}} \\
&= \frac{P\{1 \text{ event in } [0, s), 0 \text{ events in } [s, t]\}}{P\{N(t) = 1\}} \\
&= \frac{P\{1 \text{ event in } [0, s)\}P\{0 \text{ events in } [s, t]\}}{P\{N(t) = 1\}} \\
&= \frac{\lambda s e^{-\lambda s} e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\
&= \frac{s}{t}
\end{aligned}$$

This result may be generalized, but before doing so we need to introduce the concept of order statistics.

Let $Y_1, Y_2, \dots, Y_n$ be n random variables. We say that $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ are the order statistics corresponding to $Y_1, Y_2, \dots, Y_n$ if $Y_{(k)}$ is the kth smallest value among $Y_1, \dots, Y_n, k = 1, 2, \dots, n$. For instance if $n = 3$ and $Y_1 = 4$, $Y_2 = 5$, $Y_3 = 1$ then $Y_{(1)} = 1, Y_{(2)} = 4, Y_{(3)} = 5$. If the $Y_i, i = 1, \dots, n$,are independent identically distributed continuous random variables with probability density f, then the joint density of the order statistics $Y_{(1)}, Y_{(2)}, \dots, \dot{Y}_{(n)}$ is given by

$$f(y_1, y_2, \dots, y_n) = n! \prod_{i=1}^{n} f(y_i), \quad y_1 < y_2 < \dots < y_n$$

The preceding follows since

(i) $(Y_{(1)}, Y_{(2)}, ..., Y_{(n)})$ will equal $(y_1, y_2, ..., y_n)$ if $(Y_1, Y_2, ..., Y_n)$ is equal to any of the $n!$ permutations of $(y_1, y_2, \dots, y_n)$;

and

(ii) the probability density that $(Y_1, Y_2, \dots, Y_n)$ is equal to $y_{i_1}, \dots, y_{i_n}$ is $\prod_{j=1}^{n} f(y_{i_j}) = \prod_{j=1}^{n} f(y_j)$ when $i_1, \dots, i_n$ is a permutation of $1, 2, \dots, n$

If the $Y_i, i = 1, \dots, n$, are uniformly distributed over $(0, t)$, then we obtain from the preceding that the joint density function of the order statistics $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ is

$$f(y_1, y_2, \dots, y_n) = \frac{n!}{t^n}, \quad 0 < y_1 < y_2 < \dots < y_n < t$$

We are now ready for the following useful theorem.

Theorem 12.1. *Given that $N(t) = n$, the n arrival times $S_1, \dots, S_n$ have the same distribution as the order statistics corresponding to n independent random variables uniformly distributed on the interval* $(0, t)$.

Proof. To obtain the conditional density of $S_1, \dots, S_n$ given that $N(t) = n$ note that for $0 < S_1 < \dots < S_n < t$ the event that $S_1 = s_1, S_2 = s_2, ...,$ $S_n = s_n, N(t) = n$ is equivalent to the event that the first $n + 1$ interarrival times satisfy $T_1 = s_1, T_2 = s_2 - s_1, \dots, T_n = s_n - s_{n-1}, T_{n+1} > t - s_n$. Hence, using Proposition 1 in Lecture 7, we have that the conditional joint density of $S_1, \dots, S_n$ given that $N(t) = n$ is as follows:

$$\begin{aligned} f(s_1, \dots, s_n | n) &= \frac{f(s_1, \dots, s_n, n)}{P\{N(t) = n\}} \\ &= \frac{\lambda e^{-\lambda s_1} \lambda e^{-\lambda(s_2 - s_1)} \dots \lambda e^{-\lambda(s_n - s_{n-1})} e^{-\lambda(t - s_n)}}{e^{-\lambda t}((\lambda t)^n / n!)} \\ &= \frac{n!}{t^n}, \quad 0 < s_1 < \dots < s_n < t \end{aligned}$$

which proves the result. □

Remark 12.1. The preceding result is usually paraphrased as stating that, under the condition that n events have occurred in $(0, t)$, the times $S_1, \dots, S_n$ at which events occur, considered as unordered random variables, are distributed independently and uniformly in the interval $(0, t)$.

Application of Theorem 23.1 (Sampling a Poisson Process) In Proposition 1 Lecture 8 we showed that if each event of a Poisson process is independently classified as a type I event with probability p and as a type II event with probability $1 - p$ then the counting processes of type I and type II events are independent Poisson processes with respective rates

λp and $\lambda(1-p)$. Suppose now, however, that there are k possible types of events and that the probability that an event is classificd as a type i event, $i = 1, \dots, k$, depends on the time the event occurs. Specifically, suppose that if an event occurs at time y then it will be classified as a type i event, independently of anything that has previously occurred, with probability $P_i(y), i = 1, \dots, k$ where $\sum_{i=1}^{k} P_i(y) = 1$. Upon using Theorem 23.1 we can prove the following useful proposition.

Proposition 12.1. *If $N_i(t), i = 1, \dots, k$, represents the number of type i events occurring by time t then $N_i(t), i = 1, \dots, k$, are independent Poisson random variables having means*

$$E[N_i(t)] = \lambda \int_0^t P_i(s)\, ds$$

Before proving this proposition, let us first illustrate its use.

Example 12.1. (An Infinite Server Queue) Suppose that customers arrive at a service station in accordance with a Poisson process with rate λ. Upon arrival the customer is immediately served by one of an infinite number of possible servers, and the service times are assumed to be independent with a common distribution G. What is the distribution of $X(t)$, the number of customers that have completed service by time t? What is the distribution of $Y(t)$, the number of customers that are being served at time t?

To answer the preceding questions let us agree to call an entering customer a type I customer if he completes his service by time t and a type II customer if he does not complete his service by time t. Now, if the customer enters at time s, $s \leqslant t$, then he will be a type I customer if his service time is less than $t - s$. Since the service time distribution is G, the probability of this will be $G(t-s)$. Similarly, a customer entering at time s, $s \leqslant t$, will be a type II customer with probability $\bar{G}(t-s) = 1 - G(t-s)$. Hence, from Proposition 24.1 we obtain that the distribution of $X(t)$, the number of customers that have completed service by time t, is Poisson distributed with mean

$$E[X(t)] = \lambda \int_0^t G(t-s)\, ds = \lambda \int_0^t G(y)\, dy \tag{12.1}$$

Similarly, the distribution of $Y(t)$, the number of customers being served at time t is Poisson with mean

$$E[Y(t)] = \lambda \int_0^t \bar{G}(t-s)\, ds = \lambda \int_0^t \bar{G}(y)\, dy \tag{12.2}$$

Furthermore, $X(t)$ and $Y(t)$ are independent.

Suppose now that we are interested in computing the joint distribution of $Y(t)$ and $Y(t+s)$—that is, the joint distribution of the number in the system at time t and at time $t + s$. To accomplish this, say that an arrival is type l: if he arrives before time t and completes service between t and $t+s$,

type 2: if he arrives before t and completes service after $t+s$,

type 3: if he arrives between t and $t+s$ and completes service after $t+s$,

type 4: otherwise.

Hence an arrival at time y will be type i with probability $P_i(y)$ given by

$$P_1(y) = \begin{cases} G(t+s-y) - G(t-y), & \text{if } y < t \\ 0, & \text{otherwise} \end{cases}$$
$$P_2(y) = \begin{cases} \tilde{G}(t+s-y), & \text{if } y < t \\ 0, & \text{otherwise} \end{cases}$$
$$P_3(y) = \begin{cases} \bar{G}(t+s-y), & \text{if } t < y < t+s \\ 0, & \text{otherwise} \end{cases}$$
$$P_4(y) = 1 - P_1(y) - P_2(y) - P_3(y)$$

Hence, if $N_i = N_i(s+t), i = 1, 2, 3$, denotes the number of type i events that occur, then from Proposition 24.1, $N_i, i = 1, 2, 3$, are independent Poisson random variables with respective means

$$E[N_i] = \lambda \int_0^{t+s} P_i(y)\, dy, \quad i = 1, 2, 3$$

Because

$$\begin{aligned} Y(t) &= N_1 + N_2, \\ Y(t+s) &= N_2 + N_3 \end{aligned}$$

it is now an easy matter to compute the joint distribution of $Y(t)$ and $Y(t+s)$. For instance,

$$\begin{aligned} &\mathrm{Cov}[Y(t), Y(t+s)] \\ &= \mathrm{Cov}(N_1 + N_2, N_2 + N_3) \\ &= \mathrm{Cov}(N_2, N_2) \quad \text{by independence of } N_1, N_2, N_3 \\ &= \mathrm{Var}(N_2) \\ &= \lambda \int_0^t \bar{G}(t+s-y)\, dy = \lambda \int_0^t \bar{G}(u+s)\, du \end{aligned}$$

where the last equality follows since the variance of a Poisson random variable equals its mean, and from the substitution $u = t - y$. Also, the joint distribution of $Y(t)$ and $Y(t+s)$ is as follows:

$$\begin{aligned}P\{Y(t)=i,Y(t+s)=j\} &= P\{N_1+N_2=i,N_2+N_3=j\}\\ &= \sum_{l=0}^{\min(i,j)} P\{N_2=l,N_1=i-l,N_3=j-l\}\\ &= \sum_{l=0}^{\min(i,j)} P\{N_2=l\}P\{N_1=i-l\}P\{N_3=j-l\} \quad \blacksquare\end{aligned}$$

Example 12.2. (Tracking the Number of HIV Infections) There is a relatively long incubation period from the time when an individual becomes infected with the HIV virus, which causes AIDS, until the symptoms of the disease appear.

As a result, it is difficult for public health officials to be certain of the number of members of the population that are infected at any given time. We will now present a first approximation model for this phenomenon, which can be used to obtain a rough estimate of the number of infected individuals.

Let us suppose that individuals contract the HIV virus in accordance with a Poisson process whose rate λ is unknown. Suppose that the time from when an individual becomes infected until symptoms of the disease appear is a random variable having a known distribution G. Suppose also that the incubation times of different infected individuals are independent.

Let $N_1(t)$ denote the number of individuals who have shown symptoms of the disease by time t. Also, let $N_2(t)$ denote the number who are HIV positive but have not yet shown any symptoms by time t. Now, since an individual who contracts the virus at time s will have symptoms by time t with probability $G(t-s)$ and will not with probability $\bar{G}(t-s)$, it follows from Proposition 24.1 that $N_1(t)$ and $N_2(t)$ are independent Poisson random variables with respective means

$$E[N_1(t)] = \lambda\int_0^t G(t-s)ds = \lambda\int_0^t G(y)dy$$

and

$$E[N_2(t)] = \lambda\int_0^t \bar{G}(t-s)ds = \lambda\int_0^t \bar{G}(y)dy$$

Now, if we knew λ, then we could use it to estimate N2(t), the number of individuals infected but without any outward symptoms at time t , by its mean value E[N2(t)]. However, since is unknown, we must first estimate it. Now, we will presumably know the value of N1(t), and so we can use its known value as an estimate of its mean E[N1(t)]. That is, if the number of individuals who have exhibited symptoms by time t is n1, then we can estimate that

$$n_1 \approx E[N_1(t)] = \lambda\int_0^t G(y)dy$$

Therefore, we can estimate λ by the quantity $\hat{\lambda}$ given by

$$\hat{\lambda} = n_1 / \int_0^t G(y)dy$$

Using this estimate of λ, we can estimate the number of infected but symptomless individuals at time t by

$$\begin{aligned}\text{estimate of } N_2(t) &= \hat{\lambda} \int_0^t \hat{G}(y)dy \\ &= \frac{n_1 \int_0^t \hat{G}(y)dy}{\int_0^t G(y)dy}\end{aligned}$$

For example, suppose that G is exponential with mean μ. Then $\bar{G}(y) = e^{-}y/\mu$ and a simple integration gives that

$$\text{estimate of } N_2(t) = \frac{n_1\mu(1 - e^{-t/\mu})}{t - \mu(1 - e^{-t/\mu})}$$

If we suppose that $t = 16$ years, $\mu = 10$ years, and $n_1 = 220$ thousand, then the estimate of the number of infected but symptomless individuals at time 16 is

$$\text{estimate} = \frac{2,200(1 - e^{-1.6})}{16 - 10(1 - e^{-1.6})} = 218.96$$

That is, if we suppose that the foregoing model is approximately correct (and we should be aware that the assumption of a constant infection rate λ that is unchanging over time is almost certainly a weak point of the model), then if the incubation period is exponential with mean 10 years and if the total number of individuals who have exhibited AIDS symptoms during the first 16 years of the epidemic is 220 thousand, then we can expect that approximately 219 thousand individuals are HIV positive though symptomless at time 16.

Proof of Proposition 5.3

Let us compute the joint probability $P\{N_i(t) = n_i, i = 1, \dots, k\}$. To do so note first that in order for there to have been n_i type i events for $i = 1, \dots, k$ there must have been a total of $\sum_i = 1^k n_i$ events. Hence, conditioning on $N(t)$ yields

$$\begin{aligned}&P\{N_1(t) = n_1, \dots, N_k(t) = n_k\} \\ &= P\left\{ N_1(t) = n_1, \dots, N_k(t) = n_k \,\middle|\, N(t) = \sum_{i=1}^k n_i \right\} \\ &\times P\left\{ N(t) = \sum_{i=1}^k n_i \right\}\end{aligned}$$

Now consider an arbitrary event that occurred in the interval $[0, t]$. If it had occurred at time s,then the probability that it would be a type i event would be $P_i(s)$. Hence, since by

Theorem 23.1 this event will have occurred at some time uniformly distributed on $(0, t)$, it follows that the probability that this event will be a type i event is

$$P_i = \frac{1}{t}\int_0^t P_i(s)ds$$

independently of the other events. Hence,

$$P\left\{N_i(t) = n_i, i = 1, \dots, k | N(t) = \sum_{i=1}^{k} n_i\right\}$$

will just equal the multinomial probability of n_i type i outcomes for $i = 1, \dots, k$ when each of $\sum_{i=1}^{k} n_i$ independent trials results in outcome i with probability $P_i, i = 1, \dots, k$. That is,

$$P\left\{N_1(t) = n_1, \dots, N_k(t) = n_k | N(t) = \sum_{i=1}^{k} n_i\right\} = \frac{(\sum_{i=1}^{k} n_i)!}{n_1! \cdots n_k!} P_1^{n_1} \cdots P_k^{n_k}$$

Consequently,

$$\begin{aligned} P\{N_1(t) = n_1, \dots, N_k(t) = n_k\} &= \frac{(\sum_{i=1}^{k} n_i)!}{n_1! \cdots n_k!} P_1^{n_1} \cdots P_k^{n_k} e^{-\lambda t} \frac{(\lambda t) \sum_i n_i}{(\sum_i n_i)!} \\ &= \prod_{i=1}^{k} e^{-\lambda t} P_i (\lambda t P_i)^{n_i} / n_i! \end{aligned}$$

and the proof is complete

We now present an additional example of the usefulness of Theorem 23.1.

Example 12.3. Insurance claims are made at times distributed according to a Poisson process with rate λ; the successive claim amounts are independent random variables having distribution G with mean μ, and are independent of the claim arrival times. Let S_i and C_i denote, respectively, the time and the amount of the ith claim. The total discounted cost of all claims made up to time t , call it $D(t)$, is defined by

$$D(t) = \sum_{i=1}^{N(t)} e^{-\alpha S_i} C_i$$

where α is the discount rate and $N(t)$ is the number of claims made by time t. To determine the expected value of $D(t)$, we condition on $N(t)$ to obtain

$$E[D(t)] = \sum_{n=0}^{\infty} E[D(t) | N(t) = n] e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

Now, conditional on $N(t) = n$ the claim arrival times $S_1, \ldots, S_n$ are distributed as the ordered values $U_{(1)}, \ldots, U_{(n)}$ of n independent uniform $(0, t)$ random variables $U_1, \ldots, U_n$. Therefore,

$$\begin{aligned} E[D(t)|N(t) = n] &= E\left[\sum_{i=1}^{n} C_i e^{-\alpha U_{(i)}}\right] \\ &= \sum_{i=1}^{n} E\left[C_i e^{-\alpha U_{(i)}}\right] \\ &= \sum_{i=1}^{n} E\left[C_i\right] E\left[e^{-\alpha U_{(i)}}\right] \end{aligned}$$

where the final equality used the independence of the claim amounts and their arrival times. Because $E[C_i] = \mu$, continuing the preceding gives

$$\begin{aligned} E[D(t)|N(t) = n] &= \mu \sum_{i=1}^{n} E\left[e^{-\alpha U_{(i)}}\right] \\ &= \mu E\left[\sum_{i=1}^{n} e^{-\alpha U_{(i)}}\right] \\ &= \mu E\left[\sum_{i=1}^{n} e^{-\alpha U_i}\right] \end{aligned}$$

The last equality follows because $U_{(1)}, \ldots, U_{(n)}$ are the values $U_1, \ldots, U_n$ in increasing order, and so $\sum_{i=1}^{n} e^{-\alpha U_{(i)}} = \sum_{i=1}^{n} e^{-\alpha U_i}$. Continuing the string of equalities yields

$$\begin{aligned} E[D(t)|N(t) = n] &= n\mu E\left[e^{-\alpha U}\right] \\ &= n\frac{\mu}{t}\int_0^t e^{-\alpha x} dx \\ &= n\frac{\mu}{\alpha t}(1 - e^{-\alpha t}) \end{aligned}$$

Therefore,

$$E[D(t)|N(t)] = N(t)\frac{\mu}{\alpha t}(1 - e^{-\alpha t})$$

Taking expectations yields the result

$$E[D(t)] = \frac{\lambda\mu}{\alpha}(1 - e^{-\alpha t})$$

13 Introduction to Continuous-Time Markov Chains

13.1 Introduction

We consider a class of probability models that has a wide variety of applications in the real world. The members of this class are the continuous-time analogs of the discrete-time Markov chains and as such are characterized by the Markovian property that, given the present state, the future is independent of the past.

One example of a continuous-time Markov chain has already been met. This is the Poisson process. For if we let the total number of arrivals by time t [that is, $N(t)$] be the state of the process at time t, then the Poisson process is a continuous-time Markov chain having states $0, 1, 2, ...$ that always proceeds from state n to state $n+1$, where $n \geq 0$. Such a process is known as a pure *birth process* since when a transition occurs the state of the system is always increased by one. More generally, an exponential model which can go (in one transition) only from state n to either state $n-1$ or state $n+1$ is called a *birth and death model*. For such a model, transitions from state n to state $n+1$ are designated as births, and those from n to $n-1$ as deaths. Birth and death models have wide applicability in the study of biological systems and in the study of waiting line systems in which the state represents the number of customers in the system. These models will be studied extensively.

Next, we define continuous-time Markov chains and then relate them to the discrete-time Markov chains.

13.2 Continuous-Time Markov Chains

Suppose we have a continuous-time stochastic process $\{X(t), t \geq 0\}$ taking on values in the set of nonnegative integers. In analogy with the definition of a discrete-time Markov chain, we say that the process $\{X(t), t \geq 0\}$ is a continuous-time Markov chain if for all $s, t \geq 0$ and nonnegative integers $i, j, x(u)$, $0 \leq u < s$

$$\begin{aligned} P\{X(t+s) = j | X(s) = i, X(u) = x(u), 0 \leq u < s\} \\ = P\{X(t+s) = j | X(s) = i\} \end{aligned}$$

In other words, a continuous-timeMarkov chain is a stochastic process having the Markovian property that the conditional distribution of the future $X(t+s)$ given the present $X(s)$ and the past $X(u), 0 \leq u < s$, depends only on the present and is independent of the past. If, in addition,

$$P\{X(t+s) = j | X(s) = i\}$$

is independent of s, then the continuous-time Markov chain is said to have *stationary* or *homogeneous* transition probabilities.

All Markov chains considered in this course will be assumed to have stationary transition probabilities.

Suppose that a continuous-time Markov chain enters state i at some time, say, time 0, and suppose that the process does not leave state i (that is, a transition does not occur) during the next ten minutes. What is the probability that the process will not leave state i during the following five minutes? Now since the process is in state i at time 10 it follows, by the Markovian property, that the probability that it remains in that state during the interval $[10, 15]$ is just the (unconditional) probability that it stays in state i for at least five minutes. That is, if we let T_i denote the amount of time that the process stays in state i before making a transition into a different state, then

$$P\{T_i > 15 | T_i > 10\} = P\{T_i > 5\}$$

or, in general, by the same reasoning,

$$P\{T_i > s+t | T_i > s\} = P\{T_i > t\}$$

for all $s, t \geq 0$. Hence, the random variable T_i is *memoryless* and must thus be exponentially distributed.

In fact, the preceding gives us another way of defining a continuous-time Markov chain. Namely, it is a stochastic process having the properties that each time it enters state i

(i) the amount of time it spends in that state before making a transition into a different state is exponentially distributed with mean, say, $1/v_i$, and

(ii) when the process leaves state i, it next enters state j with some probability, say, P_{ij}. Of course, the P_{ij} must satisfy

$$\begin{aligned} P_{ii} &= 0, \quad \text{all } i \\ \sum_j P_{ij} &= 1, \quad \text{all } i \end{aligned}$$

In other words, a continuous-time Markov chain is a stochastic process that moves from state to state in accordance with a (discrete-time) Markov chain, but is such that the amount of time it spends in each state, before proceeding to the next state, is exponentially distributed. In addition, the amount of time the process spends in state i, and the next state visited, must be independent random variables. For if the next state visited were dependent on T_i , then information as to how long the process has already been in state i would be relevant to the prediction of the next state—and this contradicts the Markovian assumption.

Example 13.1. (A Shoeshine Shop) Consider a shoeshine establishment consisting of two chairs—chair 1 and chair 2. A customer upon arrival goes initially to chair 1 where his shoes are cleaned and polish is applied. After this is done the customer moves on to chair 2 where the polish is buffed. The service times at the two chairs are assumed to be independent random variables that are exponentially distributed with respective rates μ_1 and μ_2. Suppose that potential customers arrive in accordance with a Poisson process having rate λ, and that a potential customer will enter the system only if both chairs are empty. The preceding model can be analyzed as a continuous-time Markov chain, but first we must decide upon an appropriate state space. Since a potential customer will enter the system only if there are no other customers present, it follows that there will always either be 0 or 1 customers in the system. However, if there is 1 customer in the system, then we would also need to know which chair he was presently in. Hence, an appropriate state space might consist of the three states 0, 1, and 2 where the states have the following interpretation:

state	Interpretation
0	system is empty
1	a customer is in chair 1
2	a customer is in chair 2

We leave it as an exercise for you to verify that

$$v_0 = \lambda, \quad v_1 = \mu_1, \quad v_2 = \mu_2$$

$$P_{01} = P_{12} = P_{20} = 1$$

13.3 Birth and Death Processes

Consider a system whose state at any time is represented by the number of people in the system at that time. Suppose that whenever there are n people in the system, then (i) new arrivals enter the system at an exponential rate λ_n, and (ii) people leave the system at an exponential rate μ_n. That is, whenever there are n persons in the system, then the time until the next arrival is exponentially distributed with mean $1/\lambda_n$ and is independent of the time until the next departure which is itself exponentially distributed with mean $1/\mu_n$. Such a system is called a *birth and death process.* The parameters $\{\lambda_n\}_{n=0}^{\infty}$ and $\{\mu_n\}_{n=1}^{\infty}$ are called, respectively, the arrival (or birth) and departure (or death) rates.

Thus, a birth and death process is a continuous-time Markov chain with states $\{0, 1, \dots\}$ for which transitions from state n may go only to either state $n-1$ or state $n+1$. The relationships between the birth and death rates and the state transition rates and probabilities are

$$\begin{aligned}
v_0 &= \lambda_0, \\
v_i &= \lambda_i + \mu_i, \quad i > 0 \\
P_{01} &= 1, \\
P_{i,i+1} &= \frac{\lambda_i}{\lambda_i + \mu_i}, \quad i > 0 \\
P_{i,i-1} &= \frac{\mu_i}{\lambda_i + \mu_i}, \quad i > 0
\end{aligned}$$

The preceding follows, because if there are i in the system, then the next state will be $i+1$ if a birth occurs before a death, and the probability that an exponential random variable with rate λ_i will occur earlier than an (independent) exponential with rate μ_i is $\lambda_i/(\lambda_i+\mu_i)$. Moreover, the time until either a birth or a death occurs is exponentially distributed with rate $\lambda_i + \mu_i$ (and so, $v_i = \lambda_i + \mu_i$).

Example 13.2. (The Poisson Process) Consider a birth and death process for which

$$\begin{aligned}
\mu_n &= 0, \quad \text{for all } n \geq 0 \\
\lambda_n &= \lambda, \quad \text{for all } n \geq 0
\end{aligned}$$

This is a process in which departures never occur, and the time between successive arrivals is exponential with mean $1/\lambda$. Hence, this is just the Poisson process.

A birth and death process for which $\mu_n = 0$ for all n is called a pure birth process. Another pure birth process is given by the next example.

Example 13.3. (A Birth Process with Linear Birth Rate) Consider a population whose members can give birth to new members but cannot die. If each member acts independently of the others and takes an exponentially distributed amount of time, with mean $1/\lambda$, to give birth, then if $X(t)$ is the population size at time t, then $\{X(t), t \geq 0\}$ is a pure birth process with $\lambda_n = n\lambda$, $n \geq 0$. This follows since if the population consists of n persons and each gives birth at an exponential rate λ, then the total rate at which births occur is $n\lambda$. This pure birth process is known as a Yule process after G. Yule, who used it in his mathematical theory of evolution.

Example 13.4. (A Linear Growth Model with Immigration) A model in which

$$\begin{aligned}
\mu_n &= n\mu, \quad n \geq 1 \\
\lambda_n &= n\lambda + \theta, \quad n \geq 0
\end{aligned}$$

is called a linear growth process with immigration. Such processes occur naturally in the study of biological reproduction and population growth. Each individual in the population is assumed to give birth at an exponential rate λ; in addition, there is an exponential rate of increase θ of the population due to an external source such as immigration. Hence, the total birth rate where there are n persons in the system is $n\lambda + \theta$. Deaths are assumed to occur at an exponential rate μ for each member of the population, so $\mu_n = n\mu$.

Let $X(t)$ denote the population size at time t. Suppose that $X(0) = i$ and let

$$M(t) = E[X(t)]$$

We will determine $M(t)$ by deriving and then solving a differential equation that it satisfies. We start by deriving an equation for $M(t+h)$ by conditioning on $X(t)$. This yields

$$M(t+h) = E[X(t+h)] = E[E[X(t+h)|X(t)]]$$

Now, given the size of the population at time t then, ignoring events whose probability is $o(h)$, the population at time $t+h$ will either increase in size by 1 if a birth or an immigration occurs in $(t, t+h)$, or decrease by 1 if a death occurs in this interval, or remain the same if neither of these two possibilities occurs. That is, given $X(t)$,

$$X(t+h) = \begin{cases} X(t)+1, \quad \text{with probability } [\theta + X(t)\lambda]h + o(h) \\ X(t)-1, \quad \text{with probability } X(t)\mu h + o(h) \\ X(t), \quad \text{with probability } 1 - [\theta + X(t)\lambda + X(t)\mu \end{cases}$$

Therefore,

$$E[X(t+h)|X(t)] = X(t) + [\theta + X(t)\lambda - X(t)\mu]h + o(h)$$

Taking expectations yields

$$M(t+h) = M(t) + (\lambda - \mu)M(t)h + \theta h + o(h)$$

or, equivalently,

$$\frac{M(t+h) - M(t)}{h} = (\lambda - \mu)M(t) + \theta + \frac{o(h)}{h}$$

Taking the limit as $h \to 0$ yields the differential equation

$$M'(t) = (\lambda - \mu)M(t) + \theta \tag{13.1}$$

If we now define the function $h(t)$ by

$$h(t) = (\lambda - \mu)M(t) + \theta$$

then

$$h'(t) = (\lambda - \mu)M'(t)$$

Therefore, the differential Equation 13.1 can be rewritten as

$$\frac{h'(t)}{\lambda - \mu} = h(t)$$

Or

$$\frac{h'(t)}{h(t)} = \lambda - \mu$$

Integration yields

$$\log[h(t)] = (\lambda - \mu)t + c$$

or

$$h(t) = Ke^{(\lambda-\mu)t}$$

Putting this back in terms of $M(t)$ gives

$$\theta + (\lambda - \mu)M(t) = Ke^{(\lambda-\mu)t}$$

To determine the value of the constant K,we use the fact that $M(0) = i$ and evaluate the preceding at $t = 0$. This gives

$$\theta + (\lambda - \mu)i = K$$

Substituting this back in the preceding equation for $M(t)$ yields the following solution for $M(t)$:

$$M(t) = \frac{\theta}{\lambda - \mu}[e^{(\lambda-\mu)t} - 1] + ie^{(\lambda-\mu)t}$$

Note that we have implicitly assumed that $\lambda \neq \mu$. If $\lambda = \mu$, then the differential equation (6.1) reduces to

$$M'(t) = \theta \tag{13.2}$$

Integrating Equation 13.2 and using that $M(0) = i$ gives the solution

$$M(t) = \theta t + i \quad ■$$

Example 13.5. (The Queueing System $M/M/1$) Suppose that customers arrive at a single-server service station in accordance with a Poisson process having rate λ. That is, the times between successive arrivals are independent exponential random variables having mean $1/\lambda$. Upon arrival, each customer goes directly into service if the server is free; if not, then the customer joins the queue (that is, he waits in line). When the server finishes serving a customer, the customer leaves the system and the next customer in line, if there are any waiting, enters the service. The successive service times are assumed to be independent exponential random variables having mean $1/\mu$. The preceding is known as the $M/M/1$ queueing system. The first M refers to the fact that the interarrival process is Markovian (since it is a Poisson process) and the second to the fact that the service distribution is exponential (and, hence, Markovian).The 1 refers to the fact that there is a single server. If we let $X(t)$ denote the number in the system at time t then $\{X(t), t \geqslant 0\}$ is a birth and death process with

$$\begin{aligned} \mu_n &= \mu, \quad n \geqslant 1 \\ \lambda_n &= \lambda\,, \quad n \geqslant 0 \quad ■ \end{aligned}$$

Example 13.6. (A Multiserver Exponential Queueing System) Consider an exponential queueing system in which there are s servers available, each serving at rate μ. An entering customer first waits in line and then goes to the first free server. This is a birth and death process with parameters

$$\mu_n = \begin{cases} n\mu, & 1 \leqslant n \leqslant s \\ s\mu, & n > s \\ \lambda_n = \lambda, & n \geqslant 0 \end{cases}$$

To see why this is true, reason as follows: If there are n customers in the system, where $n \leqslant s$, then n servers will be busy. Since each of these servers works at rate μ, the total departure rate will be $n\mu$. On the other hand, if there are n customers in the system, where $n > s$, then all s of the servers will be busy, and thus the total departure rate will be $s\mu$.This is known as an $M/M/s$ queueing model.

Consider now a general birth and death process with birth rates $\{\lambda_n\}$ and death rates $\{\mu_n\}$, where $\bar{\mu}_0 = 0$, and let T_i denote the time, starting from state i, it takes for the process to enter state $i+1, i \geqslant 0$. We will recursively compute $E[T_i]$, $i \geqslant 0$, by starting with $i = 0$. Since T_0 is exponential with rate λ_0, we have that

$$E[T_0] = \frac{1}{\lambda_0}$$

For $i > 0$, we condition whether the first transition takes the process into state $i-1$ or $i+1$. That is, let

$$I_i = \begin{cases} 1, & \text{if the first transition from} i \text{is to} i+1 \\ 0, & \text{if the first transition from} i \text{is to} i-1 \end{cases}$$

and note that

$$\begin{aligned} E[T_i|I_i = 1] &= \frac{1}{\lambda_i + \mu_i}, \\ E[T_i|I_i = 0] &= \frac{1}{\lambda_i + \mu_i} + E[T_{i-1}] + E[T_i] \end{aligned} \tag{13.3}$$

This follows since, independent of whether the first transition is from a birth or death, the time until it occurs is exponential with rate $\lambda_i + \mu_i$; now if this first transition is a birth, then the population size is at $i+1$, so no additional time is needed; whereas if it is death, then the population size becomes $i-1$ and the additional time needed to reach $i+1$ is equal to the time it takes to return to state i (and this has mean $E[T_{i-1}]$) plus the additional time it then takes to reach $i+1$ (and this has mean $E[T_i]$). Hence, since the probability that the first transition is a birth is $\lambda_i/(\lambda_i + \mu_i)$, we see that

$$E[T_i] = \frac{1}{\lambda_i + \mu_i} + \frac{\mu_i}{\lambda_i + \mu_i}(E[T_{i-1}] + E[T_i])$$

or, equivalently,

$$E[T_i] = \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i} E[T_{i-1}], \quad i \geqslant 1$$

Starting with $E[T_0] = 1/\lambda_0$, the preceding yields an efficient method to successively compute $E[T_1], E[T_2]$, and so on.

Suppose now that we wanted to determine the expected time to go from state i to state j where $i < j$. This can be accomplished using the preceding by noting that this quantity will equal $E[T_i] + E[T_{i+1}] + \cdots + E[T_{j-1}]$.

Example 13.7. For the birth and death process having parameters $\lambda_i \equiv \lambda$, $\mu_i \equiv \mu$,

$$\begin{aligned} E[T_i] &= \frac{1}{\lambda} + \frac{\mu}{\lambda} E[T_{i-1}] \\ &= \frac{1}{\lambda}(1 + \mu E[T_{i-1}]) \end{aligned}$$

Starting with $E[T_0] = 1/\lambda$, we see that

$$\begin{aligned} E[T_1] &= \frac{1}{\lambda}\left(1 + \frac{\mu}{\lambda}\right), \\ E[T_2] &= \frac{1}{\lambda}\left[1 + \frac{\mu}{\lambda} + \left(\frac{\mu}{\lambda}\right)^2\right] \end{aligned}$$

and, in general,

$$\begin{aligned} E[T_i] &= \frac{1}{\lambda}\left[1 + \frac{\mu}{\lambda} + \left(\frac{\mu}{\lambda}\right)^2 + \cdots + \left(\frac{\mu}{\lambda}\right)^i\right] \\ &= \frac{1 - (\mu/\lambda)^{i+1}}{\lambda - \mu}, \quad i \geqslant 0 \end{aligned}$$

The expected time to reach state j, starting at state $k, k < j$, is

$$\begin{aligned} E[\text{time to go from } k \text{ to } j] &= \sum_i = k^{j-1} E[T_i] \\ &== \frac{j-k}{\lambda - \mu} - \frac{(\mu/\lambda)^{k+1}}{\lambda - \mu} \frac{[1 - (\mu/\lambda)^{j-k}]}{1 - \mu/\lambda} \end{aligned}$$

The foregoing assumes that $\lambda \neq \mu$. If $\lambda = \mu$, then

$$\begin{aligned} E[T_i] &= \frac{i+1}{\lambda}, \\ E[\text{ time to go from } k \text{ to } j] &= \frac{j(j+1) - k(k+1)}{21} \end{aligned}$$

We can also compute the variance of the time to go from 0 to $i+1$ by utilizing the conditional variance formula. First note that Equation 13.3 can be written as

$$E[T_i|I_i] = \frac{1}{\lambda_i + \mu_i} + (1 - I_i)(E[T_{i-1}] + E[T_i])$$

Thus

$$\begin{aligned}\operatorname{Var}(E[T_i|I_i]) &= (E[T_{i-1}] + E[T_i])^2 \operatorname{Var}(I_i) \\ &= (E[T_{i-1}] + E[T_i])^2 \frac{\mu_i\lambda_i}{(\mu_i+\lambda_i)^2}\end{aligned} \tag{13.4}$$

where $\operatorname{Var}(I_i)$ is as shown since I_i is a Bernoulli random variable with parameter $p = \lambda_i/(\lambda_i + \mu_i)$. Also, note that if we let X_i denote the time until the transition from i occurs, then

$$\begin{aligned}\operatorname{Var}(T_i|I_i = 1) &= \operatorname{Var}(X_i|I_i = 1) \\ &= \operatorname{Var}(X_i) \\ &= \frac{1}{(\lambda_i+\mu_i)^2}\end{aligned} \tag{13.5}$$

where the preceding uses the fact that the time until transition is independent of the next state visited. Also,

$$\begin{aligned}\operatorname{Var}(T_i|I_i = 0) &= \operatorname{Var}(X_i + \text{time to get back to } i + \text{time to then reach } i+1) \\ &= \operatorname{Var}(X_i) + \operatorname{Var}(T_{i-1}) + \operatorname{Var}(T_i)\end{aligned} \tag{13.6}$$

where the foregoing uses the fact that the three random variables are independent. We can rewrite Equation 13.5 and Equation 13.6 as

$$\operatorname{Var}(T_i|I_i) = \operatorname{Var}(X_i) + (1 - I_i)[\operatorname{Var}(T_{i-1}) + \operatorname{Var}(T_i)]$$

so

$$E[\operatorname{Var}(T_i|I_i)] = \frac{1}{(\mu_i+\lambda_i)^2} + \frac{\mu_i}{\mu_i+\lambda_i}[\operatorname{Var}(T_{i-1}) + \operatorname{Var}(T_i)] \tag{13.7}$$

Hence, using the conditional variance formula, which states that $\operatorname{Var}(T_i)$ is the sum of Equation 13.7 and Equation 13.4, we obtain

$$\begin{aligned}\operatorname{Var}(T_i) = {} & \frac{1}{(\mu_i+\lambda_i)^2} + \frac{\mu_i}{\mu_i+\lambda_i}[\operatorname{Var}(T_{i-1}) + \operatorname{Var}(T_i)] \\ &+ \frac{\mu_i\lambda_i}{(\mu_i+\lambda_i)^2}(E[T_{i-1}] + E[T_i])^2\end{aligned}$$

or, equivalently,

$$\operatorname{Var}(T_i) = \frac{1}{\lambda_i(\lambda_i+\mu_i)} + \frac{\mu_i}{\lambda_i}\operatorname{Var}(T_{i-1}) + \frac{\mu_i}{\mu_i+\lambda_i}(E[T_{i-1}] + E[T_i])^2$$

Starting with $\operatorname{Var}(T_0) = 1/\lambda_0^2$ and using the former recursion to obtain the expectations, we can recursively compute $\operatorname{Var}(T_i)$. In addition, if we want the variance of the time to reach state j, starting from state $k, k < j$,then this can be expressed as the time to go from k to $k+1$ plus the additional time to go from $k+1$ to $k+2$, and so on. Since, by the Markovian property, these successive random variables are independent, it follows that

$$\operatorname{Var}(\text{time to go from } k \text{ to } j) = \sum_{i=k}^{j-1} \operatorname{Var}(T_i)$$

14 Introduction to Queueing Theory

We will study a class of models in which customers arrive in some random manner at a service facility. Upon arrival they are made to wait in queue until it is their turn to be served. Once served they are generally assumed to leave the system. For such models we will be interested in determining, among other things, such quantities as the average number of customers in the system (or in the queue) and the average time a customer spends in the system (or spends waiting in the queue).

14.1 Preliminaries

We will derive certain identities which are valid in the great majority of queueing models.

14.1.1 Cost Equations

Some fundamental quantities of interest for queueing models are

L, the average number of customers in the system;
L_Q, the average number of customers waiting in queue;
W, the average amount of time a customer spends in the system;
W_Q, the average amount of time a customer spends waiting in queue.

A large number of interesting and useful relationships between the preceding and other quantities of interest can be obtained by making use of the following idea: Imagine that entering customers are forced to pay money (according to some rule) to the system. We would then have the following basic cost identity:

$$\begin{aligned}&\text{average rate at which the system earns}\\ &= \lambda_a \times \text{ average amount an entering customer pays}\end{aligned} \tag{14.1}$$

where λ_a is defined to be average arrival rate of entering customers. That is, if $N(t)$ denotes the number of customer arrivals by time t, then

$$\lambda_a = \lim_{t\to\infty} \frac{N(t)}{t}$$

We now present a heuristic proof of Equation 14.1.

Heuristic Proof of Equation 14.1 Let T be a fixed large number. In two different ways, we will compute the average amount of money the system has earned by time T. On one hand, this quantity approximately can be obtained by multiplying the average rate at which the system earns by the length of time T. On the other hand, we can approximately compute it by multiplying the average amount paid by an entering customer by the average number of customers entering by time T (and this latter factor is approximately $\lambda_a T$). Hence, both sides of Equation 14.1 when multiplied by T are approximately equal to the average amount earned by T. The result then follows by letting $T \to \infty$.

By choosing appropriate cost rules, many useful formulas can be obtained as special cases of Equation 14.1. For instance, by supposing that each customer pays \$1 per unit time while in the system, Equation 14.1 yields the so-called Little's formula,

$$L = \lambda_a W \tag{14.2}$$

This follows since, under this cost rule, the rate at which the system earns is just the number in the system, and the amount a customer pays is just equal to its time in the system.

Similarly if we suppose that each customer pays \$1 per unit time while in queue, then Equation 14.1 yields

$$L_Q = \lambda_a W_Q \tag{14.3}$$

By supposing the cost rule that each customer pays \$1 per unit time while in service we obtain from Equation 14.1 that the

$$\text{average number of customers in service } = \lambda_a E[S] \tag{14.4}$$

where $E[S]$ is defined as the average amount of time a customer spends in service.

It should be emphasized that Equation 14.1 through Equation 14.4 are valid for almost all queueing models regardless of the arrival process, the number of servers, or queue discipline.

14.1.2 Steady-State Probabilities

Let $X(t)$ denote the number of customers in the system at time t and define $P_n, n \geqslant 0$,by

$$P_n = \lim_{t \to \infty} P\{X(t) = n\}$$

where we assume the preceding limit exists. In other words, P_n is the limiting or long-run probability that there will be exactly n customers in the system. It is sometimes referred to as the *steady-state probability* of exactly n customers in the system. It also usually turns out that P_n equals the (long-run) proportion of time that the system contains exactly n customers. For example, if $P_0 = 0.3$, then in the long run, the system will be empty of customers for 30 percent of the time. Similarly, $P_1 = 0.2$ would imply that for 20 percent of the time the system would contain exactly one customer.

Two other sets of limiting probabilities are $\{a_n, n \geqslant 0\}$ and $\{d_n, n \geqslant 0\}$, where

$$\begin{aligned} a_n &= \text{proportion of customers that find } n \\ &\quad \text{in the system when they arrive, and} \\ d_n &= \text{proportion of customers leaving behind } n \\ &\quad \text{in the system when they depart} \end{aligned}$$

That is, P_n is the proportion of time during which there are n in the system; a_n is the proportion of arrivals that find n; and d_n is the proportion of departures that leave behind n. That these quantities need not always be equal is illustrated by the following example.

Example 14.1. Consider a queueing model in which all customers have service times equal to 1, and where the times between successive customers are always greater than 1 [for instance, the interarrival times could be uniformly distributed over (1,2)]. Hence, as every arrival finds the system empty and every departure leaves it empty, we have

$$a_0 = d_0 = 1$$

However,

$$P_0 \neq 1$$

as the system is not always empty of customers.

It was, however, no accident that α_n equaled d_n in the previous example. That arrivals and departures always see the same number of customers is always true as is shown in the next proposition.

Proposition 14.1. *In any system in which customers arrive and depart one at a time*

$$\textit{the rate at which arrivals find } n = \textit{ the rate at which departures leave } n$$

and

$$a_n = d_n$$

Proof. Proof An arrival will see n in the system whenever the number in the system goes from n to $n+1$; similarly, a departure will leave behind n whenever the number in the system goes from $n+1$ to n. Now in any interval of time T the number of transitions from n to $n+1$ must equal to within 1 the number from $n+1$ to n.[Between any two transitions from n to $n+1$,there must be one from $n+1$ to n, and conversely.] Hence, the rate of transitions from n to $n+1$ equals the rate from $n+1$ to n; or, equivalently, the rate at which arrivals find n

equals the rate at which departures leave n. Now a_n, the proportion of arrivals finding n, can be expressed as

$$a_n = \frac{\text{the rate at which arrivals find } n}{\text{overall arrival rate}}$$

Similarly,

$$d_n = \frac{\text{the rate at which departures leave } n}{\text{overall departure rate}}$$

Thus, if the overall arrival rate is equal to the overall departure rate, then the preceding shows that $a_n = d_n$. On the other hand, if the overall arrival rate exceeds the overall departure rate, then the queue size will go to infinity, implying that $a_n = d_n = 0$. □

Hence, on the average, arrivals and departures always see the same number of customers. However, as Example 25.1 illustrates, they do not, in general, see the time averages. One important exception where they do is in the case of Poisson arrivals.

Proposition 14.2. *Poisson arrivals always see time averages. In particular, for Poisson arrivals,*

$$P_n = a_n$$

To understand why Poisson arrivals always see time averages, consider an arbitrary Poisson arrival. If we knew that it arrived at time t, then the conditional distribution of what it sees upon arrival is the same as the unconditional distribution of the system state at time t. For knowing that an arrival occurs at time t gives us no information about what occurred prior to t. (Since the Poisson process has independent increments, knowing that an event occurred at time t does not affect the distribution of what occurred prior to t.) Hence, an arrival would just see the system according to the limiting probabilities.

Contrast the foregoing with the situation of Example 25.1 where knowing that an arrival occurred at time t tells us a great deal about the past; in particular it tells us that there have been no arrivals in $(t-1, t)$. Thus, in this case, we cannot conclude that the distribution of what an arrival at time t observes is the same as the distribution of the system state at time t.

For a second argument as to why Poisson arrivals see time averages, note that the total time the system is in state n by time T is (roughly) $P_n T$. Hence, as Poisson arrivals always arrive at rate λ no matter what the system state, it follows that the number of arrivals in $[0, T]$ that find the system in state n is (roughly) $\lambda P_n T$. In the long run, therefore, the rate at which arrivals find the system in state n is λP_n and, as λ is the overall arrival rate, it follows that $\lambda P_n / \lambda = P_n$ is the proportion of arrivals that find the system in state n.

The result that Poisson arrivals see time averages is called the $PASTA$ principle.

15 Exponential Models

15.1 A Single-Server Exponential Queueing System

Suppose that customers arrive at a single-server service station in accordance with a Poisson process having rate . That is, the time between successive arrivals are independent exponential random variables having mean $1/\lambda$. Each customer, upon arrival, goes directly into service if the server is free and, if not, the customer joins the queue. When the server finishes serving a customer, the customer leaves the system, and the next customer in line, if there is any, enters service. The successive service times are assumed to be independent exponential random variables having mean $1/\mu$.

The preceding is called the $M/M/1$ queue. The two M's refer to the fact that both the interarrival and the service distributions are exponential (and thus memoryless, or Markovian), and the 1 to the fact that there is a single server. To analyze it, we shall begin by determining the limiting probabilities P_n, for $n = 0, 1, ...$ To do so, think along the following lines. Suppose that we have an infinite number of rooms numbered $0, 1, 2, ...$, and suppose that we instruct an individual to enter room n whenever there are n customers in the system. That is, he would be in room 2 whenever there are two customers in the system; and if another were to arrive, then he would leave room 2 and enter room 3. Similarly, if a service would take place he would leave room 2 and enter room 1 (as there would now be only one customer in the system).

Now suppose that in the long run our individual is seen to have entered room 1 at the rate of ten times an hour. Then at what rate must he have left room 1? Clearly, at this same rate of ten times an hour. For the total number of times that he enters room 1 must be equal to (or one greater than) the total number of times he leaves room 1. This sort of argument thus yields the general principle which will enable us to determine the state probabilities. Namely, for each $n \geq 0$, *the rate at which the process enters state n equals the rate at which it leaves state n*. Let us now determine these rates. Consider first state 0. When in state 0 the process can leave only by an arrival as clearly there cannot be a departure when the system is empty. Since the arrival rate is λ and the proportion of time that the process is in state 0 is P_0, it follows that the rate at which the process leaves state 0 is λP_0. On the other hand, state 0 can only be reached from state 1 via a departure. That is, if there is a single customer in the system and he completes service, then the system becomes empty. Since the service rate is μ and the proportion of time that the system has exactly one customer is P_1, it follows that the rate at which the process enters state 0 is μP_1.

Hence, from our rate-equality principle we get our first equation,

$$\lambda P_0 = \mu P_1$$

Now consider state 1. The process can leave this state either by an arrival (which occurs at rate λ) or a departure (which occurs at rate μ). Hence, when in state 1, the process will leave this state at a rate of $\lambda + \mu^*$ Since the proportion of time the process is in state 1 is P_1, the rate at which the process leaves state 1 is $(\lambda+\mu)P_1$. On the other hand, state 1 can be entered either from state 0 via an arrival or from state 2 via a departure. Hence, the rate at which the process enters state 1 is $\lambda P_0 + \mu P_2$. Because the reasoning for other states is similar, we obtain the following set of equations:

$$\begin{array}{cc} \text{State} & \text{Rate at which the process leaves} = \text{rate at which it enters} \\ 0 & \lambda P_0 = \mu P_1 \\ n, n \geqslant 1 & (\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1} \end{array} \tag{15.1}$$

The set of Equation 15.1 which balances the rate at which the process enters each state with the rate at which it leaves that state is known as *balance equations*. In order to solve Equation 15.1, we rewrite them to obtain

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu} P_0, \\ P_{n+1} &= \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1} \right), \quad n \geqslant 1 \end{aligned}$$

Solving in terms of P_0 yields

$$\begin{aligned} P_0 &= P_0, \\ P_1 &= \frac{\lambda}{\mu} P_0, \\ P_2 &= \frac{\lambda}{\mu} P_1 + \left(P_1 - \frac{\lambda}{\mu} P_0 \right) = \frac{\lambda}{\mu} P_1 = \left(\frac{\lambda}{\mu} \right)^2 P_0, \\ P_3 &= \frac{\lambda}{\mu} P_2 + \left(P_2 - \frac{\lambda}{\mu} P_1 \right) = \frac{\lambda}{\mu} P_2 = \left(\frac{\lambda}{\mu} \right)^3 P_0, \\ P_4 &= \frac{\lambda}{\mu} P_3 + \left(P_3 - \frac{\lambda}{\mu} P_2 \right) = \frac{\lambda}{\mu} P_3 = \left(\frac{\lambda}{\mu} \right)^4 P_0, \\ P_{n+1} &= \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1} \right) = \frac{\lambda}{\mu} P_n = \left(\frac{\lambda}{\mu} \right)^{n+1} P_0 \end{aligned}$$

To determine P_0 we use the fact that the P_n must sum to 1, and thus

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

or

$$\begin{aligned} P_0 &= 1 - \frac{\lambda}{\mu}, \\ P_n &= \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \quad n \geqslant 1 \end{aligned} \tag{15.2}$$

Notice that for the preceding equations to make sense, it is necessary for λ/μ to be less than l. For otherwise $\sum_{n=0}^{\infty}(\lambda/\mu)^n$ would be infinite and all the P_n would be 0. Hence, we shall assume that $\lambda/\mu < 1$. Note that it is quite intuitive that there would be no limiting probabilities if $\lambda > \mu$. For suppose that $\lambda > \mu$. Since customers arrive at a Poisson rate λ, it follows that the expected total number of arrivals by time t is λt. On the other hand, what is the expected number of customers served by time t? If there were always customers present, then the number of customers served would be a Poisson process having rate μ since the time between successive services would be independent exponentials having mean $1/\mu$. Hence, the expected number of customers served by time ι is no greater than μt; and, therefore, the expected number in the system at time t is at least

$$\lambda t - \mu t = (\lambda - \mu)t$$

Now if $\lambda > \mu$,then the preceding number goes to infinity as t becomes large. That is, $\lambda/\mu > 1$, the queue size increases without limit and there will be no limiting probabilities. Note also that the condition $\lambda/\mu < 1$ is equivalent to the condition that the mean service time be less than the mean time between successive arrivals. This is the general condition that must be satisfied for limited probabilities to exist in most single-server queueing systems.

Remark 15.1.

(i) In solving the balance equations for the $M/M/1$ queue, we obtained as an intermediate step the set of equations

$$\lambda P_n = \mu P_{n+1}, \quad n \geqslant 0$$

These equations could have been directly argued from the general queueing result (shown in Proposition 1 in Lecture 14) that the rate at which arrivals find n in the system – namely λP_n–is equal to the rate at which departures leave behind n–namely, μP_{n+1}.

(ii) We can also prove that $P_n = (\lambda/\mu)^n(1-\lambda/\mu)$ by using a queueing cost identity. Suppose that, for a fixed $n > 0$, whenever there are at least n customers in the system the nth oldest customer (with age measured from when the customer arrived) pays 1 per unit time. Letting X be the steady state number of customers in the system, because the system earns 1 per unit time whenever X is at least n, it follows that

$$\text{average rate at which the system earns} = P\{X \geqslant n\}$$

Also, because a customer who finds fewer than $n-1$ in the system when it arrives will pay 0, while an arrival who finds at least $n-1$ in the system will pay 1 per unit time for an exponentially distributed time with rate μ

$$\text{average amount a customer pays} = \frac{1}{\mu}P\{X \geqslant n-1\}$$

Therefore, the queueing cost identity yields that

$$P\{X \geqslant n\} = (\lambda/\mu)P\{X \geqslant n-1\}, \quad n > 0$$

Iterating this gives

$$\begin{aligned} P\{X \geqslant n\} &= (\lambda/\mu)P\{X \geqslant n-1\} \\ &= (\lambda/\mu)^2 P\{X \geqslant n-2\} \\ &= \cdots \\ &= (\lambda/\mu)^n P\{X \geqslant 0\} \\ &= (\lambda/\mu)^n \end{aligned}$$

Therefore,

$$P\{X = n\} = P\{X \geqslant n\} - P\{X \geqslant n+1\} = (\lambda/\mu)^n(1-\lambda/\mu)$$

Now let us attempt to express the quantities L, L_Q, W, and W_Q in terms of the limiting probabilities P_n. Since P_n is the long-run probability that the system contains exactly n customers, the average number of customers in the system clearly is given by

$$\begin{aligned} \mathrm{L} &= \sum_{n=0}^{\infty} nP_n \\ &= \sum_{n=0}^{\infty} n\left(\frac{\lambda}{\mu}\right)^n\left(1-\frac{\lambda}{\mu}\right) \\ &= \frac{\lambda}{\mu-\lambda} \end{aligned} \tag{15.3}$$

where the last equation followed upon application of the algebraic identity

$$\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}$$

The quantities W, W_Q, and L_Q now can be obtained with the help of Equations (2) and (3) from Lecture 14. That is, since $\lambda_\alpha = \lambda$, we have from Equation 15.3 that

$$\begin{aligned} W &= \frac{L}{\lambda} \\ &= \frac{1}{\mu - \lambda}, \\ W_Q &= W - E[S] \\ &= W - \frac{1}{\mu} \\ &= \frac{\lambda}{\mu(\mu - \lambda)}, \\ L_Q &= \lambda W_Q \\ &= \frac{\lambda^2}{\mu(\mu - \lambda)} \end{aligned} \tag{15.4}$$

Example 15.1. Suppose that customers arrive at a Poisson rate of one per every 12 minutes, and that the service time is exponential at a rate of one service per 8 minutes. What are L and W?

Solution 15.1. Since $\lambda = \frac{1}{12}, \mu = \frac{1}{8}$, we have

$$L = 2, \quad W = 24$$

Hence, the average number of customers in the system is two, and the average time a customer spends in the system is 24 minutes.

Now suppose that the arrival rate increases 20 percent to $\lambda = \frac{1}{10}$. What is the corresponding change in L and W? Again using Equation 15.3, we get

$$L = 4, \quad W = 40$$

Hence, an increase of 20 percent in the arrival rate doubled the average number of customers in the system.

To understand this better, write Equation 15.3 as

$$\begin{aligned} L &= \frac{\lambda/\mu}{1 - \lambda/\mu}, \\ W &= \frac{1/\mu}{1 - \lambda/\mu} \end{aligned}$$

From these equations we can see that when λ/μ is near 1, a slight increase in λ/μ will lead to a large increase in L and W.

Remark 15.2. (**A Technical Remark**) We have used the fact that if one event occurs at an exponential rate λ, and another independent event at an exponential rate μ, then together they occur at an exponential rate $\lambda + \mu$. To check this formally, let T_1 be the time at which the first event occurs, and T_2 the time at which the second event occurs. Then

$$P\{T_1 \leqslant t\} = 1 - e^{-\lambda t},$$
$$P\{T_2 \leqslant t\} = 1 - e^{-\mu t}$$

Now if we are interested in the time until either T_1 or T_2 occurs, then we are interested in $T = \min(T_1, T_2)$. Now

$$\begin{aligned} P\{T \leqslant t\} &= 1 - P\{T > t\} \\ &= 1 - P\{\min(T_1, T_2) > t\} \end{aligned}$$

However, $\min(T_1, T_2) > t$ if and only if both T_1 and T_2 are greater than t; hence,

$$\begin{aligned} P\{T \leqslant t\} &= 1 - P\{T_1 > t, T_2 > t\} \\ &= 1 - P\{T_1 > t\}P\{T_2 > t\} \\ &= 1 - e^{-\lambda t}e^{-\mu t} \\ &= 1 - e^{-(\lambda+\mu)t} \end{aligned}$$

Thus, T has an exponential distribution with rate $\lambda + \mu$, and we are justified in adding the rates.

Given that an $M/M/1$ steady-state customer—that is, a customer who arrives after the system has been in operation a long time—spends a total of t time units in the system, let us determine the conditional distribution of N, the number of others that were present when that customer arrived. That is, letting W^* be the amount of time a customer spends in the system, we will find $P\{N = n|W^* = t\}$. Now,

$$\begin{aligned} P\{N = n|W^* = t\} &= \frac{f_{N,W^*}(n,t)}{f^*_W(t)} \\ &= \frac{P\{N = n\}f_{W^*|N}(t|n)}{f^*_W(t)} \end{aligned}$$

where $f^*_W|N(t|n)$ is the conditional density of W^* given that $N = n$, and $f^*_W(t)$ is the unconditional density of W^*. Now, given that $N = n$, the time that the customer spends in the system is distributed as the sum of $n + 1$ independent exponential random variables with a common rate μ, implying that the conditional distribution of W^* given that $N = n$ is the gamma distribution with parameters $n + 1$ and μ. Therefore, with $C = 1/f^*_W(t)$

$$\begin{aligned} P\{N = n|W^* = t\} &= CP\{N = n\}\mu e^{-\mu t}\frac{(\mu t)^n}{n!} \\ &= C(\lambda/\mu)^n(1 - \lambda/\mu)\mu e^{-\mu t}\frac{(\mu t)^n}{n!} \quad \text{(by PASTA)} \\ &= K\,\frac{(\lambda t)^n}{n!} \end{aligned}$$

where $K = C(1 - \lambda/\mu)\mu e^{-\mu t}$ does not depend on n. Summing over n yields

$$1 = \sum_{n=0}^{\infty} P\{N = n|T = t\} = K \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} = Ke^{\lambda t}$$

Thus, $K = e^{-\lambda I}$, showing that

$$P\{N = n|W^* = t\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

Therefore, the conditional distribution of the number seen by an arrival who spends a total of t time units in the system is the Poisson distribution with mean λt. In addition, as a by-product of our analysis, we have that

$$\begin{aligned} f_{W^*}(t) &= 1/C \\ &= \frac{1}{K}(1 - \lambda/\mu)\mu e^{-\mu t} \\ &= (\mu - \lambda)e^{-(\mu-\lambda)t} \end{aligned}$$

In other words, W^*, the amount of time a customer spends in the system, is an exponential random variable with rate $\mu - \lambda$. (As a check, we note that $E[W^*] = 1/(\mu - \lambda)$ which checks with Equation 15.4 since W=E[W*].)

Remark 15.3. Another argument as to why W^* is exponential with rate $\mu - \lambda$ is as follows. If we let N denote the number of customers in the system as seen by an arrival, then this arrival will spend $N + 1$ service times in the system before departing. Now,

$$P\{N + 1 = j\} = P\{N = j - 1\} = (\lambda/\mu)^{j-1}(1 - \lambda/\mu), \quad j \geqslant 1$$

In words, the number of services that have to be completed before the arrival departs is a geometric random variable with parameter $1 - \lambda/\mu$. Therefore, after each service completion our customer will be the one departing with probability $1 - \lambda/\mu$. Thus, no matter how long the customer has already spent in the system, the probability he will depart in the next h time units is $\mu h + o(h)$, the probability that a service ends in that time, multiplied by $1 - \lambda/\mu$. That is, the customer will depart in the next h time units with probability $(\mu - \lambda)h + o(h)$, which says that the hazard rate function of W^* is the constant $\mu - \lambda$. But only the exponential has a constant hazard rate, and so we can conclude that W^* is exponential with rate $\mu - \lambda$.

15.2 A Single-Server Exponential Queueing System Having Finite Capacity

In the previous model, we assumed that there was no limit on the number of customers that could be in the system at the same time. However, in reality there is always a finite system capacity N, in the sense that there can be no more than N customers in the system at any time. By this, we mean that if an arriving customer finds that there are already N customers present, then he does not enter the system.

As before, we let $P_n, 0 \leq n \leq N$, denote the limiting probability that there are n customers in the system. The rate-equality principle yields the following set of balance equations:

State	Rate at which the process leaves = rate at which it enters
0	$\lambda P_0 = \mu P_1$
$1 \leq n \leq N-1$	$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1}$
N	$\mu P_N = \lambda P_{N-1}$

The argument for state 0 is exactly as before. Namely, when in state 0, the process will leave only via an arrival (which occurs at rate λ and hence the rate at which the process leaves state 0 is λP_0. On the other hand, the process can enter state 0 only from state 1 via a departure; hence, the rate at which the process enters state 0 is μP_1. The equation for states n, where l⩽ $n < N$, is the same as before. The equation for state N is different because now state N can only be left via a departure since an arriving customer will not enter the system when it is in state N; also, state N can now only be entered from state $N-1$ (as there is no longer a state $N+1$) via an arrival. We could now either solve the balance equations exactly as we did for the infinite capacity model, or we could save a few lines by directly using the result that the rate at which departures leave behind $n-1$ is equal to the rate at which arrivals find $n-1$. Invoking this result yields that

$$\mu P_n = \lambda P_{n-1}, \quad n = 1, \dots, N \tag{15.5}$$

giving that

$$P_n = \frac{\lambda}{\mu} P_{n-1} = \left(\frac{\lambda}{\mu}\right)^2 P_{n-2} = \cdots = \left(\frac{\lambda}{\mu}\right)^n P_0, \quad n = 1, \dots, N \tag{15.6}$$

By using the fact that $\sum_{n=0}^{N} P_n = 1$ we obtain

$$\begin{aligned} 1 &= P_0 \sum_{n=0}^{N} \left(\frac{\lambda}{\mu}\right)^n \\ &= P_0 \left[\frac{1-(\lambda/\mu)^{N+1}}{1-\lambda/\mu}\right] \end{aligned}$$

or

$$P_0 = \frac{(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}}$$

and hence from Equation 15.6 we obtain

$$P_n = \frac{(\lambda/\mu)^n(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}}, \quad n = 0, 1, \dots, N \tag{15.7}$$

Note that in this case, there is no need to impose the condition that $\lambda/\mu < 1$. The queue size is, by definition, bounded so there is no possibility of its increasing indefinitely.

As before, L may be expressed in terms of P_n to yield

$$\begin{aligned} \mathrm{L} &= \sum_{n=0}^{N} n\, P_n \\ &= \frac{(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}} \sum_{n=0}^{N} n \left(\frac{\lambda}{\mu}\right)^n \end{aligned}$$

which after some algebra yields

$$L = \frac{\lambda[1 + N(\lambda/\mu)^{N+1} - (N+1)(\lambda/\mu)^N]}{(\mu-\lambda)(1-(\lambda/\mu)^{N+1})} \tag{15.8}$$

In deriving W, the expected amount of time a customer spends in the system, we must be a little carcful about what we mean by a customer. Specifically, are we including those "customers"who arrive to find the system full and thus do not spend any time in the system? Or, do we just want the expected time spent in the system by a customer who actually entered the system? The two questions lead, of course, to differcnt answers. In the first case, we have $\lambda_a = \lambda$; whereas in the second case, since the fraction of arrivals that actually enter the system is $1 - P_N$, it follows that $\lambda_a = \lambda(1 - P_N)$. Once it is clear what we mean by a customer, W can be obtained from

$$W = \frac{L}{\lambda_a}$$

Example 15.2. Suppose that it costs $c\mu$ dollars per hour to provide service at a rate μ. Suppose also that we incur a gross profit of A dollars for each customer served. If the system has a capacity N, what service rate μ maximizes our total profit?

Solution 15.2. To solve this, suppose that we use rate μ. Let us determine the amount of money coming in per hour and subtract from this the amount going out each hour. This will give us our profit per hour, and we can choose μ so as to maximize this.

Now, potential customers arrive at a rate λ. However, a certain proportion of them do not join the system—namely, those who arrive when there are N customers already in the system. Hence, since P_N is the proportion of time that the system is full, it follows that entering customers arrive at a rate of $\lambda(1 - P_N)$. Since each customer pays \$A, it follows that money comes in at an hourly rate of $\lambda(1-P_N)A$ and since it goes out at an hourly rate of $c\mu$,it follows that our total profit per hour is given by

$$\begin{aligned} \text{profit per hour} &= \lambda(1-P_N)A - c\mu \\ &= \lambda A\left[1 - \frac{(\lambda/\mu)^N(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}}\right] - c\mu \\ &= \frac{\lambda A[1-(\lambda/\mu)^N]}{1-(\lambda/\mu)^{N+1}} - c\mu \end{aligned}$$

For instance if $N = 2, \lambda = 1, A = 10, c = 1$, then

$$\begin{aligned} \text{profit per hour} &= \frac{10[1-(1/\mu)^2]}{1-(1/\mu)^3} - \mu \\ &= \frac{10(\mu^3-\mu)}{\mu^3-1} - \mu \end{aligned}$$

in order to maximize profit we differentiate to obtain

$$\frac{d}{d\mu}[\text{profit per hour}] = 10\frac{(2\mu^3-3\mu^2+1)}{(\mu^3-1)^2} - 1$$

The value of μ that maximizes our profit now can be obtained by equating to zero and solving numerically.

15.2.1 A Shoeshine Shop

Consider a shoeshine shop consisting of two chairs. Suppose that an entering customer first will go to chair 1. When his work is completed in chair l, he will go either to chair 2 if that chair is empty or else wait in chair 1 until chair 2 becomes empty. Suppose that a potential customer will enter this shop as long as chair 1 is empty. (Thus, for instance, a potential customer might enter even if there is a customer in chair 2.)

If we suppose that potential customers arrive in accordance with a Poisson process at rate λ, and that the service times for the two chairs are independent and have respective exponential rates of μ_1 and μ_2, then

(a) what proportion of potential customers enters the system?

(b) what is the mean number of customers in the system?

(c) what is the average amount of time that an entering customer spends in the system?

To begin we must first decide upon an appropriate state space. It is clear that the state of the system must include more information than merely the number of customers in the system. For instance, it would not be enough to specify that there is one customer in the system as we would also have to know which chair he was in. Further, if we only know that there are two customers in the system, then we would not know if the man in chair 1 is still being served or if he is just waiting for the person in chair 2 to finish. To account for these points, the following state space, consisting of the five states, $(0,0), (1,0), (0,1), (1,1)$, and $(b,1)$, will be used. The states have the following interpretation:

State	Interpretation
$(0,0)$	There are no customers in the system.
$(1,0)$	There is one customer in the system, and he is in chair 1.
$(0,1)$	There is one customer in the system, and he is in chair 2.
$(1,1)$	There are two customers in the system, and both are presently being served.
$(b,1)$	There are two customers in the system, but the customer in the first chair has completed his work in that chair and is waiting for the second chair to become free.

It should be noted that when the system is in state $(b,1)$, the person in chair 1, though not being served, is nevertheless "blocking" potential arrivals from entering the system.

As a prelude to writing down the balance equations, it is usually worthwhile to make a transition diagram. This is done by first drawing a circle for each state and then drawing an arrow labeled by the rate at which the process goes from one state to another. The transition diagram for this model is shown in the following figure. The explanation for the diagram is as follows: The arrow from state $(0,0)$ to state $(1,0)$ which is labeled λ means that when the process is in state $(0,0)$, that is, when the system is empty, then it goes to state $(1,0)$ at a rate λ, that is via an arrival. The arrow from $(0,1)$ to $(1,1)$ is similarly explained.

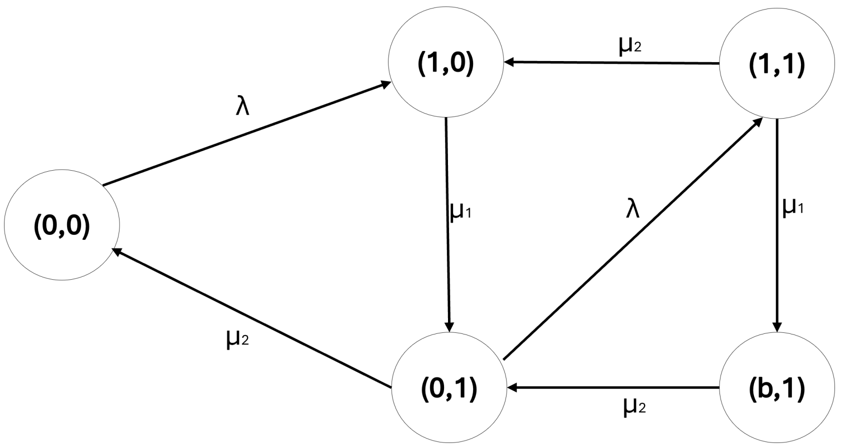


Figure 15.1: A transition diagram.

When the process is in state $(1,0)$, it will go to state $(0,1)$ when the customer in chair 1 is finished and this occurs at a rate μ_1; hence the arrow from $(1,0)$ to $(0,1)$ labeled μ_1. The arrow from $(1,1)$ to $(b,1)$ is similarly explained.

When in state $(b,1)$ the process will go to state $(0,1)$ when the customer in chair 2 completes his service (which occurs at rate μ_2); hence the arrow from $(b,1)$ to $(0,1)$ labeled μ_2. Also when in state $(1,1)$ the process will go to state $(1,0)$ when the man in chair 2 finishes and hence the arrow from $(1,1)$ to $(1,0)$ labeled μ_2. Finally, if the process is in state $(0,1)$, then it will go to state $(0,0)$ when the man in chair 2 completes his service, hence the arrow from $(0,1)$ to $(0,0)$ labeled μ_2.

Because there are no other possible transitions, this completes the transition diagram.

To write the balance equations we equate the sum of the arrows (multiplied by the probability of the states where they originate) coming into a state with the sum of the arrows (multiplied by the probability of the state) going out of that state. This gives

State	Rate that the process leaves = rate that it enters
$(0,0)$	$\lambda P_{00} = \mu_2 P_{01}$
$(1,0)$	$\mu_1 P_{10} = \lambda P_{00} + \mu_2 P_{11}$
$(0,1)$	$(\lambda + \mu_2) P_{01} = \mu_1 P_{10} + \mu_2 P_{b1}$
$(1,1)$	$(\mu_1 + \mu_2) P_{11} = \lambda P_{01}$
$(b,1)$	$\mu_2 P_{b1} = \mu_1 P_{11}$

These along with the equation

$$P_{00} + P_{10} + P_{01} + P_{11} + P_{b1} = 1$$

may be solved to determine the limiting probabilities. Though it is easy to solve the preceding equations, the resulting solutions are quite involved and hence will not be explicitly presented. However, it is easy to answer our questions in terms of these limiting probabilities. First, since a potential customer will enter the system when the state is either $(0,0)$ or $(0,1)$, it follows that the proportion of customers entering the system is $P_{00} + P_{01}$. Secondly, since there is one customer in the system whenever the state is $(0,1)$ or $(1,0)$ and two customers in the system whenever the state is $(1,1)$ or $(b,1)$, it follows that L, the average number in the system, is given by

$$L = P_{01} + P_{10} + 2(P_{11} + P_{b1})$$

To derive the average amount of time that an entering customer spends in the system, we use the relationship $W = L/\lambda_\alpha$. Since a potential customer will enter the system when in state $(0,0)$ or $(0,1)$, it follows that $\lambda_a = \lambda(P_{00} + P_{01})$ and hence

$$W = \frac{P_{01} + P_{10} + 2(P_{11} + P_{b1})}{\lambda_{(}P_{00} + P_{01})}$$

Example 15.3.

(a) If $\lambda = 1, \mu_1 = 1, \mu_2 = 2$, then calculate the preceding quantities of interest.

(b) If $\lambda = 1$, $\mu_1 = 2$, $\mu_2 = 1$, then calculate the preceding.

Solution 15.3.

(a) Solving the balance equations yields

$$P_{00} = \frac{12}{37}, \quad P_{10} = \frac{16}{37}, \quad P_{11} = \frac{2}{37}, \quad P_{01} = \frac{6}{37}, \quad P_{b1} = \frac{1}{37}$$

Hence,

$$L = \frac{28}{37}, \quad W = \frac{28}{18}$$

(b) Solving the balance equations yields

$$P_{00} = \frac{3}{11}, \quad P_{10} = \frac{2}{11}, \quad P_{11} = \frac{1}{11}, \quad P_{b1} = \frac{2}{11}, \quad P_{01} = \frac{3}{11}$$

Hence,

$$L = 1, \quad W = \frac{11}{6}$$

16 Brownian Motion

16.1 Brownian Motion

Let us start by considering the symmetric random walk, which in each time unit is equally likely to take a unit step either to the left or to the right. That is, it is a Markov chain with $P_{i,i+1} = \frac{1}{2} = P_{i,i-1}, i = 0, \pm 1, \ldots.$ Now suppose that we speed up this process by taking smaller and smaller steps in smaller and smaller time intervals. If we now go to the limit in the right manner, what we obtain is Brownian motion.

More precisely, suppose that each Δt time unit we take a step of size Δx either to the left or the right with equal probabilities. If we let $X(t)$ denote the position at time t then

$$X(t) = \Delta x(X_1 + \cdots + X_{[t/\Delta t]}) \tag{16.1}$$

where

$$X_i = \begin{cases} +1, & \text{if the } i\text{th step of length } \Delta x \text{ is to the right,} \\ -1, & \text{if it is to the left} \end{cases}$$

and $[t/\Delta t]$ is the largest integer less than or equal to $t/\Delta t$, and where the X_i are assumed independent with

$$P\{X_i = 1\} = P\{X_i = -1\} = \frac{1}{2}$$

As $E[X_i] = 0$, $\text{Var}(X_i) = E[X_i^2] = 1$, we see from Equation 16.1 that

$$\begin{aligned} E[X(t)] &= 0 \\ \text{Var}(X(t)) &= (\Delta x)^2 \left[\frac{t}{\Delta t}\right] \end{aligned} \tag{16.2}$$

We shall now let Δx and Δt go to 0. However, we must do it in a way such that the resulting limiting process is nontrivial (for instance, if we let $\Delta x = \Delta t$ and let $\Delta t \to 0$, then from the preceding we see that $E[X(t)]$ and $\text{Var}(X(t))$ would both converge to 0 and thus $X(t)$ would equal 0 with probability 1). If we let $\Delta x = \sigma\sqrt{\Delta t}$ for some positive constant σ then from Equation 16.2 we see that *as* $\Delta t \to 0$

$$E[X(t)] = 0, \text{Var}(X(t)) \to \sigma^2 t$$

We now list some intuitive properties of this limiting process obtained by taking $\Delta x = \sigma\sqrt{\Delta t}$ and then letting $\Delta t \to 0$. From Equation 16.1 and the central limit theorem the following seems reasonable:

(i) $X(t)$ is normal with mean 0 and variance $\sigma^2 t$. In addition, because the changes of value of the random walk in nonoverlapping time intervals are independent, we have

(ii) $\{X(t), t \geqslant 0\}$ has independent increments, in that for all $t_1 < t_2 < \cdots < t_n$

$$X(t_n) - X(t_{n-1}), X(t_{n-1}) - X(t_{n-2}), \ldots, X(t_2) - X(t_1), X(t_1)$$

are independent. Finally, because the distribution of the change in position of the random walk over any time interval depends only on the length of that interval, it would appear that

(iii) $\{X(t),\, t \geqslant 0\}$ has stationary increments, in that the distribution of $X(t+s) - X(t)$ does not depend on t. We are now ready for the following formal definition.

Definition 16.1. A stochastic process $\{X(t), t \geqslant 0\}$ is said to be a *Brownian motion* process if

(i) $X(0) = 0$;

(ii) $\{X(t), t \geqslant 0\}$ has stationary and independent increments;

(iii) for every $t > 0, X(t)$ is normally distributed with mean 0 and variance $\sigma^2 t$.

The Brownian motion process, sometimes called the Wiener process, is one of the most useful stochastic processes in applied probability theory. It originated in physics as a description of Brownian motion. This phenomenon, named after the English botanist Robert Brown who discovered it, is the motion exhibited by a small particle which is totally immersed in a liquid or gas. Since then, the process has been used beneficially in such areas as statistical testing of goodness of fit, analyzing the price levels on the stock market, and quantum mechanics.

The first explanation of the phenomenon of Brownian motion was given by Einstein in 1905. He showed that Brownian motion could be explained by assuming that the immersed particle was continually being subjected to bombardment by the molecules of the surrounding medium. However, the preceding concise definition of this stochastic process underlying Brownian motion was given by Wiener in a series of papers originating in 1918.

When $\sigma = 1$, the process is called *standard Brownian motion.* Because any Brownian motion can be converted to the standard process by letting $B(t) = X(t)/\sigma$ we shall, unless otherwise stated, suppose throughout this chapter that $\sigma = 1$.

The interpretation of Brownian motion as the limit of the random walks Equation 16.1 suggests that $X(t)$ should be a continuous function of t. This turns out to be the case, and it may be proven that, with probability 1, $X(t)$ is indeed a continuous function of t. This fact is quite deep, and no proof shall be attempted.

As $X(t)$ is normal with mean 0 and variance t, its density function is given by

$$f_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}$$

To obtain the joint density function of $X(t_1), X(t_2), \dots, X(t_n)$ for $t_1 < \dots < t_n$, note first that the set of equalities

$$\begin{aligned} X(t_1) &= x_1, \\ X(t_2) &= x_2, \\ &\vdots \\ X(t_n) &= x_n \end{aligned}$$

is equivalent to

$$\begin{aligned} X(t_1) &= x_1, \\ X(t_2) - X(t_1) &= x_2 - x_1, \\ &\vdots \\ X(t_n) - X(t_{n-1}) &= x_n - x_{n-1} \end{aligned}$$

However, by the independent increment assumption it follows that $X(t_1)$, $X(t_2) - X(t_1), \dots, X(t_n) - X(t_{n-1})$, are independent and, by the stationary increment assumption, that $X(t_k) - X(t_{k-1})$ is normal with mean 0 and variance $t_k - t_{k-1}$. Hence, the joint density of $X(t_1), \dots, X(t_n)$ is given by

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= f_{t_1}(x_1) f_{t_2-t_1}(x_2 - x_1) \cdots f_{t_n - t_{n-1}}(x_n - x_{n-1}) \\ &= \frac{\exp\left\{-\frac{1}{2}\left[\frac{x_1^2}{t_1} + \frac{(x_2-x_1)^2}{t_2-t_1} + \dots + \frac{(x_n-x_{n-1})^2}{t_n-t_{n-1}}\right]\right\}}{(2\pi)^{n/2}[t_1(t_2-t_1)\cdots(t_n-t_{n-1})]^{1/2}} \end{aligned} \tag{16.3}$$

From this equation, we can compute in principle any desired probabilities. For instance, suppose we require the conditional distribution of $X(s)$ given that $X(t) = B$ where $s < t$. The conditional density is

$$\begin{aligned} f_{s|t}(x|B) &= \frac{f_s(x)\, f_{t-s}(B - x)}{f_t(B)} \\ &= K_1 \exp\{-x^2/2s - (B-x)^2/2(t-s)\} \\ &= K_2 \exp\left\{-x^2\left(\frac{1}{2s} + \frac{1}{2(t-s)}\right) + \frac{Bx}{t-s}\right\} \\ &= K_2 \exp\left\{-\frac{t}{2s(t-s)}\left(x^2 - 2\frac{sB}{t}x\right)\right\} \\ &= K_3 \exp\left\{-\frac{(x - Bs/t)^2}{2s(t-s)/t}\right\} \end{aligned}$$

where K_1, K_2, and K_3 do not depend on x. Hence, we see from the preceding that the conditional distribution of $X(s)$ given that $X(t) = B$ is, for $s < t$, normal with mean and variance given by

$$\begin{aligned} E[X(s)|X(t) = B] &= \frac{s}{t}B, \\ \operatorname{Var}[X(s)|X(t) = B] &= \frac{s}{t}(t-s) \end{aligned} \tag{16.4}$$

Example 16.1. In a bicycle race between two competitors, let $Y(t)$ denote the amount of time (in seconds) by which the racer that started in the inside position is ahead when 100t percent of the race has been completed, and suppose that $\{Y(t),\ 0 \leqslant t \leqslant 1\}$ can be effectively modeled as a Brownian motion process with variance parameter σ^2.

(a) If the inside racer is leading by σ seconds at the midpoint of the race, what is the probability that she is the winner?

(b) If the inside racer wins the race by a margin of σ seconds, what is the probability that she was ahead at the midpoint?

Solution 16.1.

(a)

$$\begin{aligned} &P\{Y(1) > 0|Y(1/2) = \sigma\} \\ &= P\{Y(1) - Y(1/2) > -\sigma|Y(1/2) = \sigma\} \\ &= P\{Y(1) - Y(1/2) > -\sigma\} \quad \text{by independent increments} \\ &= P\{Y(1/2) > -\sigma\} \quad \text{by stationary increments} \\ &= P\left\{\frac{Y(1/2)}{\sigma/\sqrt{2}} > -\sqrt{2}\right\} \\ &= \Phi(\sqrt{2}) \\ &\approx 0.9213 \end{aligned}$$

where $\Phi(x) = P\{N(0,1) \leqslant x\}$ is the standard normal distribution function.

(b) Because we must compute $P\{Y(1/2) > 0|Y(1) = \sigma\}$, let us first determine the conditional distribution of $Y(s)$ given that $Y(t) = C$, when $s < t$. Now, since $\{X(t), t \geqslant 0\}$ is standard Brownian motion when $X(t) = Y(t)/\sigma$, we obtain from Equation 16.4 that the conditional distribution of $X(s)$, given that $X(t) = C/\sigma$, is normal with mean $sC/t\sigma$ and variance $s(t-s)/t$. Hence, the conditional distribution of $Y(s) = \sigma X(s)$ given that $Y(t) = C$ is normal with mean sC/t and variance $\sigma^2 s(t-s)/t$. Hence,

$$\begin{aligned} P\{Y(1/2) > 0|Y(1) = \sigma\} &= P\{N(\sigma/2, \sigma^2/4) > 0\} \\ &= \Phi(1) \\ &\approx 0.8413 \end{aligned}$$

16.2 Hitting Times, Maximum Variable, and the Gambler's Ruin Problem

Let T_a denote the first time the Brownian motion process hits a. When $a > 0$ we will compute $P\{T_a \leqslant t\}$ by considering $P\{X(t) \geqslant a\}$ and conditioning on whether or not $T_a \leqslant t$. This gives

$$\begin{aligned} P\{X(t) \geqslant a\} &= P\{X(t) \geqslant a | T_a \leqslant t\} P\{T_a \leqslant t\} \\ &\quad + P\{X(t) \geqslant a | T_a > t\} P\{T_a > t\} \end{aligned} \tag{16.5}$$

Now if $T_a \leqslant t$, then the process hits a at some point in $[0, t]$ and, by symmetry, it is just as likely to be above a or below a at time t. That is

$$P\{X(t) \geqslant a | T_a \leqslant t\} = \frac{1}{2}$$

As the second right-hand term of Equation 16.5 is clearly equal to 0 (since, by continuity, the process value cannot be greater than a without having yet hit a), we see that

$$\begin{aligned} P\{T_a \leqslant t\} &= 2P\{X(t) \geqslant a\} \\ &= \frac{2}{\sqrt{2\pi t}} \int_a^\infty e^{-x^2/2t}\, dx \\ &= \frac{2}{\sqrt{2\pi}} \int_{a/\sqrt{t}}^\infty e^{-y^2/2}\, dy, \quad a > 0 \end{aligned} \tag{16.6}$$

For $a < 0$, the distribution of T_a is, by symmetry, the same as that of T_{-a}. Hence, from Equation 16.6 we obtain

$$P\{T_a \leqslant t\} = \frac{2}{\sqrt{2\pi}} \int_{|a|/\sqrt{t}}^\infty e^{-y^2/2}\, dy \tag{16.7}$$

Another random variable of interest is the maximum value the process attains in $[0, t]$. Its distribution is obtained as follows: For $a > 0$

$$\begin{aligned} P\left\{\max_{0 \leqslant s \leqslant t} X(s) \geqslant a\right\} &= P\{T_a \leqslant t\} \quad \text{by continuity} \\ &= 2P\{X(t) \geqslant a\} \quad \text{from (6)} \\ &= \frac{2}{\sqrt{2\pi}} \int_{a/\sqrt{t}}^\infty e^{-y^2/2}\, dy \end{aligned}$$

Let us now consider the probability that Brownian motion hits A before $-B$ where $A > 0, B > 0$. To compute this we shall make use of the interpretation of Brownian motion as being a limit of the symmetric random walk. To start let us recall from the results of the gambler's ruin problem that the probability that the symmetric random walk goes up A before going down B when each step is equally likely to be either up or down a distance Δx is with $N = (A + B)/\Delta x, i = B/\Delta x]$ equal to $B\Delta x/(A + B)\Delta x = B/(A + B)$.

Hence, upon letting $\Delta x \to 0$, we see that

$$P\{\text{up } A \text{ before down } B\} = \frac{B}{A + B}$$

17 Variations on Brownian Motion and Pricing Stock Options

17.1 Brownian Motion with Drift

We say that $\{X(t), t \geqslant 0\}$ is a Brownian motion process with drift coefficient μ and variance parameter σ^2 if

(i)$X(0) = 0$;

(ii) $\{X(t),\, t \geqslant 0\}$ has stationary and independent increments;

(iii) $X(t)$ is normally distributed with mean μt and variance $t\sigma^2$.

An equivalent definition is to let $\{B(t), t \geqslant 0\}$ be standard Brownian motion and then define

$$X(t) = \sigma B(t) + \mu t$$

17.2 Geometric Brownian Motion

If $\{Y(t),\, t \geqslant 0\}$ is a Brownian motion process with drift coefficient μ and variance parameter σ^2, then the process $\{X(t),\, t \geqslant 0\}$ defined by

$$X(t) = e^{Y(t)}$$

is called *geometric Brownian motion.*

For a geometric Brownian motion process $\{X(t)\}$, let us compute the expected value of the process at time t given the history of the process up to time s. That is, for $s < t$, consider $E[X(t)|X(u), 0 \leqslant u \leqslant s]$. Now,

$$\begin{aligned} E[X(t)|X(u),\, 0 \leqslant u \leqslant s] &= E\big[e^{Y(t)}|Y(u),\, 0 \leqslant u \leqslant s\big] \\ &= E\big[e^{Y(s)+Y(t)-Y(s)}|Y(u),\, 0 \leqslant u \leqslant s\big] \\ &= e^{Y(s)}E\Big[e^{Y(t)-Y(s)}|Y(u),\, 0 \leqslant u \leqslant s\Big] \\ &= X(s)E\Big[e^{Y(t)-Y(s)}\Big] \end{aligned}$$

where the next to last equality follows from the fact that $Y(s)$ is given, and the last equality from the independent increment property of Brownian motion. Now, the moment generating function of a normal random variable W is given by

$$E[e^{aW}] = e^{aE[W]+a^2\,\mathrm{Var}(W)/2}$$

Hence, since $Y(t) - Y(s)$ is normal with mean $\mu(t-s)$ and variance $(t-s)\sigma^2$, it follows by setting $a = 1$ that

$$E\left[e^{Y(t)-Y(s)}\right] = e^{\mu(t-s)+(t-s)\sigma^2/2}$$

Thus, we obtain

$$E[X(t)|X(u),\, 0 \leqslant u \leqslant s] = X(s)e^{(t-s)(\mu+\sigma^2/2)} \tag{17.1}$$

Geometric Brownian motion is useful in the modeling of stock prices over time when you feel that the percentage changes are independent and identically distributed. For instance, suppose that X_n is the price of some stock at time n. Then, it might be reasonable to suppose that $X_n/X_{n-1}, n \geqslant 1$, are independent and identically distributed. Let

$$Y_n = X_n/X_{n-1}$$

and so

$$X_n = Y_n X_{n-1}$$

Iterating this equality gives

$$\begin{aligned} X_n &= Y_n Y_{n-1} X_{n-2} \\ &= Y_n Y_{n-1} Y_{n-2} X_{n-3} \\ &\vdots \\ &= Y_n Y_{n-1} \cdots Y_1 X_0 \end{aligned}$$

Thus,

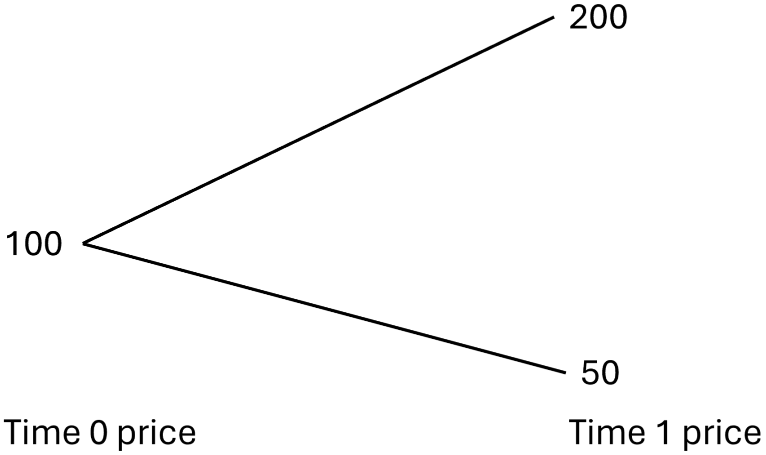
$$\log(X_n) = \sum_{i=1}^{n} \log(Y_i) + \log(X_0)$$

Since $\log(Y_i), i \geqslant 1$ are independent and identically distributed, $\{\log(X_n)\}$ will, when suitably normalized, approximately be Brownian motion with a drift, and so $\{X_n\}$ will be approximately geometric Brownian motion.

17.3 Pricing Stock Options

17.3.1 An Example in Options Pricing

In situations in which money is to be received or paid out in differing time periods, we must take into account the time value of money. That is, to be given the amount v a time t in the future is not worth as much as being given v immediately. The reason for this is that if we were immediately given v, then it could be loaned out with interest and so be worth more than v at time t. To take this into account, we will suppose that the time 0 value, also called the *present value*, of the amount v to be earned at time t is $ve^{-\alpha t}$. The quantity α is often called the discount factor. In economic terms, the assumption of the discount function $e^{-\alpha t}$ is equivalent to the assumption that we can earn interest at a continuously compounded rate of 100α percent per unit time.



We will now consider a simple model for pricing an option to purchase a stock at a future time at a fixed price.

Suppose the present price of a stock is \$100 per unit share, and suppose we know that after one time period it will be, in present value dollars, either \$200 or \$50 (see the figure above). It should be noted that the prices at time 1 are the present value (or time 0) prices. That is, if the discount factor is , then the actual possible prices at time 1 are either $200e^{\alpha}$ or $50e^{\alpha}$. To keep the notation simple, we will suppose that all prices given are time 0 prices.

Suppose that for any y, at a cost of cy, you can purchase at time 0 the option to buy y shares of the stock at time 1 at a (time 0) cost of \$150 per share. Thus, for instance, if you do purchase this option and the stock rises to \$200, then you would exercise the option at time 1 and realize a gain of \$200 - \$150 = \$50 for each of the y option units purchased. On the other hand, if the price at time 1 was \$50, then the option would be worthless at time 1. In addition, at a cost of $100x$ you can purchase x units of the stock at time 0, and this will be worth either $200x$ or $50x$ at time 1.

We will suppose that both x or y can be either positive or negative (or zero). That is, you can either buy or sell both the stock and the option. For instance, if x were negative then you would be selling $-x$ shares of the stock, yielding you a return of $-100x$, and you would then be responsible for buying $-x$ shares of the stock at time 1 at a cost of either \$200 or \$50 per share.

We are interested in determining the appropriate value of c, the unit cost of an option. Specifically, we will show that unless $c = 50/3$ there will be a combination of purchases that will always result in a positive gain.

To show this, suppose that at time 0 we

buy x units of stock, and
buy y units of options

where x and y (which can be either positive or negative) are to be determined. The value of our holding at time 1 depends on the price of the stock at that time; and it is given by the following

$$\text{value} = \begin{cases} 200x + 50y, & \text{if price is 200} \\ 50x, & \text{if price is 50} \end{cases}$$

The preceding formula follows by noting that if the price is 200 then the x units of the stock are worth $200x$, and the y units of the option to buy the stock at a unit price of 150 are worth $(200 - 150)y$. On the other hand, if the stock price is 50, then the x units are worth $50x$ and the y units of the option are worthless. Now, suppose we choose y to be such that the preceding value is the same no matter what the price at time 1. That is, we choose y so that

$$200x + 50y = 50x$$

or

$$y = -3x$$

(Note that y has the opposite sign of x, and so if x is positive and as a result x units of the stock are purchased at time 0, then $3x$ units of stock options are also sold at that time. Similarly, if x is negative, then $-x$ units of stock are sold and $-3x$ units of stock options are purchased at time 0.)

Thus, with $y = -3x$, the value of our holding at time 1 is

$$\text{value} = 50x$$

Since the original cost of purchasing x units of the stock and $-3x$ units of options is

$$\text{original cost} = 100x - 3xc$$

we see that our gain on the transaction is

$$\text{gain} = 50x - (100x - 3xc) = x(3c - 50)$$

Thus, if $3c = 50$, then the gain is 0; on the other hand if $3c \neq 50$, we can guarantee a positive gain (no matter what the price of the stock at time 1) by letting x be positive when $3c > 50$ and letting it be negative when $3c < 50$.

For instance, if the unit cost per option is $c = 20$, then purchasing 1 unit of the stock ($x = 1$) and simultaneously selling 3 units of the option ($y = -3$) initially costs us $100 - 60 = 40$. However, the value of this holding at time 1 is 50 whether the stock goes up to 200 or down to 50. Thus, a guaranteed profit of 10 is attained. Similarly, if the unit cost per option is $c = 15$, then selling 1 unit of the stock ($x = -1$) and buying 3 units of the option ($y = 3$) leads to an initial gain of $100 - 45 = 55$. On the other hand, the value of this holding at time 1 is -50. Thus, a guaranteed profit of 5 is attained.

A sure win betting scheme is called an *arbitrage*. Thus, as we have just seen, the only option cost c that does not result in an arbitrage is $c = 50/3$.

17.3.2 The Arbitrage Theorem

Consider an experiment whose set of possible outcomes is $S = \{1, 2, \dots, m\}$. Suppose that n wagers are available. If the amount x is bet on wager i, then the return $xr_i(j)$ is earned if the outcome of the experiment is j. In other words, $r_i(\cdot)$ is the return function for a unit bet on wager i. The amount bet on a wager is allowed to be either positive or negative or zero.

A betting scheme is a vector $x = (x_1, \dots, x_n)$ with the interpretation that x_1 is bet on wager 1, x_2 on wager 2, . . . , and x_n on wager n. If the outcome of the experiment is j, then the return from the betting scheme x is

$$\text{return from } x = \sum_{i=1}^{n} x_i r_i(j)$$

The following theorem states that either there exists a probability vector $p = (p_1, \dots, p_m)$ on the set of possible outcomes of the experiment under which each of the wagers has expected return 0, or else there is a betting scheme that guarantees a positive win.

Theorem 17.1. *(The Arbitrage Theorem) Exactly one of the following is true: Either*

(i) there exists a probability vector $p = (p_1, \dots, p_m)$ for which

$$\sum_{j=1}^{m} p_j r_i(j) = 0, \quad \textit{for all } i = 1, \dots, n$$

or

(ii) there exists a betting scheme $x = (x_1, \dots, x_n)$ for which

$$\sum_{i=1}^{n} x_i r_i(j) > 0, \quad \textit{for all } j = 1, \dots, m$$

In other words, if X is the outcome of the experiment, then the arbitrage theorem states that either there is a probability vector p for X such that

$$E_p[r_i(X)] = 0, \quad \text{for all } i = 1, \dots, n$$

or else there is a betting scheme that leads to a sure win.

Remark 17.1. This theorem is a consequence of the (linear algebra) theorem of the separating hyperplane, which is often used as a mechanism to prove the duality theorem of linear programming.

The theory of linear programming can be used to determine a betting strategy that guarantees the greatest return. Suppose that the absolute value of the amount bet on each wager must be less than or equal to 1. To determine the vector x that yields the greatest guaranteed win—call this win v—we need to choose x and v so as to maximize v, subject to the constraints

$$\sum_{i=1}^{n} x_i r_i(j) \geq v, \quad \text{for } j = 1, \dots, m$$
$$-1 \leq x_i \leq 1, \quad i = 1, \dots, n$$

This optimization problem is a linear program and can be solved by standard techniques (such as by using the simplex algorithm). The arbitrage theorem yields that the optimal v will be positive unless there is a probability vector p for which $\sum_{j=1}^{m} p_j r_i(j) = 0$ for all $i = 1, \dots, n$.

Example 17.1. In some situations, the only types of wagers allowed are to choose one of the outcomes $i, i = 1, \dots, m$, and bet that i is the outcome of the experiment. The return from such a bet is often quoted in terms of "odds". If the odds for outcome i are o_i (often written as "o_i to 1") then a 1 unit bet will return o_i if the outcome of the experiment is i and will return -1 otherwise. That is,

$$r_i(j) = \begin{cases} o_i, & \text{if } j = i \\ -1 & \text{otherwise} \end{cases}$$

Suppose the odds $o_1, \dots, o_m$ are posted. In order for there not to be a sure win there must be a probability vector $p = (p_1, \dots, p_m)$ such that

$$0 \equiv E_p[r_i(X)] = o_i p_i - (1 - p_i)$$

That is, we must have

$$p_i = \frac{1}{1 + o_i}$$

Since the p_i must sum to 1, this means that the condition for there not to be an arbitrage is that

$$\sum_{i=1}^{m} (1 + o_i)^{-1} = 1$$

Thus, if the posted odds are such that $\sum_{i=1}^{m}(1+o_i)^{-1} \neq 1$, then a sure win is possible. For instance, suppose there are three possible outcomes and the odds are as follows:

Outcome	Odds
1	1
2	2
3	3

That is, the odds for outcome 1 are $1 - 1$, the odds for outcome 2 are $2 - 1$, and that for outcome 3 are 3—1. Since

$$\frac{1}{2} + \frac{1}{3} + \frac{1}{4} > 1$$

a sure win is possible. One possibility is to bet \$-1 on outcome 1 (and so you either win 1 if the outcome is not 1 and lose 1 if the outcome is 1) and bet -0.7 on outcome 2, and -0.5 on outcome 3. If the experiment results in outcome 1, then we win $-1 + 0.7 + 0.5 = 0.2$; if it results in outcome 2, then we win $1 - 1.4 + 0.5 = 0.1$; if it results in outcome 3, then we win $1 + 0.7 - 1.5 = 0.2$. Hence, in all cases we win a positive amount.

Remark 17.2. If $\sum_i (1+o_i)^{-1} \neq 1$, then the betting scheme

$$x_i = \frac{(1+o_i)^{-1}}{1 - \sum_i (1+o_i)^{-1}}, \quad i = 1, \dots, n$$

will always yield a gain of exactly 1.

Example 17.2. Let us reconsider the option pricing example of the previous section, where the initial price of a stock is 100 and the present value of the price at time 1 is either 200 or 50. At a cost of c per share we can purchase at time 0 the option to buy the stock at time 1 at a present value price of 150 per share. The problem is to set the value of c so that no sure win is possible.

In the context of this section, the outcome of the experiment is the value of the stock at time 1. Thus, there are two possible outcomes. There are also two different wagers: to buy (or sell) the stock, and to buy (or sell) the option. By the arbitrage theorem, there will be no sure win if there is a probability vector $(p, 1-p)$ that makes the expected return under both wagers equal to 0.

Now, the return from purchasing 1 unit of the stock is

$$\text{return} = \begin{cases} 200 - 100 = 100, & \text{if the price is 200 at time 1} \\ 50 - 100 = -50, & \text{if the price is 50 at time 1} \end{cases}$$

Hence, if p is the probability that the price is 200 at time 1, then

$$E[\text{return}] = 100p - 50(1-p)$$

Setting this equal to 0 yields that

$$p = \frac{1}{3}$$

That is, the only probability vector $(p, 1-p)$ for which wager 1 yields an expected return 0 is the vector $(\frac{1}{3}, \frac{2}{3})$.

Now, the return from purchasing one share of the option is

$$\text{return} = \begin{cases} 50 - c, & \text{if price is } 200 \\ \\ -c, & \text{if price is } 50 \end{cases}$$

Hence, the expected return when $p = \frac{1}{3}$ is

$$\begin{aligned} E[\text{return}] &= (50 - c)\frac{1}{3} - c\frac{2}{3} \\ &= \frac{50}{3} - c \end{aligned}$$

Thus, it follows from the arbitrage theorem that the only value of c for which there will not be a sure win is $c = \frac{50}{3}$, which verifies the result of Section 17.3.1.

17.4 The Black-Scholes Option Pricing Formula

Suppose the present price of a stock is $X(0) = x_0$, and let $X(t)$ denote its price at time t. Suppose we are interested in the stock over the time interval 0 to T. Assume that the discount factor is α (equivalently, the interest rate is 100α percent compounded continuously), and so the present value of the stock price at time t is $e^{-\alpha t}X(t)$.

We can regard the evolution of the price of the stock over time as our experiment, and thus the outcome of the experiment is the value of the function $X(t)$, $0 \leqslant t \leqslant T$. The types of wagers available are that for any $s < t$ we can observe the process for a time s and then buy (or sell) shares of the stock at price $X(s)$ and then sell (or buy) these shares at time t for the price $X(t)$. In addition, we will suppose that we may purchase any of N different options at time 0. Option i, costing c_i per share, gives us the option of purchasing shares of the stock at time t_i for the fixed price of K_i per share, $i = 1, \dots, N$.

Suppose we want to determine values of the c_i for which there is no betting strategy that leads to a sure win. Assuming that the arbitrage theorem can be generalized (to handle the preceding situation, where the outcome of the experiment is a function), it follows that there will be no sure win if and only if there exists a probability measure over the set of outcomes under which all of the wagers have expected return 0. Let P be a probability measure on the set of outcomes. Consider first the wager of observing the stock for a time s and then purchasing (or selling) one share with the intention of selling (or purchasing) it at time $t, 0 \leqslant s < t \leqslant T$.

The present value of the amount paid for the stock is $e^{-\alpha s X(s)}$, whereas the present value of the amount received is $e^{-\alpha t X(t)}$. Hence, in order for the expected return of this wager to be 0 when P is the probability measure on $X(t), 0 \leqslant t \leqslant T$, we must have that

$$E_{\mathbf{P}}[e^{-\alpha t} X(t) | X(u), 0 \leqslant u \leqslant s] = e^{-\alpha s} X(s) \tag{17.2}$$

Consider now the wager of purchasing an option. Suppose the option gives us the right to buy one share of the stock at time t for a price K. At time t, the worth of this option will be as follows:

$$\text{worth of option at time } t = \begin{cases} X(t) - K, & \text{if } X(t) \geqslant K \\ 0, & \text{if } X(t) < K \end{cases}$$

That is, the time t worth of the option is $(X(t) - K)^+$. Hence, the present value of the worth of the option is $e^{-\alpha t} \left(X\left(t\right) - K\right)^+$. If c is the (time 0) cost of the option, we see that, in order for purchasing the option to have expected (present value) return 0, we must have that

$$E_{\mathbf{P}}[e^{-\alpha t} (X(t) - K)^+] = c \tag{17.3}$$

By the arbitrage theorem, if we can find a probability measure P on the set of outcomes that satisfies Equation 17.2, then if c, the cost of an option to purchase one share at time t at the fixed price K , is as given in Equation 17.3, then no arbitrage is possible. On the other hand, if for given prices $c_i, i = 1, \dots, N$, there is no probability measure P that satisfies both Equation 17.2 and the equality

$$c_i = E_{\mathbf{P}}[e^{-\alpha t_i} (X(t_i) - K_i)^+], \quad i = 1, \dots, N$$

then a sure win is possible.

We will now present a probability measure P on the outcome $X(t), 0 \leqslant t \leqslant T$, that satisfies Equation 17.2.

Suppose that

$$X(t) = x_0 e^{Y(t)}$$

where $\{Y(t), t \geqslant 0\}$ is a Brownian motion process with drift coefficient μ and variance parameter σ^2. That is, $\{X(t), t \geqslant 0\}$ is a geometric Brownian motion process. From Equation 17.1 we have that, for $s < t$,

$$E[X(t) | X(u), 0 \leqslant u \leqslant s] = X(s) e^{(t-s)(\mu + \sigma^2/2)}$$

Hence, if we choose μ and σ^2 so that

$$\mu + \sigma^2/2 = \alpha$$

then Equation 17.2 will be satisfied. That is, by letting P be the probability measure governing the stochastic process $\{x_0 e^{Y(t)}, 0 \leqslant t \leqslant T\}$, where $\{Y(t)\}$ is Brownian motion with drift parameter μ and variance parameter σ^2, and where $\mu + \sigma^2/2 = \alpha$, Equation 17.2 is satisfied.

It follows from the preceding that if we price an option to purchase a share of the stock at time t for a fixed price K by

$$c = E_{\mathbf{P}}[e^{-\alpha t}(X(t) - K)^+]$$

then no arbitrage is possible. Since $X(t) = x_0 e^{Y(t)}$, where $Y(t)$ is normal with mean μt and variance $t\sigma^2$,we see that

$$\begin{aligned} ce^{\alpha t} &= \int_{-\infty}^{\infty} (x_0 e^y - K)^+ \frac{1}{\sqrt{2\pi t\sigma^2}} e^{-(y-\mu t)^2/2t\sigma^2}\, dy \\ &= \int_{\log(K/x_0)}^{\infty} (x_0 e^y - K) \frac{1}{\sqrt{2\pi t\sigma^2}} e^{-(y-\mu t)^2/2t\sigma^2}\, dy \end{aligned}$$

Making the change of variable $w = (y - \mu t)/(\sigma t^{1/2})$ yields

$$ce^{\alpha t} = x_0 e^{\mu t} \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{\sigma w\sqrt{t}} e^{-w^2/2}\, dw - K \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{-w^2/2}\, dw \tag{17.4}$$

where

$$a = \frac{\log(K/x_0) - \mu t}{\sigma\sqrt{t}}$$

Now,

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{\sigma w\sqrt{t}} e^{-w^2/2}\, dw &= e^{t\sigma^2/2} \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{-(w-\sigma\sqrt{t})^2/2}\, dw \\ &= e^{t\sigma^2/2} P\{N(\sigma\sqrt{t}, 1) \geqslant a\} \\ &= e^{t\sigma^2/2} P\{N(0,1) \geqslant a - \sigma\sqrt{t}\} \\ &= e^{t\sigma^2/2} P\{N(0,1) \leqslant -(a - \sigma\sqrt{t})\} \\ &= e^{t\sigma^2/2} \phi\left(\sigma\sqrt{t} - a\right) \end{aligned}$$

where $N(m, v)$ is a normal random variable with mean m and variance v, and ϕ is the standard normal distribution function.

Thus, we see from Equation 17.4 that

$$ce^{\alpha t} = x_0 e^{\mu t + \sigma^2 t/2} \phi(\sigma\sqrt{t} - a) - K\phi(-a)$$

Using that

$$\mu + \sigma^2/2 = \alpha$$

and letting $b = -a$, we can write this as follows:

$$c = x_0 \phi(\sigma\sqrt{t} + b) - Ke^{-\alpha t}\phi(b) \tag{17.5}$$

where

$$b = \frac{\alpha t - \sigma^2 t/2 - \log(K/x_0)}{\sigma\sqrt{t}}$$

The option price formula given by Equation 17.5 depends on the initial price of the stock x_0, the option exercise time t, the option exercise price K, the discount (or interest rate) factor α, and the value σ^2. Note that for any value of σ^2, if the options are priced according to the formula of Equation 17.5 then no arbitrage is possible. However, as many people believe that the price of a stock actually follows a geometric Brownian motion—that is, $X(t) = x_0 e^{Y(t)}$ where $Y(t)$ is Brownian motion with parameters μ and σ^2—it has been suggested that it is natural to price the option according to the formula Equation 17.5 with the parameter σ^2 taken equal to the estimated value (see the remark that follows) of the variance parameter under the assumption of a geometric Brownian motion model. When this is done, the formula Equation 17.5 is known as the Black-Scholes option cost valuation. It is interesting that this valuation does not depend on the value of the drift parameter μ but only on the variance parameter σ^2.

If the option itself can be traded, then the formula of Equation 17.5 can be used to set its price in such a way so that no arbitrage is possible. If at time s the price of the stock is $X(s) = x_s$, then the price of a (t, K) option—that is, an option to purchase one unit of the stock at time t for a price K—should be set by replacing t by $t - s$ and x_0 by x_s in Equation 17.5.

Remark 17.3. If we observe a Brownian motion process with variance parameter σ^2 over any time interval, then we could theoretically obtain an arbitrarily precise estimate of σ^2. For suppose we observe such a process $\{Y(s)\}$ for a time t. Then, for fixed h, let $N = [t/h]$ and set

$$\begin{aligned} &W_1 = Y(h) - Y(0), \\ &W_2 = Y(2h) - Y(h), \\ &\vdots \\ &W_N = Y(Nh) - Y(Nh - h) \end{aligned}$$

Then random variables $W_1, \dots, W_N$ are independent and identically distributed normal random variables having variance $h\sigma^2$. We now use the fact that $(N-1)S^2/(\sigma^2 h)$ has a chi-squared distribution with $N - 1$ degrees of freedom, where S^2 is the sample variance defined by

$$S^2 = \sum_{i=1}^{N} (W_i - \overline{W})^2/(N-1)$$

Since the expected value and variance of a chi-squared with k degrees of freedom are equal to k and $2k$, respectively, we see that

$$E[(N-1)S^2/(\sigma^2 h)] = N - 1$$

and

$$\text{Var}[(N-1)S^2/(\sigma^2 h)] = 2(N-1)$$

From this, we see that

$$E[S^2/h] = \sigma^2$$

and

$$\mathrm{Var}[S^2/h] = 2\sigma^4/(N-1)$$

Hence, as we let h become smaller (and so $N = [t/h]$ becomes larger) the variance of the unbiased estimator of σ^2 becomes arbitrarily small.

18 Introduction to Monte Carlo Simulation

Let $\mathbf{X} = (X_1, \dots, X_n)$ denote a random vector having a given density function $f(x_1, \dots, x_n)$ and suppose we are interested in computing

$$E[g(\mathbf{X})] = \int \int \cdots \int g(x_1, \dots, x_n) f(x_1, \dots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

for some n-dimensional function g. In many situations, it is not analytically possible either to compute the preceding multiple integral exactly or even to numerically approximate it within a given accuracy. One possibility that remains is to approximate $E[g(\mathbf{X})]$ by means of simulation.

To approximate $E[g(\mathbf{X})]$, start by generating a random vector $\mathbf{X}^{(1)} = (X_1^{(1)}, \dots, X_n^{(1)})$, having the joint density $f(x_1, \dots, x_n)$ and then compute $Y^{(1)} = g(\mathbf{X}^{(1)})$. Now generate a second random vector (independent of the first) $\mathbf{X}^{(2)}$ and compute $Y^{(2)} = g(\mathbf{X}^{(2)})$. Keep on doing this until r, a fixed number, of independent and identically distributed random variables $Y^{(i)} = g(\mathbf{X}^{(i)}), i = 1, \dots, r$ have been generated. Now by the strong law of large numbers, we know that

$$\lim_{r\to\infty} \frac{Y^{(1)} + \cdots + Y^{(r)}}{r} = E[Y^{(i)}] = E[g(\mathbf{X})]$$

and so we can use the average of the generated Ys as an estimate of $E[g(\mathbf{X})]$. This approach to estimating $E[g(\mathbf{X})]$ is called the *Monte Carlo simulation* approach. Clearly there remains the problem of how to generate, or *simulate*, random vectors having a specified joint distribution. The first step in doing this is to be able to generate random variables from a uniform distribution on $(0, 1)$. One way to do this would be to take 10 identical slips of paper, numbered $0, 1, \dots, 9$, place them in a hat and then successively select n slips, with replacement, from the hat. The sequence of digits obtained (with a decimal point in front) can be regarded as the value of a uniform $(0, 1)$ random variable rounded off to the nearest $\left(\frac{1}{10}\right)^n$. For instance, if the sequence of digits selected is $3, 8, 7, 2, 1$, then the value of the uniform $(0, 1)$ random variable is 0.38721 (to the nearest 0.00001). Tables of the values of uniform $(0, 1)$ random variables, known as random number tables, have been extensively published [for instance, see The RAND Corporation, *A Million Random Digits with* 100,000 *Normal Deviates* (New York: The Free Press, 1955)].

However, this is not the way in which digital computers simulate uniform $(0, 1)$ random variables. In practice, they use pseudo random numbers instead of truly random ones. Most random number generators start with an initial value X_0, called the seed, and then recursively compute values by specifying positive integers a, c, and m, and then letting

$$X_{n+1} = (aX_n + c) \text{ modulo } m, \quad n \geq 0$$

where the preceding means that $aX_n + c$ is divided by m and the remainder is taken as the value of X_{n+1}. Thus each X_n is either $0, 1, \ldots, m-1$ and the quantity X_n/m is taken as an approximation to a uniform $(0, 1)$ random variable. It can be shown that subject to suitable choices for a, c, m, the preceding gives rise to a sequence of numbers that looks as if it were generated from independent uniform $(0, 1)$ random variables.

As our starting point in the simulation of random variables from an arbitrary distribution, we shall suppose that we can simulate from the uniform $(0, 1)$ distribution, and we shall use the term "random numbers" to mean independent random variables from this distribution. Then we will discuss how to simulate general continuous random variables, discrete random variables, jointly distributed random variables, and stochastic processes. Let us first consider two applications of simulation to combinatorial problems.

Example 18.1. (Generating a Random Permutation) Suppose we are interested in generating a permutation of the numbers $1, 2, \ldots, n$ that is such that all $n!$ possible orderings are equally likely. The following algorithm will accomplish this by first choosing one of the numbers $1, \ldots, n$ at random and then putting that number in position n; it then chooses at random one of the remaining $n-1$ numbers and puts that number in position $n-1$; it then chooses at random one of the remaining $n-2$ numbers and puts it in position $n-2$, and so on (where choosing a number at random means that each of the remaining numbers is equally likely to be chosen). However, so that we do not have to consider exactly which of the numbers remain to be positioned, it is convenient and efficient to keep the numbers in an ordered list and then randomly choose the position of the number rather than the number itself. That is, starting with any initial ordering $p_1, p_2, \ldots, p_n$, we pick one of the positions $1, \ldots, n$ at random and then interchange the number in that position with the one in position n. Now we randomly choose one of the positions $1, \ldots, n-1$ and interchange the number in this position with the one in position $n-1$, and so on.

To implement the preceding, we need to be able to generate a random variable that is equally likely to take on any of the values $1, 2, \ldots, k$. To accomplish this, let U denote a random number—that is, U is uniformly distributed over $(0, 1)$— and note that kU is uniform on $(0, k)$ and so

$$P\{i - 1 < kU < i\} = \frac{1}{k}, \quad i = 1, \ldots, k$$

Hence, if the random variable $I = [kU] + 1$ will be such that

$$P\{I = i\} = P\{[kU] = i - 1\} = P\{i - 1 < kU < i\} = \frac{1}{k}$$

The preceding algorithm for generating a random permutation can now be written as follows:

Step 1: Let $p_1, p_2, \dots, p_n$ be any permutation of $1, 2, \dots, n$ (for instance, we can choose $p_j = j, j = 1, \dots, n$).

Step 2: Set $k = n$.

Step 3: Generate a random number U and let $I = [kU] + 1$.

Step 4: Interchange the values of p_I and p_k.

Step 5: Let $k = k - 1$ and if $k > 1$ go to Step 3.

Step 6: $p_1, \dots, p_n$ is the desired random permutation

For instance, suppose $n = 4$ and the initial permutation is $1, 2, 3, 4$. If the first value of I (which is equally likely to be either $1, 2, 3, 4$) is $I = 3$, then the new permutation is $1, 2, 4, 3$. If the next value of I is $I = 2$ then the new permutation is $1, 4, 2, 3$. If the final value of I is $I = 2$, then the final permutation is $1, 4, 2, 3$, and this is the value of the random permutation.

One very important property of the preceding algorithm is that it can also be used to generate a random subset, say of size r, of the integers $1, \dots, n$. Namely, just follow the algorithm until the positions $n, n-1, \dots, n-r+1$ are filled. The elements in these positions constitute the random subset.

The algorithm can be realized with Python as follows.

```python
import numpy as np

def int_perm(n):
    """
    Generate a permutation of integers 1,2,...,n

    Input:
    n: int, the largest integers

    Output:
    perm: 1d array of size n, the array of permuted integers of 1,2,...,n
    """

    perm = np.arange(1, n+1, dtype='int')
    k = n
    while k > 1:
        U = np.random.random()
        I = int(k*U)  # No need to add 1 since Python indices start from 0
        perm[[I, k-1]] = perm[[k-1, I]]  # exchange the int at pos I and pos k-1, noting k s
        k -= 1
    return perm
```

```
n = 10
print('A permutation of integers from ', 1 , ' to ', n, ' is: ', int_perm(n))
print('Another permutation of integers from ', 1 , ' to ', n, ' is: ', int_perm(n))
```

Example 18.2. (Estimating the Number of Distinct Entries in a Large List) Consider a list of n entries where n is very large, and suppose we are interested in estimating d, the number of distinct elements in the list. If we let m_i denote the number of times that the element in position i appears on the list, then we can express d by

$$d = \sum_{i=1}^{n} \frac{1}{m_i}$$

To estimate d, suppose that we generate a random value X equally likely to be either $1, 2, \dots, n$ (that is, we take $X = [nU] + 1$) and then let $m(X)$ denote the number of times the element in position X appears on the list. Then

$$E\left[\frac{1}{m(X)}\right] = \sum_{i=1}^{n} \frac{1}{m_i}\frac{1}{n} = \frac{d}{n}$$

Hence, if we generate k such random variables $X_1, \dots, X_k$ we can estimate d by

$$d \approx \frac{n\sum_{i=1}^{k} 1/m(X_i)}{k}$$

Suppose now that each item in the list has a value attached to it—$v(i)$ being the value of the ith element. The sum of the values of the distinct items—call it v— can be expressed as

$$v = \sum_{i=1}^{n} \frac{v(i)}{m(i)}$$

Now if $X = [nU] + 1$, where U is a random number, then

$$E\left[\frac{v(X)}{m(X)}\right] = \sum_{i=1}^{n} \frac{v(i)}{m(i)}\frac{1}{n} = \frac{v}{n}$$

Hence, we can estimate v by generating $X_1, \dots, X_k$ and then estimating v by

$$v \approx \frac{n}{k}\sum_{i=1}^{k} \frac{v(X_i)}{m(X_i)}$$

19 General Techniques for Simulating Continuous Random Variables

19.1 The Inverse Transformation Method

A general method for simulating a random variable having a continuous distribution— called the *inverse transformation method*—is based on the following proposition.

Proposition 19.1. *Let U be a uniform $(0, 1)$ random variable. For any continuous distribution function F if we define the random variable X by*

$$X = F^{-1}(U)$$

then the random variable X has distribution function F. [$F^{-1}(u)$ is defined to equal that value x for which $F(x) = u$.]

Proof.

$$F_X(a) = P\{X \leq a\} = P\{F^{-1}(U) \leq a\} \tag{19.1}$$

Now, since $F(x)$ is a monotone function, it follows that $F^{-1}(U) \leq a$ if and only if $U \leq F(a)$. Hence, from Equation 19.1, we see that

$$F_X(a) = P\{U \leq F(a)\} = F(a)$$

□

Hence we can simulate a random variable X from the continuous distribution F, when F^{-1} is computable, by simulating a random number U and then setting $X = F^{-1}(U)$.

Example 19.1. (Simulating an Exponential Random Variable) If $F(x) = 1-e^{-x}$,then $F^{-1}(u)$ is that value of x such that

$$1 - e^{-x} = u$$

or

$$x = -\log(1 - u)$$

Hence, if U is a uniform $(0, 1)$ variable, then

$$F^{-1}(U) = -\log(1 - U)$$

is exponentially distributed with mean 1. Since $1 - U$ is also uniformly distributed on $(0, 1)$ it follows that $-\log U$ is exponential with mean 1. Since cX is exponential with mean c when X is exponential with mean 1, it follows that $-c\log U$ is exponential with mean c.

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import expon

def exp_invtran(sample_size):
    """
    Simulate an array of exponential variables (lambda=1) using inverse transform

    Input
    sample_size: int, number of exponential variables to generate

    Ouput
    exp_rvs: 1d array of shape (sample_size, ), the array of generated exponential variables
    """

    U = np.random.rand(sample_size)
    exp_rvs = -np.log(1-U)

    return exp_rvs
```

```python
sample_size = 10000
exp_rvs = exp_invtran(sample_size)

# density plot for the samples
plt.hist(exp_rvs, bins=50, density=True, color='b', label='hist')

# theoretical pdf
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = expon.pdf(x)
plt.plot(x, p, 'r', linewidth=2, label='density')

plt.xlabel('x')
plt.ylabel('Density')
plt.title('Histogram and PDF of Exponential Distribution')
plt.legend()

plt.show()
```

19.2 The Rejection Method

Suppose that we have a method for simulating a random variable having density function $g(x)$. We can use this as the basis for simulating from the continuous distribution having density $f(x)$ by simulating Y from g and then accepting this simulated value with a probability proportional to $f(Y)/g(Y)$. Specifically let c be a constant such that

$$\frac{f(y)}{g(y)} \leqslant c \quad \text{for all y}$$

We then have the following technique for simulating a random variable having density f.

Rejection Method

Step 1: Simulate Y having density g and simulate a random number U.

Step 2: If $U \leqslant f(Y)/cg(Y)$ set $X = Y$. Otherwise return to Step 1.

Proposition 19.2. *The random variable X generated by the rejection method has density function f.*

Proof. Let X be the value obtained, and let N denote the number of necessary iterations. Then

$$\begin{aligned} P\{X \leqslant x\} &= P\{Y_N \leqslant x\} \\ &= P\{Y \leqslant x | U \leqslant f(Y)/cg(Y)\} \\ &= \frac{P\{Y \leqslant x, U \leqslant f(Y)/cg(Y)\}}{K} \\ &= \frac{\int P\{Y \leqslant x, U \leqslant f(Y)/cg(Y) | Y = y\} g(y)\, dy}{K} \\ &= \frac{\int_{-\infty}^{x} (f(y)/cg(y)) g(y)\, dy}{K} \\ &= \frac{\int_{-\infty}^{x} f(y)\, dy}{Kc} \end{aligned}$$

where $K = P\{U \leqslant f(Y)/cg(Y)\}$. Letting $x \to \infty$ shows that $K = 1/c$ and the proof is complete. □

Example 19.2. Let us use the rejection method to generate a random variable having density function

$$f(x) = 20x(1-x)^3, \quad 0 < x < 1$$

Since this random variable (which is beta with parameters $2, 4$) is concentrated in the interval $(0, 1)$, let us consider the rejection method with

$$g(x) = 1, \quad 0 < x < 1$$

To determine the constant c such that $f(x)/g(x) \leqslant c$, we use calculus to determine the maximum value of

$$\frac{f(x)}{g(x)} = 20x(1-x)^3$$

Differentiation of this quantity yields

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = 20[(1-x)^3 - 3x(1-x)^2]$$

Setting this equal to 0 shows that the maximal value is attained when $x = \frac{1}{4}$, and thus

$$\frac{f(x)}{g(x)} \leqslant 20\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^3 = \frac{135}{64} \equiv c$$

Hence,

$$\frac{f(x)}{cg(x)} = \frac{256}{27}\, x(1-x)^3$$

and thus the rejection procedure is as follows:

Step 1: Generate random numbers U_1 and U_2.

Step 2: If $U_2 \leqslant \frac{256}{27} U_1 (1 - U_1)^3$, stop and set $X = U_1$. Otherwise return to Step 1.

The average number of times that step 1 will be performed is $c = \frac{135}{64}$.

```
sample_size = 10000  # number of samples to generate
count = 0
samples = np.zeros(sample_size)
while count < sample_size:
    U1 = np.random.random()  # generate a sample from distribution g, in this case, uniform U
    U2 = np.random.random()
    if U2 <= 256/27*U1*(1-U1)**3:
        samples[count] = U1
        count +=1
```

```
from scipy.stats import beta

# density plot for the samples
plt.hist(samples, bins=50, density=True, color='b', label='hist')

# theoretical pdf
a = 2
b = 4
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = beta.pdf(x, a, b)
plt.plot(x, p, 'r', linewidth=2, label='density')

plt.xlabel('x')
plt.ylabel('Density')
plt.title('Histogram and PDF of beta(2,4) Distribution')
plt.legend()

plt.show()
```

Example 19.3. (Simulating a Normal Random Variable) To simulate a standard normal random variable Z (that is, one with mean 0 and variance 1) note first that the absolute value of Z has density function

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad 0 < x < \infty \tag{19.2}$$

We will start by simulating from the preceding density by using the rejection method with

$$g(x) = e^{-x}, \quad 0 < x < \infty$$

Now, note that

$$\frac{f(x)}{g(x)} = \sqrt{2e/\pi} \exp\{-(x-1)^2/2\} \leqslant \sqrt{2e/\pi}$$

Hence, using the rejection method we can simulate from Equation 19.2 as follows:

(a) Generate independent random variables Y and U, Y being exponential with rate 1 and U being uniform on (0,1).

(b) If $U \leqslant \exp\{-(Y-1)^2/2\}$, or equivalently, if

$$-\log U \geqslant (Y-1)^2/2$$

set $X = Y$. Otherwise return to step (a).

Once we have simulated a random variable X having density function Equation 19.2 we can then generate a standard normal random variable Z by letting Z be equally likely to be either X or $-X$.

To improve upon the foregoing, note first that from Example 25.1 it follows that $-\log U$ will also be exponential with rate 1. Hence, steps (a) and (b) are equivalent to the following:

(a') Generate independent exponentials with rate 1, Y_1, and Y_2.

(b') Set $X = Y_1$ if $Y_2 \geqslant (Y_1 - 1)^2/2$. Otherwise return to (a').

Now suppose that we accept step (b'). It then follows by the lack of memory property of the exponential that the amount by which Y_2 exceeds $(Y_1 - 1)^2/2$ will also be exponential with rate 1.

Hence, summing up, we have the following algorithm which generates an exponential with rate 1 and an independent standard normal random variable.

Step 1: Generate Y_1, an exponential random variable with rate 1.

Step 2: Generate Y_2, an exponential with rate 1.

Step 3: If $Y_2 - (Y_1 - 1)^2/2 > 0$, set $Y = Y_2 - (Y_1 - 1)^2/2$ and go to step 4. Otherwise go to step 1.

Step 4: Generate a random number U and set

$$Z = \begin{cases} Y_1, & \text{if } U \leqslant \frac{1}{2} \\ -Y_1, & \text{if } U > \frac{1}{2} \end{cases}$$

The random variables Z and Y generated by the preceding are independent with Z being normal with mean 0 and variance 1 and Y being exponential with rate 1. (If we want the normal random variable to have mean μ and variance σ^2, just take $\mu + \sigma z$.)

20 Special Techniques for Simulating Continuous Random Variables

Special techniques have been devised to simulate from most of the common continuous distributions. We now present some of these.

20.1 The Normal Distribution (Box-Muller Method)

Let X and Y denote independent standard normal random variables and thus have the joint density function

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}, \quad -\infty < x < \infty, -\infty < y < \infty$$

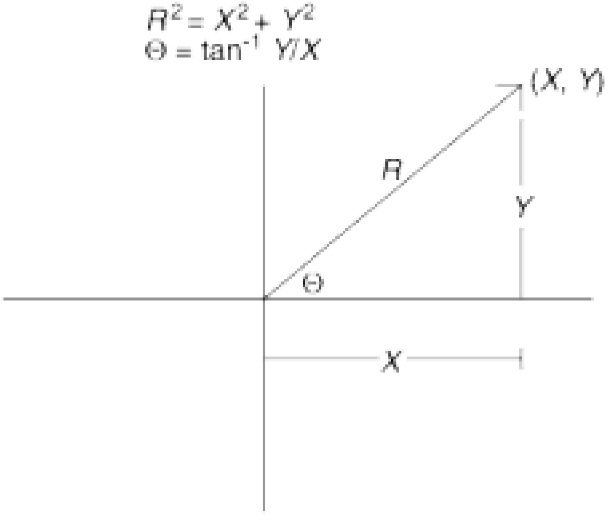


Figure 20.1: An elephant

Consider now the polar coordinates of the point (X, Y). As shown in Figure 20.1,

$$R^2 = X^2 + Y^2,$$
$$\Theta = \tan^{-1} Y/X$$

To obtain the joint density of R^2 and Θ, consider the transformation

$$d = x^2 + y^2, \quad \theta = \tan^{-1}(y/x)$$

The Jacobian of this transformation is

$$\begin{aligned} \mathrm{J} = \begin{vmatrix} \dfrac{\partial d}{\partial x} & \dfrac{\partial d}{\partial y} \\ \dfrac{\partial \theta}{\partial x} & \dfrac{\partial \theta}{\partial y} \end{vmatrix} &= \begin{vmatrix} 2x & 2y \\ \dfrac{1}{1+y^2/x^2}\left(\dfrac{-y}{x^2}\right) & \dfrac{1}{1+y^2/x^2}\left(\dfrac{1}{x}\right) \end{vmatrix} \\ &= 2\begin{vmatrix} x & y \\ -\dfrac{y}{x^2+y^2} & \dfrac{x}{x^2+y^2} \end{vmatrix} = 2 \end{aligned}$$

Hence, the joint density of R^2 and Θ is given by

$$\begin{aligned} f_{R^2,\Theta}(d,\theta) &= \frac{1}{2\pi} e^{-d/2} \frac{1}{2} \\ &= \frac{1}{2} e^{-d/2} \frac{1}{2\pi}, \quad 0 < d < \infty, 0 < \theta < 2\pi \end{aligned}$$

Thus, we can conclude that R^2 and Θ are independent with R^2 having an exponential distribution with rate $\frac{1}{2}$ and Θ being uniform on $(0, 2\pi)$.

Let us now go in reverse from the polar to the rectangular coordinates. From the preceding if we start with W, an exponential random variable with rate $\frac{1}{2}$ (W plays the role of R^2) and with V, independent of W and uniformly distributed over $(0, 2\pi)$ (V plays the role of Θ) then $X = \sqrt{W}\cos V, Y = \sqrt{W}\sin V$ will be independent standard normals. Hence using the results of Example 1 Lecture 19 we see that if U_1 and U_2 are independent uniform $(0,1)$ random numbers, then

$$\begin{aligned} X &= (-2\log\, U_1)^{1/2}\cos(2\pi U_2), \\ Y &= (-2\log\, U_1)^{1/2}\sin(2\pi U_2) \end{aligned} \tag{20.1}$$

are independent standard normal random variables.

The preceding approach to generating standard normal random variables is called the *Box-Muller* approach. Its efficiency suffers somewhat from its need to compute the preceding sine and cosine values. There is, however, a way to get around this potentially time-consuming difficulty. To begin, note that if U is uniform on $(0,1)$, then $2U$ is uniform on $(0,2)$, and so $2U - 1$ is uniform on $(-1,1)$.

Thus, if we generate random numbers U_1 and U_2 and set

$$\begin{aligned} V_1 &= 2U_1 - 1, \\ V_2 &= 2U_2 - 1 \end{aligned}$$

then (V_1, V_2) is uniformly distributed in the square of area 4 centered at $(0,0)$ (see Figure 20.2).

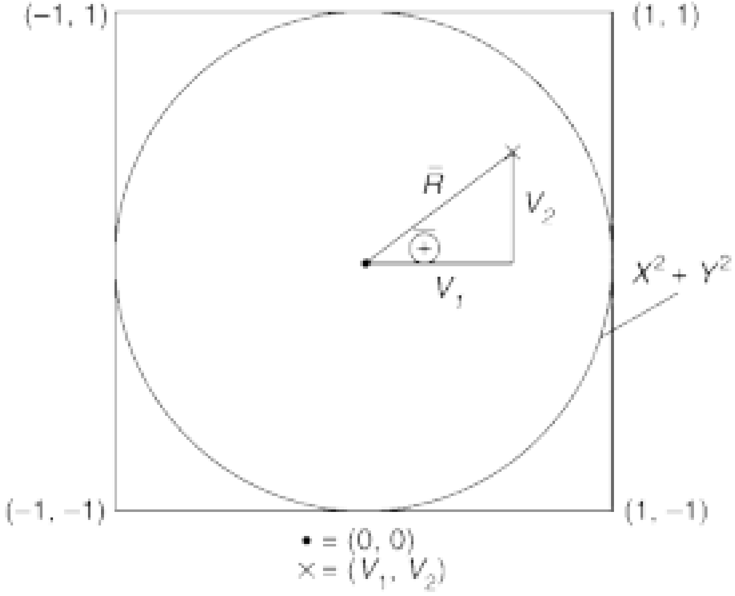


Figure 20.2

Suppose now that we continually generate such pairs (V_1, V_2) until we obtain one that is contained in the circle of radius 1 centered at $(0, 0)$—that is, until (V_1, V_2) is such that $V_1^2 + V_2^2 \leqslant 1$. It now follows that such a pair (V_1, V_2) is uniformly distributed in the circle. If we let $\bar{R}, \bar{\Theta}$ denote the polar coordinates of this pair, then it is easy to verify that $\bar{R}$ and $\bar{\Theta}$ are independent, with $\bar{R}^2$ being uniformly distributed on $(0, 1)$, and $\bar{\Theta}$ uniformly distributed on $(0, 2\pi)$.

Since

$$\sin\bar{\Theta} = V_2/\bar{R} = \frac{V_2}{\sqrt{V_1^2 + V_2^2}},$$
$$\cos\bar{\Theta} = V_1/\bar{R} = \frac{V_1}{\sqrt{V_1^2 + V_2^2}}$$

it follows from Equation 20.1 that we can generate independent standard normals X and Y by generating another random number U and setting

$$X = (-2\log U)^{1/2} V_1/\bar{R},$$
$$Y = (-2\log U)^{1/2} V_2/\bar{R}$$

In fact, since (conditional on $V_1^2 + V_2^2 \leqslant 1$) $\bar{R}^2$ is uniform on $(0, 1)$ and is independent of $\bar{\Theta}$, we can use it instead of generating a new random number U; thus showing that

$$X = (-2\log \bar{R}^2)^{1/2} V_1/\bar{R} = \sqrt{\frac{-2\log S}{S}} V_1,$$
$$Y = (-2\log \bar{R}^2)^{1/2} V_2/\bar{R} = \sqrt{\frac{-2\log S}{S}} V_2$$

are independent standard normals, where

$$S = \bar{R}^2 = V_1^2 + V_2^2$$

Summing up, we thus have the following approach to generating a pair of independent standard normals:

Step 1: Generate random numbers U_1 and U_2.

Step 2: Set $V_1 = 2U_1 - 1$, $V_2 = 2U_2 - 1$, $S = V_1^2 + V_2^2$.

Step3: If $S > 1$, return to step 1

Step 4: Return the independent unit normals

$$X = \sqrt{\frac{-2\log S}{S}} V_1, \quad Y = \sqrt{\frac{-2\log S}{S}} V_2$$

The preceding is called the *polar method*. Since the probability that a random point in the square will fall within the circle is equal to $\pi/4$ (the area of the circle divided by the area of the square), it follows that, on average, the polar method will require $4/\pi = 1.273$ iterations of step 1. Hence, it will, on average, require 2.546 random numbers, 1 logarithm, 1 square root, 1 division, and 4.546 multiplications to generate 2 independent standard normals.

```python
import numpy as np

def box_muller(sample_size):
    """
    Generate standard normal variables with the box-muller method
    input:
    sample_size: int, the sample size (should be an even number)

    output:
    samples: 1d array of size sample_size, the generated standard normal variables
    """
    samples = np.zeros(sample_size)
    for i in range(int(sample_size/2)):
        (U1, U2) = np.random.rand(2)
        X1 = np.sqrt(-2*np.log(U1))*np.cos(2*np.pi*U2)
        X2 = np.sqrt(-2*np.log(U1))*np.sin(2*np.pi*U2)
        samples[2*i:2*i+1+1] = [X1, X2]
    return samples
```

```
from scipy.stats import norm
import matplotlib.pyplot as plt

sample_size = 10000
samples = box_muller(sample_size)

# density plot for the samples
plt.hist(samples, bins=50, density=True, color='b', label='hist')

# theoretical pdf
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x)
plt.plot(x, p, 'r', linewidth=2, label='density')

plt.xlabel('x')
plt.ylabel('Density')
plt.title('Histogram and PDF of standard normal Distribution')
plt.legend()

plt.show()
```

20.2 The Gamma Distribution

To simulate from a gamma distribution with parameters (n, λ), where n is an integer, we use the fact that the sum of n independent exponential random variables each having rate λ has this distribution. Hence, if $U_1, \dots, U_n$ are independent uniform (0,1) random variables,

$$X = \frac{1}{\lambda}\sum_{i=1}^{n} \log\, U_i = -\frac{1}{\lambda}\, \log\left(\prod_{i=1}^{n} U_i\right)$$

has the desired distribution.

20.3 The Chi-Squared Distribution

The chi-squared distribution with n degrees of freedom is the distribution of $\chi_n^2 = Z_1^2 + \cdots + Z_n^2$ where $Z_i, i = 1, \ldots, n$ are independent standard normals. Using the fact that $Z_1^2 + Z_2^2$ has an exponential distribution with rate $\frac{1}{2}$. Hence, when n is even—say $n = 2k$—χ_{2k}^2 has a gamma distribution with parameters $(k, \frac{1}{2})$. Hence, $-2\log(\prod_{i=1}^k U_i)$ has a chi-squared distribution with $2k$ degrees of freedom. We can simulate a chisquared random variable with $2k+1$ degrees of freedom by first simulating a standard normal random variable Z and then adding Z^2 to the preceding. That is,

$$\chi_{2k+1}^2 = Z^2 - 2\log\left(\prod_{i=1}^k U_i\right)$$

where $Z, U_1, \ldots, U_n$ are independent with Z being a standard normal and the others being uniform $(0, 1)$ random variables.

21 Simulating from Discrete Distributions

All of the general methods for simulating from continuous distributions have analogs in the discrete case. For instance, if we want to simulate a random variable X having probability mass function

$$P\{X = x_j\} = P_j, \quad j = 1, 2, \dots, \quad \sum_j P_j = 1$$

We can use the following discrete time analog of the inverse transform technique.

To simulate X for which $P\{X = x_j\} = P_j$

let U be uniformly distributed over $(0, 1)$, and set

$$X = \begin{cases} x_1, & \text{if } U < P_1 \\ x_2, & \text{if } P_1 < U < P_1 + P_2 \\ \vdots & \\ x_j, & \text{if } \sum_1^{j-1} P_i < U < \sum_i^j P_i \\ \vdots & \end{cases}$$

As,

$$P\{X = x_j\} = P\left\{\sum_1^{j-1} P_i < U < \sum_1^j P_i\right\} = P_j$$

we see that X has the desired distribution.

Example 21.1. (The Geometric Distribution) Suppose we want to simulate X such that

$$P\{X = i\} = p(1-p)^{i-1}, \quad i \geqslant 1$$

As

$$\sum_{i=1}^{j-1} P\{X = i\} = 1 - P\{X > j-1\} = 1 - (1-p)^{j-1}$$

we can simulate such a random variable by generating a random number U and then setting X equal to that value j for which

$$1 - (1-p)^{j-1} < U < 1 - (1-p)^j$$

or, equivalently, for which

$$(1-p)^j < 1-U < (1-p)^{j-1}$$

As $1-U$ has the same distribution as U,we can thus define X by

$$\begin{aligned} \mathrm{X} &= \min\{j\colon (1-p)^j < U\} = \min\left\{j\colon j > \frac{\log U}{\log(1-p)}\right\} \\ &= 1 + \left[\frac{\log U}{\log(1-p)}\right] \end{aligned}$$

As in the continuous case, special simulation techniques have been developed for the more common discrete distributions. We now present some of these.

```
import numpy as np
from scipy.stats import geom
import matplotlib.pyplot as plt

def generate_geometric(sample_size, p):
    """
    Simulate an array of geometric variables with hyperparameter p

    Input
    sample_size: int, number of geometric variables to generate
    p: the hyperparameter p for the geometric distribution

    Ouput
    geo_rvs: 1d array of shape (sample_size, ), the array of generated geometric variables
    """
    U = np.random.rand(sample_size)
    geo_rvs = 1 + np.floor(np.log(U)/np.log(1-p))
    return geo_rvs
```

```
sample_size = 100000
p = 0.5
geo_samples = generate_geometric(sample_size, p)

fig, ax = plt.subplots(1, 1)
p = 0.5
x = np.arange(1, 7)
ax.plot(x, geom.pmf(x, p), 'bo', ms=8, label='geom pmf')
ax.vlines(x, 0, geom.pmf(x, p), colors='b', lw=5, alpha=0.5);

ax.hist(geo_samples, bins=x, density=True, color='g', label='hist')
ax.legend()
ax.set_xlabel('i')
ax.set_ylabel('probability')
```

Example 21.2. (Simulating a Binomial Random Variable) A binomial (n, p) random variable can be most easily simulated by recalling that it can be expressed as the sum of n independent Bernoulli random variables.That is, if $U_1, \ldots, U_n$ are independent uniform (0,1) variables, then letting

$$X_i = \begin{cases} 1, & \text{if } U_i < p \\ 0, & \text{otherwise} \end{cases}$$

it follows that $X \equiv \sum_{i=1}^{n} X_i$ is a binomial random variable with parameters n and p.

One difficulty with this procedure is that it requires the generation of n random numbers. To show how to reduce the number of random numbers needed, note first that this procedure does not use the actual value of a random number U but only whether or not it exceeds p. Using this and the result that the conditional distribution of U given that $U < p$ is uniform on $(0, p)$ and the conditional distribution of U given that $U > p$ is uniform on $(p, 1)$,we now show how we can simulate a binomial (n, p) random variable using only a single random number:

Step 1: Let $\alpha = 1/p, \beta = 1/(1 - p)$.

Step2: Set $k = 0$.

Step 3: Generate a uniform random number U.

Step4: If $k = n$ stop. Otherwise reset k to equal $k + 1$.

Step 5: If $U \leqslant p$ set $X_k = 1$ and reset U to equal αU. If $U > p$ set $X_k = 0$ and reset U to equal $\beta(U - p)$. Return to step 4.

This procedure generates $X_1, \ldots, X_n$ and $X = \sum_{i=1}^{n} X_i$ is the desired random variable. It works by noting whether $U_k \leqslant p$ or $U_k > p$; in the former case it takes U_{k+1} to equal U_k/p, and in the latter case it takes U_{k+1} to equal $(U_k - p)/\ (1 - p)$.

Example 21.3. (Simulating a Poisson Random Variable) To simulate a Poisson random variable with mean λ, generate independent uniform $(0, 1)$ random variables $U_1, U_2, \ldots$ stopping at

$$N + 1 = \min \left\{ n \colon \prod_{i=1}^{n} U_i < e^{-\lambda} \right\}$$

The random variable N has the desired distribution, which can be seen by noting that

$$N = \max \left\{ n \colon \sum_{i=1}^{n} -\log U_i < \lambda \right\}$$

But $-\log U_i$ is exponential with rate 1, and so if we interpret $-\log U_i, i \geqslant 1$, as the interarrival times of a Poisson process having rate 1, we see that $N = N(\lambda)$ would equal the number of events by time λ. Hence N is Poisson with mean λ.

When λ is large we can reduce the amount of computation in the preceding simulation of $N(\lambda)$, the number of events by time λ of a Poisson process having rate 1, by first choosing an integer m and simulating S_m, the time of the mth event of the Poisson process and then simulating $N(\lambda)$ according to the conditional distribution of $N(\lambda)$ given S_m. Now the conditional distribution of $N(\lambda)$ given S_m is as follows:

$$\begin{aligned} N(\lambda)|S_m = s &\sim m + \text{Poisson}(\lambda - s) \qquad \text{if } s < \lambda \\ N(\lambda)|S_m = s &\sim \text{Binomial}\left(m-1, \tfrac{\lambda}{s}\right) \qquad \text{if } s > \lambda \end{aligned}$$

where~ means "has the distribution of." This follows since if the mth event occurs at time s, where $s < \lambda$, then the number of events by time λ is m plus the number of events in (s, λ). On the other hand given that $S_m = s$ the set of times at which the first $m-1$ events occur has the same distribution as a set of $m-1$ uniform $(0, s)$ random variables. Hence, when $\lambda < s$, the number of these which occur by time λ is binomial with parameters $m-1$ and λ/s. Hence, we can simulate $N(\lambda)$ by first simulating S_m and then simulate either $P(\lambda - S_m)$, a Poisson random variable with mean $\lambda - S_m$ when $S_m < \lambda$, or simulate $\text{Bin}(m-1, \lambda/S_m)$, a binomial random variable with parameters $m-1$, and λ/S_m, when $S_m > \lambda$; and then setting

$$N(\lambda) = \begin{cases} m + P(\lambda - S_m), & \text{if } S_m < \lambda \\ \text{Bin}(m-1, \lambda/S_m), & \text{if } S_m > \lambda \end{cases}$$

In the preceding it has been found computationally effective to let m be approximately $\frac{7}{8}\lambda$. Of course, S_m is simulated by simulating from a gamma(m, λ) distribution via an approach that is computationally fast when m is large.

22 Simulating Stochastic Processes

We can easily simulate a stochastic process by simulating a sequence of random variables. For instance, We have seen that a Poisson process is a counting process for which the times between successive events are independent and identically distributed exponential random variables. One possible generalization is to consider a counting process for which the times between successive events are independent and identically distributed with an arbitrary distribution. Such a counting process is called a *renewal process*. Let $\{N(t), t \geq 0\}$ be a counting process and let X_n denote the time between the $(n-1)$st and the nth event of this process, $n \geq 1$.

Definition 22.1. If the sequence of nonnegative random variables $\{X_1 X_2, \dots\}$ is independent and identically distributed, then the counting process $\{N(t), t \geq 0\}$ is said to be a renewal process.

To simulate the first t time units of a renewal process having interarrival distribution F we can simulate independent random variables $X_1, X_2, \dots$ having distribution F, stopping at

$$N = \min\{n : X_1 + \cdots + X_n > t\}$$

The $X_i, i \geq 1$, represent the interarrival times of the renewal process and so the preceding simulation yields $N-1$ events by time t—the events occurring at times $X_1, X_1 + X_2, \dots, X_1 + \cdots + X_{N-1}$.

Actually there is another approach for simulating a Poisson process that is quite efficient. Suppose we want to simulate the first t time units of a Poisson process having rate λ. To do so, we can first simulate $N(t)$, the number of events by t , and then use the result that given the value of $N(t)$, the set of $N(t)$ event times is distributed as a set of n independent uniform $(0, t)$ random variables. Hence, we start by simulating $N(t)$, a Poisson random variable with mean λt (by one of the methods given in Example 2 Lecture 21). Then, if $N(t) = n$, generate a new set of n random numbers—call them $U_1, \dots, U_n$—and $tU_1, \dots, tU_n$ will represent the set of $N(t)$ event times. If we could stop here this would be much more efficient than simulating the exponentially distributed interarrival times. However, we usually desire the event times in increasing order—for instance, for $s < t$,

$$N(s) = \text{ number of } U_i : tU_i \leq s$$

and so to compute the function $N(s), s \leq t$, it is best to first order the values $U_i, i = 1, \dots, n$ before multiplying by t. However, in doing so you should not use an all-purpose sorting algorithm, such as quick sort, but rather one that takes into account that the elements to be sorted come from a uniform $(0, 1)$ population. Such a sorting algorithm, of n uniform $(0, 1)$ variables, is as follows: Rather than a single list to be sorted of length n we will consider n ordered, or linked, lists of random size. The value U will be put in list i if its value is between $(i-1)/n$ and i/n—that is, U is put in list $[nU]+1$. The individual lists are then ordered, and the total linkage of all the lists is the desired ordering. As almost all of the n lists will be of relatively small size [for instance, if $n = 1000$ the mean number of lists of size greater than 4 is (using the Poisson approximation to the binomial) approximately equal to $1000(1-\frac{65}{24}e^{-1}) \approx 4$] the sorting of individual lists will be quite quick, and so the running time of such an algorithm will be proportional to n (rather than to $n \log n$ as in the best all-purpose sorting algorithms).

An extremely important counting process for modeling purposes is the nonhomogeneous Poisson process, which relaxes the Poisson process assumption of stationary increments. Thus it allows for the possibility that the arrival rate need not be constant but can vary with time. However, there are few analytical studies that assume a nonhomogeneous Poisson arrival process for the simple reason that such models are not usually mathematically tractable. (For example, there is no known expression for the average customer delay in the single-server exponential service distribution queueing model which assumes a nonhomogeneous arrival process.) Clearly such models are strong candidates for simulation studies.

22.1 Simulating a Nonhomogeneous Poisson Process

We now present three methods for simulating a nonhomogeneous Poisson process having intensity function $\lambda(t), 0 \leq t < \infty$.

22.1.1 Conditional Distribution of the Arrival Times

Recall the result for a Poisson process having rate λ that given the number of events by time T the set of event times are independent and identically distributed uniform $(0, T)$ random variables. Now suppose that each of these events is independently counted with a probability that is equal to $\lambda(t)/\lambda$ when the event occurred at time t. Hence, given the number of counted events, it follows that the set of times of these counted events are independent with a common distribution given by $F(s)$, where

$$\begin{aligned} F(s) &= P\{\text{time} \leqslant s|\text{counted}\} \\ &= \frac{P\{\text{time} \leqslant s, \text{counted}\}}{P\{\text{counted}\}} \\ &= \frac{\int_0^T P\{\text{time} \leqslant s, \text{counted}|\text{time} = x\}\, dx/T}{P\{\text{counted}\}} \\ &= \frac{\int_0^s \lambda(x)\, dx}{\int_0^T \lambda(x)\, dx} \end{aligned}$$

The preceding (somewhat heuristic) argument thus shows that given n events of a nonhomogeneous Poisson process by time T the n event times are independent with a common density function

$$f(s) = \frac{\lambda(s)}{m(T)}, \quad 0 < s < T, \quad m(T) = \int_0^T \lambda(s)\, ds \tag{22.1}$$

Since $N(T)$, the number of events by time T, is Poisson distributed with mean $m(T)$, we can simulate the nonhomogeneous Poisson process by first simulating $N(T)$ and then simulating $N(T)$ random variables from the density Equation 22.1.

Example 22.1. If $\lambda(s) = cs$, then we can simulate the first T time units of the nonhomogeneous Poisson process by first simulating $N(T)$, a Poisson random variable having mean $m(T) = \int_0^T cs\; ds = CT^2/2$, and then simulating $N(T)$ random variables having distribution

$$F(s) = \frac{s^2}{T^2}, \quad 0 < s < T$$

Random variables having the preceding distribution either can be simulated by use of the inverse transform method (since $F^{-1}(U) = T\sqrt{U}$) or by noting that F is the distribution function of $\max(TU_1, TU_2)$ when U_1 and U_2 are independent random numbers.

If the distribution function specified by Equation 22.1 is not easily invertible, we can always simulate from Equation 22.1 by using the rejection method where we either accept or reject simulated values of uniform $(0, T)$ random variables. That is, let $h(s) = 1/T, 0 < s < T$. Then

$$\frac{f(s)}{h(s)} = \frac{T\lambda(s)}{m(T)} \leqslant \frac{\lambda T}{m(T)} \equiv C$$

where λ is a bound on $\lambda(s), 0 \leqslant s \leqslant T$. Hence, the rejection method is to generate random numbers U_1 and U_2 then accept TU_1 if

$$U_2 \leqslant \frac{f(TU_1)}{Ch(TU_1)}$$

or, equivalently, if

$$U_2 \leqslant \frac{\lambda(TU_1)}{\lambda}$$

22.2 Simulating a Two-Dimensional Poisson Process

A point process consisting of randomly occurring points in the plane is said to be a two-dimensional Poisson process having rate λ if

(a) the number of points in any given region of area A is Poisson distributed with mean λA; and

(b) the numbers of points in disjoint regions are independent.

For a given fixed point **O** in the plane, we now show how to simulate events occurring according to a two-dimensional Poisson process with rate λ in a circular region of radius r centered about **O**. Let $R_i, i \geqslant 1$, denote the distance between **O** and its ith nearest Poisson point, and let $C(a)$ denote the circle of radius a centered at **O**. Then

$$P\{\pi R_1^2 > b\} = P\left\{R_1 > \sqrt{\frac{b}{\pi}}\right\} = P\{\text{no points in } C(\sqrt{b/\pi})\} = e^{-\lambda b}$$

Also, with $C(a_2) - C(a_1)$ denoting the region between $C(a_2)$ and $C(a_1)$;

$$\begin{aligned}
&P\{\pi R_2^2 - \pi R_1^2 > b | R_1 = r\} \\
&= P\left\{R_2 > \sqrt{(b + \pi r^2)/\pi} | R_1 = r\right\} \\
&= P\left\{\text{no points in } C\left(\sqrt{(b + \pi r^2)/\pi}\right) - C(r) | R_1 = r\right\} \\
&= P\left\{\text{no points in } C\left(\sqrt{(b + \pi r^2)/\pi}\right) - C(r)\right\} \quad \text{by (b)} \\
&= e^{-\lambda b}
\end{aligned}$$

In fact, the same argument can be repeated to obtain the following

Proposition 22.1. *With $R_0 = 0$,*

$$\pi R_i^2 - \pi R_{i-1}^2, \quad i \geqslant 1,$$

are independent exponentials with rate λ.

```
import numpy as np
from scipy.stats import expon

def Poisson_circle(lam, r):
    """
    Generate a 2D Poisson process for a circular region

    Inputs:
    lam: float, the lambda
    r: float, the radius of the circle

    Ouputs:
    poisson_points: 2d array of shape (N-1, 2), (where N is the number of Poisson points to b
    """
    sum = 0.0
    N = 0
    X = np.array([])
    while sum <= r**2:
        Xi = expon.rvs()
        sum += Xi/(lam*np.pi)
        N += 1
        X = np.append(X, Xi)
    if N == 1:
        print('There are no points in the circle!')
        return
    else:
        poisson_points = np.zeros((N-1, 2))
        R = np.zeros(N-1)
        for i in range(N-1):
            R[i] = np.sqrt(X[:i+1].sum()/(lam*np.pi))
        U = np.random.rand(N-1)
        poisson_points[:, 0] = R
        poisson_points[:, 1] = 2*np.pi*U
        return poisson_points
```

```
import matplotlib.pyplot as plt

lam = 1.0
r = 10.0

pts = Poisson_circle(lam, r)
plt.plot(pts[:,0]*np.cos(pts[:,1]), pts[:,0]*np.sin(pts[:,1]), 'bo')
plt.title('Realization of a 2D Poisson porocess in a circular regions')
plt.show()
```

In other words, the amount of area that needs to be traversed to encompass a Poisson point is exponential with rate λ. Since, by symmetry, the respective angles of the Poisson points are independent and uniformly distributed over $(0, 2\pi)$, we thus have the following algorithm for simulating the Poisson process over a circular region of radius r about $\mathbf{O}$:

Step 1: Generate independent exponentials with rate 1, $X_1, X_2, ...$, stopping at

$$N = \min\left\{n \colon \frac{X_1 + \cdots + X_n}{\lambda\pi} > r^2\right\}$$

Step2: If $N = 1$, stop. There are no points in $C(r)$. Otherwise, for $i = 1, \dots, N-1$, set

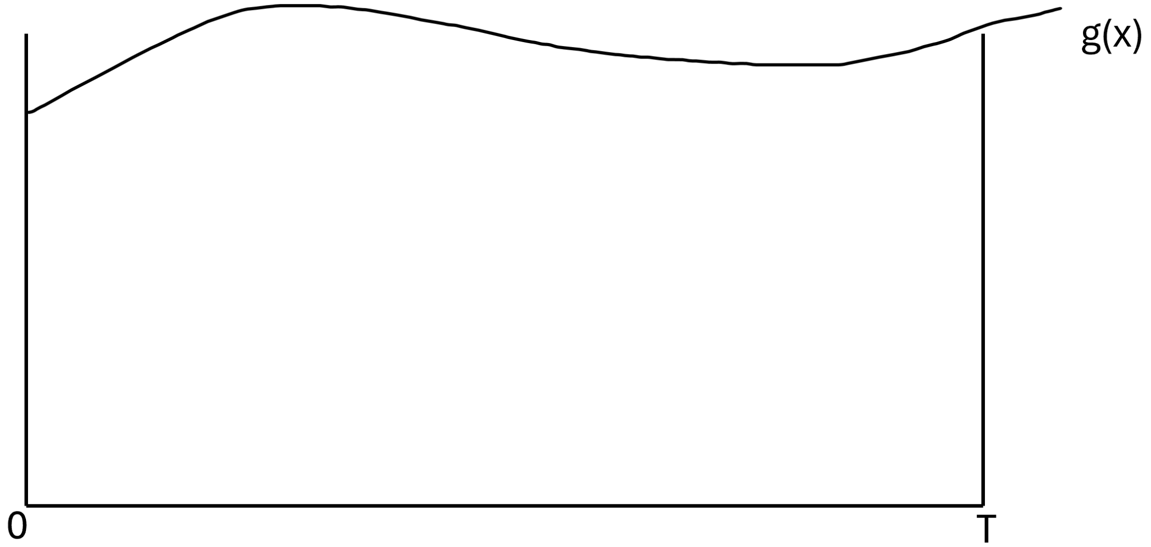
$$R_i = \sqrt{(X_1 + \cdots + X_i)/\lambda\pi}$$

Step 3: Generate independent uniform $(0, 1)$ random variables $U_1, \dots, U_{N-1}$.

Step 4: Return the $N - 1$ Poisson points in $C(r)$ whose polar coordinates are

$$(R_i, 2\pi U_i), \quad i = 1, \dots, N-1$$

The preceding algorithm requires, on average, $1 + \lambda\pi r^2$ exponentials and an equal number of uniform random numbers. Another approach to simulating points in $C(r)$ is to first simulate N, the number of such points, and then use the fact that, given N, the points are uniformly distributed in $C(r)$. This latter procedure requires the simulation of N, a Poisson random variable with mean $\lambda\pi r^2$; we must then simulate N uniform points on $C(r)$, by simulating R from the distribution $F_R(a) = a^2/r^2$ and θ from uniform $(0, 2\pi)$ and must then sort these N uniform values in increasing order of R. The main advantage of the first procedure is that it eliminates the need to sort.



The preceding algorithm can be thought of as the fanning out of a circle centered at **O** with a radius that expands continuously from 0 to r. The successive radii at which Poisson points are encountered is simulated by noting that the additional area necessary to encompass a Poisson point is always, independent of the past, exponential with rate λ. This technique can be used to simulate the process over noncircular regions. For instance, consider a nonnegative function $g(x)$, and suppose we are interested in simulating the Poisson process in the region between the x-axis and g with x going from 0 to T (see Figure above). To do so we can start at the left-hand end and fan vertically to the right by considering the successive areas $\int_0^a g(x)dx$. Now if $X_1 < X_2 < \cdots$ denote the successive projections of the Poisson points on the x-axis, then analogous to Proposition 24.1, it will follow that (with $X_0 = 0$)$\lambda \int_{X_{i-1}}^{X_i} g(x)dx, i \geqslant 1$, will be independent exponentials with rate 1. Hence, we should simulate $\epsilon_1, \epsilon_2, \ldots$, independent exponentials with rate 1, stopping at

$$N = \min\left\{n \colon \epsilon_1 + \cdots + \epsilon_n > \lambda \int_0^T g(x)\,dx\right\}$$

and determine $X_1, \ldots, X_{N-1}$ by

$$\lambda \int_0^{X_1} g(x)\,dx = \epsilon_1,$$
$$\lambda \int_{X_1}^{X_2} g(x)\,dx = \epsilon_2,$$
$$\vdots$$
$$\lambda \int_{X_{N-2}}^{X_{N-1}} g(x)\,dx = \epsilon_{N-1}$$

If we now simulate $U_1, \ldots, U_{N-1}$—independent uniform $(0, 1)$ random numbers —then as the projection on the y-axis of the Poisson point whose x-coordinate is X_i, is uniform on $(0, g(X_i))$, it follows that the simulated Poisson points in the interval are $(X_i, U_i g(X_i)), i = 1, \ldots, N-1$.

Of course, the preceding technique is most useful when g is regular enough so that the foregoing equations can be solved for the X_i. For instance, if $g(x) = y$ (and so the region of interest is a rectangle), then

$$X_i = \frac{\epsilon_1 + \cdots + \epsilon_i}{\lambda y}, \quad i = 1, \ldots, N-1$$

and the Poisson points are

$$(X_i, yU_i), \quad i = 1, \ldots, N-1$$

23 Variance Reduction Techniques-Part I

Let $X_1, \dots, X_n$ have a given joint distribution, and suppose we are interested in computing

$$\theta \equiv E[g(X_1, \dots, X_n)]$$

where g is some specified function. It is often the case that it is not possible to analytically compute the preceding, and when such is the case we can attempt to use simulation to estimate θ. This is done as follows: Generate $X_1^{(1)}, \dots, X_n^{(1)}$ having the same joint distribution as $X_1, \dots, X_n$ and set

$$Y_1 = g(X_1^{(1)}, \dots, X_n^{(1)})$$

Now, simulate a second set of random variables (independent of the first set) $X_1^{(2)}, \dots, X_n^{(2)}$ having the distribution of $X_1, \dots, X_n$ and set

$$Y_2 = g(X_1^{(2)}, \dots, X_n^{(2)})$$

Continue this until you have generated k (some predetermined number) sets, and so have also computed $Y_1, Y_2, \dots, Y_k$. Now, $Y_1, \dots, Y_k$ are independent and identically distributed random variables each having the same distribution of $g(X_1, \dots, X_n)$. Thus, if we let $\bar{Y}$ denote the average of these k random variables—that is,

$$\bar{Y} = \sum_{i=1}^{k} Y_i/k$$

then

$$E[\bar{Y}] = \theta,$$
$$E\Big[(\bar{Y} - \theta)^2\Big] = \mathrm{Var}(\bar{Y})$$

Hence, we can use $\bar{Y}$ as an estimate of θ. As the expected square of the difference between $\bar{Y}$ and θ is equal to the variance of $\bar{Y}$, we would like this quantity to be as small as possible. [In the preceding situation, $\mathrm{Var}(\bar{Y}) = \mathrm{Var}(Y_i)/k$, usually not known in advance but must be estimated from the generated values $Y_1, \dots, Y_n$.] We now present three general techniques for reducing the variance of our estimator.

23.0.1 Use of Antithetic Variables

In the preceding situation, suppose that we have generated Y_1 and Y_2, identically distributed random variables having mean θ. Now,

$$\begin{aligned}\operatorname{Var}\left(\frac{Y_1+Y_2}{2}\right) &= \frac{1}{4}[\operatorname{Var}(Y_1)+\operatorname{Var}(Y_2)+2\operatorname{Cov}(Y_1,Y_2)] \\ &= \frac{\operatorname{Var}(Y_1)}{2}+\frac{\operatorname{Cov}(Y_1,Y_2)}{2}\end{aligned}$$

Hence, it would be advantageous (in the sense that the variance would be reduced) if Y_1 and Y_2 rather than being independent were negatively correlated. To see how we could arrange this, let us suppose that the random variables $X_1,\dots,X_n$ are independent and, in addition, that each is simulated via the inverse transform technique. That is, X_i is simulated from $F_i^{-1}(U_i)$ where U_i is a random number and F_i is the distribution of X_i. Hence, Y_1 can be expressed as

$$Y_1 = g(F_1^{-1}(U_1),\dots,F_n^{-1}(U_n))$$

Now, since $1-U$ is also uniform over $(0,1)$ whenever U is a random number (and is negatively correlated with U) it follows that Y_2 defined by

$$Y_2 = g(F_1^{-1}(1-U_1),\dots,F_n^{-1}(1-U_n))$$

will have the same distribution as Y_1. Hence, if Y_1 and Y_2 were negatively correlated, then generating Y_2 by this means would lead to a smaller variance than if it were generated by a new set of random numbers. (In addition, there is a computational savings since rather than having to generate n additional random numbers, we need only subtract each of the previous n from 1.) The following theorem will be the key to showing that this technique—known as the use of antithetic variables—will lead to a reduction in variance whenever g is a monotone function.

Theorem 23.1. *If $X_1,\dots,X_n$ are independent, then, for any increasing functions f and g of n variables,*

$$E[f(\mathbf{X})g(\mathbf{X})] \geqslant E[f(\mathbf{X})]E[g(\mathbf{X})] \tag{23.1}$$

where $\mathbf{X}=(X_1,\dots,X_n)$.

Proof. The proof is by induction on n. To prove it when $n=1$, let f and g be increasing functions of a single variable. Then, for any x and y,

$$(f(x)-f(y))(g(x)-g(y)) \geqslant 0$$

since if $x \geqslant y$ ($x \leqslant y$) then both factors are nonnegative (nonpositive). Hence, for any random variables X and Y,

$$(f(X)-f(Y))(g(X)-g(Y)) \geqslant 0$$

implying that

$$E[(f(X)-f(Y))(g(X)-g(Y))] \geqslant 0$$

or, equivalently,

$$E[f(X)g(X)]+E[f(Y)g(Y)] \geqslant E[f(X)g(Y)]+E[f(Y)g(X)]$$

If we suppose that X and Y are independent and identically distributed then, as in this case,

$$\begin{aligned} E[f(X)g(X)] &= E[f(Y)g(Y)], \\ E[f(X)g(Y)] &= E[f(Y)g(X)] = E[f(X)]E[g(X)] \end{aligned}$$

we obtain the result when $n = 1$.

So assume that Equation 23.1 holds for $n-1$ variables, and now suppose that $X_1, \dots, X_n$ are independent and f and g are increasing functions. Then

$$\begin{aligned} & E[f(\mathbf{X})g(\mathbf{X})|X_n = x_n] \\ & = E[f(X_1,\dots,X_{n-1},x_n)g(X_1,\dots,X_{n-1},x_n)|X_n = x] \\ & = E[f(X_1,\dots,X_{n-1},x_n)g(X_1,\dots,X_{n-1},x_n)] \quad \text{by independence} \\ & \geqslant E[f(X_1,\dots,X_{n-1},x_n)]E[g(X_1,\dots,X_{n-1},x_n)] quad \text{by the induction hypothesis} \\ & = E[f(\mathbf{X})|X_n = x_n]E[g(\mathbf{X})|X_n = x_n] \end{aligned}$$

Hence,

$$E[f(\mathbf{X})g(\mathbf{X})|X_n] \geqslant E[f(\mathbf{X})|X_n]E[g(\mathbf{X})|X_n]$$

and, upon taking expectations of both sides,

$$\begin{aligned} E[f(\mathbf{X})g(\mathbf{X})] &\geqslant E[E[f(\mathbf{X})|X_n]E[g(\mathbf{X})|X_n]] \\ &\geqslant E[f(\mathbf{X})]E[g(\mathbf{X})] \end{aligned}$$

The last inequality follows because $E[f(\mathbf{X})|X_n]$ and $E[g(\mathbf{X})|X_n]$ are both increasing functions of X_n, and so, by the result for $n = 1$,

$$\begin{aligned} E\Big[E[f(\mathbf{X})|X_n]E[g(\mathbf{X})|X_n]\Big] &\geqslant E\Big[E[f(\mathbf{X})|X_n]\Big]E\Big[E[g(\mathbf{X})|X_n]\Big] \\ &= E[f(\mathbf{X})]E[g(\mathbf{X})] \end{aligned}$$

□

Corollary 23.1. *If $U_1, \dots, U_n$ are independent, and k is either an increasing or decreasing function, then*

$$\mathrm{Cov}(k(U_1,\dots,U_n), k(1-U_1,\dots,1-U_n)) \leqslant 0$$

Proof. Suppose k is increasing. As $-k(1-U_1, \dots, 1-U_n)$ is increasing in $U_1, \dots, U_n$, then, from Theorem 23.1,

$$\mathrm{Cov}(k(U_1, \dots, U_n), -k(1-U_1, \dots, 1-U_n)) \geqslant 0$$

When k is decreasing just replace k by its negative. When k is decreasing just replace k by its negative. □

Since $F_i^{-1}(U_i)$ is increasing in U_i (as F_i, being a distribution function, is increasing) it follows that $g(F_1^{-1}(U_1), \dots, F_n^{-1}(U_n))$ is a monotone function of $U_1, \dots, U_n$ whenever g is monotone. Hence, if g is monotone the antithetic variable approach of twice using each set of random numbers $U_1, \dots, U_n$ by first computing $g(F_1^{-1}(U_1), \dots, F_n^{-1}(U_n))$ and then $g(F_1^{-1}(1-U_1), \dots, F_n^{-1}(1-U_n))$ will reduce the variance of the estimate of $E[g(X_1, \dots, X_n)]$. That is, rather than generating k sets of n random numbers, we should generate $k/2$ sets and use each set twice.

Example 23.1. (Simulating the Reliability Function) Consider a system of n components in which component i, independently of other components, works with probability $p_i, i = 1, \dots, n$. Letting

$$X_i = \begin{cases} 1, & \text{if component } i \text{ works} \\ 0, & \text{otherwise} \end{cases}$$

suppose there is a monotone structure function ϕ such that

$$\phi(X_1, \dots, X_n) = \begin{cases} 1, & \text{if the system works under } X_1, \dots, X_n \\ 0, & \text{otherwise} \end{cases}$$

We are interested in using simulation to estimate

$$r(p_1, \dots, p_n) \equiv E[\phi(X_1, \dots, X_n)] = P\{\phi(X_1, \dots, X_n) = 1\}$$

Now, we can simulate the X_i by generating uniform random numbers $U_1, \dots, U_n$ and then setting

$$X_i = \begin{cases} 1, & \text{if } U_i < p_i \\ 0, & \text{otherwise} \end{cases}$$

Hence, we see that

$$\phi(X_1, \dots, X_n) = k(U_1, \dots, U_n)$$

where k is a decreasing function of $U_1, \dots, U_n$. Hence,

$$\mathrm{Cov}(k(\mathbf{U}), k(\mathbf{1}-\mathbf{U})) \leqslant 0$$

and so the antithetic variable approach of using $U_1, \dots, U_n$ to generate both $k(U_1, \dots, U_n)$ and $k(1-U_1, \dots, 1-U_n)$ results in a smaller variance than if an independent set of random numbers was used to generate the second k.

Example 23.2. (Simulating a Queueing System) Consider a given queueing system, and let D_i denote the delay in queue of the ith arriving customer, and suppose we are interested in simulating the system so as to estimate

$$\theta = E[D_1 + \cdots + D_n]$$

Let $X_1, \ldots, X_n$ denote the first n interarrival times and $S_1, \ldots, S_n$ the first n service times of this system, and suppose these random variables are all independent. Now in most systems $D_1 + \cdots + D_n$ will be a function of $X_1, \ldots, X_n$, $S_1, \ldots, S_n$—say,

$$D_1 + \cdots + D_n = g(X_1, \ldots, X_n, S_1, \ldots, S_n)$$

Also g will usually be increasing in S_i and decreasing in $X_i, i = 1, \ldots, n$. If we use the inverse transform method to simulate $X_i, S_i, i = 1, \ldots, n$—say, $X_i = F_i^{-1}(1 - U_i), S_i = G_i^{-1}(\bar{U}_i)$ where $U_1, \ldots, U_n, \bar{U}_1, \ldots, \bar{U}_n$, are independent uniform random numbers—then we may write

$$D_1 + \cdots + D_n = k(U_1, \ldots, U_n, \bar{U}_1, \ldots, \bar{U}_n)$$

where k is increasing in its variates. Hence, the antithetic variable approach will reduce the variance of the estimator of θ. [Thus, we would generate $U_i, \bar{U}_i, i = 1, \ldots, n$ and set $X_i = F_i^{-1}(1 - U_i)$ and $Y_i = G_i^{-1}(\bar{U}_i)$ for the first run, and $X_i = F_i^{-1}(U_i)$ and $Y_i = G_i^{-1}(1 - \bar{U}_i)$ for the second.] As all the U_i and $\bar{U}_i$ are independent, however, this is equivalent to setting $X_i = F_i^{-1}(U_i), Y_i = G_i^{-1}(\bar{U}_i)$ in the first run and using $1 - U_i$ for U_i and $1 - \bar{U}_i$ for $\bar{U}_i$ in the second.

Example 23.3. Use the crude Monte Carlo and antithetic variable estimators to estimate the standard normal cdf

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Since the integration cover an unbounded interval, we break this problem into two cases: $x \geq 0$ and $x < 0$, and use the symmetry of the normal density to handle the second case. So

$$\Phi(x) = 0.5 + \int_{0}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \ x \geq 0$$

To estimate $\theta = \int_0^x e^{-t^2/2} dt$ for $x > 0$, we can generate random $U(0, x)$ numbers, but it would change the parameters of uniform distribution for each different value.

We prefer an algorithm that always samples from $U(0, 1)$ via a change of variables. Making the substitution $y = t/x$, we have $dt = xdy$ and

$$\theta = \int_0^1 x e^{-(xy)^2/2} dy$$

Thus, $\theta = E_Y[xe^{-(xY)^2/2}]$, where $Y \sim U(0,1)$. Finally the Monte Carlo estimator for $\Phi(x)$ is $0.5 + \frac{1}{\sqrt{2\pi}}\theta$.

By restricting the simulation to the upper tail, the function $g(\cdot)$ is monotone, so the antithetic variable approach works. Generate random numbers $y_1, \dots, y_{n/2} \sim U(0,1)$ and compute half of the replicates using

$$g(y_j) = xe^{-(xy_j)^2/2},\ j = 1, \dots, n/2$$

and compute the remaining half of the replicates using

$$g(y_j) = xe^{-(x(1-y_j))^2/2},\ j = 1, \dots, n/2$$

Then the antithetic variable estimator is

$$\theta^{\text{anvar}} = \frac{1}{n}\sum_{j=1}^{n/2}\left(xe^{-(xy_j)^2/2} + xe^{-(x(1-y_j))^2/2}\right)$$

```python
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt

nsamples = [2**10, 2**11, 2**12, 2**13, 2**14]
ncopy = 64
l_cmc = np.zeros(len(nsamples))
l_av = np.zeros(len(nsamples))
x = 1

np.random.seed(64)
# Crude Monte Carlo
for i in range(len(nsamples)):
    for j in range(ncopy):
        samples = np.random.rand(nsamples[i])
        l_cmc[i] += np.mean(x*np.exp(-(x*samples)**2/2))
    l_cmc[i] /= ncopy
    l_cmc[i] = 0.5 + l_cmc[i]/np.sqrt(2*np.pi)

# Antithetic variables
for i in range(len(nsamples)):
    for j in range(ncopy):
        samples = np.random.rand(int(nsamples[i]/2))
        tmp = np.sum(x*np.exp(-(x*samples)**2/2))
        tmp += np.sum(x*np.exp(-(x*(1-samples))**2/2))
        l_av[i] += tmp/nsamples[i]
    l_av[i] /= ncopy
    l_av[i] = 0.5 + l_av[i]/np.sqrt(2*np.pi)

print('Estimates for crude MC: ', l_cmc)
print('Estimates for antithetic variable estimator: ', l_av)
```

We can see the convergence behavior of the two estimators from the following figure.

```
plt.loglog(nsamples, abs((sp.stats.norm.cdf(1)-l_cmc)/sp.stats.norm.cdf(1)), 'b-', label='cru
plt.loglog(nsamples, abs((sp.stats.norm.cdf(1)-l_av)/sp.stats.norm.cdf(1)), 'r-.', label='ant
plt.legend()
plt.xlabel('Sample size')
plt.ylabel('Relative error')
plt.title('$\Phi(1)$ using Monte Carlo methods')
plt.show()
```

24 Variance Reduction Techniques-Part II

24.1 Variance Reduction by Conditioning

Let us start by recalling the conditional variance formula

$$\operatorname{Var}(Y) = E[\operatorname{Var}(Y|Z)] + \operatorname{Var}(E[Y|Z]) \tag{24.1}$$

Now suppose we are interested in estimating $E[g(X_1, \dots, X_n)]$ by simulating $\mathbf{X} = (X_1, \dots, X_n)$ and then computing $Y = g(X_1, \dots, X_n)$. Now, if for some random variable Z we can compute $E[Y|Z]$ then, as $\operatorname{Var}(Y|Z) \geqslant 0$, it follows from the conditional variance formula that

$$\operatorname{Var}(E[Y|Z]) \leqslant \operatorname{Var}(Y)$$

implying, since $E[E[Y|Z]] = E[Y]$, that $E[Y|Z]$ is a better estimator of $E[Y]$ than is Y.

In many situations, there are a variety of Z_i that can be conditioned on to obtain an improved estimator. Each of these estimators $E[Y|Z_i]$ will have mean $E[Y]$ and smaller variance than does the raw estimator Y. We now show that for any choice of weights $\lambda_i, \lambda_i \geqslant 0, \sum_i \lambda_i = 1, \sum_i \lambda_i E[Y|Z_i]$ is also an improvement over Y.

Proposition 24.1. *For any $\lambda_i \geqslant 0, \sum_{i=1}^{\infty} \lambda_i = 1$,*

(a) $E[\sum_i \lambda_i E[Y|Z_i]] = E[Y]$,

(b) $\operatorname{Var}(\sum_i \lambda_i E[Y|Z_i]) \leqslant \operatorname{Var}(Y)$.

Proof. The proof of (a) is immediate. To prove (b), let N denote an integer valued random variable independent of all the other random variables under consideration and such that

$$P\{N = i\} = \lambda_i, \quad i \geqslant 1$$

Applying the conditional variance formula twice yields

$$\begin{aligned}\operatorname{Var}(Y) &\geqslant \operatorname{Var}(E[Y|N, Z_N]) \\ &\geqslant \operatorname{Var}(E[E[Y|N, Z_N]|Z_1, \dots]) \\ &= \operatorname{Var} \sum_i \lambda_i E[Y|Z_i]\end{aligned}$$

□

Example 24.1. Considers a queueing system having Poisson arrivals and suppose that any customer arriving when there are already N others in the system is lost. Suppose that we are interested in using simulation to estimate the expected number of lost customers by time t. The raw simulation approach would be to simulate the system up to time t and determine L, the number of lost customers for that run. A better estimate, however, can be obtained by conditioning on the total time in $[0, t]$ that the system is at capacity. Indeed, if we let T denote the time in $[0, t]$ that there are N in the system, then

$$E[L|T] = \lambda T$$

where λ is the Poisson arrival rate. Hence, a better estimate for $E[L]$ than the average value of L over all simulation runs can be obtained by multiplying the average value of T per simulation run by λ. If the arrival process were a nonhomogeneous Poisson process, then we could improve over the raw estimator L by keeping track of those time periods for which the system is at capacity. If we let $I_1, \ldots, I_C$ denote the time intervals in $[0, t]$ in which there are N in the system, then

$$E[L|I_1, \ldots, I_C] = \sum_{i=1}^{C} \int_{I_i} \lambda(s)\, ds$$

where $\lambda(s)$ is the intensity function of the nonhomogeneous Poisson arrival process. The use of the right side of the preceding would thus lead to a better estimate of $E[L]$ than the raw estimator L.

Example 24.2. Suppose that we wanted to estimate the expected sum of the times in the system of the first n customers in a queueing system. That is, if W_i is the time that the ith customer spends in the system, then we are interested in estimating

$$\theta = E\left[\sum_{i=1}^{n} W_i\right]$$

Let Y_i denote the "state of the system" at the moment at which the ith customer arrives. It can be shown that for a wide class of models the estimator $\sum_{i=1}^{n} E[W_i|Y_i]$ has (the same mean and) a smaller variance than the estimator $\sum_{i=1}^{n} W_i$. (It should be noted that whereas it is immediate that $E[W_i|Y_i]$ has smaller variance than W_i, because of the covariance terms involved it is not immediately apparent that $\sum_{i=1}^{n} E[W_i|Y_i]$ has smaller variance than $\sum_{i=1}^{n} W_i$.)

24.2 Control Variates

Again suppose we want to use simulation to estimate $E[g(\mathbf{X})]$ where $\mathbf{X} = (X_1, \ldots, X_n)$. But now suppose that for some function f the expected value of $f(\mathbf{X})$ is known—say, $E[f(\mathbf{X})] = \mu$. Then for any constant a we can also use

$$W = g(\mathbf{X}) + a(f(\mathbf{X}) - \mu)$$

as an estimator of $E[g(\mathbf{X})]$. Now,

$$\operatorname{Var}(W) = \operatorname{Var}(g(\mathbf{X})) + a^2 \operatorname{Var}(f(\mathbf{X})) + 2a \operatorname{Cov}(g(\mathbf{X}), f(\mathbf{X}))$$

Simple calculus shows that the preceding is minimized when

$$a = \frac{-\operatorname{Cov}(f(\mathbf{X}), g(\mathbf{X}))}{\operatorname{Var}(f(\mathbf{X}))}$$

and, for this value of a,

$$\operatorname{Var}(W) = \operatorname{Var}(g(\mathbf{X})) - \frac{[\operatorname{Cov}(f(\mathbf{X}), g(\mathbf{X}))]^2}{\operatorname{Var}(f(\mathbf{X}))}$$

Because $\operatorname{Var}(f(\mathbf{X}))$ and $\operatorname{Cov}(f(\mathbf{X}), g(\mathbf{X}))$ are usually unknown, the simulated data should be used to estimate these quantities.

Dividing the preceding equation by $\operatorname{Var}(g(\mathbf{X}))$ shows that

$$\frac{\operatorname{Var}(W)}{\operatorname{Var}(g(\mathbf{X}))} = 1 - \operatorname{Corr}^2(f(\mathbf{X}), g(\mathbf{X}))$$

where $\operatorname{Corr}(X, Y)$ is the correlation between X and Y. Consequently, the use of a control variate will greatly reduce the variance of the simulation estimator whenever $f(\mathbf{X})$ and $g(\mathbf{X})$ are strongly correlated.

Example 24.3. Consider a continuous time Markov chain which, upon entering state i, spends an exponential time with rate v_i in that state before making a transition into some other state, with the transition being into state j with probability $P_{i,j}, i \geqslant 0, j \neq i$. Suppose that costs are incurred at rate $C(i) \geqslant 0$ per unit time whenever the chain is in state $i, i \geqslant 0$. With $X(t)$ equal to the state at time t, and α being a constant such that $0 < \alpha < 1$, the quantity

$$W = \int_0^\infty e^{-\alpha t} C(X(t))\, dt$$

represents the total discounted cost. For a given initial state, suppose we want to use simulation to estimate $E[W]$. Whereas at first it might seem that we cannot obtain an unbiased estimator without simulating the continuous time Markov chain for an infinite amount of time (which is clearly impossible), we can make use of the results of Example 1 in Lecture 8 which gives the equivalent expression for $E[W]$:

$$E[W] = E\left[\int_0^T C(X(t))\, dt\right]$$

where T is an exponential random variable with rate α that is independent of the continuous time Markov chain. Therefore, we can first generate the value of T, then generate the states of the continuous time Markov chain up to time T, to obtain the unbiased estimator $\int_0^T C(X(t))dt$. Because all the costs rates are nonnegative this estimator is strongly positively correlated with T, which will thus make an effective control variate.

Example 24.4. (A Queueing System) Let D_{n+1} denote the delay in queue of the $n+1$ customer in a queueing system in which the interarrival times are independent and identically distributed (i.i.d.) with distribution F having mean μ_F and are independent of the service times which are i.i.d. with distribution G having mean μ_G. If X_i is the interarrival time between arrival i and $i+1$, and if S_i is the service time of customer $i, i \geqslant 1$, we may write

$$D_{n+1} = g(X_1, \dots, X_n, S_1, \dots, S_n)$$

To take into account the possibility that the simulated variables X_i, S_i may by chance be quite different from what might be expected we can let

$$f(X_1, \dots, X_n, S_1, \dots, S_n) = \sum_{i=1}^{n} (S_i - X_i)$$

As $E[f(\mathbf{X}, \mathbf{S})] = n(\mu_G - \mu_F)$ we could use

$$g(\mathbf{X}, \mathbf{S}) + a[f(\mathbf{X}, \mathbf{S}) - n(\mu_G - \mu_F)]$$

as an estimator of $E[D_{n+1}]$. Since D_{n+1} and f are both increasing functions of $S_i, -X_i, i = 1, \dots, n$ it follows from Theorem 1 Lecture 23 that $f(\mathbf{X}, \mathbf{S})$ and $D_n + 1$ are positively correlated, and so the simulated estimate of a should turn out to be negative.

If we wanted to estimate the expected sum of the delays in queue of the first $N(T)$ arrivals, then we could use $\sum_{i=1}^{N(T)} S_i$ as our control variable. Indeed as the arrival process is usually assumed independent of the service times, it follows that

$$E\left[\sum_{i=1}^{N(T)} S_i\right] = E[S]E[N(T)]$$

This control variable could also be used if the arrival process were a nonhomogeneous Poisson with rate $\lambda(t)$; in this case,

$$E[N(T)] = \int_0^T \lambda(t)\, dt$$

25 Variance Reduction Techniques-Part III

25.1 Importance Sampling

Let $\mathbf{X} = (X_1, \dots, X_n)$ denote a vector of random variables having a joint density function (or joint mass function in the discrete case) $f(\mathbf{x}) = f(x_1, \dots, x_n)$, and suppose that we are interested in estimating

$$\theta = E[h(\mathbf{X})] = \int h(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}$$

where the preceding is an n-dimensional integral. (If the X_i are discrete, then interpret the integral as an n-fold summation.)

Suppose that a direct simulation of the random vector $\mathbf{X}$, so as to compute values of $h(\mathbf{X})$, is inefficient, possibly because (a) it is difficult to simulate a random vector having density function $f(\mathbf{x})$, or (b) the variance of $h(\mathbf{X})$ is large, or (c) a combination of (a) and (b).

Another way in which we can use simulation to estimate θ is to note that if $g(\mathbf{X})$ is another probability density such that $f(\mathbf{x}) = 0$ whenever $g(\mathbf{x}) = 0$, then we can express θ as

$$\begin{aligned} \theta &= \int \frac{h(\mathbf{x}) f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\, d\mathbf{x} \\ &= E_g\left[\frac{h(\mathbf{X}) f(\mathbf{X})}{g(\mathbf{X})}\right] \end{aligned} \tag{25.1}$$

where we have written E_g to emphasize that the random vector $\mathbf{X}$ has joint density $g(\mathbf{x})$.

It follows from Equation 25.1 that θ can be estimated by successively generating values of a random vector $\mathbf{X}$ having density function $g(\mathbf{x})$ and then using as the estimator the average of the values of $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$. If a density function $g(\mathbf{x})$ can be chosen so that the random variable $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ has a small variance then this approach—referred to as *importance sampling*—can result in an efficient estimator of θ .

Let us now try to obtain a feel for why importance sampling can be useful. To begin, note that $f(\mathbf{X})$ and $g(\mathbf{X})$ represent the respective likelihoods of obtaining the vector $\mathbf{X}$ when $\mathbf{X}$ is a random vector with respective densities f and g. Hence, if $\mathbf{X}$ is distributed according to g, then it will usually be the case that $f(\mathbf{X})$ will be small in relation to $g(\mathbf{X})$ and thus when $\mathbf{X}$ is simulated according to g the likelihood ratio $f(\mathbf{X})/g(\mathbf{X})$ will usually be small in comparison to 1. However, it is easy to check that its mean is 1:

$$E_g\left[\frac{f(\mathbf{X})}{g(\mathbf{X})}\right] = \int \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\,d\mathbf{x} = \int f(\mathbf{x})\,d\mathbf{x} = 1$$

Thus we see that even though $f(\mathbf{X})/g(\mathbf{X})$ is usually smaller than 1, its mean is equal to 1; thus implying that it is occasionally large and so will tend to have a large variance. So how can $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ have a small variance? The answer is that we can sometimes arrange to choose a density g such that those values of $\mathbf{x}$ for which $f(\mathbf{x})/g(\mathbf{x})$ is large are precisely the values for which $h(\mathbf{x})$ is exceedingly small, and thus the ratio $h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$ is always small. Since this will require that $h(\mathbf{x})$ sometimes be small, importance sampling seems to work best when estimating a small probability; for in this case the function $h(\mathbf{x})$ is equal to 1 when $\mathbf{x}$ lies in some set and is equal to 0 otherwise.

We will now consider how to select an appropriate density g. We will find that the so-called tilted densities are useful. Let $M(t) = E_f[e^{tX}] = \int e^{tx} f(x)dx$ be the moment generating function corresponding to a one-dimensional density f.

Definition 25.1. A density function

$$f_t(x) = \frac{e^{tx} f(x)}{M(t)}$$

is called a *tilted* density of $f, -\infty < t < \infty$.

A random variable with density f_t tends to be larger than one with density f when $t > 0$ and tends to be smaller when $t < 0$.

In certain cases the tilted distributions f_t have the same parametric form as does f.

Example 25.1. If f is the exponential density with rate λ then

$$f_t(x) = Ce^{tx}\lambda e^{-\lambda x} = \lambda C e^{-(\lambda - t)x}$$

where $C = 1/M(t)$ does not depend on x. Therefore, for $t \leqslant \lambda, f_t$ is an exponential density with rate $\lambda - t$.

If f is a Bernoulli probability mass function with parameter p, then

$$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Hence, $M(t) = E_f[e^{tX}] = pe^t + 1 - p$ and so

$$\begin{aligned} f_t(x) &= \frac{1}{M(t)}(pe^t)^x(1-p)^{1-x} \\ &= \left(\frac{pe^t}{pe^t + 1 - p}\right)^x \left(\frac{1-p}{pe^t + 1 - p}\right)^{1-x} \end{aligned}$$

That is, f_t is the probability mass function of a Bernoulli random variable with parameter

$$p_t = \frac{pe^t}{pe^t + 1 - p}$$

We leave it as an exercise to show that if f is a normal density with parameters μ and σ^2 then f_t is a normal density mean $\mu + \sigma^2 t$ and variance σ^2.

In certain situations the quantity of interest is the sum of the independent random variables $X_1, \dots, X_n$. In this case the joint density f is the product of one-dimensional densities. That is,

$$f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$$

where f_i is the density function of X_i. In this situation it is often useful to generate the X_i according to their tilted densities, with a common choice of t employed.

Example 25.2. Let $X_1, \dots, X_n$ be independent random variables having respective probability density (or mass) functions f_i, for $i = 1, \dots, n$. Suppose we are interested in approximating the probability that their sum is at least as large as a, where a is much larger than the mean of the sum. That is, we are interested in

$$\theta = P\{S \geqslant a\}$$

where $S = \sum_{i=1}^n X_i$, and where $a > \sum_{i=1}^n E[X_i]$. Letting $I\{S \geqslant a\}$ equal 1 if $S \geqslant a$ and letting it be 0 otherwise, we have that

$$\theta = E_{\mathrm{f}}[I\{S \geqslant a\}]$$

where $\mathbf{f} = (f_1, \dots, f_n)$. Suppose now that we simulate X_i according to the tilted mass function $f_{i,t}, i = 1, \dots, n$, with the value of $t, t > 0$ left to be determined. The importance sampling estimator of θ would then be

$$\hat{\theta} = I\{S \geqslant a\} \prod \frac{f_i(X_i)}{f_{i,t}(X_i)}$$

Now,

$$\frac{f_i(X_i)}{f_{i,t}(X_i)} = M_i(t) e^{-tX_i}$$

and so

$$\hat{\theta} = I\{S \geqslant a\} M(t) e^{-tS}$$

where $M(t) = \prod M_i(t)$ is the moment generating function of S. Since $t > 0$ and $I\{S \geqslant a\}$ is equal to 0 when $S < a$, it follows that

$$I\{S \geqslant a\} e^{-tS} \leqslant e^{-ta}$$

and so

$$\hat{\theta} \leqslant M(t) e^{-ta}$$

To make the bound on the estimator as small as possible we thus choose $t, t > 0$,to minimize $M(t)e^{-ta}$. In doing so, we will obtain an estimator whose value on each iteration is between 0 and $\min_t M(t)e^{-ta}$. It can be shown that the minimizing t, call it t^*, is such that

$$E_{t^*}[S] = E_{t^*}\left[\sum_{i=1}^{n} X_i\right] = a$$

where, in the preceding, we mean that the expected value is to be taken under the assumption that the distribution of X_i is f_{i,t^*} for $i = 1, \dots, n$.

For instance, suppose that $X_1, \dots, X_n$ are independent Bernoulli random variables having respective parameters p_i, for $i = 1, \dots, n$. Then, if we generate the X_i according to their tilted mass functions $p_{i,t}, i = 1, \dots, n$ then the importance sampling estimator of $\theta = P\{S \geqslant a\}$ is

$$\hat{\theta} = I\{S \geqslant a\}e^{-tS}\prod_{i=1}^{n}(p_i e^t + 1 - p_i)$$

Since $p_{i,t}$ is the mass function of a Bernoulli random variable with parameter $p_i e^t/(p_i e^t + 1 - p_i)$ it follows that

$$E_t\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\frac{p_i e^t}{p_i e^t + 1 - p_i}$$

The value of t that makes the preceding equal to a can be numerically approximated and then utilized in the simulation.

As an illustration, suppose that $n = 20, p_i = 0.4$, and $a = 16$. Then

$$E_t[S] = 20\frac{0.4e^t}{0.4e^t + 0.6}$$

Setting this equal to 16 yields, after a little algebra,

$$e^{t^*} = 6$$

Thus, if we generate the Bernoullis using the parameter

$$\frac{0.4e^{t^*}}{0.4e^{t^*} + 0.6} = 0.8$$

then because

$$M(t^*) = (0.4e^{t^*} + 0.6)^{20} \quad \text{and} \quad e^{-t^*S} = (1/6)^S$$

we see that the importance sampling estimator is

$$\hat{\theta} = I\{S \geqslant 16\}(1/6)^S 3^{20}$$

It follows from the preceding that

$$\hat{\theta} \leqslant (1/6)^{16}3^{20} = 81/2^{16} = 0.001236$$

That is, on each iteration the value of the estimator is between 0 and 0.001236. Since, in this case, θ is the probability that a binomial random variable with parameters $20, 0.4$ is at least 16, it can be explicitly computed with the result $\theta = 0.0000317$. Hence, the raw simulation estimator I, which on each iteration takes the value 0 if the sum of the Bernoullis with parameter 0.4 is less than 16 and takes the value 1 otherwise, will have variance

$$\mathrm{Var}(I) = \theta(1-\theta) = 3.169 \times 10^{-4}$$

On the other hand, it follows from the fact that $0 \leqslant \hat{\theta} \leqslant 0.001236$ that

$$\mathrm{Var}(\hat{\theta}) \leqslant 2.9131 \times 10^{-7}$$

Example 25.3. Consider a single-server queue in which the times between successive customer arrivals have density function f and the service times have density g. Let D_n denote the amount of time that the nth arrival spends waiting in queue and suppose we are interested in estimating $\alpha = P\{D_n \geqslant a\}$ when a is much larger than $E[D_n]$. Rather than generating the successive interarrival and service times according to f and g, respectively, they should be generated according to the densities f_{-t} and g_t, where t is a positive number to be determined. Note that using these distributions as opposed to f and g will result in smaller interarrival times (since $-t < 0$) and larger service times. Hence, there will be a greater chance that $D_n > a$ than if we had simulated using the densities f and g. The importance sampling estimator of α would then be

$$\hat{\alpha} = I\{D_n > a\}e^{t(S_n - Y_n)}[M_f(-t)M_g(t)]^n$$

where S_n is the sum of the first n interarrival times, Y_n is the sum of the first n service times, and M_f and M_g are the moment generating functions of the densities f and g, respectively. The value of t used should be determined by experimenting with a variety of different choices.

Example 25.4. Find the probability that a randomly chosen variable X from the standard normal distribution is greater than 3. We know that one way to solve this is by solving the following integral:

$$P\{X > 3\} = \int_3^\infty f_X(t)\,dt = \frac{1}{\sqrt{2\pi}}\int_3^\infty e^{-t^2/2}\,dt$$

We use a normal distribution with $\mu = 4$ and $\sigma = 1$ as g.

```
import numpy as np
import scipy.stats as stats

h = lambda x : x > 3
f = lambda x : stats.norm().pdf(x)
g = lambda x : stats.norm(loc=4,scale=1).pdf(x)

# Sample from the N(4,1).
```

```
N = 10**4
X = np.random.normal(4,scale=1,size=N)

# Calculate the estimate for importance sampling
est_IS = 1./N * np.sum(h(X)*f(X)/g(X))
# Calculate the estimate for crude Monte Carlo
X = np.random.normal(size=N)
est_MC = 1./N * np.sum(h(X))
# Calculate the true probability
true_prob = True_value = stats.norm.cdf(-3)

print('Importance sampling error: ', np.abs(est_IS-true_prob),
'Crude MC error is: ', np.abs(est_MC-true_prob))
```

Now we visualize the performance of crude Monte Carlo and importance sampling using the following Python code.

```
import matplotlib.pyplot as plt

True_value = stats.norm.cdf(-3)

num_experiments = 32  # Run the simulations 32 times for each sample size for both crude MC a
sample_sizes = 2**(np.arange(5,12))  # 32, 64, ..., 2028

err_MC = np.zeros(len(sample_sizes))
err_IS = np.zeros(len(sample_sizes))
for i in range(len(sample_sizes)):
    X = np.random.normal(size=sample_sizes[i])
    err_MC[i] += 1./N * np.sum(h(X))
    X = np.random.normal(4,scale=1,size=N)
    err_IS[i] += 1./N * np.sum(h(X)*f(X)/g(X))
    err_MC[i] = np.abs(err_MC[i] - True_value)
    err_IS[i] = np.abs(err_IS[i] - True_value)

plt.loglog(sample_sizes, err_MC, 'g--', label='MC')
plt.loglog(sample_sizes, err_IS, 'b-', label='IS')
plt.xlabel('Sample size')
plt.ylabel('Error')
plt.title('Estimating P(X>3) where X is a standard normal with MC and IS')
plt.legend();
```