Chapter 4

Discrete-time Markov Chains and Applications to Population Genetics

A *stochastic process* is a quantity that varies randomly from point to point of an index set; for example, your total earnings after n plays of a game of chance, as a function of the index n. Chapter 4 is about a class of stochastic processes called *discrete-time Markov chains* and their application to population genetics. While Chapter 3 was driven by the scientific questions of population genetics, Chapter 4 is organized instead around its mathematical content. The reason for this change of perspective is the incredible importance of Markov chains in scientific and engineering, whenever stochastic models are needed. In particular, they are applied in many areas of quantitative biology beyond population genetics. This chapter is meant to give the reader enough basic theory of Markov chains to understand their use in current biological literature, whatever the area. Nevertheless, in this chapter population genetics will serve as the primary motivation and as the source of most examples.

To understand why stochastic processes are important in population genetics, think back to Chapter 3. All the models there impose the infinite population assumption: they set the frequency of a genotype, which is really random, equal to the probability a mating produces that genotype, which is not. When this is done, genotype frequencies become deterministic functions of time, for which we were able to derive difference equations.

However, when the population is small, fluctuations of genotype frequencies due to chance cannot be ignored. For example, consider one locus with two alleles, A and a, in a parent population for which $f_A = 0.9$. Then, as we learned from Chapter 3, the probability a random mating produces an AA offspring is $f_A^2 = 0.81$. If the next generation consists of only 10 individuals produced by 10 independent random matings, the event all are AA has the small, but not insignificant, probability

$(0.81)^{10} \approx 0.107$. So, allele a can disappear in the offspring population, not because it is less fit, but simply by chance, by what biologists call *random genetic drift*. Likewise, there is a chance that allele A could disappear. In fact, as we will see later, when the alleles experience no mutation and the population maintains a constant size, one of the alleles will eventually disappear. This is completely different from what happens in the infinite population model, for which f_A is constant from generation to generation. Thus, in small populations, randomness is so important a feature of the evolution of gene frequencies that they must be treated as stochastic processes.

The theory of Markov chains relies heavily on the use of conditional probabilities and conditional expectations. The reader may consult Sections 2.1.5 and 2.3.6 of Chapter 2 for the relevant background. This chapter will also make use of elementary linear algebra.

4.1 Markov Chains: Introduction

4.1.1 Discrete-time, stochastic process models.

A *discrete-time stochastic processes* is a sequence of random variables, $\{X(0), X(1), X(2), \ldots\}$, indexed by a discrete 'time' parameter t, $t = 0, 1, 2, \ldots$. Whether or not the index t represents time depends on the application, but the order of the sequence is important. Also, it is not essential that the first time be $t = 0$ or that time increase by unit steps. These are just conventions that we use when discussing stochastic processes in general. Throughout the chapter, we denote a generic term of a stochastic process by $X(t)$ and think of it as the random value of the process at time t. We denote the process itself by $\{X(t);\, t \geq 0\}$, or, simply, $\{X(t)\}$.

In practice, $X(t)$ will represent the 'state' of some system at time t. The set of all possible states is called the *state space* of the process, and we usually denote it by $\mathcal{E}$. For example, in population genetics models, $X(t)$ may be the number of individuals with a certain genotype at time t. If the population maintains a constant size N, the appropriate state space is $\mathcal{E} = \{0, 1, \ldots, N\}$. In later chapters, DNA is treated as a random sequence of bases $X(1), X(2), \ldots, X(T)$; then $X(t)$ represents the base at site t along the DNA segment—hence t denotes a position rather than a time—and the state space is the DNA alphabet $\{A, T, C, G\}$. When $\mathcal{E}$ is finite, as in these examples, or countably infinite, we say $\mathcal{E}$ is *discrete*. This will be the case for most models in this text.

Given a stochastic processes $X = \{X(t);\, t \geq 0\}$, the expression,

$$\mathbb{P}\left(X(0) = x_0, X(1) = x_1, \ldots, X(t) = x_t\right), \tag{4.1}$$

as a function of $(x_0, x_1, \ldots, x_t)$, is the joint probability mass function of $X(0), \ldots, X(t)$, and is called a *finite-dimensional distribution* of $\{X(t);\, t \geq 0\}$. It is often helpful to think of $\mathbb{P}\left(X(0) = x_0, X(1) = x_1, \ldots, X(t) = x_t\right)$ as the probability that the process

follows the *path*, $\{X(0) = x_0, X(1) = x_1, \ldots, X(t) = x_t\}$, up to time t, and so we sometimes call it a *path probability*.

The collection of all finite dimensional distributions of a process determines how to compute the probability of any path, and thus provides a complete statistical description of the process. A *model* of a stochastic process is therefore any set of assumptions that determines, at least in principle, all its finite dimensional distributions. A model is allowed to contain unspecified parameters, in which case it is sometimes called a *parametric model*.

To illustrate the terminology, let $\{X(0), X(1), \ldots\}$ represent a sequence of coin tosses, with $X(t) = 1$, if toss t results in heads, and $X(t) = 0$, if tails. The assumption, "$X(0), X(1), \cdots$ are independent tosses of a fair coin," constitutes a model, because it implies

$$\mathbb{P}\left(X(0)=x_0, X(1)=x_1, \ldots, X(t)=x_t\right) = (1/2)^{t+1},$$

for any t and any sequence $(x_0, \ldots, x_t)$ of 0's and 1's, and hence it completely determines all joint probabilities. The slightly more general assumption, "$X(0), X(1), \ldots$ are independent, identically distributed Bernoulli variables," is a parametric model in which the parameter is the unspecified probability, $p := \mathbb{P}(X(t)=1)$, because in this case,

$$\mathbb{P}\left(X(0)=x_0, X(1)=x_1, \ldots, X(t)=x_t\right) = \prod_{i=0}^{t} P(X(i)=x_i) = p^{\hat{n}}(1-p)^{t+1-\hat{n}},$$

where $\hat{n} = \sum_0^t x_i$ is the number of 1's in the sequence $(x_0, \ldots, x_t)$. Hence, the finite dimensional distributions are all determined up to the unspecified value of p.

How does one model a stochastic process observed in nature? This requires translating assumptions about the physical causes of its behavior into conditions determining its finite dimensional distributions. When the process evolves in time, it is often most natural to use conditional probabilities. For any time $t \geq 0$,

$$\begin{aligned} &\mathbb{P}\Big(X(0)=x_0, \ldots, X(t)=x_t, X(t+1) = x_{t+1}\Big) \qquad (4.2)\\ &\quad = \mathbb{P}\Big(X(t+1)=x_{t+1}\Big|X(0)=x_0, \ldots, X(t)=x_t\Big)\mathbb{P}\Big(X(0)=x_0, \ldots, X(t)=x_t\Big). \end{aligned}$$

This is a consequence of the simple identity, $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$, when $A = \{X(t+1) = x_{t+1}\}$ and $B = \{X(0) = x_0, \ldots, X(t) = x_t\}$. If t is thought of as the present time, the expression

$$\mathbb{P}\Big(X(t+1) = x_{t+1} \mid (X(0), \ldots, X(t)) = (x_0, \ldots, x_t)\Big), \quad t \geq 0.$$

in (4.2) is a conditional probability for the value of the process one step ahead into the future, given its entire past history; we call it a *one-step-ahead conditional probability*. As we see from equation (4.2), these conditional probabilities link the

joint distribution of $(X(0),\ldots,X(t),X(t{+}1))$ to that of $(X(0),\ldots,X(t))$, and thus they determine how joint distributions evolve as time goes forward. Hence, we can model a stochastic process by specifying its one-step-ahead conditional probabilities. In practice, they can often be deduced directly from assumptions on the physical nature of the process.

Example 4.1.1 Transitions between two states at random times.

Imagine moving between two stations, labeled 0 and 1, and let $X(t)$ denote your station at time t. At each time, t, you flip a coin that lands heads up with probability 0.25. If the flip comes up heads you move to the opposite station at time $t{+}1$; otherwise you stay put. All coin flips are independent. Then $X(t)$ is a stochastic process with state space $\{0,1\}$.

Its one-step-ahead conditional probabilities are easy to calculate. Consider any history, $\{X(0)=x_0, X(1)=x_1,\ldots,X(t{-}1)=x_{t-1}, X(t)=0\}$, that ends up in station 0 at time t. Then, you will move to station 1 at the next time if your coin flip comes up heads. Since this flip is independent of all previous events and heads has probability 0.25,

$$\begin{aligned}\mathbb{P}\Big(X(t{+}1){=}1\Big|X(t){=}0, X(t{-}1){=}x_{t-1},\ldots,X(0){=}x_0\Big)&\\ = \mathbb{P}\Big(X(t{+}1){=}1\Big|X(t){=}0\Big) = 0.25,&\\ \mathbb{P}\Big(X(t{+}1){=}0\Big|X(t){=}0, X(t{-}1){=}x_{t-1},\ldots,X(0){=}x_0\Big)&\\ = \mathbb{P}\Big(X(t{+}1){=}0\Big|X(t){=}0\Big) = 0.75.&\end{aligned}$$

Note that the past history, $\{X(t{-}1){=}x_{t-1},\ldots,X(0){=}x_0\}$, does not affect these conditional probabilities. Only the fact that you are at station 0 at time t is relevant. The one-step ahead probabilities when $X(t)=1$ are easily computed in the same manner, and again they do not depend on what the process did before time t. $\diamond$.

4.1.2 Markov Chains: Definition and basic properties

The one-step-ahead conditional probabilities in Example 4.1.1 depended on the past only through the given value $X(t)=x_t$ at the present time t. Markov chains are defined by requiring this property to hold generally.

Definition. A stochastic process $X=\{X(t);\, t\geq 0\}$ taking values in (a discrete) state space $\mathcal{E}$ is called a *Markov chain* if

$$P\Big(X(t{+}1)=x_{t+1}\mid X(0){=}x_0,\ldots,X(t){=}x_t\Big) = P\Big(X(t{+}1)=x_{t+1}\mid X(t){=}x_t\Big) \tag{4.3}$$

for every $t \geq 0$ and for all sequences of states, $x_0, \ldots, x_t$, for which both conditional probabilities are well-defined. (Both will be well-defined if and only if $\mathbb{P}\Big(X(0)=x_0, \ldots, X(t)=x_t\Big) > 0$.)

Markov chains are named in honor of the Russian mathematician, A. Markov(1856-1922), who did pioneering work in the definition and analysis of this class of processes.

We call (4.3) the *Markov chain condition*. The whole theory of Markov chains flows out of this innocent-looking assumption. In particular, it implies a vastly more general property called the *Markov property*, which says that at any time t, the past and future of the process are conditionally independent, given the present. The Markov property is explained in Section 4.1.4.

Later in the text, we will also define continuous-time Markov chains. Until then, the term 'Markov chain' always means a discrete-time Markov chain.

If X is a Markov chain, $P\Big(X(t+1) = j \mid X(t) = i\Big)$ is called the *probability of transition from state i to state j at time t*. If this probability does not depend on t, it is denoted by p_{ij}, and X is said to be *time-homogeneous*. The process of Example 4.1.1 is clearly an example of a time-homogeneous Markov chain. Its state space is $\mathcal{E} = \{0, 1\}$, and its transition probabilities are

$$p_{00} = 0.75, \quad p_{01} = 0.25, \quad p_{10} = 0.25, \quad p_{11} = 0.75,$$

because the chain moves from the state it's in to the other state if and only if the coin toss results in heads, which has probability 0.25.

The default assumption in this chapter is that all Markov chains are time-homogeneous, and the term Markov chain *should always be interpreted as* time-homogeneous Markov chain.

Example 4.1.2. The two-state Markov chain. The general, two-state Markov chain is not much more complicated than Example 4.1.1. The two states can be almost everything, but for convenience label them again as 0 and 1. A Markov chain moving between 0 and 1 is defined by four transition probabilities, p_{00}, p_{01}, p_{10}, and p_{11}. However, since 0 and 1 are the only states,

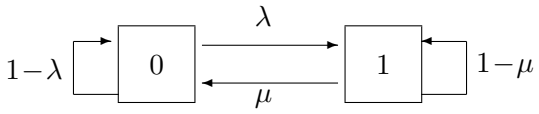
$$\begin{aligned} p_{00} + p_{01} &= \mathbb{P}\Big(X(t+1)=0|X(t)=0) + \mathbb{P}\Big(X(t+1)=1|X(t)=0) \\ &= \mathbb{P}\Big(X(t+1) \in \{0,1\}|X(t)=0) = 1. \end{aligned}$$

Likewise, $p_{10} + p_{11} = 1$. Thus, the two transition probabilities, p_{01} and p_{10}, completely determine the two-state Markov chain. It is standard to denote p_{01} by λ and p_{10} by μ, so that

$$p_{00} = 1 - \lambda, \quad p_{01} = \lambda, \quad p_{11} = \mu \quad \text{and} \quad p_{10} = 1 - \mu.$$

This is only a mathematical description of a two-state chain. Actual examples can be generated by a number of mechanisms. For instance, a variant of the coin flipping procedure of Example 4.1.1 will always work. Imagine again that states 0 and 1 represent two stations, and that we decide whether to move from one station to the other by flipping a coin and moving if the result is heads. But now suppose there is a different coin at each station, and the coin at station 0 comes up heads with probability λ, that at station 1 with probability μ. Assume all coin flips are independent. Then the process of moving between stations is a Markov chain, because for any t, the state at time $t+1$, conditional on its past history, depends only on which station it's at time t and on a coin flip, using the coin of that station, independent of all previous events. Clearly, $p_{01} = \lambda$, since we move from 0 to 1 only if the coin at station 0 comes up heads, and likewise $p_{10} = \mu$.

The following figure presents a nice visual representation of the two-state chain that immediately reveals its structure:



This is called a *state transition diagram.* The labeled boxes represent the states of the Markov chain, the arrows indicate the transitions that occur with positive probability, and each arrow is labeled by the probability of its associated transition. ◇

Given a Markov chain, it is convenient to organize its transition probabilities in a matrix, called the *state transition matrix.* If the state space is finite, and an order $\{s_1, \ldots, s_N\}$ for $\mathcal{E}$ is chosen, the state transition matrix (with respect to this order) is

$$A = \begin{pmatrix} p_{s_1 s_1} & p_{s_1 s_2} & \cdots & p_{s_1 s_N} \\ p_{s_1 s_1} & p_{s_2 s_2} & \cdots & p_{s_2 s_N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{s_N s_1} & p_{s_N s_2} & \cdots & p_{s_N s_N} \end{pmatrix}. \tag{4.4}$$

Often, we abbreviate this matrix by writing $A = [p_{ij}]_{i,j \in \mathcal{E}}$. (It may seem strange to use different letters to denote a matrix and its elements, but P has other uses in probability(!), so we avoid employing it also for state transition matrices.) The two-state Markov chain of Example 4.1.2 provides a simple example; its state transition matrix is

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

This can be read off directly from the state transition diagram.

The row indexed by s_i in (4.4) is the vector of transition probabilities starting from state s_i. Since the process must end up in some state, the sum of these probabilities over all states equals 1. That is,

$$\sum_{j=0}^{N} p_{s_i s_j} = \sum_{j=0}^{N} \mathbb{P}(X(t+1)=s_j | X(t)=s_i) = 1$$

for every row of A. Any square matrix of non-negative entries each of whose rows sums to 1 is called a *stochastic matrix*. Thus, state transition matrices are stochastic matrices. Conversely, any stochastic matrix is a valid model for the transition probabilities of a Markov chain.

The next theorem is the main result of this section. It tells us how to compute any finite dimensional distribution of a Markov chain.

Theorem 1 *Let $\{X_t;\, t \geq 0\}$ be a time-homogeneous Markov chain with transition probabilities $\{p_{ij};\, i,j \in \mathcal{E}\}$. For any $t \geq 0$ and any path $\{x_0, \ldots, x_t\}$,*

$$\mathbb{P}\Big(X(0)=x_0, X(1)=x_1, \ldots, X(t)=x_t\Big) = \mathbb{P}(X(0)=x_0) p_{x_0x_1} p_{x_1x_2} \cdots p_{x_{t-1}x_t} \quad (4.5)$$

or, equivalently,

$$\mathbb{P}\Big(X(1)=x_1, \ldots, X(t)=x_t \Big| X(0)=x_0\Big) = p_{x_0x_1} p_{x_1x_2} \cdots p_{x_{t-1}x_t} \quad (4.6)$$

Conversely, if either of these is always true, $\{X(t);\, t \geq 0\}$ is a Markov chain with transition probabilities $\{p_{ij};\, i,j \in \mathcal{E}\}$.

Formula (4.5) is easy to remember. Think of a path $(x_0, x_1, \ldots, x_t)$ as an initial state x_0 followed by a series of transitions $x_0 \to x_1 \to x_2 \to \cdots \to x_t$; the probability a Markov chain follows this path is the probability that it starts at x_0, times the product of the transition probabilities between successive states along the path.

The probability mass function,

$$\mathbb{P}(X(0) = x), \qquad x \in \mathcal{E},$$

is called the *initial (probability) law* of $\{X(t);\, t \geq 0\}$, because it is the law of $X(0)$ at the initial time $t = 0$. Theorem 1 implies that an initial law and a state transition matrix completely determine a model of a Markov chain, because, by equation (4.5), they specify all finite dimensional distributions. In practice the initial law is often incidental, and then we regard the state transition matrix alone as a model, one which allows arbitrary initial laws.

Proof of Theorem 1. The proof of (4.5) is by iterated conditioning. For convenience, denote $\mathbb{P}\Big(X(0)=x_0, X(1)=x_1, \ldots, X(t)=x_t\Big)$ by $P_t(x_0, x_1, \ldots, x_t)$. By conditioning

on the event $\{X(0)=x_0,\ldots,X(t-1)=x_{t-1}\}$ using (4.2), and then applying the Markov condition,

$$\begin{aligned} P_t(x_0,\ldots,x_{t-1},x_t) &= \mathbb{P}\Big(X(0)=x_0, X(1)=x_1,\ldots,X(t-1)=x_{t-1}\Big) \\ &\qquad \times \mathbb{P}\Big(X(t)=x_t \;\Big|\; X(0)=x_0,\ldots,X(t-1)=x_{t-1}\Big) \\ &= P_{t-1}(x_0,\ldots,,x_{t-1})p_{x_{t-1}x_t} \end{aligned} \tag{4.7}$$

The parameter $t \geq 1$ in this argument is arbitrary, and if we apply it with t replaced by $t-1$, $P_{t-1}(x_0,\ldots,x_{t-1},x_{t-1}) = P_{t-2}(x_0,\ldots,x_{t-2})p_{x_{t-2}x_{t-1}}$. Substituting back into (4.7),

$$P_t(x_0,\ldots,x_t) = P_{t-2}(x_0,\ldots,x_{t-2})p_{x_{t-2}x_{t-1}}p_{x_{t-1}x_t}.$$

By continuing this procedure backwards in time, we arrive finally at (4.5).

Equation (4.6) is just the same as (4.5) if both sides of (4.5) are divided by $\mathbb{P}(X(0)=x_0)$. Conversely if (4.5) is true,

$$\begin{aligned} &\mathbb{P}\Big(X(t+1)=x_{t+1}\Big|X(0)=x_0,\ldots,X(t)=x_t\Big) \\ &\qquad = \frac{\mathbb{P}\Big(X(0)=x_0,\ldots,X(t)=x_t,X(t+1)=x_{t+1}\Big)}{\mathbb{P}\Big(X(0)=x_0,\ldots,X(t)=x_t\Big)} \\ &\qquad = \frac{\mathbb{P}(X(0)=x(0))p_{x_0x_1}\cdots p_{x_{t-1}x_t}p_{x_tx_{t+1}}}{\mathbb{P}(X(0)=x(0))p_{x_0x_1}\cdots p_{x_{t-1}x_t}} = p_{x_tx_{t+1}}, \end{aligned}$$

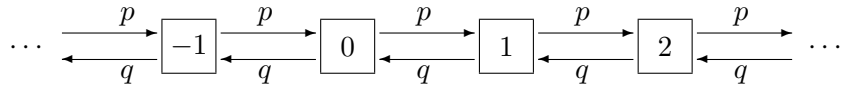
which proves the Markov condition. ◇

Example 4.1.3. Consider the two state Markov chain of Example 4.1.2 and suppose at time $t=0$, it is equally likely to be in either of the two states, i.e., $\rho_0(0) = \rho_1(0) = 1/2$. Then

$$\begin{aligned} \mathbb{P}\Big((X(0),\ldots,X(5))=(1,1,0,0,0,1)\Big) &= \rho_1(0)p_{11}p_{10}p_{00}p_{00}p_{01} \\ &= (1/2)(1-\mu)\mu(1-\lambda)^2\lambda. \end{aligned}$$ ◇

4.1.3 More Examples.

Example 4.1.4. Simple random walk. Simple random walk is the integer-valued Markov chain defined by the state transition diagram:



Equivalently, if i and j are any integers,

$$p_{ij} = \begin{cases} p, & \text{if } j = i+1; \\ q, & \text{if } j = i-1; \\ 0, & \text{if } |j-i| > 1. \end{cases} \tag{4.8}$$

Here, $0 < p < 1$ and $p + q = 1$. Thus, the simple random walk moves from integer to integer by unit steps only, at each time taking one step to the right with probability p and one step to the left with probability $q = 1 - p$, independently of its present position and past history. When $p = q = 1/2$, simple random walk is called *symmetric, simple random walk.*

There is an easy way to construct a simple random walk. Let $X(0)$ denote a random, integer-valued, initial position. Assume that $\xi_1, \xi_2, \ldots$ are independent random variables, independent of $X(0)$, and all having the distribution,

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

Then $X(t) = X(0) + \sum_{i=1} \xi_i$, $t \geq 0$, defines a simple random walk with transition probabilities as in (4.8). This is easy to see. For each t, $X(t) - X(t-1) = \xi_t$, and thus the ξ_t represents the t^{th} step the process takes. Suppose $X(t) = i$ is given. Since the $(t+1)^{\text{st}}$ step, ξ_{t+1}, is independent of all other steps and of $X(0)$, it clear that $X(t+1) = X(t) + \xi_{t+1} = i + \xi_{t+1}$ does not depend on the history of the process before time t, and so $\{X(t)\}$ is a Markov chain. It has the correct transition probabilities because $p_{i,i+1} = \mathbb{P}(X(t+1) = i+1 | X(t) = i) = \mathbb{P}(\xi_{t+1} = 1) = p$ and $p_{i,i-1} = \mathbb{P}(X(t+1) = i-1 | X(t) = i) = \mathbb{P}(\xi_{t+1} = -1) = q$.

Simple random walk is sometimes described picturesquely as the 'drunkard's walk.' A drunk walks out of a bar—no, this is not the start of a joke!—and staggers up and down the street, taking unit-sized steps. The poor fellow is so inebriated that each new step is independent of all his previous efforts. If all steps have the same probability of moving him up or down the street, he performs a simple random walk.

There is also a gambling interpretation. Imagine successive, independent plays, on each of which you win a dollar with probability p and lose a dollar with probability q. If ξ_t represents what you win on play t, then $X(t) = X(0) + \sum_1^t \xi_i$ is your total fortune after t plays, when $X(0)$ is the amount of money you start with.

The adjective 'simple' in 'simple random walk' means that the process moves by integer steps only. The term, 'random walk', is also used for any process of the form

$$Y(t) = Y(0) + \sum_{1}^{t} \eta_i,$$

where $Y(0)$ is a random variable, and $\eta_1, \eta_2, \ldots$ are independent, identically distributed random variables, which are also independent of $Y(0)$. See Exercise 4.10 for an example. $\diamond$

Example 4.1.5. The Wright-Fisher model for genotype evolution in a finite population: selectively neutral, one locus case.

Study this example closely! It is the basic stochastic model of population genetics. It is named after R.A. Fisher (1890-1962), a British statistician, and Sewall Wright (1889-1988), an American geneticist, who made pioneering contributions to the discipline.

Let A be an allele, and let $X(t)$ denote the number of A's in the allele pool of generation t, in a finite population. The selectively neutral, Wright-Fisher model is a model for how $X(t)$ evolves randomly. It assumes: (i) the species is monecious and diploid, and the size of the population remains constant; (ii) generations are non-overlapping; (iii) each generation is produced from the previous one by independent random matings; (iv) there is no selection, mutation, or migration. Compare these hypotheses to the infinite population model defined in Section 3.3.1. The only differences are that the population is finite in size and that random matings are explicitly assumed to be independent. It is not important to the model how many other alleles there are at the locus of A.

We will show how assumptions (i)—(iv) translate directly into a Markov chain model for $\{X(t)\}$. Of course, $X(t)$ is always a non-negative integer, and if N denotes the (constant) size of the population, it can be no larger than $2N$. Thus, the state space for the process $\mathcal{E} = \{0, 1, \ldots, 2N\}$.

Now, let t be any time, and fix a past history, $\{X(0) = i_0, \ldots, X(t-1) = i_{t-1}, X(t) = i\}$, ending up with i alleles of type A in generation t. We want to compute the conditional distribution of $X(t+1)$. By assumptions (i) and (iii), generation $t+1$ is created by N, independent, random matings of generation t parents. From Chapter 3, the genotype produced by a random mating is the result of two independent, random samples of the allele pool of generation t. Thus, the allele pool of generation $t+1$ is created by $2N$ independent random samples. Now, the probability of randomly selecting A from the allele pool of generation t is its frequency, and since $X(t) = i$ is given, this frequency is $\frac{i}{2N}$. Recall that the number of successes in n independent trials, when p is the probability of success, has the binomial distribution with parameters n and p. Counting the selection of A as a success, it follows that, given $X(0) = i_0, \ldots, X(t-1) = i_{t-1}$, and $X(t) = i$,

$$X(t+1) \text{ is binomial with parameters } n = 2N \text{ and } p = i/2N.$$

This has two consequences. First, this distribution depends only on $X(t) = i$, not on the values of $X(s)$ for times s strictly prior to t, and therefore $\{X(t)\}$ does indeed satisfy the Markov condition. Second, it tells that that the transitions probabilities are given by the binomial distribution:

$$p_{ij} = \mathbb{P}\Big(X(t+1) = j \Big| X(t) = i\Big) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}, \quad 0 \le i, j \le 2N. \tag{4.9}$$

Two states deserve special attention, states 0 and $2N$. State 0 means there are no alleles A, and if this is the case, they will never appear in future generations, since there is no mutation in this model. Thus $p_{00} = 1$ and $p_{0j} = 0$ for all $1 \le j \le 2N$. Likewise, $p_{2N,2N} = 1$, because once the chain enters state $2N$, no a alleles are left and it must stay in state $2N$ ever after.

In general, a state i of a Markov chain is said to be *absorbing* if $p_{ii} = 1$, because once the chain hits state i, it remains there forever. Thus, states 0 and $2N$ of the Wright-Fisher chain are absorbing.

When $0 < i < 2N$, the formula of (4.9) shows that $p_{ij} > 0$ for any other state j, and so the chain can move from i to any other state j in one step. Thus states 0 and $2N$ are the only absorbing states. It is of little use to write down a state transition diagram for a Wright-Fisher chain, even for fairly small values of N. The diagram will be crowded with arrows going from each non-absorbing state to all others and will not help us understand the chain better.

Let

$$Y(t) = \frac{X(t)}{2N}.$$

$Y(t)$ is the frequency of allele t in generation t and it, rather than $X(t)$, is the actual object of interest. $\{Y(t)\}$ is also called the Wright-Fisher process and is really the one we should work with. We have defined the equivalent process, $\{X(t)\}$, for notational convenience. It is much cleaner to write the transition probability p_{ij} for the transition of $X(t)$ from i to j, then the corresponding transition probability, $p_{i/2N,j/2N}$, for $Y(t)$! And as long as N is fixed, it makes no difference whether we analyze $\{X(t)\}$ or $\{Y(t)\}$. However, when we wish to compare the Wright-Fisher chain for different values of N, it is much better to use $\{Y(t)\}$, because it represents a frequency no matter what N is.

Example 4.1.6. The Moran model. This is named in honor of the Australian applied probabilist P.A.P. Moran (1917-1988). It is a finite population model with overlapping generations, but without selection, mutation or migration. However, it assumes that reproduction is asexual, by simple duplication. The model treats a population divided into two types, which shall be denoted by the letters A and a. One might think of these letters as denoting genotypes or alleles, but that is not necessary, and a may just represent 'not A.'

The population changes according to the following mechanism. At each time t, two individuals are selected from the current population by independent random sampling *with* replacement. The first individual gives birth to a copy of itself, which joins the population together with its parent. The second individual is removed from the population (it dies). The result of these two steps is the population at time $t+1$. The random samplings at different times are all independent.

Note the following features of this model. First, since one individual is added and one removed at each stage, the population remains constant in size. Second, the choice of who reproduces and who dies is made with replacement, and so the individual chosen to reproduce may be the same as the one chosen to die, with the end result that it is just replaced with a copy of itself. Finally, there is no selection in the model, because in every generation each individual has the same chance to reproduce or die, irrespective of type.

Let $X(t)$ denote the number of individuals of type A at stage t, and let N denote the total size of the population. We will argue that $X = \{X(t)\}_{t\geq 0}$ is a Markov chain and calculate its transition probabilities.

Suppose $X(t) = i$ and $0 < i < N$, and perform the random selection, copy and replacement experiment described above. The probability of selecting type A is then i/N and of selecting a is $(N-i)/N$ in each of the two samples. There are three possibilities. The first is that a type A is chosen to reproduce and a type a to die. This happens with probability $\dfrac{i}{N}\dfrac{(N-i)}{N}$ and increases the number of A's by one. Thus,

$$\mathbb{P}\Big(X(t+1)=i+1 \;\Big|\; X(t)=i\Big) = \frac{i(N-i)}{N^2}.$$

The second possibility is that the opposite occurs, a type a is chosen to reproduce and a type A to die, in which case $X(t+1) = i-1$. Again,

$$\mathbb{P}\Big(X(t+1)=i-1\Big|X(t)=i\Big) = \frac{i(N-i)}{N^2}.$$

Finally, if both individuals selected are the same type, the overall number of type A's remains unchanged. As this is the only remaining case,

$$\mathbb{P}\Big(X(t+1)=i\Big|X(t)=i\Big) = 1 - 2\frac{i(N-i)}{N^2}.$$

This can also be expressed as the sum of the probability of choosing A twice and the probability of choosing a twice:

$$\mathbb{P}\Big(X(t+1)=i\Big|X(t)=i\Big) = \left(\frac{i}{N}\right)^2 + \left(\frac{N-i}{N}\right)^2 = \frac{i^2+(N-i)^2}{N^2}.$$

Since these are the only possibilities, $p_{ij} = 0$ if $|j-i| > 1$.

If $i = 0$, there are no type A's, and since no mutation occurs, type A's cannot enter the population at the next, or at any future, time. Thus, $p_{00} = 1$ and 0 is

absorbing. Similarly, if $i = N$, type a's can never enter the population and N is absorbing.

This entire analysis is unchanged if values of $X(s)$ at times s previous to t are also given—only $X(t) = i$ is relevant. Thus X satisfies the Markov chain condition, and so it is a Markov chain.

As an example, we write down the state transition matrix for $N = 4$ when the states are in the order $0, 1, 2, 3, 4$; the reader may easily verify the values of the entries:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{3}{16} & \frac{10}{16} & \frac{3}{16} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{3}{16} & \frac{10}{16} & \frac{3}{16} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

◇

Example 4.1.7. Birth-and-death chains. These are a class of Markov chains generalizing random walk. The Moran model is a particular example.

Let $\mathcal{E}$ be a set of consecutive integers; $\mathcal{E}$ could be finite or infinite. A birth-and-death chain on $\mathcal{E}$ is a Markov chain which only makes transitions of at most unit size, but which allows the transition probabilities to depend on the present state. Thus, for each $i \in \mathcal{E}$, there are non-negative numbers, p_i, q_i, and r_i, with $p_i + q_i + r_i = 1$, such that

$$p_{ij} = \begin{cases} p_i, & \text{if } j = i+1 \text{ and } i+1 \in \mathcal{E}; \\ r_i, & \text{if } j = i; \\ q_i, & \text{if } j = i-1, \text{ and } i-1 \in \mathcal{E}; \\ 0, & \text{otherwise.} \end{cases}$$

Because of this structure, the state transition matrix of a birth-and-death chain has a characteristic form; its only non-zero entries are along, above, and below the diagonal. For example, the state transition matrix of the general birth-and-death chain on $\{0, 1, \ldots, N\}$ is

$$A = \begin{pmatrix} 1-p_0 & p_0 & 0 & 0 & \cdots & \cdots & 0 \\ q_1 & r_1 & p_1 & 0 & \cdots & \cdots & 0 \\ 0 & q_2 & r_2 & p_2 & \cdots & \cdots & 0 \\ \vdots & & & & & & \vdots \\ \vdots & & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & q_{N-1} & r_{N-1} & p_{N-1} \\ 0 & \cdots & \cdots & \cdots & 0 & q_N & 1-q_N \end{pmatrix}. \qquad \diamond$$

In all examples so far, the Markovian nature of each process was fairly obvious from its definition. Sometimes, the process of direct interest is not Markovian. When

this happens, it is often possible to embed it in a Markov process by expanding the state space, and this may be helpful in analyzing the model. The next example illustrates this point.

Example 4.1.8. Modeling DNA. Take a sample of DNA and let $X(1), X(2), \ldots, X(K)$ represent the successive bases along one of its strands, read in the $5'$ to $3'$ direction. This will be a string of letters from the DNA alphabet, $\{A, T, C, G\}$.

We are interested in stochastic models for $X(1), \ldots, X(K)$. This may seem odd. What is random about a string of DNA? After all, it just is what it is. However, many experiments involve *sampling* DNA from a population. Since DNA sequences vary from member to member, the sampled sequence is effectively random. In fact, the probability distribution of this random sample is precisely the object of interest in studies of the variation DNA sequences across a population.

Markov chain models of DNA sequences are often used, and we will discuss some later in the text. But how reasonable are they? The Markov condition would require that the conditional distribution of the base $X(t+1)$ at position $t+1$, given the bases $X(1), \ldots, X(t)$, depends only on the base $X(t)$ at position t. This seems unlikely *a priori*, especially for DNA in coding regions. DNA codons are three base pairs long and the sequence in which they occur codes for proteins. Thus, for example, one might expect that the conditional distribution of $X(t+1)$ given say, $X(t-1) = A, X(t) = A$ could be different than the conditional distribution given $X(t-1) = C, X(t) = A$. A better model for DNA might be to assume that $\{X(t)\}$ is a k-Markov chain, for some $k > 1$. This means that, for any t, k previous bases always suffice to characterize a one-step-ahead conditional probability: that is,

$$\begin{aligned} &P\Big(X(t+1) = x_{t+1} \mid X(0) = x_0, \ldots, X(t) = x_t\Big) \\ &\quad = \quad P\Big(X(t+1) = x_{t+1} \mid X(t) = x_t, \ldots, X(t-k+1) = x_{t-k+1}\Big). \end{aligned} \tag{4.10}$$

(For this to make sense, $t \geq k - 1$.)

k-Markov chains can be analyzed on their own terms, but it is just as easy to embed them in a Markov chain. We show how when $k = 2$; the general case is an easy extension of the same method. The idea is to let the state be two adjacent bases, and to define a new process

$$Z(t) := (X(t-1), X(t)), \quad \text{for times } t \geq 1.$$

We will argue that if $\{X(t)\}$ is 2-Markov, then $\{Z(t)\}$ is a Markov chain. For suppose we know the values of $Z(1), \ldots, Z(t-1), Z(t)$, and we want to calculate a conditional probability concerning $Z(t+1) = (X(t), X(t+1))$. This is equivalent to calculating a conditional probability concerning $X(t)$ and $X(t+1)$ given values of $X(0), X(1), \ldots, X(t-1), X(t)$. But as $X(t)$ is already given, this is the same as computing a conditional probability of $X(t+1)$ given the history of X up to time t. The 2-Markov property of $\{X(t)\}$ implies this conditional probability depends

only on the values of $X(t-1)$ and $X(t)$, and hence only on the value of $Z(t) = (X(t-1), X(t))$, irrespective of the values of $Z(s)$ for $s \leq t-1$. This is the Markov chain condition for $\{Z(t); t \geq 1\}$, as we wanted to show. (See Exercise 4.1.9 for the rudiments of a direct analysis of 2-Markov chains.)

Of course, whether a k-Markov chain for $k > 1$ is a better model for DNA than a Markov chain is a question to be resolved by fitting alternative models to data and seeing which works best. The purpose of this example was to emphasize the meaning and scope of the Markov condition, (4.3), and how its use might guide the correct choice of a state space. ◇

4.1.4 The Markov property.

According to the Markov condition, (4.3), each new state of a Markov chain is determined randomly by a mechanism depending only on the current state and otherwise *independent* of the previous history of the chain. We can imagine there is a separate machine for each state. When the chain gets to state i, we push the button on its machine, which returns a random state j to move to next, where the probability of j is p_{ij} for all states j. (For instance, the 'machine' we used for two-state chains was coin flipping.) The intuitive consequences of this picture go beyond one-step-ahead conditional probabilities. For example, it ought to be true that

$$P\Big(X(t{+}1){=}x_{t+1} \mid X(t{-}1){=}x_{t-1}, X(t){=}x_t\Big) = P\Big(X(t{+}1){=}x_{t+1} \mid X(t){=}x_t\Big) \quad (4.11)$$

always holds. The Markov condition, (4.3), does not say this directly, because it is stated for the full past history, and in this instance, the conditioning event on the left-hand side includes only partial information about the process prior to time t. But, as we shall see, (4.11) is in fact true and is a consequence of the Markov condition.

Similarly, Markov-type conditions should extend more than one-step into the future. For example, we would expect,

$$P\Big(X(t{+}2) = x_{t+2} \mid X(0){=}x_0, \ldots, X(t){=}x_t\Big) = P\Big(X(t{+}2) = x_{t+2} \mid X(t){=}x_t\Big),$$

and, likewise,

$$P\Big(X(t{+}2) = x_{t+2} \mid X(t{-}1) = x_{t-1}, X(t){=}x_t\Big) = P\Big(X(t{+}2) = x_{t+2} \mid X(t){=}x_t\Big), \quad (4.12)$$

because, whatever the history before t, where the chain moves two steps ahead depends on first pushing the button for the machine of state x_t, which determines the state at time $t+1$, and then pushing the button for the machine at the new state $X(t{+}1)$. None of this will depend on additional information from the past. Again, the Markov chain condition does not state these new identities directly, but, as we will see shortly, it does imply them.

The next theorem makes a very general statement, that incorporates both conditioning on partial past history and looking more than one step into the future.

Theorem 2 *Let $\{X(t)\}$ be a Markov chain. Let t be any time and i be any state, let U be* any *event defined in terms of $\{X(t), X(t+1), X(t+2), \ldots\}$, and let V be* any *event defined in terms of $\{X(0), \ldots, X(t-1)\}$. Then,*

$$P(U|\{X(t)=i\} \cap V) = P(U|X(t)=i)) \tag{4.13}$$

In this theorem statement, if we regard t as the present time, U is any event concerning the future of the process from time t onward, V is any event concerning past behavior prior to t, and (4.13) says that U is conditionally independent of V, given the present value $X(t) = i$. The fact that (4.13) holds is called the *Markov property* of $\{X(t)\}$.

The statement of the Markov property is somewhat abstract, so let us consider some examples to understand its scope. First, by taking $U = \{X(t+1) = x_{t+1}$ and $V = \{X(0) = x_0, \ldots, X(t-1) = x_{t-1}\}$, we recover the Markov condition, (4.3), defining a Markov chain:

$$P\Big(X(t+1) = x_{t+1} \mid X(0)=x_0, \ldots, X(t)=x_t\Big) = P\Big(X(t+1) = x_{t+1} \mid X(t)=x_t\Big)$$

Next, by letting $U = \{X(t+2) = x_{t+2}\}$ and $V = \{X(t-1) = x_{t-1}\}$, we recover the identity stated in (4.12).

Finally, suppose, $\{X(t)\}$ is the Moran model for a population of size N, let U be the event that the process hits the absorbing state 0 at some time t or later, and let $V = \{X(0) = x_0, \ldots, X(t-1) = x_{t-1}\}$. Then, (4.12) says that the conditional probability of absorption into state 0, when $X(t) = i$, is independent of the previous history of the process.

We have seen that the Markov condition (4.3) is a special case of the Markov property. But a process is a Markov chain if and only if it satisfies the Markov condition, and hence Theorem 2 is really saying that the Markov condition, which is a statement about one-step-ahead conditional probabilities only, implies the much more general Markov property. We shall not prove Theorem 2. The proof is not complicated, but it's messy to write out and does not add to the intuitive understanding of what the Markov property says and why it is true.

The Markov property is used over and over in the analysis of Markov chains, which is one reason we have tried to state it carefully and generally. We give two, basic applications to illustrate its use. The first generalizes Theorem 1 to any starting time. We will show that if $\{X(t);\, t \geq 0\}$ is a time-homogeneous, Markov chain, then for any nonnegative integers s and $t \geq 0$,

$$\mathbb{P}\Big(X(s+1)=x_1, \ldots, X(s+t)=x_t \Big| X(s)=x_0\Big) = p_{x_0x_1}p_{x_1x_2}\cdots p_{x_{t-1}x_t} \tag{4.14}$$

This is exactly the same formula derived in (4.6) of Theorem 2. Thus, referring back to that formula,

$$\mathbb{P}\Big(X(s+1)=x_1, \ldots, X(s+t)=x_t \Big| X(s)=x_0\Big) = \mathbb{P}\Big(X(1)=x_1, \ldots, X(t)=x_t \Big| X(0)=x_0\Big). \tag{4.15}$$

This is a natural expression of the time-homogeneity of the process. It says that the process starting from i at time s behaves the same as the process starting from i at time 0.

To derive (4.14), we shall use the general identity:

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

Taking $A = \{X(s+t) = x_t\}$, $B = \{X(s+1) = x_1, \ldots, X(s+t-1) = x_{t-1}\}$, and $C = \{X(s) = x_0\}$, we get

$$\begin{aligned}
&\mathbb{P}\Big(X(s+1)=x_1,\ldots,X(s+t)=x_t\Big|X(s)=x_0\Big)\\
&\quad=\mathbb{P}\Big(X(s+t)=x_t\Big|X(s)=x_0,X(s+1)=x_1,\ldots,X(s+t-1)=x_{t-1}\Big)\\
&\qquad\times\mathbb{P}\Big(X(s+1)=x_1,\ldots,X(s+t-1)=x_{t-1}\Big|X(s)=x_0\Big).
\end{aligned}$$

But by the Markov property, the first term on the right-hand side is simply the transition probability $p_{x_{t-1}x_t}$, and thus, reversing the order of the product,

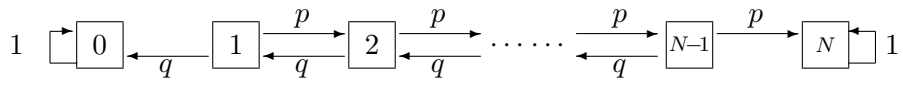
$$\begin{aligned}
&\mathbb{P}\Big(X(s+1)=x_1,\ldots,X(s+t)=x_t\Big|X(s)=x_0\Big)\\
&\qquad=\mathbb{P}\Big(X(s+1)=x_1,\ldots,X(s+t-1)=x_{t-1}\Big|X(s)=x_0\Big)\cdot p_{x_{t-1}x_t}.
\end{aligned}$$

Now apply the same method again to the first term on the right-hand side: then,

$$\begin{aligned}
&\mathbb{P}\Big(X(s+1)=x_1,\ldots,X(s+t)=x_t\Big|X(s)=x_0\Big)\\
&\qquad=\mathbb{P}\Big(X(s+1)=x_1,\ldots,X(s+t-2)=x_{t-2}\Big|X(s)=x_0\Big)\cdot p_{x_{t-2}x_{t-1}}p_{x_{t-1}x_t}.
\end{aligned}$$

Continuing in this fashion will eventually bring you to equation (4.14).

The second application illustrates a technique used again and again in the analysis of Markov chains: using conditioning and the Markov property to derive difference equations for probabilities of events. To keep the discussion concrete, consider the symmetric random walk on the set of integers $\{0, 1, 2, \ldots, N\}$, with absorption at the 'boundary' states 0 and N. Thus, starting, at some integer i, $0 < i < N$, it behaves as a symmetric random walk until it hits either 0 or N, when it stops moving. The state transition diagram is:



Let U be the event that the chain eventually gets absorbed into the state 0. We are interested in computing the probability that U occurs starting from $X(0) = i$,

$0 \leq i \leq N$. This is called the gambler's ruin problem, the idea being that $X(t)$ is the fortune after t plays of a game of chance in which the gambler wins a dollar with probability p and loses with probability q. N is the total amount of money at stake between him and the casino and i is the amount he starts with. He will play either until he wins all N and breaks the bank, or he goes broke (ruin). What is the probability of ruin?

To answer this question, we will let $g_i = \mathbb{P}(U|X(0)=i)$ and derive a difference equation for g_i, $0 \leq i \leq N$. First note that, by definition,

$$g_0 = 1 \quad \text{and} \quad g_N = 0 \tag{4.16}$$

Now suppose $X(0) = i$, where $0 < i < N$. The central idea is to condition on where the chain moves next. There only two possibilities: it moves to $i+1$, which happens with probability p, or to $i-1$, which happens with probability q. By the rule of total probabilities,

$$g_i = \mathbb{P}\Big(U|X(0)=i, X(1)=i+1\Big) \cdot p + \mathbb{P}\Big(U|X(0)=i, X(1)=i-1\Big) \cdot q. \tag{4.17}$$

Now U occurs if and only if the sequence $\{X(1), X(2), \ldots\}$ hits 0 before N. Thus, by the Markov property,

$$\begin{aligned}
\mathbb{P}\Big(U|X(0)=i, X(1)=i+1\Big) &= \mathbb{P}\Big(\{X(1), X(2), \ldots\} \text{ hits 0 before } N \Big| X(0)=i, X(1)=i+1\Big) \\
&= \mathbb{P}\Big(\{X(1), X(2), \ldots\} \text{ hits 0 before } N \Big| X(1)=i+1\Big)
\end{aligned}$$

This last expression is the probability of ruin starting in state $i+1$ at time $s=1$. But, as just observed, a time-homogeneous chain starting in a state x at time s behaves just like the chain starting at x at time 0. Thus, the probability of ruin starting in state $i+1$ at time $s=1$ is the same is the same as the probability of ruin starting in state $i+1$ at time $s=0$, and this is, by definition, g_{i+1}. By similar reasoning,

$$\mathbb{P}\Big(U|X(0)=i, X(1)=i-1\Big) = g_{i-1}.$$

Using these results in equation (4.17), we arrive at the difference equation,

$$g_i = pg_{i+1} + qg_{i-1}, \quad 0 < i < n. \tag{4.18}$$

To find g_i for all i, we need to solve this equation with the 'boundary' conditions in (4.16). The solution, which you are asked to verify in Exercise 4.1.11, is:

$$g_i = \frac{1-(q/p)^i}{1-(q/p)^N}, \text{ if } p \neq q, \qquad g_i = \frac{i}{N}, \text{ if } p = q = \tfrac{1}{2}. \tag{4.19}$$

Later, we will solve similar problems for the Moran and Wright-Fisher chains using a different method.

4.1.5 Exercises

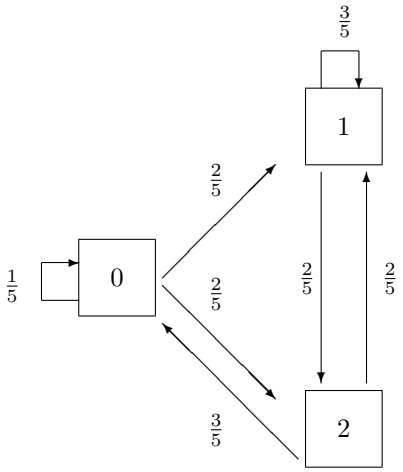
4.1.1. Consider a three state Markov chain with transition probability matrix:

	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 $\mathbb{P}(X(0) = 1) = 1$. Find the probability that $\mathbb{P}(X(2) = 0)$ by adding the probabilities of all paths that lead from state 1 to state 0 in two steps.

4.1.2. Consider the following state transition diagram for a Markov chain.



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

b) Assume that $\mathbb{P}(X(0)=0) = 0.4$, $\mathbb{P}(X(0)=1) = 0.5$, and $\mathbb{P}(X(0)=2) = 0.1$. Find $\mathbb{P}\left(X(0)=1, X(1)=1, X(2)=2, X(3)=0, X(4)=0\right)$.

c) What is $\mathbb{P}(X(5)=2 \mid X(4)=1, X(3)=0, X(2)=2)$?

d) What is $\mathbb{P}(X(6)=2 \mid X(4)=1, X(3)=0, X(2)=2)$? (You will need to use (4.14).)

e) Suppose $\mathbb{P}(X(0) = 0) = 1$, that is, the chain starts in state 0. Let L be the first time it leaves 0. (L is a random variable and $X(0) = 0, X(1) = 0, \ldots, X(L-1) = 0, X(L) \neq 0$.) For any positive integer k, what is $\mathbb{P}(L = k)$?

f) Generalize (d). If s is a state of a Markov chain, and L is the first time the chain leaves s, find $\mathbb{P}(L = k \mid X(0)=s)$ in terms of p_{ss}.

4.1.3. Let $X(t) = 0$ if it rains on day t, and let $X(t) = 1$ 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 rains on

only one of two successive 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.

a) Explain why $\{X(t);\, t \geq 0\}$ is not a Markov chain, but is a 2-Markov chain.

b) Let $Z(t) = (X(t-1), X(t))$. As explained in the text, this will be a Markov chain. List its states and calculate its state transition matrix.

4.1.4. (a) Consider the Wright-Fisher chain with $N = 10$. Suppose $P(X(0)=5) = 1$. Find $P(X(2)=10)$. (Follow the procedure of Exercise 4.1.1.)

(b) For the Moran model with $N = 4$ and $\mathbb{P}(X(0)=1) = 1$, calculate $\mathbb{P}(X(3)=0)$ and $\mathbb{P}(X(3)=1)$.

4.1.5. In this problem, we consider a process in the spirit of the Moran chain, but it is not a model of reproduction. Let $X(n)$ denote the number of black marbles at time t in an urn containing 100 black and white marbles. The number of black marbles $X(t+1)$ will be determined by the following experiment. Draw two marbles at random (random selection without replacement). 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. The experiments producing $X(n+1)$ from $X(n)$ for different n are independent. Show that $\{X(t)\}$ is a birth and death chain and compute its transition probabilities. Does this chain have absorbing states?

4.1.6. (Moran model with mutation). Suppose that in the Moran model, if a type A is chosen to reproduce its offspring mutates from A to a with probability u, and if a type a reproduces its offspring mutates to type A with probability v. Show that the transition probabilities are

$$\begin{aligned} p_{i,i+1} &= \frac{i(N-i)}{N^2}(1-u) + \frac{(N-i)^2}{N^2} \cdot v \\ p_{i,i-1} &= \frac{i(N-i)}{N^2}(1-v) + \frac{i^2}{N^2} \cdot u \\ p_{ii} &= \frac{i(N-i)}{N^2}(u+v) + \frac{(N-i)^2}{N^2}(1-v) + \frac{i^2}{N^2}(1-u) \end{aligned}$$

b) In the Moran model, at each time step a sample of size 2 is selected sequentially with replacement, and the first individual of the sample is copied and the second eliminated. Suppose instead that the sampling is done without replacement. Find the transition probabilities of the Markov chain in this case.

4.1.7. (a) (Wright-Fisher with mutation) Consider a locus with two alleles A and a. Suppose we assume all the hypotheses of the Wright-Fisher model, but we allow mutation, and A mutates to a with probability u, while a mutates to A with probability v, in the course of transmission to the next generation. Derive the transition probabilities for this model. (To be clear, first an allele is chosen from the allele pool of generation t. If it is A, it mutates to a with probability u or stays the same with probability $1-u$, and if it is a, it mutates to A with probability v or stays the same with probability v. The resulting allele is then placed in the allele pool of generation $t+1$. Remember, the population stays constant, so there $2N$ alleles in each generation.)

(b) (With selection) Using the approach of Section 3.4.2, modify the Wright-Fisher model so that it includes selection. Again assume just two alleles. Do not allow mutation.

In this problem, take as the state, the number of A alleles in the allele pool of generation t at its time of birth, and assume that N individuals are born in each new generation. To deduce the transition probabilities, we need to know the probability that when a parent is chosen from generation t, it contributes an allele A. As in Chapter 3, this is the probability it contributes A given it has survived, and you compute this as in Chapter 3 in terms of the selection coefficients.

4.1.8. What is the average lifetime of an individual in the Moran model? (Once an individual is born into the population, how long can it expect to survive on average? Observe that at each time step all surviving individuals are equally likely to die.)

4.1.9. Assume that $\{X(t);\, t \geq 0\}$ is a 2-Markov chain, as defined in Example 4.1.8. Thus

$$\begin{aligned}&\mathbb{P}\left(X(t+1)=x \;\middle|\; X(t)=x_t, X(t-1)=x_{t-1}, \ldots, X(0)=x_0\right)\\ &\quad = \mathbb{P}\left(X(t+1)=x \;\middle|\; X(t)=x_t, X(t-1)=x_{t-1}\right). \end{aligned} \tag{4.20}$$

We discussed how to embed $\{X(t);\, t \geq 0\}$ in a Markov chain in Example 4.1.8. Here we consider a direct analysis. Define the transition probabilities

$$a_{yz,u} \stackrel{\triangle}{=} \mathbb{P}\left(X(t)=u \;\middle|\; X(t)=z, X(t-1)=y\right).$$

We can still compute the probability of a path relatively easily assuming (4.20). Show that

$$\begin{aligned}&\mathbb{P}\left(X(t)=j_t, X(t-1)=j_{t-1}, \ldots, X(0)=j_0\right) =\\ &\quad \mathbb{P}\left(X(0)=j_0, X(1)=j_1\right) a_{j_0 j_1, j_2} a_{j_1 j_2, j_3} a_{j_2 j_3, j_4} \cdots a_{j_{t-2} j_{t-1}, j_t}.\end{aligned}$$

(Hint: start your calculation applying $\mathbb{P}(A \cap B) = \mathbb{P}(A \;\middle|\; B)\mathbb{P}(B)$ and property (4.20) to write

$$\begin{aligned}&\mathbb{P}\left(X(t)=j_t, X(t-1)=j_{t-1}, \ldots, X(0)=j_0\right)\\ &\quad = \mathbb{P}\left(X(t)=j_t \;\middle|\; X(t-1)=j_{t-1}, \ldots, X(0)=j_0\right) \mathbb{P}\left(X(t-1)=j_{t-1}, \ldots, X(0)=j_0\right)\end{aligned}$$

The first term can be expressed more simply using the transition probabilities. Continue the calculation using this technique.)

4.1.10. (A generalization of simple random walk.) Let $Y_1, Y_2, \ldots$ be independent, integer-valued random variables all with the same probability mass function

$$P(Y_i=m) = q_m, \qquad m \in \{\cdots, -2, -1, 0, 1, 2, \cdots\}.$$

Let $X(0) = 0$ and $X(t) = \sum_{i=1}^{t} Y_i$ for positive integers t. Show that X is an integer-valued Markov chain and find a formula for the transition probability p_{ij} in terms of $\{q_m; m = \cdots, -2, -1, 0, 1, 2, \cdots\}$.

*4.1.11.*Verify the solution in (4.19) to the gambler's ruin problem.

4.2 Theory of Markov Chains, Part I: Computation

Let $\{X(t);\, t \geq 0\}$ be a Markov chain evolving in a state space $\mathcal{E}$. We shall denote the probability mass function of $X(t)$ by $\rho(t) = \{\rho_i(t);\, i \in \mathcal{E}\}$; it is the function,

$$\rho_i(t) := \mathbb{P}\Big(X(t)=i\Big), \qquad i \in \mathcal{E},$$

returning, for each i, the probability that $X(t)$ is in state i. We also refer to $\rho(t)$ as the *probability distribution* of $X(t)$. One of the main goals of Markov chain theory is understanding how $\rho(t)$ evolves as time progresses. This is directly relevant to applications. To test a model against reality, we must know what it predicts, but the predictions of a stochastic model are not specific values of future states—these will be random. Instead they are the probabilities distributions of future states. Thus, for example, to see if the Wright-Fisher model fits data, we would like to be able to now the probabilities of different allele frequencies in future generations.

A closely related issue is computing expectations of the general form, $E[h(X(t)]$. For instance, when $X(t)$ represents a genotype frequency, we may wish to calculate its mean, $E[X(t)]$, and variance, $\mathrm{Var}(X(t))$. By definition of expected value,

$$E[h(X(t)] = \sum_{i\in\mathcal{E}} h(i)\mathbb{P}(X(t)=i) = \sum_{i\in\mathcal{E}} h(i)\rho_i(t).$$

Thus, formulas for the distribution $\rho(t)$ will translate to formulas for expectations.

In this section, we will derive simple formulae for computing probability distributions and expectations for Markov chains. These formulae will make essential use of matrix algebra. The next section, Section 4.3, will address the consequences of these formulae for the qualitative behavior of Markov chains.

4.2.1 Computing n-step transition probabilities and $\rho(t)$.

Consider a generic Markov chain evolving in a state-space $\mathcal{E}$. We are free to label the states of $\mathcal{E}$ as we like, and, for a general discussion, it is convenient to let $\mathcal{E} = \{1, 2, 3, \ldots, N\}$, when $\mathcal{E}$ has a finite number, N, of elements, and to let $\mathcal{E} = \{1, 2, 3, \ldots\}$, when $\mathcal{E}$ is countably infinite. The state transition matrix is then $A = [p_{ij}]_{1\leq i,j\leq\mathcal{E}}$, where the states are taken in their natural order. (When $\mathcal{E}$ is infinite, we can think of A as a matrix with an infinite number of rows, each infinitely long.) In the rest of the section, assume $\mathcal{E} = \{1, \ldots, N\}$. The extension of the conclusion to the case of infinite $\mathcal{E}$ is straightforward.

In what follows, it will be important to interpret the probability mass function, $\rho(t)$, as a *row* vector: in the case of a finite state space,

$$\begin{aligned} \rho(t) &:= \Big(\rho_1(t), \rho_2(t), \ldots, \rho_N(t)\Big) \\ &= \Big(\mathbb{P}(X(t)=1), \ldots, \mathbb{P}(X(t)=N)\Big). \end{aligned}$$

(When, $\mathcal{E}$ is infinite, think of this as a row vector of infinite length.)

Consider $\rho_j(X(t)) = \mathbb{P}(X(t)=j)$, and what happens if we try to compute this probability by conditioning on $X(t-1)$. One of the disjoint events $\{X(t-1)=k\}$, $k \in \mathcal{E}$, must occur, and, so by the rule of total probability (see (2.18) in Chapter 2),

$$\rho_j(t) = \mathbb{P}(X(t)=j) = \sum_{k=1}^{N} \mathbb{P}\Big(X(t)=j|X(t-1)=k\Big)\mathbb{P}\Big(X(t-1)=k\Big).$$

By definition, $\mathbb{P}\Big(X(t)=j|X(t-1)=k\Big) = p_{kj}$ and $\mathbb{P}\Big(X(t-1)=k\Big) = \rho_k(t-1)$, and so

$$\rho_j(t) = \sum_{k=1}^{N} \rho_k(t-1)p_{kj}, \quad \text{for each } j \in \mathcal{E}. \tag{4.21}$$

This sum is the j^{th} component of the product

$$\rho(t-1)\cdot A = \Big(\rho_1(t-1),\ldots,\rho_N(t-1)\Big)\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix}.$$

If follows that

$$\rho(t) = \rho(t-1)\cdot A, \quad t \geq 1. \tag{4.22}$$

(There is no problem in making sense of this formula when $\rho(t)$ and A are infinite, as long as we define $[\rho(t-1)\cdot A]_j$ as $\sum_{k\in\mathcal{E}} \rho_k(t-1)p_{kj}$ for all j.) The following theorem collects some consequences of this important identity.

Theorem 3 *Let $\{X(t)\}$ be a time-homogeneous Markov chain with state transition matrix A.*

a) For any $0 \leq s < t$,

$$\rho(t) = \rho(t-s)\cdot A^s. \tag{4.23}$$

In particular, for any $t \geq 1$,

$$\rho(t) = \rho(0)\cdot A^t. \tag{4.24}$$

b) For any states i and j, the t-step-ahead transition probability is

$$P\Big(X(s+t)=j \;\Big|\; X(s)=i\Big) = [A^t]_{ij}. \tag{4.25}$$

Proof: Equation (4.23) is proved by repeated application of (4.22):

$$\begin{aligned} \rho(t) &= \rho(t-1)\cdot A = [\rho(t-2)\cdot A]\cdot A = \rho(t-2)\cdot A^2 \\ &= [\rho(t-3)\cdot A]\cdot A^2 = \rho(t-3)\cdot A^3 \\ &= \cdots = \rho(t-s)\cdot A^s. \end{aligned}$$

To prove part b), suppose $\mathbb{P}(X(0)=i)=1$; thus, $\rho_i(0)=1$ and $\rho_j(0)=0$ if $j\neq i$. Then, on the one hand,

$$\mathbb{P}(X(t)=j)=[\rho(0)\cdot A^t]_j=\sum_{k\in\mathcal{E}}\rho_k(0)[A^t]_{kj}=[A^t]_{ij}.$$

On the other, by the rule of total probability

$$\mathbb{P}(X(t)=j)=\sum_{k\in\mathcal{E}}\mathbb{P}(X(t)=j|X(0)=k)\mathbb{P}(X(0)=k)=\mathbb{P}(X(t)=j|X(0)=i),$$

and thus b) follows. ◇

It is worth emphasizing what part b) of this theorem says: for every positive integer t, A^t is the matrix of t-step-ahead probabilities of a Markov chain with transition matrix A. Parts a) and b) of Theorem 3 are really aspects of the same fact, and are stated separately only for clarity. We have derived b) from a). Conversely, b) implies a) by the following calculation, which first employs the rule of total probabilities and then uses b),

$$\begin{aligned}\rho_j(t) &= \sum_{i\in\mathcal{E}}P(X(t)=j|X(0)=i)\rho_i(0)\\ &= \sum_{i\in\mathcal{E}}\rho_i(0)[A^t]_{ij}=[\rho(0)\cdot A^t]_j.\end{aligned}$$

The formulas of Theorem 3 are wonderful results. Consider their implications for Markov population genetics models. No matter how complicated the underlying dynamics, whether they incorporate selection and mutation or not, there is an explicit formula for the probability distribution of any future genotype frequency and it requires only matrix multiplication. In comparison, the infinite population model with selection in Chapter 3 led to a nonlinear difference equation without an explicit solution.

While the explicit formulae of Theorem 3 are very nice, they still require work to interpret. The computation of powers of large and complicated matrices is not trivial, either explicitly or numerically. For example, there are no known explicit formulae for the entries of A^t, when A is the transition matrix of the Wright-Fisher or Moran chains, even for small population sizes. Nevertheless, the behavior of A^t as $t\to\infty$ is understood in great detail. The main aspects of this limit theory are discussed in Section 4.3.

As a first example using Theorem 3, it is instructive to study a case in which explicit calculations *are* possible.

Example 4.2.1. Let $\mathcal{E}=\{0,1\}$ and consider the state transition matrix

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

If $\lambda = 0 = \mu$, the chain stays in its initial state for all time. If $\lambda = \mu = 1$, the chain moves deterministically and periodically, alternately visiting states 0 and 1. To avoid these uninteresting and non-stochastic cases, assume $0 < \lambda + \mu < 2$. Our object is to compute A^t for all integers $t \geq 1$. This is not entirely trivial. Calculation by hand (well, actually I used the Maple software package), gives

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

and

$$A^3 = \begin{pmatrix} (1-\lambda)^3 + \lambda\mu(3-2\lambda-\mu) & 1-(1-\lambda)^3 - \lambda\mu(3-2\lambda-\mu) \\ 1-(1-\mu)^3 - \lambda\mu(3-\lambda-2\mu) & (1-\mu)^3 + \lambda\mu(3-\lambda-2\mu) \end{pmatrix}.$$

No simple pattern is immediately evident.

However, by being more astute, a simple formula can be found:

$$A^t = \frac{1}{\lambda+\mu} \begin{pmatrix} \mu+\lambda\alpha^t & \lambda - \lambda\alpha^t \\ \mu - \mu\alpha^t & \lambda + \mu\alpha^t \end{pmatrix} \quad \text{where } \alpha = 1-\lambda-\mu. \tag{4.26}$$

Rather than derive this from scratch, let us just check that it works. Let $B(t)$ denote the matrix on the right-hand-side of (4.26). The reader should verify the following—it requires only straightforward calculation—that

$$B(0) = I \ \text{ (the } 2\times 2 \text{ identity matrix), and } \ B(t+1) = A \cdot B(t).$$

From these two facts it follows that: $B(1) = A \cdot B(0) = A \cdot I = A$; then, that $B(2) = A \cdot B(1) = A \cdot A = A^2$; then, that $B(3) = A \cdot B(2) = A^3$; and, continuing in this fashion, that $B(t) = A^t$ for all $t \geq 1$.

The explicit formula (4.26) allows one to compute $\lim_{t\to\infty} A^t$. Under the assumption that $0 < \lambda + \mu < 2$, the constant $\alpha = 1 - \lambda - \mu$, satisfies $-1 < \alpha < 1$, and hence $\lim_{t\to\infty} \alpha^t = 0$. Thus,

$$\lim_{t\to\infty} A^t = \begin{pmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \\ \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{pmatrix}.$$

This is an interesting result. Since $P(X(t+1)=j|X(0)=i) = [A^t]_{ij}$ by Theorem 3, it says

$$\lim_{t\to\infty} P(X(t)=0|X(0)=0) = \frac{\mu}{\lambda+\mu} = \lim_{t\to\infty} P(X(t)=0|X(0)=1),$$
$$\lim_{t\to\infty} P(X(t)=1|X(0)=0) = \frac{\lambda}{\lambda+\mu} = \lim_{t\to\infty} P(X(t)=1|X(0)=1).$$

The limiting conditional probabilities of $X(t)$ are independent of $X(0)$! In fact, no matter what the initial distribution $\rho(0) = (\rho_0(0), \rho_1(0))$ is,

$$\begin{aligned}
\lim_{t\to\infty} \; & (P(X(t)=0), P(X(t)=1)) = \lim_{t\to\infty} \rho(t) = \lim_{t\to\infty} \rho(0) \cdot A^t \\
&= \lim_{t\to\infty} (\rho_0(0), \rho_1(0)) \cdot \begin{pmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \\ \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{pmatrix} \\
&= \left((\rho_0(0) + \rho_1(0)) \frac{\mu}{\lambda+\mu}, (\rho_0(0) + \rho_1(0)) \frac{\lambda}{\lambda+\mu} \right) \\
&= \left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu} \right). \qquad (4.27)
\end{aligned}$$

Thus the limiting probabilities are independent of initial distribution altogether. This is not an accident. Suppose the chain starts in state 0; eventually it will enter state 1, and from that point on, it forgets how it got to 1—this is the Markov property—and behaves like the chain starting in 1. Thus the distribution of $X(t)$, in the limit as $t \to \infty$, should not depend on its distribution at time 0. In Section 4.4, we discuss conditions on general, finite-state-space chains that guarantee the existence of a limit, $\lim_{t\to\infty} \rho(t)$, which is the same for all initial distributions. $\diamond$

Example 4.2.2. When $0 < \lambda, \mu < 1$ in the two-state chain example, neither state is absorbing. It is instructive to compute powers of chain with an absorbing state to see how the t-step-ahead transition probabilities evolve. For simplicity, we shall consider the Moran model with $N = 4$, In Example 4.1.6 we calculated its state transition matrix to be

$$A := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{3}{16} & \frac{10}{16} & \frac{3}{16} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{3}{16} & \frac{10}{16} & \frac{3}{16} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Two of its powers, computed using the Maple software package, are shown below, rounded to three significant digits:

$$A^4 := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .448 & .253 & 175 & .100 & .024 \\ .149 & .233 & .237 & .233 & .149 \\ .024 & .100 & 175 & .253 & .448 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$A^{16} := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .697 & .036 & .035 & .035 & .196 \\ .429 & .047 & .047 & .047 & .429 \\ .196 & .035 & .035 & .036 & .697 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Any power of a stochastic matrix is a stochastic matrix, and so A^4 and A^{16} are stochastic matrices. A few of the rows in these examples do not quite add to one, but this is due only to round-off error.

Several features of A and its powers are immediately apparent. First, if you rotate them by 180 degrees you get the same matrices back. This reflects a symmetry of the Moran chain. Since $p_{i,i+1} = p_{i,i-1}$ for $i = 1, 2, 3$ and since states 0 and 4 are both absorbing, the chain would look the same if we relabeled the states in reverse order.

Secondly, notice that the first and last rows are the same in A, A^4 and A^{16}. In fact, they are the same in all powers, A^t. This is easy to see. The first row of A is the row of transition probabilities out of state 0. But, because 0 is an absorbing state and the chain can never leave state 0, we have for any $t \geq 1$ that $[A^t]_{00} = \mathbb{P}(X(t) = 0 | X(0) = 0) = 1$ and $[A^t]_{0j} = \mathbb{P}(X(t) = j | X(0) = 0) = 0$, if $j \neq 0$. The same analysis applies to the absorbing state 4, corresponding to the last row of A.

Consider next the entries of the first column in A^t,

$$[A^t]_{i0} = \mathbb{P}(X(t) = 0 | X(0) = i)$$

for $i = 1, 2, 3$. These are the probabilities to be in state 0 at time t starting from a non-absorbing state. But since a chain never leaves an absorbing state once it gets there, $[A^t]_{i0}$ is the same as the probability the chain hits state 0 *at or before* time t. The probability can only increase with t, since additional time gives only more opportunity to reach 0. This is clear for the matrices we computed; for instance, $[A^4]_{10} = 0.448 < 0.697 < [A^{16}]_{10}$.

Finally, consider starting in a non-absorbing state, for example, state 1. The probability that the chain has not hit either of the absorbing states, 0 or 4, by time t is the sum, $[A^t]_{11} + [A^t]_{12} + [A^t]_{13}$, of the probabilities to be in the non-absorbing states. This probability can only decrease as t increases. For $t = 4$, it is approximately $0.253 + 0.175 + 0.100 = 0.528$, while for $t = 16$, it is $0.036 + 0.035 + 0.035 = 0.106$. The probability $[A^t]_{21} + [A^t]_{22} + [A^t]_{23} = \mathbb{P}\Big(X(t) \neq 0, X(t) \neq 1 | X(0) = 2\Big)$ experiences a similar rapid decline as t increases from 4 to 16. On this evidence, it is natural to guess that

$$\lim_{t \to \infty} \mathbb{P}(X(t) = j | X(t) = i) = \lim_{t \to \infty} [A^t]_{ij} = 0$$

if $1 \leq i, j \leq 3$. Equivalently, the Moran chain eventually hits an absorbing state, no matter where it starts. In fact this is true, not only for Moran chains, but much more generally. We will state a precise result about reaching absorbing state in Section 4.3, where we we also show how to compute the probability to end up in a particular absorbing state.

4.2.2 Expectations for Markov chains.

Conditional expectations and expectations are also simple to compute for Markov chains, and the formulas for doing so can again be expressed in the language of matrix algebra.

If g is a function defined on the state space $\mathcal{E}$ of $\{X(t)\}$, then, by definition of expectation,

$$\begin{aligned} E\left[g(X(t+1)) \mid X(t)=i\right] &= \sum_{j\in\mathcal{E}} g(j)\mathbb{P}\left(X(t+1)=j|X(t)=i\right) \\ &= \sum_{j\in\mathcal{E}} p_{ij}g(j), \qquad \text{for all } i\in\mathcal{E}. \end{aligned} \tag{4.28}$$

Let us again work with the generic state space $\mathcal{E}=\{1,2,\ldots,N\}$. Then the right-hand side of the last expression is the sum $\sum_{j=1}^{N} p_{ij}g(j)$, and this is precisely the i^{th} component of the product of $A=[p_{ij}]_{1\le i,j\le N}$ with the *column* vector

$$\mathbf{g} := \begin{pmatrix} g(1) \\ g(2) \\ \vdots \\ g(N) \end{pmatrix}.$$

Since (4.9) is true for each state i, it follows that

$$\begin{aligned} \begin{pmatrix} E\left[g(X(t+1)) \mid X(t)=1\right] \\ \vdots \\ E\left[g(X(t+1)) \mid X(t)=N\right] \end{pmatrix} &= \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix} \begin{pmatrix} g(1) \\ g(2) \\ \vdots \\ g(N) \end{pmatrix} \\ &= A\cdot\mathbf{g} \end{aligned} \tag{4.29}$$

This has an easy extension to t-step ahead conditional expectations, derived by the same reasoning:

$$\begin{pmatrix} E\left[g(X(s+t)) \mid X(s)=1\right] \\ \vdots \\ E\left[g(X(s+t)) \mid X(s)=N\right] \end{pmatrix} = A^t\cdot\mathbf{g} \tag{4.30}$$

These formulas show us how to compute conditional expectations given the present state. It is also easy to compute straight expectations, $E[g(X(t)]$. By definition

$$E[g(X(t))] = \sum_{i\in\mathcal{E}} g(i)\mathbb{P}(X(t)=i) = \sum_{i\in\mathcal{E}} \rho_i(t)g(i),$$

where we have used our notation $\rho_i(t)$ for $\mathbb{P}(X(t)=i)$. But this last sum is just the product

$$(\rho_1(t),\ldots,\rho_N(t))\begin{pmatrix} g(1) \\ g(2) \\ \vdots \\ g(N) \end{pmatrix} = \rho(t)\cdot\mathbf{g}.$$

It follows that

$$E[g(X(t))] = \sum_{i\in\mathcal{E}} \rho_i(t)g(i) = \sum_{i\in\mathcal{E}} \rho_i(t)[\mathbf{g}]_i = \rho(t)\cdot\mathbf{g} = \rho(0)\cdot A^t\cdot\mathbf{g} \tag{4.31}$$

Example 4.2.3. Expectation for the selectively neutral Wright-Fisher chain.

Allele numbers in the Wright-Fisher model fluctuate randomly; this is called *genetic drift*. What about the *average* numbers of alleles? In this example, we will study the Wright-Fisher chain with no mutation or selection, as defined in Example 4.1.5, and we will show that it is constant in expectation. That is, we will show that

$$E[X(t+1)|X(t)=i] = i, \tag{4.32}$$

and, whatever the initial distribution is,

$$E[X(t)] = E[X(0)], \quad \text{for all } t \geq 1. \tag{4.33}$$

This result reflects the absence of mutation and selection in the model, and it is the close cousin of the fact that allele frequencies are constant in infinite population models when mutation and selection do not act. Indeed, since the infinite population assumption means equating allele frequencies with their expected values, the two results are essentially the same. Despite this, the Wright-Fisher chain behaves in a strikingly different fashion as time increases. As we will see in the next section, the Wright-Fisher chain must eventually end up in one of the absorbing states, 0 or $2N$. Thus, after a long time, the chain will be in either state 0 or state $2N$ with probability close to one, but balanced in probability between these states so that $E[X(t)]$ is always $E[X(0)]$!

To see why (4.32) is true, recall the definition of the Wright-Fisher chain. We found that, *given* $X(t) = i$, the number, $X(t+1)$, of A alleles in the next generation is a binomial random variable with parameters $n = 2N$ and $p = i/2N$. The expectation of a binomial random variable with parameters n and p is np (see Example 2.14 of Chapter 2). Thus $E[X(t+1)|X(t)=i] = (2N)(i/2N) = i$, as claimed.

We use this fact to prove $E[X(t+1)] = E[X(t)]$. By conditioning on $X(t)$ and using Theorem 9 of Chapter 2,

$$E[X(t+1)] = \sum_{i=0}^{2N} E[X(t+1)|X(t)=i]\mathbb{P}(X(t)=i) = \sum_{i=0}^{2N} i\mathbb{P}(X(t)=i) = E[X(t)].$$

It is interesting to look at this calculation from the point of view of formulas (4.29) and (4.31). Let h denote the identity function, $h(i) = i$ on the state space $\mathcal{E} = \{0, 1, \ldots, 2N\}$ of the Wright-Fisher chain. Let

$$\mathbf{h} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ \vdots \\ 2N \end{pmatrix}$$

be the associated vector. Then

$$E[h(X(t+1))|X(t)=i] = E[X(t+1)|X(t)=i] = i = h(i),$$

and, from (4.29), this is the same as

$$\mathbf{h} = A \cdot \mathbf{h},$$

where A is the state transition matrix of the Wright-Fisher chain. Thus $\mathbf{h}$ is an eigenvector of A with eigenvalue 1, and as a consequence, $A^t \cdot \mathbf{h} = \mathbf{h}$ for all $t \geq 1$. (For example, $A^2\mathbf{h} = A(A\mathbf{h}) = A\mathbf{h} = \mathbf{h}$; $A^3\mathbf{h} = A(A^2\mathbf{h}) = A\mathbf{h} = \mathbf{h}$, etc.) From equation (4.31),

$$\begin{aligned} E[X(t)] = E[h(X(t))] &= \rho(0) \cdot A^t \cdot \mathbf{h} = \rho(0) \cdot \mathbf{h}. \\ &= \sum_{i=0}^{2N} h(i)\rho_i(0) = E[h(X(0))] = E[X(0)], \end{aligned}$$

for all t.

The expectation, $E[X(t)]$, is also constant for the Moran model with no mutation or selection. The reader is asked to show this in Exercise 4.2.4. ◇

4.2.3 Exercises

4.2.1 Let $B(t)$ be the matrix in equation (4.26). To complete the proof in Example 4.2.1 that $A^t = B(t)$, show that $B(0) = I$ and that $B(t+1) = A \cdot B(t)$.

4.2.2 Suppose $\mathbb{P}(X(0) = 0) = 0.4$ and $\mathbb{P}(X(0) = 1) = 0.6$ for the two step chain of example 4.2.1. Calculate (in terms of λ and μ),

a) $\mathbb{P}(X(4) = 1)$.

b) $E[g(X(4)]$, where $g(0) = 1$ and $g(1) = 2$.

4.2.3. a) Verify that for the selectively neutral, mutation free Wright-Fisher model

with $2N = 4$,

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{81}{256} & \frac{27}{64} & \frac{27}{128} & \frac{3}{64} & \frac{1}{256} \\ \frac{1}{16} & \frac{1}{4} & \frac{3}{8} & \frac{1}{4} & \frac{1}{16} \\ \frac{1}{256} & \frac{3}{64} & \frac{27}{128} & \frac{27}{64} & \frac{81}{256} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

b) Calculation with Maple shows

$$A^2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .463 & .233 & .178 & .092 & .034 \\ .166 & .211 & .246 & .211 & .166 \\ .034 & .092 & .178 & .233 & .463 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

$$A^4 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .604 & .100 & .102 & .080 & .114 \\ .312 & .121 & .136 & .121 & .312 \\ .114 & .080 & .102 & .100 & .604 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

$$A^{10} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .725 & .016 & .018 & .016 & .225 \\ ..466 & .022 & .024 & .022 & .466 \\ .225 & .016 & .018 & .016 & .725 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

(Some rounding error enters these calculations.) Use these matrices to answer the following questions.

(i) Determine $P(X(5) = 2|X(3) = 4)$.

(ii) Determine $P(X(4) = 0|X(0) = 3)$.

(iii) Suppose the distribution of $X(0)$ is given by $P(X(0) = 1) = 0.2$, $P(X(0) = 2) = 0.3$, $P(X(0) = 3) = 0.5$, and $P(X(0) \in \{0, 4\}) = 0$. Find the distribution of $X(10)$. Find $E[X^2(5)]$.

4.2.4. Prove (4.32) and (4.33) for the Moran model defined in Example 4.1.6.

4.2.5. In the Wright-Fisher model model with mutation derived in Exercise 4.1.7, the conditional distribution of $X(t+1)$ given $X(t) = i$ is binomial with parameters $n = 2N$ and $p = v + (1 - u - v)(i/2N)$. Here u is the probability A mutates to a and v is the probability a mutates to A. Let $Y(t) = X(t)/2n$ be the frequency of A.

Show that $E[Y(t+1)] = v + (1-u-v)E[Y(t)]$. (Compare to the mutation model of Section 3.3.5). Find $\lim_{t\to\infty} E[Y(t)]$.

4.2.6. Consider a Markov chain evolving in $\{0, 1, 2\}$ with state transition matrix,

$$A = \begin{pmatrix} \frac{1-\lambda}{1+\delta} & \frac{\lambda}{1+\delta} & \frac{\delta}{1+\delta} \\ \\ \frac{\mu}{1+\delta} & \frac{1-\mu}{1+\delta} & \frac{\delta}{1+\delta} \\ \\ 0 & 0 & 1 \end{pmatrix}.$$

Assume $\delta > 0$ and $0 < \lambda, \mu < 1$. Let $\alpha = 1 - \lambda - \mu$. Show that

$$A^t = \frac{1}{(1+\delta)^n} \begin{pmatrix} \frac{\mu+\lambda\alpha^t}{\lambda+\mu} & \frac{\lambda-\lambda\alpha^t}{\lambda+\mu} & (1+\delta)^n - 1 \\ \frac{\mu-\mu\alpha^t}{\lambda+\mu} & \frac{\lambda+\mu\alpha^t}{\lambda+\mu} & (1+\delta)^n - 1 \\ 0 & 0 & (1+\delta)^n \end{pmatrix}$$

Find $\lim_{t\to\infty} A^t$ and interpret what this says about the behavior of the corresponding Markov chain.

4.3 Limit behavior of Markov chains

In the last section, we raised the problem of calculating the distribution, $\rho(t)$, of $X(t)$, and exploring how it evolves with time. We know that $\rho(t)$ depends only on the initial distribution, $\rho(0)$, of the chain and on the state transition matrix, A, through the formula $\rho(t) = \rho(0) \cdot A^t$. So, to analyze $\rho(t)$ for any possible $\rho(0)$, we need to look at A^t as a function of t. The major results of Markov chain theory give us a very complete understanding of the limiting behavior of A^t as $t \to \infty$. As it turns out, $\lim_{t\to\infty} A^t$ exists quite generally, given only natural and generic conditions on the chain, and there are simple, algebraic characterizations of this limit. In this section, we will describe some of these results and illustrate how they are applied. We will focus on important cases of applied interest, rather than present the most general theory, and we will omit most proofs.

In what follows, the reader should keep in mind the two examples of the previous section: the Moran model with absorbing states (Example 4.2.2), and the two-state Markov chain (Example 4.2.1), especially when $\lambda > 0$ and $\mu > 0$. These were simple enough to explore by direct computation and we discovered some interesting results. The Moran model has the property that either absorbing state can be reached from any non-absorbing state. Numerical evidence from computing A^t for some higher powers of t suggested that the Moran chain must eventually eventually hit one of its absorbing states, no matter where its starts from. This turns out to be true and reflects a general fact: if a Markov chain in a finite state space can get from any state to an absorbing state, then it must eventually end up in an absorbing state. It

cannot knock about in the non-absorbing states forever. We discuss this in Section 4.3.2, where we also show how to compute the probability to end up in a particular absorbing state.

In contrast, the two-state chain, when both $p_{01} = \lambda > 0$ and $p_{10} = \mu > 0$, has no absorbing states. In this case we were able to to compute A^t exactly and to calculate $\lim_{t\to\infty} A^t$. We found as a consequence that a limiting distribution $\lim_{t\to\infty}\Big(\mathbb{P}(X(0)=0), \mathbb{P}(X(0)=0)\Big)$ exists and is independent of the initial state of the chain. This also reflects a general fact: in a finite state space chain which can move between any two states, the distribution of $X(t)$ settles down to a limit independent of its initial distribution. More is true: the average amount of time spent by the chain in state i tends to the limiting probability to be in i, in the infinite time limit. This is a generalization to Markov chains of the law of large numbers and is known as an *ergodic theorem*. This theory is treated in Section 4.3.3.

4.3.1 Classification of states.

The long-time behavior of a Markov chain will obviously depend on the possible ways it can move among its states. We have seen this already in the difference between the Moran model with absorbing states of Example 4.2.2 and the two state chain without absorbing states. This section introduces more refined distinctions between states, based on the way in which the chain visits them.

The definitions of this section will be based on the concept of the *hitting time* of a state. For any state i, let T_i denote the first time t, $t = 1, 2, \ldots$ at which $X(t) = i$; if the chain never hits i for $t \geq 1$, set $T_i = \infty$. Mathematically,

$$T_i := \min\{t;\, t \geq 1 \text{ and } X(t) = i\}.$$

We emphasize that when $X(0) = i$, T_i is the first time *after* $t = 0$, at which the chain *returns* to i.

When i and j are different states. j is said to be *accessible* from i, written $i \to j$, if

$$\mathbb{P}(T_j < \infty | X(0) = i) > 0.$$

In words, j is accessible from i if, when starting from i, the chain eventually hits j with a positive probability. It is fairly easy to check accessibility. Suppose there is a path, $\{i = i_0, i_1, \ldots, i_{m-1}, i_m = j\}$, connecting i to j, along which $p_{i_k, i_{k+1}} > 0$ for every pair of successive states. Then, using the formula for the probability of a path from Theorem 1,

$$\mathbb{P}\Big(X(m)=j, X(m-1)=i_{m-1}, \ldots, X(1)=i_1 \Big| X(0)=i\Big) = p_{i_0,i_1} \ldots p_{i_{m-1},j} > 0,$$

and hence j is accessible from i. In fact, it is clear that $i \to j$ if and only if such a path exists. Equivalently $i \to j$ if and only if j can be reached from i by following arrows in the state transition diagram.

If $i \to j$ *and* $j \to i$, then we say states i and j *communicate* and we write $i \sim j$. As a matter of convention we say that i communicates with itself ($i \sim i$).

Examples 4.3.1. (a) If i is absorbing, then no other state is accessible from i, and so no other states can communicate with i.

(b) Consider the Moran model without mutation in the state space $\{0, 1, \ldots, N\}$. States 0 and N are absorbing states. Let i and j be any non-absorbing states. Without loss of generality, assume $0 < i < j < N$. They are connected by the simple paths

$$\{i, i+1, \ldots, j-1, j\} \text{ and, in reverse, } \{j, j-1, \ldots, i+1, i\}.$$

But for every non-absorbing state k in the Moran models, $p_{k,k+1}$ and $p_{k,k-1}$ are both positive. Hence $p_{i,i+1} \cdots p_{j-1,j}$ and $p_{j,j-1} \cdots p_{i+1,i}$ are both positive, and $i \sim j$. From every state i, except $i = N$, $\{i, i-1, \ldots, 1, 0\}$ is a path of positive probability. Thus $i \to 0$ as long as $i \neq N$. Similarly, $i \to N$, as long as $i \neq 0$.

The situation is similar for the Wright-Fisher chain without mutation. In this case the state space is $\{0, 1, \ldots, 2N\}$, states 0 and $2N$ are absorbing, and $p_{ij} > 0$ as long as i is not 0 or $2N$. Thus, if $1 \leq i \leq 2N-1$ and j is any other state, $i \to j$, and any two, non-absorbing states communicate. ◇

A state i is said to be *recurrent* if

$$\mathbb{P}(T_i < \infty | X(0) = i) = 1;$$

in words, when i is recurrent, a chain starting out in state i must return to i at a later time with probability one, (since $\{T_i < \infty\}$ is the event of at least one return to i). A state that is not recurrent is said to be *transient*. The terms *recurrent* and *transient* are justified by the following theorem.

Theorem 4 *If i is transient, $\{X(t)\}$ visits i only finitely many times with probability one.*

If $X(0) = i$, and i is recurrent, then $\{X(t)\}$ visits i infinitely often with probability one.

A full proof of this result is given at the end of this section, but it is easy to understand intuitively. Because of the Markov property and the assumption of time-homogeneity, the probabilistic behavior of the chain *after* a visit to i is just the same as if it started in i at time $t = 0$. In other words, the chain starts over as chain beginning in state i, after each return to state i. If i is recurrent and $X(0) = i$, the chain must return to i at least once, by definition. But then once it gets to i, it restarts as a chain beginning in i, and, by recurrence, must return again. By repeating this argument *ad infinitum*, it must return infinitely often.

If a state i is recurrent, it is interesting to ask how long the first return to i takes on average; this is the conditional expectation, $E[T_i | X(0) = i]$. It can happen

that this expected time is infinite, even though a return to i must take place. This may seem counterintuitive, but it is the case, for example, with simple, symmetric random walk, as we will show later. To distinguish between finite and infinite expected return times, we say that a state i is *positive recurrent* if it is recurrent *and* if $E[T_i|X(0)=i] < \infty$. On the other hand a state i is *null recurrent* if it is recurrent and $E[T_i|X(0)=i] = \infty$.

If all the states of a Markov chain are positive (respectively, null) recurrent, then the chain itself is said to be positive (respectively, null) recurrent.

Examples 4.3.2. (a) Any absorbing state i, is positive recurrent, because then $\mathbb{P}(T_i = 1|X(0)=i) = 1$ and so $E[T_i|X(0)=i] = 1$.

(b) All the non-absorbing states in the Moran model without mutation are transient. For suppose the chain starts in any non-absorbing state, i . Clearly, there is a positive probability it ends up in an an absorbing state, and once that happens it cannot return to i. Hence, the probability of return to i is less than 1, and i is transient.

(c) Consider the two-state chain with $p_{01} = \lambda > 0$ and $p_{10} = \mu > 0$. These conditions mean the probabilities to reach 1 from 0 and 0 from 1 are both positive, and hence $0 \sim 1$. We will show later in this section that

$$E[T_0|X(0)=0] = \frac{\lambda+\mu}{\mu}. \tag{4.34}$$

By interchanging the roles of μ and λ, it follows likewise that

$$E[T_1|X(0)=1] = \frac{\lambda+\mu}{\lambda}. \tag{4.35}$$

Thus, both i and j are positive recurrent.

It is interesting to relate these mean return times to the limiting probabilities derived in Example 4.2.1. There we showed that

$$\lim_{t\to\infty} \Big(P(X(t)=0|X(0)=0), P(X(t)=1|X(0)=1)\Big) = \Big(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu}\Big)$$

From equations (4.34) and (4.35), it follows that

$$\lim_{t\to\infty} P(X(t)=0) = \frac{1}{E[T_0|X(0)=0]} \quad \text{and} \quad \lim_{t\to\infty} P(X(t)=0) = \frac{1}{E[T_1|X(0)=1]}. \tag{4.36}$$

Thus, the limiting probability to be in each state is is the inverse of the mean return time for that state. This makes perfect sense. Intuitively, as time elapses the chain will visit i once in every $E[T_i|X(0)=i]$ time steps, on average, and so the probability to be in state i should be $(E[T_i|X(0)=i])^{-1}$. This relationship between mean return time and limiting probability holds in general in chains with communicating states, a fact we shall explain in Section 4.3.3. ◇

The following theorem collects some other basic facts about transient and recurrent states that are important to know.

Theorem 5 *(i) A recurrent state i in a Markov chain evolving in a* finite *state space is automatically positive recurrent.*

(ii) If $i \sim j$, then states i and j are either both transient, both positive recurrent, or both null recurrent.

(iii) If all the states of a finite state space Markov chain communicate, the chain is positive recurrent.

This theorem helps make life easy, at least when life requires the analysis of Markov chains! For example, consider the Wright-Fisher or Moran model with mutation, when the probability of mutation from A to a and from a to A are both positive. The transition probabilities for these cases were derived in Exercises 4.1.6 and 4.1.7. The possibility of mutation prevents the possibility of absorbing states, and it is easy to see that $i \sim j$ for any two states. Without further work, we deduce from Theorem 5 that both these chains are positive recurrent.

We will not prove all the statements of Theorem 5, but they are mostly intuitive. For example, consider two communicating state i and j. Starting from i, we know from Theorem 4 that the chain returns to i infinitely often. In each excursion from i back to i, there is some positive chance to hit j. Just as an infinite sequence of independent coin tosses must produce an infinite number of heads if the probability of heads is positive, an infinite number of the excursions from i back to i must include a visit to j. Thus the chain visits j infinitely often, and so j is also recurrent.

Moreover, statement (iii) of the theorem is direct consequence of the first two. A Markov chain evolving in a finite state space must visit at least one of its states infinitely often, and by statements (i) and (ii) of the theorem, that state must be positive recurrent. But if all states communicate, then they must all be positive recurrent.

The remainder of this section supplies the derivation of equations (4.34) and (4.35), and the proof of Theorem 4. It is optional, but recommended.

Derivation of equations (4.34) and (4.35).

Suppose $X(0) = 0$. Since 0 and 1 are the only states, the first time, T_1, to hit state 1 is also the first time the chain exists state 0. Now, $T_1 > s$ if and only if $X(1) = 0, X(2) = 0, \ldots, X(s) = 0$. Thus

$$\begin{aligned} \mathbb{P}(T_1 > s \big| X(0) = 0) &= \mathbb{P}\Big(X(1) = 0, X(2) = 0, \ldots, X(s) = 0 \Big| X(0) = 0) \\ &= p_{00}^s = (1-\lambda)^s. \end{aligned} \tag{4.37}$$

This means that, given $X(0) = 0$, T_1 is a geometric random variable with parameter $p = \lambda$—see, Chapter 2, pages 17 and 18. In particular, since $0 < \lambda$ by assumption, and hence $0 \leq 1 - \lambda < 1$, it follows that

$$\mathbb{P}\big(T_1 = \infty \big| X(0) = 0\big) = \lim_{s \to \infty} \mathbb{P}\big(T_1 > s \big| X(0) = 0\big) = \lim_{s \to \infty} (1-\lambda)^s = 0.$$

This means that the chain cannot stay in state 0 forever; it must eventually move to state 1. In addition, because the expected value of a geometric random variable with parameter p is $1/p$ (see Chapter 2, page 34),

$$E[T_1 \big| X(0)=0] = \frac{1}{\lambda}. \tag{4.38}$$

The same type of reasoning applied using the assumption that $\mu > 0$, shows that, given $X(0) = 1$, the first time T_0 to hit 0 is a geometric random variable, and hence $\mathbb{P}(T_0 < \infty \big| X(0)=1) = 1$, and

$$E[T_0 \big| X(0)=1] = \frac{1}{\mu}. \tag{4.39}$$

Now consider the chain starting in state 0. It must eventually hit 1. But we have just seen the chain must leave state 1 eventually and return to 0, because $\mu > 0$. It follows that state 0 is recurrent. Likewise, state 1 is recurrent, since $\lambda > 0$ implies that the chain must return to 1 if it ever hits 0.

The first step in computing $E[T_0 \big| X(0)=0]$ explicitly, is to condition on the state $X(1)$ that the Markov chain moves to starting from $X(0) = 0$. Since it moves from 0 to 1 with probability λ and stays at 0 with probability $1 - \lambda$,

$$\begin{aligned} & E[T_0 \big| X(0)=0] \\ &= \quad E[T_0 \big| X(1)=0, X(0)=0](1-\lambda) + E[T_0 \big| X(1)=1, X(0)=0]\lambda. \end{aligned}$$

Now, in the last expression, $E[T_0 \big| X(1)=0, X(0)=0] = 1$, since $\{X(1)=0, X(0)=0\}$ implies $T_1 = 1$. On the other hand,

$$E[T_0 \big| X(1)=1, X(0)=0] = 1 + \frac{1}{\mu},$$

because, according to (4.39), the expected time to return from 1 to 0 is $1/\mu$ and this must be added to the first time step used in going from 0 to 1. Therefore,

$$E[T_0 \big| X(0)=0] = 1 - \lambda + \lambda\left(1 + \frac{1}{\mu}\right) = \frac{\lambda+\mu}{\mu},$$

as we wanted to prove. The same reasoning, but interchanging the roles of μ and λ, shows

$$E[T_1 \big| X(0)=1] = \frac{\lambda+\mu}{\lambda}.$$

◇

Proof of Theorem 4.

Let V_i denote the number visits the chain makes to state i strictly after time 0; to emphasize, if $X(0)=i$, the visit to i at time $t = 0$ is not counted in V_i.

$$r_i = \mathbb{P}(T_i < \infty \big| X(0)=i).$$

By definition, when $0 \le r_i < 1$ when i is transient, and $r_i = 1$ when i is recurrent. We shall show that

$$\mathbb{P}(V_i \ge n) = r_i^n \quad \text{for all } n. \tag{4.40}$$

As an immediate consequence,

$$\begin{aligned}\mathbb{P}(V_i=\infty\big|X(0)=i) &= \lim_{n\to\infty}\mathbb{P}(V_i\geq n\big|X(0)=i)=\lim_{n\to\infty}r_i^n\\ &= \begin{cases}0, & \text{if } 0\leq r_i<1;\\ 1, & \text{if } r_i=1.\end{cases}\end{aligned}$$

It remains to show why (4.40) is true. (Incidentally, when i is transient, (4.40) implies that V_i+1 is a geometric random variable.) The case, $n=1$, is easy, because $\{V_i\geq 1\}$, the event of at least one return to i, means exactly the same thing as $\{T_i<\infty\}$. Thus

$$\mathbb{P}(V_i\geq 1\big|X(0)=i)=\mathbb{P}(T_i<\infty\big|X(0)=i)=r_i.$$

Next consider the event $\{V_i\geq n+1\}$. If this happens, the time T_i of first return to i is finite, and so from the rule of total probabilities,

$$\mathbb{P}\big(V_i\geq n+1\big|X(0)=i\big)=\sum_{t=1}^{\infty}\mathbb{P}\big(V_i\geq n+1\big|T_i=t,X(0)=i\big)\mathbb{P}(T_i=t\big|X(0)=i). \tag{4.41}$$

Now, given $\{T_i=t\}$, the chain returns to i at least $n+1$ times if and only if it makes at least n visits to i *after* time t: Thus,

$$\mathbb{P}\Big(V_i\geq n+1\big|T_i=t,X(0)=i\Big)=\mathbb{P}\Big(\text{at least } n \text{ visits to } i \text{ after time } t\Big|T_i=t,X(0)=i\Big).$$

Here, we can use the Markov property: the event $\{T_i=t,X(0)=i\}$ can be expressed as $\{X(t)=i,X(0)=i\}\cap\{X(s)\neq i, \text{ if } 0<s<t\}$, and the event of at least n visits to i after time t is an event in the future beyond time t. Therefore, by the Markov property stated in Theorem 2, only $\{X(t)=i\}$ is relevant to the conditional probability, and

$$\begin{aligned}&\mathbb{P}\Big(\text{at least } n \text{ visits to } i \text{ after time } t\Big|T_i=t,X(0)=i\Big)\\ &\qquad=\mathbb{P}\Big(\text{at least } n \text{ visits to } i \text{ after time } t\Big|X(t)=i\Big).\end{aligned}$$

But, because $\{X(t)\}$ is a time-homogeneous chain, the probability it makes at least n visits to i after time t, given $X(t)=i$, is the same as the probability it makes at least n visits to i after $t=0$, given $X(0)=0$. In other words,

$$\mathbb{P}\Big(\text{at least } n \text{ visits to } i \text{ after time } t\Big|X(t)=i\Big)=\mathbb{P}\Big(V_i\geq n\big|X(0)=0\Big).$$

Note that the right-hand side does not depend on t! Therefore, it follows from equation (4.41) that

$$\begin{aligned}\mathbb{P}\big(V_i\geq n+1\big|X(0)=i\big) &= \sum_{t=1}^{\infty}\mathbb{P}\big(V_i\geq n\big|X(0)=i\big)\mathbb{P}(T_i=t\big|X(0)=i)\\ &= \mathbb{P}\big(V_i\geq n\big|X(0)=i\big)\sum_{t=1}^{\infty}\mathbb{P}(T_i=t\big|X(0)=i)\\ &= \mathbb{P}\big(V_i\geq n\big|X(0)=i\big)\mathbb{P}\big(T_i<\infty\big|X(0)=i\big)\\ &= r_i\mathbb{P}\big(V_i\geq n\big|X(0)=i\big).\end{aligned}$$

This says that if we let $a_n:=\mathbb{P}\big(V_i\geq n\big|X(0)=i\big)$, $\{a_n\}$ satisfies the simple difference equation $a_{n+1}=r_ia_n$. Since $a_1=\mathbb{P}\big(V_i\geq 1\big|X(0)=i\big)=r_i$, the solution is $\mathbb{P}\big(V_i\geq n\big|X(0)=i\big)=a_n=r_i^n$, and this is what we wanted to prove.

4.3.2 Chains with absorbing states

In this section we discuss we discuss finite state-space Markov chains with the property that it is possible to hit an absorbing state no matter where the chain starts. The guiding examples are the Moran and Wright-Fisher models without mutation.

In the discussion, $\mathcal{E}$ will denote the (finite) state space of a chain and $\mathcal{S}$ will denote the subset of its absorbing states. As usual, A denotes its state transition matrix. If i is any state and if a is an absorbing state, let

$$g_i^a = \mathbb{P}\Big(X(t)=a \text{ for some } t \geq 0 \Big| X(0)=i\Big).$$

This is the probability the chain eventually gets absorbed by state a given that it starts in state i. For the Wright-Fisher model in a population of size N, the two absorbing states are 0 and $2N$, and the event of getting absorbed into one of them is called *fixation*, because when it occurs, either allele A disappears (absorption into state 0) or takes over (absorption into state $2N$). Then g_i^0 and g_i^{2N} are the probabilities of absorption into these two alternatives starting from i, and are interesting to know from the biological point of view.

If a is an absorbing state, the probability the chain gets absorbed into state a *by time* t, when it starts in i, is simply the probability

$$\mathbb{P}(X(t)=a|X(0)=i) = [A^t]_{ia}.$$

because once the chain arrives at a, it stays there. The probability to be eventually absorbed in a is the limit as $t \to \infty$ of the probability to be absorbed by time t. Thus

$$g_i^a = \lim_{t\to\infty} \mathbb{P}(X(t)=a|X(0)=i) = \lim_{t\to\infty} [A^t]_{ia}.$$

The probability of starting in i and eventually getting to some absorbing state is

$$\mathbb{P}\Big(\text{eventual absorption}\Big|X(0)=i) = \sum_{a\in\mathcal{S}} g_i^a.$$

Theorem 6 *Let $\{X(t)\}$ be a Markov chain in a finite state space $\mathcal{E}$. Assume that the set of absorbing state $\mathcal{S}$ is not empty and assume that for every non-absorbing state i, there exists an absorbing state $a \in \mathcal{S}$, such that a is accessible from i. Let $\mathcal{T} = \mathcal{E} - \mathcal{S}$ be the set of all non-absorbing states.*

Then

(i) Every state $i \in \mathcal{T}$ is transient.

(ii) For every $i \in \mathcal{T}$, $\sum_{a\in\mathcal{S}} g_i^a = 1$. (Absorption occurs with probability one no matter where the chain starts.)

(iii) If $j \in \mathcal{T}$, then $\lim_{t\to\infty} [A^t]_{ij} = 0$, for all $i \in \mathcal{E}$.

(iv) For each a, $\{g_i^a;\, i \in \mathcal{E}\}$ is the unique solution to

$$g_a^a = 1 \quad \textit{and} \quad g_s^a = 0 \quad \textit{if } s \in \mathcal{S} \textit{ and } s \neq a; \tag{4.42}$$

$$g_i^a = \sum_{j \in \mathcal{E}} p_{ij} g_j^a = \sum_{j \in \mathcal{T}} p_{ij} g_j^a + p_{ia} \tag{4.43}$$

The important qualitative result of this theorem is that a chain must eventually enter an absorbing state, if it evolves in a finite state space with only transient and absorbing states. The Wright-Fisher and Moran models without mutation both satisfy the hypotheses of Theorem 6, and thus fixation must ultimately occur in both models. Thus, in the absence of mutation, an allele will eventually either take over or disappear from the population, just due to chance.

The important quantitative result of the theorem is the system of linear equations for absorption probabilities stated in part (iv). They enable us, in principle, to find the probability with which the chain ends up in a particular absorbing state. In fact, the derivation in Section 4.1.4 of the probability of ruin in the gambler's ruin problem is an example of the equations in part (iv).

Statements (i) of the theorem is immediate from the definition of transience. By assumption, if i is not an absorbing state, a chain starting in state i can reach an absorbing state with positive probability, and therefore the probability that it returns to i is less than one. From Theorem 4, we know that the chain can visit a transient state only finite many times, and, since there are only a finite number of transient state—because the state space is finite—the chain starting in any transient state must eventually hit an absorbing state. This implies statement (ii), because $\sum_{a \in \mathcal{S}} g_i^a$ is the probability to hit eventually an absorbing state, starting from i, and it also implies statement (iii), because if the probability of hitting some absorbing state eventually is one, the limiting probability to be in any transient state goes to 0 as $t \to \infty$.

Rather than justify statement (iv) generally, we will explain and use equations (4.42) and (4.43) in an example.

Example 4.3.3. Absorption probabilities for the Moran Model with no mutation when $N = 4$.

The state space for this example is $\mathcal{E} = \{0, 1, 2, 3, 4\}$ and the state transition matrix, which was computed in Example 4.2.6, is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{3}{16} & \frac{10}{16} & \frac{3}{16} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{3}{16} & \frac{10}{16} & \frac{3}{16} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We are interested in computing g_i^N, the probability that the chain eventually hits the absorbing state N, given that $X(0) = i$. We have right away that,

$$g_0^4 = 0 \quad \text{and} \quad g_4^4 = 1, \tag{4.44}$$

because if $X(0) = 0$, the chain stays forever in state O and thus cannot reach 4, and if $X(0) = 4$, the chain is already in state 4 and hence never leaves. These equations are precisely what equation (4.42) says for the case of $a = 4$.

Let V be the event the chain eventually hits N; thus, $g_i^4 = \mathbb{P}(V|X(0) = i)$, by definition. Consider the case, $i = 1$. We will explore what happens if we try to compute g_1^N by conditioning on the value of $X(1)$. Note first that

$$\mathbb{P}(V|X(1) = j, X(0) = 1) = \mathbb{P}(V|X(1) = j) = g_j^4,$$

because, once $X(1) = j$ is given, the Markov property implies $X(0) = 1$ is no longer relevant, and we are looking at the probability of eventually hitting state 4 starting from state j.

Now, if $X(0) = 1$, there are three possibilities for $X(1)$; it is 0 with probability p_{10}, it is 1 with probability p_{11} and it is 2 with probability p_{12}, (note, $p_{13} = p_{14} = 0$). From the previous paragraph, if the chain arrives at 0, the probability of V is then $g_0^4 = 0$; if it arrives at 1 the probability of V is then g_1^4, and so on. Thus, by the rule of total probabilities,

$$g_1^4 = p_{10}g_0^4 + p_{11}g_1^4 + p_{12}g_2^4 = (10/16)g_1^4 + (3/16)g_2^4 \tag{4.45}$$

This is the version, for this example, of equation (4.43) of (iv) in Theorem 6 in the case $i = 1$.

The reader should repeat the same type of reasoning to derive similar equations for $i = 2$ and $i = 3$. They are:

$$g_2^4 = p_{21}g_1^4 + p_{22}g_2^4 + p_{23}g_3^4 = (1/4)g_1^4 + (1/2)g_2^4 + (1/4)g_3^4 \quad \text{and} \tag{4.46}$$

$$g_3^4 = p_{32}g_2^4 + p_{33}g_3^4 + p_{34}g_4^4 = (3/16)g_2^4 + (10/16)g_2^4 + (2/16). \tag{4.47}$$

$$\tag{4.48}$$

Thus, we have derived three, linear equations for the three unknowns g_1^4, g_2^4, and g_3^4. The reader may check that the solution is $g_1^4 = 1/4$, $g_2^4 = 1/2$, and $g_3^4 = 3/4$. $\diamond$

The method of the previous example can always be used and the reader should understand how it is derived. However, there are sometimes simpler methods. In the following example, we again consider the Moran model, but this time for general N, and we derive an expression for the probability of absorption in state N. The derivation will use the fact, proved earlier, that expectation is constant for the Moran chain, and statement (iii) of Theorem 6.

Example 4.3.4 Absorption probabilities for the Moran model without mutation, general N.

We showed in Exercise 4.2.4 that the expected value of the Moran chain does not change with time. In particular,

$$i = E[X(t)|X(0)=i] \quad \text{for all } t \geq 1.$$

Let us write out the expectation in terms of the conditional distribution of $X(t)$. Thus,

$$i = \sum_{j=0}^{N} j\mathbb{P}(X(t)=j|X(0))=i) = \sum_{j=0}^{N} j[A^t]_{ij} = \sum_{j=1}^{N} j[A^t]_{ij}.$$

In the last sum, we omitted the term corresponding to $j = 0$ because it is automatically 0. If $1 \leq j \leq N-1$, then j is transient, and part (iii)of Theorem 6 says that $\lim_{t\to\infty}[A^t]_{ij} = 0$. On the other hand, we know from the discussion at the beginning of this section that $\lim_{t\to\infty}[A^t]_{iN} = g_i^N$. Thus, by taking limits as $t \to \infty$,

$$i = \lim_{t\to\infty} \sum_{j=1}^{N} j[A^t]_{ij} = Ng_i^N.$$

Hence

$$g_i^N = \frac{i}{N}, \quad \text{for all } 0 \leq i \leq N.$$

Since 0 and N are the only absorbing states, and since absorption must take place, $g_i^0 = 1 - g_i^N = 1 - \frac{i}{N}$.

When put in words, we get a very pleasing result: the probability that allele A eventually takes over the population is equal to its initial frequency in the population. ◇

Example 4.3.5. Absorption probabilities for the Wright-Fisher chain with mutation. Consider the Wright-Fisher chain in a population of size $2N$ when there is no mutation. As we showed in the previous section, the expectation of this chain is also constant in time. The same argument we just used for the Moran model shows that

$$g_i^{2N} = \frac{i}{2N} \quad \text{for all } 0 \leq i \leq 2N.$$

The reader should repeat the argument for this case as an exercise. ◇

4.3.3 Ergodic theorem for recurrent Markov Chains

Stationary distributions

Let $\{X(t)\}$ be a stochastic process evolving in the discrete state space $\mathcal{E}$, and let $\rho(t) = \{\rho_i(t);\, i \in \mathcal{E}\}$ denote the distribution of $X(t)$ for each $t \geq 0$; recall that, by

definition, $\rho_i(t) := \mathbb{P}(X(t) = i)$ for every $i \in \mathcal{E}$. If $\rho(t) = \rho(0)$ for all $t \geq 0$, then we say that $\rho(0)$ is a *stationary distribution* for $\{X(t)\}$, because it is the same for all t. Thus, while the process itself is moving around randomly from state to state as time progresses, the probability of $X(t)$ being in any particular state does not change. We we usually denote the stationary distribution by $\pi = \{\pi_i;\, i \in \mathcal{E}\}$.

When $\{X(t)\}$ is a Markov chain evolving in the state space $\mathcal{E} = \{0, 1, \ldots, N\}$ we have been interpreting $\rho(t)$ as a row vector $\rho(t) = (\rho_0(t), \rho_1(t), \ldots, \rho_N(t))$. Likewise, we think of a stationary distribution as row vector $\pi = (\pi_0, \pi_1, \ldots, \pi_N)$, where π_i represents the unchanging probability that $X(t) = i$. Notice that π must a *probability vector*; that is $\pi_{\geq} 0$ for all state i and $\sum_1^n \pi_i = 1$.

Now let A be a state transition matrix. We have the following very important result. A vector π is a stationary distribution for a Markov process with state transition matrix A if and only if

$$\pi = \pi \cdot A \tag{4.49}$$

$$\sum_{i \in \mathcal{E}} \pi_i = 1 \text{ and } \pi_i \geq 0 \text{ for all } i \text{ in } \mathcal{E}. \tag{4.50}$$

The second equation, (4.50), is just the statement that π must be a probability vector. Suppose that it is, and assume π is a stationary distribution; that is, $\rho(t) = \pi$ for all t. Then, by using the identity $\rho(t+1) = \rho(t) \cdot A$ (see Theorem 3), $\pi = \rho(t+1) = \rho(t) \cdot A = \pi \cdot A$, and hence π must be a solution to (4.49). Conversely, if $\pi = \pi \cdot A$, if π is a probability vector and if $\rho(0) = \pi$, then $\rho(1) = \pi \cdot A = \pi$, hence $\rho(2) = \rho(1)A = \pi A = \pi$, and continuing in this way by induction, $\rho(t) = \pi$ for all positive integers t. Thus, π is a stationary distribution.

From this characterization, finding a stationary distribution for a Markov chain is a matter of solving the system of linear equations given by (4.49) and (4.50).

Example 4.3.5. Consider the two state chain with $\lambda + \mu > 0$. Then (4.49) and (4.50) become

$$(\pi_0, \pi_1) = (\pi_0, \pi_1) \cdot \begin{pmatrix} 1-\lambda & \lambda \\ \mu & 1-\mu \end{pmatrix} \quad \text{and} \quad \pi_0 + \pi_1 = 1.$$

This matrix equation is equivalent to the two equations: $\pi_0 = (1-\lambda)\pi_0 + \mu\pi_1$, and $\pi_1 = \lambda\pi_0 + (1-\mu)\pi_1$. Both of these reduce to $\lambda\pi_0 = \mu\pi_1$. The reader may show that the unique solution to $\lambda\pi_0 = \mu\pi_1$ and $\pi_0 + \pi_1 = 1$ is

$$(\pi_0, \pi_1) = \Big(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu}\Big). \qquad \diamond$$

In the study of the two-state chain in Example 4.2.1, we found that, whatever the initial law,

$$\lim_{t \to \infty} \Big(\mathbb{P}(X(t) = 0), \mathbb{P}(X(t) = 1)\Big) = \Big(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu}\Big).$$

This is exactly the unique, stationary distribution we just found for the two-state chain. This is not an accident, but an illustration of an important and general fact. *For any Markov chain, if $\pi = \lim_{t\to\infty} \rho(t)$ exists, then π is a stationary distribution.* Indeed, let A be the state transition matrix of the chain, and assume $\pi = \lim_{t\to\infty} \rho(t)$ exists. Of course, if $t \to \infty$, then $t-1 \to \infty$ also, and so $\pi = \lim_{t\to\infty} \rho(t-1)$. But we know for a Markov chain that $\rho(t) = \rho(t-1) \cdot A$. Thus,

$$\pi = \lim_{t\to\infty} \rho(t) = \lim_{t\to\infty} \rho(t-1) \cdot A = \pi \cdot A,$$

and hence π is stationary.

Periodicity of states

Consider a Markov chain $\{X(t)\}$, that, starting from state i, we can only return to i after an even number of steps, and in fact $\mathbb{P}(X(2t) = i | X(0) = i) > 0$ for all t. In this case, we say that state i has period 2. Periodicity is an important consideration in the study of limits of distributions. If i has period 2, then it really only makes sense to look at $\mathbb{P}(X(t) = i | X(0) = i)$ along the sequence of even times, because $\mathbb{P}(X(t) = i | X(0) = i) = 0$ if t is odd.

More generally, the *period* of i is the *largest* integer d such that the process can return to i in some multiple of d time steps; mathematically, the period is the largest integer d such that

$$\mathbb{P}\Big(T_i \in \{d, 2d, 3d, \ldots\} \Big| X(0) = i\Big) = 1.$$

If the period of a state is $d = 1$, the state is said to be aperiodic.

It is clear that if $p_{ii} = \mathbb{P}(X(1) = i | X(0) = i) > 0$, then state i is aperiodic. Since $p_{ii} > 0$ for every i in the Moran and Wright-Fisher models, with mutation or not, all the states in these chains are aperiodic. In fact, aperiodicity is the norm for most applied models, and will be the only case of real interest to us. An exception is the symmetric random walk. In this case, the chain must move up or down one unit in every time step and so all states have period 2.

The reader should be aware of the facts stated in the following theorem. Its proof is omitted.

Theorem 7 *a) If $i \sim j$, then i and j have the same period.*

b) If there is a positive integer t_0, such that every entry of the matrix A^{t_0} is positive, then all states communicate and are aperiodic.

The Ergodic Theorem

This is the most important theorem in the theory of Markov chains. It generalizes to positive recurrent chains all the nice properties we discovered by explicit computation for two-state, recurrent chains.

Theorem 8 *Let A be the state transition matrix of a Markov chain. Assume all states communicate with one another and are positive recurrent. Then*

(a)) There is admits a unique stationary distribution π for the chain.

(b) If, in addition, the chain is aperiodic, then

$$\lim_{t\to\infty}[A^t]_{ij} = \pi_j, \quad \textit{for all states } i \textit{ and } j; \tag{4.51}$$

Moreover, the stationary distribution has the following two probabilistic interpretations for a Markov chain $\{X(t)\}$ whose state transition matrix is A:

$$\pi_j = \frac{1}{E[T_j|X(0)=j]} \quad \textit{for each state } j, \textit{ and;} \tag{4.52}$$

for any state j and any initial distribution of $X(0)$,

$$\pi_j = \lim_{t\to\infty} \frac{\textit{number of visits of } X(1), X(2), \ldots, X(t) \textit{ to state } j}{t} \tag{4.53}$$

with probability one.

Finally, no matter what the initial law of $X(t)$ is

$$\lim_{t\to\infty} \rho(t) = \pi. \tag{4.54}$$

Remarks.

1. The statement, (4.53), in this theorem is a generalization of the law of large numbers to Markov chains. In words, it says that the long run average number of visits to a state j is equal to the limiting probability, π_j, to be in state j. Statements of this type are called *ergodic theorems*, whence the name given to Theorem 8.

2. It should be emphasized that the statements in equations (4.51), (4.52), and (4.54) require the assumption of aperiodicity. Modified versions of these statements are true in the periodic case, but they are not necessary for our applications.

3. No assumption about the finiteness of the state space is made in Theorem 8. It is true whenever the state space is discrete. When the state space is finite and all states communicate, then, as we have discussed, the chain is positive recurrent, and the theorem applies.

4. When the states of a chain all communicate and are all null recurrent, it can be shown that a stationary distribution does not exist and $\lim_{t\to\infty}[A^t]_{ij} = 0$ for all j.

5. The matrix interpretation of (4.51) is worth noting. Let the state space be $\mathcal{E} = \{0, 1, \ldots, N\}$. Then,

$$\lim_{t\to\infty} A^t = \begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \cdots & \pi_N \\ \pi_0 & \pi_1 & \pi_2 & \cdots & \pi_N \\ \vdots & \vdots & \vdots & & \vdots \\ \pi_0 & \pi_1 & \pi_2 & \cdots & \pi_N \end{pmatrix}. \tag{4.55}$$

This form makes it very clear that $\lim_{t\to\infty}\mathbb{P}\Big(X(t)=j|X(0)=i\Big) = \lim_{t\to\infty}[A^t]_{ij}$ is independent of the starting state, i.

We have demonstrated all the claims of Theorem 8, except for statement (4.53), for the two-state chain, and I recommend that the reader review this analysis. A A proof of Theorem 8 for the general case is beyond the scope of this text, and is omitted. We only note here that (4.54) is a consequence of (4.51). Indeed, from formula (4.24), we know that for a general initial law, $\rho(0)$ of $X(0)$,

$$\mathbb{P}(X(t)=j) = \sum_{i\in\mathcal{E}}\rho(i)[A^t]_{ij}.$$

By taking a limit on both sides and exchanging limit and summation,

$$\begin{aligned}\lim_{t\to\infty}\mathbb{P}(X(t)=j) &= \sum_{i\in\mathcal{E}}\rho_i(0)\lim_{t\to\infty}[A^t]_{ij}\\ &= \sum_{i\in\mathcal{E}}\rho_i(0)\pi_j = \pi_j\sum_{i\in\mathcal{E}}\rho_i(0)\\ &= \pi_j.\end{aligned}$$

(The last step follows because $\rho(0)$ is a probability distribution on $\mathcal{E}$, and hence $\sum_{i\in\mathcal{E}}\rho(i)=1$.) This gives (4.54).

Example 4.3.6. Consider the Moran model with mutation as defined in Example 4.6. Assume that $N=3$ and $u=v=1/4$. Then the probability transition matrix is

$$A = \begin{pmatrix} 3/4 & 1/4 & 0 & 0\\ 7/36 & 19/36 & 10/36 & 0\\ 0 & 10/36 & 19/36 & 7/36\\ 0 & 0 & 1/4 & 3/4\end{pmatrix}.$$

The state space is finite and all state communicate, and hence the chain is positive recurrent and Theorem 8 applies. The probability to be in state j will tend in the limit as $t\to\infty$ to π_j, where $\pi=(\pi_0,\pi_1,\pi_2,\pi_3)$ is the unique solution to

$$(\pi_0,\pi_1,\pi_2,\pi_3) = (\pi_0,\pi_1,\pi_2,\pi_3)\begin{pmatrix} 3/4 & 1/4 & 0 & 0\\ 7/36 & 19/36 & 10/36 & 0\\ 0 & 10/36 & 19/36 & 7/36\\ 0 & 0 & 1/4 & 3/4\end{pmatrix},$$

and $\pi_0+\pi_1+\pi_2+\pi_3=1$. The first system of equation is equivalent to

$$\begin{aligned}\pi_0 &= \frac{3}{4}\pi_0 + \frac{7}{36}\pi_1\\ \pi_1 &= \frac{1}{4}\pi_0 + \frac{19}{36}\pi_1 + \frac{10}{36}\pi_2\\ \pi_2 &= \frac{10}{36}\pi_1 + \frac{19}{36}\pi_2 + \frac{1}{4}\pi_3\\ \pi_3 &= \frac{7}{36}\pi_2 + \frac{3}{4}\pi_3\end{aligned}$$

Use these equations in succession to solve for π_1, π_2, π_3 in terms of π_0. The result is $\pi = \pi_0(1, \frac{9}{7}, \frac{9}{7}, 1)$. Since the terms of π must sum to one, $1 = \pi_0[1 + (9/7) + (9/7) + 1] = (32/7)\pi_0$. Thus $\pi = (7/32, 9/32, 9/32, 7/32)$. Theorem 8 give us additional information. For example, the average time it takes the chain to return to state 0 given that it starts there is

$$E[T_0|X(0)=0] = \frac{1}{\pi_0} = \frac{32}{7}. \qquad \diamond$$

Detailed balance equations and stationary distributions

Let A be a state transition matrix for a chain on $\{0, 1, \ldots, N\}$. We say that $\pi = (\pi(0), \pi(1), \ldots, \pi(N))$ satisfies the detailed balance equations if

$$\pi(i)p_{ij} = \pi(j)p_{ji}, \quad \text{for all } i \text{ and } j.$$

If π satisfies the detailed balance equations and if $\pi(i) \geq 0$ for each i and $\sum_{i=0}^{N} \pi(i) = 1$, then π is a stationary distribution for A. This is easy to see because if π satisfies the detailed balance equations, then by summing both sides of the detailed balance equations in the index j and using the fact that $\sum_{j=0}^{N} p_{ij} = 1$,

$$\pi(i) = \sum_{j=0}^{N} \pi(i)p_{ij} = \sum_{j=0}^{N} \pi(j)p_{ji}.$$

for each i, and these are the equations defining a stationary distribution.

If you are trying to calculate a stationary distribution, it is a good idea to first look for a solution to the detailed balance equations. When such a solution exists—and it does for many applied models—it is usually easy to find. However, failure to find a solution of the detailed balance equations does not mean there is no stationary distribution. A stationary distribution does not necessarily satisfy the detailed balance equations.

One class of examples for which detailed balance equations have a solutions is birth-and-death chains. This class includes the Moran model with mutation. Recall from Example 4.1.7 that a finite birth-and-death chain has a transition matrix of the form

$$A = \begin{pmatrix} 1-p_0 & p_0 & 0 & 0 & \cdots & \cdots & 0 \\ q_1 & r_1 & p_1 & 0 & \cdots & \cdots & 0 \\ 0 & q_2 & r_2 & p_2 & \cdots & \cdots & 0 \\ \vdots & & & & & & \vdots \\ \vdots & & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & q_{N-1} & r_{N-1} & p_{N-1} \\ 0 & \cdots & \cdots & \cdots & 0 & q_N & 1-q_N \end{pmatrix}.$$

Thus, for $1 \leq i \leq N-1$, $p_{i,i-1} = q_i$, $p_{ii} = r_i$, $p_{i,i+1} = p_i$. This transition matrix defines an irreducible chain if $p_i > 0$ and $q_i > 0$ for all i. It is aperiodic if at least one

of the probabilities, $1-p_0, r-1, \ldots, r_{N-1}, 1-q_N$, on the diagonal of A is positive. Assume these conditions are satisfied.

The detailed balance equations for this transition matrix reduce to

$$\pi(i)p_i = \pi(i+1)q_{i+1}, \quad 0 \le i \le n-1.$$

This means for each i that $\pi(i+1) = \dfrac{p_i}{q_{i+1}}\pi(i)$ for each i, $0 \le i < N$. By iterating,

$$\pi(i+1) = \frac{p_i p_{i-1} \cdots p_0}{q_{i+1} q_i \cdots q_1}\pi(0) = \pi(0)\prod_{k=0}^{i}\frac{p_k}{q_{k+1}}.$$

For π to be a stationary distribution its components must sum to 1:

$$1 = \sum_{i=0}^{N}\pi(i) = \pi(0)\Big[1 + \sum_{1}^{N}\prod_{k=0}^{i-1}\frac{p_k}{q_{k+1}}\Big].$$

Therefore

$$\pi(0) = \Big[1 + \sum_{1}^{N}\prod_{k=0}^{i-1}\frac{p_k}{q_{k+1}}\Big]^{-1}.$$

and for $i \ge 1$,

$$\pi(i) = \Big[1 + \sum_{1}^{N}\prod_{k=0}^{i-1}\frac{p_k}{q_{k+1}}\Big]^{-1}\prod_{k=0}^{i-1}\frac{p_k}{q_{k+1}}.$$

4.3.4 Exercises

Exercise 4.3.1. Consider the Wright-Fisher chain with no mutation and a population of size N. Show that the probability that $X(t)$ eventually equals $2N$, given $X(0) = i$, is $i/2N$.

Exercise 4.3.2. Consider the Markov chain on the state space $\{0, 1, 2, 3\}$ given by

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 7/36 & 19/36 & 10/36 & 0 \\ 0 & 10/36 & 19/36 & 7/36 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Let g be defined by $g(0) = 27$, $g(1) = 17$, $g(2) = 10$, $g(3) = 0$.

(a) Show that $E[g(X(t)) \,\big|\, X(0)=i] = g(i)$ for all i, $0 \le i \le t$, and $t \ge 1$.

(b) Find the probability that, starting from $i = 1$, the Markov chain gets absorbed in state 3.

Exercise 4.3.3 Consider the Moran model with mutation as defined in Example 4.6. Assume that $N = 3$ and $u = v = 1/4$. Then the probability transition matrix is

$$A = \begin{pmatrix} 3/4 & 1/4 & 0 & 0 \\ 7/36 & 19/36 & 10/36 & 0 \\ 0 & 10/36 & 19/36 & 7/36 \\ 0 & 0 & 1/4 & 3/4 \end{pmatrix}.$$

Let $\mathbf{g}$ defined by $g(i) = i$ for $0 \leq i \leq 3$.

(a) Use equation (4.29) in the text to compute $E[g(X(1)) \mid X(0)=i]$ for $0 \leq i \leq 3$. Represent your answer as a vector.

(b) Compute $E[g(X(2)) \mid X(0)=i]$ for $0 \leq i \leq 3$. (Hint; because of equation (4.30) in the text, you only need to multiply your answer to (a) by A. Why?)

Exercise 4.3.4. 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

Assume that the initial distribution is $\rho(0) = (1/3, 1/4, 5/12)$ and find $\rho(\infty) = \lim_{t\to\infty} \rho(t)$. If the initial distribution is different will the limit exist and be the same, or might it be different?

Exercise 4.3.5. Consider the Moran model with $N = 4$ in which A mutates to a with probability $1/3$ and a mutates to A with probability probability $1/3$ also. Find the stationary distribution.

Exercise 4.3.6. (a) Use the method of detailed balance equations to re-derive the invariant density for the chain of Example 4.3.6.

Answer the following questions. For large times, what approximately is the probability that $X(t) = 1$?

If the chain is in state 1, how long on average does it take to return to state 1?

(b) Write down the Moran model for $N = 4$ with mutation rates of $1/4$ for A mutating to a and of $1/2$ for a mutating to A. Find the stationary distribution of this chain.

Exercise 4.3.7. This problem requires you use numerical software: Maple or Matlab will work fine. Let A be the transition matrix of the Moran model in problem 4.3.5. How many time steps t are needed so that all entries of A^t are within .05 of their limiting values as $t \to \infty$? Display a few computed values of A^t to justify your answer.