4

Stochastic Processes (i): Poisson Processes and Markov Chains

4.1 The Homogeneous Poisson Process and the Poisson Distribution

In this section we state the fundamental properties that define a Poisson process, and from these properties we derive the Poisson distribution, introduced in Section 1.3.7.

Suppose that a sequence of events occurs during some time interval. These events form a homogeneous Poisson process if the following two conditions are met:

(1) The occurrence of any event in the time interval (a, b) is independent of the occurrence of any event in the time interval (c, d), where (a, b) and (c, d) do not overlap.

(2) There is a constant $\lambda > 0$ such that for any sufficiently small time interval $(t, t + h)$, $h > 0$, the probability that one event occurs in $(t, t + h)$ is independent of t, and is $\lambda h + o(h)$ (the $o(h)$ notation is discussed in Appendix B.8), and the probability that more than one event occurs in the interval $(t, t + h)$ is $o(h)$.

Condition 2 has two implications. The first is *time homogeneity*: The probability of an event in the time interval $(t, t+h)$ is independent of t. Second, this condition means that the probability of an event occurring in a small time interval is (up to a small-order term) proportional to the length of the interval (with fixed proportionality constant λ). Thus the probability of no

events in the interval $(t, t+h)$ is

$$1 - \lambda h + o(h), \tag{4.1}$$

and the probability of one *or more* events in the interval $(t, t+h)$ is

$$\lambda h + o(h).$$

The two conditions listed above are often taken as defining "randomness" of the occurrences of the events in question. Various naturally occurring phenomena follow, or very nearly follow, these two conditions. Continuing the example in Section 1.13, suppose a cellular protein degrades spontaneously, and the quantity of this protein in the cell is maintained at a constant level by the continual generation of new proteins at approximately the degradation rate. The number of proteins that degrade in any given time interval approximately satisfies conditions 1 and 2. The justification that condition 1 can be assumed in the model is that the number of proteins in the cell is essentially constant and that the spontaneous nature of the degradation process makes the independence assumption reasonable. The discussion leading to (1.116) and that following equation (2.87) makes condition 2 reasonable for spontaneously-occurring phenomena. This condition also follows from the same logic discussed in the example on page 10 that when np is small, the probability of at least one success in n Bernoulli trials is approximately np. For a precise treatment of this issue, see Feller (1968).

We now show that under conditions 1 and 2, the number N of events that occur up to any arbitrary time t has a Poisson distribution with parameter λt.

At time 0 the value of N is necessarily 0, and at any later time t, the possible values of N are 0, 1, 2, 3, ... We denote the probability that $N = j$ at any given time t by $P_j(t)$. Note that this is a departure from our standard notational convention, which would be $P_N(j)$ with t an implicit parameter. This notational change is made because the main interest here is in assessing how $P_j(t)$ behaves as a function of j and t.

The event that $N = 0$ at time $t + h$ occurs only if no events occur in $(0, t)$ and also no events occur in $(t, t+h)$. Thus for small h,

$$P_0(t+h) = P_0(t)(1 - \lambda h + o(h)) = P_0(t)(1 - \lambda h) + o(h). \tag{4.2}$$

The first equality follows from conditions 1 and 2.

The event that $N = 1$ at time $t + h$ can occur in two ways. The first is that $N = 1$ at time t and that no event occurs in the time interval $(t, t+h)$, the second is that $N = 0$ at time t and that exactly one event occurs in the time interval $(t, t+h)$. This gives

$$P_1(t+h) = P_0(t)(\lambda h) + P_1(t)(1 - \lambda h) + o(h), \tag{4.3}$$

where the term $o(h)$ is the sum of two terms, both of which are $o(h)$. Finally, for $j = 2, 3, \ldots$, the event that $N = j$ at time $t + h$ can occur in three different ways. The first is that $N = j$ at time t and that no event

occurs in the time interval $(t, t+h)$. The second is that $N = j-1$ at time t and that exactly one event occurs in $(t, t+h)$. The final possibility is that $N \leq j-2$ at time t and that two or more events occur in $(t, t+h)$. Thus for $j = 2, 3, \ldots,$

$$P_j(t+h) = P_{j-1}(t)(\lambda h) + P_j(t)(1-\lambda h) + o(h). \tag{4.4}$$

Equations (4.3) and (4.4) look identical, and the difference between them relates only to terms of order $o(h)$. Therefore, we can take (4.4) to hold for all $j \geq 1$. Subtracting $P_j(t)$ $(j \geq 1)$ from both sides of equation (4.4) and $P_0(t)$ from both sides of equation (4.2), and then dividing through by h, we get

$$\frac{P_0(t+h) - P_0(t)}{h} = -\frac{P_0(t)(\lambda h) + o(h)}{h}$$

$$\frac{P_j(t+h) - P_j(t)}{h} = \frac{P_{j-1}(t)(\lambda h) - P_j(t)(\lambda h) + o(h)}{h},$$

$j = 1, 2, 3, \ldots$. Letting $h \to 0$, we get

$$\frac{d}{dt}P_0(t) = -\lambda P_0(t), \tag{4.5}$$

and

$$\frac{d}{dt}P_j(t) = \lambda P_{j-1}(t) - \lambda P_j(t), \quad j = 1, 2, 3, \ldots. \tag{4.6}$$

The $P_j(t)$ are subject to the conditions

$$P_0(0) = 1, \quad P_j(0) = 0, \quad j = 1, 2, 3, \ldots. \tag{4.7}$$

Equation (4.5) is one of the most fundamental of differential equations, and has the solution

$$P_0(t) = Ce^{-\lambda t}. \tag{4.8}$$

The condition $P_0(0) = 1$ implies $C = 1$, leading to

$$P_0(t) = e^{-\lambda t}. \tag{4.9}$$

Given this, we now show that the set of equations (4.6) has the solution

$$P_j(t) = \frac{e^{-\lambda t}(\lambda t)^j}{j!}, \quad j = 1, 2, 3, \ldots. \tag{4.10}$$

The method for solving these equations follows the induction procedure described in Section B.18. Equation (4.9) shows (4.10) is true for $j = 0$. It must next be shown that the assumption that (4.10) holds for $j-1$ implies that it holds for j. Assuming that (4.10) is true for $j-1$, equation (4.6) gives

$$\frac{d}{dt}P_j(t) = \frac{\lambda e^{-\lambda t}(\lambda t)^{j-1}}{(j-1)!} - \lambda P_j(t).$$

From this,

$$e^{\lambda t}\left(\frac{d}{dt}P_j(t)+\lambda P_j(t)\right)=\frac{\lambda(\lambda t)^{j-1}}{(j-1)!}.$$

This equation may be rewritten as

$$\frac{d}{dt}\left(P_j(t)e^{\lambda t}\right)=\frac{\lambda(\lambda t)^{j-1}}{(j-1)!},$$

and integration of both sides of this equation gives

$$P_j(t)e^{\lambda t}=\frac{(\lambda t)^j}{j!}+C,$$

for some constant C. From (4.7) it follows that $C=0$. Thus

$$P_j(t)=\frac{e^{-\lambda t}(\lambda t)^j}{j!},\quad j=0,1,2,\ldots. \tag{4.11}$$

This completes the induction, showing that at time t the random variable N has a Poisson distribution with parameter λt.

Conditions 1 and 2 are often taken as giving a mathematical definition of the concept of "randomness," and since many calculations in bioinformatics, some of which are described later in this book, make the randomness assumption, the Poisson distribution arises often.

4.2 The Poisson and the Binomial Distributions

An informal statement concerning the way in which the Poisson distribution arises as a limiting case of the binomial was made in Section 1.3.7. A more formally correct version of this statement is as follows. If in the binomial distribution (1.8) we let $n\to+\infty$, $p\to 0$, with the product np held constant at λ, then for any y, the binomial probability in (1.8) approaches the Poisson probability in (1.22). This may be proved by writing the binomial probability (1.8) as

$$\frac{1}{y!}(np)((n-1)p)\cdots((n-y+1)p)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-y}. \tag{4.12}$$

Fix y and λ and write $p=\lambda/n$. Then as $n\to\infty$, each term in the above product has a finite limit as $n\to+\infty$: Terms of the form $(n-i)\lambda/n$ approach λ for any i, and

$$\left(1-\frac{\lambda}{n}\right)^n\to e^{-\lambda},$$

(see (B.3)), and finally,

$$\left(1-\frac{\lambda}{n}\right)^{-y}\to 1.$$

Therefore, the expression in (4.12) approaches

$$\lambda^y e^{-\lambda}/y! \tag{4.13}$$

as $n \to +\infty$, and this is the Poisson probability (1.22).

4.3 The Poisson and the Gamma Distributions

There is an intimate connection, implied by equation (4.9), between the Poisson distribution and the exponential distribution. The (random) time until the first event occurs in a Poisson process with parameter λ is given by the exponential distribution with parameter λ. To see this, let $F(t)$ be the probability that the first event occurs before time t. Then the density function for the time until the first occurrence is the derivative $\frac{d}{dt}F(t)$. From (4.9), $F(t) = 1 - P_0(t) = 1 - e^{-\lambda t}$. Therefore, $\frac{d}{dt}F(t) = \lambda e^{-\lambda t}$. This is the exponential distribution (1.66), with notation changed from x to t.

It can also be shown that the distribution of the time between successive events is given by the exponential distribution. Thus the (random) time until the kth event occurs is the sum of k independent exponentially distributed times. The material surrounding (2.23) shows that this sum has the gamma distribution (1.75). Let t_0 be some fixed value of t. Then if the time until the kth event occurs exceeds t_0, the number of events occurring before time t_0 is less than k, and conversely. This means that the probability that $k-1$ or fewer events occur before time t_0 must be identical to the probability that the time until the kth event occurs exceeds t_0. In other words it must be true that

$$e^{-\lambda t_0}\left(1 + (\lambda t_0) + \frac{(\lambda t_0)^2}{2!} + \cdots + \frac{(\lambda t_0)^{k-1}}{(k-1)!}\right) = \frac{\lambda^k}{\Gamma(k)}\int_{t_0}^{+\infty} x^{k-1}\, e^{-\lambda x} dx. \tag{4.14}$$

This equation can also be established by repeated integration by parts of the right-hand side.

4.4 The Pure-Birth Process

In the derivation of the Poisson distribution (4.10) it was assumed that the probability of an event in any time interval $(t, t+h)$ is independent of the number of events that have occurred up to time t. In several cases of biological interest this is not a reasonable assumption, and the concept of an event is not quite appropriate. Instead, some random variable is considered whose initial value is k (usually $k \geq 1$). In the pure-birth process it is assumed that, given that the value of the random variable at time t is j, the probability that it increases to $j+1$ in a short time interval $(t, t+h)$ is $\lambda_j h$, where (as with the Poisson process) we ignore terms of order $o(h)$.

(The Poisson case arises when $\lambda_j = \lambda$, that is λ_j is independent of j.) Following the same procedures as those carried out above for the Poisson process, we arrive at the infinite set of differential equations

$$\frac{d}{dt}P_k(t) = -\lambda_k P_k(t), \tag{4.15}$$

$$\frac{d}{dt}P_j(t) = \lambda_{j-1}P_{j-1}(t) - \lambda_j P_j(t), \quad j = k+1, k+2, k+3, \ldots. \tag{4.16}$$

for the probability that the random variable of interest takes the value j at time t. We now outline two examples of this process, and observe that in neither case does the value of the random variable at time t follow the Poisson distribution.

The Yule process. In the Yule process it is assumed that $\lambda_j = j\lambda$, for some constant λ. The motivation for this choice is that in some populations it is reasonable to assume that if the current size of the population is j, the probability that it increases to size $j+1$ in a short time interval is essentially proportional to j. It is easy to see that in this case the solution of equations (4.15) and (4.16) is

$$P_j(t) = \binom{j-1}{j-k} e^{-k\lambda t}\left(1 - e^{\lambda t}\right)^{j-k}, \quad j = k, k+1, \ldots. \tag{4.17}$$

The polymerase chain reaction (PCR). The polymerase chain reaction is a very important method used widely to amplify a comparatively short sequence of DNA. Here we model the length of the product, or "amplicon," of this reaction, considering a simplified version of the process described by Velikanov and Krapal (1999).

A primer, of initial length k (usually about $20-30$ base pairs), is used to initiate the reaction. The product of the reaction is formed by sequential additions of single base pair units to the primer. This addition forms a pure-birth process, and the function λ_j in the pure birth process is assumed here, for simplicity, to be of the form $m - j$. (This assumption involves a re-scaling of the time axis: Velikanov and Krapal (1999) provide a more general treatment.) The form of this function implies that, once the length of the product reaches the value m, no further increase in its length is possible. Equation (4.16) now becomes

$$\frac{d}{dt}P_j(t) = (m-j-1)P_{j-1}(t) - (m-j)P_j(t), \quad j = k+1, k+2, \ldots \tag{4.18}$$

The joint solution of this equation and equation (4.15), subject to the condition $P_k(0) = 1$, is

$$P_j(t) = \binom{m-k}{j-k}\left(1 - e^{-t}\right)^{j-k} e^{-(m-j)t}, \quad j = k, k+1, \ldots \tag{4.19}$$

The length of the additional material formed from the primer is $j-k$, and writing $i=j-k, n=m-k$, equation (4.19) can be written as

$$P_i(t)=\binom{n}{i}\left(1-e^{-t}\right)^i e^{-(n-i)t}, \quad i=0,1,2,\ldots,n. \tag{4.20}$$

It is interesting that this solution for i and the solution to the Yule process are both in the form of the binomial distribution. Equation (4.20), the moments of the binomial distribution as given in Table 1.1 and the linearity properties of means and variances listed in Sections 1.4 and 1.5 show, for example, that at time t, $E(j)=k+(m-k)(1-e^{-t})$ and that $\mathrm{Var}(j)=(m-k)(e^{-t}-e^{-2t})$. This variance assumes its maximum value when $t=\log 2$.

4.5 Introduction to Finite Markov Chains

In this section we give a brief outline of the theory of a simple case of a discrete-time finite Markov chain. The focus is on material needed to discuss the construction of PAM matrices as described in Section 6.5.3. Further developments of Markov chain theory suitable for other applications, in particular for the evolutionary applications discussed in Chapter 14, are given in Chapter 11.

We introduce the simple discrete-time finite Markov chain in abstract terms as follows. Consider some finite discrete set S of possible "states," labeled $\{E_1, E_2, \ldots, E_s\}$. At each of the unit time points $t=1,2,3,\ldots$, a Markov chain process occupies one of these states. In each time step t to $t+1$, the process either stays in the same state or moves to some other state in S. Further, it does this in a probabilistic, or stochastic, way rather than in a deterministic way. That is, if at time t the process is in state E_j, then at time $t+1$ it either stays in this state or moves to some other state E_k according to some well-defined probabilistic rule described in more detail below. This process follows the requirements of a simple Markov chain if it has the following properties.

(i) The *Markov* property. If at some time t the process is in state E_j, the probability that one time unit later it is in state E_k depends only on E_j, and not on the past history of the states it was in before time t. That is, the current state is all that matters in determining the probabilities for the states that the process will occupy in the future.

(ii) The *temporally homogeneous transition probabilities* property. Given that at time t the process is in state E_j, the probability that one time unit later it is in state E_k is independent of t.

More general Markov processes relax one or both requirements, but we assume throughout this chapter that the above properties hold.

The concept of "time" used above is appropriate if, for example, we consider the evolution through time of the nucleotide at a given site in some population. Aspects of this process are discussed later in this book. However, the concept of time is sometimes replaced by that of "space." As an example, we may consider a DNA sequence read from left to right. Here there would be a Markov dependence between nucleotides if the nucleotide type at some site depended in some way on the type at the site immediately to its left. Aspects of the Markov chains describing this process are also discussed later in this book. Because Markov chains are widely applicable to many different situations, it is useful to describe the properties of these chains in abstract terms rather than in the concrete terms appropriate to one specific application.

In many cases the Markov chain process describes the behavior of the value of a random variable changing through time. For example, in reading a DNA sequence from left to right this random variable might be the excess of purines over pyrimidines so far observed at any point. Because of this it is often convenient to adopt a different terminology and to say that the value of the random variable is j rather than saying that the state occupied by the process is E_j. We use both forms of expression below, and also, when no confusion should arise, we abuse terminology by using expressions like "the random variable is in state E_j."

4.6 Transition Probabilities and the Transition Probability Matrix

Suppose that at time t a Markovian random variable is in state E_j. We denote the probability that at time $t+1$ it is in state E_k by p_{jk}, called the *transition probability* from E_j to E_k. In writing this probability in this form we are already using the two Markov assumptions described above: First, no mention is made in the notation p_{jk} of the states that the random variable was in before time t (the memoryless property), and second, t does not occur in the notation p_{jk} (the time homogeneity property).

It is convenient to group the transition probabilities p_{jk} into the so-called *transition probability matrix*, or more simply the transition matrix, of the Markov chain. We denote this matrix by P, and write it as

$$P = \begin{array}{c} \\ (\text{from } E_1) \\ (\text{from } E_2) \\ \vdots \\ (\text{from } E_s) \end{array} \begin{array}{c} \begin{array}{cccccc} (\text{to } E_1) & (\text{to } E_2) & (\text{to } E_3) & \cdots & (\text{to } E_s) \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1s} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{s1} & p_{s2} & p_{s3} & \cdots & p_{ss} \end{bmatrix} \end{array}. \tag{4.21}$$

The rows and columns of P are in correspondence with the states E_1, $E_2, \ldots, E_s$, so these states being understood, P is usually written in the simpler form

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1s} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{s1} & p_{s2} & p_{s3} & \cdots & p_{ss} \end{bmatrix}. \tag{4.22}$$

Any row in the matrix corresponds to the state *from* which the transition is made, and any column in the matrix corresponds to the state *to* which the transition is made. Thus the probabilities in any particular row in the transition matrix must sum to 1. However, the probabilities in any given column do not have to sum to anything in particular.

It is also assumed that there is some *initial* probability distribution for the various states in the Markov chain. That is, it is assumed that there is some probability π_i that at the initial time point the Markovian random variable is in state E_i. A particular case of such an initial distribution arises when it is known that the random variable starts in state E_i, in which case $\pi_i = 1$, $\pi_j = 0$ for $j \neq i$. In principle the initial probability distribution and the transition matrix P jointly determine all the properties the entire process. In practice, many properties are not found easily, or if found are obtained by special methods.

The probability that the Markov chain process moves from state E_i to state E_j after two steps can be found by matrix multiplication. It is this fact that makes much of Markov chain theory an application of linear algebra. The argument is as follows.

Let $p_{ij}^{(2)}$ be the probability that if the Markovian random variable is in state E_i at time t, then it is in state E_j at time $t+2$. We call this a *two-step* transition probability. Since the random variable must be in some state k at the intermediate time $t+1$, summation over all possible states at time $t+1$ and use of equation (1.94) gives

$$p_{ij}^{(2)} = \sum_k p_{ik} p_{kj}.$$

The right-hand side in this equation is the (i, j) element in the matrix P^2. Thus if the matrix $P^{(2)}$ is defined as the matrix whose (i, j) element is $p_{ij}^{(2)}$, then the (i, j) element in $P^{(2)}$ is equal to the (i, j) element in P^2. This leads to the identity

$$P^{(2)} = P^2.$$

Extension of this argument to an arbitrary number n of steps gives

$$P^{(n)} = P^n. \tag{4.23}$$

That is, the "n-step" transition probabilities are given by the entries in the nth power of P.

4.7 Markov Chains with Absorbing States

Some Markov chains have absorbing states. These can be recognized by the appearance of one or more 1's on the main diagonal of the transition matrix. If there are no 1's on the main diagonal, then there are no absorbing states. For the Markov chains with absorbing states that we consider, sooner or later some absorbing state will be entered, never thereafter to be left. The two questions we are most interested in regarding these Markov chains are:

(i) If there are two or more absorbing states, what is the probability that a specified absorbing state is the one eventually entered?

(ii) What is the mean time until one or another absorbing state is eventually entered?

We shall address these questions in detail in Chapter 11. In the remainder of this chapter we discuss only certain aspects of the theory of Markov chains with no absorbing states, focusing on the theory needed for the construction of substitution matrices, to be discussed in more detail in Chapter 6.

4.8 Markov Chains with No Absorbing States

The questions of interest about a Markov chain with no absorbing state are quite different from those asked when there are absorbing states.

In order to simplify the discussion, we assume in the remainder of this chapter that all Markov chains discussed are *finite*, *aperiodic*, and *irreducible*.

Finiteness means that there is a finite number of possible states. The aperiodicity assumption is that there is no state such that a return to that state is possible only at t_0, $2t_0$, $3t_0, \ldots$ transitions later, where t_0 is an integer exceeding 1. If the transition matrix of a Markov chain with states E_1, E_2, E_3, E_4 is, for example,

$$P = \begin{bmatrix} 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.3 & 0.7 \\ 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \end{bmatrix}, \tag{4.24}$$

then the Markov chain is periodic. If the Markovian random variable starts (at time 0) in E_1, then at time 1 it must be either in E_3 or E_4, at time 2

it must be in either E_1 or E_2, and in general it can visit only E_1 at times $2, 4, 6, \ldots$. It is therefore periodic. The aperiodicity assumption holds for essentially all applications of Markov chains in bioinformatics, and we often take aperiodicity for granted without any explicit statement being made.

The irreducibility assumption implies that any state can eventually be reached from any other state, if not in one step then after several steps. Except for the case of Markov chains with absorbing states, the irreducibility assumption also holds for essentially all applications in bioinformatics.

4.8.1 Stationary Distributions

Suppose that a Markov chain has transition matrix P and that at time t the probability that the process is in state E_j is φ_j, $j = 1, 2, \ldots, s$. This implies that the probability that at time $t + 1$ the process is in state j is $\sum_{k=1}^{s} \varphi_k p_{kj}$. Suppose that for every j these two probabilities are equal, so that

$$\varphi_j = \sum_{k=1}^{s} \varphi_k p_{kj}, \quad j = 1, 2, \ldots, s. \tag{4.25}$$

In this case the probability distribution $(\varphi_1, \varphi_2, \ldots, \varphi_s)$ is said to be *stationary*; that is, the probability that the process is in state E_j has not changed between times t and $t + 1$, and therefore will never change. Despite this, the state occupied by the process can of course change from one time point to the next.

It will be shown in Chapter 11 that for finite aperiodic irreducible Markov chains there is a unique distribution satisfying (4.25). This is then called the *stationary distribution* of the Markov chain. When we discuss stationary distributions in this book, we assume that they relate to finite aperiodic irreducible Markov chains, and thus exist and are unique.

If the row vector $\boldsymbol{\varphi}'$ is defined by

$$\boldsymbol{\varphi}' = (\varphi_1, \varphi_2, \ldots, \varphi_s), \tag{4.26}$$

then in matrix and vector notation, the set of equations in (4.25) becomes

$$\boldsymbol{\varphi}' = \boldsymbol{\varphi}' P. \tag{4.27}$$

The prime here is used to indicate the transposition of the row vector into a column vector. The vector $(\varphi_1, \varphi_2, \ldots, \varphi_s)$ must also satisfy the equation $\sum_k \varphi_k = 1$. In vector notation, this is the equation

$$\boldsymbol{\varphi}' \mathbf{1} = 1, \tag{4.28}$$

where $\mathbf{1} = (1, 1, \ldots, 1)'$. Equations (4.27) and (4.28) can then be used to find the stationary distribution when it exists. In this process one of the equations in (4.27) is redundant and can be omitted. An example is given in the next section.

In Chapter 11 we shall show that if the Markov chain is finite, aperiodic, and irreducible, then as n increases, $P^{(n)}$ approaches the matrix

$$\begin{bmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_s \\ \varphi_1 & \varphi_2 & \cdots & \varphi_s \\ \varphi_1 & \varphi_2 & \cdots & \varphi_s \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1 & \varphi_2 & \cdots & \varphi_s \end{bmatrix}, \tag{4.29}$$

where $(\varphi_1, \varphi_2, \ldots, \varphi_s)$ is the stationary distribution of the Markov chain.

The form of this matrix shows that no matter what the starting state was, or what was the initial probability distribution of the starting state, the probability that n time units later the process is in state j is increasingly closely approximated, as $n \to \infty$, by the value φ_j.

There is another implication, relating to long-term averages, of the calculations above. That is, if a Markov chain is observed for a very long time, then the proportion of times that it is observed to be in state E_j should be approximately φ_j, for all j.

4.8.2 Example

Consider the Markov chain with transition probability matrix given by

$$P = \begin{bmatrix} 0.6 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.5 & 0.1 \\ 0.1 & 0.3 & 0.1 & 0.5 \end{bmatrix}. \tag{4.30}$$

For this example the vector equation (4.27) consists of four separate linear equations in four unknowns. As noted above, when used jointly with (4.28) they form a redundant set of equations and any one of them can be discarded. Omission of the last equation in (4.27) leads to

$$\begin{aligned} 0.6\varphi_1 + 0.1\varphi_2 + 0.2\varphi_3 + 0.1\varphi_4 &= \varphi_1, \\ 0.1\varphi_1 + 0.7\varphi_2 + 0.2\varphi_3 + 0.3\varphi_4 &= \varphi_2, \\ 0.2\varphi_1 + 0.1\varphi_2 + 0.5\varphi_3 + 0.1\varphi_4 &= \varphi_3, \\ \varphi_1 + \varphi_2 + \varphi_3 + \varphi_4 &= 1. \end{aligned}$$

To four decimal place accuracy, these four simultaneous equations have the solution

$$\boldsymbol{\varphi}' = (0.2414, 0.3851, 0.2069, 0.1667). \tag{4.31}$$

This is the stationary distribution corresponding to the matrix P given in (4.30). In informal terms, from the point of view of long-term averages, over a long time period the random variable should spend about 24.14% of the time in state E_1, about 38.51% of the time in state E_2, and so on.

The rate at which the rows in $P^{(n)}$ approach this stationary distribution can be assessed from the following values:

$$P^{(2)} = \begin{bmatrix} 0.42 & 0.20 & 0.24 & 0.14 \\ 0.16 & 0.55 & 0.15 & 0.14 \\ 0.25 & 0.29 & 0.32 & 0.14 \\ 0.16 & 0.39 & 0.15 & 0.30 \end{bmatrix}, \tag{4.32}$$

$$P^{(4)} \cong \begin{bmatrix} 0.2908 & 0.3182 & 0.2286 & 0.1624 \\ 0.2151 & 0.4326 & 0.1899 & 0.1624 \\ 0.2538 & 0.3569 & 0.2269 & 0.1624 \\ 0.2151 & 0.4070 & 0.1899 & 0.1880 \end{bmatrix}, \tag{4.33}$$

$$P^{(8)} \cong \begin{bmatrix} 0.24596 & 0.37787 & 0.20961 & 0.16656 \\ 0.23873 & 0.38946 & 0.20525 & 0.16656 \\ 0.24309 & 0.38223 & 0.20812 & 0.16656 \\ 0.23873 & 0.38880 & 0.20525 & 0.16721 \end{bmatrix}, \tag{4.34}$$

$$P^{(16)} \cong \begin{bmatrix} 0.24142 & 0.38494 & 0.20692 & 0.16667 \\ 0.24135 & 0.38510 & 0.20688 & 0.16667 \\ 0.24140 & 0.38503 & 0.20691 & 0.16667 \\ 0.24135 & 0.38510 & 0.20688 & 0.16667 \end{bmatrix}. \tag{4.35}$$

After 16 time units, the stationary distribution has, for all practical purposes, been reached. The discussion in Chapter 11 shows how the rate at which this convergence occurs can be calculated in a more informative manner.

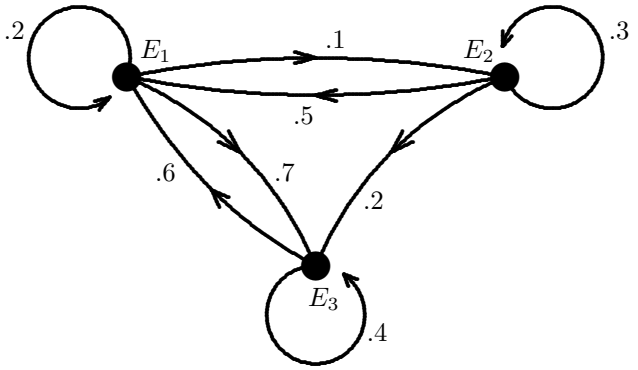
4.9 The Graphical Representation of a Markov Chain

It is often convenient to represent a Markov chain by a directed graph. A directed graph is a set of "nodes" and a set of "edges" connecting these nodes. The edges are "directed," that is, they are marked with arrows giving each edge an orientation from one node to another.

We represent a Markov chain by identifying the states with nodes and the transition probabilities with edges. Consider, for example, the Markov chain with states E_1, E_2, and E_3 and with probability transition matrix

$$\begin{bmatrix} .2 & .1 & .7 \\ .5 & .3 & .2 \\ .6 & 0 & .4 \end{bmatrix}.$$

This Markov chain is represented by the following graph:



Notice that we do not draw the edge if its corresponding transition probability is known to be zero, as is the case in this example with the transition from E_3 to E_2.

A graph helps us capture information at a glance that might not be so apparent from the transition matrix itself. Sometimes it is also convenient to include a *start state*; this is a dummy state that is visited only once, at the beginning. Therefore, all transition probabilities into the start state are zero. The transition probabilities out of the start state are given by the initial distribution of the Markov chain. If the Markov chain starts at time $t = 0$, we can think of the start state as being visited in time $t = -1$. We can further have an *end state*, which stops the Markov chain when visited.

We refer to the graph structure, without any probabilities, as the *topology* of the graph. Sometimes the topology of a model is known, but the various probabilities are unknown.

We will use these definitions when we discuss hidden Markov models in Chapter 12.

4.10 Modeling

There are many applications of the homogeneous Poisson process in bioinformatics. However, the two key assumptions made in the derivation of the Poisson distribution formula (4.10), namely homogeneity and independence, do not always hold in practice. Similarly, there are many applications of Markov chains in the literature, in particular in the evolutionary processes discussed in Chapter 14. Many of these applications also make assumptions, specifically the two Markov assumptions stated in Section 4.5. The modeling assumptions made in the evolutionary context are discussed further in Section 15.9.

In the context of DNA or protein sequence analysis, where time is replaced by position along the sequence, it is very likely that neither of these two Markov assumptions is correct. The data from chromosome 22 in humans (Dunham et al. (1999)) makes it apparent that the probability that the nucleotide a is next followed by g depends to some extent on the current location in the chromosome. Further, it is likely that the Markov chain memoryless assumption does not hold: The probability that the nucleotide a is next followed by g might well depend on the nucleotide (or nucleotides) immediately preceding a. Tests for this possibility are discussed in Section 5.2: Nevertheless these tests are often not applied, and the memoryless Markov chain theory is often assumed when its applicability is uncertain.

The construction of phylogenetic trees, both by algorithmic methods and by methods involving Markov chain evolutionary models, involves many assumptions, both explicit and implicit. An example of quite different phylogenetic trees arising from different models is given in Section 15.8. A discussion of various statistical tests for appropriate evolutionary models is discussed in Section 15.9.

Given that modeling assumptions made for both Poisson processes and Markov chains often do not hold exactly, one might ask why they are made and why there is such an extensive Poisson process and Markov chain literature. Mathematical models generally make simplifying assumptions about properties of the phenomena being modelled. This concern opens up the question of why we model natural phenomena with mathematics if we cannot do so with complete accuracy. In fact, it is not necessarily desirable that we attempt to make an extremely accurate model of reality. The more closely any phenomenon is modelled, the more complicated the model becomes. If a mathematical model becomes too complicated, then solving the equations necessary to find answers to the questions we wish to ask can become intractable. Therefore, we are almost always faced with the task of finding a middle ground between tractable simple models and intractable complex models. The key point is that a model need only capture enough of the true complexity of a situation to serve our purposes, whatever they might be.

Finding this middle ground, however, is not an easy task: Being able to extract the essence of a complex reality in a simplified model that then allows a successful mathematical analysis requires some skill and experience.

In biology it is not always possible to evaluate a model's efficacy directly. Rather, a model is often tested on how well it performs its job. Sometimes benchmarks can be well defined, but often efficacy is not easy to verify empirically, and subjectivity can enter in. This is an unfortunate but usually unavoidable problem.

To illustrate this we consider the early versions of the widely used BLAST procedure discussed in more detail in Chapter 10. One of the simplifying assumptions used is that nucleotides (or amino acids) are identically and

independently distributed along a DNA (or protein) sequence. Current data show that this assumption is false. Nevertheless, this simple BLAST procedure does work, in that the model captures enough of biological reality to be effective.

A further aspect of the modeling process is that we do not expect any model to be the final one used. Any given process is often initially modelled using several simplifying assumptions, and then more refined models are introduced as time goes on. Indeed, applications of the simple models often indicate those areas in which more precise modeling is needed. Various updates of the BLAST procedure exemplify this. Recent versions of BLAST remove some of the simplifying assumptions made in earlier versions and provide an example of the joint evolution of models and data analysis. Unfortunately, the mathematical theory involved in these more sophisticated versions is far more complicated than that in the simpler BLAST theory, and we shall only outline it in this book.

Not every problem we might wish to solve with a model has a happy middle ground where our assumptions find a workable balance between tractability and reality. Thus while we should be willing to accept simplifying assumptions, we should always be on the lookout for oversimplifications, especially those that are not sufficiently backed up by testing for the efficacy of the model used. Model testing is an active area of statistical research in bioinformatics, and aspects of model testing, especially in the evolutionary and phylogenetic tree contexts, are discussed further in Section 15.9.

Problems

4.1. Prove (4.14) by repeated integration by parts of the right-hand side.

4.2. Events occur in a Poisson process with parameter λ. Given that 10 events occur in the time period $[0, 2]$, what is the probability that 6 of these events occur in the time period $[0, 1]$? Given that 6 of these events occur in the time period $[0, 1]$, what is the probability that 10 of these events occur in the time period $[0, 2]$?

4.3. ("Competing Poissons.") Suppose that events occur as described in Section 4.1, but that now each event is of one of k types, labeled types $1, 2, \ldots, k$. The type of any event is independent of the type of any other event. The probability that any event is of type i is p_i. Equation (4.10) continues to give the probability that exactly j events occur in the time period $(0, t)$. Assuming this,

(i) Find the (marginal) probability that j_i events of type i occur in the time period $(0, t)$.

(ii) Find the probability that j_i $(i = 1, \ldots, k)$ events of type i occur occur in the time period $(0, t)$.

(iii) Find the joint conditional probability that j_i $(i = 1, \ldots, k)$ events of type i occur in the time period $(0, t)$, given that j events in total occur in this time period. Relate your answer to expression (2.45).

4.4. The transition matrix of a Markov chain is

$$\begin{bmatrix} .7 & .3 \\ .4 & .6 \end{bmatrix}.$$

Find the stationary distribution of this Markov chain.

4.5. *Continuation.* If the initial probability distribution (at time 0) is $(.8, .2)$, what is the probability that at time 3 the state occupied is E_1?

4.6. The transition matrix of a Markov chain is

$$\begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}.$$

Find the stationary distribution of this Markov chain in terms of a and b, and interpret your result.

4.7. The transition matrix of a Markov chain is

$$\begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}.$$

Use induction on n (see Section B.18) to show that the probability that the Markov chain revisits the initial state at the nth transition is

$$p_{ii}^{(n)} = \frac{1}{4} + \frac{3}{4}(-\frac{1}{3})^n.$$

(This result is needed for Problem 14.7.)

4.8. Use equations (4.27) and (4.23) to show that the stationary distribution $\boldsymbol{\varphi}'$ satisfies the equation

$$\boldsymbol{\varphi}' = \boldsymbol{\varphi}' P^{(n)}, \tag{4.36}$$

for any positive integer n. For the numerical example in Section 4.8.2, use the expression for $P^{(2)}$ given in equation (4.32) and the expression for $\boldsymbol{\varphi}'$ given in (4.31) to check this claim for the case $n = 2$.

Why does equation (4.36) "make sense"?

4.9. Show that if the transition matrix of an irreducible, aperiodic, finite Markov chain is symmetric, then the stationary distribution is a (discrete) uniform distribution.

4.10. Show that if the transition matrix P of a Markov chain is of the circulant form

$$P = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & \cdots & a_{s-3} & a_{s-2} & a_{s-1} & a_s \\ a_s & a_1 & a_2 & a_3 & \cdots & a_{s-4} & a_{s-3} & a_{s-2} & a_{s-1} \\ a_{s-1} & a_s & a_1 & a_2 & \cdots & a_{s-5} & a_{s-4} & a_{s-3} & a_{s-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ a_4 & a_5 & a_6 & a_7 & \cdots & a_s & a_1 & a_2 & a_3 \\ a_3 & a_4 & a_5 & a_6 & \cdots & a_{s-1} & a_s & a_1 & a_2 \\ a_2 & a_3 & a_4 & a_5 & \cdots & a_{s-2} & a_{s-1} & a_s & a_1 \end{bmatrix}, \tag{4.37}$$

where $a_j > 0$ for all j, then the stationary distribution is a (discrete) uniform distribution.

4.11. Suppose that the transition matrix of a Markov chain is

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & q & 0 & p & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{4.38}$$

Show that this Markov chain is periodic. Despite this fact, solve equations (4.27) and (4.28) for the case $p = q$.

4.12. Suppose that a transition matrix P is of size $2s \times 2s$ and can be written in the partitioned form

$$P = \left[\begin{array}{c|c} 0 & A \\ \hline B & 0 \end{array}\right],$$

where A and B are both $s \times s$ matrices. Use this expression to find formulae for (i) $P^{(2n)}$, (ii) $P^{(2n+1)}$ in terms of the matrices A and B, and interpret your results.

4.13. Suppose that a finite Markov chain is irreducible and that there exists at least one state E_i such that $p_{ii} > 0$. Show that the Markov chain is aperiodic.

4.14. (More difficult). Any $s \times s$ matrix of non-negative numbers for which all rows sum to 1 can be regarded as the transition matrix of some Markov

chain. Is it true that any such matrix can be the two-step transition matrix of some Markov chain? *Hint:* Consider the case $s = 2$. Write down the general form of a 2×2 Markov chain matrix and find when the sum of the diagonal terms is minimized.