Chapter 5

Markov Chain Models

5.1 Introduction

A model for a random process, $X = \{X_1, X_2, \ldots, X_N\}$, is a set of rules that characterize the joint probability distribution between all its random variables. So far, the primary model we have used is the simple one in which $X_1, X_2, \ldots, X_N$ are i.i.d. This is the model for repeated independent random samples and for a sequence of coin tosses. We have proposed it also as a model, which we called the IID sites model, for the bases along a DNA segment, and this was the foundation for our counting process models of cut site locations of restriction enzymes. However, it is clear that many applications require models with dependencies between the X_i's. The IID sites model, for example, does not seem realistic; coding DNA comes in codon triplets that spell out the order of amino acids in proteins, so one should expect dependencies, at least between nearby sites. This chapter takes the first step away from the i.i.d. set-up. It presents the definition and the elementary properties of a class of processes called Markov chains. The probabilistic behavior of a Markov chain is determined just by the dependencies between successive random variables—between X_1 and X_2, between X_2 and X_3, etcetera. Despite this restriction, Markov chains are rich in behavior, amenable to analysis, and adaptable to many applications; hence they are centrally important to applied and theoretical probability.

In this chapter, the random processes we study will be finite or infinite sequences $X \stackrel{\triangle}{=} \{X_t\}_{t\geq 1} = \{X_1, X_2, \ldots\}$ of random variables taking values in a discrete set $\mathcal{E}$. The elements of $\mathcal{E}$ will be called *states*, and, accordingly, $\mathcal{E}$ will be refered to as the *state space.* In Markov chain applications, the index t in X_t is usually thought of as a time index, and X_t represents the

state of the process at "time" t.

Here are some situations in which models with dependent X_i's are natural:

Example 5.1. (Games of chance) We want to model a player's fortune in a game of chance in which the stake on each play is one dollar. We let X_t denote the player's fortune at the end of play number t. In this case, a convenient state space is the set of all integers, $\mathcal{E} \stackrel{\triangle}{=} \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. The fortune X_{t+1} at time $t+1$ will clearly depend on the fortune at the previous plays, so an i.i.d. model is not appropriate. $\diamond$

Example 5.2. (Population dynamics) We are interested in studying population growth in a model that includes random effects. In this case, t might label the generation, and X_t is the size of the population in generation t. An appropriate state space is the set $\mathcal{E} \stackrel{\triangle}{=} \{0, 1, 2, \ldots\}$ of all non-negative integers. Since the population evolves, its values in different generations, even if random, will exhibit dependencies. $\diamond$

Example 5.3. (Population genetics) Suppose we are studying the evolution of the frequency of an allele A in a population which has the same size N in each generation. We let t index the generation and take Y_t to be the number of A alleles in generation t. The state space is $\mathcal{E} = \{0, 1, \ldots, 2N\}$ since the possible number of alleles can range from 0 to $2N$. This is precisely the framework of the Wright-Fisher and Moran models discussed in Chapter 3, except there the process was the frequency $X_t = Y_t/(2N)$ of allele A, evolving in the state space $\{0, 1/2N, 2/2N, \ldots, 1\}$. The Wright-Fisher and Moran models are Markov chain models in which the the number Y_t of A alleles in generation t determines the probability distribution of Y_{t+1}.

$\diamond$

Example 5.4. (Sequence models) We wish to model the bases appearing along a randomly drawn DNA segment as we read it starting from the $5'$ end. In this case, the index t stands for the t^{th} site from the $5'$ end, and X_t is the base at site t. Here, $\mathcal{E}$ is the DNA alphabet $\{A, T, C, G\}$. In this example, t is not a time index, but it does capture the idea of reading the sequence in a specific direction.

Example 5.5 (Protein evolution) Remember from Chapter 1 that a protein is a linear chain of amino acids and that genes carry the instructions for making proteins in the order in which the bases appear in coding DNA. Consider a protein in a species that existed in the distant past. Suppose that we follow a line of descent from this species down to a present day

species. As generation succeeds generation, random mutations in the DNA will cause mutations in the protein. A mutation can be a substitution of one amino acid for another, an insertion of a new amino acid in the protein sequence, or a deletion. If a mutation leads to a selective advantage or is selectively neutral it may become fixed in the population. In this way, proteins evolve, and since mutations are random events, we can think of this evolution as a sequence of random variables. To simplify, let us focus only on one site along the protein and assume that the amino acid there evolves only by substitution. Substitutions occur slowly, so we don't want to track the amino acid in every generation. Instead, we wait for a time long enough so that there is some reasonable chance of a substitution. For example, a unit of 1 PAM (an abbreviation, in modified order, of Accepted Point Mutation) is defined to be an interval in which one can expect about 1% of a protein's amino acid to change by evolution. (It is not a universal or exact unit, but depends on the protein under study.) Then we let $X_1, X_2, X_3, \ldots$ represent the amino acid at our site after successive 1 PAM intervals. It is a random process whose state space is the 20 letter amino acid alphabet—see Chapter 1. ◇

5.2 Markov chains; definition

Let $X = \{X_1, X_2, \ldots\}$ be a random process in the discrete state space $\mathcal{E}$. It is called a Markov chain if the conditional probabilities between the outcomes at different times satisfy the *Markov property*, which we now explain. Consider a time t and the event

$$\{X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_1 = x_1\},$$

for some sequence of states $x_t, x_{t-1}, \ldots, x_t$. This is a record of the entire history of the process up to and including the time t. We have written it in reverse-time order because we want to think of t as the present time and to express the event starting from the present and moving back into the past. The conditional probability

$$\mathbb{P}\left(X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_1 = x_1\right)$$

thus represents the probability of an event one step into the future beyond time t, conditioned on the entire past of the process up to t. On the other hand

$$\mathbb{P}\left(X_{t+1} = x_{t+1} \mid X_t = x_t\right)$$

is the conditional probability of the future event given just the present. The Markov property is satisfied when these two conditional probabilities are equal.

Definition. The sequence $\{X_1, X_2, \ldots\}$ of $\mathcal{E}$-valued random variables is said to have the *Markov property* if

$$\mathbb{P}\left(X_{t+1}=x_{t+1} \mid X_t=x_t, X_{t-1}=x_{t-1}, \ldots, X_1=x_1\right) = \mathbb{P}\left(X_{t+1}=x_{t+1} \mid X_t=x_t\right), \tag{5.1}$$

for every sequence $x_1, \ldots, x_t, x_{t+1}$ of elements of $\mathcal{E}$ and for every $t \geq 1$.

A sequence of random variables with the Markov property is called a *Markov chain.*

If $X_1, X_2, \ldots$ is a Markov chain, and i and j are states in $\{\mathcal{E}\}$, the conditional probability

$$p_{ij}(t) \stackrel{\triangle}{=} \mathbb{P}\left(X_{t+1}=j \mid X_t=i\right)$$

is called the transition probability from i to j at time t. If the transition probabilities do not depend on time, we write them simply as p_{ij}, $i, j \in \mathcal{E}$ and we say that the Markov chain is *time-homogeneous.*

The Markov property says once the value X_t at the present time is known, all further past values of the process have no relevance to conditional probabilities of future values. This is commonly expressed in words as "the future is independent of the past given the present."

We shall deal exclusively with time-homogeneous Markov chains. The notation p_{ij} will be our standard notation for the transition probabilities.

Suppose that the state space $\mathcal{E}$ is finite and let us write it as $\mathcal{E} \stackrel{\triangle}{=} \{0, 1, \ldots, S\}$. Given a set of transition probabilities, it is often useful to collect them in a matrix,

$$A = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0S} \\ p_{10} & p_{11} & \cdots & p_{1S} \\ p_{20} & p_{21} & \cdots & p_{2S} \\ \vdots & \vdots & \vdots & \vdots \\ p_{S0} & p_{S1} & \cdots & p_{SS} \end{pmatrix}.$$

The matrix A is called, logically enough, a *transition probability matrix.* Notice that the row $[p_{i0}\, p_{i1}\, \cdots\, p_{iS}]$ represents all the transition probabilities out of state i. Therefore, the probabilities in the row must sum to 1. In

general, any matrix of non-negative numbers in which every row sums to one is called a transition probability matrix, or, sometimes, a *stochastic matrix.*

Example 5.6, (Random walk). We define a specific game of chance in which the total fortune of the player is a Markov chain. Imagine that on each play a coin with probability p of heads is tossed and that all tosses are independent. The player wins a dollar if the toss turns up heads and loses a dollar if the toss is tails. Denote the player's winning on play t by ξ_t. Then, summarizing what we have said mathematically, $\xi_1, \xi_2, \ldots$ are i.i.d. with distribution

$$\mathbb{P}(\xi_t = 1) = p \quad \text{and} \quad \mathbb{P}(\xi_t = -1) = 1 - p.$$

Suppose the player starts with some fixed initial fortune of X_0 dollars. Then the player's fortune immediately after play t is

$$X_t = X_0 + \sum_{s=1}^{t} \xi_s.$$

We claim that $X = \{X_1, X_2, \ldots\}$ is a Markov chain with transition probabilities

$$p_{ij} = \begin{cases} p, & \text{if } j = i+1; \\ 1-p, & \text{if } j = i-1; \\ 0, & \text{otherwise.} \end{cases}$$

The reason this is true is almost obvious once you observe that

$$X_{t+1} = X_t + \xi_{t+1},$$

and that ξ_{t+1} and X_t are independent. The former fact is just a restatement of the definition of the game: the total winnings X_{t+1} at time $t+1$ are the winnings X_t at time t plus whatever is lost or won on play $t+1$. The latter fact is just a consequence of the independence of ξ_{t+1} and $\xi_1, \xi_2, \ldots, \xi_t$, because X_t depends only on the plays up to time t. Together these two facts imply that if X_t is known, X_{t+1} depends only on the outcome of the next play, which is independent of the past, and this gives the Markov property. For example,

$$\begin{aligned} \mathbb{P}(X_{t+1} = i_t + 1 \mid X_t = i_t, \ldots, X_1 = i_1) &= \mathbb{P}(i_t + \xi_{t+1} = i_t + 1 \mid X_t = i_t, \ldots, X_1 = i_1) \\ &= \mathbb{P}(\xi_{t+1} = 1 \mid X_t = i_t, \ldots, X_1 = i_1) \\ &= \mathbb{P}(\xi_{t+1} = 1) = p. \end{aligned}$$

The first equality follows from $X_{t+1} = X_t + \xi_{t+1}$, the third from the substitution of the value for X_t, and the last from the independence of ξ_{t+1} and all previous fortunes. This calculation verifies the Markov property for the transition from a state i to state $i+1$ and shows that the transition probability $p_{i,i+1}$ is indeed p. The case of a decrease in fortune is handled by the same argument.

Althought we have used the language of games of chance, this model is usually called *random walk*. The process X_t represents the position of a walker at time integer times t. At each time, the walker moves to the right ($\xi_{t+1} = 1$) with probability p and to the left with probability q. $\diamond$

Example 5.3 (continued) In Chapter 2, we proposed the Wright-Fisher and Moran models for the evolution of allele frequencies in a finite population. Consider here that Wright-Fisher model, which was explicitly defined to be a Markov chain. We showed there that if i labels the state corresponding to i alleles in the population,

$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}, \qquad 0 \leq i, j \leq 2N.$$

The Markov chain property makes sense. The population of generation $t+1$ is obtained from generation t by a random mating, so once the number of A alleles in generation t is known, all further information about past populations is irrelevant to the conditional probability of states in generation $t+1$. $\diamond$

Example 5.4 (continued). A Markov chain sequence model for DNA. As an alternative to the IID sites model, we can introduce dependencies between sites by hypothesizing that the bases $X_1, X_2, \ldots$ form a Markov chain. To fully specify the model we need to give the transition probability matrix:

$$A = \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}.$$

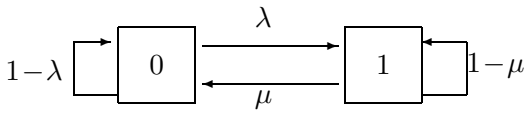
There is no *a priori* natural choice for the transition probabilities. We will discuss later how to estimate them from data. $\diamond$

Example 5.5 (continued). Let $X_1, X_2, \ldots$ denote the amino acid at a fixed site in a protein of an organism after successive intervals of 1 PAM, as we

follow a line of evolutionary descent. A Markov chain model for $X_1, X_2, \ldots$ is very natural. The Markov property says the conditional probabilities for the value of X_{t+1} given the past values of $X_t, X_{t-1}, \ldots, X_1$, should depend only on X_t. This makes sense; only the amino acid actually present at the site at time t should influence what substitutions are likely next. As in the previous example, to specify the model one must give values for the transition probabilities p_{xy}, which are the probabilities that amino acid y substitutes for amino acid x after 1 PAM unit of time. Again, these are estimated from data by a method we shall discuss later. ◇

There is a very convenient, short-hand, visual representation of a Markov chain called the *state transition diagram*. This is a graph whose vertices represent the different possible states in the state space. Arrows between vertices represent a transition between states and are labelled with the probability of that transition.

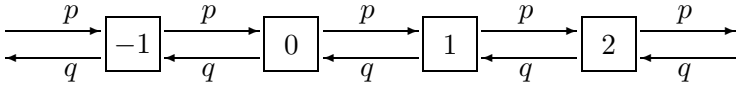
Example 5.7. Consider the diagram



This is a diagram for a *two-state Markov chain*. The two states are labeled by 0 and 1. The diagram shows that the probability p_{01} of a transition from state 0 to state 1 is λ, from state 0 back to 0 is $p_{00} = 1-\lambda$, etcetera. The state transition matrix for this chain is then,

$$A = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1-\lambda & \lambda \\ \mu & 1-\mu \end{pmatrix}. \qquad \diamond$$

Example 5.6, continued. Here is a piece of the state transition diagram for a random walk on the integers as defined in Example 5.6. The diagram continues to the left and right *ad infinitum.*



5.3 Estimating transition probabilities

Consider a Markov chain model of a random sequence $\{X_i\}$ with state space $\mathcal{E} = \{0, 1, \ldots, S\}$. The parameters of the Markov model are its transition probability matrix A and its initial distribution $p_i^{(1)}$, $i \in \mathcal{E}$. Sometimes the model prescribes an explicit set of transition probabilities. This was the case with the Wright-Fisher and Moran models of Chapter 3. There, formulas for the transition probabilities are derived from the assumption of random mating. However, in other types of applications, there may be no theoretical grounds for guessing numerical values of the transition probabilities. This is the case, for example, in setting up a Markov model of the sequence of bases along a DNA segment, as introduced in Example 5.4. In the absence of a theory that would predict how the probabilities of each succeeding base depend on the base at the previous site, one cannot propose transition probabilities *a priori*. Instead, it is necessary to collect data and use them to estimate the transition probabilities. How should this estimation be done? There is an obvious suggestion: estimate p_{ij} by the frequency of transitions from i to j in the data. We formalize and illustrate this estimate in this section. We will justify the estimate in terms of statistical theory later in the text after introducing the idea of a maximum likelihood estimate.

Imagine then that we have a process we are modeling as a Markov chain, and we observe M independent runs of the process, perhaps of varying length:

$$
\begin{aligned}
&x_{11}x_{12}x_{13}\ldots x_{1m_1}\\
&x_{21}x_{22}x_{23}\ldots x_{2m_2}\\
&\ldots\ldots\ldots\\
&x_{M1}x_{M2}x_{M3}\ldots x_{Mm_M}.
\end{aligned}
$$

We shall define the following statistics dependent on the data:

$$
\begin{aligned}
N_{ij} &\stackrel{\triangle}{=} \text{number of transitions from state } i \text{ to state } j \text{ in the data}\\
N_{i\cdot} &\stackrel{\triangle}{=} \sum_{t\in\mathcal{E}} N_{st} = \text{number of transitions starting from } i \text{ in data}\\
N_{\cdot j} &\stackrel{\triangle}{=} \sum_{s\in\mathcal{E}} N_{st} = \text{number of transitions ending in } j \text{ in data}\\
N &\stackrel{\triangle}{=} \sum_{ij\in\mathcal{E}} N_{ij} = \text{total number of transitions in data}\\
M_i &\stackrel{\triangle}{=} \text{number of times state } i \text{ appears as the first state in one of the runs}
\end{aligned}
$$

Notice that $M = \sum_{i \in \mathcal{E}} M_i$, the total number of runs. Also,

$$\begin{aligned} N_{i\cdot} &= \sum_j N_{ij}, \\ N_{\cdot j} &= \sum_i N_{ij}, \\ N &= \sum_{i \in \mathcal{E}} N_{i\cdot} = \sum_{j \in \mathcal{E}} N_{\cdot j}. \end{aligned}$$

In terms of these numbers, we will estimate the transition probability p_{ij} by

$$\hat{p}_{ij} = \frac{N_{ij}}{N_{i\cdot}}. \tag{5.2}$$

We will estimate the initial probability $\mathbb{P}(X_1 = i)$ by

$$\hat{\rho}_i^{(1)} = \frac{M_i}{M}. \tag{5.3}$$

We shall denote the matrix of estimated transition probabilities by $\hat{A}$. We can only effectively estimate the initial probabilities if M is reasonably large. For small M we must stick to estimating only transition probabilities, but for most purposes this is all we need anyway.

Example 5.8. We are building a Markov chain model of single-stranded DNA. Suppose we sample and sequence 6 DNA segments and obtain:

$$\begin{aligned} &AAACCCTGGCAATTCAGT \\ &ACCTGCGCCGTATATTATCAT \\ &GGCTCTCCAAG \\ &CCTTATATGGAAGAGG \\ &TTATTGC \\ &CCATGGC \end{aligned}$$

In this data it is easy to check that $M_A = 2$, $M_C = 2$, $M_G = 1$, and $M_T = 1$. Here, of course, $M = 6$.

Consider transitions from A to A. There are 3 in the first sequence, 1 in the third, 1 in the fourth, and none in the other sequences. Thus $N_{AA} = 5$. Continuing like this we can make a table of the counts , as follows (assuming I did not miscount!). (In this table the first row gives the counts for transitions out of A, that is, from A to A, from A to C, etc.; the second

row gives transitions out of state C, and so on.) Notice that the values of $N_{i\cdot}$ are obtain by summing rows, the values of $N_{\cdot j}$ by summing columns. Also, the total number N of transitions is the sum both of the last row and of the last column.

	A	C	G	T	$N_{i\cdot}$
A	5	2	4	9	20
C	4	7	3	5	19
G	2	6	5	2	15
T	6	4	5	5	20
$N_{\cdot j}$	17	19	17	21	$N=74$

The matrix of estimated transition probabilities is obtained easily by dividing each element in this data matrix by the sum of the elements in its row:

	A	C	G	T
A	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{9}{20}$
C	$\frac{4}{19}$	$\frac{7}{19}$	$\frac{3}{19}$	$\frac{5}{19}$
G	$\frac{2}{15}$	$\frac{6}{15}$	$\frac{5}{15}$	$\frac{2}{15}$
T	$\frac{6}{20}$	$\frac{4}{20}$	$\frac{5}{20}$	$\frac{5}{20}$

The number $M = 6$ is really too small to get a good estimate of the initial distribution, but for the sake of illustration, we write the estimates down anyway

$$\widehat{\rho}_A^{(1)} = \frac{1}{3}, \qquad \widehat{\rho}_C^{(1)} = \frac{1}{3}, \qquad \widehat{\rho}_G^{(1)} = \frac{1}{6}, \qquad \widehat{\rho}_T^{(1)} = \frac{1}{6}. \qquad \diamond$$

Example 5.5, continued: Dayhoff's $PAM1$ matrix. In example 5.5 above, we modeled the evolution of the amino acid at a site in a protein as a Markov chain. The transition matrix of this chain gives the probabilities p_{xy} that amino acid x is replaced by amino acid y in one PAM unit of evolutionary time, which was defined to be an interval in which one expects to see changes in about 1% of the amino acids in the protein. The transition matrix is also called the substitution matrix and is referred to as the $PAM1$ matrix. The use and estimation of PAM substitution matrices was pioneered by Margaret Dayhoff in the late 1970's. To create a $PAM1$ matrix they identified

71 groups of closely related proteins, each group being presumed descendants of a common ancestor, and they aligned the proteinsin each group amino acid by amino acid (allowing gaps). They reconstructed an estimated evolutionary history for each group from a common ancestor and then counted substitutions in each history, and used these substitution counts to form the estimates. The exact method they used involves some technicalities, which we outline here for completeness; but the point we want the reader to take home is that the $PAM1$ matrix is calculated from the frequency of transitions observed in the evolutionary histories inferred from observed groups of related proteins. Dayhoff did not directly apply (5.2), because she made an additional assumption about the structure of the transition probabilities. To understand this assumption we define a mutability factor for each amino acid, quantifying the probability that it does in fact suffer a substitution in one PAM unit of time:

$$m_x \stackrel{\triangle}{=} \mathbb{P}\left(X_{t+1} \neq x \;\middle|\; X_t = x\right).$$

We leave it to the reader to show that

$$p_{xy} = \mathbb{P}\left(X_{t+1} = y \;\middle|\; X_t = x\right) = m_x \mathbb{P}\left(X_{t+1} = y \;\middle|\; X_t = x, X_{t+1} \neq x\right).$$

The factor $\mathbb{P}\left(X_{t+1} = y \;\middle|\; X_t = x, X_{t+1} \neq x\right)$ is the probability of moving to y at time $t+1$, given that the process is at x at time t *and* that a substitution takes place ($\{X_{t+1} \neq x\}$). Dayhoff additional assumption is that these latter substitution probabilities are symmetric: that is, the probability of substitution of y for x given a substitution takes place equals the probability of a substitution of x for y given that a substitution takes place:

$$\mathbb{P}\left(X_{t+1} = y \;\middle|\; X_t = x, X_{t+1} \neq x\right) = \mathbb{P}\left(X_{t+1} = x \;\middle|\; X_t = y, X_{t+1} \neq y\right).$$

She then estimates m_x and $\mathbb{P}\left(X_{t+1} = y \;\middle|\; X_t = x, X_{t+1} \neq x\right)$ separately using frequency-type estimates and combines them to find $\hat{p}_{xy}$. Here is a small submatrix of the $PAM1$ matrix that may be found in Figure 82, Supplement 3, pages 345-352, of M.O. Dayhoff (1978), *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington. To simplify the appearance, the transition matrix is shown multiplied by $10,000$. The top and left sides are labeled by the letters for the amino acids whose substitution probabilities are shown—see Table 1.1 in Chapter 1 for the amino

acids these letters stand for.

	A	R	N	D
A	9867	2	9	10
R	1	9913	1	0
N	4	1	9822	36
D	6	0	42	9859

$\diamond$

5.4 Path Probabilities

One may think of a realization of a Markov chain as a path in time through its state space. Accordingly we shall refer to a sequence $\{x_1, x_2, \ldots, x_t\}$ of states at successive times as a path. The probability

$$\mathbb{P}\left((X_1, \ldots, X_t) = (x_1, \ldots, x_t)\right)$$

of a path is just a joint probability of $(X_1, \ldots, X_t)$. We shall show that because of the Markov property, any path probability can be computed in terms of transition probabilities and the probability mass function of X_1. Thus, the transition probabilities, plus the distribution of X_1 give a full probabilistic description of a Markov chain.

First, let us look at the conditional probability,

$$\mathbb{P}\left((X_2, \ldots, X_t) = (x_2, \ldots, x_t) \;\middle|\; X_1 = x_1\right).$$

Write the event $\{(X_2, \ldots, X_t) = (x_2, \ldots, x_t)\}$ in the form

$$\{(X_2, \ldots, X_{t-1}) = (x_2, \ldots, x_{t-1})\} \cap \{X_t = x_t\}.$$

Now we will use the identity (which the reader may easily check from the definition of conditional probability)

$$\mathbb{P}(B \cap C \mid A) = \mathbb{P}(B \mid A)\mathbb{P}(C \mid B \cap A),$$

with $A = \{X_1 = x_1\}$, $B = \{(X_2, \ldots, X_{t-1}) = (x_2, \ldots, x_{t-1})\}$ and $C = \{X_t = x_t\}$. Notice that

$$A \cap B = \{(X_1, \ldots, X_{t-1}) = (x_1, \ldots, x_{t-1})\}$$

is the history of the process up to time $t-1$. We obtain,

$$\begin{aligned}\mathbb{P}\left((X_2,\ldots,X_t)=(x_2,\ldots,x_t)\;\middle|\;X_1=x_1\right)=\\ \mathbb{P}\left((X_2,\ldots,X_{t-1})=(x_2,\ldots,x_{t-1})\;\middle|\;X_1=x_1\right)\\ \times\mathbb{P}\left(X_t=x_t\;\middle|\;(X_1,\ldots,X_{t-1})=(x_1,\ldots,x_{t-1})\right).\end{aligned}$$

But, by the Markov property, the second factor on the right is simply $p_{x_{t-1}x_t}$, and so

$$\begin{aligned}\mathbb{P}\left((X_2,\ldots,X_t)=(x_2,\ldots,x_t)\;\middle|\;X_1=x_1\right)=\\ \mathbb{P}\left((X_2,\ldots,X_{t-1})=(x_2,\ldots,x_{t-1})\;\middle|\;X_1=x_1\right)p_{x_{t-1}x_t}.\end{aligned}$$

Now we can apply the same argument to the first term on the right-hand side and obtain,

$$\begin{aligned}\mathbb{P}\left((X_2,\ldots,X_t)=(x_2,\ldots,x_t)\;\middle|\;X_1=x_1\right)=\\ \mathbb{P}\left((X_2,\ldots,X_{t-2})=(x_2,\ldots,x_{t-2})\;\middle|\;X_1=x_1\right)p_{x_{t-2}x_{t-1}}p_{x_{t-1}x_t}.\end{aligned}$$

Continuing in this manner, we end up with the nice formula expressed in the following theorem. The theorem also states, as a consequence, a formula for the unconditioned path probability.

Theorem 1 *For a Markov chain $\{X_t\}_{t\geq 1}$ and for any path $\{x_1,\ldots,x_t\}$,*

$$\mathbb{P}\left((X_2,\ldots,X_t)=(x_2,\ldots,x_t)\;\middle|\;X_1=x_1\right)=p_{x_1x_2}p_{x_2x_3}\cdots p_{x_{t-1}x_t}. \tag{5.4}$$

In words, the conditional probability of a path, conditioned on the first value, is the product of the transition probabilities between successive states of the path.

The probability of a path is computed by the formula

$$\mathbb{P}\left((X_1,X_2\ldots,X_t)=(x_1,x_2,\ldots,x_t)\right)=\mathbb{P}(X_1=x_1)p_{x_1x_2}p_{x_2x_3}\cdots p_{x_{t-1}x_t}. \tag{5.5}$$

The proof of the second formula (5.5) is an application of the identity $\mathbb{P}(A\cap B)=\mathbb{P}(B\;\big|\;A)\mathbb{P}(A)$ and formula (5.4). Thus,

$$\begin{aligned}\mathbb{P}\left((X_1,X_2\ldots,X_t)=(x_1,x_2,\ldots,x_t)\right)=\\ \mathbb{P}(X_1=x_1)\mathbb{P}\left((X_2,\ldots,X_t)=(x_2,\ldots,x_t)\;\middle|\;X_1=x_1\right)\end{aligned}$$

and using (5.4) for the second factor on the right gives the path probability formula (5.5).

The probability distribution

$$\rho_i^{(1)} \stackrel{\triangle}{=} \mathbb{P}(X=i), \qquad i \in \mathcal{E},$$

is called the *initial distribution* of the Markov chain. Theorem 1 shows that any path probability of a Markov chain is determined by its initial distribution and its transition matrix. Hence the transition matrix plus the initial distribution comprise a complete probabilistic description of the chain.

Example 5.7, continued. Consider again the two state chain with transition matrix

$$\begin{pmatrix} 1-\lambda & \lambda \\ \mu & 1-\mu \end{pmatrix}.$$

Suppose that we flip a fair coin in order to decide whether X_1 is 0 or 1. That is, $\mathbb{P}(X_1 = 0) = \rho(1)_0 = .5$ and $\mathbb{P}(X_1 = 1) = \rho_1^{1)} = .5$. Then the probability of the path 000110 is

$$\begin{aligned}\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 1, X_5 = 1, X_6 = 0) \\ = .5 a_{00} p_{00} p_{01} p_{11} p_{10} = .5(1-\lambda)(1-\lambda)\lambda(1-\mu)\mu\end{aligned}$$

5.5 Problems

Exercise 5.1. In a Markov chain model for DNA sequences, assume that the probability that each base appears in the first site is 0.25 (equal probabilities). Assume that the bases along the sequence then form a Markov chain with transition probability matrix,

	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Find the probability of the sequence $AACTTTGGATCCG$. Leave your answer as a product of numbers. What would be the probability of this sequence under the i.i.d. site model with equal probabilities?

Exercise 5.2. Consider a three state Markov chain with transition probability matrix:

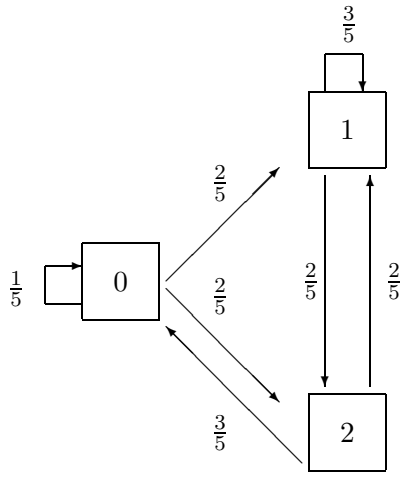
	0	1	2
0	1/4	1/4	1/2
1	1/2	1/4	1/4
2	1/4	1/2	1/4

a) Write down a state transition diagram for this chain.

b) Suppose $\mathbb{P}(X_1 = 1) = 1$. Find the probability that $\mathbb{P}(X_3 = 0) = 0$. (Consider all paths that lead to the result $X_3 = 0$.

Exercise 5.3. Write down the state transition matrix of the Moran model (see Chapter 3).

Exercise 5.4. Here is a state transition diagram for a Markov chain.



Assume that

$$\mathbb{P}(X_1 = 0) = 0.4 \qquad \mathbb{P}(X_1 = 1) = 0.5 \qquad \mathbb{P}(X_1 = 2) = 0.1$$

a) Write down the transition probability matrix for the chain.

b) Find $\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 2, X_4 = 2, X_5 = 0)$.

c) Find $\mathbb{P}(X_5=2 \mid X_4=0, X_3=2, X_2=2, X_1=2)$.

d) Suppose $\mathbb{P}(X_1 = 0) = 1$, that is, the chain starts in state 0. Let L be the first time it leaves 0. (L is a random variable and $X_1 = 0, X_2 = 0, \ldots, X_{L-1} = 0, X_L \neq 0$. What is $\mathbb{P}(L = k)$? (Find $\mathbb{P}(L = k)$ as a function of k.)

Exercise 5.5. Two white and two black tokens are distributed in two urns in such a way that each urn contains two tokens. The system is in state $s = 0, 1, 2$ at time i $(X_i = s)$ if the first urn contains s white tokens. At each step, one token is drawn from each urn, and the token from urn 1 is placed in urn 2 and conversely. Explain why $\{X_i; 1 = 0, 1, 2, \ldots\}$ is a Markov chain. Calculate its transition probability matrix.

Exercise 5.6. Let $X_i = 0$ if it rains on day i, and let $X_i = 1$ if it does not. Suppose that if it rains for two days in a row, the probability of rain on the third day is 0.1, but if it rained on only one of the first two days, the probability of rain on the third is 0.2, and if it is sunny for two days in a row, the probability of rain on the third day is 0.3. Is $X_1, X_2, \ldots$ a Markov chain? Explain mathematically.

Exercise 5.7. This is a model in the spirit of the Moran model discussed in class, but it does not apply to population genetics. Let X_n denote the number of black marbles at step n in an urn containing 100 black and white marbles. The number of black marbles X_{n+1} at the next step will be determined as follows Draw two marbles at random. If two black marbles are drawn, return a black and white marble to the urn (you have a collection of spare black and white marbles to use). If a white and a black are drawn, simply return them to the urn. If two white marbles are drawn, return a white and a black to the urn. X_{n+1} is the number of black marbles in the urn after this procedure; it is either the same, one more, or one less than X_n. If we repeat this procedure independently to obtain, successively $X_{n+1}, X_{n+2}, \ldots$ we will have a Markov chain. Determine the transition probabilities

$$\mathbb{P}(X_{n+1}=j \mid X_n=i) \qquad \text{for } j = i-1,\ j = i,\ j = i+1.$$

5.6 Multi-step transitions. Evolution of distributions.

Consider a Markov chain evolving in the state space $\mathcal{E}$ with transition probability matrix A. It is assumed that $\mathcal{E}$ is finite, although one can also write a

version of the results of this section for infinite state spaces. Let m be a positive integer. Then an m-step ahead transition probability is a conditional probability of the form:

$$p_{ij}^{(m)} \stackrel{\triangle}{=} \mathbb{P}\left(X_{t+m}=j \;\middle|\; X_t=i\right).$$

When $m = 1$, this is just the ordinary transition probability. One nice thing about Markov chains is a simple formula for computing m-step ahead transition probabilities.

Theorem 2 *$p_{ij}^{(m)} = A_{ij}^m$, where A_{ij}^m is the (i,j) element of the m^{th} power of the transition matrix A.*

Example 5.7, (continued). Consider again the two-state Markov chain discussed above. We calculate,

$$A^2 = \begin{pmatrix} (1-\lambda)^2+\lambda\mu & \lambda(2-\lambda-\mu) \\ \mu(2-\lambda-\mu) & (1-\mu)^2+\lambda\mu \end{pmatrix}$$

Thus, for example

$$\begin{aligned} \mathbb{P}(X_3=1 \mid\ X_1=0) &= \lambda(2-\lambda-\mu), \\ \mathbb{P}(X_3=1 \mid\ X_1=1) &= (1-\mu)^2+\lambda\mu. \end{aligned}$$ ◇

We shall demonstrate why Theorem 2 is true for the case $m = 2$. The proof for all m can be done by induction. Let the state space be denoted $\mathcal{E} = \{0, 1, \ldots, S\}$ We are conditioning on $X_t = i$. The event that $X_{t+2} = j$ can be viewed as the union of events

$$\{X_{t+1}=j\} = \bigcup_{k=0}^{S} \{X_{t+1}=k, X_{t+2}=j\},$$

because in the union, we have exhausted all possible values of X_{t+1}. Therefore,

$$\begin{aligned} p_{ij}^{(2)} &= \mathbb{P}\left(\bigcup_{k=0}^{S}\{X_{t+1}=k, X_{t+2}=j\} \;\middle|\; X_t=i\right) \\ &= \sum_{k=0}^{S} \mathbb{P}\left(X_{t+1}=k, X_{t+2}=j \;\middle|\; X_t=i\right) \\ &= \sum_{k=0}^{S} p_{ik}p_{kj}. \end{aligned} \tag{5.6}$$

In the last step we employed formula (5.4). But this last expression is just the (i, j) entry of A^2, which proves Theorem 2 for $m = 2$.

Now let us suppose we are given a Markov chain with transition probability matrix A, state space $\mathcal{E} = \{0, 1, \ldots, S\}$, and an initial distribution

$$\rho_i^{(1)} \stackrel{\triangle}{=} \mathbb{P}(X_1 = i), \qquad i \in \mathcal{E}.$$

For each $t \geq 2$, let us use $\rho^{(t)}$ to denote the distribution of X_t:

$$\rho_i^{(t)} \stackrel{\triangle}{=} \mathbb{P}(X_t = i), \qquad i \in \mathcal{E}.$$

A fundamental question of Markov chains is how $\rho^{(t)}$ evolves as time increases. In many applied problems, this is a main concern. For example, population genetics models try to explain the evolution of allele frequencies in natural populations. In Chapter 3, we saw that allele frequencies in finite population models are random. This means that we cannot predict the allele frequency evolution with certainty. But we can ask for the next best thing, the probability distribution of the frequency at each time and how it evolves.

We show next how to calculate $\rho^{(t)}$, $t = 1, 2, \ldots$, and in the next section we will discuss its possible limiting behaviors as $t \to \infty$. For our discussion, it is very helpful mathematically to interpret each $\rho^{(t)}$ as a row vector:

$$\rho^{(t)} \stackrel{\triangle}{=} \left(\rho_0^{(t)}, \rho_1^{(t)}, \ldots, \rho_S^{(t)}\right).$$

The reason for this is the following simple formula for the evolution of $\rho^{(t)}$:

$$\rho^{(t)} = \rho^{(t-1)} A, \tag{5.7}$$

where, as usual, A is the transition matrix. The proof of this is simple. By the rule of total probabilities:

$$\begin{aligned} \rho_i^{(t)} &= \mathbb{P}(X_t = i) \\ &= \sum_{k=0}^{S} \mathbb{P}(X_t = i \;\big|\; X_{t-1} = k)\mathbb{P}(X_{t-1} = k) \\ &= \sum_{k=0}^{S} p_{ki}\rho_k^{(t-1)} = \sum_{k=0} \rho_k^{(t-1)} p_{ki}. \end{aligned}$$

But this last expression is just the i^{th} element of $\rho^{(t-1)}A$.

Now, we can iterate equation (5.7) to solve for $\rho^{(t)}$:

$$\rho^{(t)} = \rho^{(t-1)}A = \rho^{(t-2)}A^2 = \cdots = \rho^{(1)}A^{t-1}.$$

The evolution of the distribution is governed by the evolution of the successive powers of the transition matrix. This formula is actually a generalization of the formula of Theorem 2 interpreting the elements of A^t as t-step ahead transition probabilities. Indeed suppose the initial distribution $\rho^{(1)}$ gives probability 1 for X_1 to be in state i; thus $\rho_i^{(1)} = 1$, and $\rho^{(1)}(k) = 0$, if $k \neq i$. Then, since we know $X_1 = i$, $\rho_j^{(t)} = \mathbb{P}(X_t = j) = \mathbb{P}(X_t = j \mid X_1 = i) = p_{ij}^{(t)}$, which is just the (i, j) element of A^t.

Example 5.9. Consider the two state Markov chain whose transition matrix is defined in Example 5.7 above. Assume that at the initial time the state is equally likely to be 0 or 1: $\rho^{(1)} = (0.5, 0.5)$. Then,

$$\rho^{(2)} = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} 1-\lambda & \lambda \\ \mu & 1-\mu \end{pmatrix} = ((1+\mu-\lambda)/2, (1+\lambda-\mu)/2)\,.$$

Similarily,

$$\begin{aligned} \rho^{(2)} &= \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} (1-\lambda)^2+\lambda\mu & \lambda(2-\lambda-\mu) \\ \mu(2-\lambda-\mu) & (1-\mu)^2+\lambda\mu \end{pmatrix} \\ &= \Big(((1-\lambda)^2+\mu(2-\mu))/2, ((1-\mu)^2+\lambda(2-\lambda))/2\Big) \qquad \diamond \end{aligned}$$

The calculations in the preceding example look as if they will produce more and more complicated looking expressions in λ and μ as t increases. This is true, but in fact the limiting behavior as $t \to \infty$ will be simple. Actually, the limiting behavior of A^t as $t \to \infty$ has a simple form for a large class of Markov chains, but we will discuss this in the next section.

Example 5.5, continued further; more on PAM matrices. In Example 5.5, we discussed modeling amino acid substitutions in protein evolution as a Markov chain, and, in the follow up in section 5.3, we discussed constructing a $PAM1$ matrix from data. Recall that $PAM1$ is the transition matrix corresponding to an evolutionary distance for which approximately 1% of a proteins amino acids change. In applications, biologists are often interested in protein families that have evolved over far longer times than one PAM unit. They would like to have a quanitative idea of the probability of substitutions for these longer times, because they use these long-time substitution

probabilities to compare newly discovered proteins to known protein families. A standard unit of time is 250 PAM. The $PAM250$ matrix is just the matrix of substitution probabilities for 250 PAM units of evolutionary time. According to Theorem 2,

$$PAM250 = (PAM1)^{250},$$

and in practice $PAM250$ is calculated in just this way from $PAM1$.