Chapter 4

Discrete-time Markov Chains and Applications to Population Genetics

Quantities that vary randomly over time, or space, or both are called *stochastic processes*. Chapter 4 is an introduction to a class of stochastic processes known as *discrete-time Markov chains*, which are used to model randomness in just about every discipline of science and engineering. In biology, the study of genetics, sequence analysis, evolutionary biology, ecology, cell dynamics, and cancer all employ Markov chain models. Chapter 4 discusses the theory of Markov chains in enough generality to prepare the reader to explore its applications in any of these areas. But, in continuation of the theme of Chapter 3, the focus will be on applications to population genetics.

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 described by 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 in which the proportion of A's is $f_A = 0.9$. 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 geneticists 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 allele or the other must eventually disappear. This is completely different from what happens in the infinite population model, for which allele frequency is constant from generation to generation. In small populations, random fluctuations are significant enough that genotype and allele frequencies must be treated as stochastic processes.

Chapter 4 makes deeper use of probability than Chapter 3; conditional probability and conditional expectation are particularly important, both conceptually and computationally. The reader may consult Sections 2.1.5 and 2.3.6 of Chapter 2 for the relevant background. Elementary linear algebra will also be used extensively.

4.1 Markov Chains: Introduction

4.1.1 Discrete-time stochastic process models.

A *discrete-time stochastic processes* is a sequence of random variables, $\{X(0), X(1), \ldots, X(t), \ldots\}$, indexed by a discrete parameter, t, $t = 0, 1, 2, \ldots$. The order of the random variables in the process is important, and so we will think of t generically as a time parameter and of $X(t)$ as the value of the process at time t. In many applications, t does measure the passing of time, but not always. Also, it is not necessary that the first time be $t=0$ or that time increase by unit steps, as here: these are just conventions used when discussing stochastic processes generally. Finally, the notation $\{X(t);\, t \geq 0\}$, or, simply, $\{X(t)\}$, will be used to denote a process in its entirety.

In practice, $X(t)$ will model the 'state' of some system at time t. The set of possible states—the set of possible values of $X(t)$—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, then 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*. In this chapter, the state space will always be discrete.

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

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

as a function of $(x_0, x_1, \ldots, x_t)$, is the joint probability mass function of $X(0), \ldots, X(t)$. It is often helpful to think of (4.1) as the probability that the process follows the *path*, $\big\{X(0)=x_0, X(1)=x_1, \ldots, X(t)=x_t\big\}$, up to time t.

A *model* of a stochastic process is a set of assumptions that, at least in principle, specifies the joint probability mass function of $X(0), \ldots, X(t)$ for any $t \geq 0$. In effect, a model is a complete statistical description of a stochastic process. A model

is allowed to contain unspecified parameters, in which case it is sometimes called a *parametric model*.

To illustrate the terminology, let $\{X(0), X(1), \dots\}$ 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

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

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

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

where $\hat{n} = \sum_0^t x_i$ is the number of 1's in the sequence $(x_0, \dots, x_t)$.

How does one turn assumptions about the physical mechanisms driving a physical stochastic process into a model? When the process evolves in time, conditional probabilities are a natural tool. Think of time t as the 'present'. The conditional probability distribution,

$$\mathbb{P}\Big(X(t+1) = x \;\Big|\; (X(0), \dots, X(t)) = (x_0, \dots, x_t)\Big), \qquad x \in \mathcal{E},$$

describes how the full history, $\{X(0) = x_0, \dots, X(t) = x_t\}$, of the process up to time t affects the probability distribution of its next value, $X(t+1)$. It is called a *one-step-ahead, conditional probability*. As we shall see, any joint distribution, $\mathbb{P}\Big(X(0)=x_0, X(1)=x_1, \dots, X(t)=x_t\Big)$, can be computed from the probability distribution of $X(0)$ and one-step-ahead conditional probabilities. *Therefore, modeling a stochastic process can be reduced to prescribing its one-step-ahead conditional probabilities.* (Usually, the initial law is incidental and can be treated as a model parameter.) This is very useful in practice, because hypotheses about how a process works physically often translate easily into formulas for its one-step-ahead conditional probabilities.

To see why one-step-ahead probabilities and the distribution of $X(0)$ determine a model, consider computing $\mathbb{P}(X(0) = x_0, X(1) = x_1, X(2) = x_2)$. The argument for the joint distribution of any number of variables will be clear from this case. The only fact used is the identity, $\mathbb{P}(A \cap B) = \mathbb{P}(A\big|B)\mathbb{P}(B)$, which comes from the definition of conditional probability. If this is applied with $A = \{X(2) = x_1\}$ and $B = \{X(0)=x_0, X(1)=x_1\}$, the result is

$$\begin{aligned}
&\mathbb{P}\Big(X(0)=x_0, X(1)=x_1, X(2)=x_2\Big) \\
&\quad = \quad \mathbb{P}\Big(X(2)=x_2\big|X(0)=x_0, X(1)=x_1\Big)\mathbb{P}\Big(X(0)=x_0, X(1)=x_1\Big)\,.
\end{aligned}$$

The same reasoning shows

$$\mathbb{P}\Big(X(0)=x_0, X(1)=x_1\Big) = \mathbb{P}\Big(X(1)=x_1\Big|X(0)=x_0\Big)\mathbb{P}\Big(X(0)=x_0\Big).$$

Putting the two formulas together,

$$\begin{aligned}&\mathbb{P}(X(0)=x_0, X(1)=x_1, X(2)=x_2)\\ &\quad = \mathbb{P}\Big(X(0)=x_0 \;\Big|\; X(1)=x_1, X(2)=x_2\Big)\mathbb{P}\Big(X(1)=x_1\Big|X(0)=x_0\Big)\mathbb{P}\Big(X(0)=x_0\Big).\end{aligned}$$

In the next section, we will write down the extension of this formula to the joint probability of $(X(0),\ldots,X(t))$ for any t, when $\{X(t)\}$ is a Markov chain—see (4.4).

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.

The one-step-ahead, conditional probabilities of $\{X(t)\}$ 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,

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

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

For each t, the one-step-ahead conditional probability in Example 4.1.1 depended on the past only through the value of the process at the time t. Stochastic processes possessing this general property are called Markov chains, in honor of the Russian mathematician, A. Markov(1856-1922).

Definition 1 *A stochastic process* $\{X(t);\, t \geq 0\}$ *taking values in a state space* $\mathcal{E}$ *is called a* Markov chain *if*

$$P\Big(X(t+1) = x_{t+1} \;\Big|\; X(0)=x_0, \ldots, X(t)=x_t\Big) = P\Big(X(t+1) = x_{t+1} \;\Big|\; X(t)=x_t\Big) \tag{4.2}$$

for every $t \geq 0$ *and for all sequences of states,* $x_0, \ldots, x_t$*, whenever both conditional probabilities are well-defined, (that is, whenever* $\mathbb{P}\Big(X(0)=x_0, \ldots, X(t)=x_t\Big) > 0$*.)*

We call (4.2) the *Markov condition*. The whole theory of Markov chains flows out of this innocent-looking assumption.

By assumption in this chapter, the time parameter and state space are discrete, and so Definition 1 really only defines *discrete-time* Markov chains. The notion of Markov chain extends also to processes with a continuous-time parameter and more general state spaces.

If X is a Markov chain, $P\Big(X(t+1) = j \;\Big|\; X(t)=i\Big)$ 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 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 4.1.1 above is a simple example of a Markov chain because at each moment, independent of the past, the process changes state at the next moment with probability 0.25 and stays put with probability 0.75. Thus it satisfies the Markov condition, 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 its present state to the opposite state if and only if the coin toss results in heads, which has probability 0.25 independent of time. This example has an important generalization.

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,

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

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

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

Physical realizations of two-state chains can be generated by a number of mechanisms, for instance, a variant of the coin flipping procedure of Example 4.1.1. Imagine again that states 0 and 1 represent two stations, and $X(t)$ is the station you occupy at time t. But now suppose there is a coin at each station, and, when flipped, the coin at station 0 comes up heads with probability λ, the coin at station 1 with probability μ. At each time t, you flip the coin at your current station and move to the opposite station if the flip is heads and stay put if it's tails. Assume all coin flips are independent. Then, the probability of switching states or staying put from t to $t+1$ depends only your present station, $X(t)$, and not on the sequence of states visited prior to t, and you will move from 0 to 1 with probability λ and from 1 to 0 with probability μ. Thus $X(t)$ is a Markov chain with the prescribed transition probabilities.

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

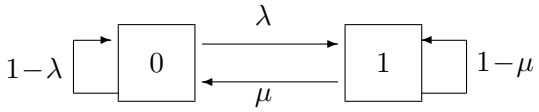


Figure 4.1: Two state Markov chain

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, $\mathcal{E}$, is finite, and an order $\{s_1, \ldots, s_N\}$ of its states is fixed, the state transition matrix (with respect to this order) is

$$A = \begin{pmatrix} p_{s_1 s_1} & p_{s_1 s_2} & \cdots & p_{s_1 s_N} \\ p_{s_1 s_1} & p_{s_2 s_2} & \cdots & p_{s_2 s_N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{s_N s_1} & p_{s_N s_2} & \cdots & p_{s_N s_N} \end{pmatrix}. \tag{4.3}$$

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 will use A instead.) 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 (4.3) 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,

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

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.

Let $\{X(t);\, t \geq 0\}$ be a time-homogeneous Markov chain with transition probabilities $\{p_{ij};\, i,j \in \mathcal{E}\}$. We showed in the last section how to compute joint distributions of a process from one-step-ahead, conditional probabilities and the distribution of $X(0)$. Applying this this argument to Markov chains leads to the following, very important formula. For any $t \geq 0$ and any path $\{x_0, \ldots, x_t\}$,

$$\mathbb{P}\Big(X(0)=x_0, X(1)=x_1, \ldots, X(t)=x_t\Big) = \mathbb{P}(X(0)=x_0) p_{x_0x_1} p_{x_1x_2} \cdots p_{x_{t-1}x_t} \quad (4.4)$$

or, equivalently,

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

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 (4.4) 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.

Formula (4.4) shows that a state transition matrix and an initial law completely specify all joint probabilities of the process and so are all that are needed to define a Markov chain model. In practice the initial law is often incidental and one one regards a state transition matrix as a model that allows arbitrary initial laws.

Equation (4.4) is derived by iterated conditioning. For notational 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 $\{X(0) = x_0, \ldots, X(t-1) = x_{t-1}\}$, and then applying the Markov condition,

$$\begin{aligned} P_t(x_0, \ldots, x_{t-1}, x_t) &= \mathbb{P}\Big(X(0){=}x_0, X(1){=}x_1, \ldots, X(t{-}1){=}x_{t-1}\Big) \\ &\qquad \times \mathbb{P}\Big(X(t){=}x_t \;\Big|\; X(0){=}x_0, \ldots, X(t{-}1){=}x_{t-1}\Big) \\ &= P_{t-1}(x_0, \ldots, , x_{t-1}) p_{x_{t-1}x_t} \end{aligned} \tag{4.6}$$

Now repeat this same calculation, but with t–1 in place of t: then $P_{t-1}(x_0, \ldots, x_{t-1}, x_{t-1}) = P_{t-2}(x_0, \ldots, x_{t-2}) p_{x_{t-2}x_{t-1}}$. Substituting back into (4.6),

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

By continuing this procedure backwards in time, one arrives finally at (4.4).

Equation (4.5) is just the same as (4.4) if both sides of (4.4) are divided by $\mathbb{P}(X(0){=}x_0)$. Conversely if (4.4) is true,

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

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

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

Formula (4.4) and (4.5) are true starting from any time positive time s:

$$\mathbb{P}\Big(X(s){=}x_0, X(s{+}1){=}x_1, \ldots, X(s{+}t){=}x_t\Big) = \mathbb{P}(X(s){=}x_0) p_{x_0x_1} p_{x_1x_2} \cdots p_{x_{t-1}x_t}\,; \tag{4.7}$$

$$\mathbb{P}\Big(X(s){=}x_1, \ldots, X(s{+}t){=}x_t\Big|X(s){=}x_0\Big) = p_{x_0x_1} p_{x_1x_2} \cdots p_{x_{t-1}x_t}\,. \tag{4.8}$$

This is proved using (4.4) itself; the derivation is left to Exercise 4.1.12. As a consequence of this formula, if $Y(t) := X(s{+}t)$, $t \geq 0$, then $\{Y(t)\}$ is a Markov chain with the same transition probabilities as $\{X(t)\}$.

4.1.3 Markov Chain Models.

This section describes several basic Markov chain models, including two important for population genetics, the Wright-Fisher and the Moran models. Examples 4.1.8 and 4.1.9 may be skipped on a first reading and returned to when needed. Also, it is possible to start reading the next section at this point.

Example 4.1.4. Simple random walk. Random walks are the most basic examples of Markov chains, and they have wide application. Simple random walk is a Markov chain that moves from integer to integer by random unit steps. The state transition diagram shows immediately how it evolves:

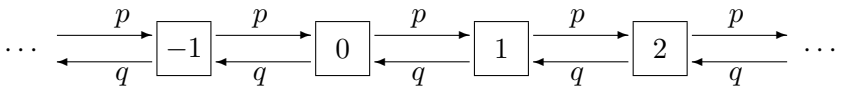


Figure 4.2: Simple random walk on the integers.

Here p is a fixed parameter, $0 < p < 1$, and $q = 1-p$. Thus, at each step the chain takes a step one integer to the right with probability p and a step one integer to the left with probability q. We can read off the transition probabilities directly from the diagram:

$$p_{ij} = \begin{cases} p, & \text{if } j = i+1; \\ q, & \text{if } j = i-1; \\ 0, & \text{if } |j-i| > 1. \end{cases} \tag{4.9}$$

Notice that they are the same for all states i. When $p = q = 1/2$, this chain is called *symmetric simple random walk.*

There is an easy way to construct a simple random walk. Let $\xi_1, \xi_2, \ldots$ be independent, identically distributed random variables, with

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

Let $X(0)$ be a random, integer-valued, initial position independent of $\xi_1, \xi_2, \ldots$, and define

$$X(t) = X(0) + \sum_{i=1}^{t} \xi_i, \quad \text{for all } t \geq 0. \tag{4.10}$$

We claim $\{X(t)\}$ is a simple random walk. Certainly, $X(t)$ is integer-valued for each t, because $X(0)$ and $\xi_1, \xi_2, \ldots$ are all integer-valued. Now,

$$X(t+1) = \big[X(0) + \sum_{i=1}^{t} \xi_i\big] + \xi_{t+1} = X(t) + \xi_{t+1}, \quad \text{for each } t \geq 0,$$

so ξ_{t+1} is just size of the step taken moving from $X(t)$ to $X(t+1)$. Since each step ξ_{t+1} equals one with probability p and -1 with probability q, we get a simple random walk. It is worthwhile proving in detail that $\{X(t)\}$ defined in this way satisfies the Markov property. The crucial point of the argument is that ξ_{t+1} is independent of $X(0), \ldots, X(t)$. This is true because, by construction, $X(0), \ldots, X(t)$ are completely determined by $X(0), \xi_1, \ldots, \xi_t$, which, by assumption, are independent of ξ_{t+1}. Using this independence,

$$\begin{aligned}
&\mathbb{P}\Big(X(t+1) = i+1 \Big| X(t)+i, \ldots X(0) = x_0\Big) \\
&\quad = \mathbb{P}\Big(X(t) + \xi_{t+1} = i+1 \Big| X(t)+i, \ldots X(0) = x_0\Big) \\
&\quad = \mathbb{P}(\xi_{t+1} = 1 \Big| X(t)+i, \ldots X(0) = x_0\Big) = \mathbb{P}\Big(\xi_{t+1} = 1\Big) = p\,.
\end{aligned}$$

Since a step to the left is the only other possibility,

$$\mathbb{P}\Big(X(t+1) = i-1 \Big| X(t) = i, \ldots X(0) = x_0\Big) = \mathbb{P}\Big(\xi_{t+1} = 1\Big) = q.$$

These identities show that $\{X(t)\}$ is a Markov chain with the correct transition probabilities.

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. ◇

Example 4.1.5. Random walk with absorbing boundaries and stem cell dynamics.

The cells of the intestinal epithelium (lining) are replenished from stem cells which reside at the bottom of small sacs, called crypts, in the intestine. These stem cells are always replacing themselves by division, and there is evidence that

the population in each crypt evolves stochastically; some lineages die away and others persist, essentially at random. The dynamics of this process are important to understanding the development of colorectal cancer, which can arise ultimately from mutated stem cells, whose daughter cells become tumor cells in the intestine.

There is a simple model using random walk for the evolution of mutant stem cell numbers in a crypt. It appears in the article, *Defining Stem Cell Dynamics in Models of Intestinal Tumor Initiation*, by L. Vermeulen, et al, which appeared in *Science* **342**, 995 (2013); see also an accompanying review of the article, entitled *Unwanted Evolution* by I. Bosic and M.A. Nowak, on page 932 the same issue.

The model assumes that the number of stem cells at the bottom of the crypt, denoted by N, remains constant in time, For the proximal small intestine, Vermeulen et al, find experimentally that $N = 5$ is most likely. They suppose, with good evidence, that the stem cells are arranged in a ring and imagine that if you watch a pair of neighboring cells in this ring, then, after a random time, one will replace another, that is, one will die out and be replaced by a copy of its neighbor. Call this a replacement event. Imagine that at some point a mutation arises in one of the stem cells of a crypt, and, since mutation is rare, that no further mutations occur in that crypt. The number of mutant stem cells can then grow or shrink, because at each replacement event between neighboring mutant and wild types, the mutant could replace the wild type or vice versa. Vermeulen, et al, propose a continuous-time Markov model for the timing and direction of replacement events. To simplify, we shall consider instead the process $X(t)$ which counts the number of mutant types in the crypt at the t^{th} replacement event. As a consequence of their model, $\{X(t)\}$ behaves like a simple random walk until it hits one of the 'boundary states', state 0 or state N. If it hits 0, the mutant type has died out, and assuming no further mutations, the process remains in state 0 thereafter. Likewise, if it hits N, the wild type population has disappeared and the process stays in state N. The process $\{X(t)\}$ is called a random walk on $\{0, 1, 2, \ldots, N\}$ with *absorbing boundaries*. Its state transition diagram is

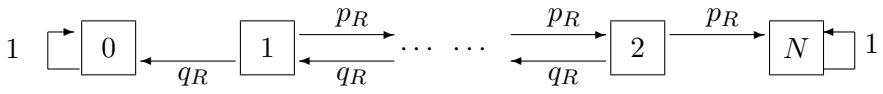


Figure 4.3: Simple random walk on a finite state space

The model depends on a single parameter, p_R, which is the probability a mutant type replaces a wild type in a replacement event. It seems simplistic, but is very useful. Vermeulen, et al, are able to confirm its approximate validity in experiments on mice and they estimate p_R for a type of mutant stem cell. They write: "interpreting the altered size distribution of clones [that is the clones of a mutant type] in terms of individual stem cell fates provides an intuitive estimate of the potency of

oncogenic mutations in the context of a stochastic model, by integrating the effects of proliferation rate, cell death frequencies, and self-renewal properties into a single parameter (p_R)."

Prior to the paper by Vermeulen, et al, Isawa, Michor, Nowak and others, had hypothesized that tissue architecture could help suppress development of mutated lineages. The random walk model supports this idea. Even if mutant types are favored ($p_R > 1/2$) there is a high probability that they will disappear when N is small, simply by chance. Experiments confirm this is the case. That stem cells reside in crypts with small cell numbers and the fact that one lineage can replace another by chance, helps retard the fixation of mutants. ◇

For the Markov chain of Figure 4.3, once the chain reaches state 0 or state N, it stays there for all future time, because the transition probability out of each of these states is zero. For this reason 0 and N are called *absorbing* states. More generally, a state i of a Markov chain is said to be *absorbing* if $p_{ii} = 1$—and hence $p_{ij} = 0$ if $j \neq i$. We shall encounter absorbing states in other models and learn how to calculate the probability that a chain eventually lands in a specific absorbing state. For the mutant stem cell model, calculating the probability of getting absorbed in state 0 is more commonly known as the *gambler's ruin* problem. In this case the chain is interpreted as the fortune of a gambler who wins a dollar or loses a dollar on each bet and plays until either beating the house (reaching state N) or going broke (reaching state 0). The solution is given in Example 4.4.1.

Example 4.1.6. 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. (It is not important how many other alleles there are at the locus of A.)

Compare these hypotheses to the infinite population model defined in Section 3.3.1. The only differences here are that the population is finite in size and that random matings are explicitly independent. The Wright-Fisher process is therefore the finite population version of the basic neutral model.

Assumptions (i)—(iv) translate directly into a Markov chain model for $\{X(t)\}$. Let N denote the (constant) size of the diploid population. The size of the allele

population is then $2N$, and so the state space for $\{X(t)\}$ is $\{0, 1, \ldots, 2N\}$. Let t be a non-negative integer, and fix a past history, ending in generation t with $X(t) = i$. By assumptions (i) and (iii), generation $t+1$ is created by N, independent, random matings of generation t. As we learned in Chapter 3, the genotype produced by each 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 of the allele pool of generation t. Given $X(t) = i$—meaning there are i type A alleles in an allelet pool of size $2N$ in generation t—the probability of drawing allele i is $\frac{i}{2N}$. Therefore $X(t+1)$ is a binomial random variable with parameters $n = 2N$ and $p = \frac{i}{2N}$. This is true no matter what the history of the process was strictly before generation t. It follows that $\{X(t)\}$ is a Markov chain with transition probabilities,

$$p_{ij} = \mathbb{P}\Big(X(t+1) = j \Big| X(t) = i\Big) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}, \quad 0 \le i, j \le 2N\,. \tag{4.11}$$

We call the model with these transition probabilities a neutral Wright-Fisher model for a population of size $2N$. More generally, a Markov chain evolving in the state space $\{0, 1, \ldots, M\}$ with transition probabilities,

$$p_{ij} = \binom{M}{j} \left(\frac{i}{M}\right)^j \left(1 - \frac{i}{M}\right)^{M-j}, \quad 0 \le i, j \le M\,, \tag{4.12}$$

is called a neutral Wright-Fisher chain for an (allele) population of size M, whether M is even or not; 'neutral' here means 'without selection or mutation.' Equivalently, a Markov chain evolving in the state space $\{0, 1, \ldots, M\}$ is a neutral Wright-Fisher chain, if, conditional on $X(t) = i$, $X(t+1)$ is a binomial random variable with parameters $n = M$ and $p = i/M$, for all t. We shall often use this characerization.

Although we motivated the Wright-Fisher model for an allele population of size $M = 2N$, in a population of N diploid individuals reproducing sexually, it applies also to haploid populations for any M, even or not. Imagine M individuals classified into two types, A and 'not A', and let $X(t)$ be the number of type A's in generation t. If generation $t+1$ is produced by M independent samples of generation t, the result is the Wright-Fisher chain with transition probabilities as in (4.12), because, conditional on $X(t) = i$, the probability of choosing a type A individual is i/M and and so the number $X(t+1)$ of type A's in the M samples is a binomial random variable with parameters $n = M$ and $p = i/M$.

Notice that the boundary states 0 and M in the neutral Wright-Fisher model are both absorbing. A population in state 0 has no type A's, and, without mutation, none can appear in future generations. Thus $p_{00} = 1$ and $p_{0j} = 0$ for all $1 \le j \le M$. Likewise, $p_{M,M} = 1$, because once the chain enters state M, there only type A's. When $0 < i < M$, (4.12) 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 M are the only absorbing states.

Remark: Wright-Fisher and the basic infinite population model. As noted, the Wright-Fisher model is the finite population version of the neutral infinite population model presented in Section 3.2.1. In fact, the infinite population model can be regarded as the limit of the Wright-Fisher model obtained as the population size, M, tends to infinity. This is a consequence of the law of large numbers for independent samples. Fix an initial frequency, $f_A(0)$ of A, and for each positive integer M let $\{X^{(M)}(t)\}$ denote the Wright-Fisher model for a population of size M starting with initial frequency $X^{(M)}(0)/M = f_A(0)$. Define the infinite population limiting frequency of A in generation $t = 1$ by

$$f_A(1) = \lim_{M\to\infty} \frac{X^{(M)}(1)}{M}\,.$$

By the definition of the Wright-Fisher chain, $X^{(M)}(1)/M$ is the frequency of allele A obtained after M independent random samples of generation 0. By the law of large numbers (see Section 2.1.5, Theorem 2.1), its limit as $M \to \infty$ is the frequency of A in generation 0, which is $f_A(0)$. Thus, $f_A(1) = f_A(0)$, and we recover the basic infinite population model in which allele frequency is constant. $\diamond$

Example 4.1.7. 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. Also, 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. Because one individual is added and one removed at each stage, the population remains constant in size. Also, 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 show 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, irrespective of all other past history. 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 $\frac{i}{N}\frac{(N-i)}{N}$ 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.

These calculations are unaffected by any history of the process previous to t—only $X(t) = i$ is relevant. Thus, X satisfies the Markov condition.

As an example, the reader should verify that the state transition matrix of the Moran model when $N = 4$ is:

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

$\diamond$

Example 4.1.8. Discrete-time 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 on a subset of integers which makes transitions of at most unit size; the transition probabilities are allowed to depend on the present state i. 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

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

Example 4.1.9. Randomly perturbed difference equations.

Consider the difference equation with random inputs:

$$X(t{+}1) = f(X(t), \xi(t{+}1)) \,. \tag{4.13}$$

Here f is a given function and $\{\xi(t);\, t \geq 1\}$ is a sequence of random inputs. For any given initial value, $X(0)$, which may also be random, there is a unique solution determined by $X(0)$ and the random inputs. Indeed, by recursively applying the equation: $X(1) = f(X(0), \xi(1))$, $X(2) = f(X(1), \xi(2))$, $X(3) = f(X(2), \xi(3))$, and so on. Clearly, for each t, $X(t)$ is is completely determined by $X(0)$ and the first t random inputs, $\xi(1), \ldots, \xi(t)$.

Equation (4.13) and higher order versions of it appear often in applications as a consequence of adding exogenous stochastic fluctuations—the random inputs—to a difference equation model. Random walk, for which $X(t{+}1) = X(t) + \xi(t{+}1)$, is an example. Another example, prominent in time series analysis, is the autoregressive process of order 1, defined by

$$X(t{+}1) = \alpha X(t) + \beta + \xi(t{+}1) \,. \tag{4.14}$$

Assume that the solution to (4.13) evolves in a discrete state space, in keeping with the framework of this chapter. Suppose also:

(i) $\{\xi(t);\, t \geq 1\}$ is a sequence of independent, identically distributed random variables; and

(ii) $X(0)$ is independent of $\{\xi(t);\, t \geq 1\}$.

Under these assumptions, the solution, $\{X(t)\}$, to (4.13) will be a Markov chain with transition probabilities,

$$p_{ij} = \mathbb{P}\big((f(i,\xi(t))=j\big)\,.$$

The right-hand side does not depend on t because it is assumed in (i) that $\xi(t)$ has the same probability distribution for each t. (The definition of Markov chain can easily be extended to more general state spaces and it will still be true that conditions (i) and (ii) imply the solution to (4.13) is Markov.)

To derive this transition probability, observe first that, since $X(0), X(1), \ldots, X(t)$ are determined completely by $X(0), \xi(1), \ldots, \xi(t)$, which by assumptions (i) and (ii) are independent of $\xi(t{+}1)$, all of $X(0), X(1), \ldots, X(t)$ are also independent of $\xi(t{+}1)$. Now if $X(t){=}i$, then $X(t{+}1) = f(i, \xi(t{+}1))$. Thus

$$\begin{aligned}
&\mathbb{P}\big(X(t{+}1){=}j \,\big|\, X(t){=}i, X(t{-}1) = x_{t-1}, \ldots, X(0){=}x_0\big) \\
&\quad = \mathbb{P}\big(f(i,\xi(t{+}1)){=}j \,\big|\, X(t){=}i, X(t{-}1) = x_{t-1}, \ldots, X(0){=}x_0\big) \\
&\quad = \mathbb{P}\big(f(i,\xi(t{+}1)){=}j\big)\,, \qquad \text{(by independence).}
\end{aligned}$$

As this last expression depends only on i and j, it proves the Markov property and shows $p_{ij} = \mathbb{P}\big(f(i,\xi(t{+}1)){=}j\big)$.

A converse is true. Given any stochastic matrix, A, it is possible to construct a function f and identically distributed random variables $\{\xi(t)\ t \geq 1\}$ such that the solution to (4.13) is a Markov chain with transition matrix A. We will show this in a later section dealing with simulation of Markov chains. ◇

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, because more than just the present value of the state affects the one-step-ahead conditional probability. It is still possible to use Markov process ideas in this circumstance, as the next example illustrates.

Example 4.1.10. k-Markov chains and 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 of the population, 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 of 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. Thus, for example, one should expect that the conditional distribution of $X(t+1)$ given say, $X(t-1) = A, X(t) = A$ will be different than the conditional distribution given $X(t-1) = C, X(t) = A$. It might even be true that that the past k bases, where $k > 2$, affect the probability of the next base.

To take into account the influence of more past variables on the one-step-ahead transition probability, one can use k-Markov chains. A process $\{X(t)]$ is a k-Markov chain if, for any t, k previous bases always suffice to characterize one-step-ahead conditional probabilities: that is,

$$\begin{aligned} &P\Big(X(t+1) = x_{t+1} \;\Big|\; X(0) = x_0, \ldots, X(t) = x_t\Big) \\ &\quad = \quad P\Big(X(t+1) = x_{t+1} \;\Big|\; X(t) = x_t, \ldots, X(t-k+1) = x_{t-k+1}\Big). \end{aligned} \tag{4.15}$$

(For this to make sense, $t \geq k-1$.) In fact k-Markov chains for moderate values of k better fit DNA sequence data.

k-Markov chains can be analyzed on their own terms, but there is a trick that turns them into Markov chain models, so working with them requires no new theory or techniques. We explain the trick for $k = 2$; the general case is an easy extension of the same method. The idea is to look at a new process

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

This has a more complicated state space than the original process $\{X(t)\}$. But it turns out that if $\{X(t)\}$ is 2-Markov, then $\{Z(t)\}$ is a Markov chain in the ordinary sense. For suppose we know the values of $Z(1), \ldots, Z(t-1), Z(t)$, and we want to calculate a conditional probability concerning $Z(t+1) = (X(t), X(t+1))$. Knowing $\{Z(0), \ldots, Z(t)\}$ is the same as knowing $\{X(0), \ldots, X(t)\}$. If we know this, we already know the component $X(t)$ of $Z(t+1)$ and so computing the conditional probability of $Z(t+1)$ given the past history of the Z-process reduces to computing a conditional probability of $X(t+1)$ given $\{X(0), \ldots, X(t)\}$. But since the X-process is 2-Markov, this conditional probability depends only on $Z(t) = (X(t-1), X(t))$, and not on any previous values of the Z-process. Thus $\{Z(t)\}$ is a Markov chain.

See Exercise 4.1.3 for a problem computing the transition probabilities of $\{Z(t)\}$ from those of $\{X(t)\}$ and see Exercise 4.1.9 for the rudiments of a direct analysis of 2-Markov chains.

$\diamond$

4.1.4 Exercises

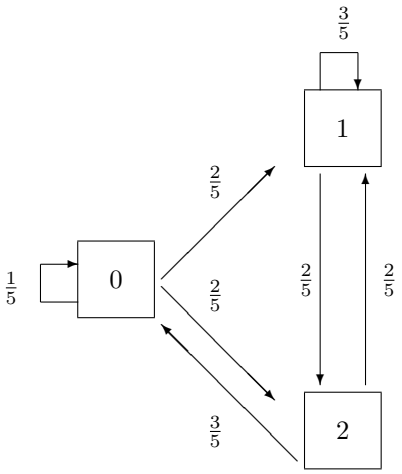
4.1.1. Consider a three state Markov chain with transition probability matrix:

	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) 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 $\mathbb{P}(L = k)$?

e) 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}. What type of random variable is L (conditional on $X(0)=s$)?

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

$$\begin{aligned}&\mathbb{P}\left(X(t+1)=x \;\middle|\; X(t)=x_t, X(t-1)=x_{t-1}, \ldots, X(0)=x_0\right)\\ &\quad = \mathbb{P}\left(X(t+1)=x \;\middle|\; X(t)=x_t, X(t-1)=x_{t-1}\right). \end{aligned} \tag{4.16}$$

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

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

We can still compute the probability of a path relatively easily assuming (4.16). 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 (4.16) 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. Let $\xi(2), \xi(3), \dots$ be independent and identically distributed and suppose they are also independent of $X(0)$ and $X(1)$. Show that

$$X(t+2) = \phi\Big(X(t+1), X(t), \xi(t+2)\Big)$$

defines a 2-Markov chain.

4.1.12. Let $\{X(t)\}$ be a Markov chain. Verify)4.7) and (4.8) using (4.4). Use them to show, $\{Y(t) = X(s+t);\, t = 0, 1, \dots\}$, is a Markov chain with the same transition probabilities as $\{X(t)\}$. Hint:

$$\begin{aligned}&\mathbb{P}\Big(X(s)=x_0, X(s+1)=x_1, \dots, X(s+t)=x_t\Big)\\ &\quad= \sum_{k_0,\dots,k_{s-1}} \mathbb{P}\Big(X(0)=k_0, \dots, X(s-1)=k_{s-1}, X(s)=x_0, X(s+1)=x_1, \dots, X(s+t)=x_t\Big).\end{aligned}$$

4.2 Theory of Markov Chains, Part I: Computation

This section is about computing probabilities and expectations of Markov chains with finite state spaces. Throughout, except in some examples, the state space is

$$\mathcal{E} = \{1, 2, \dots, N\}.$$

This entails no loss of generality because the states can always be relabeled. We use $A = [p_{ij}]_{i,j\in\mathcal{E}}$ to denote a generic state transition matrix.

4.2.1 Computing the distribution of $X(t)$.

For each state, i, and time, $t \geq 0$, let $\rho_i(t) := \mathbb{P}\Big(X(t)=i\Big)$ denote the probability that $X(t)$ is in state i, and let

$$\rho(t) \;:=\; \Big(\rho_1(t), \rho_2(t), \dots, \rho_N(t)\Big) = \Big(\mathbb{P}(X(t)=1), \dots, \mathbb{P}(X(t)=N)\Big).$$

This is a row vector representation of the probability mass function of $X(t)$, and so $\rho(t)$ is called the probability *law* or probability *distribution* of $X(t)$. The vector $\rho(0)$ is called the *initial law*.

Related quantities of interest are the t-step ahead conditional probabilities, defined by

$$p_{ij}^{(t)} = P\Big(X(s+t)=j \;\Big|\; X(s)=i\Big).$$

For any fixed i,

$$\Big(p_{i1}^{(t)}, p_{i2}^{(t)}, \dots, p_{iN}^{(t)}\Big)$$

represents the probability distribution of $X(t)$ when the chain starts in state i with probability one. Indeed, if $\mathbb{P}(X(0)=i)=1$,

$$\mathbb{P}(X(t)=j)=\mathbb{P}(X(t)=j,X(0)=i)=P\Big(X(t)=j\;\Big|\;X(0)=i\Big)\mathbb{P}(X(0)=i)=p_{ij}^{(t)}\,,\;\; i,j\in\mathcal{E}\,.$$

How does one compute $\rho(t)$? Consider $\rho_j(t)=\mathbb{P}(X(t)=j)$ for a fixed j. One of the disjoint events, $\{X(t-1)=1\},\ldots,\{X(t-1)=N\}$, must occur, and, so, by the rule of total probability,

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

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

$$\rho_j(t)=\sum_{k=1}^{N}\rho_k(t-1)p_{kj},\quad\text{for each } j\in\mathcal{E}. \tag{4.17}$$

This sum is the j^{th} component of the vector-matrix 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

$$\rho(t)=\rho(t-1)\cdot A,\quad t\geq 1. \tag{4.18}$$

This elegant formula is the reason for expressing $\rho(t)$ as a row vector. The following theorem collects some important consequences.

Theorem 1 *Let $\{X(t)\}$ be a time-homogeneous Markov chain with state transition matrix A.*

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

$$\rho(t)=\rho(t-s)\cdot A^s. \tag{4.19}$$

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

$$\rho(t)=\rho(0)\cdot A^t. \tag{4.20}$$

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

$$p_{ij}^{(t)}=P\Big(X(s+t)=j\;\Big|\;X(s)=i\Big)=[A^t]_{ij}. \tag{4.21}$$

Proof: Equation (4.19) is derived by repeated application of (4.18):

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

Part b) is a consequence of part a). Suppose, the chain starts in state i with probability one. As we saw above, this implies $P\big(X(t)=j\big) = p_{ij}^{(t)}$. Because, in this case, $\rho_i(0) = 1$ and $\rho_j(0) = 0$ if $j \neq i$, part a) implies

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

◇

The formulas of Theorem 1 are wonderful results, because they reduce the calculation of probability laws and t-step-ahead probabilities for Markov chains to computing powers of the state transition matrix. This is true even for Markov population models with complex mechanisms of selection and mutation. Contrast this with the infinite population model with selection studied in Section 3.4, which involved a nonlinear difference equation without an explicit solution.

One of the main objects of Markov chain theory is to understand how $\rho(t)$ behaves in the limit as $t \to \infty$, which is also often the main thing we want to deduce from an applied model. Because of Theorem 1, this problem reduces to studying the evolution, as $t \to \infty$, of a matrix product, A^t. Under suitable assumptions, this limiting behavior has a simple form that is relatively easy to characterize. This is the ergodic theory of Markov chains, which is treated and applied in Section 4.5. Meanwhile, we illustrate Theorem 1 by two examples. These will reveal simple instances of limit behavior that typify properties of more general Markov chains.

Example 4.2.1. The two-state chain. The powers of the state transition matrix can be computed explicitly for the two-state chain, and the results are very interesting.

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. Straight calculation of the first few powers by hand, or even using mathematical software,

gets complicated rapidly and reveals no simple patterns. However, by being more astute, a simple formula can be found:

$$A^t = \frac{1}{\lambda+\mu}\begin{pmatrix} \mu+\lambda\alpha^t & \lambda-\lambda\alpha^t \\ \mu-\mu\alpha^t & \lambda+\mu\alpha^t \end{pmatrix} \quad \text{where } \alpha = 1-\lambda-\mu. \tag{4.22}$$

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 (4.22). The reader should verify the following—it requires only straightforward calculation:

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

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 (4.22) 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 $P(X(t+1)=j\big|X(0)=i) = [A^t]_{ij}$ by Theorem 1, it says

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

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,

$$\begin{aligned}\lim_{t\to\infty}\ (P(X(t)=0), P(X(t)=1)) &= \lim_{t\to\infty}\rho(t) = \lim_{t\to\infty}\rho(0)\cdot A^t \\ &= \lim_{t\to\infty}(\rho_0(0),\rho_1(0))\cdot\begin{pmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \\ \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{pmatrix} \\ &= \left((\rho_0(0)+\rho_1(0))\frac{\mu}{\lambda+\mu}, (\rho_0(0)+\rho_1(0))\frac{\lambda}{\lambda+\mu}\right) \\ &= \left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu}\right). \end{aligned} \tag{4.23}$$

Thus the limiting probabilities to be in state 0 or 1 are independent of the 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 condition—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 4.5, we discuss conditions on general, finite-state-space chains that guarantee the existence of $\lim_{t\to\infty} \rho(t)$ and its independence of the initial distribution. $\diamond$

Example 4.2.2. The Moran model with $N = 4$. When $N = 4$ the state transition matrix for the Moran model, calculated in Example 4.1.6, is:

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

Calculating higher and higher powers of A reveals how the Moran model behaves and will provide an interesting contrast to the behavior of the two-state chain discussed in Example 4.2.1. When $0 < \lambda, \mu < 1$ neither state of the two-state chain is absorbing, and the probabilities for $X(t)$ to be in each state settle down to a limit independent of the initial distribution as $t \to \infty$. In contrast, the Moran model contains two absorbing states, both of which can be reached from any non-absorbing state. In this example we will explore the behavior of A^t numerically. The conclusions we draw will be verified theoretically in the next section of this chapter.

Three representative powers of the state transition matrix, computed using the Maple software package, are shown below, rounded to three significant digits:

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

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

$$A^{34} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .745 & .003 & .003 & .003 & .245 \\ .494 & .004 & .004 & .004 & .494 \\ .245 & .003 & .003 & .003 & .745 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

(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, due to round-off error.)

Several features of A and its powers are immediately apparent. For one, the first and last rows never change; in fact, they are the same in all powers, A^t. This is easy to see. The first row of A^t contains the transition probabilities out of state 0 in t steps: $[A^t]_{0i} = p_{0j}^{(t)} = \mathbb{P}(X(t) = j \big| X(0) = 0) = 1$. But 0 is an absorbing state, which the chain never leaves, and thus, $[A^t]_{0j} = 0$, if $j \neq 0$, for all t. The same analysis applies to the last row of A^t, containing the transition probabilities out of the absorbing state 4.

Each power also exhibits the same symmetry: the fourth row is the second in reverse and the third reads the same forward or back. In fact, for all t,

$$p_{ij}^t = p_{4-i,4-j}^t \quad \text{for each } i, j \in \{0, 1, 2, 3, 4\} \text{ for all } t.$$

This can be verified directly for $t = 1$ and it implies that $4 - X(t)$ also follows the Moran model with $N = 4$. Indeed,

$$\mathbb{P}\Big(4{-}X(t{+}1){=}j \;\Big|\; 4{-}X(t){=}i\Big) = \mathbb{P}\Big(X(t{+}1){=}4{-}j \;\Big|\; X(t){=}4{-}i\Big) = p_{4-i,4-j} = p_{ij}\,.$$

It follows that the transition probability in t steps from $4 - i$ to $4 - j$ should be the same as that from i to j, for any t. This fact just expresses the selective neutrality of the Moran model. $X(t)$ counts the number of type A individuals and $4 - X(t)$ the number of type a's, but neither has a selective advantage, so $X(t)$ and $4 - X(t)$ are both Moran chains. The Wright-Fisher model is also selective neutral and the powers of its transition matrix will exhibit the same symmetry.

Finally, consider the entries of the first and last columns of A^t. The entries of the first column are $[A^t]_{i0} = \mathbb{P}(X(t) = 0 \big| X(0) = i)$. 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, p_{i0}^t is the same as the probability the chain hits state 0 *at or before* time t. Similarly, the entries of the last column of A^t are the probabilities of getting absorbed in state 4 by step t, starting from the different states. All these can only increase with t, since additional time gives only more opportunity to reach 0 or 4. We see this clearly in the powers of A calculated above.

Since a bounded increasing sequence has a limit, $p_{i0}^\infty := \lim_{t\to\infty} p_{i0}^t$ and $p_{i4}^\infty := \lim_{t\to\infty} p_{i4}^t$ must exist; the limit p_{i0}^∞ is the probability that, starting in i, the chain eventually enters the absorbing state 0, while p_{i4}^∞ is the probability it eventually enters state 4. It follows from the values obtained for A^{34} that

$$\begin{aligned} p_{30}^\infty + p_{34}^\infty \;&=\; p_{10}^\infty + p_{14}^\infty > p_{10}^{34} + p_{14}^{34} = 0.99\,, \\ &\qquad p_{20}^\infty + p_{24}^\infty > p_{20}^{34} + p_{14}^{24} = 0.99\,. \end{aligned}$$

This suggests strongly that the probability of eventually hitting an absorbing state is one. (In genetics, hitting one of these absorbing states is called *fixation*.) This is

indeed true. We will show in the next section that

$$\lim_{t\to\infty} A^t = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .75 & 0 & 0 & 0 & .25 \\ .5 & 0 & 0 & 0 & .5 \\ .25 & 0 & 0 & 0 & .75 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \diamond$$

4.2.2 Expectations for Markov chains.

Conditional expectations and expectations for Markov chains may also be computed using matrix algebra. This section presents the basic identities and shows how they are used.

Let g be a function defined on the state space $\mathcal{E}$ of $\{X(t)\}$. We first derive a formula for the t-step ahead conditional expectation, $E[g(X(s+t))|X(s)=i]$. This is simple. By the definition of conditional expectation,

$$\begin{aligned} E\left[g(X(s+t)) \,\middle|\, X(s)=i\right] &= \sum_{j\in\mathcal{E}} g(j)\mathbb{P}\left(X(s+t)=j \middle| X(s)=i\right) \\ &= \sum_{j\in\mathcal{E}} p_{ij}^t g(j) \qquad \text{for all } i\in\mathcal{E}. \end{aligned} \tag{4.24}$$

It is helpful to express this using matrix and vector products. To this end, assume that the state space is $\mathcal{E} = \{1, 2, \ldots, N\}$. We will identify a function g on this state space with the column vector,

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

Then the right-hand side of (4.24) is the vector product

$$\sum_{j\in\mathcal{E}} p_{ij}^t g(j) = \left(p_{i1}^t, p_{i2}^t, \ldots, p_{iN}^t\right) \cdot \begin{pmatrix} g(1) \\ g(2) \\ \vdots \\ g(N) \end{pmatrix},$$

and this is the i^{th} component of the matrix vector product,

$$\begin{pmatrix} p_{11} & p_{12}^t & \cdots & p_{1N}^t \\ p_{21}^t & p_{22} & \cdots & p_{2N}^t \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1}^t & p_{N2}^t & \cdots & p_{NN}^t \end{pmatrix} \cdot \begin{pmatrix} g(1) \\ g(2) \\ \vdots \\ g(N) \end{pmatrix} = A \cdot \mathbf{g}.$$

Since (4.24) is true for each state i, it follows that

$$\begin{pmatrix} E\left[g(X(t+1)) \,\middle|\, X(t)=1\right] \\ \vdots \\ E\left[g(X(t+1)) \,\middle|\, X(t)=N\right] \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix} \begin{pmatrix} g(1) \\ g(2) \\ \vdots \\ g(N) \end{pmatrix}$$

$$= A \cdot \mathbf{g} \tag{4.25}$$

The same reasoning shows:

$$\begin{pmatrix} E\left[g(X(s+t)) \,\middle|\, X(s)=1\right] \\ \vdots \\ E\left[g(X(s+t)) \,\middle|\, X(s)=N\right] \end{pmatrix} = A^t \cdot \mathbf{g} \tag{4.26}$$

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

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

From formula (4.20) in Theorem 1, $\rho(t) = \rho(0)\cdot A^t$. Thus

$$E[g(X(t))] = \rho(0)\cdot A^t \cdot \mathbf{g} \tag{4.27}$$

Again, all expectation calculations are reduced to products of A and products with row and column vectors.

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

Recall the neutral Wright-Fisher chain for a population of N alleles; its transition probabilities are

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

Allele numbers in this chain fluctuate randomly; this is called *genetic drift*. What about the *average* numbers of alleles? In this example, we will show it is constant in expectation, that is,

$$E\big[X(t+1)\big|X(t)=i\big] = i, \tag{4.28}$$

and, whatever the initial distribution is,

$$E[X(t)] = E[X(0)], \quad \text{for all } t \geq 1. \tag{4.29}$$

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

To derive identities (4.28) and (4.29), recall the definition of the Wright-Fisher chain: *given* $X(t) = i$, the next value, $X(t+1)$, is a binomial random variable with parameters $n = N$ and $p = i/N$. The expectation of a binomial random variable with parameters n and p is np. Thus $E[X(t+1)\big|X(t)=i] = N(i/N) = i$, as claimed. That the expected number of A alleles is constant follows easily, by conditioning on $X(t)$ and using the definition of expectation:

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

It is interesting to look at this calculation from the point of view of formulas (4.25) and (4.27). 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

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

and, from (4.25), 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.) As a consequence, $A^2\mathbf{h} = A(A\mathbf{h}) = A\mathbf{h} = \mathbf{h}$; $A^3\mathbf{h} = A(A^2\mathbf{h}) = A\mathbf{h} = \mathbf{h}$, and by induction,

$$A^t\mathbf{h} = \mathbf{h} \quad \text{for all positive integers } t.$$

From equation (4.27),

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

for all t, which establishes again that the expected value does not change with t.

$E[X(t)]$ is also constant for the Moran model with no mutation or selection; see Exercise 4.2.4. ◇

4.2.3 Exercises

4.2.1 Let $B(t)$ be the matrix in equation (4.22). 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 $P(X(5) = 2 \big| X(3) = 4)$.

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

(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 $E[X^2(10)]$.

4.2.4. Prove (4.28) and (4.29) 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 Simulation of Markov Chains

Realistic Markov chain models in applications are often too complicated to analyze theoretically, and to learn how they behave it is then necessary to turn to simulation. Given a transition probability matrix A, one needs an algorithm that takes a sequence of random variables as an input and produces a Markov chain with transition probability matrix A as output. This algorithm can be run in a mathematical software program using the random number generator it provides, which typically simulates a sequence of independent uniform random variables. By running the simulation multiple times, one can build up an empirical picture of how the chain evolves.

There are different strategies for constructing a simulation algorithm. The one discussed here takes a sequence $U(0), U(1), \ldots$ of independent random random variables, all uniformly distributed on the interval $(0, 1)$, as input. Given an $N \times N$

stochastic matrix A on the state $\mathcal{E} = \{1, \ldots, N\}$ and a probability mass function $\rho(0) = \big(\rho_1(0), \ldots, \rho_N(0)\big)$ on $\mathcal{E}$, we will show how to construct functions $\phi(i, u)$ and $\psi(u)$ such that

$$X(t+1) = \phi\big(X(t), U(t+1)\big), \qquad X(0) = \psi(U(0)) \tag{4.30}$$

defines a Markov chain with transition probability matrix A and initial law $\rho(0)$. A system like this is easy to code and run using numerical software. This algorithm will also justify claim made in Section 4.1 that any Markov chain can be simulated by a random difference equation of the type in equation (4.13).

They key to the construction is a method for obtaining any discrete random variable we like from a uniformly distributed random variable. Say we are given probabilities $p_1, \ldots, p_N$ adding up to one and we want to construct a random variable Y for which

$$\mathbb{P}\big(Y = k\big) = p_k \quad \text{ for } k = 1, 2, \ldots, N,$$

Define a partition of the interval $[0, 1]$ as follows:

$$\begin{aligned} s_0 &= 0, \\ s_1 &= p_1, \\ s_2 &= p_1 + p_2 \\ &\cdots \\ s_k &= p_1 + \cdots + p_k, \\ &\cdots \\ s_n &= p_1 + \cdots + p_N = 1 \end{aligned}$$

The k^{th} subinterval of this partition is $[s_{k-1}, s_k]$, and its length is $s_k - s_{k-1} = p_k$. Let the function $\eta(u)$ on $[0, 1)$ be defined by the rule:

$$\eta(u) = k \quad \text{if } s_{k-1} \le u < s_k.$$

If U be uniformly distributed on $(0, 1)$, then

$$\mathbb{P}\big(\eta(U) = k\big) = \mathbb{P}\big(\eta(U) = k\big) = \mathbb{P}\big(s_{k-1} \le U < s_k\big) = s_k - s_{k-1} = p_k,$$

hence $\eta(U)$ is the desired random variable, Y.

Let $A = \big[p_{ij}\big]_{1 \le i,j \le N}$ be a given stochastic matrix. We will apply the construction just defined to each row of A and to the initial law $\rho(0) = \big(\rho_1(0), \ldots, \rho_n(0)\big)$. For each row i, define

$$\phi(i, u) = j \quad \text{ if } p_{11} + \cdots + p_{1,j-1} \le u < p_{11} + \cdots + p_{1,j}.$$

Also define

$$\psi(u) = j \quad \text{ if } \rho_1(0) + \cdots + \rho_{j-1}(0) \le u < \rho_1(0) + \cdots + \rho_j(0).$$

Let $U(0), U(1), \ldots$ be independent random variables each of which is uniformly distributed on $(0,1)$, and let $\{X(t)\}$ be the solution of the solution of (4.30) with ϕ and ψ as just defined. Then $\{X(t)\}$ will be a Markov chain with transition probability matrix A and initial law $\rho(0)$. Indeed, that $X(0) = \psi(U(0))$ has distribution $\rho(0)$ is an immediate consequence of our construction. It was shown in Example 4.1.9 that $\{X(t)\}_{t\geq 0}$ is a Markov chain. The following calculation, using the independence of $X(t)$ and $U(t+1)$), shows that it has the correct transition probabilities:

$$\begin{aligned}\mathbb{P}\Big(X(t+1)=j\Big|X(t)=i\Big) &= \mathbb{P}\big(\phi(X(t),Y(t+1))=j\big|X(t)=i\Big)\\ &= \mathbb{P}\big(\phi(i,U(t+1))=j\Big) = p_{ij}.\end{aligned}$$

4.4 The Markov property

The Markov condition of equation (4.2), which defined Markov chains, is a statement about the conditional probability of the chain *one step* into the future, conditioning on its *entire* past history. It implies a much more general fact, called the *Markov property,* concerning conditional probabilities of future events on past history. We say an event U belongs to the *past of the process,* $\{X(t)\}$, *at time* t if it is defined completely in terms of the path, $\{X(0), \ldots, X(t)\}$, up to time t; we say an event V belongs to the *future of* $\{X(t)\}$ *at time* t if it is defined completely by the future evolution of the process, $\{X(t), X(t+1), X(t+2), \ldots\}$. For example, the event $U = \{X(2)=4, 10 \leq X(5) < 20\}$ belongs to the past of the process at time $t=6$. The event V that there exist an $s \geq 6$ such that $X(s) = 0$ belongs to the future at $t=6$.

Theorem 2 *(The Markov property.) Let $\{X(t)\}$ be a Markov chain. For any t, any state i, any event U belonging to the past of the process t, and any event V belonging to the future of the process at t,*

$$\mathbb{P}\big(V \;\big|\; X(t)=i, U\big) = \mathbb{P}(V \;|\; X(t)=i)\,. \tag{4.31}$$

The Markov condition, (4.2), is the special case in which $V = \{X(t+1) = j\}$ and $U = \{X(0) = x_1, \ldots, X(t-1) = x_{i-1}\}$, but it implies the more general Markov property. The proof of this theorem is omitted. Exercise 4.4.1, which asks for a derivation for a special case of V and U, gives an inkling of why the result is true and how the proof goes in general.

The Markov property has a frequently used consequence. When $\{X(t)\}$ is a time-homogeneous Markov chain with transition probabilities p_{ij}, the Markov property implies that for any event in the past at time t,

$$\begin{aligned}&\mathbb{P}\Big(X(t+1)=j_1, X(t+2)=j_2, \ldots, X(t+s)=j_s\Big|X(t)=i, U\Big)\\ &\quad = p_{i,j_1}p_{j_1,j_2}\cdots p_{j_{s-1},j_s}\\ &\quad = \mathbb{P}\Big(X(1)=j_1, X(2)=j_2, \ldots, X(s)=j_s\Big|X(0)=i\Big)\,.\end{aligned} \tag{4.32}$$

Thus, given that $X(t)=i$, the chain going forward into the future, namely,

$$Y(0)=X(t), Y(1)=X(t+1), \ldots, Y(s)=X(t+s), \ldots ,$$

behaves like $\{X(t)\}$ starting in state i at time 0, *independently of any past event* U. As an illustration,

$$\begin{aligned}\mathbb{P}\Big(X(s) \text{ visits state } b \text{ for some } s \geq t \;\Big|\; X(t)=i, X(t-1)=k\Big) \\ = \mathbb{P}\Big(X(t) \text{ visits state } b \text{ for some } t \geq 0 \;\Big|\; X(0)=i\Big)\end{aligned} \tag{4.33}$$

This identity will be used in the next example.

Example 4.4.1. Gambler's ruin. Consider the random walk with absorbing boundaries, as defined in Example 4.1.4. States 0 and N are absorbing ($p_{00} = p_{NN} = 1$), but from any other state i, the chain moves either to state $i + 1$ with probability p or to state $i - 1$ with probability $q = 1 - p$. (We drop the subscript R used in Example 4.1.4.) Let a_i denote the probability that the chain eventually hits state N, given that it starts in state i. For the application of Example 4.4.1, this is the probability that eventually all the cells in the crypt are the mutant type. For the gambling interpretation, it is the probability that the gambler eventually wins all N dollars at stake. We will use the Markov property to find an explicit formula for a_i as a function of i. (In the gambling interpretation, the game stops when either the gambler acquires a fortune of size N, meaning the casino goes broke, or when his fortune drops to 0, which is his ruin. We will see that $1 - a_i$ is the probability of ruin.)

Since N is an absorbing state and the chain stays in state N once there,

$$a_i = \mathbb{P}\big(B \;\big|\; X(0)=i\big) ,$$

where $B = \Big\{X(s) \text{ visits } N \text{ for some } s \geq 1\Big\}$. If the random walk is at state i, where $1 < i < N$, it can only move to state $i+1$ or $i-1$ in the next step. The result derived in equation (4.33) implies:

$$\begin{aligned}\mathbb{P}\big(B \;\big|\; X(1)=i+1, X(0)=i\big) &= \mathbb{P}\Big(X(t)=N \text{ for some } t \geq 0 \;\Big|\; X(0)=i+1\Big) = a_{i+1} ; \\ \mathbb{P}\big(B \;\big|\; X(1)=i-1, X(0)=i\big) &= \mathbb{P}\Big(X(t)=N \text{ for some } t \geq 0 \;\Big|\; X(0)=i-1\Big) = a_{i-1} .\end{aligned}$$

The probability of moving out of state i to $i+1$ is p and to $i-1$ is q. Therefore, by conditioning on on the first transition out of state i,

$$\begin{aligned}a_i \;&=\; \mathbb{P}\big(B \;\big|\; X(1)=i+1, X(0)=i\big)\mathbb{P}(X(1))=i+1|X(0)=i) \\ &\qquad + \mathbb{P}\big(B \;\big|\; X(1)=i-1, X(0)=i\big)\mathbb{P}(X(1))=i-1|X(0)=i) \\ &= a_{i+1}p + a_{i-1}q, \quad 0 < i < N .\end{aligned} \tag{4.34}$$

This is a homogeneous, linear, second order difference equation for $\{a_i\}$ which we know how to solve—see the Appendix of Chapter 3. Its general solution is $a_i = c_1 + c_2(q/p)^i$. The constants c_1 and c_2 are found by applying the boundary conditions,

$a_0 = 0$ (because the chain can never reach N if it starts in state 0); and
$a_N = 1$ (because the chain stays in N forever if it starts there).

When $p \neq q$, the solution is

$$a_i = \frac{1 - \left(\frac{q}{p}\right)^i}{1 - \left(\frac{q}{p}\right)^N}.$$

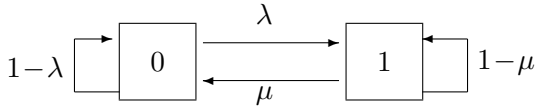
If $p = q = 1/2$, the solution is instead $a_i = i/N$. The derivations are left as exercises for the reader.

Suppose we wish to compute the probability, b_i, that the random walk gets absorbed in state 0 when starting from i. By relabeling states $0, 1, \ldots, N$ in the reverse order, we see that this is just the probability of absorption in state N, starting from $N - i$, when the probability of moving up a step is q. Thus

$$b_i = \frac{1 - \left(\frac{p}{q}\right)^{n-i}}{1 - \left(\frac{p}{q}\right)^N} = \frac{\left(\frac{q}{p}\right)^i - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N}.$$

The second expression is obtained by multiplying the first by $(q/p)^N$. It is easy to check that $a_i + b_i = 1$, which is a significant result, because it says the random walk must eventually hit one of the absorbing states 0 or N; it cannot bounce around forever among the non-absorbing states. ◇

Example 4.4.2. Expected return times in the two state chain. Consider the two-state chain defined in Example 4.1.2.



Assume that $\lambda > 0$ and $\mu > 0$, so that neither state is absorbing. A chain in a non-absorbing state must eventually leave that state. For example, the probability that the two-state chain stays in state 0 for at least k consecutive time steps is the probability that it returns to itself k times in a row, which is $(1 - \lambda)^k$; see Exercise

4.1.2 e). If $\lambda > 0$, then $\lim_{k\to\infty}(1-\lambda)^k = 0$ and hence the probability of staying in 0 forever is 0.

Let T_0 be the first time strictly after time $t = 0$ at which the chain visits state 0. We are interested in computing $E\big[T_0\big|X(0)=0\big]$, the expected time of first return to 0, starting from 0. By conditioning on what happens at the first step,

$$E\big[T_0\big|X(0)=0\big] = (1-\lambda)E[T_0\big|X(1)=0, X(0)=0] + \lambda E[T_0\big|X(1)=1, X(0)=0].$$

Now, $T_0 = 1$ if $X(1) = 0$. And by the Markov property, $E[T_0\big|X(1)=1, X(0)=0] = 1 + E[T_0\big|X(0)=1]$, because we have taken one step to get to state 1 and now we have to add the additional expected time to wait before returning to i. The time spent in state 1 is a geometric random variable with expectation $1/\mu$, and hence $E[T_0\big|X(0)=1] = 1/\mu$. Therefore

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

The same argument, obtained by switching the roles of λ and μ, shows that

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

4.4.1 Exercises

Exercise 4.4.1. Show directly, using only the Markov condition and it consequences, (4.7) and (4.8):

$$\mathbb{P}\Big(X(t{+}1){=}j_1, X(t{+}2{=}j_2 \;\big|\; X(t){=}i, X(t{-}1){=}k\Big) = \mathbb{P}\Big(X(t{+}1){=}j_1, X(t{+}2{=}j_2 \;\big|\; X(t){=}i\Big)\,.$$

Exercise 4.4.2. Use the Markov property to show: if $\{X(t)\}$ is a Markov chain with state transition matrix A, then $\{Y(t) = X(2t);\, t \geq 0\}$ is a Markov chain with state transition matrix A^2.

4.5 Limit behavior of Markov chains

Two related questions dominate Markov chain theory. How does the distribution, $\rho(t)$, of $X(t)$ evolve in the limit as $t \to \infty$? And what can one say about sample paths of Markov chains as they evolve through time? These are also vital questions in applications. What long-time behavior does an applied model predict and does it match what is observed?

Because $\rho(t) = \rho(0)\cdot A^t$, as we derived in Section 4.2, the study of Markov chain behavior reduces to understanding powers of stochastic matrices and their limits. The results of the theory are very elegant; $\lim_{t\to\infty} A^t$ exists quite generally, given

only natural and generic conditions on the chain, and there are simple algebraic and also probabilistic characterizations of this limit. This section presents the most important results and illustrates how they are applied. The focus is on important cases of applied interest, rather than the most general situation, and proofs are mostly omitted. More complete treatments may be found in standard texts, for example S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, volume 1, Elsevier, San Diego, 1975, or J.R. Norris, *Markov Chains*, Cambridge, New York, 1997.

The two example analyses treated in Section 4.2.1, the study of the two-state chain in Example 4.2.1 and the study of the neutral Moran model in Example 4.2.2, already illustrate the main theoretical results and are useful to keep in mind. For example, the Moran model has the property that either absorbing state can be reached from any non-absorbing state, and numerical evidence obtained by computing A^t for several t suggest the Moran chain must eventually eventually hit one of its absorbing states, no matter where its starts from. We proved this is true for random walk with absorbing boundaries in Example 4.4.1, by computing explicit expressions for the probability of absorption. It turns out to be 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. In Section 4.5.2, we show how to compute the probability of ending up in any particular absorbing state.

In contrast,the two-state chain with both $p_{01} = \lambda > 0$ and $p_{10} = \mu > 0$, has no absorbing states. By computing A^t and $\lim_{t\to\infty} A^t$ exactly we saw that a limit distribution, $\lim_{t\to\infty} \Big(\mathbb{P}(X(t)=0), \mathbb{P}(X(t)=1)\Big)$, exists and is the same no matter what the distribution of $X(0)$ is. The ergodic theorem discussed in Section 4.5.3 generalizes this fact and also establishes a law of large numbers: in the infinite time limit, the average amount of time spent by the chain in state i tends to the limiting probability to be in i.

4.5.1 Classification of states.

The long-time behavior of a Markov chain obviously depends on the possible ways it can move among its states. To describe these possibilities, it is useful to introduce *hitting times*. If i is a state, let T_i denote the first time t, strictly after $t = 0$, at which $X(t) = i$:

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

If the chain never hits i, set $T_i = \infty$. (To emphasize: when $X(0) = i$, T_i is not 0; it is the first time after 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

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

In words, j is accessible from i if, when starting from i, the probability the chain will hit j in the future is positive. It is easy to check accessibility: $i \to j$ if and

only if 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. 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 write $i \sim j$. As a matter of convention, we say that i communicates with itself ($i \sim i$). A *communicating class* is a maximal set of communicating states: all states in the class communicate with each other and there are no other states communicating with any of its members. A Markov chain is called *irreducible* if all its states communicate, that is, if its states form one communicating class.

Examples 4.5.1. (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 on the state space $\{0, 1, \ldots, N\}$. We know states 0 and N are absorbing. If i is any non-absorbing state and j is another state, the path which moves by unit steps from i to j has positive probability, because the probability of each unit step, either the transition probabilities $p_{k,k+1}$ and $p_{k,k-1}$ is positive for all k, $0 < k < N$. Thus, all other states are accessible from i. In particular, if both i and j are non-absorbing, $i \to j$ and $j \to i$ and hence i and j communicate. Therefore $\{1, \ldots, N-1\}$ is a communicating class.

The situation is similar for the Wright-Fisher chain without mutation on the state space $\{0, 1, \ldots, M\}$. The transition probability and $p_{ij} > 0$ as long as i is not one of the absorbing states, 0 or M. Thus, if $1 \le i, j \le M - 1$, then $i \sim j$. ◇

(c) The two-state Markov chain is irreducible if $\lambda > 0$ and $\mu > 0$. The Moran model and the Wright-Fisher model with mutation, as presented in Exercises 4.1.6 and 4.1.7, are also irreducible if both mutation probabilities, of A to a and a to A, are positive. ◇

Several more definitions are needed. The event $\{T_i < \infty\}$ is the event of at least one visit to i. A state i is said to be *recurrent* if, starting in state i, the chain returns to i with probability one: in other words, if

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

If all the states of a Markov chain are recurrent we say the chain is recurrent. A state that is not recurrent is said to be *transient*. The terms *recurrent* and *transient* are justified by the following theorem.

Theorem 3 *Let V_i be the number of visits of a chain to i.*

If i is transient, $\mathbb{P}(V_i < \infty) = 1$.

If i is recurrent, $\mathbb{P}(V_i = \infty \mid T_i < \infty) = 1$. (If the chain visits i once, it must visit i infinitely often.)

It turns out also that transience and recurrence are properties of communicating classes of states.

Theorem 4 *If $i \sim j$, then states i and j are either both transient or both recurrent.*

We shall not write out a rigorous proof of Theorem 3, but the intuition behind it is easy to understand. Because of the Markov property and the assumption of time-homogeneity, once a chain visits i, it behaves in the future just like the chain starting at i at time 0. By definition, if i is recurrent, the chain starting at i at time 0 must eventually return to i. Thus, it must return to i after every visit to i and so must return an infinite number of times. On the other hand, if i is transient, the probability of not returning to i, starting from i, is $1-\mathbb{P}\big(T_i < \infty \big| X(0)=i\big)$, which is positive. By the Markov property, the event of returning or not to i is independent of the past up to the last visit to i. It's as if one is flipping a coin that comes up heads with probability $1 - \mathbb{P}\big(T_i < \infty \big| X(0)=i\big) > 0$ to decide whether to return or not. With probability one, heads must eventually appear, so eventually the chain does not return to i.

A detailed proof of Theorem 4 will also be omitted, but again the idea behind it is simple. Let states i and j communicate, let i be recurrent, and suppose the chain starts in state i. The chain will return to i infinitely often. In each excursion between visits to state i, the probability to visit j is some positive number p—if p were equal to 0 it would not be possible to get to j, contrary to assuming $i \sim j$. Again, it's like coin tossing an infinite number of times independently. If the probability of heads is p, heads will come up infinitely often. Likewise, among the infinite number of excursions from i back to itself, an infinite number will include a visit to j. But by Theorem 3 an infinite number of visits to j can occur only if j is recurrent.

Examples 4.3.2. (a) Any absorbing state i, is recurrent, because $\mathbb{P}(T_i = 1 \big| X(0) = i) = 1$.

(b) All the non-absorbing states in the Moran and Wright-Fisher models 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 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) If i is a state in a chain and $p_{ii} < 1$, the chain cannot stay in i forever. For, if $X(0) = i$ and the chain stays in i forever, then for any n, $(X(1), \ldots, X(n)) = (i, i, \ldots, i)$. But the probability of this path, consisting of n transitions from state i to itself is p_{ii}^n. Thus the probability of staying in i forever is less than or equal to p_{ii}^n for all n; but $\lim_{n\to\infty} p_{ii}^n = 0$, so the probability to stay in i forever is zero. (In fact, part e) of Exercise 4.1.2 shows that the first time to leave i is a geometric random variable.)

Consider the two-state chain with $p_{01} = \lambda > 0$ and $p_{10} = \mu > 0$. Then $p_{00} = 1 - \lambda < 1$ and $p_{11} = 1 - \mu <$, and so the chain cannot stay in either state forever. Therefore, both states are recurrent, because, for example, if the chain leaves state 0 for state 1 it must eventually leave state 1 and return to state 0. $\diamond$

4.5.2 Chains with absorbing states

This section concerns finite state-space Markov chains for which an absorbing state is accessible from every non-absorbing state; thus all non-absorbing states are transient. The guiding examples are the Moran and Wright-Fisher models without mutation and the random walk on a finite state space with absorbing boundaries.

Let $\mathcal{E}$ denote the (finite) state space of the Markov chain, and $\mathcal{S}$ the set of its absorbing states, assumed to be non-empty. As usual, $A = [p_{ij}]_{i,j\in\mathcal{E}}$ denotes its state transition matrix. If i is any state and if a is an absorbing state, let

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

denote the probability the chain eventually gets absorbed by state a, given that it starts in state i. This is the limit, as $t \to \infty$ of the probability to be absorbed by state a by time t:

$$g_i^a = \lim_{t\to\infty} \mathbb{P}\Big(X(s) = a \text{ for some } 0 \leq s \leq t \Big| X(0)=i\Big)\,.$$

But since the chain stays in state a once it hits a, $X(s) = a$ for some $0 \leq s \leq t$ if and only if $X(t) = a$. Thus,

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

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

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

Theorem 5 *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 let $\mathcal{T} = \mathcal{E} - \mathcal{S}$ be the set of all non-absorbing states. Assume that for every $i \in \mathcal{T}$, there exists an absorbing state $a \in \mathcal{S}$, such that a is accessible from i—that is, $g_i^a > 0$.*

Then

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

(ii) For every $i \in \mathcal{T}$, the probability that the chain eventually hits a state in $\mathcal{S}$, given that it starts in state i, is one. (The chain must eventually land in an absorbing state.)

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

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

$$g_a^a = 1 \quad \text{and} \quad g_s^a = 0 \quad \text{if } s \in \mathcal{S} \text{ and } s \neq a; \tag{4.35}$$

$$g_i^a = \sum_{j\in\mathcal{E}} p_{ij} g_j^a = \sum_{j\in\mathcal{T}} p_{ij} g_j^a + p_{ia} \tag{4.36}$$

The important qualitative result in this theorem is that a Markov 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 5. Thus, in the absence of mutation, an allele will eventually either take over or disappear from the population, just due to chance, for both models. In population genetics, this is called *fixation*. (There are examples of Markov chains with an infinite number of transient states, in which the state may remain among transient states forever. Thus, the finite state assumption is vital for statement (i) of Theorem 5.)

The important quantitative result is the system of linear equations for absorption probabilities stated in part (iv). They enable one to find the probability with which a chain ends up in a particular absorbing state.

Statement (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 3, we know that the chain can visit a transient state only finitely many times, and, since there are only a finite number of transient states—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 $t \to \infty$.

The equalities in (4.35) are true by definition because if $s \neq a$ and s is absorbing, the chain can never reach a. For each state i, the equation in (4.36) for g_i^a is derived by conditioning on the next state j the chain visits after i and applying the Markov property, as we did to derive (4.34) in analyzing gambler's ruin. Instead of presenting a detailed, general argument, we illustrate by an example.

Example 4.5.3. Absorption probabilities for the neutral Moran Model 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

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

This model allows no mutation. We are interested in computing g_i^4, the probability that the chain eventually hits the absorbing state 4, given that $X(0) = i$.

Let V be the event the chain eventually hits 4; thus, $g_i^4 = \mathbb{P}(V \,|\, X(0) = i)$. Let $0 < i < 4$. Then

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

because, once $X(1) = j$ is given, the Markov property implies $X(0) = 1$ is no longer relevant. If $X(0) = i$, $X(1)$ can only be one of $i-1$, i, or $i+1$, which happen with respective probabilities, $p_{i,i-1}$, p_{ii}, and $p_{i,i+1}$. By the rule of total probabilities,

$$\begin{aligned} g_i^4 &= \mathbb{P}(V \,|\, X(0) = i) \\ &= \mathbb{P}(V \,|\, X(1) = i-1, X(0) = i) p_{i,i-1} + \mathbb{P}(V \,|\, X(1) = i, X(0) = i) p_{ii} \\ &\qquad + \mathbb{P}(V \,|\, X(1) = i+1, X(0) = i) p_{i,i+1} \\ &= g_{i-1}^4 \, p_{i,i-1} + g_i^4 \, p_{ii} + g_{i+1}^4 \, p_{i,i+1} \, . \end{aligned}$$

We shall write out these equations for $i = 1, 2, 3$ using (4.35) of Theorem 5, which says

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

Again, these are true by definition because if $X(0) = 0$, the chain stays forever in state 0 and thus cannot reach 4, and if $X(0) = 4$, the chain is already in state 4 and never leaves. The result is:

$$\begin{aligned} g_1^4 &= p_{10} g_0^4 + p_{11} g_1^4 + p_{12} g_2^4 = (10/16) g_1^4 + (3/16) g_2^4 && (4.37) \\ g_2^4 &= p_{21} g_1^4 + p_{22} g_2^4 + p_{23} g_3^4 = (1/4) g_1^4 + (1/2) g_2^4 + (1/4) g_3^4 && \text{and } (4.38) \\ g_3^4 &= p_{32} g_2^4 + p_{33} g_3^4 + p_{34} g_4^4 = (3/16) g_2^4 + (10/16) g_2^4 + (2/16). && (4.39) \\ & && (4.40) \end{aligned}$$

These are three, linear equations for the three unknowns g_1^4, g_2^4, and g_3^4. The solution is $g_1^4 = 1/4$, $g_2^4 = 1/2$, and $g_3^4 = 3/4$. This is a suggestive result: the probability of absorption into state 4 starting from i alleles of type A is just their initial frequency, $i/4$. The next example shows that this is true in general. $\diamond$

Example 4.5.4 Absorption probabilities for the neutral Wright-Fisher model with no mutation.

Consider the neutral Wright-Fisher model with no mutation for a haploid population of size N. We will show

$$g_i^N = \frac{i}{N} \quad \text{for every } i,\ 0 \le i \le N. \tag{4.41}$$

In words: the probability that allele A eventually takes over the population is equal to its initial frequency in the population. Of course, this is automatic for $i = 0$ and

$i = N$. It remains to treat the other cases. Instead of deriving linear equations for g_i^N, $0 \le i \le N$, which leads to a very complicated looking set of equations, we illustrate a different method. It takes advantage of identity (4.28),

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

which was proved earlier in showing that the expected value of the Wright-Fisher chain is constant from generation to generation. Let us write out $E[X(t)\big|X(0)=i]$ in this equation as a sum using the definition of conditional expectation:

$$i = \sum_{j=0}^{N} j\,\mathbb{P}\big(X(t)=j\big|X(0))=i\big)\,.$$

In this sum, the term corresponding to $j = 0$ is automatically 0. If $1 \le j \le N-1$, then j is transient, and part (iii) of Theorem 5 says that $\lim_{t\to\infty}\mathbb{P}\big(X(t)=j\big|X(0))=i\big) = 0$. On the other hand, $g_i^N = \lim_{t\to\infty}\mathbb{P}(X(t)=N \;\big|\; X(0)=i)$. Thus, by taking the limit of both sides of the previous equation as $t \to \infty$,

$$i = \lim_{t\to\infty}\sum_{j=1}^{N} j\,\mathbb{P}\big(X(t)=j\big|X(0)) = Ng_i^N\,.$$

It follows immediately that $g_i^n = i/N$, as claimed.

Since 0 and N are the only absorbing states, and since Theorem 5 says absorption must take place, $g_i^0 = 1 - g_i^N = 1 - \frac{i}{N}$.

Since (4.28) is true for the general, neutral Moran model without mutation—see Exercise 4.2.4—the exact same argument shows $g_i^N = i/N$ for the Moran model as well. ◇

4.5.3 The ergodic theorem for recurrent Markov Chains

Irreducible, recurrent Markov chains, such as the Moran and Wright-Fisher models with mutation, have no absorbing states and their behavior is much different. For simplicity we focus mostly on the case of a finite state space, $\{1,\dots,N\}$. As usual, A will denote a state transition matrix and the row vector

$$\rho(t) = \Big(\mathbb{P}(X(t)=1), \mathbb{P}(X(t)=2), \dots, \mathbb{P}(X(t)=N)\Big)\,.$$

denotes the law of the process at time t. We will comment on extensions to the infinite state space case as we go along.

Stationary distributions

Consider a Markov chain on $\{1,\dots,N\}$ with transition probability matrix A. If

$$\rho(t) = (\pi_1, \pi_2, \dots, \pi_N) \;\text{ for every } t = 0, 1, \dots,$$

we say that $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ is a stationary distribution for the chain. Thus, although as time progresses the processes itself is moving around randomly from state to state, the probability distribution of $X(t)$ does not change.

Recall that $\rho(t) = \rho(t–1) \cdot A$ and $\rho(t) = \rho(0) \cdot A^t$—see (4.18) and (4.20) in Section 4.2.1. Thus, if π is a stationary distribution,

$$\pi = \rho(1) = \rho(0) \cdot A = \pi \cdot A\,.$$

Conversely, if π defines a probability distribution on $\{1, 2, \ldots, N\}$—that is, $\pi_i \geq 0$ for each i and $\sum_1^N \pi_i = 1$—and satisfies this equation, then it is a stationary distribution. Indeed, if $\pi = \pi \cdot A$, then

$$\pi = \pi A = (\pi A)A = \pi A^2 = (\pi A)A^2 = \pi A^3 = \cdots = \pi A^t \quad \text{for all } t \geq 1.$$

Thus, if $\rho(0) = \pi$, then,

$$\rho(t) = \rho(0) \cdot A^t = \pi \cdot A^4 = \pi\,,$$

and so π is a stationary distribution.

Therefore, a vector $\pi = (\pi_1, \ldots, \pi_N)$ is a stationary distribution for a Markov process with state transition matrix A if and only if

$$\pi = \pi \cdot A \tag{4.42}$$

$$\sum_{i \in \mathcal{E}} \pi_i = 1 \text{ and } \pi_i \geq 0 \text{ for all } i \text{ in } \mathcal{E}. \tag{4.43}$$

If π solves (4.42), so does $c\,\pi$ for any scalar c—simply multiply both sides of (4.42) by c. Therefore (4.43) must be used when employing these equations to find a stationary distribution.

Example 4.5.5. Let s_0 be an absorbing state. If $\mathbb{P}(X(0) = s_0) = 1$, then clearly, $\mathbb{P}(X(t) = s_0) = 1$ for all t. Thus the degenerate probability distribution assigning a probability of one to be in state s_0, is a stationary distribution. If s_1 is a second absorbing state, assigning probability one to s_1 gives a different stationary distribution.

You are asked to show in Exercise 4.5.8 that if π and η are stationary distributions, so is $a\pi + b\eta$, whenever $a, b \geq 0$, and $a + b = 1$. If π is the stationary distribution assigning probability one to absorbing state s_0 and η assigns probability one to absorbing state s_1, $a\pi + b\eta$ is the distribution which assigns probability a to state s_0 and probability b to s_1, as may easily be checked. This example shows stationary distributions are not necessarily unique. ◇

Example 4.5.6. Consider the two state chain with $p_{01} = \lambda$ and $p_{10} = \mu$. Then (4.42) and (4.43) become

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

The first equation is a system of two linear equations: $\pi_0 = (1-\lambda)\pi_0 + \mu\pi_1$, and $\pi_1 = \lambda\pi_0 + (1-\mu)\pi_1$. But both of these reduce to the same equation,

$$\lambda\pi_0 = \mu\pi_1 \,.$$

(We have already pointed out that $\pi = \pi \cdot A$ cannot determine π uniquely.) Thus, to be an invariant distribution, (π_0, π_1) must solve

$$\lambda\pi_0 = \mu\pi_1 \quad \text{and} \quad \pi_0 + \pi_1 = 1 \,.$$

The reader may easily check that when $\lambda + \mu > 0$, there is a unique solution,

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

The concept of stationary distribution generalizes without change to the case of an infinite state spaces, $\mathcal{E}$. A probability distribution, $\pi = \{\pi_i;\, i \in \mathcal{E}\}$ is a stationary distribution for a Markov chain if $\mathbb{P}(X(t)=i) = \pi_i$ for all times t and state i. The algebraic characterization in equations (4.42) and (4.43) also generalizes: $\{\pi_i;\, i \in \mathcal{E}\}$ is a stationary distribution if and only if

$$\pi_j \;=\; \sum_{i\in\mathcal{E}} \pi_j \, p_{ij} \quad \text{for all } j \in \mathcal{E}, \text{ and;} \tag{4.44}$$

$$\sum_{i\in\mathcal{E}} \pi_i \;=\; 1 \;\text{ and } \pi_i \geq 0 \text{ for all } i \in \mathcal{E}. \tag{4.45}$$

Infinite-time limits, stationary distributions and periodicity of states

In Example 4.2.1, we found that when $\lambda + \mu > 0$ in the two-state model, then whatever the initial distribution,

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

which is exactly the unique, stationary distribution we just calculated. 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. The converse question is also interesting. Suppose there is a unique stationary distribution, π. Does $\lim_{t\to\infty} \rho(t)$ exist? The next example shows it may not and why.

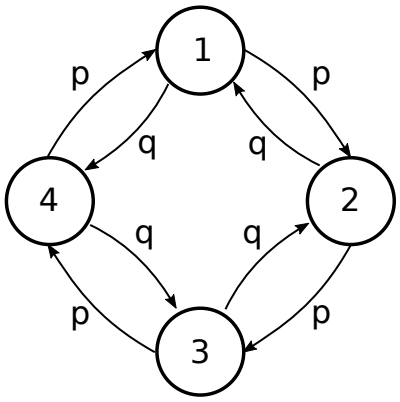


Figure 4.4: A periodic chain.

Example 4.5.7. Consider the Markov chain with the state transition diagram illustrated in Figure 4.4, and assume $p > 0$ and $q > 0$, and, of course, $p + q = 1$. This is a version of simple random walk, but on a circular lattice instead of on the integers. The state transition diagram is symmetric—the chain looks the same from the viewpoint of any state—which suggests that the only stationary distribution, would be the one in which all states are equally likely: $\pi = (0.25, 0.25, 0.25, 0.25)$. It is easy to check that π is in fact the unique stationary measure. Since the components of π add to one, it is only necessary to check that it is the unique solution to (4.42). We leave this easy calculation to the reader: the state transition matrix, A, is

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} = \begin{bmatrix} 0 & p & 0 & q \\ q & 0 & p & 0 \\ 0 & q & 0 & p \\ p & 0 & q & 0 \end{bmatrix} .$$

However, it is not true that $\lim_{t\to\infty} \rho(t)$ exists for *any* initial distribution. Since the chain must move one state clockwise or counterclockwise at each step and since there are four states, it can only return to its starting point after an even number of steps. You can convince yourself of this by trying a few paths. Thus $\mathbb{P}(X(2s+1) = 1 \mid X(0) = 1) = 0$ for every nonnegative integer s. On the other hand and $\mathbb{P}(X(2s) = 1 \mid X(0) = 1) > 0$ for every nonnegative integer s, and in fact one can show,

$$\lim_{s\to\infty} \mathbb{P}(X(2s)=1 \mid X(0)=1) = 0.5 .$$

(This is not immediately obvious; Exercise 4.5.8 explains why.) Since the limit of $\mathbb{P}(X(t)=1 \mid X(0)=1)$ as t tends to infinity through even times differs from the limit through odd times, $\lim_{t\to\infty} \mathbb{P}(X(t)=1 \mid X(0)=1)$ does not exist. ◇

Because the chain of Figure 4.4 can return to its starting point with positive probability for any even number of time steps, but not for any odd number of

time steps, its states are said to be *periodic* with period 2. This definition can be generalized. Let $\{X(t)\}$ be a Markov chain and let i be one of its states. The *period* of i is the greatest common divisor of the set of all times t such that $\mathbb{P}(X(t) = i|X(0) = i) > 0$. Put in another way, given $X(0) = i$, the period of i is the largest number d such that the return time T_i takes values in the set $\{d, 2d, 3d, \dots\}$.

A state whose period is 1 is said to be *aperiodic.* It is clear that if $p_{ii} = \mathbb{P}(X(1) = i\big|X(0) = i) > 0$, 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 the models we encounter.

The reader should be aware of the facts stated in the following theorem. Their proofs are omitted.

Theorem 6 *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. This condition is also necessary when the state space is finite.

The Ergodic Theorem

This is the most important theorem in the theory of Markov chains. It addresses the existence, uniqueness, and interpretation of stationary distributions for irreducible, recurrent Markov chains. In this theory, the mean return time

$$E[T_i \mid X(0) = i],$$

the expected time it takes for the chain to return to state i when it starts from i, plays a role. If we look at the mean return times we calculated for the two-state chain when $\lambda > 0$ and $\mu > 0$, we find a connection to its stationary distribution. Recall from Example 4.4.2 that

$$E[T_0\big|X(0) = 0] = \frac{\lambda + \mu}{\mu} \quad \text{and} E[T_1\big|X(0) = 1] = \frac{\lambda + \mu}{\lambda}.$$

Thus

$$\left(\frac{1}{E[T_0\big|X(0) = 0]}, \frac{1}{E[T_1\big|X(0) = 1]}\right) = \left(\frac{\mu}{\lambda + \mu}, \frac{\lambda}{\lambda + \mu}\right),$$

which is the invariant distribution of the two-state chain—see Example 4.5.6. We shall see that this is no accident but reflects a general fact relating mean return times and stationary distributions. It is part of the main theorem, the ergodic theorem for Markov chains, which follows next.

Theorem 7 *Consider an irreducible, recurrent Markov chain evolving in a finite state space, with state transition matrix A.*

(a) The chain admits a unique stationary distribution π;

(b) The stationary distribution, π, satisfies:

$$\pi_j = \frac{1}{E[T_j \big| X(0)=j]} \qquad \textit{for each state } j, \textit{ and;} \tag{4.46}$$

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

$$\pi_j = \lim_{t\to\infty} \frac{\textit{number of visits of } X(1), X(2), \ldots, X(t) \textit{ to state } j}{t} \tag{4.47}$$

with probability one;

(c) Suppose, in addition, that the chain is aperiodic. Then

$$\lim_{t\to\infty} [A^t]_{ij} = \pi_j, \qquad \textit{for all states } i \textit{ and } j\,, \tag{4.48}$$

and it follows that no matter what the initial law of $X(t)$ is

$$\lim_{t\to\infty} \rho(t) = \pi\,. \tag{4.49}$$

Remarks.

1. The conclusions drawn in (4.46) and (4.47) are especially interesting because they relate the stationary distribution to the long-run behavior of paths. In particular, (4.46) generalizes the relationship between stationary distribution and mean return times we discovered for two-state chains. What these results say is very intuitive. Since $E[T_j \big| X(0)=j]$ is the average time it takes between visits to state j, one expects the chain to spend

$$\frac{1}{E[T_j \big| X(0)=j]}$$

of its time in state j, and this fraction is the same as π_j. Identity (4.47) basically says the same thing, only more directly. In fact, it is a generalization to Markov chains of the law of large numbers, because it equates the long run average number of visits to a state j with 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 7.

2. In matrix form, (4.48) is

$$\lim_{t\to\infty} A^t = \begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \cdots & \pi_N \\ \pi_0 & \pi_1 & \pi_2 & \cdots & \pi_N \\ \vdots & \vdots & \vdots & & \vdots \\ \pi_0 & \pi_1 & \pi_2 & \cdots & \pi_N \end{pmatrix}. \tag{4.50}$$

This is a useful way to remember (4.48). It makes it visually clear that $\lim_{t\to\infty} \mathbb{P}\Big(X(t)=j \big| X(0)=i\Big) = \lim_{t\to\infty}[A^t]_{ij} = \pi_j$ is independent of the starting state, i in the aperiodic case. And it suggests a convenient numerical method for approximating the

stationary distribution: compute higher and higher powers of A until the entries converge to the decimal accuracy desired for the approximation.

3. One consequence of (4.46) is that the mean return times of all states in an irreducible recurrent chain are finite. This seems obvious because there are only a finite number of states to visit. But for general chains with an infinite state space, it is possible that a state is recurrent, so $\mathbb{P}(T_i < \infty \big| X(0) = i) = 1$, yet $E[T_i \big| X(0) = i] = \infty$. For example this is true for all states of symmetric random walk, although this requires some work (omitted) to show. The possibility of infinite mean return times has motivated a definition: a recurrent state i such that $E[T_i \big| X(0) = i] < \infty$ is said to be positive recurrent, while if $E[T_i \big| X(0) = i] = \infty$ it is said to be null recurrent. It turns out that if $i \sim j$ and states i and j are recurrent, they are either both positive recurrent or both null recurrent. It turns out that positive recurrence is the key to extending the ergodic theorem to the infinite state space case.

Theorem 8 *All the statements of Theorem 7 hold for an irreducible chain with an infinite (discrete) state space if its states are positive recurrent. When the states of a chain all communicate and are all null recurrent, a stationary distribution does not exist and* $\lim_{t\to\infty}[A^t]_{ij} = 0$ *for all* j.

We have illustrated all the claims of Theorem 7 for the two-state chain, except for statement (4.47). A proof of Theorem 7 for the general case is beyond the scope of this text and is omitted. We only note that (4.49) is a consequence of (4.48). Indeed, from formula (4.20), 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 (4.49).

Example 4.5.6. Consider the Moran model with mutation as defined in Example 4.6. Assume that $N = 3$ and $u = v = 1/4$. Then the state 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, all states communicate, and hence the chain is positive recurrent and Theorem 7 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 7 give us additional information. For example, the average time it takes the chain to return to state 0 given that it starts there is

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

Detailed balance equations and stationary distributions

In some circumstances, the stationary distribution can be found using a system of simpler linear equations called the *detailed balance equations.* Let A be a state transition matrix on a state space $\mathcal{E}$. We say that a probability distribution $\pi = \{\pi_i, ; i \in \mathcal{E}\}$ satisfies the detailed balance equations if

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \text{for all states } i \text{ and } j.$$

A solution to the detailed balance equations, provided it defines a probability distribution, is a stationary distribution. This is easy to see by summing both sides of the detailed balance equations in the index j and using $\sum_{j=0}^{N} p_{ij} = 1$: for each i

$$\pi_i = \sum_{j=0}^{N} \pi_i p_{ij} = \sum_{j=0}^{N} \pi_j p_{ji},$$

which is precisely the equation defining a stationary distribution.

If you are trying to calculate a stationary distribution, it is a good idea to start with the detailed balance equations, because of their simplicity. *However, in many cases the stationary distribution does not satisfy the detailed balance equations, which therefore will not have a solution.* Then one needs to return to the original equation, $\pi = \pi \cdot A$.

Birth-and-death chains, which include Moran models with mutation, are one class for which the detailed balance approach works. Recall from Example 4.1.7 the transition matrix for a general birth-and-death chain:

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

Thus, for $1 \le i \le N-1$, $p_{i,i-1} = q_i$, $p_{ii} = r_i$, $p_{i,i+1} = p_i$. 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

$$\pi_i p_i = \pi_{i-1} q_{i+1}, \qquad 0 \le i \le n-1.$$

This means 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:

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

Therefore

$$\pi_0 = \left[1 + \sum_{1}^{N} \prod_{k=0}^{i-1} \frac{p_k}{q_{k+1}}\right]^{-1}.$$

and for $i \ge 1$,

$$\pi_i = \left[1 + \sum_{1}^{N} \prod_{k=0}^{i-1} \frac{p_k}{q_{k+1}}\right]^{-1} \prod_{k=0}^{i-1} \frac{p_k}{q_{k+1}}. \tag{4.51}$$

Numerical Approximation of the stationary distribution

Often, one cannot find the stationary distribution explicitly. In such cases, a numerical approximation can be calculated by applying standard, numerical linear algebra software to solve the system, $\pi = \pi \cdot A$ and $\pi_1 + \cdots + \pi_N = 1$. For example, one can use Gaussian elimination to solve $\pi = \pi \cdot A$ up to an unknown constant, and then use $\pi_1 + \cdots + \pi_N = 1$ to determine that constant. A discussion of other algorithms, and ways to streamline them for Markov chain calculations, can be found, for instance, in the book by William J. Stewart, *Probability, Markov Chains, Queues, and Simulation*, Princeton University Press (2009). We shall not pursue these methods here.

There is another method which is easy to implement. Assume the chain is aperiodic and all states communicate. By Remark 2 after the statement of Theorem 7, A^t converges to a matrix, each of whose rows is the stationary distribution. Thus, to approximate π, we can simply calculate higher and higher powers of A^t, stopping when the entries stabilize to the degree of accuracy we desire. One has to be careful about round-off error, which can build up with successive powers. But the final approximation $\bar{\pi}$ can always be checked easily, simply by comparing $\bar{\pi}$ to $\bar{\pi} \cdot A$.

To illustrate, consider the state transition matrix, A, of Example 4.5.6. Using the Maple software package, and calculating successively, $B_1 = A^4$, $B_2 = B_1^4 = A^{16}$, and finally $B_3 = B_2^4 = A^{64}$, one obtains (in Maple), after rounding to 6 decimal places,

$$A^{64} = \begin{bmatrix} 0.218754 & 0.281252 & 0.281248 & 0.218746 \\ 0.218751 & 0.281251 & 0.281249 & 0.218749 \\ 0.218749 & 0.281249 & 0.281251 & 0.218751 \\ 0.218746 & 0.281248 & 0.281252 & 0.218754 \end{bmatrix},$$

and A^{65} agrees with this to 5 decimal places. (The rows sum exactly to one; this will not always be the case because of round-off error.) Compare this to the exact stationary distribution computed in Example 4.5.6, which is

$$\big(0.21875, 0.28125, 0.28125, 0.21875\big),$$

with no roundoff error. This agrees to 5 decimal places with the rows of A^{64}. (Of course if we used the numerical method we would not know rounding to 5 decimal places gives the exact answer unless we checked $\pi = \pi \cdot A$ by exact calculation.)

4.5.4 Exercises

Exercise 4.5.1. 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 4.5.2. 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 $E[g(X(t)) \mid X(0) = i] = g(i)$ for all i, $0 \leq i \leq t$, and $t \geq 1$.

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

Exercise 4.5.3 Consider the Moran model with mutation as defined in Exercise 4.1.6. Assume that $N = 3$ and $u = v = 1/4$. Then the state 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 4.5.4. 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 4.5.5. 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 4.5.6. (a) Use the method of detailed balance equations to re-derive the invariant density for the chain of Example 4.5.6.

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 4.5.7. This problem requires numerical software: Maple or Matlab will work fine. Let A be the transition matrix of the Moran model in problem 4.5.5. 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.

Exercise 4.5.8. Show that if π and η are stationary distributions of a Markov chain, then so is $a\pi + b\eta$ for any non-negative a and b such that $a + b = 1$.

Exercise 4.5.8. For the periodic Markov chain of Example 4.5.7 it was claimed that

$$\lim_{s\to\infty} \mathbb{P}(X(2s)=1 \,\big|\, X(0)=1) = 0.5\,.$$

Show this by showing treating $Y(s) = X(2s)$ as a Markov chain—see the result of Exercise 4.4.2—calculating its state transition matrix, and applying Theorem 7.

4.6 Stationary Distributions for the Moran and Wright-Fisher Models with Mutation

4.6.1 Moran and Wright-Fisher with mutation

The Moran and Wright-Fisher models with mutation were defined in Exercises 4.1.6 and 4.1.7. There are three parameters in both models: the population size N, the probability u that allele A mutates to a, and the probability v that allele a mutates to A. In this section, the Wright-Fisher model is be the haploid model, for which N is the size of the allele pool, not the number of diploid individuals. In this section we will study the stationary distributions of these models numerically and analytically.

Up to now, the state in each model has been the *number* of alleles of type A in the population. Since allele frequency is the real quantity of interest in population genetics, and is a measure independent of population size, and since we wish to compare models with different population sizes, we shall use allele frequency as the state. To emphasize this difference, we use the notation $Y(t)$, for allele frequency, instead of the notation $X(t)$ used before for allele number. Thus, if $X(t)$ denotes the number of type A's in the Moran model or Wright-Fisher model for a population of size N, $Y(t) = X(t)/N$. The state transition probabilities between frequencies follow easily from the formulas for transition probabilities between allele counts stated in Examples 4.1.6 and 4.1.7. The non-zero transition probabilities between frequencies in the Moran chain are:

$$\begin{aligned}
p_{i/N,i+1/N} &= \frac{i(N-i)}{N^2}(1-u) + \frac{(N-i)^2}{N^2}\cdot v \\
p_{i/N,i-1/N} &= \frac{i(N-i)}{N^2}(1-v) + \frac{i^2}{N^2}\cdot u \\
p_{i/N,i/N} &= \frac{i(N-i)}{N^2}(u+v) + \frac{(N-i)^2}{N^2}(1-v) + \frac{i^2}{N^2}(1-u)
\end{aligned} \tag{4.52}$$

The Wright-Fisher transition probabilities are:

$$p_{i/N,j/N} = \binom{N}{j} r_i^j \,(1-r_i)^{N-j}\,, \quad \text{where } r_i = \frac{i}{N}(1-u) + \left(1-\frac{i}{N}\right) v. \tag{4.53}$$

These formulas are easier to write down and read when generic frequencies are denoted by by x or y. For example, when x replaces i/N, the Moran transition probabilities up and down are:

$$p_{x,x+1/N} = x(1-x)(1-u) + (1-x)^2 v \quad \text{and} \quad p_{x,x+1/N} = x(1-x)(1-v) + x^2 u,$$

Similarly, the Wright-Fisher transition probability from frequency x to frequency y is

$$p_{x,y} = \binom{N}{Ny} \big(x(1-u) + (1-x)v\big)^{Ny} \big(xu + (1-x)(1-v)\big)^{N(1-y)}.$$

When both $u > 0$ and $v > 0$, neither chain has absorbing states and so both are irreducible. Both chains are also aperiodic.

Mutation exerts pressure on the allele frequency to move in one direction or the other. We can measure this effect at different frequencies by the *expected displacement function*

$$E\big[Y(t+1)\big|Y(t)=x\big] - x.$$

When no mutation is present, the expected displacement is always zero for both Moran and Wright-Fisher chains. We showed this in Example 4.2.3 and Exercise 4.2.4, and used it to conclude that expected allele frequency does not change with time when there is no mutation. When there is mutation, displacement is no longer non-zero. Consider first the Wright-Fisher chain. If $Y(t) = x$, then $Y(t+1) = X(t{+}1)/N$, where $X(t{+}1)$ is a Bernoulli random variable with parameters $n = N$ and $p = x(1-u)+(1-x)xv$. Thus $E[X(t{+}1)|Y(t){=}x]$ equals $Np = N[x(1-u)+(1-x)v]$, and the expected displacement is

$$E\Big[Y(t+1)\Big|Y(t){=}x\Big] - x = x(1-u) + (1-x)v - x = (1-x)v - xu, \tag{4.54}$$

For the Moran model,

$$\begin{aligned} E\big[Y(t+1)\big|Y(t){=}x\big] - x &= (x - \frac{1}{N})p_{x,x-1/N} + xp_{xx} + (x+\frac{1}{N})p_{x,x+1} - x \\ &= \frac{p_{x,x+1} - p_{x,x-1/N}}{N} = \frac{(1-x)v - xu}{N}. \end{aligned} \tag{4.55}$$

Remarkably, aside from the factor $1/N$, this is the same as for the Wright-Fisher chain. But remember that time scales differ between models. In the Wright-Fisher case, time is measured in generations, so $(1-x)v - xu$ is the displacement *per generation*. In the Moran case, time is measured by the number of individual birth and death events. It takes N of these events to create N new individuals, and hence N units of time may be considered the equivalent of one generation. When this change of time scaled is factored out, the expected displacements per generation are the same for both chains.

4.6.2 The stationary distribution problem

How do allele frequencies evolve as $t \to \infty$ in these models? If u and v are strictly positive, then Theorem 7 applies, and the answer is given by the unique stationary distributions of the chains. What do these distributions look like and how do their shapes depend on U and V?

This is a challenging mathematical problem. There is not an explicit formula for the Wright-Fisher stationary distribution, at least that I know. There is an explicit formula for the Moran model, because it is a birth and death type chain; but it is so complicated that it provides little insight.

Often, simple approximate invariant distributions emerge from studying an appropriate limiting case, in this case, as $N \to \infty$. An analysis of expected displacement suggests how this limit should be set up. In real populations, allele frequencies do not change by large or erratic jumps over time scales of generations. To enforce this, the expected displacement of the Wright-Fisher for a population of size N should be of the order of $1/N$, which is the the spacing between adjacent frequencies in the state space. By formula (4.54), this will be true for the Wright-Fisher model u and v are of order $1/N$. To this end we will fix U and V, and study the Wright-Fisher model with mutation probabilities

$$u_N = \frac{U}{N} \quad \text{and} \quad v_N = \frac{V}{N}\,,$$

as the population size N increases. Then the model will remain biologically meaningful as for all N. The same scaling will also be used for the Moran model.

Remarkably, it turns out than when the mutation probabilities are scaled in this way, the stationary distributions of both the Wright-Fisher and Moran models converge to beta distributions depending on U and V. Moreover, the approach to the limit is typically rapid, so that limit distribution is usually a good approximation even for relatively small N. We will illustrate this by numerical experiments in the next section, and then formulate a precise mathematical statement of the limit result.

That an interesting limit exists is perhaps counter-intuitive. After all, we introduced stochastic models precisely for the purpose of treating finite populations. Won't taking the population size to infinity lead right back to infinite population models that evolve deterministically? The answer is no for a subtle reason. If we let $N \to \infty$ while looking just one step ahead into the future, the limit is in fact the infinite population model with mutation. But for each fixed N, we are taking $t \to \infty$, allowing the process to settle into a stationary distribution first. This stationary distribution has a meaningful limit, different from the infinite population model of Chapter 3, as $N \to \infty$.

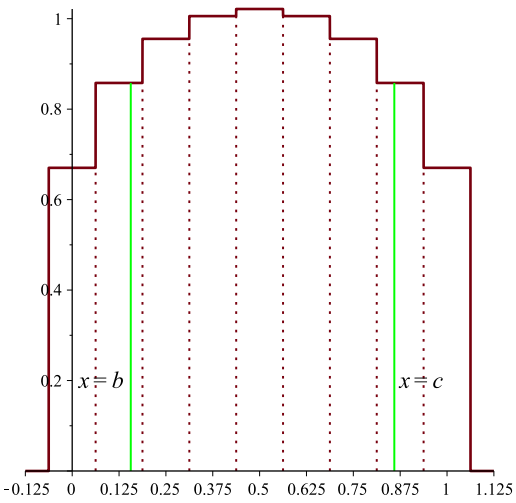


Figure 4.5: Representing distributions of frequencies

4.6.3 Numerical experiments on stationary distributions

The state space of the Moran model when $N = 8$ is the set of frequencies 0, 1/8, 2/8, $\ldots$, 7/8, and 1; when $N = 10$ it is a different set, namely, $0, 1/10, 2/10, \ldots, 9/10, 1$, To compare one model to the other, we need a way to compare probability distributions defined on different discrete subsets of $[0, 1]$. The solution is to use a bar graph that represents discrete probability distributions by areas under a curve; the function defining the bar graph is effectively a probability density that approximates the discrete distribution. For more on approximation of continuous random variables by discrete random variables, see Chapter 2.

It is easiest to explain the method by example. For the Moran Model with $N = 8$ and $u = v = 1/8$ (hence $\mathcal{U} = \mathcal{V} = 1$), the invariant distribution, rounded to three decimal places, is given by the vector,

$$\pi = \Big[0.084, 0.107, 0.119, 0.126, 0.128, 0.126, 0.119, 0.107, 0.084\Big].$$

This means that $\pi_0 = 0.084$ is the probability under the stationary distribution that the frequency of A is 0, $\pi_{1/8} = 0.107$ is the probability it is 1/8, $\pi_{2/8} = 0.119$ is the probability it is 2/8, etc. (The symmetry in π is due to the equality of mutation rates.) To represent this probability distribution graphically, construct over each state $i/8$ a rectangle of width 1/8, centered at $i/8$ whose area is $\pi_{i/8}$. To do this, simply set the height of this rectangle equal to $8 \cdot \pi_{1/8}$. Figure 4.1 below displays the result of this construction. Let f denote the function defined by the upper boundary of these rectangles; its graph is the solid curve of the figure.

By virtue of this graph, we can understand the discrete distribution, π in terms of areas under f. In effect, f is a probability *density* function approximating the

discrete distribution defined by π. For example, the probability assigned by π to frequencies between the values b and c in Figure 4.1. is the sum of the areas of the rectangles centered on the frequencies, $\{1/4, 3/8, 1/2, 5/8, 3/4\}$, lying between b and c. The probability assigned by f is $\int_b^c f(y)\,dy$, which gives approximately the same value In this case, the integral overestimates because the area it computes includes pieces of rectangles that belong to frequencies adjacent to, but not inside, $[b, c]$. For other choices of b and c, the integral might underestimate the probability. In any case, it is not hard to see that the error is never more than the average area of the two rectangles in which b and c actually reside, that is, never more than $\dfrac{f(b) + f(c)}{16}$. The approximation is rough in this figure because $N = 8$ is not very large. Nevertheless f captures qualitatively how the discrete distribution π assigns probability mass over its state space.

The general construction is a simple extension. Let π be any probability distribution on the points $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}$ of the unit interval. For each point $\frac{i}{N}$, place a rectangle of width $\frac{1}{N}$ and height $N\pi_{i/N}$, centered on the point itself. The upper boundary of these rectangles defines a function, f, which serves as a probability density approximating the distribution defined by π; if (b, c) is an interval which is large compared to the spacing $1/N$, then $\displaystyle\int_b^c f(y)\,dy$ approximates the probability π assigns to the points i/N between b and c, and the error of the approximation is less than or equal to

$$\frac{f(b) + f(c)}{N}.$$

For applications in which $f(b)$ and $f(c)$ remain below a finite bound, this error tends to 0 as N gets larger and larger. Even for moderately large N, f gives a much clearer picture of the distribution defined by π, than writing out π as a vector.

Now it is easy to compare distributions defined for different values of N. Simply superimpose the rectangular approximations on the same graph. We will use this method to experiment numerically with stationary distributions of the Moran chain and Wright-Fisher chains. Let $f_N(x; U, V)$ denote the approximating density (the upper boundary of the rectangular region) of the invariant distribution for the Moran model, when the N is the population size and the mutation probabilities are

$$u_N = \frac{U}{N} \quad \text{and} \quad v_N = \frac{V}{N}.$$

In keeping with the strategy explained in the previous section, we want to study what happens as N increases while U and V are held fixed. The next few figures show results for selected cases. Figures 4.2 and 4.3 treat two cases with equal mutation probabilities. In Figure 4.2 , $U = V = 1$, and graphs of of $f_N(x; 1, 1)$ are plotted as N increases through the values $N = 8$ (red) (the case of Figure 4.1), $N = 20$ (blue), $N = 60$ (green) and $N = 120$ (black). (If your version of

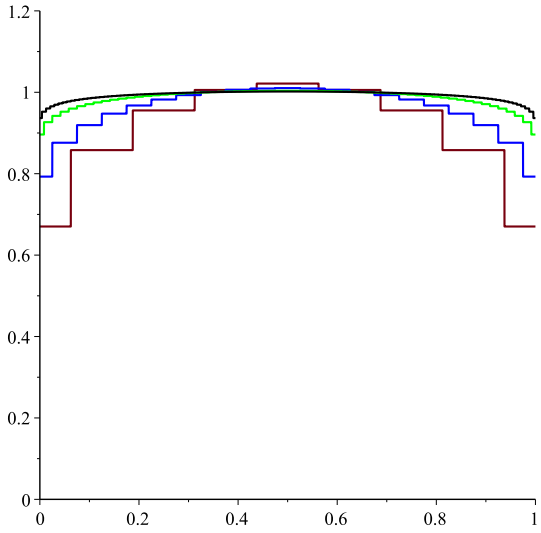


Figure 4.6: $f_N(x;,1,1)$ for $N = 8$, 20, 60, and 120.

the manuscript does not show color, note that the smaller the piecewise constant portions of the graph are, the larger N is.) In Figure 4.2, $U = V = 0.5$, and graphs of $f_N(x;,0.5,0.5)$ are plotted for $N = 20$ (red), $N = 40$ (blue), and $N = 80$ (green). Finally Figure 4.4 treats the non-symmetric case, $U = 2$, and $V = 0.5$. In displaying the plots, I have graphed the functions $f_N(x;U,V)$ only over the interval $[0,1]$ for clarity. As in Figure 4.1, the rectangles centered at 0 and 1 actually stick out a distance $1/(N)$ beyond $[0,1]$ on both sides. This is a minor nuisance because the the discrete distributions are always restricted to $[0,1]$, but is not relevant to the limiting picture. These figures were prepared using the Maple software package. The calculations were based on the explicit formula for the stationary distribution of a discrete-time birth and death chain given in equation (4.51).

Figures 4.5 and 4.6 illustrate calculations for the Wright-Fisher chain. In these figures, $g_N(x;U,V)$ denotes the approximating density for the stationary distribution of a Wright-Fisher chain, when N is the population size (and thus there are N alleles), and

$$u_n = \frac{U}{N} \quad \text{and} \quad v_N = \frac{V}{N}$$

are the mutation probabilities. In Figure 4.5, $U = V = 0.05$ and graphs for $g_N(x;1,1)$ for $N = 16$ (red), $N = 32$ (blue), and $N = 48$ are plotted. In Figure 4.6, $U = 1$, $V = 0.25$, and the approximate density is plotted also for $N = 8$, $N = 16$, and $N = 24$. Because there is no explicit formula for the invariant density of the Wright-Fisher chain, a linear equation solver in Maple was used to calculate these graphs.

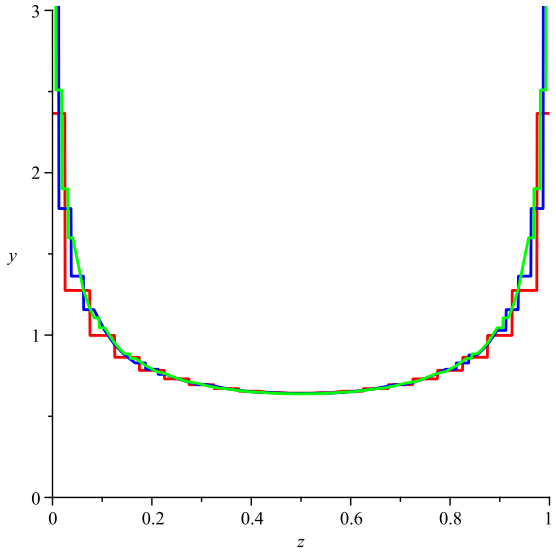


Figure 4.7: Graphs of $f_N(x;,0.5,0.5)$ for $N = 20$, $N = 40$, and $N = 80$.

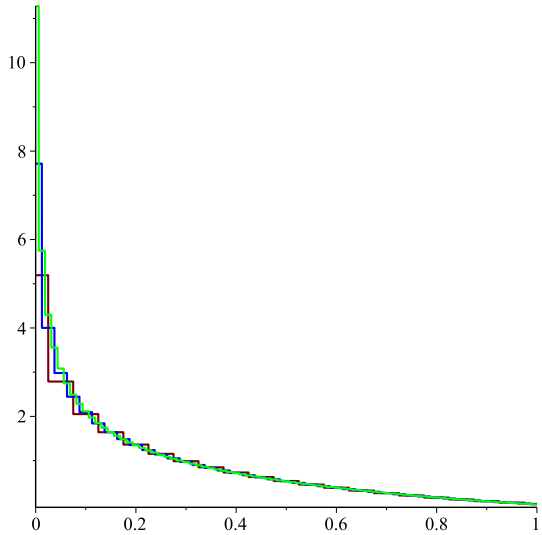


Figure 4.8: Graphs of $f_N(x;,2,0.5)$ for $N = 20$, $N = 40$, and $N = 80$.

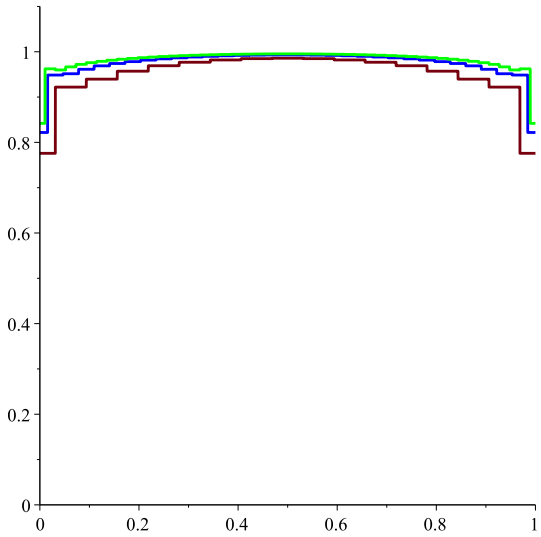


Figure 4.9: $g_N(x; 0.5, 0.5)$ for $N = 16$, $N = 32$, and $N = 48$.

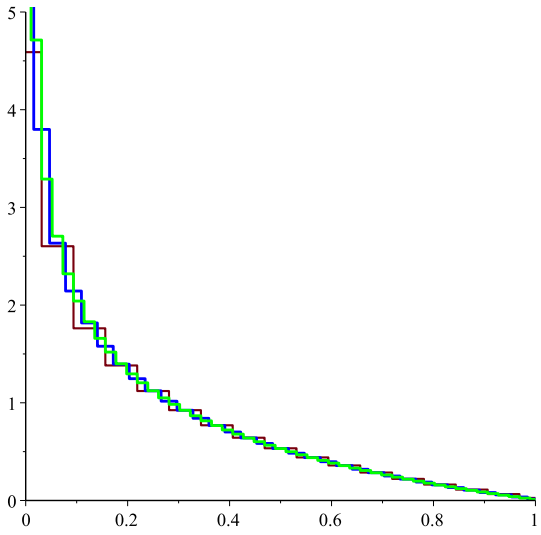


Figure 4.10: $g_N(x; 1, 0.25)$ for $N = 16$, $N = 32$, and $N = 48$.

The results in all cases are striking and resoundingly vindicate the approach of scaling the mutation probabilities by dividing by the population size. The graphs seem to settle down to a simple limit that is a good approximation to the invariant distributions even for relatively small values of N. The distributions of Figures 4.2 (Moran with $U = V = 1$) and Figure 4.5 (Wright-Fisher with $U = V = 0.5$) are approximately uniform over $[0, 1]$, and seem to be approaching the uniform distribution exactly as N increases. As we shall see, it is not an accident that the same limit occurs in both cases when Moran model mutation probabilities are twice those of the Wright-Fisher.

In Figure 4.3 (Moran), the symmetric mutation probabilities are smaller by one-half than in Figure 4.2. On the numerical evidence, a limit again exists. It is symmetric, since $U = V$, but no longer uniform, and assigns relatively higher probabilities to frequencies closer to 0 and 1. This makes sense, because the smaller mutation probabilities exert less displacement pressure, so when the chain enters states of relatively low or relatively high frequencies, it will tend to stay in these states longer.

The mutation probabilities for the Moran model in Figure 4.4 are again twice those of the Wright-Fisher model in Figure 4.6. shows that the existence of a limit does not require symmetric mutation probabilities. Again, limits appear to exist in both cases, convergence to the limit is rapid, and the limits appear to be the same. At the scale of the figure, the graphs for the relatively small values of N shown are almost indistinguishable over much of $(0, 1)$. The shape of the limit makes intuitive sense. In these examples, the chance A mutates to a is four times greater than the reverse. If the process happens to move into a region of high frequency of A, it will move out of it by mutation relatively rapidly and tend to stay in the regions of low value.

From the numerical experiments it appears that for any positive U and V,

$$g(x; U, V) = \lim_{N\to\infty} g_N(x; U, V) \quad \text{and} \quad f(x; U, V) := \lim_{N\to\infty} f_N(x; U, V)$$

exist for all $x \in (0, 1)$. The implication for invariant distributions may be stated as follows. For definiteness, consider the Wright-Fisher case. Let Y_N be the random frequency of the Wright-Fisher chain in stationarity when N is the population size; that is, the probability distribution of Y_N is the stationary distribution. when the the mutation probabilities are $u = U/N$ and $v = V/N$. Then for any $0 \le a < b \le 1$,

$$\lim_{N\to\infty} \mathbb{P}\Big(a \le Y_N \le b\Big) = \int_a^b g(x; U, V)\, dx. \tag{4.56}$$

In the language of probability theory, Y_N converges in distribution to a random variable with probability density $g(x; U, V)$.

A theorem can be proved verifying the picture suggested by the numerical calculations we have presented and giving an explicit and simple formula for the limiting densities.

Theorem 9 *For $U > 0$ and $V > 0$, and $0 < x < 1$*

$$g(x : U, V) := \lim_{N\to\infty} g_N(x; U, V) = \frac{\Gamma(2(U+V))}{\Gamma(2U)\Gamma(2V)}\, x^{2V-1}(1-x)^{2U-1}\,. \tag{4.57}$$

In addition,

$$\lim_{N\to\infty} f_N(x; U, V) = g(x; U/2, V/2)\,. \tag{4.58}$$

For x outside $(0, 1)$, the density is of course equal to zero. In the definition of $g(x; U, V)$, $\Gamma(\nu) = \int_0^\infty x^{\nu-1}e^{-x}\,dx$ is the Gamma function; the ratio of Gamma functions in (4.57) is precisely the constant needed to make $g(x; U, V)$ integrate to one.

The proof of this theorem is technical and advanced. Roughly, the strategy is to show that $g(x; U, V)$ should satisfy the differential equation,

$$\frac{1}{2}\frac{d^2}{dx^2}\Big[x(1-x)g(x)\Big] - \frac{d}{dx}\Big[((1-x)V - xU)g(x)\Big] = 0, \tag{4.59}$$

and then to solve this equation. The reader might wish to check, by direct substitution, that the function given in (4.57) is indeed a solution.

The terms in the differential equation (4.59) have physical interpretations. Recall the expected displacement function, $E[Y(t+1)|Y(t) = x] - x$, which was defined and calculated for the Wright-Fisher chain with mutation in equation (4.54). When the mutation rates are $u_N = U/N$ and $v_N = V/N$, where N is the population size, it follows from (4.54) that $N(E[Y(t+1)|Y(t) = x] - x) = (1-x)V - xU$. This is the origin and meaning of the factor, $(1-x)V - xU$, in the first derivative term of (4.59). The variance of the displacement may also be calculated easily and it is,

$$N\mathrm{Var}\big(Y(t+1)|Y(t)=x\big) = x(1-x) + \text{ terms of order } 1\,.$$

Dividing by N and letting $N \to \infty$, this converges to $x(1-x)$ which is the factor in the term $\frac{1}{2}\frac{d^2}{dx^2}\Big[x(1-x)g(x)\Big]$ in (4.59). Thus the shape of the stationary distribution is determined in the by the scaled expected displacement and the scaled variance of the displacement.

What happens if we formally compute the expected displacement and its variance for the Moran model and use them in (4.59) instead? We saw in (4.55) that the expected displacement function for the Moran model for the Moran model with the same mutation rates. $N^2(E[Y(t+1)|Y(t) = x] - x) = (1-x)V - xU$, differing from the Wright-Fisher case only in an additional factor of N. It can be shown additionally that for the Moran model $N^2\mathrm{Var}(Y(t+1)|Y(t)=x) = 2x(1-x) + \text{terms of order 1}$. This differs from the Wright-Fisher case by a factor of 2 and suggests correctly that the equation for $f(x; U, V)$ should be,

$$\frac{1}{2}\frac{d^2}{dx^2}\Big[2x(1-x)f(x)\Big] - \frac{d}{dx}\Big[((1-x)V - xU)f(x)\Big] = 0\,.$$

Dividing by 2 this, becomes

$$\frac{1}{2}\frac{d^2}{dx^2}\Big[x(1-x)g(x)\Big] - \frac{d}{dx}\Big[((1-x)(V/2) - x(U/2))f(x)\Big] = 0$$

which is the equation (4.59) with U and V replaced by $U/2$ and $V/2$, respectively. This analysis thus suggests that $f(x;U,V) = g(x;U/2,V/2)$, and Theorem 9 asserts that this is indeed the case.

In conclusion, if Y_N is the frequency of a Wright-Fisher or Moran model with mutation after it has evolved a long time, it is approximately a beta random variable. The numerical results show that this approximation can be very good even when N is only of moderate size.

It is instructive to calculate the limiting densities for the cases studied in Figures 4.2—4.6. For Figures 4.2 and 4.5, the limit, from formula (4.57) is the function,

$$f(x;1,1) = g(x;,0.5,0.5) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}x^{2(.5)-1)}(1-x)^{2(.5)-1)} = 1, \quad 0 < x < 1\,.$$

(Note: $\Gamma(n) = (n-1)!$ for positive integers n.) Just as we expected from the graphs, this is the density for the uniform distribution on $(0,1)$. For Figure 4.4 and 4.6, the limit is

$$f(x;2,0.5) = g(x;1,0.25) = \frac{\Gamma(2.5)}{\Gamma(2)\Gamma(0.5)}x^{-1/2}(1-x) = (3/4)\cdot x^{-1/2}(1-x).$$

(The constant $3/4$ can be derived by requiring $g(x;1,0.25)$ to integrate to one.) To illustrate the validity of formula (4.57) and how well it works, we graph $g(x;1.0.25)$ and $g_{24}(x;1,0.25)$ together in Figure 4.7. The graphs match exceptionally well from roughly $x = 0.2$ to $x = 1$ and only differ significantly very close to 0.

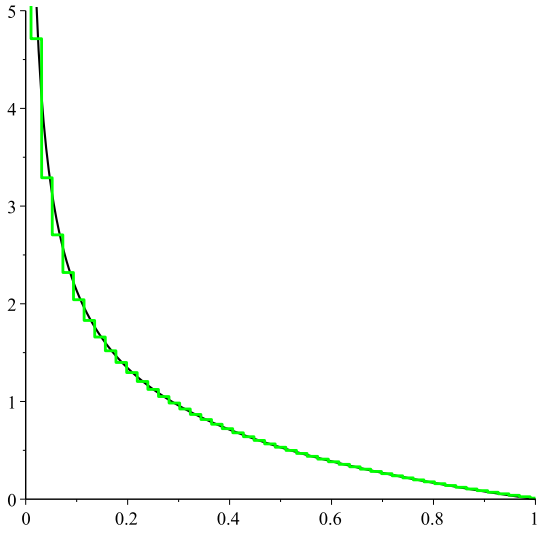


Figure 4.11: $g_{24}(x; 1, 0.25)$ compared to $g(x; 1, 0.25) = (3/4) \cdot x^{-1/2}(1 - x)$.