CHAPTER 2

Discrete time Markov chains

Starting from a very simple model of daily precipitation, we build some theory for discrete time Markov chains. We develop some estimation and testing theory. and look at the goodness of fit of this simple model in different ways. A parsimonious model for higher order of dependence is applied to some meteorological data. Harmonic functions are introduced as a tool towards computing how long it takes a random walk on a graph to hit a subset of the boundary states. We analyze some problems in epidemiology and genetics using branching processes. A hidden Markov model categorizing atmospheric variables yields an improved fit to the precipitation data. A similar model is used to describe whether a chemical transmission channel in a nerve cell is open or closed.

2.1. Precipitation at Snoqualmie Falls

The US Weather Service maintains a large number of precipitation monitors throughout the United States. One station is located at the Snoqualmie Falls in the foothills of the Cascade Mountains in western Washington. A day is defined as wet if at least 0.01 inches of precipitation falls during a precipitation day: 8 a.m. through 8 a.m. the following calendar day. To start with, we shall ignore the amounts of rainfall, and just look at the pattern of wet and dry days. Using data from 1948 through 1983, and looking at January rainfall only, there were 325 dry and 791 wet days. Let $X_{ij}=1(\text{day } i \text{ of year } j \text{ wet})$, where $1(A)$ is 1 if the event A occurs, and 0 otherwise. A very simple model, which we can call the Bernoulli model, is that $X_{ij}\sim\text{Bin}(1,p)$, with the X_{ij} independent, i.e., an iid model, and with p being the probability of rain at Snoqualmie Falls on a January day. The **likelihood** (probability of the observed data as a function of p) is

$$L(p) \propto p^{791}(1-p)^{325}. \tag{2.1}$$

Appendix A contains a brief review of likelihood theory for multinomial data to illustrate some of the central ideas. Edwards (1985) is a good reference for more general likelihood theory. The maximum point of $L(p)$ is the maximum

CHAPTER 2

Discrete time Markov chains

Starting from a very simple model of daily precipitation, we build some theory for discrete time Markov chains. We develop some estimation and testing theory. and look at the goodness of fit of this simple model in different ways. A parsimonious model for higher order of dependence is applied to some meteorological data. Harmonic functions are introduced as a tool towards computing how long it takes a random walk on a graph to hit a subset of the boundary states. We analyze some problems in epidemiology and genetics using branching processes. A hidden Markov model categorizing atmospheric variables yields an improved fit to the precipitation data. A similar model is used to describe whether a chemical transmission channel in a nerve cell is open or closed.

2.1. Precipitation at Snoqualmie Falls

The US Weather Service maintains a large number of precipitation monitors throughout the United States. One station is located at the Snoqualmie Falls in the foothills of the Cascade Mountains in western Washington. A day is defined as wet if at least 0.01 inches of precipitation falls during a precipitation day: 8 a.m. through 8 a.m. the following calendar day. To start with, we shall ignore the amounts of rainfall, and just look at the pattern of wet and dry days. Using data from 1948 through 1983, and looking at January rainfall only, there were 325 dry and 791 wet days. Let $X_{ij}=1$(day i of year j wet), where $1(A)$ is 1 if the event A occurs, and 0 otherwise. A very simple model, which we can call the Bernoulli model, is that X_{ij}~Bin$(1,p)$, with the X_{ij} independent, i.e., an iid model, and with p being the probability of rain at Snoqualmie Falls on a January day. The **likelihood** (probability of the observed data as a function of p) is

$$L(p) \propto p^{791}(1-p)^{325}. \tag{2.1}$$

Appendix A contains a brief review of likelihood theory for multinomial data to illustrate some of the central ideas. Edwards (1985) is a good reference for more general likelihood theory. The maximum point of $L(p)$ is the maximum

likelihood estimate (mle) $\hat{p}$ of p. Letting n be the number of days observed, it is easy to see that $\hat{p}=\sum x_{ij}/n=0.709$. A standard error for this estimate is $(p(1-p)/n)^{½}$ which we estimate (using $\hat{p}$ in place of p) to be 0.014.

In order to assess the fit of the binomial model for rainfall, we first try to see if the independence assumption seems reasonable. We may suspect a certain amount of persistence, i.e., stretches of like weather, in the data. This would be induced by the relatively slow movement of large weather systems through an area. In the winter, a typical front may take up to three days to pass through from the Pacific Ocean. In order to study this hypothesis, let us look at consecutive pairs of days. Figure 2.1 shows the pattern of rainfall.

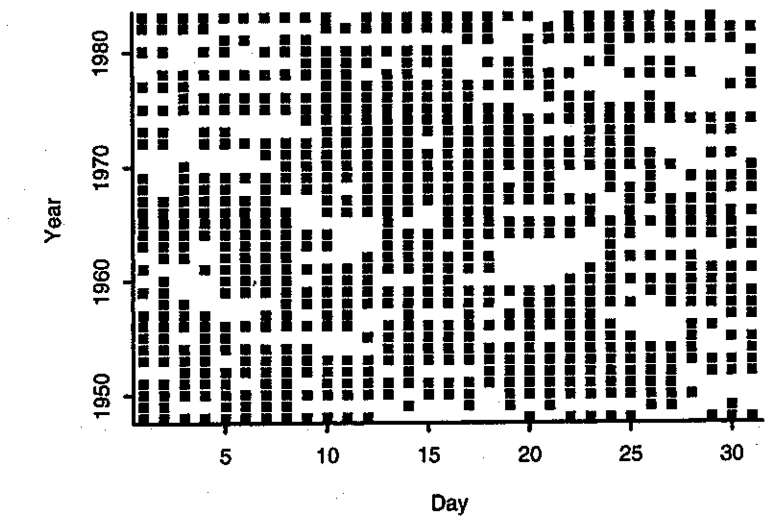

Figure 2.1. The pattern of January precipitation at Snoqualmie Falls. Each square is a day with measurable precipitation. Rows correspond to years, columns to days.

If the independence model is correct, we would expect to see $36\times30\times\hat{p}(1-\hat{p})=223$ dry days following wet days, since we have 36 years of data, and 30 consecutive pairs of days for each January. Table 2.1 contains the total counts, with expected counts under the independence assumption shown in parenthesis. There seems to be a lot more dry days followed by dry days, and wet days followed by wet days, than what the simple iid model predicts. To build a better model of this phenomenon, let us introduce two parameters:

$$p_w = \mathbf{P}(\text{wet today} \mid \text{wet yesterday}) \tag{2.2}$$

$$p_d = \mathbf{P}(\text{wet today} \mid \text{dry yesterday}). \tag{2.3}$$

Table 2.1 Observed precipitation

	Today dry		Today wet		Total
Yesterday dry	186	(91)	123	(223)	309
Yesterday wet	128	(223)	643	(543)	771
Total	314		766		1080

If the X_i are not independent, we must specify the conditional probabilities

$$\mathbf{P}(X_{i+1}=l \mid X_0=k_0, \ldots, X_i=k_i) \tag{2.4}$$

for all i, l, and $k_1, \ldots, k_i$. Note that we will assume unless otherwise specified that the process is observed from time 0. A simple (and perhaps natural) way to specify the probabilities in (2.4) is to assume that the conditional probability only depends on what happened at the previous time point. This assumption was first studied systematically by the Russian probabilist Markov[1] in a sequence of papers, starting in 1907, on generalizing various limit laws to dependent data. Formally we write the **Markov assumption** for a random process (X_n) with discrete state space

$$\begin{aligned}\mathbf{P}(X_{n+1}=l \mid X_0=k_0, \ldots, X_n=k_n) \\ = \mathbf{P}(X_{n+1}=l \mid X_n=k_n) = p_{k_n l}(n).\end{aligned} \tag{2.5}$$

If (X_n) satisfies (2.5) it is called a **Markov chain**. Two seemingly more general forms of (2.5) are outlined in Exercise **1**: in part (a) we show that the conditional distribution of the process at any set of future times, given any set of times up to and possibly including the present, only depends on the last of the times in the condition, and in part (b) we show that an equivalent, and rather colorful, way of stating the Markov property is that the future is independent of the past, given the present.

The functions $p_{ij}(n)$ are called **transition probabilities**. We can write the transition probabilities in matrix form. The matrices $\mathbb{P}(n) = (p_{ij}(n))$ are called **transition matrices**.

In order to prove the existence of a Markov chain with a given set of transition matrices and distribution of X_0 one has to verify the Kolmogorov consistency condition (1.22). This is made precise, e.g., in Freedman (1983, pp. 7–8). Here is a simple fact about transition matrices:

[1] Markov, Andrei Andreevich (1856–1922). Russian probabilist in the St Petersburg School. He was a student of Chebyshev, and proved the law of large numbers rigorously in a variety of cases, including dependent sequences.

ved precipitation

	Today wet		Total
1)	123	(223)	309
3)	643	(543)	771
	766		1080

ecify the conditional probabilities

$i)$ (2.4)

will assume unless otherwise specified
A simple (and perhaps natural) way to
ssume that the conditional probability
evious time point. This assumption was
n probabilist Markov[1] in a sequence of
; various limit laws to dependent data.
ption for a random process (X_n) with

$k_n)$

$_n) = p_{k_n l}(n).$ (2.5)

v chain. Two seemingly more general
1: in part (a) we show that the condi-
set of future times, given any set of
resent, only depends on the last of the
ve show that an equivalent, and rather
erty is that the future is independent of

ansition probabilities. We can write
orm. The matrices $\mathbb{P}(n) = (p_{ij}(n))$ are

Markov chain with a given set of tran-
ne has to verify the Kolmogorov con-
precise, e.g., in Freedman (1983, pp.
on matrices:

ian probabilist in the St Petersburg School. He
of large numbers rigorously in a variety of cases,

Proposition 2.1 If $\mathbb{P}$ is a sequence of transition matrices for a Markov chain with state space $S=\{0,\ldots,K\}$, where K may be finite or infinite, then $\sum_{j=0}^{K} p_{ij}(n)=1$ for any n.

Proof We have that $p_{ij}(n)=\mathbf{P}(X_{n+1}=j \mid X_n=i)$, so

$$\sum_{j=0}^{K} p_{ij}(n) = \sum_{j=0}^{K} \mathbf{P}(X_{n+1}=j \mid X_n=i)$$

$$= \mathbf{P}\Big(\bigcup_{j=0}^{K} \{X_{n+1}=j\} \mid X_n=i\Big) = \mathbf{P}(X_{n+1}\in S \mid X_n=i) = 1 \qquad (2.6)$$

since the process must go somewhere. □

It is often a reasonable simplifying assumption that the transition probabilities are independent of time; such Markov chains are said to have **stationary transition probabilities.** In that case we just need a single transition matrix $\mathbb{P}=\mathbb{P}(1)$. For our rainfall model, we are only considering January. This makes the assumption of stationary transition probabilities reasonable, if we believe (at least approximately) that this month is meteorologically homogeneous. The state space is {dry,wet}, which we can map into $\{0,1\}$. Then, using (2.2) and (2.3), $p_{00}=1-p_d$, $p_{01}=p_d$, $p_{10}=1-p_w$, and $p_{11}=p_w$. In matrix notation,

$$\mathbb{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1-p_d & p_d \\ 1-p_w & p_w \end{pmatrix}. \qquad (2.7)$$

A matrix of non-negative elements with all row sums equal to one is often called a **stochastic** matrix. From now on we will, unless specifically stating otherwise, assume that all transition probabilities are stationary. Here are some elementary properties of stochastic matrices.

Proposition 2.2 (i) A stochastic matrix has at least one eigenvalue equal to one.

(ii) If $\mathbb{P}$ is stochastic, then $\mathbb{P}^k$ is also stochastic for all $k=1,2,3,\ldots$.

Proof (i) is a consequence of the definition of a stochastic matrix, which can be written $\mathbb{P}\mathbf{1}^T = \mathbf{1}^T$, where $\mathbf{1}$ is a row vector of ones (recall that all vectors are assumed to be row vectors). Hence $(\mathbb{I}-\mathbb{P})\mathbf{1}^T = \mathbf{0}$, where $\mathbb{I}$ is the identity matrix, so $\mathbf{1}$ is a right eigenvector of $\mathbb{P}$ corresponding to the eigenvalue 1. Now (ii) follows easily, writing

$$\mathbb{P}^k\mathbf{1}^T = \mathbb{P}^{k-1}\mathbb{P}\mathbf{1}^T = \mathbb{P}^{k-1}\mathbf{1}^T = \cdots = \mathbf{1}^T. \qquad (2.8)$$

□

The likelihood for a Markov chain can be written, using (2.5) and (1.13), as

$$\begin{aligned} L(\mathbb{P}) &= \mathbf{P}(X_0=x_0)\prod_{i=0}^{n-1}\mathbf{P}(X_{i+1}=x_{i+1}\mid X_i=x_i) \\ &= \mathbf{P}(X_0=x_0)\prod_{i=0}^{n-1}p_{x_i x_{i+1}} = \mathbf{P}(X_0=x_0)\prod_{k,l=0}^{K}p_{kl}^{n_{kl}}, \end{aligned} \tag{2.9}$$

where n_{kl} is the number of transitions from k to l observed in the chain. In our example we have the additional complication that we are considering 36 years. A simple model is to assume that years are independent. While quasi-periodic large-scale meteorological oscillations such as El Niño may make this hypothesis somewhat suspect (cf. Woolhiser, 1992), it nevertheless allows us to proceed. Furthermore, we shall be able to test it later (Exercise **D1**). Under the assumption of year-to-year independence the likelihood is a product of 36 factors, each of the form (2.9). Clearly, the product collapses, and we can use the data in Table 2.1 to compute

$$L(\mathbb{P}) = L(p_{01},p_{11}) = \left\{\prod_{i=1}^{36}\mathbf{P}(X_0^i=x_0^i)\right\}p_{00}^{186}p_{01}^{123}p_{10}^{128}p_{11}^{643} \tag{2.10}$$

Assuming that the starting values X_0^i for each year i are fixed (this assumption will be discussed in more detail in section 2.7), so that the beginning term in the right-hand side of (2.10) is 1, we find that L is maximized by

$$\hat{p}_{01}=\frac{123}{309}=0.398 \quad \hat{p}_{11}=\frac{643}{771}=0.834 \tag{2.11}$$

so

$$\hat{\mathbb{P}} = \begin{pmatrix}0.398 & 0.602\\ 0.166 & 0.834\end{pmatrix}. \tag{2.12}$$

These estimates are substantially different from the estimate $\hat{p}=0.709$ from the iid model. However, we may question whether such a difference could occur by chance. At a first glance this seems very unlikely, since $\hat{p}_{01}$ is 22 standard errors (of $\hat{p}$) away from $\hat{p}$. For a formal test of significance we use the likelihood ratio test. Recall (or see Appendix A) that under suitable regularity conditions, the **log likelihood ratio** $2(\log L(\hat{\mathbb{P}})-\log L(\hat{p}))$ has a χ^2 distribution with degrees of freedom equal to the difference in the dimension of the parameter spaces; in this case 2-1=1. Although this result was developed for iid processes, it is also true in the Markov chain case. We will return to it in section 2.7. In order to be able to compare the likelihoods we need to exclude the January 1 measurements when computing the iid mle, since those observations cannot be used to compute the Markov chain mle's. This yields $\hat{p}=771/1080=0.714$, slightly higher than the 0.709 we obtained from the full data set. Computing the log likelihood ratio we get e written, using (2.5) and (1.13), as

$=x_{i+1} \mid X_i=x_i)$

$$\mathbf{P}(X_0=x_0)\prod_{k,l=0}^{K} p_{kl}^{n_{kl}}, \qquad (2.9)$$

om k to l observed in the chain. In our ation that we are considering 36 years. are independent. While quasi-periodic such as El Niño may make this iser, 1992), it nevertheless allows us to o test it later (Exercise **D1**). Under the e the likelihood is a product of 36 fac- product collapses, and we can use the

$$X_0^i=x_0^i)\Bigg\} p_{00}^{186}p_{01}^{123}p_{10}^{128}p_{11}^{643} \qquad (2.10)$$

each year i are fixed (this assumption n 2.7), so that the beginning term in the t L is maximized by

$$\frac{43}{71} = 0.834 \qquad (2.11)$$

$$(2.12)$$

nt from the estimate $\hat{p}$=0.709 from the hether such a difference could occur by unlikely, since $\hat{p}_{01}$ is 22 standard errors significance we use the likelihood ratio nder suitable regularity conditions, the)) has a χ^2 distribution with degrees of nension of the parameter spaces; in this veloped for iid processes, it is also true to it in section 2.7. In order to be able exclude the January 1 measurements e observations cannot be used to com- ds $\hat{p}$=771/1080=0.714, slightly higher data set. Computing the log likelihood

$$\begin{aligned}2(\log L(\hat{\mathbb{P}})-\log L(\hat{p})) = 2(&643\log 0.834+128\log 0.166\\ &+123\log 0.398+186\log 0.602 \qquad (2.13)\\ &-771\log 0.714-309\log 0.286) = 184.5.\end{aligned}$$

which under the null hypothesis of the Bernoulli model is distributed $\chi^2(1)$, corresponding to a P-value of 0. We therefore reject the iid model at all reasonable levels.

2.2. The marginal distribution

Although the Markov assumption tells us how to compute conditional probabilities, one often wants marginal probabilities. It is relatively straightforward to compute these. For example, in a 0-1 chain we have that

$$\begin{aligned}\mathbf{P}(X_{n+1}=1) &= \mathbf{P}(X_{n+1}=1, X_n=0) + \mathbf{P}(X_{n+1}=1, X_n=1)\\ &= \mathbf{P}(X_n=0)\,p_{01} + \mathbf{P}(X_n=1)\,p_{11} \qquad (2.14)\\ &= \mathbf{P}(X_n=1)\,(p_{11}-p_{01}) + p_{01}.\end{aligned}$$

Define the **initial distribution** $\mathbf{p}_0 = (p_0(0),\ldots,p_0(K))$ where $p_0(i) = \mathbf{P}(X_0=i)$. In the 0-1 case we write $p_0(1)\equiv p_1$. Then (2.14) can be written

$$\begin{aligned}\mathbf{P}(X_1=1) &= p_1(p_{11}-p_{01}) + p_{01},\\ \mathbf{P}(X_2=1) &= (p_{11}-p_{01})\mathbf{P}(X_1=1) + p_{01}\\ &= p_1(p_{11}-p_{01})^2 + p_{01}(1+(p_{11}-p_{01})) \qquad (2.15)\\ &\cdots\\ \mathbf{P}(X_n=1) &= (p_{11}-p_{01})^n p_1 + p_{01}\sum_{j=0}^{n-1}(p_{11}-p_{01})^j.\end{aligned}$$

If $p_{00}=p_{11}=1$ we have $\mathbf{P}(X_n=1)=p_1$. If $p_{01}\neq p_{11}$ we can write

$$\mathbf{P}(X_n=1) = \frac{p_{01}}{1-(p_{11}-p_{01})} + \left[p_1-\frac{p_{01}}{1-(p_{11}-p_{01})}\right](p_{11}-p_{01})^n. \qquad (2.16)$$

Notice that the effect of the initial distribution p_1 is dampened exponentially, and disappears completely when $p_1=p_{01}/(1-(p_{11}-p_{01}))$. In that situation $\mathbf{P}(X_n=1)$ is the same for each n. This choice of p_1 is called the **stationary initial distribution**. We will return to this in section 2.4.

More generally, let the state space S be an arbitrary countable set, which we identify with the integers $\mathbf{Z}$, and define $p_{jk}^{(n)} = \mathbf{P}(X_n=k \mid X_0=j)$. Here is an important computation, called the **Chapman[1]–Kolmogorov equation,**

[1] Chapman, Sydney (1888–1970). Leading British astro- and geophysicist. Major contributions to the understanding of the aurora; space physics; and convection in the atmosphere.

although it was discovered independently by many workers, including Bachelier (1900) and Einstein (1905).

Lemma 2.1

$$p_{jk}^{(n)} = \sum_{l \in S} p_{jl}^{(m)} p_{lk}^{(n-m)}, \quad 1 \le m \le n-1. \tag{2.17}$$

Proof Since the process must be somewhere in S at time m, we have

$$\begin{aligned} p_{jk}^{(n)} &= \mathbf{P}(\{X_n=k\} \cap \bigcup_{l \in S} \{X_m=l\} \mid X_0=j) \\ &= \sum_{l \in S} \mathbf{P}(X_n=k, X_m=l \mid X_0=j) \\ &= \sum_{l \in S} \mathbf{P}(X_n=k \mid X_m=l, X_0=j) \mathbf{P}(X_m=l \mid X_0=j) \\ &= \sum_{l \in S} \mathbf{P}(X_n=k \mid X_m=l) \mathbf{P}(X_m=l \mid X_0=j). \end{aligned} \tag{2.18}$$

□

In matrix notation we rewrite (2.17) as

$$\mathbb{P}_n \equiv (p_{jk}^{(n)}) = \mathbb{P}_{n-m} \mathbb{P}_m. \tag{2.19}$$

But $\mathbb{P}_1 = \mathbb{P}$ so $\mathbb{P}_n = \mathbb{P}^n$. Let $\mathbf{p}_n = (\ldots, \mathbf{P}(X_n=0), \ldots, \mathbf{P}(X_n=k), \ldots)$ denote the probability distribution of X_n. Since

$$\mathbf{p}_n = \mathbf{p}_{n-1} \mathbb{P} \tag{2.20}$$

(recall the computation of $\mathbf{P}(X_n=1)$ earlier) we see that

$$\mathbf{p}_n = \mathbf{p}_0 \mathbb{P}^n. \tag{2.21}$$

Application (Snoqualmie Falls precipitation) Suppose that we accept the Markov chain model developed in section 2.1 for the Snoqualmie Falls precipitation data, and that it happened to rain on January 1 this year. What would be the probability of rain on January 6, i.e., five days hence? To compute this probability, we need to determine $\hat{\mathbb{P}}^5$, where

$$\hat{\mathbb{P}} = \begin{bmatrix} 0.602 & 0.398 \\ 0.166 & 0.834 \end{bmatrix} \tag{2.22}$$

so that

$$\hat{\mathbb{P}}^5 = \begin{bmatrix} 0.305 & 0.695 \\ 0.290 & 0.710 \end{bmatrix}. \tag{2.23}$$

Notice that the two rows of $\hat{\mathbb{P}}^5$ are much more similar than those of $\hat{\mathbb{P}}$. This will be explained in section 2.5. The desired probability is obtained by setting ndently by many workers, including

$\leq n-1.$ (2.17)

mewhere in S at time m, we have

$l\} \mid X_0=j)$

$_0=j)$

$_0=j)\mathbf{P}(X_m=l \mid X_0=j)$ (2.18)

$\mathbf{P}(X_m=l \mid X_0=j).$

□

s

(2.19)

$_n=0), \ldots, \mathbf{P}(X_n=k), \ldots)$ denote the proba-

(2.20)

lier) we see that

(2.21)

precipitation) Suppose that we accept section 2.1 for the Snoqualmie Falls pre- rain on January 1 this year. What would 6, i.e., five days hence? To compute this where

(2.22)

(2.23)

nuch more similar than those of $\hat{\mathbb{P}}$. This desired probability is obtained by setting

$\mathbf{p}_0=(0,1)$ in (2.21), so that

$$\hat{\mathbf{p}}_5=(0,1)\hat{\mathbb{P}}^5=(0.29,0.71). \tag{2.24}$$

In other words, the probability of rain at Snoqualmie Falls on January 6, given rain on January 1, is 0.71. □

A different type of question is how long a Markov chain would be expected to stay in a given state. Clearly, if $p_{ii}=0$ it is certain not to stay. If $p_{ii}>0$, the time spent in the state has a geometric distribution with mean $1/(1-p_{ii})$(Exercise **2**). For the Snoqualmie Falls application this translates to a mean of 2.5 consecutive dry days and 6.0 consecutive wet days in January. We return to this in section 2.9.

2.3. Classification of states

Let $A \subset S$. The **hitting time** T_A of A is

$$T_A = \begin{cases} \min\{n>0: X_n \in A\} & \text{if } X_n \text{ ever hits } A \\ \infty & \text{otherwise} \end{cases} \tag{2.25}$$

If $A=\{a\}$ we write T_a. Denote the distribution of the chain, starting from the state x (i.e., $p_0(x)=1$ and $p_0(y)=0$ for any $y \neq x$), by $\mathbf{P}^x$. More generally, we write the distribution of the chain starting from the initial distribution $\mathbf{p}_0$ as $\mathbf{P}^{\mathbf{p}_0}$, and compute it using the formula

$$\mathbf{P}^{\mathbf{p}_0}(A) = \sum_{i \in S} p_0(i)\mathbf{P}^i(A). \tag{2.26}$$

This amounts to first choosing the initial state i at random from $\mathbf{p}_0$, and then running the chain starting from state i.

Proposition 2.3 $\quad p_{jk}^{(n)} = \sum_{m=1}^{n} \mathbf{P}^j(T_k=m)p_{kk}^{(n-m)}.$

Proof Write $\{X_n=k\} = \sum_{m=1}^{n}\{T_k=m, X_n=k\}$, where the summation sign stands for a union of disjoint sets. Now

$$\begin{aligned} p_{jk}^{(n)} = \mathbf{P}^j(X_n=k) &= \sum_{m=1}^{n} \mathbf{P}^j(T_k=m,\, X_n=k) \\ &= \sum_{m=1}^{n} \mathbf{P}^j(T_k=m)\mathbf{P}^j(X_n=k \mid T_k=m) \qquad (2.27) \\ &= \sum \mathbf{P}^j(T_k=m)\mathbf{P}(X_n=k \mid X_0=j, X_1 \neq k, \ldots, X_{m-1} \neq k, X_m=k) \\ &= \sum \mathbf{P}^j(T_k=m)\mathbf{P}(X_n=k \mid X_m=k) = \sum \mathbf{P}^j(T_k=m)p_{kk}^{(n-m)}. \end{aligned}$$

□

Call a state **absorbing** if $p_{kk}=1$. If the chain ever reaches k it stays there forever.

Corollary For an absorbing state k we have that $p_{jk}^{(n)}=\mathbf{P}^j(T_k\le n)$.

Proof The content of this equation is really trivial: in order to go from j to k in n steps we need to hit k no later than time n. A formal proof follows from the observation that $p_{kk}^{(n-m)}=1$ for all $m<n$ and Proposition 2.3. □

T_k is an example of a particularly interesting class of random times. Call the random time τ a **Markov time** if the event $\{\tau=n\}$ is completely determined by the values of $X_0,\ldots,X_n$. The **strong Markov property** asserts that the Markov property holds also at Markov times. More formally, let $f_i(k)=\mathbf{P}^k(X_1=i)$. Then

$$\mathbf{P}(X_{\tau+1}=i \mid X_0,\ldots,X_\tau)=f_i(X_\tau). \tag{2.28}$$

A proof of this can be found, e.g., in Freedman (1983, Theorem 1:21).

Say that i **reaches** j, written $i\to j$, if there is an n such that $p_{ij}^{(n)}>0$. If $i\to j$ and $j\to i$ we say that i and j **communicate**, denoted $i\leftrightarrow j$.

Theorem 2.1 $\leftrightarrow$ is an equivalence relation.

Proof $i\leftrightarrow i$ since $p_{ii}^{(0)}=\mathbf{P}(X_n=i \mid X_n=i)=1$. Next, $i\leftrightarrow j$ implies that $j\leftrightarrow i$ by definition. Finally, if $i\leftrightarrow j$ and $j\leftrightarrow k$ there are integers m and n such that $p_{ij}^{(n)}>0$ and $p_{jk}^{(m)}>0$. Thus

$$p_{ik}^{(n+m)}=\sum_r p_{ir}^{(n)}p_{rk}^{(m)}\ge p_{ij}^{(n)}p_{jk}^{(m)}>0 \tag{2.29}$$

and $i\to k$. To show that $k\to i$ uses a similar argument. □

We can partition all states into equivalence classes with respect to the relation $\leftrightarrow$. A Markov chain is **irreducible** if there is only one equivalence class, i.e., if all states communicate.

Example (A model for radiation damage) A finite **birth and death chain** is a Markov chain on $\{0,\ldots,K\}$ in which a particle in state i can either stay or move to one of the neighboring states $i+1$ or $i-1$. Reid and Landau (1951) proposed this chain as a model for the transmission of radiation damage following the initial damage due to the absorption of radiation quanta. The mechanism by which this transmission takes place was assumed to be the depolymerization of macromolecules associated with the sensitive volume of the organism. State 0 corresponds to a healthy organism, and state K to one with visible radiation damage. The intermediate states correspond to amplification or healing of the initial damage, which is taken to be state 1. The extreme states are assumed absorbing, so the transition matrix for this process is

$$\mathbb{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q_1 & r_1 & p_1 & 0 & \cdots & 0 & 0 \\ 0 & q_2 & r_2 & p_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & r_{K-1} & p_{K-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{2.30}$$

where r_i is the conditional probability of staying in state i, p_i is the conditional probability of moving to state $i+1$ (amplification of damage), and q_i the conditional probability of moving to state $i-1$ (recovery). Reid and Landau suggested to use $r_i=0$, $p_i=i/K$, and $q_i=1-i/K$. This chain has three classes: $\{0\}$, $\{K\}$, and $\{1,\ldots,K-1\}$. Starting from state 1, we may want to compute the recovery probability (λ_0), i.e., the probability of reaching state 0 before state K. By conditioning on the last step, which must be from 1 to 0, we can write

$$\lambda_0 = p_{10} \sum_{n=0}^{\infty} p_{11}^{(n)}. \tag{2.31}$$

For example, if $K=3$ so

$$\mathbb{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2/3 & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.32}$$

we find that $p_{11}^{(2n)} = (\frac{1}{3} \times \frac{1}{3})^n$ while $p_{11}^{(2n+1)} = 0$ (note that the only way to achieve a transition from 1 to 1 in $2n$ steps is to go 1-2-1-2-1 $\cdots$). Hence $\lambda_0 = 3/4$. Generally, $\lambda_0 = 1-2^{-(K-1)}$ (Exercise 3). □

We say that a state i has **period** d if $p_{ii}^{(n)}=0$ for all n not divisible by d, and d is the greatest such integer. This means that if the chain is in state i at time n it can only return there at times of the form $n+kd$ for some integer k. If $p_{ii}^{(n)}=0$ for all n, we say that state i has infinite period. A state with period 1 is called **aperiodic**.

Theorem 2.2 Periodicity is an equivalence class property, i.e., if $i \leftrightarrow j$ then $d(i)=d(j)$.

Proof Let m, n be such that $p_{ij}^{(m)}>0$, $p_{ji}^{(n)}>0$, and assume that $p_{ii}^{(s)}>0$. Then

$$p_{jj}^{(m+n)} \geq p_{ji}^{(n)} p_{ij}^{(m)} > 0 \tag{2.33}$$

and

$$p_{jj}^{(m+n+s)} \geq p_{ji}^{(n)} p_{ii}^{(s)} p_{ij}^{(m)} > 0 \tag{2.34}$$

so $d(j)$ must divide $m+n$ and $m+n+s$. Hence it must divide their difference s for any s such that $p_{ii}^{(s)}>0$. Therefore $d(j)$ divides $d(i)$. Similarly, $d(i)$ is seen to divide $d(j)$, so the two numbers must be equal. □

Example (A model for radiation damage, continued) In the Reid–Landau radiation damage model described earlier we have

$$\mathbb{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q_1 & r_1 & p_1 & 0 & \cdots & 0 & 0 \\ 0 & q_2 & r_2 & p_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & r_{K-1} & p_{K-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{2.35}$$

Therefore the period of the class $\{1,\ldots,K-1\}$ is 2. □

To prove the next result, we need a number-theoretic lemma:

Lemma 2.2 Given positive integers n_1 and n_2 with greatest common divider (gcd) 1, any integer $n>n_1n_2$ can be written $n=ln_1+kn_2$ for non-negative integers l and k.

Proof Consider the modulo n_2 residue classes of the n_2 distinct positive integers n, $n-n_1$, $n-2n_1, \ldots, n-(n_2-1)n_1$. Either these residue classes are all different, in which case one residue class must be 0, so the corresponding number $n-kn_1$ is divisible by n_2, i.e., $n=kn_1+ln_2$, or at least two residue classes are the same. If the common residue class is 0 the preceding argument applies. Otherwise we can write $n-sn_1=a+bn_2$ and $n-tn_1=a+cn_2$ for $0 \le t<s \le n_2-1$, $b<c$, and $0<a<n_2$. Hence

$$n-sn_1-(n-tn_1)=(s-t)n_1=(c-b)n_2. \tag{2.36}$$

Since $\gcd(n_1,n_2)=1$ we must have $s-t$ containing all prime factors of n_2. But then $s-t \ge n_2$ which is a contradiction. □

Proposition 2.4 If i and j are states of an irreducible aperiodic chain, then there is an integer $N=N(i,j)$ such that $p_{ij}^{(n)}>0$ for all $n \ge N$.

Proof Since $d(j)=1$ there are integers n_1, n_2 with gcd 1 such that $p_{jj}(n_k)>0$, $k=1,2$. From Lemma 2.2 we see that any sufficiently large n can be written ln_1+kn_2, whence

$$p_{jj}^{(n)}=p_{jj}^{(ln_1+kn_2)} \ge \left[p_{jj}^{(n_1)}\right]^l \left[p_{jj}^{(n_2)}\right]^k>0. \tag{2.37}$$

Finally, for each pair i,j there is an n_0 such that $p_{ij}^{(n_0)}>0$. Hence

$$p_{ij}^{(n+n_0)} \ge p_{ij}^{(n_0)} p_{jj}^{(n)}>0. \tag{2.38}$$

□

Corollary Let X and Y be iid irreducible aperiodic Markov chains. Then $Z=(X,Y)$ is an irreducible Markov chain.

Proof It is clear that Z is Markov, with transition probabilities

$$\tilde{p}_{ij,kl} = \mathbf{P}(Z_{t+1}=(k,l) \mid Z_t=(i,j)) = \mathbf{P}(X_{t+1}=k, Y_{t+1}=l \mid X_t=i, Y_t=j)$$
$$= \frac{\mathbf{P}(X_{t+1}=k, X_t=i)}{\mathbf{P}(X_t=i)} \frac{\mathbf{P}(Y_{t+1}=l, Y_t=j)}{\mathbf{P}(Y_t=j)} = p_{ik}p_{jl}. \tag{2.39}$$

By the Proposition we can find an $N=N(i,j,k,l)$ such that $p_{ik}^{(n)}>0$ and $p_{jl}^{(n)}>0$ for all $n>N$. Thus $\tilde{p}_{ij,kl}^{(n)}>0$ and Z is therefore irreducible. □

Let $f_{ij}^{(n)}=\mathbf{P}^i(T_j=n)$ be the **first passage distribution** from state i to state j. We have $f_{ij}^{(0)}=0$ and

$$f_{ij}^{(n)} = \mathbf{P}(X_n=j, X_k\neq j, k=1,..,n-1 \mid X_0=i). \tag{2.40}$$

Define $f_{ij}=\sum_{n=0}^{\infty} f_{ij}^{(n)}=\mathbf{P}^i(T_j<\infty)$. The state i is called **persistent** (also called **recurrent** by some authors) if $f_{ii}=1$, **transient** otherwise. Think of a persistent state as one that the process will eventually return to, while a transient state is one with positive probability of no return.

Theorem 2.3 A state i is persistent iff $\sum_{n\geq 1} p_{ii}^{(n)}=\infty$.

Proof If i is transient, let M be the number of returns to i. Then $M=\sum_n 1(X_n=i)$. By the strong Markov property (2.28) $\mathbf{P}^i(M\geq k)=f_{ii}^k$, so $\mathbf{E}^i M=\sum_{k\geq 1}\mathbf{P}^i(M\geq k)=f_{ii}/(1-f_{ii})$, where $\mathbf{E}^i$ is expectation with respect to $\mathbf{P}^i$. Since $f_{ii}<1$, $\mathbf{E}^i M<\infty$. But

$$\infty > \mathbf{E}^i M = \sum_{n=1}^{\infty} \mathbf{E}^i 1(X_n=i) = \sum_{n=1}^{\infty} p_{ii}^{(n)}. \tag{2.41}$$

Conversely, if i is persistent it returns with probability 1. By the strong Markov property it starts over again, and hence returns with probability one. Thus it returns an infinite number of times with probability one, so $\mathbf{P}^i(M=\infty)=1$, i.e., $\mathbf{E}^i M=\infty$. □

Remark The proof of Theorem 2.3 shows that $M-1$ has a geometric distribution with parameter f_{ii}.

Example (A simple random walk) A **simple random walk** is a birth and death chain on the integers with $p_j\equiv p$, $r_j\equiv 0$ and $q_j\equiv q$, so $q=1-p$. This is an irreducible Markov chain with countable state space. One interpretation is the chain corresponding to the number of heads in successive tosses of a coin with probability p of heads. We will look at the persistence or transience of state 0. A binomial computation shows that

$$p_{00}^{(2n)} = \binom{2n}{n} p^n q^n \sim \frac{(4pq)^n}{(\pi n)^{½}} \tag{2.42}$$

using Stirling's formula $n! \sim n^{n+½} e^{-n} (2\pi)^{½}$. Hence $\sum p_{00}^{(2n)} = \infty$ iff $p = q = ½$. In other words, 0 is a persistent state iff the coin is fair. □

Remark One can define a simple random walk in higher dimensions by requiring that at any point on the k-dimensional integer lattice the process has the same probabilities of going to its nearest neighbors, regardless of which point it is at. The process is **fair** if the probability is the same to go to each of its neighbors. A similar computation (Exercise 4) to the one in the example above shows that if $k=2$ the origin (and thus any state) is persistent. However, if $k>2$ it is transient. In three dimensions, with probability 1/6 of going up, down, east, west, north, or south, we get

$$\begin{aligned}
p_{00}^{(2n)} &= \left(\frac{1}{6}\right)^{2n} \sum_{j+k\le n} \frac{2n!}{(j!)^2 (k!)^2 (n-j-k)!^2} \\
&= \left(\frac{1}{2}\right)^{2n} \binom{2n}{n} \sum_{j+k\le n} \left(\frac{1}{3^n} \frac{n!}{j!k!(n-j-k)!}\right)^2 \qquad (2.43) \\
&\le \left(\frac{1}{2}\right)^{2n} \binom{2n}{n} \max_{j+k\le n} \frac{1}{3^n} \frac{n!}{j!k!(n-j-k)!} \sum_{j+k\le n} \frac{1}{3^n} \frac{n!}{j!k!(n-j-k)!}.
\end{aligned}$$

The sum is one, being the sum of all probabilities in a trinomial distribution with probability ⅓ of each category, and the maximum is obtained when $j=k=(n-j-k)=n/3$ (or as close as possible to this if n is not divisible by 3). Applying Stirling's formula we see that an upper bound to $p_{00}^{(2n)}$ is, to within an order of n^{-2},

$$2^{-2n} \times \frac{2^{2n}}{\sqrt{n\pi}} \times 3^{-n} \frac{3^{n+3/2}}{2\pi n} = ½ \left(\frac{3}{\pi n}\right)^{3/2}, \tag{2.44}$$

so $\sum p_{00}^{(2n)} < \infty$, whence the walk has positive probability not to return. In fact, the probability of return is about 0.35. In other words, the three-dimensional lattice is a huge place, in which it is easy to get lost. We return to more general random walks in section 2.10. □

Corollary For a transient state i, $p_{ii}^{(n)} \to 0$.

Proof Immediate since $\sum p_{ii}^{(n)} < \infty$. □

For a persistent state $f_{ii}^{(n)}$ is a probability distribution with mean $\mu_i = \sum_n n f_{ii}^{(n)}$, the **mean recurrence time**. If $\mu_i = \infty$, state i is called **null**, otherwise it is called **positive**. This somewhat puzzling nomenclature will be explained in section 2.5. An irreducible aperiodic positive chain is called **ergodic**.

Application (Snoqualmie Falls precipitation, continued) For $n \geq 2$

$$\begin{aligned} f_{11}^{(n)} &= \mathbf{P}^1(X_n=1, X_{n-1}=0, \ldots, X_1=0) \\ &= \mathbf{P}^1(X_n=1 \mid X_{n-1}=0, \ldots, X_1=0)\mathbf{P}^1(X_{n-1}=0, \ldots, X_1=0) \\ &= p_{01}(1-p_{11})(1-p_{01})^{n-2}. \end{aligned} \tag{2.45}$$

Also, $f_{11}^{(1)} = p_{11}$, so the mean recurrence time is $\mu_1 = \sum k f_{11}^{(k)} = 1 + (1-p_{11})/p_{01}$, which we estimate to be 1.42, using $\hat{p}_{01} = 0.398$ and $\hat{p}_{11} = 0.834$. Given a wet day, the mean number of dry days to follow is $\mu_1 - 1$, which we estimate to be 0.42 days. This is a weighted average of wet days inside a wet spell (with no dry days following) and starts of dry spells (with mean duration $1/p_{01}$; cf. Exercise 2). The variance of the recurrence time is $(1-p_{11})(1-p_{01})/p_{01}^2$. Plugging in the estimated transition probabilities and taking the square root we compute a standard deviation of 0.79. Looking at the actual data, eliminating dry periods that overlap Jan. 1 or 31, the mean dry spell length is 2.21, with a standard deviation of 1.64. In order to compare this to the model estimate of $\mu_1 - 1$, we must multiply by the observed proportion of wet–dry transitions, or 0.166, yielding 0.37, only slightly below the model estimate. □

Theorem 2.4 Persistence is an equivalence class property.

Proof Let $i \leftrightarrow j$ and assume that j is persistent. Then there are integers n and m so that $p_{ij}^{(n)} > 0$ and $p_{ji}^{(m)} > 0$. For any $s \geq 0$

$$p_{ii}^{(n+m+s)} \geq p_{ij}^{(n)} p_{jj}^{(s)} p_{ji}^{(m)} \tag{2.46}$$

so

$$\sum_s p_{ii}^{(n+m+s)} \geq p_{ij}^{(n)} p_{ji}^{(m)} \sum_s p_{jj}^{(s)} = \infty, \tag{2.47}$$

and the result follows from Theorem 2.3. □

Let $a_0,a_1,\ldots$ be a sequence of real numbers. If $A(s)=\sum_{k=0}^{\infty} a_k s^k$ converges in some interval $|s|<s_0$ we call $A(s)$ the **generating function** of (a_i). It is easy to show that if $\sum a_k<\infty$ then $A(1-)=\lim_{s\uparrow 1} A(s)=\sum_k a_k$, and for $a_k\geq 0$, if $A(1-)=a<\infty$, then $\sum a_k=a$.

Example (Probability generating functions) If $(p_k;k\geq 0)$ is a probability distribution, the generating function $P(s)=\sum_k^{\infty} p_k s^k$ converges for all $|s|\leq 1$. P is called a **probability generating function** (pgf). If X has pgf P, define the kth **factorial moment** $m_{(k)}=\mathbf{E}X(X-1)\cdots(X-k+1)$. By differentiating under the summation sign in the definition of P we see that

$$m_{(k)} = \sum_{i=k}^{\infty} i(i-1)\cdots(i-k+1)p_k = \frac{d^k}{ds^k}P(1-). \tag{2.48}$$

A similar computation shows that we can recover the probabilities from either the pgf or the factorial moments:

$$p_k = \frac{d^k}{ds^k}\frac{P(s)}{k!}\Big|_{s=0} = \sum_{i=k}^{\infty}(-)^{i-k}\frac{m_{(i)}}{k!(i-k)!}. \tag{2.49}$$

□

Given two sequences (a_k) and (b_k) with generating functions A and B, respectively, we define the **convolution** of the sequences as the sequence (c_k) given by

$$c_k = \sum_{i=0}^{k} a_i b_{k-i}. \tag{2.50}$$

It is easy to see that (c_k) has generating function $C(s)=A(s)B(s)$.

Example (Probability generating functions, continued) If $X_1,\ldots,X_n$ are iid positive random variables with pgf P(s), then the sum $S_n=\sum_1^n X_i$ has pgf $P(s)^n$. For example, $\mathbf{P}(S_n=0)=P(0)^n=p_0^n$, and

$$\mathbf{E}S_n = \frac{dP(s)^n}{ds}\Big|_{s=1} = nP'(1)P(1)^{n-1} = n\mathbf{E}X \tag{2.51}$$

since $P(1)=1$. □

Recall from Proposition 2.4 that

$$p_{ii}^{(n)} = \sum_{k=0}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} \tag{2.52}$$

for any $n\geq 1$. Define the generating functions

$$P_{ij}(s) = \sum_{n=0}^{\infty} p_{ij}^{(n)} s^n \tag{2.53}$$

and

$$F_{ij}(s) = \sum_{n=0}^{\infty} f_{ij}^{(n)} s^n. \tag{2.54}$$

Then, noting that (2.52) is a convolution, we see that

$$F_{ii}(s)P_{ii}(s) = P_{ii}(s) - 1 \tag{2.55}$$

since $p_{ii}^{(0)}=1$. Thus

$$P_{ii}(s) = \frac{1}{1-F_{ii}(s)}. \tag{2.56}$$

Likewise

$$P_{ij}(s) = F_{ij}(s)P_{jj}(s). \tag{2.57}$$

It is worth noting that

$$F_{ii}'(1-) = \mu_i. \tag{2.58}$$

Remark We can use this to give an alternative proof of the result in Theorem 2.3 that i is persistent iff $\sum p_{ii}^{(n)}=\infty$. Assume first that $\sum f_{ii}^{(n)}=1$. Then $F_{ii}(1-)=1$ so $P_{ii}(1-)=\infty$, or $\sum p_{ii}^{(n)}=\infty$. Conversely, if $\sum f_{ii}^{(n)}<1$ we have that $F_{ii}(1-)<1$, so $P_{ii}(1-)<\infty$, and $\sum p_{ii}^{(n)}<\infty$. We can interpret $P_{ii}(1-)$ as the expected number of visits to i, starting from i. □

Example (Coin-tossing) The computation above can be modified to show that for the fair coin-tossing random walk, state 0 (and hence any state) is null persistent. Since $F_{00}(s)=1-P_{00}(s)^{-1}$ we see that $F_{00}'(1-)=P_{00}'(1-)/P_{00}^2(1-)$. Now notice that

$$A_N = \sum_{n=0}^{N} n p_{00}^{(n)} = O(N^{3/2}), \tag{2.59}$$

$$B_N = \sum_{n=0}^{N} p_{00}^{(n)} = O(N^{1/2}), \tag{2.60}$$

and $A_N \to P_{00}'(1-)$, $B_N \to P_{00}(1-)$, so that

$$F_{00}'(1-) = \lim_{N\to\infty} \frac{A_N}{B_N^2} = \lim_{N\to\infty} O(N^{1/2}) = \infty, \tag{2.61}$$

showing that the mean recurrence time is infinite. □

Lemma 2.3 Suppose that j is persistent. Then it is positive persistent iff $\pi_j = \lim_{s \uparrow 1}(1-s)P_{jj}(s) > 0$, and then $\pi_j = 1/\mu_j$.

Proof From (2.56) we have

$$1 - F_{jj}(s) = P_{jj}(s)^{-1} \tag{2.62}$$

so

$$\frac{1-F_{jj}(s)}{1-s} = \frac{1}{(1-s)P_{jj}(s)}. \tag{2.63}$$

But the left-hand side of (2.63) converges to $F_{jj}'(1) = \mu_j$ as $s \uparrow 1$. Hence the limit of the right-hand side is the same, and the result follows. □

Remark This almost proves the convergence of averages of transition probabilities, namely

$$(1/n)\sum_1^n p_{jj}^{(k)} \to \mu_j^{-1} \text{ as } n \to \infty. \tag{2.64}$$

Consider

$$(1-s)P_{jj}(s) = \frac{\sum_{k=0}^{\infty} s^k p_{jj}^{(k)}}{\sum_{k=0}^{\infty} s^k} \tag{2.65}$$

for any $s \leq 1$. If we could take the limit as $s \uparrow 1$ under the summations, the right-hand side would converge to $(1/n)\sum_1^n p_{jj}^{(k)}$, while the left-hand side would converge to $1/\mu_j$ by the Lemma. There is, however, no elementary result allowing us to take the limit under the summation sign. We need a so-called **Tauberian** theorem, such as that given in Feller (1971, Theorem 5 in section XIII.V). We will be able to deduce the result using less difficult mathematics in section 2.5. □

Call a set C of states **closed** if $f_{jk} = F_{jk}(1-) = 0$ for $j \in C$, $k \notin C$. Then $P_{jk}(1-) = 0 = \sum p_{jk}^{(n)}$ and we must have $p_{jk}^{(n)} = 0$ for all n. In fact, in order to verify that a set of states is closed we need only show that $p_{jk} = 0$ for $j \in C$, $k \notin C$, since, e.g.,

$$p_{jk}^{(2)} = \sum_{s \in S} p_{js} p_{sk} = \sum_{s \in C} p_{js} p_{sk} = 0, \tag{2.66}$$

the case for general n following by induction. If C is closed and the process starts in C it will never leave it. An absorbing state is closed. We call a set C of states **irreducible** if $x \leftrightarrow y$ for all $x, y \in C$. The irreducible closed sets are precisely the equivalence classes under $\leftrightarrow$. If A and B are disjoint sets we write $A+B$ for their union.

Theorem 2.5 If S_T={transient states} and S_P={persistent states}, we have that

$$S = S_T + S_P \tag{2.67}$$

and

$$S_P = \sum C_i \text{ of disjoint, irreducible, closed sets.} \tag{2.68}$$

Proof Let $x \in S_P$, and define $C=\{y \in S_P : x \to y\}$. By persistence $f_{xx}=1$, so $x \in C$. We first show that C is closed. Suppose that $y \in C$, $y \to z$. Since y is persistent, z must also be persistent. Since $x \to y \to z$ we have $z \in C$ so that C is closed.

Next we show that C is irreducible. Choose y and z in C. We need to show that $z \leftrightarrow y$. Since $x \to y$, $y \to x$ by persistence. But $x \to z$ by definition of C, so $y \to x \to z$. The same argument, with y and z transposed, shows that $z \to y$.

Now let C and D be irreducible closed subsets of S_P, and let $x \in C \cap D$. Take $y \in C$. Since C is irreducible, $x \to y$. Since D is closed, $x \in D$, and $x \to y$ we have that $y \in D$. Thus $C \subset D$. Similarly $D \subset C$, so they are equal. □

It follows from this theorem that if a chain starts in C_i it will stay there forever (and we may as well let $S=C_i$). On the other hand, if it starts in S_T it either stays there forever, or moves into one of the C_i in which it stays forever.

Theorem 2.6 Within a persistent class either all the μ_i are finite or all are infinite.

Proof As before we can find k, m such that $p_{ij}^{(k)}>0$, $p_{ji}^{(m)}>0$. Since

$$p_{jj}^{(n+k+m)} \geq p_{ji}^{(m)} p_{ii}^{(n)} p_{ij}^{(k)} \tag{2.69}$$

we see by averaging and anticipating the result (2.64) that

$$\pi_j \geq p_{ji}^{(m)} \pi_i p_{ij}^{(k)}, \tag{2.70}$$

so if $\pi_i>0$ then $\pi_j>0$, while if $\pi_j=0$ then $\pi_i=0$. The converse obtains by interchanging i and j in the argument. □

Proposition 2.5 If S is finite, then at least one state is persistent, and all persistent states are positive.

Proof Assume that all states are transient. Then $\sum_j p_{ij}^{(n)}=1$ for all n. In particular,

$$\lim_{n\to\infty}\sum_{j\in S} p_{ij}^{(n)} = 1.$$

But from the corollary to Theorem 2.3, we have that each term in the sum, and therefore the entire sum, goes to zero. Hence at least one state is persistent. Assume that one such is state j. Consider $C_j = \{i{:}j\to i\}$. According to Theorem 2.5, once the process the process enters C_j it will stay there forever. For every $i\in C_j$ we can find a finite n with $p_{ij}^{(n)}>0$. For $i\neq j$ let ν_i denote the expected number of visits to i between two visits to j, i.e.,

$$\nu_i = \mathbf{E}^j\sum_{n=0}^{T_j-1} 1(X_n=i) = \sum_{n=0}^{\infty}\mathbf{P}^j(X_n=i,T_j>n). \tag{2.71}$$

Define $\nu_j=1$ in accordance with the definition of ν_i. Let $i\neq j$, and note that $\{X_n=i,T_k>n\}$ then is the same as $\{X_n=i,T_j>n-1\}$. Hence compute

$$\begin{aligned}
\nu_i &= \sum_{n=1}^{\infty}\mathbf{P}^j(X_n=i,T_j>n) = \sum_{n=1}^{\infty}\mathbf{P}^j(X_n=i,T_j>n-1)\\
&= \sum_{n=1}^{\infty}\sum_{k\in S}\mathbf{P}^j(X_n=k,T_j>n-1,X_{n-1}=i)\\
&= \sum_{n=1}^{\infty}\sum_{k\in S}\mathbf{P}^j(X_n=k\mid T_j>n-1,X_{n-1}=k)\,\mathbf{P}^j(T_j>n-1,X_{n-1}=k) \qquad (2.72)\\
&= \sum_{k\in S}p_{kj}\sum_{n=1}^{\infty}\mathbf{P}^j(X_n=k,T_j>n-1) = \sum_{k\in S}\sum_{m=0}^{\infty}\mathbf{P}^j(X_m=k,T_j>m)\\
&= \sum_{k\in S}p_{kj}\nu_k.
\end{aligned}$$

Since C_j is closed, the sum over $k\in S$ only has contributions from the states in C_j. For $i=j$ we have, since j is persistent, that

$$\begin{aligned}
\nu_j \equiv 1 &= \sum_{n=1}^{\infty}\mathbf{P}^j(T_j=n) = \sum_{n=1}^{\infty}\sum_{k\in S}\mathbf{P}^j(T_j=n,X_{n-1}=k)\\
&= \sum_{n=1}^{\infty}\sum_{k\in S}\mathbf{P}^j(T_j>n-1,X_n=j,X_{n-1}=k) \qquad (2.73)\\
&= \sum_{n=1}^{\infty}\sum_{k\in S}p_{kj}\mathbf{P}^j(T_j>n-1,X_n=k) = \sum_{k\in S}p_{kj}\nu_k.
\end{aligned}$$

Writing $\boldsymbol{\nu}=(\nu_1,\nu_2,\ldots)$ we have shown that $\boldsymbol{\nu}\mathbb{P}=\boldsymbol{\nu}$. By iterating we see that $\boldsymbol{\nu}\mathbb{P}^n=\boldsymbol{\nu}$ for all $n=1,2,\cdots$ In particular, for $i\in C_j$,

$$\nu_i p_{ij}^{(n)} \le \nu_j \equiv 1, \tag{2.74}$$

so $\nu_i \le 1/p_{ij}^{(n)} < \infty$ for some n. Finally we compute

$$\mu_j = \sum_{n=0}^{\infty} \mathbf{P}^j(T_j > n) = \sum_{i\in S} \sum_{n=0}^{\infty} \mathbf{P}^j(X_n = i, T_j > n) = \sum_{i\in S} \nu_i. \tag{2.75}$$

Again, the sum over $i \in S$ is really only over $i \in C_j$. Since S is finite, C_j must be finite, and we can pick n so large that $\nu_i < \infty$ for all i in C_j. The final sum in (2.75) therefore is a finite sum of finite elements, so $\mu_j < \infty$. □

Proposition 2.6 If i is a null persistent state, then $p_{ii}^{(n)} \to 0$ as $n \to \infty$.

This result was first proved by Erdös, Feller and Pollard (1949) using generating function techniques. The details are somewhat involved, and not of a probabilistic nature, so we shall omit the proof which can be found, e.g., in Feller (1968, sec. XIII.11). Incidentally, this proposition explains how null persistent states were named. Correspondingly, positive persistent states have $p_{ii}^{(n)} > 0$ for all n large enough.

2.4. Stationary distribution

A large portion of the theory of stochastic processes focuses on processes that have marginal distributions that are not time-dependent. Looking back at equation (2.16) we see that if we choose $p_1 = p_{01}/(1-(p_{11}-p_{01}))$, we obtain a marginal distribution which is independent of n, and simply equal to the initial distribution. We will denote such a distribution (when it exists) by $\boldsymbol{\pi}$. By letting $\mathbf{p}_n = \boldsymbol{\pi}$ in relation (2.20) we must have $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$, or equivalently

$$\boldsymbol{\pi}(\mathbb{I} - \mathbb{P}) = 0. \tag{2.76}$$

Thus $\boldsymbol{\pi}$ is a left eigenvector of $\mathbb{P}$, corresponding to the eigenvalue 1 (recall from Proposition 2.2 that such an eigenvalue always exists). The solution to (2.76) is called the **stationary distribution** of the Markov chain. If $S = \{0,1\}$ we saw that $\pi_1 = p_{01}/(1-(p_{11}-p_{01}))$. Thus if $p_{11} = p_{01}$, so that the occurrence of state 1 is independent of the previous state, then $\pi_1 = p_{01}$. Otherwise π_1 is between the smaller and the larger of p_{01} and p_{11}. To see that this choice of $\boldsymbol{\pi}$ indeed satisfies (2.76), note that

$$\left[\frac{1-p_{11}}{1-(p_{11}-p_{01})}, \frac{p_{01}}{1-(p_{11}-p_{01})}\right] \begin{bmatrix} p_{01} & -p_{01} \\ -(1-p_{11}) & 1-p_{11} \end{bmatrix} = (0,0). \tag{2.77}$$

When we use the stationary distribution as initial distribution we see that

$$\mathbf{p}_1 = \boldsymbol{\pi}\mathbb{P} = \boldsymbol{\pi} \tag{2.78}$$

$$\mathbf{p}_2 = \mathbf{p}_1\mathbb{P} = \boldsymbol{\pi},$$

etc., so that, indeed,

$$\mathbf{p}_n = \boldsymbol{\pi} \quad \text{for all } n. \tag{2.79}$$

We then say that the one-dimensional distributions $\mathbf{p}_n$ are **time invariant** (another name for this is stationary). Therefore $\boldsymbol{\pi}$ is also known as the **stationary initial distribution.** In fact, starting from $\boldsymbol{\pi}$ all finite-dimensional distributions are time invariant, in the sense that

$$(X_{k_1},X_{k_2},\ldots,X_{k_n}) \overset{d}{=} (X_{k_1+k},X_{k_2+k},\ldots,X_{k_n+k}) \tag{2.80}$$

for all non-negative integers n, k, $k_1,\ldots,k_n$ (Exercise **5**). Processes satisfying (2.80) are called **strictly stationary.**

The strength of the dependence in a Markov chain can be computed from the transition matrix. By repeated conditioning we see that

$$\mathbf{E}X_nX_{n+k} = \mathbf{E}(\mathbf{E}(X_{n+k} \mid X_n)X_n) = \sum_{j,l\in S} jlp_{lj}^{(k)}p_n(l). \tag{2.81}$$

If the chain is strictly stationary the right-hand side of (2.81) simplifies to $\sum jlp_{lj}^{(k)}\pi_l$, so the covariance between X_n and X_{n+k} is

$$\mathbf{Cov}(X_n,X_{n+k}) = \sum_{j,l\in S} jlp_{lj}^{(k)}\pi_l - (\sum_{j\in S} j\pi_j)^2. \tag{2.82}$$

Anticipating the next section we see that if the chain has a limiting distribution, so $p_{lj}^{(k)}\to\pi_j$, the covariance goes to zero as k goes to infinity.

Application (Snoqualmie Falls precipitation, continued) In the 0-1 case the sums in (2.82) only have one term. The correlation function for a two-state Markov chain thus becomes

$$\mathbf{Corr}(X_n,X_{n+k}) = \frac{p_{11}^{(k)}-\pi_1}{1-\pi_1} = (p_{11}-p_{01})^k \tag{2.83}$$

using the following induction argument. If $k=1$ we have

$$\frac{p_{11}-\pi_1}{1-\pi_1} = \frac{p_{11}(1-(p_{11}-p_{01}))-p_{01}}{1-p_{11}} = p_{11}-p_{01} \tag{2.84}$$

as required. Assuming the formula (2.83) is correct for $k=n$, then we can write it as

$$p_{11}^{(n)} = \pi_1 + (1-\pi_1)(p_{11}-p_{01})^n. \tag{2.85}$$

Since $p_{11}^{(n+1)}=(1-p_{11}^{(n)})p_{01}+p_{11}^{(n)}p_{11}$ we have

$$\begin{aligned}\frac{p_{11}^{(n+1)}-\pi_1}{1-\pi_1} &= \frac{p_{01}+(1-\pi_1)(p_{11}-p_{01})^{n+1}+\pi_1(p_{11}-p_{01})-\pi_1}{1-\pi_1} \\ &= (p_{11}-p_{01})^{n+1} + \frac{p_{01}-\pi_1(1-p_{11}+p_{01})}{1-\pi_1} \qquad (2.86)\\ &= (p_{11}-p_{01})^{n+1}\end{aligned}$$

where the last equality is from the definition of π_1. This completes the induction. We see that the correlation function is geometrically decreasing. For the Snoqualmie Falls data, where $\hat{p}_{11}-\hat{p}_{01}$ is 0.436, we get the estimated correlations given in Table 22.

Table 2.2 Estimated correlations for Snoqualmie Falls data

Lag	1	2	3	4	5	6	7
Corr	0.436	0.190	0.083	0.036	0.016	0.007	0.003

□

Equation (2.76) shows that if a chain has a stationary distribution, it must be the eigenvector of $\mathbb{P}$ corresponding to the eigenvalue 1. Sometimes we can be a bit more explicit. Recall that if k is persistent, then ν_j is the expected number of visits to j before returning to k.

Lemma 2.4 An irreducible positive persistent chain has a stationary distribution given by $\pi_i=\nu_i/\mu_k$ for a fixed state k.

Proof We need to show that $\boldsymbol{\pi}\mathbb{P}=\boldsymbol{\pi}$, i.e., that $\sum_{i\in S}\nu_i p_{ij}=\nu_j$ and that $\sum_{i\in S}\nu_i=\mu_k$ (in order for $\boldsymbol{\pi}$ to be a probability distribution). But this was established in the proof of Proposition 2.5 (without using the assumption of a finite state space). □

Which Markov chains have a stationary distribution? The answer is quite simple. We will restrict attention to irreducible chains, since any other chain can be be decomposed into irreducible subclasses. The quantity $\pi_i=1/\mu_i$, which arose in our criterion for positive persistence in Lemma 2.3, now assumes a more important role.

Theorem 2.7 An irreducible chain has a stationary distribution if and only if it is positive persistent. The stationary distribution is unique and given by $\pi_i=\mu_i^{-1}$.

Proof Suppose that $\boldsymbol{\pi}$ is a stationary distribution and the chain is transient or null persistent. Then $p_{ij}^{(n)}\to 0$ as $n\to\infty$ by the corollary to Theorem 2.3 and by Proposition 2.6, respectively. Hence for any i and j, if we are allowed to take limits under the summation sign,

$$\pi_j=\sum_{i\in S}\pi_i p_{ij}^{(n)}\to 0 \quad \text{as } n\to\infty \tag{2.87}$$

so $\boldsymbol{\pi}$ is not a distribution. To see that this argument is valid, let (S_m) be a sequence of finite subsets of S, such that $S_m\uparrow S$ as $m\to\infty$. Then

$$\pi_j = \sum_{i \in S_m} \pi_i p_{ij}^{(n)} + \sum_{i \notin S_m} \pi_i p_{ij}^{(n)} \le \sum_{i \in S_m} \pi_i p_{ij}^{(n)} + \sum_{i \notin S_m} \pi_i. \tag{2.88}$$

For each m the first term goes to zero as $n \to \infty$, since we can always take the elementwise limit of a finite sum. The second term can be made arbitrarily small by taking m large, since $\boldsymbol{\pi}$ is summable. The key to this argument is that the $p_{ij}^{(n)}$ are bounded. Thus the existence of a stationary distribution implies that all states are persistent.

Let the initial distribution be $\boldsymbol{\pi}$. Using Exercise **5**, all finite-dimensional probabilities are time invariant. Thus

$$\begin{aligned}
\mathbf{P}^{\pi}(X_n=i, T_j \ge n+1) &= \mathbf{P}^{\pi}(X_n=i, X_1, \ldots, X_n \ne j) \\
&= \mathbf{P}^{\pi}(X_{n-1}=i, X_0, \ldots, X_{n-1} \ne j) \qquad (2.89)\\
&= \mathbf{P}^{\pi}(X_{n-1}=i, T_j \ge n) - \mathbf{P}^{\pi}(X_{n-1}=i, X_0=j, T_j \ge n) \\
&= \mathbf{P}^{\pi}(X_{n-1}=i, T_j \ge n) - \pi_j \mathbf{P}^j(X_{n-1}=i, T_j \ge n).
\end{aligned}$$

Summing over $n \le N$ and rearranging terms yields

$$\begin{aligned}
\sum_{n=1}^{N} \pi_j \mathbf{P}^j(X_{n-1}=i, T_j \ge n) &= \sum_{n=1}^{N} (\mathbf{P}^{\pi}(X_{n-1}=i, T_j \ge n) - \mathbf{P}^{\pi}(X_n=i, T_j \ge n+1)) \\
&= \pi_i - \mathbf{P}^{\pi}(X_N=i, T_j \ge N+1) \qquad (2.90)
\end{aligned}$$

since the sum telescopes. Letting $N \to \infty$ the last term on the right-hand side of (2.90) disappears since j is ergodic, and we get

$$\sum_{n=1}^{\infty} \pi_j \mathbf{P}^j(X_{n-1}=i, T_j \ge n) = \pi_i. \tag{2.91}$$

Summing (2.91) over all $i \in S$ we see that

$$\pi_j \sum_{n=1}^{\infty} \mathbf{P}^j(T_j \ge n) = 1. \tag{2.92}$$

But $\sum P^j(T_j \ge n) = \mu_j$ and we see that $\pi_j \mu_j = 1$. Suppose that $\pi_i = 0$. Then

$$0 = \pi_i = \sum_j \pi_j p_{ji}^{(n)} \ge \pi_j p_{ji}^{(n)}, \tag{2.93}$$

so whenever $j \to i$ we have $\pi_j = 0$. But then all the π_i are 0 by irreducibility, and $\boldsymbol{\pi}$ is not a distribution. Hence all the π_i are positive, so $\mu_j < \infty$. Therefore, the existence of a stationary distribution for an irreducible chain implies that it is uniquely given by $\pi_i = \mu_i^{-1}$, and that the chain must be positive persistent. Conversely, for a positive persistent chain the distribution in Lemma 2.4 is a stationary distribution. □

We can now evaluate the expected number of visits to i between successive visits to k.

Corollary $\quad \nu_i = \mathbf{E}^k \sum_{n=0}^{T_k-1} 1(X_n=i) = \mu_k/\mu_i = \pi_i/\pi_k$

Proof This follows directly from Lemma 2.4 and the uniqueness in Theorem 2.7. □

The equation (2.76) for the stationary distribution can be written

$$\pi_j = \sum_{i\in S} \pi_j p_{ji} = \sum_{i\in S} p_{ij}\pi_i. \tag{2.94}$$

We can interpret $\sum_i \pi_j p_{ji}$ as the probability flux out of state j, and $\sum_i \pi_i p_{ij}$ as the probability flux into state j. Consider a large number of independent particles following the same Markov chain. Then, if the system is in equilibrium, the number of particles moving into and out of state i at any time should be approximately the same. In other words, the proportion of particles moving out (the flux out of the state) should be the same as the proportion of particles moving in (the flux into the state). In this interpretation, it is natural to think of (2.94) as an equation of **full balance**.

Many physical systems, obeying classical mechanics, have a physical description that is symmetric with respect to past and future. In the context of stochastic processes, the corresponding requirement is that the probabilistic structure of the process run forward in time must be the same as the structure of the process run backward in time.

Let $(X_k, k\in\mathbf{Z})$ be an ergodic chain, defined for both positive and negative time. We may consider the chain Y defined by $Y_k=X_{-k}$. Then Y is a Markov chain, although not necessarily with stationary transition probabilities:

$$\begin{aligned}
&\mathbf{P}(Y_{k+1}=j \mid Y_k=i, Y_{k-1}=i_1, \ldots, Y_{k-n}=i_n) \\
&= \mathbf{P}(X_{-(k+1)}=j \mid X_{-k}=i, X_{-(k-1)}=i_1, \ldots, X_{-(k-n)}=i_n) \\
&= \frac{\mathbf{P}(X_{-(k+1)}=j, X_{-k}=i, X_{-(k-1)}=i_1, \ldots, X_{-(k-n)}=i_n)}{\mathbf{P}X_{-k}=i, X_{-(k-1)}=i_1, \ldots, X_{-(k-n)}=i_n)} \\
&= \frac{\mathbf{P}(X_{-(k-n)}=i_n, \ldots, X_{-(k-1)}=i_1 \mid X_{-k}=i)\mathbf{P}(X_{-k}=i \mid X_{-(k+1)}=j)}{\mathbf{P}(X_{-(k-n)}=i_n, \ldots, X_{-(k-1)}=i_1 \mid X_{-k}=i)} \\
&\times \frac{\mathbf{P}(X_{-(k+1)}=j)}{\mathbf{P}(X_{-k}=i)} = p_{ji}\frac{p_j^{(-(k+1))}}{p_i^{(-k)}}.
\end{aligned} \tag{2.95}$$

If X has the stationary marginal distribution $\boldsymbol{\pi}$ for all k we see that Y has stationary transition probabilities q_{ij} given by

$$q_{ij} = \mathbf{P}(Y_{k+1}=j \mid Y_k=i) = p_{ji}\frac{\pi_j}{\pi_i}. \tag{2.96}$$

Note that since X is defined for all $k \in \mathbf{Z}$, it is not enough to set $\mathbf{p}^{(0)}=\boldsymbol{\pi}$ in order to have marginal distribution $\boldsymbol{\pi}$ for all k. This only works for $k \geq 0$; e.g., X_{-1} does not necessarily yield the right distribution. We call X **reversible** if X and Y have the same transition matrix, i.e. if

$$p_{ij} = p_{ji}\frac{\pi_j}{\pi_i} \tag{2.97}$$

or, equivalently,

$$\pi_i p_{ij} = p_{ji}\pi_j. \tag{2.98}$$

This is called the law of **detailed balance**, stating that the probability flux from i to j in equilibrium is the same as that from j to i. Detailed balance is a property of isolated systems in both classical and quantum mechanics. It was first noted in chemical reaction kinetics. A proof of the detailed balance property for closed classical systems is in Van Kampen (1981, section V.6). The conditions of detailed balance can sometimes be used to find the stationary distribution of a chain.

Theorem 2.8 If, for an irreducible Markov chain, a distribution $\boldsymbol{\pi}$ exists, satisfying the law of detailed balance (2.98) for all $i,j \in S$, then the chain is reversible and positive persistent with stationary distribution $\boldsymbol{\pi}$.

Proof Using Theorem 2.7 we need only show that $\boldsymbol{\pi}$ is a stationary distribution. But

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j, \tag{2.99}$$

so $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$. □

Example (A Birth and death chain) Assume that $p_j>0$, $q_{j+1}>0$ for all $j \geq 0$, while $q_0=0$, so that all states communicate. Then we will show that the detailed balance equation

$$p_j\pi_j = q_{j+1}\pi_{j+1} \tag{2.100}$$

holds, and that the equilibrium distribution is given by

$$\pi_j = \pi_0 \prod_{i=1}^{j} \frac{p_{i-1}}{q_i} \tag{2.101}$$

where

$$\pi_0^{-1} = \sum_{j=0}^{\infty} \prod_{i=1}^{j} \frac{p_{i-1}}{q_i} \tag{2.102}$$

provided that the sum converges. This convergence is a condition for the persistence of the chain as well as for its reversibility.

To see that (2.100) holds, note first that the full balance equation for π_j is

$$\pi_j = p_{j-1}\pi_{j-1} + r_j\pi_j + q_{j+1}\pi_{j+1}, \ j>0, \tag{2.103}$$

while for $j=0$

$$\pi_0 = r_0\pi_0 + q_1\pi_1. \tag{2.104}$$

Since $q_0=0$, $r_0=1-p_0$ and (2.104) becomes

$$p_0\pi_0 = q_1\pi_1 \tag{2.105}$$

which is the detailed balance equation for $j=0$. Assume that (2.100) holds for $j=k$. From (2.104) we see that

$$\pi_{k+1} = p_k\pi_k + r_{k+1}\pi_{k+1} + q_{k+2}\pi_{k+2}. \tag{2.106}$$

Since by the induction hypothesis $p_k\pi_k=q_{k+1}\pi_{k+1}$ (2.106) becomes

$$p_{k+1}\pi_{k+1} = q_{k+2}\pi_{k+2} \tag{2.107}$$

whence the detailed balance equation holds. The evaluation of π is now immediate.

The particular case of a random walk reflected at the origin has $p_i=1-q_i=p$, so

$$\pi_0^{-1} = \sum \left[\frac{p}{1-p}\right]^j = \frac{1-p}{2-p}, \tag{2.108}$$

provided that $p<½$. The stationary distribution is then geometric. If $p\geq½$ the process is transient. □

Example (The Ehrenfest model for diffusion) Consider two containers, labeled 0 and 1, in contact with each other. We have N molecules that move between the containers. At each time one molecule is chosen at random, and moved to the other container. We can describe this system using a binary number of N digits, for a total of 2^N possible states. The transition probabilities for this **micro-level** process X are

$$p_{x,x'} = \begin{cases} 1/N & \text{if } x \text{ and } x' \text{ only differ in one location} \\ 0 & \text{otherwise.} \end{cases} \tag{2.109}$$

Consider the case $N=2$, so the states are 00, 01, 10, and 11 (or in decimal notation 0,...,3). Then

$$\mathbb{P} = \begin{bmatrix} 0 & ½ & ½ & 0 \\ ½ & 0 & 0 & ½ \\ ½ & 0 & 0 & ½ \\ 0 & ½ & ½ & 0 \end{bmatrix}. \tag{2.110}$$

Clearly $\mathbb{P}$ is doubly stochastic, whence the stationary distribution is $\pi_i=2^{-N}$, $i=0,\ldots,2^N-1$. The chain is periodic with period 3. It satisfies not only the detailed balance equation, but also the stronger **micro-reversibility** condition that

$$p_{x,x'} = p_{x',x} \quad \text{for all } x,x'. \tag{2.111}$$

If we regard the molecules as indistinguishable, we get a **macro-level** description of the process. Let Y_k be the number of molecules in container 0. Then Y_k is a Markov chain with non-zero transition probabilities

$$p_{Y;i,i-1} = \frac{i}{N} \quad p_{Y;i,i+1} = \frac{N-i}{N}. \tag{2.112}$$

Since this is a birth and death chain we know from the previous example that the process is reversible and the stationary distribution satisfies

$$\pi_j = \pi_0 \prod_{i=1}^{j} \frac{N-i+1}{i} = \pi_0 \binom{N}{j}, \tag{2.113}$$

so $\pi_0=2^{-N}$. In other words, the stationary distribution for Y is obtained from that of X by summing over the number of micro-states corresponding to a given macro-state.

This model was introduced by Ehrenfest and Ehrenfest (1906) to explain a paradox in thermodynamics, exposed by Loschmidt (1876). The paradox is that although statistical mechanics can be derived from classical mechanics, the laws of classical mechanics are time-reversible while thermodynamics contains irreversible processes: entropy must increase with time. This physical sense of reversibility would require that for given micro-states x and x', with corresponding macro-states y and y' we have both

$$\mathbf{P}(X_k=x \mid X_0=x') = \mathbf{P}(X_k=x' \mid X_0=x) \tag{2.114}$$

and

$$\mathbf{P}(Y_k=y \mid Y_0=y') = \mathbf{P}(Y_k=y' \mid Y_0=y). \tag{2.115}$$

If now y is small, and y' is nearly $N/2$, (2.114) holds by micro-reversibility, but (2.115) would not hold. Rather, the right-hand side would be much larger than the left-hand side, because of a tendency for the process to veer towards its stationary mean (we are anticipating the results of the next section here, in that the process in the long run tends towards its stationary distribution). The statistical sense of reversibility involves equilibrium behavior, which the classical mechanics laws do not explicitly mention. Our explanation of the Loschmidt paradox, therefore, will be that the process is not micro-reversible at the macro-level. Chandrasekhar (1943, section III.4) and Whittle (1986) contain more material pertinent to this type of question. □

ice the stationary distribution is $\pi_i=2^{-N}$,
: with period 3. It satisfies not only the
e stronger **micro-reversibility** condition

(2.111)

guishable, we get a **macro-level** descrip-
oer of molecules in container 0. Then Y_k is
on probabilities

$$\frac{N-i}{N}. \quad (2.112)$$

we know from the previous example that
ary distribution satisfies

$$\left.\begin{matrix} N \\ j \end{matrix}\right), \quad (2.113)$$

ry distribution for Y is obtained from that
of micro-states corresponding to a given

Ehrenfest and Ehrenfest (1906) to explain
ed by Loschmidt (1876). The paradox is
be derived from classical mechanics, the
eversible while thermodynamics contains
ncrease with time. This physical sense of
r given micro-states x and x', with
e have both

$$= x' \mid X_0 = x) \quad (2.114)$$

$$= y' \mid Y_0 = y). \quad (2.115)$$

(2.114) holds by micro-reversibility, but
ght-hand side would be much larger than
cy for the process to veer towards its sta-
esults of the next section here, in that the
ts stationary distribution). The statistical
librium behavior, which the classical
ation. Our explanation of the Loschmidt
process is not micro-reversible at the
ection III.4) and Whittle (1986) contain
question. □

2.5. Long term behavior

As we have discussed before, many physical systems tend to settle down to an **equilibrium state**, where the state occupation probabilities are independent of the initial probabilities. Recall that when we powered up $\hat{\mathbb{P}}$ for the Snoqualmie Falls precipitation model, the rows got more and more similar. Figure 2.2 illustrates this.

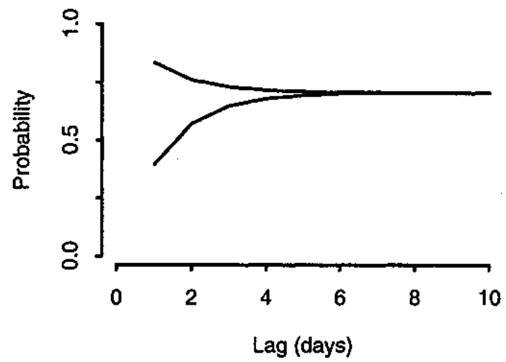

Figure 2.2. n-step transition probabilities for Snoqualmie Falls model. The upper curve is $p_{11}^{(n)}$ while the lower is $p_{01}^{(n)}$.

In fact, under suitable conditions

$$\mathbb{P}^n \to \begin{bmatrix} \pi \\ \pi \\ \cdots \\ \pi \end{bmatrix}. \quad (2.116)$$

We say that the chain has a **limiting distribution.** What this means is that if the chain is left running for a long time, it reaches an equilibrium situation regardless of its initial distribution. In this equilibrium situation the state occupancy probabilities are equal to the stationary distribution. Note namely that

$$\mathbf{p}_n = \mathbf{p}_0 \mathbb{P}_n \to \mathbf{p}_0 \begin{bmatrix} \pi \\ \pi \\ \cdots \\ \pi \end{bmatrix} = \pi$$

regardless of $\mathbf{p}_0$. As the next example shows, there may be a stationary distribution without the chain having a limiting distribution.

Example (A chain without a limiting distribution) Let

$$\mathbb{P} = \begin{bmatrix} 0\,1\,0 \\ 0\,0\,1 \\ 1\,0\,0 \end{bmatrix}. \quad (2.117)$$

Then $\pi=(1,1,1)/3$, but $\mathbb{P}^n$ does not converge. Rather, it cycles through three different matrices. Notice that the period of this chain is 3. This particular $\mathbb{P}$ also is doubly stochastic, and has a uniform stationary distribution (Exercise **6**).

In fact, this trivial example is in many ways typical for what happens in a periodic chain. Assume that the irreducible chain X has period d. Then for every state k we can find integers l,m such that $p_{0k}^{(l)}>0$ and $p_{k0}^{(m)}>0$. Hence $p_{00}^{(l+m)}\geq p_{0k}^{(l)}p_{k0}^{(m)}>0$, so d must divide $l+m$, i.e., $l+m=rd$ for some integer r. Fixing m we see that $l=-m+rd\equiv s+td$ where s is the residue class of m modulo d. Thus we can for every state k find an integer s_k, $0\leq s_k<d$, such that $p_{0k}^{(n)}=0$ unless $n\equiv s_k \bmod d$. Let $G_s=\{k: s_k=s\}$. Then

$$S = G_0+\cdots+G_{d-1}. \tag{2.118}$$

One-step transitions are only possible from states in G_k to states in G_{k+1} (where $G_d\equiv G_0$), and going d steps out of G_k leads back to G_k. Hence for a chain with transition matrix $\mathbb{P}^d$, each G_k is a closed irreducible set. For $d=3$ we have

$$\mathbb{P} = \begin{bmatrix} 0 & A & 0 \\ 0 & 0 & B \\ C & 0 & 0 \end{bmatrix}, \tag{2.119}$$

where A consists of transition probabilities from G_0 to G_1, etc. □

We shall now ascertain the long term behavior of some aspects of a Markov chain. Again we restrict attention to the irreducible case. We start with the asymptotic behavior of n-step transition probabilities.

Theorem 2.9 Let k be an aperiodic state of an irreducible Markov chain with mean recurrence time $\mu_k\leq\infty$. Then

$$\lim_{n\to\infty} p_{kk}^{(n)} = \frac{1}{\mu_k}. \tag{2.120}$$

Proof The transient case is immediate from the corollary to Theorem 2.3, and the null persistent case is Proposition 2.6. To prove the positive persistent case we shall use a technique called **coupling**. Let X and Y be iid copies of the chain, and let $Z=(X,Y)$. Recall from the corollary to Proposition 2.4 that Z is an irreducible Markov chain with transition probabilities $p_{ij,kl}=p_{ik}p_{jl}$. Since X is positive persistent, it has a unique stationary distribution π. Then Z has stationary distribution η with $\eta_{i,j}=\pi_i\pi_j$. Therefore Z is positive persistent by Theorem 2.7. Let $Z_0=(i,j)$, choose $s\in S$, and let $T_{s,s}$ be the hitting time of (s,s) for Z. Since Z is persistent, $\mathbf{P}(T_{s,s}<\infty)=1$. Suppose that $m\leq n$ and that $X_m=Y_m$. Then X_n and Y_n are identically distributed by the strong Markov property. Thus, conditional on $\{T_{s,s}\leq n\}$, the random variables X_n and Y_n have the same conditional distribution. Compute

$$p_{ik}^{(n)} = \mathbf{P}^{i,j}(X_n=k) = \mathbf{P}^{i,j}(X_n=k,T_{s,s}\leq n) + \mathbf{P}^{i,j}(X_n=k,T_{s,s}>n)$$

iod of this chain is 3. This particular $\mathbb{P}$
form stationary distribution (Exercise 6).

many ways typical for what happens in a
cible chain X has period d. Then for every
such that $p^{(l)}_{0k}>0$ and $p^{(m)}_{k0}>0$. Hence
$+m$, i.e., $l+m=rd$ for some integer r. Fix-
ere s is the residue class of m modulo d.
nteger s_k, $0\le s_k<d$, such that $p^{(n)}_{0k}=0$ unless

(2.118)

from states in G_k to states in G_{k+1} (where
leads back to G_k. Hence for a chain with
ed irreducible set. For $d=3$ we have

(2.119)

ities from G_0 to G_1, etc. □

n behavior of some aspects of a Markov
the irreducible case. We start with the
n probabilities.

dic state of an irreducible Markov chain
en

(2.120)

ediate from the corollary to Theorem 2.3,
ition 2.6. To prove the positive persistent
oupling. Let X and Y be iid copies of the
he corollary to Proposition 2.4 that Z is an
tion probabilities $p_{ij,kl}=p_{ik}p_{jl}$. Since X is
tionary distribution $\boldsymbol{\pi}$. Then Z has station-
refore Z is positive persistent by Theorem
let $T_{s,s}$ be the hitting time of (s,s) for Z.
ppose that $m\le n$ and that $X_m=Y_m$. Then X_n
the strong Markov property. Thus, condi-
bles X_n and Y_n have the same conditional

$${}_n=k,T_{s,s}\le n) + \mathbf{P}^{i,j}(X_n=k,T_{s,s}>n)$$

$$= \mathbf{P}^{i,j}(Y_n=k,T_{s,s}\le n) + \mathbf{P}^{i,j}(X_n=k,T_{s,s}>n) \tag{2.121}$$
$$\le \mathbf{P}^{i,j}(Y_n=k) + \mathbf{P}^{i,j}(T_{s,s}>n) = p^{(n)}_{jk} + \mathbf{P}^{i,j}(T_{s,s}>n).$$

Interchanging i and j, we get a similar result, so that

$$|\,p^{(n)}_{ik} - p^{(n)}_{jk}\,| \le \mathbf{P}^{i,j}(T_{s,s}>n) + \mathbf{P}^{j,i}(T_{s,s}>n) \to 0 \text{ as } n\to\infty \tag{2.122}$$

since $\mathbf{P}^{i,j}(T_{s,s}<\infty)=1$ for any i,j. Thus

$$p^{(n)}_{ik} - p^{(n)}_{jk} \to 0 \tag{2.123}$$

as $n\to\infty$ for any i, j, and k. Consequently, if $p^{(n)}_{ik}$ has a limit it does not depend on i. But

$$\pi_k - p^{(n)}_{jk} = \sum_{i\in S}\pi_i(p^{(n)}_{ik} - p^{n}_{jk}) \to 0 \tag{2.124}$$

by bounded convergence. □

Example (The Pólya urn model) Quite a few stochastic processes were originally thought of using colored balls in urns. A paper by Eggenberger and Pólya[1] (1923) dealt with epidemic data for contagious diseases. Given that an individual has a disease, such as smallpox, the probability that other individuals who are in contact with the diseased one themselves become infected is higher than for people who have had no such contact. Hence individuals do not act independently as far as epidemics are concerned.

Eggenberger and Pólya proposed the following urn scheme: consider an urn with N balls, R of which are red and B are black. A ball is pulled out of the urn at random and replaced with $1+d$ balls of the same color. Clearly $d=0$ corresponds to drawing with replacement, and $d=-1$ to drawing without replacement. After the kth replacement the urn has $R+B+dk$ balls. If the draws yield r red and b black balls ($r+b=k$), there are $R+rd$ red and $B+bd$ black balls, whence the probability of a red ball drawn at the $(k+1)$th draw is $(R+rd)/(N+kd)$. Let $X_k=1$(red ball drawn at trial k). Then, with $\mathbf{X}=(X_1,\ldots,X_n)$ and $r_n=\sum_1^n x_i$,

$$\mathbf{P}(\mathbf{X}=\mathbf{x}) = \frac{\prod_{j=1}^{r_n}(R+(r_n-j)d)\prod_{k=1}^{n-r_n}(B+(n-r_n-k)d)}{(N+(n-1)d)(N+(n-2)d)\cdots N}. \tag{2.125}$$

Conditioning on the past yields

$$\mathbf{P}(X_n=x_n\,|\,X_0^{n-1}=x_0^{n-1}) = \begin{cases} \dfrac{R+r_{n-1}d}{N+(n-1)d} & \text{if } x_n=1, \\[2ex] \dfrac{B+(n-1-r_{n-1})d}{N+(n-1)d} & \text{if } x_n=0. \end{cases} \tag{2.126}$$

[1] Pólya, György (1887–1985). Hungarian mathematician. Invented the term "random walk". Perhaps best known for his ideas about general heuristics for solving problems.

Thus X_n is not a Markov chain, unless $d=0$. However, the number of red balls drawn, $R_n=\sum_1^n X_i$, is a Markov chain:

$$\mathbf{P}(R_n=r \mid R_0^{n-1}=r_0^{n-1}) = \begin{cases} \dfrac{R+dr_{n-1}}{N+d(n-1)} & \text{if } r=r_{n-1}+1 \\ \dfrac{B+d(n-1-r_{n-1})}{N+d(n-1)} & \text{if } r=r_{n-1} \end{cases}. \tag{2.127}$$

Note, however, that the transition probabilities for R_n are time-dependent, since they depend explicitly on $n-1$, not only on r_{n-1}. It is not hard (Exercise 7) to derive the marginal distribution of R_n, which is

$$\begin{aligned}\mathbf{P}(R_n=r) = &\binom{n}{r}\times R(R+d)\cdots(R+(r-1)d)\\ &\times\frac{B(B+d)\cdots(B+(n-r)d)}{N(N+d)\cdots(N+(n-1)d)}.\end{aligned} \tag{2.128}$$

Consider the case where n is large and R small relative to N, corresponding to a rare disease. In particular, let $R=hN/n$ and $d=cN/n$. By taking limits in (2.128) we see that

$$\lim_{n\to\infty}\mathbf{P}(R_n=r) = \frac{1}{r!}(1+c)^{-(h/c+r)}h(h+c)\cdots(h+(r-1)c) \tag{2.129}$$

which is a negative binomial distributions with parameters h/c and $c/(1+c)$, and mean h. When $c\to 0$ this limit is just the Poisson approximation to the binomial. □

Corollary For an irreducible aperiodic chain

$$\lim_{n\to\infty} p_{jk}^{(n)} = \frac{f_{jk}}{\mu_k}. \tag{2.130}$$

Proof Recall that $p_{jk}^{(n)}=\sum_{l=0}^n f_{jk}^{(l)}p_{kk}^{(n-l)}$. Taking limits under the summation sign we get

$$p_{jk}^{(n)}\to(1/\mu_k)\sum f_{jk}^{(l)}=f_{jk}/\mu_l. \tag{2.131}$$

To verify that we can take limits under the summation sign, write

$$p_{jk}^{(n)} = \sum_{l=0}^{m-1} f_{jk}^{(l)}p_{kk}^{(n-l)} + \sum_{l=m}^{n} f_{jk}^{(l)}p_{kk}^{(n-l)}. \tag{2.132}$$

Since $p_{kk}^{(n)}\le 1$ for all n we have

$$\sum_{l=0}^{m-1} f_{jk}^{(l)}p_{kk}^{(n-l)} \le p_{jk}^{(n)} \le \sum_{l=0}^{m-1} f_{jk}^{(l)}p_{kk}^{(n-l)} + \sum_{l=m}^{n} f_{jk}^{(l)}. \tag{2.133}$$

Now let $n\to\infty$ to see that

$$\sum_{l=0}^{m-1} f_{jk}^{(l)}/\mu_k \le \liminf_{n\to\infty} p_{jk}^{(n)} \le \limsup_{n\to\infty} p_{jk}^{(n)} \le \sum_{l=0}^{m-1} f_{jk}^{(l)}/\mu_k + \sum_{l=m}^{\infty} f_{jk}^{(l)}, \quad (2.134)$$

and finally let $m \to \infty$ to obtain the result. □

Let $N_k(n)=\sum_{i=1}^{n} 1(X_i=k)$ count the time spent in state k.

Corollary If k is a persistent aperiodic state, then

$$\lim_{n\to\infty} \mathbf{E} N_k(n)/n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} p_{kk}^{(i)} = \frac{1}{\mu_k} \quad (2.135)$$

for any starting state communicating with k.

Remark The limit in the corollary is called a **Cesàro limit** of the $p_{kk}^{(i)}$. The existence of a Cesàro limit is implied by, but does not imply, the existence of a limit of the sequence. □

Proof By Theorem 2.9 we have $p_{kk}^{(i)} \to 1/\mu_k$. By the remark above, this implies that the Cesàro limit is the same. Now notice that

$$\mathbf{E}^k N_k(n) = \sum_{i=1}^{n} \mathbf{P}^k(X_i=k) = \sum_{i=1}^{n} p_{kk}^{(i)}. \quad (2.136)$$

If we instead start at j, communicating with k, we get

$$\mathbf{E}^j N_k(n) = \sum_{i=1}^{n} \mathbf{P}^j(X_i=k) = \sum_{i=1}^{n} p_{jk}^{(i)} \quad (2.137)$$

and by Corollary 1 above $p_{jk}^{(n)} \to f_{jk}/\mu_k$, so the same holds for the Cesàro limit. Since j communicates with k we have $f_{jk}=1$. □

Consider a persistent state k. The **limiting occupation probability** is the proportion of time spent in that state in an infinitely long realization, i.e., $\lim_{n\to\infty} N_k(n)/n$. In Corollary 2 we computed the expected value of this average. The next result yields a law of large numbers.

Theorem 2.10 The limiting occupation probability of an ergodic state is $1/\mu_k$ (with probability 1).

Proof Suppose that the chain starts in state k. Let $T_k(1), T_k(2), \cdots$ be the successive times when the chain reaches k. By the strong Markov property $T_k(1), T_k(2)-T_k(1), T_k(3)-T_k(2), \ldots$ are iid random variables with pgf $F_{kk}(s)$ and mean $\mu_k<\infty$. By the strong law of large numbers we have with probability one that

$$\lim_{r\to\infty}\frac{T_k(1)+(T_k(2)-T_k(1))+\cdots+(T_k(r)-T_k(r-1))}{r}$$

$$=\lim_{r\to\infty}\frac{T_k(r)}{r}=\mu_k. \tag{2.138}$$

Recall that $N_k(n)/n$ is the proportion of time spent in state k up to time n. Thus

$$T_k(N_k(n)) \le n \le T_k(N_k(n)+1). \tag{2.139}$$

In addition, $N_k(n)\to\infty$ as $n\to\infty$, again with probability one, since k is revisited infinitely often. Thus

$$\frac{N_k(n)}{n} \le \frac{N_k(n)}{T_k(N_k(n))} \to \frac{1}{\mu_k} \text{ with probability 1} \tag{2.140}$$

and

$$\frac{N_k(n)+1}{n} \ge \frac{N_k(n)+1}{T_k(N_k(n)+1)} \to \frac{1}{\mu_k} \text{ with probability 1} \tag{2.141}$$

so that $N_k(n)/n \to 1/\mu_k$ a.s. The case when the process starts from a state other than k is left as Exercise **8**. □

Example (The Hardy–Weinberg law) Consider a large population of individuals, each of whom possesses a particular pair of genes. We classify each gene as type A or type a. Assume that when two individuals mate, each contributes a randomly chosen gene to the resulting offspring, and assume also that mates are selected at random from the population. Write the proportion of individuals in the population with AA, Aa, and aa genes, respectively, as p, q, and r. Then the proportion of A-genes in the population is $P=p+q/2$ and the proportion of a-genes is $Q=q/2+r$. Under random mating, therefore, an individual will have probability P^2 of receiving the gene combination AA, probability $2PQ$ of receiving Aa, and probability Q^2 of receiving aa. Hence in the next generation the proportion of A-genes is $P^2+PQ=P$, and the proportion of a-genes is Q. We see that the proportions of gene types as well as the proportion of gene pairs remain stable after the first mating. This is called the **Hardy–Weinberg law** (Hardy[1], 1908; Weinberg, 1908). Assume now that we have a population with $P^2{:}2PQ{:}Q^2$ gene pair ratio, and consider the genetic history of a single individual, assuming for simplicity that each individual has exactly one offspring. If X_n is the genetic state of the nth descendant we have a Markov chain with state space {AA,Aa,aa}, and transition matrix

[1] Hardy, Godfrey Harold (1877–1947). English pure mathematician. His main contributions came through his long collaboration with Littlewood on problems in number theory, inequalities, and complex analysis. He was apparently not very fond of this non-theoretical paper, which he published in an obscure American journal.

$\cdot + (T_k(r) - T_k(r-1))$

(2.138)

ne spent in state k up to time n. Thus

(2.139)

th probability one, since k is revisited

with probability 1 (2.140)

$\frac{1}{\mu_k}$ with probability 1 (2.141)

n the process starts from a state other □

w) Consider a large population of
ticular pair of genes. We classify each
ien two individuals mate, each contri-
ulting offspring, and assume also that
pulation. Write the proportion of indi-
d aa genes, respectively, as p, q, and r.
pulation is $P=p+q/2$ and the propor-
dom mating, therefore, an individual
e gene combination AA, probability
of receiving aa. Hence in the next gen-
$Q=P$, and the proportion of a-genes is
ypes as well as the proportion of gene
This is called the **Hardy–Weinberg**
sume now that we have a population
nsider the genetic history of a single
at each individual has exactly one
e nth descendant we have a Markov
ansition matrix

e mathematician. His main contributions came
n problems in number theory, inequalities, and
fond of this non-theoretical paper, which he

$$\mathbb{P} = \begin{bmatrix} P & Q & 0 \\ P/2 & (P+Q)/2 & Q/2 \\ 0 & P & Q \end{bmatrix}. \tag{2.142}$$

From the Hardy–Weinberg law it would seem natural that the stationary distribution for this chain, which, by the theorem, is also the proportion of descendants in each genetic state in the long run, should be $(P^2, 2PQ, Q^2)$. This is indeed the case (Exercise **9**). □

It is possible to deduce more general laws of large numbers. The following, which we shall find particularly useful later, is often called the **ergodic theorem for Markov chains.**

Theorem 2.11 Let X be a positive persistent chain. Then, regardless of the initial distribution, if $f: S \to \mathbf{R}$ satisfies $\mathbf{E}^{\pi} |f(X_1)| < \infty$, where π is the stationary distribution, then

$$\frac{1}{n}\sum_{j=1}^{n} f(X_j) \to \mathbf{E}^{\pi} f(X_1) \tag{2.143}$$

in probability.

Remark This result holds with probability one. See Bhattacharya and Waymire (1990, section II.9) for details. □

Proof We divide up the time axis using the random times $T_k(l)$ of successive returns to state k. Write

$$Z_l = \sum_{T_k(l)+1}^{T_k(l+1)} f(X_j) \tag{2.144}$$

where $T_k(0) \equiv 0$. Then $Z_0, Z_1, \ldots$ are independent by the strong Markov property, and $Z_1, Z_2, \ldots$ are also identically distributed. Decompose

$$\sum_{1}^{n} f(X_j) = Z_0 + \sum_{j=1}^{N_k(n)} Z_i - \sum_{j=n+1}^{T_k(N_k^{(n)}+1)} f(X_j) \equiv Z_0 + S_{N_k(n)} - R_n. \tag{2.145}$$

We deal with each of these terms separately. First note that, since the chain is positive persistent, Z_0 is a sum of a finite number of random variables (with probability one). Hence $Z_n/n \to 0$ with probability one, and so in probability.

By persistence we have $\mathbf{P}(N_k^{(n)} \to \infty) = 1$, so using the law of large numbers we deduce, provided that $\mathbf{E}|Z_1| < \infty$, that $S_{N_k(n)}/N_k(n) \to \mathbf{E}Z_1$ in probability. Also, $N_k(n)/n \to \pi_k$ with probability one according to Theorem 2.10. Hence $S_{N_k(n)}/n \to \pi_k \mathbf{E}Z_1$ in probability.

Next note that

$$|R_n| \le \sum_{j=n+1}^{T_k(N_k^{(n)}+1)} |f(X_j)| \le \sum_{j=T_k(N_k^{(n)})+1}^{T_k(N_k^{(n)}+1)} |f(X_j)| \equiv \xi_n. \tag{2.146}$$

By the strong Markov property, again, $\xi_1,\xi_2,\ldots$ are iid, and by Markov's inequality $\xi_n/n \to 0$ in probability provided $\mathbf{E}\xi_1<\infty$. Hence

$$\mathbf{P}(|R_n| > n\varepsilon) \le \mathbf{P}(\xi_n > n\varepsilon) \to 0. \tag{2.147}$$

Clearly, if $\mathbf{E}\xi_1<\infty$ then $\mathbf{E}|Z_1|<\infty$. To see that the former holds, note that if ν_i as before is the expected number of visits to i between successive visits to k we have

$$\mathbf{E}\sum_{j=T_k(1)+1}^{T_k(2)} |f(X_j)| = \sum_{i\in S} |F(i)|\,\nu_i = \sum_{i\in S} |f(i)| \frac{\pi_i}{\pi_k}, \tag{2.148}$$

using the corollary to Theorem 2.7. Finally, compute

$$\mathbf{E}Z_1 = \sum_{i\in S} f(i)\nu_i = \frac{1}{\pi_k}\sum_{i\in S} f(i)\pi_i, \tag{2.149}$$

so that

$$\frac{S_{N_k(n)}}{n} \to \sum_{i\in S} f(i)\pi_i = \mathbf{E}^{\pi} f(X_1) \tag{2.150}$$

in probability. □

In the next section we shall find an important use of this result. There are central limit type results for ergodic chains as well. We write $\xi_n \sim \mathrm{AsN}(\mu_n,\sigma_n^2)$ if $(\xi_n-\mu_n)/\sigma_n$ converges in distribution to the standard normal distribution. Define $T_k(0)\equiv 0$ and $U_k(m)=T_k(m)-T_k(m-1)$.

Theorem 2.12 Let k be an ergodic state. Suppose that $\sigma_k^2=\sum_1^\infty (n-\mu_k)^2 f_{kk}^{(n)}$ satisfies $0<\sigma_k^2<\infty$, and that the distribution of $U_k(m)$ is non-degenerate. Then

$$N_k(n) \sim \mathrm{AsN}\left[\frac{n}{\mu_k}, \frac{n\sigma_k^2}{\mu_k^3}\right]. \tag{2.151}$$

Proof Assume that we start from k. Since the $U_k(l)$ are iid we have by the central limit theorem that $\sum_{l=1}^m U_k(l) \sim \mathrm{AsN}(m\mu_k, m\sigma_k^2)$. Write

$$\{N_k(n)<m\} = \{T_k(m)>n\} \tag{2.152}$$

and choose $m=[n/\mu_k + x(n\sigma_k^2/\mu_k^3)^{1/2}]$, where $[y]$ stands for the integer part of y, and x is an arbitrary real number. Now note that $T_k(m)=\sum_{l=1}^m U_k(l)$, so

$$\mathbf{P}\left\{\frac{N_k(n)-n/\mu_k}{(n\sigma_k^2/\mu_k^3)^{1/2}}<x\right\}=\mathbf{P}\Big(\sum_{l=1}^{m}U_k(l)>n\Big)$$
$$=\mathbf{P}\left[\sum_{l=1}^{m}U_k(l)-m\mu_k(m\sigma_k^2)^{-1/2}>n-m\mu_k(m\sigma_k^2)^{-1/2}\right] \qquad (2.153)$$
$$=\mathbf{P}\left[\sum_{l=1}^{m}U_k(m)-m\mu_k(m\sigma_k^2)^{-1/2}>\frac{-x}{1+o(n^{-1/4})}\right]\xrightarrow{d}\Phi(x),$$

where $\Phi(x)$ is the standard normal cdf. The case when we start from another state only changes the distribution of $U_k(1)$, which is asymptotically negligible. □

So far we have concentrated on aperiodic chains. The periodic case can be dealt with by looking at an imbedded aperiodic chain. Here is a version of Theorem 2.9 for periodic chains.

Theorem 2.13 Let X be an irreducible persistent Markov chain of period d. Then

$$\lim_{n\to\infty} p_{kk}^{(nd)}=\frac{d}{\mu_k} \qquad (2.154)$$

and writing $r_{jk}=\min\{r:p_{jk}^{(r)}>0\}$ we also have

$$\lim_{n\to\infty} p_{jk}^{(r_{jk}+nd)}=\frac{df_{jk}}{\mu_k}. \qquad (2.155)$$

Proof Let $Y_k=X_{dk}$. Then Y is ergodic with transition matrix $\mathbb{P}_Y=\mathbb{P}^d$. Hence

$$P_{Y;jk}(s)=\sum_n p_{Y;jk}^{(n)}s^n=\sum_n p_{jk}^{(nd)}s^n=P_{jk}(s^{1/d}) \qquad (2.156)$$

since $p_{jk}^{(l)}=0$ for $l\neq nd$. Rewriting equation (2.56) we have

$$F_{Y;kk}(s)=\frac{P_{Y;kk}(s)-1}{P_{Y;kk}(s)}=\frac{P_{kk}(s^{1/d})-1}{P_{kk}(s^{1/d})}=F_{kk}(s^{1/d}), \qquad (2.157)$$

so by Theorem 2.9

$$p_{Y;kk}^{(n)}\to\frac{1}{\frac{d}{ds}F_{kk}(s^{1/d})\big|_{s=1}}. \qquad (2.158)$$

The left-hand side is $p_{kk}^{(nd)}$, while the right-hand side is $(F_{kk}'(1-)/d)^{-1}=d/\mu_k$. The second part follows just as did the first corollary of Theorem 2.9. □

Example (Limiting behavior of a particular chain) Let

$$\mathbb{P} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.25 & 0.75 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.159}$$

Here $S=\{0,1,2,3\}$ with $S_T=\{2\}$ and $S_P=\{0,1\}+\{3\}$. Starting from state 2 where do we go? Let $u=\mathbf{P}^2$(absorption in $\{0,1\}$). By partitioning the sum into the possible values of the first step we get

$$\begin{aligned} u &= \sum_{k=0}^{3} \mathbf{P}^2(X_1=k, \text{ absorption in } \{0,1\}) \\ &= \sum_{k=0}^{3} \mathbf{P}^2(Z_1=k)\mathbf{P}^k(\text{absorption in } \{0,1\}) \\ &= (0.25+0.25)\times 1 + 0.25u + 0.25\times 0 = 0.5 + 0.25u \end{aligned} \tag{2.160}$$

whence $u=2/3$. The stationary distribution for the subclass $\{0,1\}$ is $(1/3,2/3)$. Therefore

$$\lim_{n\to\infty} p_{20}^{(n)} = \frac{2}{3}\times\frac{1}{3} = \frac{2}{9}. \tag{2.161}$$

Similarly $\lim_{n\to\infty} p_{21}^{(n)}=4/9$. In summary

$$\lim_{n\to\infty} \mathbb{P}^n = \begin{bmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 2/3 & 0 & 0 \\ 2/9 & 4/9 & 0 & 1/3 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.162}$$

The technique used in this argument, namely conditioning on the first step, often proves very useful. □

2.6. Markov chain Monte Carlo methods

An interesting recent application of the asymptotic theory of Markov chains is to Monte Carlo calculation of complicated integrals. There is a variety of problems that reduce to needing to compute such an integral.

Example (Likelihood) Let $L_{\mathbf{x}}(\theta)$ be a likelihood function based on an observation $\mathbf{x}$ of a random vector $\mathbf{X}$. We make no particular assumptions of the structure of $\mathbf{X}$: it could be a sequence of iid random variables, or a realization of a stochastic process. Frequently we can write

$$L_{\mathbf{x}}(\theta) = h(\mathbf{x};\theta)/c(\theta) \tag{2.163}$$

where h is known, but the normalizing constant $c(\theta)=\int h(\mathbf{x};\theta)d\mathbf{x}$ is too complicated to compute explicitly.

Example (Mixture distribution)

Suppose that we have iid observations from a mixture of exponential distributions with density

$$f(x;\boldsymbol{\theta}) = \sum_{j=1}^{k} p_j \lambda_j e^{-x\lambda_j}. \tag{2.164}$$

Here k is assumed known, so the unknown parameter is $\boldsymbol{\theta}=(p_1,\ldots,p_k,\lambda_1,\ldots,\lambda_k)$. One can, of course, write out the likelihood as the product of terms of the form (2.164), but the maximization problem can be unpleasant due to difficulties in the numerical evaluation of some of the terms. We can put it in the form needed for Markov chain Monte Carlo (or MCMC for short) by letting Λ be a random variable, taking on the value λ_i with probability p_i. Then

$$f(x;\boldsymbol{\theta}) = \mathbf{E}\Lambda e^{-\Lambda x}. \tag{2.165}$$

Considering a iid sequence Λ_i, the likelihood can be written

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathbf{E}\Lambda_i e^{-\Lambda_i x_i} = \mathbf{E}\prod_{i=1}^{n} \Lambda_i e^{-\Lambda_i x_i}. \tag{2.166}$$

□

Example (Posterior distribution) Suppose that $\boldsymbol{\theta}$, instead of being an unknown constant, is a random variable with a distribution $\pi(\boldsymbol{\theta})$, often called the **prior** distribution. If we have data $\mathbf{x}$ that conditionally upon $\boldsymbol{\theta}$ are drawn from a joint distribution $f(\mathbf{x}\mid\boldsymbol{\theta})$, we can use Bayes' theorem to compute the conditional distribution

$$\pi(\boldsymbol{\theta}\mid\mathbf{x}) = \frac{f(\mathbf{x}\mid\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{x}\mid\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{2.167}$$

called the **posterior** distribution, since it is the distribution of $\boldsymbol{\theta}$ after $\mathbf{x}$ was observed. The integral in the denominator is often difficult to compute, as is the ratio of integrals (called **posterior expectation**)

$$\int \boldsymbol{\theta}\pi(\boldsymbol{\theta}\mid\mathbf{x}) = \frac{\int \boldsymbol{\theta} f(\mathbf{x}\mid\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int f(\mathbf{x}\mid\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.168}$$

□

Example (Monte Carlo testing) Let H_0 be a simple hypothesis about the distribution of a multidimensional random variable X. Suppose that we have a continuous test statistic $T=T(X)$, and we reject H_0 for large observed values t of T. Let f be the density of T, and assume that we can simulate a random sample $t_2, \ldots, t_n$ from f. We base the observed significance level, or P-value, of t on its rank among the n values $t, t_2, \ldots, t_n$. If the rank of t is k, we reject H_0 at the k/n-level, since the rank is uniformly distributed on $1, \ldots, n$ when H_0 is true (Bickel and Doksum, 1977, p. 347). In fact, all that is needed for this to hold is that the T_i have a joint distribution which is invariant under permutations of the indices. Such distributions are called **exchangeable**, and arise, e.g., when the random variables are conditionally independent, given another random variable.

We can extend this procedure to the case of a composite null hypothesis, provided that the problem admits a sufficient statistic. We then merely simulate from the conditional distribution, given the observed values of the sufficient statistics. Of course, this simulation problem can be quite hard. □

Example (The Rasch model of item analysis) Consider r individuals responding to c test items each. Let $X_{ij}=1$(individual i answered item j correctly). Rasch (1960) suggested the model

$$\mathbf{P}(X_{ij}=1) = \exp(\alpha_i+\beta_j)/(1+\exp(\alpha_i+\beta_j)) \tag{2.169}$$

where $\sum_i \alpha_i = \sum_j \beta_j = 0$. The likelihood can be written

$$\prod_{i,j} \frac{\exp(x_{ij}(\alpha_i+\beta_j))}{(1+\exp(\alpha_i+\beta_j))} = \frac{\prod_i \exp(x_{i\bullet}\alpha_i) \prod_j \exp(x_{\bullet j}\beta_j)}{\prod_{i,j}(1+\exp(\alpha_i+\beta_j))}, \tag{2.170}$$

and we see that the totals $x_{i\bullet}=\sum_j x_{ij}$ and $x_{\bullet j}=\sum_i x_{ij}$ are sufficient statistics (cf. Appendix A). Hence, given these totals, all possible binary tables have the same probability. The problem is to device an enumeration scheme for all these tables. It is a very hard combinatorial problem. □

As it happens, it is often possible to construct a Markov chain with limiting distribution proportional to a given function $f(\mathbf{u})$. One can then estimate $\int f(\mathbf{u})d\mathbf{u}$ by running a Monte Carlo simulation of the Markov chain long enough to reach equilibrium. Exactly how long that is depends on the problem at hand.

Example (Likelihood, continued) Let $f(\mathbf{x})=g(\mathbf{x})/c$ be a fixed density, chosen so that $h(\mathbf{x};\theta)>0$ implies that $f(\mathbf{x})>0$. The mle of θ maximizes

$$L_{\mathbf{x}}(\theta) = \frac{h(\mathbf{x};\theta)/g(\mathbf{x})}{c(\theta)/c}. \tag{2.171}$$

For any θ we can evaluate $h(\mathbf{x};\theta)/g(\mathbf{x})$, but not $c(\theta)/c$. Note that

$$\frac{c(\theta)}{c} = \int_S \frac{h(\mathbf{x};\theta)}{c} d\mathbf{x} = \int_S \frac{h(\mathbf{x};\theta)}{g(\mathbf{x})} f(\mathbf{x}) d\mathbf{x} = \mathbf{E}_f\left[h(\mathbf{X};\theta)/g(\mathbf{X})\right]. \tag{2.172}$$

If we can generate samples from f we can estimate the expectation on the right-hand side of (2.172). The classical Monte Carlo method is to draw N observations $\mathbf{x}_i$ iid from f and then compute $(1/n)\sum h(\mathbf{x}_i;\boldsymbol{\theta})/g(\mathbf{x}_i)$. But f is a multivariate distribution, and it may not be easy to generate random samples from this distribution. □

The MCMC method, instead of generating iid observations, generates dependent samples from a Markov chain with stationary distribution $\mathbf{f}=(f(\mathbf{x});\mathbf{x}\in S)$ and uses Theorem 2.11 to obtain the convergence. How can this be done? One approach, the **Gibbs sampler**, was introduced into the statistical literature by Geman and Geman (1984), although it originates in statistical physics where it is called the **heat bath** method. The Gibbs sampler computes successive values of the vector $\mathbf{x}$. At stage t we have a current vector $\mathbf{x}(t)$. At the next stage we update each component of $\mathbf{x}$ in turn. Suppose we have updated $x_1,\ldots,x_{i-1}$ with new values $x_1(t+1),\ldots,x_{i-1}(t+1)$. The new value at component i, $x_i(t+1)$, is drawn at random from $f_i(\bullet \mid \mathbf{x}_1^{i-1}(t+1),\mathbf{x}_{i+1}^m(t))$ (recall the notation from section 1.2). At each stage, each component is updated just once. Variants of the Gibbs sampler have the order of updating change from stage to stage, e.g., by going through the components in the order of a random permutation, chosen anew at each iteration.

Proposition 2.7 If $f(\mathbf{x})$ satisfies the positivity condition (1.14), the Gibbs sampler is an ergodic Markov chain with stationary distribution $\mathbf{f}=(f(\mathbf{x});\mathbf{x}\in S)$.

Proof It is clear from the construction that the conditional distribution of $\mathbf{x}(t+1)$ given the past only depends on $\mathbf{x}(t)$, so the process is Markovian. The transition matrix has elements

$$p_{\mathbf{x},\mathbf{y}} = f_1(y_1 \mid \mathbf{x}_2^m)f_2(y_2 \mid y_1,\mathbf{x}_3^m)f_3(y_3,\mathbf{y}_1^2,\mathbf{x}_4^m)\cdots f_m(y_m \mid \mathbf{y}_1^{m-1}). \quad (2.173)$$

The positivity assumption guarantees that $p_{\mathbf{x},\mathbf{y}}>0$ for all $\mathbf{x},\mathbf{y}\in S=\{\mathbf{x}:f(\mathbf{x})>0\}$. Now note that $\mathbb{P}=\mathbb{P}_1\mathbb{P}_2\cdots\mathbf{P}_m$ where $\mathbb{P}_i$ has $(\mathbf{x},\mathbf{y})$-element

$$p_{i;\mathbf{x},\mathbf{y}}=f_i(y_i \mid \mathbf{x}_{-i})1(\mathbf{y}_{-i}=\mathbf{x}_{-i}). \quad (2.174)$$

To see this, it is perhaps easiest to do the case $m=2$, from which the general argument follows by a similar argument. Write

$$\begin{aligned}\left[\mathbb{P}_1\mathbb{P}_2\right]_{\mathbf{x},\mathbf{y}} &= \sum_{\mathbf{z}} p_{1;\mathbf{x},\mathbf{z}}p_{2;\mathbf{z},\mathbf{y}} \\ &= \sum_{\mathbf{z}} f_1(z_1 \mid \mathbf{x}_{-1})1(\mathbf{z}_{-1}=\mathbf{x}_{-1}) \\ &\quad \times f_2(y_2 \mid \mathbf{z}_{-2})1(\mathbf{y}_{-2}=\mathbf{z}_{-2}).\end{aligned} \quad (2.175)$$

The only $\mathbf{z}$'s for which the summands do not vanish have $z_1=y_1$, $z_2=x_2$ and

$\mathbf{z}_3^m = \mathbf{x}_3^m = \mathbf{y}_3^m$. Hence the sum is

$$\left[\mathbb{P}_1 \mathbb{P}_2\right]_{\mathbf{x},\mathbf{y}} = f_1(y_1 \mid \mathbf{x}_{-1}) f_2(y_2 \mid y_1, x_3^m) \tag{2.176}$$

as was to be shown. Hence

$$\frac{p_{i;\mathbf{x},\mathbf{y}}}{p_{i;\mathbf{y},\mathbf{x}}} = \frac{f_i(y_i \mid \mathbf{x}_{-i})}{f_i(x_i \mid \mathbf{y}_{-i})} = \frac{f(\mathbf{y})}{f(\mathbf{x})}. \tag{2.177}$$

Recall from section 2.4 that this means that $\mathbb{P}_i$ is a reversible Markov chain with stationary distribution $\mathbf{f}$. Therefore

$$\mathbf{f}\,\mathbb{P} = \mathbf{f}\,\mathbb{P}_1 \mathbb{P}_2 \cdots \mathbf{P}_m = \mathbf{f}\,\mathbb{P}_2 \cdots \mathbf{P}_m = \mathbf{f}\,\mathbb{P}_m = \mathbf{f}, \tag{2.178}$$

verifying that the chain has stationary distribution $\mathbf{f}$. By positivity it is irreducible, so the result follows from Theorem 2.7. □

Example (Mixture distribution, continued) The Gibbs sampler draws, given $\boldsymbol{\theta}$, vectors $\boldsymbol{\Lambda}=(\Lambda_1, \ldots, \Lambda_n)$. Since in this very simple case the Λ_i are iid, the Gibbs sampler just repeatedly generates iid $\boldsymbol{\Lambda}^{(i)}$, $i=1, \ldots, N$, and then estimates the likelihood by averaging

$$\hat{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{j=1}^{N} \prod_{i=1}^{n} \Lambda_i^{(j)} e^{-\Lambda_i^{(j)} x_i}. \tag{2.179}$$

Rather large values of N may be needed to evaluate the likelihood precisely enough. Of course, in this simple case one can compute the likelihood exactly. Standard optimization routines can be used to find the mle of $\boldsymbol{\theta}$. □

Ideally, in order to obtain observations from the stationary distribution $\boldsymbol{\pi}$ of the Gibbs sampler, we should choose a starting value from $\boldsymbol{\pi}$. But if we knew how to do this there would be no need to run the Gibbs sampler! As outlined in Exercise **14**, the convergence to the stationary distribution is exponentially fast, so we first run the Gibbs sampler for a **burn-in** period in order to get close enough to the stationary distribution. Only after the burn-in period do we actually start to collect observations. The proper length of the burn-in period is a subject of current research.

Example (Monte Carlo testing, continued) Since the Gibbs sampler maintains detailed balance it is reversible. The reverse chain must have the same stationary distribution as the forward chain. We use this to create exchangeable paths. Starting from the observed value $X_0=x$, we run the Gibbs sampler backwards n steps, yielding $X_{-n}=y$, say. The we simulate $N-1$ paths n steps forward in time, all starting from y, yielding observations $X_0^{(i+1)}=x^{(i+1)}$, $i=1, \ldots, N-1$. Since y (at least very nearly) is an observation from $\boldsymbol{\pi}$, the same is true for $x^{(2)}, \ldots, x^{(N)}$. Given y, $X_0, X_0^{(2)}, \ldots, X_0^{(N)}$ are independent, so they form an exchangeable sequence, and the earlier discussion of Monte Carlo testing from exchangeable sequences applies. □

Example (The Rasch model of item analysis, continued) Let S be the set of $r\times c$-tables having marginals $x_{i\cdot}$ and $x_{\cdot j}$. We need to construct a Markov chain having the uniform distribution on S as its stationary distribution. Let $\mathbf{y}=(y_{ij})$ be a configuration in S. Consider any sub-rectangle of $\mathbf{y}$ having ones in two diagonally opposite corners and zeros in the other opposite corners. Exchanging the zeros with ones, and the ones with zeros, does not change the margins, so it yields another $r\times c$-table $\mathbf{z}$ in S. We call this procedure a **switch**, and the sub-rectangle **switchable**. Figure 2.3 shows this concept.

1	0	0	1	0	2
0	1	1	1	0	3
0	0	1	0	0	1
1	0	0	0	1	2
1	1	0	0	0	2
3	2	2	2	1	10

Figure 2.3. Two switchable sub-rectangles in a 5×5 table.

Any table $\mathbf{z}\in S$ can be reached from $\mathbf{y}$ by a series of switches. To produce our Markov chain, pick a non-empty rectangle at random, and switch it if it is switchable. Clearly this preserves the margins. To see that it has a uniform stationary distribution, we just need to check that the transition matrix is symmetric. But that is easy to see: if we can go from $\mathbf{y}$ to $\mathbf{z}$ in one step, we must do that by switching a single rectangle. Thus $P_{\mathbf{y},\mathbf{z}}=P_{\mathbf{z},\mathbf{y}}$, since the opposite switch of the same rectangle brings us back.

In order to apply this procedure to the (very small) table in Figure 2.3, we need to define an appropriate test statistic. This should reflect the type of alternative model we have in mind. Here one may consider the idea that there is an interaction between difficulty and ability. We can rearrange the table so that individuals are ordered by increasing total score (row sum), and questions are ordered by decreasing success rate (column sum). One possible such reordering (ties make it non-unique) is given in Figure 2.4. A measure of interaction could be the χ^2 statistic for independence in the 2×2 table given by summing the two lowest and the two highest scores in the two most solved and the two least solved questions. This statistic is $(N_{11}-c_1r_1/N)^2/(c_1r_1/N)$, which for this table is $(1-2\times3/5)^2/(2\times3/5)=.033$. Simulating this using the Monte Carlo testing method outlined above, by first moving 2,000 steps backwards, and then 99 times move 2,000 steps forward, yields 16 that were larger than and 7 that were

0	0	1	0	0	1
1	1	0	0	0	2
1	0	0	0	1	2
1	0	0	1	0	2
0	1	1	1	0	3
3	2	2	2	1	10

Figure 2.4. Reordering of Figure 2.3.

equal to 0.333, for a P-value between 0.16 and 0.23. We find no evidence against the Rasch model based on this test statistic. □

The first Markov chain Monte Carlo method was developed by Metropolis et al. (1953). The algorithm, called the **Metropolis algorithm**, employs an auxiliary symmetric transition matrix $(q_{x,y})$ (having $q_{x,y}=q_{y,x}$). As before, we want to find a Markov chain with stationary distribution f. The next value of the Markov chain, when the present value is x, is generated by the following update method:

1. Simulate y from the distribution $q_{x,\cdot}$.
2. Calculate the odds ratio $r=f(y)/f(x)$.
3. If $r\geq 1$ the next value is y.
4. If $r<1$ go to y with probability r, and stay at x with probability $1-r$.

It should be clear that the next state only depends on the previous state, so that this is, indeed, a Markov chain. As for the Gibbs sampler, the simplest way to see that it has stationary distribution f is to note that it satisfies detailed balance. Consider a finite state space $\{1,\ldots,K\}$, and order the values so that $f(i)\leq f(j)$ for $i<j$. Then we have $p_{ij}=q_{ij}$, while $p_{ji}=q_{ji}f(i)/f(j)=p_{ij}f(i)/f(j)$, using the symmetry of the auxiliary transition matrix. A generalization of both the Metropolis algorithm and the Gibbs sampler is due to Hastings (1971), and outlined in Exercise **10**. We have demonstrated the following result.

Proposition 2.8 If $f(\mathbf{x})$ satisfies the positivity condition (1.14), the Metropolis algorithm generates an ergodic Markov chain with stationary distribution $\mathbf{f}=(f(\mathbf{x});\mathbf{x}\in S)$.

2.7. Likelihood theory for Markov chains

Given a set of observations from a two-state Markov chain, we saw in section 2.1 how it is possible to estimate the transition matrix, and thus any function thereof, using the method of maximum likelihood. In this section we study the general finite-state Markov chain, and discuss the likelihood theory for both estimation and testing. We will first look at the **nonparametric** case, where the parameter of interest is a point in the space of all transition matrices. Let $N_{ij}(n)=\sum_{l=1}^{n} 1(X_{l-1}=i, X_l=j)$ count the number of i,j-transitions. If $N_{ij}(n)=n_{ij}$, the likelihood (2.9) takes the form

$$L(\mathbf{p}_0,\mathbb{P}) = p_0(x_0)\prod_{l=1}^{n} p_{x_{l-1},x_l}$$

$$= p_0(x_0)\prod_{i\in S}\prod_{j\in S} p_{ij}^{n_{ij}} = p_0(x_0)\prod_{i\in S} L_i(\mathbb{P}) \tag{2.180}$$

where $L_i(\mathbb{P}) = \prod_{j\in S} p_{ij}^{n_{ij}}$ depends only on the elements in the ith row $\mathbb{P}_{i\bullet}$ of $\mathbb{P}$. In other words, we are estimating $|S|$ independent probability distributions. Let $l(\mathbf{p}_0,\mathbb{P})=\log L(\mathbf{p}_0,\mathbb{P})$. Then (2.180) corresponds, with obvious notation, to

$$l(\mathbf{p}_0,\mathbb{P}) = l_0(\mathbf{p}_0) + \sum_{i\in S} l_i(\mathbb{P}_{i\bullet}). \tag{2.181}$$

We want to maximize l subject to the constraints that $\mathbf{p}_0\mathbf{1}^T=1$, where $\mathbf{1}$ is a vector of ones, and that $\mathbb{P}_{i\bullet}\mathbf{1}^T=1$. Each of these maximizations can be done separately using Lagrange multipliers by differentiating a term of the form

$$l_i(\mathbb{P}_{i\bullet}) + \lambda(\mathbb{P}_{i\bullet}\mathbf{1}^T - 1) = \sum_{j\in S} n_{ij}\log p_{ij} + \lambda\Big(\sum_{j\in S} p_{ij} - 1\Big). \tag{2.182}$$

Setting the derivatives equal to zero and writing $n_i = \sum_{j\in S} n_{ij}$ we get

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i} \text{ when } n_i>0 \text{ and } \hat{p}_0(i) = 1(i=x_0). \tag{2.183}$$

We can think of this as multinomial likelihoods with random sample sizes. The estimates are very reasonable: $\hat{p}_{ij}$ is just the observed proportion of i,j-transitions among all transitions out of i. If $n_i=0$ there are no exits from state i. The likelihood is then flat as a function of p_{ij} for any j in S, and we conventionally set $\hat{p}_{ij}=0$, $i\neq j$.

Let $\hat{S}=\{i\in S: n_i\geq 1\}$ be the observed part of the state space. Obviously, $\hat{S}$ is always finite. We will, for simplicity, ignore the possibility that $T_i=n$, i.e., that state i is reached for the first time at time n, since we then cannot estimate any transitions out of state i (this problem can usually be solved by taking one more observation). Notice that $(\hat{p}_{ij}, i,j\in\hat{S})$ is a stochastic matrix over $\hat{S}$. The class structure of $\hat{S}$ is determined by $\hat{\mathbb{P}}$.

Proposition 2.9 The Markov chain on $\hat{S}$, governed by $\hat{\mathbb{P}}$, has a class of transient states, and precisely one closed class $\hat{S}_P$ of persistent states.

Proof Using Theorem 2.5, we need to show that $\hat{S}_P$ is closed. First note that $x_m \to x_{m'}$ whenever $m<m'$. Choose m_0 so that $x_{m_0}\in\hat{S}_P$ but $x_m\notin\hat{S}_P$ for $m<m_0$. Then $\{x_{m_0},\ldots,x_n\}$ is closed. □

Remark In particular, $\hat{\mathbb{P}}$ is irreducible on $\hat{S}$ if $\hat{S}_P=\hat{S}$. On the other extreme, $\hat{S}_P$ could be empty, making the estimated chain transient. This would happen if no state entered was ever returned to. □

With only one observation of the initial distribution one cannot learn much about it. There are two possible approaches. One is to condition on $X_0=x_0$, and study the conditional likelihood

$$L^c(\mathbb{P}) = \prod_{i\in S} L_i(\mathbb{P}). \tag{2.184}$$

The conditional mle's are the same as the unconditional ones. The other possibility, appropriate if the chain has been running for a long time, is to use the stationary initial distribution. This is equivalent to maximizing $L(\mathbf{p}_0,\mathbb{P})$ subject to the additional constraint that $\mathbf{p}_0=\mathbf{p}_0\mathbb{P}$. A drawback is that the nice factorization of the likelihood into terms that only depend on rows of $\mathbb{P}$ no longer obtains.

Application (Snoqualmie Falls precipitation, continued) For a two-state chain $\pi=(1-p_{11},p_{11})/(1-(p_{11}-p_{01}))$. If $X_0=0$ we have

$$L(\pi,\mathbb{P}) = \frac{(1-p_{01})^{n_{00}}p_{01}^{n_{01}}(1-p_{11})^{n_{10}+1}p_{11}^{n_{11}}}{1-(p_{11}-p_{01})}. \tag{2.185}$$

Taking logarithms (note that we have parametrized the model so that the rows sum to one), we obtain the likelihood equations

$$\begin{aligned} -\frac{n_{10}+1}{(1-p_{11})}+\frac{n_{11}}{p_{11}} &= -\frac{1}{1-(p_{11}-p_{01})} \\ -\frac{n_{00}}{(1-p_{01})}+\frac{n_{01}}{p_{01}} &= \frac{1}{1-(p_{11}-p_{01})} \end{aligned} \tag{2.186}$$

which are mixed polynomial equations of second order. Clearly, as the n_{ij} increase, the effect of the initial distribution diminishes.

For the Snoqualmie Falls data there were 11 dry and 25 rainy January 1. Hence the likelihood becomes

$$L(\pi,\mathbb{P}) = (1-p_{01})^{186}p_{01}^{123+25}(1-p_{11})^{128+11}p_{11}^{643}(1-(p_{11}-p_{01}))^{-36}$$

which is maximized by $\hat{p}_{01}=0.397$ and $\hat{p}_{11}=0.834$, virtually the same estimates as for the conditional method, namely $\hat{p}_{01}=0.398$ and $\hat{p}_{11}=0.834$. □

In terms of long term behavior of the mle's, we cannot hope to estimate $\mathbb{P}$ well if it does not correspond to an irreducible chain, since we need a large number of (i,j)-transitions for all i and j. If there is more than one persistent class we only get to see one of the classes. Therefore we assume that we are dealing with an ergodic chain. It is convenient to introduce the **step chain** $(Y_n,n\geq 0)$ defined by $Y_n=(X_n,X_{n+1})$. If we know Y_n, we know where X_n is going next.

Lemma 2.5 (a) (Y_n) is a Markov chain with state space $\tilde{S}=\{(i,j)\in S^2: p_{ij}>0\}$, initial distribution $\tilde{p}_0$ given by $\tilde{p}_0(i,j)=p_0(i)p_{ij}$ and transition matrix $\mathbb{P}=(\tilde{p}_{ij,kl})$ given by

$$\tilde{p}_{kl,ij} = 1(i=l)p_{ij}. \tag{2.187}$$

(b) If (X_n) is ergodic and $\mathbf{p}_0=\boldsymbol{\pi}$, then (Y_n) is also ergodic with stationary initial distribution $\tilde{\boldsymbol{\pi}}$ given by $\tilde{\pi}(i,j)=\pi(i)p_{ij}$.

Proof

$$\begin{aligned}
&\mathbf{P}(Y_n=(i,j)\mid Y_{n-1}=(k_1,l_1),\ldots,Y_0=(k_n,l_n))\\
&\quad= \mathbf{P}(X_{n+1}=j,X_n=i \mid X_n=l_1,X_{n-1}=k_1,\ldots,X_0=k_n)\\
&\quad= \mathbf{P}(X_{n+1}=j.X_n=i \mid X_n=l_1,X_{n-1}=k_1) \qquad (2.188)\\
&\quad= \mathbf{P}(Y_n=(i,j)\mid Y_{n-1}=(k_1,l_1))
\end{aligned}$$

verifying the Markov property. Furthermore, (2.188) can be evaluated as

$$\begin{aligned}
\mathbf{P}(Y_n=(i,j)\mid Y_{n-1}=(k,l)) &= \mathbf{P}(X_{n+1}=j \mid X_n=i,X_n=l,X_{n-1}=k)\\
&\quad\times \mathbf{P}(X_n=i \mid X_n=l,X_{n-1}=k) \qquad (2.189)\\
&= \begin{cases}\mathbf{P}(X_{n+1}=j\mid X_n=i) & \text{if } i=l\\ 0 & \text{otherwise.}\end{cases}
\end{aligned}$$

The remaining parts are proved using similar computations. □

We now use the step chain and the ergodic theorem to show strong consistency of the estimator $\hat{\mathbb{P}}$. Let $\hat{p}_{ij}(n)$ be the mle of p_{ij} based on observing the chain up to time n.

Theorem 2.14 If (X_n) is an ergodic chain, then $\hat{p}_{ij}(n)\to p_{ij}$ with probability 1 as $n\to\infty$ for all $i,j\in S$, regardless of the initial distribution.

Proof If $p_{ij}=0$ then $\mathbf{P}(\hat{p}_{ij}(n)=0)=1$, so we only need to consider $(i,j)\in\tilde{S}$. But

$$N_{ij}(n) = \sum_{k=1}^{n} 1(Y_k=(i,j)) \tag{2.190}$$

and since the step chain is ergodic we have from Theorem 2.10 and Lemma 2.5 (a) that

$$\frac{1}{n}N_{ij}(n) \to \pi_i p_{ij} \text{ with probability 1.} \tag{2.191}$$

Using Theorem 2.10 again we see that

$$\frac{1}{n}N_i(n) \to \pi_i \text{ with probability 1,} \tag{2.192}$$

whence the result follows. □

The mle's are asymptotically normally distributed. To show this, we first need a technical result. Let $W_i^{(m)}=X_{1+T_i(m)}$ be the state entered directly after the mth return to i, and write $Q_{ij}(n)=\sum_{m=1}^{[n\pi_i]}1(W_i^{(m)}=j)$. Finally let $Q_i(n)=(Q_{ij}(n),j\in S)$.

Lemma 2.6 The vectors $Q_i(n)$ for $i\in S$ are independent having multinomial distributions with sample size $[n\pi_i]$ and success probabilities $\mathbb{P}_{i\cdot}$.

Proof We need to show that the $W_i^{(m)}$ are independent with $\mathbf{P}(W_i^{(m)}=j)=p_{ij}$. But this follows from the strong Markov property, since given that $X_{T_i(m)}=i$ the future and the past are independent. □

We are now able to establish the asymptotic normality of the mle.

Theorem 2.15 Let X_n be an ergodic process. Then, regardless of the initial distribution,

$$\left[N_i(n)^{½}(\hat{p}_{ij}(n)-p_{ij}),\ i,j\in S\right] \xrightarrow{d} \mathrm{N}(\mathbf{0},\Sigma) \tag{2.193}$$

where

$$\Sigma_{ij,kl} = \begin{cases} p_{ij}(1-p_{ij}) & (i,j)=(k,l) \\ -p_{ij}p_{il} & i=k, j\neq l \\ 0 & \text{otherwise} \end{cases} \tag{2.194}$$

Remark The asymptotic covariance has multinomial structure within rows and independence between rows. Note, however, that we have to use a **random norming**, which is quite different from asymptotics for iid sequences. □

Proof Since $n\pi_i/N_i(n)\xrightarrow{\text{a.s.}}1$ we need only show that

$$\left[\frac{N_{ij}(n)-N_i(n)p_{ij}}{(n\pi_i)^{½}},\ i,j\in S\right] \xrightarrow{d} \mathrm{N}(\mathbf{0},\Sigma). \tag{2.195}$$

The basic idea is that $N_{ij}(n)$ is about the same as $Q_{ij}(n)$, and $N_i(n)$ is about $[n\pi_i]$. From the results in Appendix A and the lemma we know that

$$\left[\frac{Q_{ij}(n)-[n\pi_i]p_{ij}}{[n\pi_i]^{½}},\ i,j\in S\right] \xrightarrow{d} \mathrm{N}(\mathbf{0},\Sigma). \tag{2.196}$$

ity 1, (2.192)

□

istributed. To show this, we first need a
the state entered directly after the mth
$^{n)}=j$). Finally let $Q_i(n)=(Q_{ij}(n), j\in S)$.

$\in S$ are independent having multinomial
success probabilities $\mathbb{P}_{i\cdot}$.

$^{n)}$ are independent with $\mathbf{P}(W_i^{(m)}=j)=p_{ij}$.
property, since given that $X_{T_i(m)}=i$ the
□

otic normality of the mle.

process. Then, regardless of the initial

$\xrightarrow{d} \mathrm{N}(0,\Sigma)$ (2.193)

$l)$

(2.194)

se

has multinomial structure within rows
however, that we have to use a **random**
symptotics for iid sequences. □

nly show that

$\xrightarrow{d} \mathrm{N}(0,\Sigma)$. (2.195)

the same as $Q_{ij}(n)$, and $N_i(n)$ is about
d the lemma we know that

$\xrightarrow{d} \mathrm{N}(0,\Sigma)$. (2.196)

Hence we just need to show that these approximations are adequate, in the sense that

$$D_n = (n\pi_i)^{-1/2}(N_{ij}(n)-N_i(n)p_{ij}-Q_{ij}(n)+[n\pi_i]p_{ij}) \xrightarrow{\mathbf{P}} 0. \tag{2.197}$$

For fixed i,j let $Z_n=1(W_i^{(m)}=j)-p_{ij}$ and $S_n=\sum_1^n Z_i$. The Z_i are iid with mean zero, variance σ^2, and fourth moment κ. We can write D_n from (2.197) in terms of S_n as

$$D_n=(n\pi_i)^{-1/2}(S_{N_i(n)} - S_{[n\pi_i]}). \tag{2.198}$$

Then

$$\mathbf{P}(\,|D_n|>\varepsilon) \le \mathbf{P}(\,|D_n>\varepsilon, |N_i(n)-n\pi_i| \le xn^{1/2}) + \mathbf{P}(\,|N_i(n)-n\pi_i|>xn^{1/2}) \tag{2.199}$$

where x is a number to be chosen below. The first term of the right-hand side of (2.199) can be written as a sum over the possible values M of $N_i(n)$ satisfying the inequality $|m-n\pi_i| \le xn^{1/2}$. Using Chebyshev's inequality twice yields an upper bound of

$$\sum_{m\in M} \frac{1}{n^2\pi_i^2\varepsilon^4}\mathbf{E}(S_m-S_{[n\pi_i]})^4. \tag{2.200}$$

Since $S_m-S_{[n\pi_i]}$ is a sum of $|m-[n\pi_i]+1|$ of the Z_i we have, using $\mathbf{E}(\sum_1^n Z_i)^4 = n\kappa + 3n(n-1)\sigma^4$, that

$$\mathbf{E}(S_m-S_{[n\pi_i]})^4 \le (xn^{1/2}+1)\kappa+3(xn^{1/2}+1)^2\sigma^4. \tag{2.201}$$

The sum in (2.200) has at most $2xn^{1/2}+1$ terms, so we get

$$\mathbf{P}(\,|D_n|>\varepsilon) \le \frac{2xn^{1/2}+1}{n^2\pi_i^2\varepsilon^4}((xn^{1/2}+1)\kappa+3(xn^{1/2}+1)^2\sigma^4) + \mathbf{P}(\,|N_i(n)-n\pi_i|>xn^{1/2}). \tag{2.202}$$

The first term on the right-hand side goes to zero, while the second can be made arbitrarily small by making x large and using Theorem 2.10. □

Application (Snoqualmie Falls precipitation, continued) Using the result of Theorem 2.15, we see that $\hat{p}_{01}$ and $\hat{p}_{11}$ are asymptotically independent. Furthermore, $\hat{p}_{11}$ is approximately normally distributed with mean p_{11} and variance $p_{11}((1-p_{11})/n\pi_1$. We estimate that variance using $\hat{p}_{11}=n_{11}/n_1$ and $\hat{\pi}_1=n_1/n$ where $n_{11}=\sum_{i=1}^{36} n_{11}^{(i)}$, etc. Since $n_{11}=643$, $n_1=771$ and $n=1080$, an asymptotic 95% confidence band for p_{11} is (0.808,0.860), while one for p_{01} is (0.343,0.453) using $n_{01}=123$ and $n_0=309$. These are individual confidence bands, and the asymptotic joint coverage probability of the rectangle formed by these intervals is, using the asymptotic independence, $0.95^2 = 0.903$. To find an asymptotic 95% joint confidence set we can use individual 97.5% intervals, which yield the rectangle $(0.775,0.893)\times(0.272,0.524)$. □

Sometimes it is natural to look at a smaller model than the full nonparametric model. One may have some relatively simple model in mind, which is parametrized in some fashion.

Application (Russian linguistics) One of Markov's own examples of a Markov chain is in his 1924 probability text. The reconstruction here is using the description by Maistrov (1974). Markov studied a piece of text from Puškin's "Eugen Onegin", and classified 20,000 consecutive characters as vowels or consonants. The data are given below:

Table 2.3 Eugen Onegin characters

	Vowel next	Consonant next	Total
Vowel	1106	7532	8638
Consonant	7533	3829	11362
Total	8639	11361	20000

It is quite clear that the choice of vowel or consonant following a given letter is not independent of the letter. A very simple linguistic model is to assume a constant probability p of switching from one type of character to another. The transition matrix for this hypothesis is

$$\mathbb{P} = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}, \tag{2.203}$$

i.e., a one-dimensional subset of the space of stochastic matrices. □

For simplicity we will look only at the case of a finite state space. Assume that the transition probabilities $p_{ij}=p_{ij}(\boldsymbol{\theta})$ depend on an unknown parameter $\boldsymbol{\theta}$, taking values in Θ, an open subset of R^r. We will need some regularity conditions:

Conditions A: (i) $D=\{i,j : p_{ij}(\boldsymbol{\theta})>0\}$ is independent of $\boldsymbol{\theta}$.

(ii) Each $p_{ij}(\boldsymbol{\theta})\in C^3$, i.e., each $p_{ij}(\boldsymbol{\theta})$ is three times continuously differentiable.

(iii) The $d\times r$-matrix $\partial p_{ij}(\boldsymbol{\theta})/\partial\theta_k$, $i,j\in D$, $k=1,\ldots,r$, where d is the cardinality of D, has rank r.

(iv) For each $\boldsymbol{\theta}$ there is only one ergodic class and no transient states.