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 \( X_{ij} \sim \text{Bin}(1, p) \), 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

\[
L(p) = p^{791} (1-p)^{325}.
\]  

(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 $X_j=1$ if (year $j$) is 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 $X_j$ is iid. with $X_i$ 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

$$L(p) = p^{X_j}(1-p)^{1-X_j}. \tag{2.1}$$

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

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 360(0.33)(1-0.33) = 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:

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

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

Precipitation at Snoqualmie Falls
Table 2.1 Observed precipitation

<table>
<thead>
<tr>
<th></th>
<th>Today dry</th>
<th>Today wet</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yesterday dry</td>
<td>186 (91)</td>
<td>123 (223)</td>
<td>309</td>
</tr>
<tr>
<td>Yesterday wet</td>
<td>128 (223)</td>
<td>643 (543)</td>
<td>771</td>
</tr>
<tr>
<td>Total</td>
<td>314</td>
<td>766</td>
<td>1080</td>
</tr>
</tbody>
</table>

If the $X_t$ are not independent, we must specify the conditional probabilities

$$P(X_{t+1} = l \mid X_0 = k_0, \ldots, X_t = k_t)$$

for all $i$, $l$, and $k_1, \ldots, k_t$. 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

$$P(X_{n+1} = l \mid X_0 = k_0, \ldots, X_n = k_n)$$

$$= P(X_{n+1} = l \mid X_n = k_n) = p_{k,l}(n).$$

(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 $P(n) = (p_{ij}(n))$ 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 Andreyevich (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.
Precipitation at Snoqualmie Falls

Proposition 2.1 If $P$ 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 $p_{ij}(n) = P(X_{n+1} = j | X_n = i)$, so

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

$$
= P(\bigcup_{j=0}^{K} \{X_{n+1} = j\} | X_n = i) = P(X_{n+1} \in S | X_n = i) = 1
$$

(2.6)

since the process must go somewhere. \qed

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 $P = P(1)$. 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), $p_{00} = 1-p_d$, $p_{01} = p_p$, $p_{10} = 1-p_w$, and $p_{11} = p_d$. In matrix notation,

$$
P = \begin{pmatrix}
p_{00} & p_{01} \\
p_{10} & p_{11}
\end{pmatrix} = \begin{pmatrix}
1-p_d & p_p \\
p_w & 1-p_w
\end{pmatrix}.
$$

(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 $P$ is stochastic, then $P^k$ 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 $P^T = 1^T$, where 1 is a row vector of ones (recall that all vectors are assumed to be row vectors). Hence $(I - P)1^T = 0$, where $I$ is the identity matrix, so 1 is a right eigenvector of $P$ corresponding to the eigenvalue 1. Now (ii) follows easily, writing

$$
P^k 1^T = P^{k-1} P^T 1^T = P^{k-1} 1^T = \ldots = 1^T.
$$

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

\[
L(\mathbf{p}) = \prod_{i=0}^{n-1} P(X_{i+1} = x_{i+1} \mid X_i = x_i) \\
= P(X_0 = x_0) \prod_{i=0}^{n-1} P_{X_{i+1} = x_{i+1}} = P(X_0 = x_0) \prod_{k,l=0}^{k} P_{kl}^{x_l}.
\]

(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

\[
L(\mathbf{p}) = L(p_{01}, p_{11}) = \left[ \prod_{i=1}^{36} P(X_i = x_i) \right] p_{01}^{13} p_{11}^{123} p_{10}^{30} p_{00}. 
\]

(2.10)

Assuming that the starting values \( X_0 \) 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

\[
\hat{p}_{01} = \frac{123}{309} = 0.398 \quad \hat{p}_{11} = \frac{643}{771} = 0.834
\]

(2.11)

so

\[
\hat{p} = \begin{bmatrix} 0.398 & 0.602 \\ 0.166 & 0.834 \end{bmatrix}
\]

(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 \( p_{01} \) 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 \( 2(\log L(\mathbf{p}) - \log L(\hat{p})) \) has a \( \chi^2 \) 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
The marginal distribution

\[ 2 (\log L(\hat{p}) - \log L(\hat{p})) = 2 (643 \log 0.834 + 128 \log 0.166) \]
\[ + 123 \log 0.398 + 186 \log 0.602 \]
\[ - 771 \log 0.714 - 309 \log 0.286 = 184.5 \]

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

\[ P(X_{n+1} = 1) = P(X_{n+1} = 1, X_n = 0) + P(X_{n+1} = 1, X_n = 1) \]
\[ = P(X_n = 0) p_{01} + P(X_n = 1) p_{11} \]
\[ = P(X_n = 1) (p_{11} - p_{01}) + p_{01}. \]

Define the initial distribution \( p_0 = (p_0(0), \ldots, p_0(K)) \) where \( p_0(i) = P(X_0 = i) \).

In the 0-1 case we write \( p_0(1) = p_1 \). Then (2.14) can be written

\[ P(X_1 = 1) = p_1 (p_{11} - p_{01}) + p_{01}, \]
\[ P(X_2 = 1) = (p_{11} - p_{01}) P(X_1 = 1) + p_{01} \]
\[ = p_1 (p_{11} - p_{01})^2 + p_{01} (1 + (p_{11} - p_{01})) \]
\[ \ldots \]
\[ P(X_n = 1) = (p_{11} - p_{01})^n p_1 + p_{01} \sum_{j=0}^{n-1} (p_{11} - p_{01})^j. \]

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

\[ P(X_n = 1) = \frac{p_{01}}{1 - (p_{11} - p_{01})} + \frac{p_1 - \frac{p_{01}}{1 - (p_{11} - p_{01})}}{1 - (p_{11} - p_{01})} (p_{11} - p_{01})^n. \]

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 \( P(X_n = 1) \) 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 \( \mathbb{Z} \), and define \( p_0^n = P(X_n = k | X_0 = j) \). Here is an important computation, called the Chapman–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**

\[ p_{ij}^{(n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n-m)}, \quad 1 \leq m \leq n-1. \]  \hfill (2.17)

**Proof** Since the process must be somewhere in \( S \) at time \( m \), we have

\[
p_{ij}^{(n)} = \mathbb{P}(\{X_n = k\} \cap \bigcup_{l \in S} \{X_m = l\} \mid X_0 = j)
= \sum_{l \in S} \mathbb{P}(X_n = k, X_m = l \mid X_0 = j)
= \sum_{l \in S} \mathbb{P}(X_n = k \mid X_m = l, X_0 = j) \mathbb{P}(X_m = l \mid X_0 = j)
= \sum_{l \in S} \mathbb{P}(X_n = k \mid X_m = l) \mathbb{P}(X_m = l \mid X_0 = j).
\]  \hfill (2.18)

In matrix notation we rewrite (2.17) as

\[
\mathbb{P}_n = (p_{ij}^{(n)}) = \mathbb{P}_{n-1} \mathbb{P}.
\]  \hfill (2.19)

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

\[
p_n = p_{n-1} \mathbb{P}
\]  \hfill (2.20)

(recall the computation of \( \mathbb{P}(X_n = l) \) earlier) we see that

\[
p_n = p_0 \mathbb{P}^n.
\]  \hfill (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

\[
\hat{\mathbb{P}} = \begin{bmatrix} 0.602 & 0.398 \\ 0.166 & 0.834 \end{bmatrix}
\]  \hfill (2.22)

so that

\[
\hat{\mathbb{P}}^5 = \begin{bmatrix} 0.305 & 0.695 \\ 0.290 & 0.710 \end{bmatrix}.
\]  \hfill (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
Classification of states

$p_0=(0,1)$ in (2.21), so that

$$\hat{p}_5=(0,1)\hat{p}^5=(0.29,0.71).$$ (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 $p_{ij}=0$ it is certain not to stay. If $p_{ij}>0$, the time spent in the state has a geometric distribution with mean $1/(1-p_{ij})$ (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

$$T_A = \begin{cases} \min\{n > 0 : X_n \in A\} & \text{if } X_n \text{ ever hits } A \\ \infty & \text{otherwise} \end{cases}$$ (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 $P_x$. More generally, we write the distribution of the chain starting from the initial distribution $p_0$ as $P^{p_0}$, and compute it using the formula

$$P^{p_0}(A) = \sum_{i \in S} p_0(i)P_x(A).$$ (2.26)

This amounts to first choosing the initial state $i$ at random from $p_0$, and then running the chain starting from state $i$.

Proposition 2.3

$$p_m^{(k)} = \sum_{m=1}^{k} p^m(T_k=m)p^{(k-m)}. \tag{2.27}$$

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

$$p_m^{(k)} = p^m(X_n=k) = \sum_{m=1}^{n} p^m(T_k=m, X_n=k)$$

$$= \sum_{m=1}^{n} p^m(T_k=m)p^{(k)}(X_n=k \mid T_k=m)$$

$$= \sum_{m=1}^{n} p^m(T_k=m)p^{(k)}(X_n=k \mid T_k=m) = \sum_{m=1}^{n} p^m(T_k=m)p^{(k-m)}. \tag{2.27}$$
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 $p_{kk}^{(n)} = P(T_k \leq n)$.

**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 $p_{kk}^{(n)} = 1$ 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 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 $f_j(k) = P^k(X_1 = i)$. Then

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

A proof of this can be found, e.g., in Freedman (1983, Theorem 1.21).

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

**Theorem 2.1**  $\leftrightarrow$ is an equivalence relation.

**Proof**  $i \leftrightarrow i$ since $p_{ii}^{(0)} = P(X_0 = i \mid X_0 = i) = 1$. 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

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

and $i \rightarrow k$. To show that $k \rightarrow i$ 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 $i-1$. Reed 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
Classification of states

\[
\mathbf{P} = \begin{bmatrix}
1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
q_1 & r_1 & p_1 & 0 & \cdots & 0 & 0 \\
0 & q_2 & r_2 & p_2 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & r_{k-1} & p_{k-1} \\
0 & 0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}
\]  

(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 \( q_i \) is 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 \( \lambda_0 \), 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

\[
\lambda_0 = p_1 \sum_{n=0}^{\infty} p_n^{(q_i)}.
\]  

(2.31)

For example, if \( K=3 \) so

\[
\mathbf{P} = \begin{bmatrix}
1 & 0 & 0 & 0 \\
2/3 & 0 & 1/3 & 0 \\
0 & 1/3 & 0 & 2/3 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]  

(2.32)

we find that \( p_n^{(q_i)} = (\frac{1}{3} \times \frac{1}{3})^n \) while \( p_n^{(q_i+1)} = 0 \) (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 \( p_n^{(q_i)} = 0 \) 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 \( p_n^{(q_i)} = 0 \) 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 \( p_{ij}^{(m)} > 0, p_{ij}^{(n)} > 0 \), and assume that \( p_{ij}^{(q_i)} > 0 \). Then

\[
p_{ij}^{(m+n)} \geq p_{ij}^{(q_i)} p_{ij}^{(m)} > 0
\]  

and

\[
p_{ij}^{(m+n+s)} \geq p_{ij}^{(q_i)} p_{ij}^{(s)} p_{ij}^{(m)} > 0
\]  

(2.33)

(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 \( p_{ij}^{(q_i)} > 0 \). 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

\[
P = \begin{bmatrix}
1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
q_1 & r_1 & p_1 & 0 & \cdots & 0 & 0 \\
0 & q_2 & r_2 & p_2 & \cdots & 0 & 0 \\
& \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & 0 & \cdots & r_{K-1} & p_{K-1} \\
0 & 0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}
\] (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 \(n\) is divisible by \(n_1\) or \(n_2\). Any integer \(n\) can be written \(n = l n_1 + k n_2\) 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 \(0 \leq s \leq n_2-1, b < c, \) and \(0 < a < n_2\). Hence

\[
n - sn_1 - (n - tn_1) = (s - t)n_1 = (c - b)n_2.
\] (2.36)

Since gcd\((n_1, n_2) = 1\) we must have \(s - t\) containing all prime factors of \(n_2\). But then \(s - t \geq n_2\) 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 \(n \geq N\).

**Proof** Since \(d(j) = 1\) there are integers \(n_1, n_2\) with gcd 1 such that

\[
p_{ij}(n_k) > 0, \quad k = 1, 2.
\] From Lemma 2.2 we see that any sufficiently large \(n\) can be written \(ln_1 + kn_2\), whence

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

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

\[
p_{ij}^{(n + n_0)} \geq p_{ij}^{(n_0)} p_{ij}^{(n)} > 0.
\] (2.38) □
Discrete time Markov chains

(Continued) In the described earlier we have

\[
\begin{align*}
\cdots & \quad 0 \quad 0 \\
\cdots & \quad 0 \quad 0 \\
\cdots & \quad 0 \quad 0 \\
\cdots & \quad r_{K-1} \quad p_{K-1} \\
\cdots & \quad 0 \quad 1 \\
\end{align*}
\]

(2.35)

Then \( r_{K-1} p_{K-1} = 2 \).

□

Lemmas:

1. If \( p_{ij} \) and \( p_{ij}^{(1)} \) denote the passage probabilities from state \( i \) to state \( j \) of the Markov chain and \( p_{ij}^{(n)} \), respectively, then for \( n \geq 1 \),

\[
p_{ij}^{(n)} = p_{ij}^{(1)} p_{ji}^{(n-1)}.
\]

2. Let \( n_1 \) and \( n_2 \) with greatest common divisor \( n \mid n_1 \) and \( n \mid n_2 \) can be written \( n = in_1 + kn_2 \) for non-negative integers \( i \) and \( k \).

3. Let \( n_1 \) and \( n_2 \) with greatest common divisor \( n \mid n_1 \). Either these residue classes are all distinct positive integers \( n_1 \) or two residue classes are 0, so the corresponding residues are \( n = kn_1 + ln_2 \) or at least two residue classes are 0.

4. Let \( n_1 \) and \( n_2 \) with gcd \( n \mid n_1 \). Either these residue classes are all distinct positive integers \( n_1 \) or two residue classes are 0, so the preceding argument can be informally stated as \( n = (c-b)n_1 \).

□

Classification of states

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

\[
\begin{align*}
p_{ij}^{(n)} &= P(X_{t+1} = (k, l) \mid Z_t = (i, j)) = P(X_{t+1} = k, Y_{t+1} = l \mid X_t = i, Y_t = j) \\
&= \frac{P(X_{t+1} = k, X_t = i) P(Y_{t+1} = l, Y_t = j)}{P(X_t = i) P(Y_t = j)} = p_{ik} p_{jl}.
\end{align*}
\]

(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 \( p_{ij}^{(n)}>0 \) and \( Z \) is therefore irreducible.

□

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

\[
f_{ij}^{(n)} = P(X_n = j, X_k = j, k = 1, \ldots, n-1 \mid X_0 = i).
\]

(2.40)

Define \( f_{ij} = \sum_{n=0}^{\infty} f_{ij}^{(n)} = P(T_j < \infty) \). 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 \( \sum_{n=2}^{\infty} p_{ii}^{(n)} = 0 \).

Proof If \( i \) is transient, let \( M \) be the number of returns to \( i \). Then \( \mathbb{E}M = \sum_{n=1}^{\infty} P(M = n) = f_{ii} \), so \( \mathbb{E}M = \sum_{n=2}^{\infty} P(M > n) = f_{ii} / (1 - f_{ii}) \), where \( \mathbb{E} \) is expectation with respect to \( \mathbb{P} \).

Since \( f_{ii} < 1 \), \( \mathbb{E}M < \infty \). But

\[
\mathbb{E}M = \sum_{n=1}^{\infty} \mathbb{E}[1 \mid X_n = i] = \sum_{n=1}^{\infty} p_{ii}^{(n)}.
\]

(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 \( \mathbb{P}(M = \infty) = 1 \), i.e., \( \mathbb{E}M = \infty \).

□

Remark The proof of Theorem 2.3 shows that \( M - 1 \) has a geometric distribution with parameter \( p_{ii} \).

Example (A simple random walk) A simple random walk is a birth and death chain on the integers with \( p_i = p \), \( r_i = 0 \) and \( q_i = q \), 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

\[
p^{(2n)}_0 \approx \left(\frac{2n}{n}\right) p^n q^n - \left(\frac{4pq}{n}\right)^n \tag{2.42}
\]

using Stirling’s formula \( n! \approx n^{n+1/2} e^{-n} (2\pi)^{1/2} \). Hence \( \sum p^{(2n)}_0 = \infty \) iff \( p = q = \frac{1}{2} \). In other words, 0 is a persistent state iff the coin is fair. \( \square \)

**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

\[
p^{(2n)}_0 \approx \frac{1}{6} \sum_{j+k=0} \frac{(2n)!}{j!k!(n-j-k)!} \frac{1}{3^n} \sum_{j+k=0} \frac{1}{j!k!(n-j-k)!} \frac{n!}{n}\tag{2.43}
\]

The sum is one, being the sum of all probabilities in a trinomial distribution with probability \( \frac{1}{3} \) 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 \( p^{(2n)}_0 \) is, to within an order of \( n^2 \),

\[
2^{-2n} \times \frac{2n!}{\sqrt{2\pi n}} \times 3 \times 3^{3n/3} \frac{3^{n+1}}{2nn} = \frac{3}{\pi n}^{3/2} , \tag{2.44}
\]

so \( \sum p^{(2n)}_0 \approx \infty \), 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. \( \square \)
Discrete time Markov chains

The persistence or transience of state 0.

\[ p \in \{0, 1\}^n. \]  

Hence \( \sum p = 1 \) if and only if \( p = \frac{1}{2} \). In other words, the coin is fair.

Random walk in higher dimensions by the consideration of a three-dimensional integer lattice the process has nearest neighbors, regardless of which probability is the same to go to each of its neighbors. The transition of any state is persistent. However, if \( k > 2 \) and probability 1/6 of going up, down, east, or west, then the transition of any state is transient.

\[ f^{(n)}(i) = \sum_{j \in \mathbb{Z}^n} P(X_n = j | X_{n-1} = i) \]  

(2.42)

Corollary: For a transient state \( i, \) \( p^{(n)}_{ii} \to 0. \)

Proof: Immediate since \( \sum p^{(n)}_i < \infty. \)

For a persistent state \( f^{(n)}(i) \) is a probability distribution with mean \( \mu_i = \sum_n n f^{(n)}(i), \) 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 \( n \geq 2 \)

\[ f^{(n)}(i) = P(X_n = i | X_{n-1} = i, \ldots, X_{n-2} = 0) \]

(2.43)

Also, \( f^{(1)}(i) = p_{ii} \), so the mean recurrence time is \( \mu_i = \sum k f^{(k)}(i) = 1 \cdot (1 - p_{11}) P_{01} \), which we estimate to be 1.42, using \( p_{01} = 0.398 \) and \( p_{11} = 0.834. \) Given a wet day, the mean number of dry days to follow is \( \mu_i - 1, \) 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 \( (1-p_{11})(1-p_{01}) \). 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 \( \mu_i - 1, \) 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 \( i \to j \) and assume that \( j \) is persistent. Then there are integers \( n \) and \( m \) so that \( p^{(n)}_{ij} > 0 \) and \( p^{(m)}_{ji} > 0. \) For any \( s \geq 0 \)

\[ p^{(n+s)}_{ii} \geq p^{(n)}_{ij} p^{(s)}_{ji} p^{(m)}_{ii} \]

(2.46)

so

\[ \sum_s p^{(n+s)}_{ii} \geq p^{(n)}_{ij} p^{(s)}_{ji} \sum_s p^{(m)}_{ii} = \infty, \]

(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 $(a_k)$. It is easy to show that if $\sum a_k < \infty$ then $A(1-) = \lim_{s \uparrow 1} A(s) = \sum a_k$, 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 $P(s) = \sum_k p_k s^k$ converges for all $|s| \leq 1$. $P$ is called a probability generating function (pgf). If $X$ has pgf $P$, define the $k$th factorial moment $m_{(k)} = E(X-1) \cdots (X-k+1)$. By differentiating under the summation sign in the definition of $P$ we see that

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

(2.48)

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

$$p_k = \frac{d^k}{ds^k} P(s) \bigg|_{s=0} = \sum_{i=k}^{\infty} \frac{(-1)^{i-k} m_{(i)}}{k! (i-k)!}$$

(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

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

(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, $P(S_n = 0) = P(0)^n = p_0^n$, and

$$E S_n = \frac{dP(s)^n}{ds} \bigg|_{s=1} = n P'(1) P(1)^{n-1} = n E X$$

(2.51)

since $P(1) = 1$.

Recall from Proposition 2.4 that

$$P_{(n)}(k) = \sum_{k=0}^{n} f(k) p_{(n-k)}$$

(2.52)

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

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

(2.53)

and

$$F_{(n)}(s) = \sum_{n=0}^{\infty} f_{(n)} s^n$$

such that

$$F_{(n)}(s) = (e^s - 1)^n$$

and

$$P_{(n)}(s) = \frac{1}{n!} (\log(1+s))^n$$

showing that

$$e^s = \sum_{n=0}^{\infty} \frac{s^n}{n!}$$

and

$$\log(1+s) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1} s^n}{n}$$

are convergent for $|s| < 1$.
Discrete time Markov chains

numbers. If \( A(x) = \sum_{k=0}^{\infty} a_k x^k \) converges in the generating function of \((a_k)\). It is easy to see that \( A(x) = \lim_{t \to 1} A(s) = \sum_k a_k \), and for \( a_k \geq 0 \), if

**Generating functions** If \((p_k : k \geq 0)\) is a probability sequence, \( P(s) = \sum_{k=0}^{\infty} p_k s^k \) converges for all \( |s| \leq 1 \). Differentiating under the sum, we get

\[
\frac{d}{ds} P(s) = \sum_{k=1}^{\infty} kp_k s^{k-1} = \int_0^s \frac{d}{ds} P(s) \, ds.
\]

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

\[
F_{ij}(s) = \sum_{n=0}^{\infty} f_{ij}(n) s^n.
\]  

(2.54)

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

\[
F_{ij}(s) P_{ii}(s) = P_{ii}(s) - 1
\]  

(2.55)

since \( p_i(0) = 1 \). Thus

\[
P_{ii}(s) = \frac{1}{1 - F_{ii}(s)}.
\]  

(2.56)

Likewise

\[
P_{ij}(s) = F_{ij}(s) P_{jj}(s).
\]  

(2.57)

It is worth noting that

\[
m_i = \frac{F_{ii}(1)}{\mu_i}.
\]  

(2.58)

**Remark** We can use this to give an alternative proof of the result in Theorem 2.3 that \( i \) is persistent iff \( \sum p_{ii} = \infty \). Assume first that \( \sum p_{ii} = 1 \). Then \( \mu_i = 1 \) so \( P_{ii}(1) = \infty \), \( \sum p_{ii} = \infty \). Conversely, if \( \sum p_{ii} < 1 \) we have that \( \mu_i < 1 \), \( P_{ii}(1) < \infty \), and \( \sum p_{ii} = \infty \). 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 \( F_{00}(s) = 1 - P_{00}(s) \) we see that \( F_{00}(1) = \frac{1}{1 - P_{00}(1)} \).

Now notice that

\[
A_N = \sum_{n=0}^{N} np_{00}(s) = O(N^{3/2}),
\]  

(2.59)

\[
B_N = \sum_{n=0}^{N} p_{00}(s) = O(N^{1/2}),
\]  

(2.60)

and \( A_N \to F_{00}(1) \), \( B_N \to P_{00}(1) \), so that

\[
F_{00}(1) = \lim_{N \to \infty} \frac{A_N}{B_N} = \lim_{N \to \infty} O(N^{3/2}) = \infty,
\]  

(2.61)

showing that the mean recurrence time is infinite.
Lemma 2.3  Suppose that \( j \) is persistent. Then it is positive persistent iff 
\[
\pi_j = \lim_{t \to \infty} (1-s)P_{jj}(s) > 0, \text{ and then } \pi_j = 1/\mu_j. 
\]

Proof  From (2.56) we have 
\[
1 - F_{jj}(s) = P_{jj}(s)^{-1}
\]
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 \( s \to 1 \). Hence the limit of the right-hand side is the same, and the result follows. \( \square \)

Remark  This almost proves the convergence of averages of transition probabilities, namely 
\[
(1/n) \sum_{i=1}^n p_j(i) \to \mu_j^{-1} \text{ as } n \to \infty. \tag{2.64}
\]
Consider 
\[
(1-s)P_{jj}(s) = \frac{\sum_{k=0}^{\infty} s^k p_j(k)}{\sum_{k=0}^{\infty} s^k} \tag{2.65}
\]
for any \( s \leq 1 \). If we could take the limit as \( s \to 1 \) under the summations, the right-hand side would converge to \((1/n) \sum_{i=1}^n p_j(i)\), 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. \( \square \)

Call a set \( C \) of states closed if \( f_{jk}=F_{jk}(1-) = 0 \) for \( j \in C, \ k \notin C \). Then 
\[
P_{jk}(1-) = 0 = \sum_{n \geq 0} p_{jk}(n) \]
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., 
\[
p_{jk} = \sum_{s \in S} p_{js}p_{sk} = \sum_{s \in C} p_{js}p_{sk} = 0, \tag{2.66}
\]
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
Classification of states

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 \( S_T = \{ \text{transient states} \} \) and \( S_P = \{ \text{persistent states} \} \), we have that
\[
S = S_T + S_P
\]
and
\[
S_P = \sum C_i \quad \text{of disjoint, irreducible, closed sets.}
\]

**Proof** Let \( x \in S_P \), and define \( C = \{ y \in S_P : x \rightarrow y \} \). By persistence \( f_{xy} = 1 \), so \( x \in C \). We first show that \( C \) is closed. Suppose that \( y \in C, y \rightarrow z \). Since \( y \) is persistent, \( z \) must also be persistent. Since \( x \rightarrow y \rightarrow z \) 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 \( z \leftarrow y \). Since \( x \rightarrow y, y \rightarrow x \) by persistence. But \( x \rightarrow z \) by definition of \( C \), so \( y \rightarrow x \rightarrow z \). The same argument, with \( y \) and \( z \) transposed, shows that \( z \rightarrow y \).

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, \( x \rightarrow y \). Since \( D \) is closed, \( x \in D \) and \( x \rightarrow y \) we have that \( y \in D \). Thus \( C \subseteq D \). Similarly \( D \subseteq C \), so they are equal.

It follows from this theorem that if a chain starts in \( C \), it will stay there forever (and we may as well let \( S = C \)). 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 \( \mu_i \) are finite or all are infinite.

**Proof** As before we can find \( k, m \) such that \( p_{ij}^{(k)} > 0, p_{ij}^{(m)} > 0 \). Since
\[
p_{ij}^{(n+k+m)} \geq p_{ij}^{(m)} p_{ij}^{(k)}
\]
we see by averaging and anticipating the result (2.64) that
\[
\pi_i \geq p_{ij}^{(m)} p_{ij}^{(k)}
\]
so if \( \pi_i > 0 \) then \( \pi_j > 0 \), while if \( \pi_i = 0 \) then \( \pi_j = 0 \). 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 \( \sum p_{ij}^{(n)} = 1 \) 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 \( v_i \) denote the expected number of visits to \( i \) between two visits to \( j \), i.e.,

\[
v_i = \mathbb{E} \left[ \sum_{n=0}^{T_j-1} 1(X_n = i) \right] = \sum_{n=0}^{\infty} \mathbb{P}(X_n = i, T_j > n).
\]  

(2.71)

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

\[
v_i = \sum_{n=1}^{\infty} \mathbb{P}(X_n = i, T_j > n) = \sum_{n=1}^{\infty} \mathbb{P}(X_n = i, T_j > n-1)
\]

\[
= \sum_{n=1}^{\infty} \sum_{k \in S} \mathbb{P}(X_n = k, T_j > n-1, X_{n-1} = i)
\]

\[
= \sum_{n=1}^{\infty} \sum_{k \in S} \mathbb{P}(X_n = k | T_j > n-1, X_{n-1} = i) \mathbb{P}(T_j > n-1, X_{n-1} = i)
\]

(2.72)

\[
= \sum_{k \in S} p_{ik} \sum_{n=1}^{\infty} \mathbb{P}(X_n = k, T_j > n-1) = \sum_{k \in S} \sum_{m=0}^{\infty} \mathbb{P}(X_n = k, T_j > m)
\]

\[
= \sum_{k \in S} p_{ik} v_k.
\]

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

\[
v_j = 1 = \sum_{n=1}^{\infty} \mathbb{P}(T_j = n) = \sum_{n=1}^{\infty} \sum_{k \in S} \mathbb{P}(T_j = n, X_{n-1} = k)
\]

(2.73)

\[
= \sum_{n=1}^{\infty} \sum_{k \in S} \mathbb{P}(T_j > n-1, X = k)
\]

\[
= \sum_{n=1}^{\infty} \sum_{k \in S} p_{ik} \mathbb{P}(T_j > n-1, X = k) = \sum_{k \in S} p_{ik} v_k.
\]

Writing \( \mathbf{v} = (v_1, v_2, \ldots) \) we have shown that \( \mathbf{vP} = \mathbf{v} \). By iterating we see that \( \mathbf{vP}^n = \mathbf{v} \) for all \( n = 1, 2, \ldots \). In particular, for \( i \in C_j \),

\[
v_i p_{ij}^{(n)} \leq v_j = 1,
\]

(2.74)
Discrete time Markov chains

3. We have that each term in the sum, and hence at least one state is persistent.
3. In order to compute \( x_n = i, T_j > n \).

For \( n \neq j \) let \( v_i \) denote the expected

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 \( n \) large enough such that \( v_i \leq \delta \) for all \( i \) in \( C_j \). The final sum in

(2.75) therefore is a finite sum of finite elements, so \( \mu_j \leq \delta \).

\( i \leq j \) for some \( n \). Finally we compute

\[
\mu_j = \sum_{n=0}^{\infty} \sum_{i \in S} P_i^j (X_n = i, T_j > n) = \sum_{i \in S} \sum_{n=0}^{\infty} P_i^j (X_n = i, T_j > n) = \sum_{i \in S} v_i.
\]

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

This result was first proved by Erdős, 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 \( p_{ij}^{(n)} > 0 \) 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 \( p_1 = p_0_1 / (1 - (p_{11} - p_{01})) \), we define 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 \( \pi \). By letting \( p_1 = \pi \) in relation (2.20) we must have \( \pi = \pi \Pi \), or equivalently

\[
\pi = \Pi \Pi = 0.
\]

Thus \( \pi \) is a left eigenvector of \( \Pi \), 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 \( \pi_1 = p_{00} / (1 - (p_{11} - p_{01})) \). 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 \( \pi_1 \) is between the smaller and the larger of \( p_{01} \) and \( p_{11} \). To see that this choice of \( \pi \) indeed satisfies (2.76), note that

\[
\begin{pmatrix}
1 - p_{11} & p_{01} \\
1 - (p_{11} - p_{01}) & 1 - p_{01}
\end{pmatrix}
\begin{pmatrix}
p_{01} \\
-p_{01}
\end{pmatrix}
= \begin{pmatrix}
0 \\
0
\end{pmatrix}.
\]

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

\[
p_1 = \pi \Pi = \pi
\]

\( p_1 = p_0 \Pi = \pi \),

eq \pi, \quad \text{for all } n.
\]

\[
\pi = \pi \quad \text{for all } n.
\]
We then say that the one-dimensional distributions $p_n$ are time invariant (another name for this is stationary). Therefore $\pi$ is also known as the stationary initial distribution. In fact, starting from $\pi$ all finite-dimensional distributions are time invariant, in the sense that

$$(X_{k_1}, X_{k_2}, \ldots, X_{k_n}) = (X_{k_1+k}, X_{k_1+k}, \ldots, X_{k_1+k})$$

(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

$$\mathbb{E}X_nX_{n+k} = \mathbb{E}(\mathbb{E}(X_nX_{n+k} | X_n)X_n) = \sum_{j, l \in S} j l p_l^{(j)} p_n(l).$$

(2.81)

If the chain is strictly stationary the right-hand side of (2.81) simplifies to $\sum_{j, l \in S} j l p_l^{(j)} \pi_l$, so the covariance between $X_n$ and $X_{n+k}$ is

$$\text{Cov}(X_n, X_{n+k}) = \sum_{j, l \in S} j l p_l^{(j)} \pi_l - (\sum_{j \in S} j \pi_j)^2.$$  

(2.82)

Anticipating the next section we see that if the chain has a limiting distribution, so $p_l^{(j)} \to \pi_l$, 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

$$\text{Corr}(X_n, X_{n+k}) = \frac{p_l^{(j)} - \pi_l}{1 - \pi_l} = (p_{11} - p_{01})^k$$

(2.83)

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

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

(2.84)

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

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

(2.85)

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

$$\frac{p_l^{(j+1)} - \pi_1}{1 - \pi_1} = \frac{p_{01} + (1 - \pi_1)(p_{11} - p_{01})^n + \pi_1(p_{11} - p_{01}) - \pi_1}{1 - \pi_1}$$

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

(2.86)
Discrete time Markov chains

formal distributions $p_n$ are time invariant therefore $\pi$ is also known as the station-
ing from $\pi$ all finite-dimensional distributions

$$X_{k, k+h}, \ldots, k_n$$ (Exercise 5). Processes satisfying

$$X_n = \sum_{i \in S} j \pi p_j^i p_n(i).$$

right-hand side of (2.81) simplifies to

$$X_n = \sum_{i \in S} j \pi p_j^i p_n(i).$$

that if the chain has a limiting distribution, this

precipitation, continued In the 0-1

variable $X = (p_{11} - p_{00})^k$

If $k = 1$ we have

$$n_k = \begin{cases} 0, & \text{if } k=1; \\ \pi_0 (p_{11} - p_{00})^{k-1}, & \text{otherwise}. \end{cases}$$

We need to show that $\pi x = \pi$, i.e., that $\sum_{i \in S} v_i p_{ij} = v_j$ and that $\sum_{i \in S} v_i = \mu_k$ (in order for $\pi$ 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 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 $\pi_i = \mu_i^{-1}.$

where the last equality is from the definition of $\pi_1$. This completes the induction. We see that the correlation function is geometrically decreasing. For the Snoqualmie Falls data, where $\hat{\rho}_{11} - \hat{\rho}_{01}$ is 0.436, we get the estimated correlations given in Table 2.2.

<table>
<thead>
<tr>
<th>Lag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corr</td>
<td>0.436</td>
<td>0.190</td>
<td>0.083</td>
<td>0.036</td>
<td>0.016</td>
<td>0.007</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Equation (2.76) shows that if a chain has a stationary distribution, it must be the eigenvector of $H$ corresponding to the eigenvalue $1$. Sometimes we can be a bit more explicit. Recall that if $k$ is persistent, then $v_j$ 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 $\pi_i = v_i/\mu_i$ for a fixed state $k$.

Proof We need to show that $\pi x = \pi$, i.e., that $\sum_{i \in S} v_i p_{ij} = v_j$ and that $\sum_{i \in S} v_i = \mu_k$ (in order for $\pi$ 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).

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 $\pi$ 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,

$$\pi_j = \sum_{i \in S} \pi_i p_{ij} \to 0 \text{ as } n \to \infty$$

so $\pi$ 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

$$\sum_{i \in S} \pi_i = \sum_{i \in S_m} \pi_i \to \sum_{i \in S} \pi_i \text{ as } m \to \infty$$

where the last equality is from the definition of $\pi_1$. This completes the induction. We see that the correlation function is geometrically decreasing. For the Snoqualmie Falls data, where $\hat{\rho}_{11} - \hat{\rho}_{01}$ is 0.436, we get the estimated correlations given in Table 2.2.

Table 2.2 Estimated correlations for Snoqualmie Falls data

<table>
<thead>
<tr>
<th>Lag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corr</td>
<td>0.436</td>
<td>0.190</td>
<td>0.083</td>
<td>0.036</td>
<td>0.016</td>
<td>0.007</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Equation (2.76) shows that if a chain has a stationary distribution, it must be the eigenvector of $H$ corresponding to the eigenvalue $1$. Sometimes we can be a bit more explicit. Recall that if $k$ is persistent, then $v_j$ 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 $\pi_i = v_i/\mu_i$ for a fixed state $k$.

Proof We need to show that $\pi x = \pi$, i.e., that $\sum_{i \in S} v_i p_{ij} = v_j$ and that $\sum_{i \in S} v_i = \mu_k$ (in order for $\pi$ 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 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 $\pi$ 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,

$$\pi_j = \sum_{i \in S} \pi_i p_{ij} \to 0 \text{ as } n \to \infty$$

so $\pi$ 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
\[
\pi_j = \sum_{i \in S_n} \pi_i p_{ij}^{(n)} + \sum_{i \in S_n} \pi_i p_{ij}^{(0)} \leq \sum_{i \in S_n} \pi_i p_{ij}^{(n)} + \sum_{i \in S_n} \pi_j. \tag{2.88}
\]

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 \( \pi \) is summable. The key to this argument is that the \( p_{ij}^{(0)} \) are bounded. Thus the existence of a stationary distribution implies that all states are persistent.

Let the initial distribution be \( \pi \). Using Exercise 5, all finite-dimensional probabilities are time invariant. Thus

\[
P^\pi(X_n = i, T_j \geq n + 1) = P^\pi(X_n = i, X_1, \ldots, X_n \neq j)
= P^\pi(X_{n-1} = i, X_0, \ldots, X_{n-1} \neq j) \tag{2.89}
= P^\pi(X_{n-1} = i, T_j \geq n) - P^\pi(X_{n-1} = i, X_0 = j, T_j \geq n)
= P^\pi(X_{n-1} = i, T_j \geq n) - \pi_j P^\pi(X_{n-1} = i, T_j \geq n).
\]

Summing over \( n \leq N \) and rearranging terms yields

\[
\sum_{n = 1}^N \pi_j P^\pi(X_{n-1} = i, T_j \geq n) = \sum_{n = 1}^N (P^\pi(X_{n-1} = i, T_j \geq n) - P^\pi(X_{n} = i, T_j \geq n + 1))
= \pi_j - P^\pi(X_{N} = i, T_j \geq N + 1) \tag{2.90}
\]

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

\[
\sum_{n = 1}^\infty \pi_j P^\pi(X_{n-1} = i, T_j \geq n) = \pi_i. \tag{2.91}
\]

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

\[
\pi_j \sum_{n = 1}^\infty P^\pi(T_j \geq n) = 1. \tag{2.92}
\]

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

\[
0 = \pi_i = \sum_{j} \pi_j p_{ji}^{(0)} \geq \pi_j p_{ji}^{(0)}, \tag{2.93}
\]

so whenever \( j \to i \) we have \( \pi_j = 0 \). But then all the \( \pi_i \) are 0 by irreducibility, and \( \pi \) is not a distribution. Hence all the \( \pi_i \) 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.
Discrete time Markov chains

\[ \pi_j = \sum_{i \in S} \pi_i p_{ij} \]

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

\[ \nu_i = E^x \sum_{n=0}^{T_i-1} 1(X_n = i) = \mu_i / \lambda_i = \pi_i / \pi_k \]

**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

\[ \pi_j = \sum_{i \in S} \pi_i p_{ij} = \sum_{i \in S} \pi_i \pi_{ij} \]

We can interpret \( \sum \pi_i p_{ij} \) as the probability flux out of state \( j \), and \( \sum \pi_i \pi_{ij} \) 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 \( (X_k, k \in Z) \) 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:

\[ P(Y_{k+1} = j \mid Y_k = i, Y_{k-1} = i_1, \ldots, Y_{k-n} = i_n) = P(X_{-(k+1)} = j \mid X_{-k} = i, X_{-(k-1)} = i_1, \ldots, X_{-(k-n)} = i_n) \]

\[ = P(X_{-(k+1)} = j \mid X_{-k} = i, X_{-(k-1)} = i_1, \ldots, X_{-(k-n)} = i_n) \frac{P(X_{-(k+1)} = j \mid X_{-k} = i, X_{-(k-1)} = i_1, \ldots, X_{-(k-n)} = i_n)}{P(X_{-(k+1)} = j \mid X_{-k} = i, X_{-(k-1)} = i_1, \ldots, X_{-(k-n)} = i_n)} \]

\[ \times \frac{P(X_{-(k+1)} = j \mid X_{-k} = i, X_{-(k-1)} = i_1, \ldots, X_{-(k-n)} = i_n)}{P(X_{-k} = i)} \]

If \( X \) has the stationary marginal distribution \( \pi \) for all \( k \) we see that \( Y \) has stationary transition probabilities \( q_{ij} \) given by
Discrete time Markov chains

\[ q_{ij} = P(Y_{k+1} = j \mid Y_k = i) = \frac{p_{ji}}{\pi_i}. \]  
(2.96)

Note that since \( X \) is defined for all \( k \in \mathbb{Z} \), it is not enough to set \( p^{(0)} = \pi \) in order to have marginal distribution \( \pi \) 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

\[ p_{ij} = \frac{\pi_j}{\pi_i} \]  
(2.97)
or, equivalently,

\[ \pi_i p_{ij} = p_{ji} \pi_j. \]  
(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 \( \pi \) 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 \( \pi \).

**Proof** Using Theorem 2.7 we need only show that \( \pi \) is a stationary distribution. But

\[ \sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j, \]  
(2.99)

so \( \pi = \pi P \).

\[ \mathbb{Q} \]

**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

\[ p_j \pi_j = q_{j+1} \pi_{j+1} \]  
(2.100)

holds, and that the equilibrium distribution is given by

\[ \pi_j = \pi_0 \prod_{i=1}^{j-1} \frac{p_{i-1}}{q_i} \]  
(2.101)

where

\[ \pi_0^{-1} = \sum_{j=0}^{\infty} \prod_{i=1}^{j} \frac{p_{i-1}}{q_i} \]  
(2.102)
Discrete time Markov chains

\[ \pi_j \]
\[ \pi_i \]
\[ (2.96) \]

In Z, it is not enough to set \( p^{(0)} = \pi \) in order to have \( \pi \) as a stationary distribution, stating that the probability flux from \( j \) to \( i \) is not well-defined. Detailed balance is a property that \( \pi \) and quantum mechanics. It was first noted in the context of the detailed balance property for Hamiltonian systems (Koppen, 1981, section V.6). The conditions needed to find the stationary distribution of a Markov chain are:

\[ \pi_i \sum_j p_{ij} \pi_j = \pi_i \sum_i p_{ij} \pi_j \]
\[ \pi_i \]
\[ (2.98) \]

Since the Markov chain, a distribution \( \pi \) exists, only show that \( \pi \) is a stationary distribution.

\[ \pi_i = \pi_j \]
\[ (2.99) \]

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 \( \pi_j \) is

\[ \pi_j = p_{j-1} \pi_{j-1} + r_j \pi_j + q_{j+1} \pi_{j+1}, \quad j > 0, \]
\[ (2.103) \]

while for \( j = 0 \)

\[ \pi_0 = r_0 \pi_0 + q_1 \pi_1. \]
\[ (2.104) \]

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

\[ p_0 \pi_0 = q_1 \pi_1 \]
\[ (2.105) \]

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

\[ \pi_{k+1} = p_k \pi_k + r_{k+1} \pi_{k+1} + q_{k+2} \pi_{k+2}. \]
\[ (2.106) \]

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

\[ p_{k+1} \pi_{k+1} = q_{k+2} \pi_{k+2} \]
\[ (2.107) \]

whence the detailed balance equation holds. The evaluation of \( \pi \) is now immediate.

The particular case of a random walk reflected at the origin has \( p_1 = 1 - q_1 = p \), so

\[ \pi_0^{-1} = \sum_j \left[ \frac{p}{1-p} \right]^j = \frac{1-p}{2-p}, \]
\[ (2.108) \]

provided that \( p < \frac{1}{2} \). The stationary distribution is then geometric. If \( p \geq \frac{1}{2} \) 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

\[ p_{x,x'} = \begin{cases} \frac{1}{N} & \text{if } x \text{ and } x' \text{ only differ in one location;} \\ 0 & \text{otherwise.} \end{cases} \]
\[ (2.109) \]

Consider the case \( N = 2 \), so the states are 00, 01, 10, and 11 (or in decimal notation 0, 1, 2, 3). Then

\[ P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}. \]
\[ (2.110) \]
Clearly $P$ 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

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

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

$$P_{Y_{i+1}} = \frac{i}{N}, \quad P_{Y_{i-1}} = \frac{N-i}{N}. \quad (2.112)$$

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

$$\pi_Y = \pi_0 \prod_{i=1}^j \frac{N-i+1}{i} = \pi_0 \left[ \begin{array}{c} N \\ j \end{array} \right] , \quad (2.113)$$

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

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

and

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

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.
Discrete time Markov chains

Since the stationary distribution is \( \pi_i = 2^{-N} \) with period 3. It satisfies not only the weaker micro-reversibility condition

\[
\frac{N-i}{N}.
\]

(2.111)

we know from the previous example that the stationary distribution satisfies

\[
\{N\}
\]

(2.112)

We obtain the stationary distribution for \( Y \) is obtained from that of micro-states corresponding to a given

(2.113)

Ehrenfest and Ehrenfest (1906) to explain the paradox observed by Loschmidt (1876). The paradox is that, even though it can be derived from classical mechanics, the process is not micro-reversible while thermodynamics contains microstates that increase with time. This physical sense of micro-reversibility for given micro-states \( x \) and \( x' \), with initial state \( X_0 = x \) and \( X_0 = x' \) have both

\[
\pi(x') = \frac{N-x'}{N} | X_0 = x
\]

(2.114)

\[
\pi(x') = y' | Y_0 = y).
\]

(2.115)

(2.114) holds by micro-reversibility, but the right-hand side would be much larger than the actual for the process to ever reach stationary distribution. The statistical equilibrium behavior, which the classical mechanics cannot predict.

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 \( P \) for the Snoqualmie Falls precipitation model, the rows got more and more similar. Figure 2.2 illustrates this.

![Probability vs Lag (days)](image)

Figure 2.2. \( n \)-step transition probabilities for Snoqualmie Falls model. The upper curve is \( p_1^n \) while the lower is \( p_0^n \).

In fact, under suitable conditions

\[
\mathbb{I}P^n \rightarrow \begin{bmatrix}
\pi \\
\pi \\
\pi
\end{bmatrix}
\]

(2.116)

We say that the chain has a limiting distribution. This means 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

\[
p_n = p_0P_n \rightarrow \begin{bmatrix}
\pi \\
\pi \\
\pi
\end{bmatrix} = \pi
\]

regardless of \( p_0 \). 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

\[
P = \begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{bmatrix}
\]

(2.117)

Then \( \pi = (1, 1, 1)/3 \), but \( P^n \) 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 \( p_{ik}^{(l)} > 0 \) and \( p_{kl}^{(m)} > 0 \). Hence \( p_{ik}^{(l+m)} = p_{ik}^{(l)} p_{kl}^{(m)} > 0 \), so \( d \) must divide \( l + m \), i.e., \( l + m = rd \) for some integer \( r \). Fixing \( m \) we see that \( l = -m + rd \) for some integer \( r \). Thus we can find an integer \( s_k \), \( 0 \leq s_k < d \), such that \( p_{ik}^{(l)} = 0 \) unless \( n \equiv s_k \mod d \). Let \( G_k = \{ k : s_k = s \} \). Then

\[
S = G_0 + \cdots + G_{d-1}.
\]

One-step transitions are only possible from states in \( G_k \) to states in \( G_{k+1} \) (where \( G_{d} = G_0 \)), 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

\[
\mathbb{P} = \begin{bmatrix}
0 & A & 0 \\
0 & 0 & B \\
C & 0 & 0
\end{bmatrix},
\]

where \( A \) consists of transition probabilities from \( G_0 \) to \( G_1 \), etc. \( \Box \)

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

\[
\lim_{n \to \infty} p_{ik}^{(n)} = \frac{1}{\mu_k}.
\]

**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 \( p_{ij} \). Since \( X \) is positive persistent, it has a unique stationary distribution \( \pi \). Then \( Z \) has stationary distribution \( \eta \) with \( \eta_i = \pi_i \). Therefore \( Z \) is positive persistent by Theorem 2.7. Let \( Z_0 = (i,j) \), choose \( s \in S \), and let \( T_{ss} \) be the hitting time of \( (s,s) \) for \( Z \). Since \( Z \) is persistent, \( \mathbb{P}(T_{ss} < \infty) = 1 \). 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 \( \{ T_{ss} \leq n \} \), the random variables \( X_n \) and \( Y_n \) have the same conditional distribution. Compute

\[
p_{ik}^{(n)} = \mathbb{P}^{ij}(X_n = k) = \mathbb{P}^{ij}(X_n = k, T_{ss} \leq n) + \mathbb{P}^{ij}(X_n = k, T_{ss} > n)
\]
Discrete time Markov chains

The period of this chain is 3. This particular P
permits stationary distribution (Exercise 6).

We may derive many ways typical for what happens in a
irreducible chain X has period d. Then for every
state k, d divides m + n, i.e., d divides m = rd for some integer r. Fix
integer m, 0 ≤ m < d, such that d divides m unless

\[ \pi_k \neq 0 \]

from states in Gk to states in Gk+1 (where
leads back to Gk. Hence for a chain with
an irreducible set. For d = 3 we have

\[ \pi_0 \neq 0 \]

\[ \pi_1 \neq 0 \]
\[ \pi_2 \neq 0 \]

by bounded convergence.

Example (The Polya urn model) Quite a few stochastic processes were
originally thought of using colored balls in urns. A paper by Eggenberger and
Polya (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 ones 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 Polya 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

\[ N\left(\frac{R}{N}\right)^2\left(\frac{B}{N}\right)^{b-1}\]
Thus $X_n$ is not a Markov chain, unless $d = 0$. However, the number of red balls drawn, $R_n = \sum_{i=1}^{n} X_i$, is a Markov chain:

$$P(R_n = r \mid R_{n-1} = r_{n-1}) = \begin{cases} \frac{R + dr_{n-1}}{N + d(n-1)} & \text{if } r = r_{n-1} + 1 \\ \frac{B + d(n-1 - r_{n-1})}{N + d(n-1)} & \text{if } r = r_{n-1} \end{cases} \quad (2.127)$$

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

$$P(R_n = r) = \left[ \begin{array}{c} n \\ r \end{array} \right] \frac{R + d}{N + d} \cdots \frac{R + (r-1)d}{N + (r-1)d} \frac{B + d}{N + d} \cdots \frac{B + (n-r)d}{N + (n-r)d} \frac{N}{N + d} \cdots \frac{N + (n-1)d}{N + (n-1)d}. \quad (2.128)$$

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

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

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

$$\lim_{n \to \infty} \pi_n(k) = \frac{f_k}{\mu_k}. \quad (2.130)$$

**Proof** Recall that $p_n(k) = \sum_{l=0}^{n} f_{|k|} p_{k,l}^{(n-l)}$. Taking limits under the summation sign we get

$$p_{k}^{(n)} \to (1/\mu_k) \sum f_{|k|} = f_k/\mu_k. \quad (2.131)$$

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

$$p_{k}^{(n)} = \sum_{l=0}^{m-1} f_{|k|} p_{k,l}^{(n-l)} + \sum_{l=m}^{n} f_{|k|} p_{k,l}^{(n-l)}. \quad (2.132)$$

Since $p_{k}^{(n)} \leq 1$ for all $n$ we have

$$\sum_{l=0}^{n-1} f_{|k|} p_{k,l}^{(n-l)} \leq p_{k}^{(n)} \leq \sum_{l=0}^{m-1} f_{|k|} p_{k,l}^{(n-l)} + \sum_{l=m}^{n} f_{|k|}. \quad (2.133)$$

Now let $n \to \infty$ to see that
Discrete time Markov chains

Theorem 2.9 Let $d = 0$. However, the number of red balls

$$
\sum_{i=0}^{m-1} f_{ik}(i) \mu_i \leq \lim_{n \to \infty} p_{ik}^{(n)} \leq \lim_{n \to \infty} p_{ik}^{(n)} \leq \sum_{i=0}^{m-1} f_{ik}(i) \mu_i + \sum_{i=m}^{\infty} f_{ik}(i),
$$

(2.134)

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

\[ \square \]

Corollary If $k$ is a persistent aperiodic state, then

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

(2.135)

for any starting state communicating with $k$.

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

\[ \square \]

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

$$
E^k N_k(n) = \sum_{i=1}^{n} P^k(X_i = k) = \sum_{i=1}^{n} p_{ik}^{(n)}.
$$

(2.136)

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

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

(2.137)

and by Corollary 1 above $p_{jk}^{(n)} \to I/\mu_k$, 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 $I/\mu_k$ (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 p.g.f. $F_{kk}(s)$ and mean $\mu_k < \infty$. By the strong law of large numbers we have with probability one that
Discrete time Markov chains

\[
\lim_{r \to \infty} \frac{T_k(1) + (T_k(2) - T_k(1)) + \cdots + (T_k(r) - T_k(r-1))}{r} = \lim_{r \to \infty} \frac{T_k(r)}{r} = \mu_k. \tag{2.138}
\]

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

\[
T_k(N_k(n)) \leq n \leq T_k(N_k(n) + 1). \tag{2.139}
\]

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

\[
\frac{N_k(n)}{n} \leq \frac{N_k(n)}{T_k(N_k(n))} \to \frac{1}{\mu_k} \quad \text{with probability 1} \tag{2.140}
\]

and

\[
\frac{N_k(n) + 1}{n} \geq \frac{N_k(n) + 1}{T_k(N_k(n) + 1)} \to \frac{1}{\mu_k} \quad \text{with probability 1} \tag{2.141}
\]

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, 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 \( n \)-th descendant we have a Markov chain with state space \( \{AA, Aa, aa\} \), and transition matrix

\[\text{Hardy, Godfrey Harold (1877–1947). English pure mathematician. His main contributions came through his long collaboration with Littlewood on problems in number theory, inequalities, and complex analysis. He was apparently not very fond of this non-theoretical paper, which he published in an obscure American journal.}\]
The long-term behavior of the Markov chain can be described by the transition matrix $P$ defined as:

$$P = \begin{bmatrix} 0 & \frac{P}{P Q} & \frac{Q}{P} \\ \frac{P}{P+Q} & 0 & \frac{Q}{P+Q} \\ \frac{Q}{P+Q} & \frac{Q}{P+Q} & 0 \end{bmatrix}. \quad (2.142)$$

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 $(P, Q, P Q, Q^2)$. This is indeed the case (Exercise 9). \qed

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 $f: S \to \mathbb{R}$ satisfies $\mathbb{E}^\pi \left[ f(X_1) \right] < \infty$, where $\pi$ is the stationary distribution, then

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

in probability.

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

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

$$Z_k = \sum_{T_k(l)+1}^n f(X_j) \quad (2.144)$$

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

$$\sum_{j=1}^{n} f(X_j) = Z_0 + \sum_{j=1}^{n} Z_j - \sum_{j=n+1}^{\infty} f(X_j) = Z_0 + S_{N(n)} - R_n. \quad (2.145)$$

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 $Z_0/n \to 0$ with probability one, and so in probability.

By persistence we have $P(N_k^0 \to \infty) = 1$, so using the law of large numbers we deduce, provided that $\mathbb{E} \left| Z_1 \right| < \infty$, that $S_{N(n)}/n \to 0$ in probability. Also, $N_k(n)/n \to \pi_k$ with probability one according to Theorem 2.10. Hence $S_{N(n)}/n \to \pi_k EZ_1$ in probability.
Next note that
\[ |R_n| \leq \sum_{j=\tau_0(n^+)+1}^{T_0(n^+)+1} |f(X_j)| \leq \sum_{j=\tau_0(n^+)+1}^{T_0(n^+)+1} \xi_n = \xi_n. \quad (2.146) \]

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 \( E \xi_1 < \infty \). Hence
\[ P(\{ R_n > n\varepsilon \}) \leq P(\xi_n > n\varepsilon) \to 0. \quad (2.147) \]

Clearly, if \( E\xi_1 < \infty \) then \( E |Z_1| < \infty \). To see that the former holds, note that if \( v_i \) as before is the expected number of visits to \( i \) between successive visits to \( k \) we have
\[ E \sum_{j=\tau_i(k)+1}^{T_i(k)} |f(X_j)| = \sum_{i \in S} |F(i)| v_i = \sum_{i \in S} |f(i)| \frac{\tau_i}{\tau_k}, \quad (2.148) \]

using the corollary to Theorem 2.7. Finally, compute
\[ E Z_1 = \sum_{i \in S} f(i) v_i = \frac{1}{\tau_k} \sum_{i \in S} f(i) \tau_i \]
so that
\[ \frac{S_{N_k(n)}}{n} \to \sum_{i \in S} f(i) \tau_i = E f(X_1) \quad (2.150) \]
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 \( \xi_n - \text{as} N(\mu_n, \sigma_n^2) \) if \( (\xi_n - \mu_n) / \sigma_n \) converges in distribution to the standard normal distribution. Define \( T_k(0) = 0 \) and \( U_k(m) = T_k(m) - T_k(m-1) \).

**Theorem 2.12** Let \( k \) be an ergodic state. Suppose that \( \sigma_k^2 = \sum_{i=1}^m (n-mu_k)^2 f(l)^2 \) satisfies \( 0 < \sigma_k^2 < \infty \), and that the distribution of \( U_k(m) \) is non-degenerate. Then
\[ N_k(n) - \text{as} N \left[ \frac{n}{\mu_k}, \frac{n \sigma_k^2}{\mu_k^3} \right]. \quad (2.151) \]

**Proof** Assume that we start from \( k \). Since the \( U_k(l) \) are iid we have by the central limit theorem that \( \sum_{l=1}^m U_k(l) - \text{as} N(m \mu_k, m \sigma_k^2) \). Write
\[ \{ N_k(n) < m \} = \{ T_k(m) > n \} \quad (2.152) \]
and choose \( n = [n \mu_k + x(n \sigma_k^2/\mu_k)^2] \), 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
Long term behavior

\[
P \left( \frac{N_k(n) - n \mu_k}{(n \sigma_k^2)^{1/2}} < x \right) = P \left( \sum_{i=1}^{n} U_k(l) > n \right)
\]

\[
= P \left( \sum_{i=1}^{n} U_k(l) - m \mu_k (m \sigma_k^2)^{-1/2} > n - m \mu_k (m \sigma_k^2)^{-1/2} \right)
\]

\[
= P \left( \sum_{i=1}^{n} U_k(m) - m \mu_k (m \sigma_k^2)^{-1/2} > \frac{-x}{1 + o(n^{-1/2})} \right) \to \Phi(x),
\]

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. \( \square \)

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

\[
\lim_{n \to \infty} p_{kk}^{(nd)} = \frac{\mu_k}{d},
\]

and writing \( r_k = \min \{ r : p_{kl}(r > 0) \} \) we also have

\[
\lim_{n \to \infty} \frac{p_{lk}^{(r_k + nd)}}{\mu_k} = \frac{d_{lk}}{\mu_k},
\]

**Proof** Let \( Y_k = X_{dk} \). Then \( Y \) is ergodic with transition matrix \( P_Y = P^d \). Hence

\[
P_{Y,dk}(s) = \sum_n p_{Y,dk}^{(n)} s^n = \sum_n p_{dk}^{(nd)} s^n = P_{dk}(s^{1/d})
\]

since \( p_{dk}^{(n)} = 0 \) for \( n \neq d \). Rewriting equation (2.56) we have

\[
F_{Y,dk}(s) = \frac{P_{Y,dk}(s) - 1}{P_{Y,dk}(s)} = \frac{P_{dk}(s^{1/d}) - 1}{P_{dk}(s^{1/d})} = F_{dk}(s^{1/d}),
\]

so by Theorem 2.9

\[
p_{dk}^{(s)} \to \frac{1}{d} \left| \frac{d}{ds} F_{kk}(s^{1/d}) \right|_{s=1}
\]

The left-hand side is \( p_{dk}^{(s)} \), while the right-hand side is \( (F_{kk}(1-yd))^{-1} = d/\mu_k \). The second part follows just as did the first corollary of Theorem 2.9. \( \square \)
Example (Limiting behavior of a particular chain) Let

\[
P = \begin{bmatrix}
0.5 & 0.5 & 0 & 0 \\
0.25 & 0.75 & 0 & 0 \\
0.25 & 0.25 & 0.25 & 0.25 \\
0 & 0 & 0 & 0
\end{bmatrix}
\]  \quad (2.159)

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

\[
u = \sum_{k=0}^{3} P^2(X_1 = k, \text{ absorption in } \{0, 1\})
= \sum_{k=0}^{3} P^2(Z_1 = k) P^k(\text{absorption in } \{0, 1\})
= (0.25 + 0.25) \times 1 + 0.25 u + 0.25 \times 0 = 0.5 + 0.25 u
\]  \quad (2.160)

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

\[
\lim_{n \to \infty} P_0^{(n)} = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9}
\]  \quad (2.161)

Similarly \( \lim_{n \to \infty} P_2^{(n)} = 4/9 \). In summary

\[
\lim_{n \to \infty} P = \begin{bmatrix}
1/3 & 2/3 & 0 & 0 \\
1/3 & 2/3 & 0 & 0 \\
2/9 & 4/9 & 0 & 1/3 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]  \quad (2.162)

The technique used in this argument, namely conditioning on the first step, often proves very useful. \( \square \)

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

\[
L_x(\theta) = h(x; \theta) / c(\theta)
\]  \quad (2.163)

where \( h \) is known, but the normalizing constant \( c(\theta) = \int h(x; \theta) dx \) is too complicated to compute explicitly.
Markov chain Monte Carlo methods

Example (Mixture distribution) Suppose that we have iid observations from a mixture of exponential distributions with density

\[ f(x; \theta) = \sum_{j=1}^{k} p_j \lambda_j e^{-\lambda_j x}. \]  

(2.164)

Here \( k \) is assumed known, so the unknown parameter is \( \theta = (p_1, \ldots, p_k, \lambda_1, \ldots, \lambda_k) \). 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 \( \Lambda \) be a random variable, taking on the value \( \lambda_i \) with probability \( p_i \). Then

\[ f(x; \theta) = E \Lambda e^{-\Lambda x}. \]  

(2.165)

Considering a iid sequence \( \Lambda_i \), the likelihood can be written

\[ L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} E \Lambda_i e^{-\Lambda_i x_i} = E \prod_{i=1}^{n} \Lambda_i e^{-\Lambda_i x_i}. \]  

(2.166)

\[ \square \]

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

\[ \pi(\theta \mid x) = \frac{f(x \mid \theta) \pi(\theta)}{\int f(x \mid \theta) \pi(\theta) d\theta} \]  

(2.167)

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

\[ \int \theta \pi(\theta \mid x) = \frac{\int \theta f(x \mid \theta) \pi(\theta) d\theta}{\int f(x \mid \theta) \pi(\theta) d\theta}. \]  

(2.168)

\[ \square \]

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 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

\[
P(X_{ij}=1) = \frac{\exp(\alpha_i+\beta_j)}{1+\exp(\alpha_i+\beta_j)}
\]

(2.169)

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

\[
\prod_{i,j} \frac{\exp(x_{ij}(\alpha_i+\beta_j))}{(1+\exp(\alpha_i+\beta_j))} = \frac{\prod_i \exp(x_i, \alpha_i) \prod_j \exp(x_j, \beta_j)}{\prod_i (1+\exp(\alpha_i+\beta_j))}.
\]

(2.170)

and we see that the totals \( x_i=\sum_j x_{ij} \) and \( x_j=\sum_i x_{ij} \) are sufficient statistics (cf. Appendix A). Hence, given these totals, all possible binary tables have the same probability. The problem is to devise 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 \( f(u) \). One can then estimate \( f(u)/u \) 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 \( f(x)=g(x)/c \) be a fixed density, chosen so that \( h(x;\theta)>0 \) implies that \( f(x)>0 \). The mle of \( \theta \) maximizes

\[
L_x(\theta) = \frac{h(x;\theta)g(x)}{c(\theta)/c}.
\]

(2.171)

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

\[
c(\theta) = \int h(x;\theta)dx = \int \frac{h(x;\theta)}{g(x)}g(x)dx = E_f \left[ h(X;\theta)g(X) \right].
\]

(2.172)

Markov chain

If we can generate a random vector \( x \) from the distribution \( x_i \), iid from the joint distribution, and then update it, that is called the heuristics of the vector \( x \).

The MCMC method used by the authors is called the Metropolis-Hastings algorithm. To update each component, draw new values \( x_i(\theta) \) from some distribution (often a Normal distribution) and then update the component if the move is accepted. The acceptance probability is given by

\[
P_{x,y} = \min \left( 1, \frac{f(x)g(y)}{f(y)g(x)} \right).
\]

The positivity assumption is now weaker. To see this, it is enough to show that

\[
P_{x,z} = \min \left( 1, \frac{f(z)g(x)}{f(x)g(z)} \right).
\]

The only z's for which this is true are the z's for which f(z) < f(x)g(z)/g(x).
Discrete time Markov chains

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 \( x_i \) iid from \( f \) and then compute \( \frac{1}{N} \sum h(x_i; \theta) g(x_i) \). But \( f \) is a multivariate distribution, and it may not be easy to generate random samples from this distribution. \( \square \)

Markov chain Monte Carlo methods

The MCMC method, instead of generating iid observations, generates dependent samples from a Markov chain with stationary distribution \( f = (f(x); x \in S) \) 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 \( x \). At stage \( t \) we have a current vector \( x(t) \). At the next stage we update each component of \( x \) in turn. Suppose we have updated \( x_1, \ldots, x_{t-1} \) with new values \( x_1(t+1), \ldots, x_{t-1}(t+1) \). The new value at component \( i \), \( x_i(t+1) \), is drawn at random from \( f_i(x_i(t), x_{-i}(t+1), \theta) \) (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 \( f(x) \) satisfies the positivity condition (1.14), the Gibbs sampler is an ergodic Markov chain with stationary distribution \( f = (f(x); x \in S) \).

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

\[
P_{x,y} = f_i(y_1 | x_1^{(t)}, \ldots, y_{i-1}) f_i(y_{i+1}, \ldots, y_n | x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t)}, \ldots, x_n^{(t+1)})
\]

(2.173)

The positivity assumption guarantees that \( p_{x,y} > 0 \) for all \( x, y \in S = \{ x : f(x) > 0 \} \). Now note that \( P = P_1 P_2 \cdots P_n \) where \( P_i \) has \( (x, y) \)-element

\[
P_i(x, y) = f_i(y_1 | x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t)}, \ldots, x_n^{(t+1)})
\]

(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

\[
(P_1 P_2)_{x,y} = \sum_z P_{1,x,z} P_{2,z,y}
\]

(2.175)

The only \( z \)'s for which the summands do not vanish have \( z_1 = y_1, z_2 = x_2 \) and
Discrete time Markov chains

\[ z_3^{x^2} = x_3^y = y_3^x. \] Hence the sum is

\[ \left[ P_1 P_2 \right]_{x,y} = f_1(y_1 | x_1) f_2(y_2 | y_1, x_2^3) \]

(2.176)

as was to be shown. Hence

\[ P_{i;2,x} = f_i(y_i | x_i); \quad P_{i;2,x} = f(x). \]

(2.177)

Recall from section 2.4 that this means that \( P_i \) is a reversible Markov chain with stationary distribution \( f \). Therefore

\[ f \mathbb{P} = f P_1 P_2 \cdots P_m = f P_2 \cdots P_m = f \mathbb{P}_m = f, \]

(2.178)

verifying that the chain has stationary distribution \( f \). By positivity it is irreducible, so the result follows from Theorem 2.7.

**Example (Mixture distribution, continued)**

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

\[ \hat{L}(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{n} \Lambda_j^{(j)} e^{-\lambda_0 x_j}. \]

(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 \( \Theta \).

Ideally, in order to obtain observations from the stationary distribution \( \pi \) of the Gibbs sampler, we should choose a starting value from \( \pi \). 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 \( X_0^{(i+1)} = x^{(i+1)}, \quad i = 1, \ldots, N - 1 \). Since \( y \) (at least very nearly) is an observation from \( \pi \), 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 model.

**Example (The Reversibility property)**

The set of \( r \times c \)-tables that constitute a Markov chain having the table \( y = (y_{ij}) \) as configuration have two diagonally opposite \( r \times c \)-sub-tables. Exchanging the zero columns and margins, so it yields another \( r \times c \)-table, and the sub-rectangle.

**Figure 2.3.** Two tables.

Any table \( x \in S \) can be obtained by flipping a Markov chain, picking a row or column switchable. Clearly, the stationary distribution is symmetric. But that is equivalent to saying that by switching any two rows in the table, we can get to the same rectangle.

In order to apply the Gibbs sampler, we need to define an appropriate model of interaction between the locations. For each individual we order the locations by increasing order (ties make it non-unique). The \( \chi^2 \) statistic is found in the lowest and the two back solved questions. This is \( (1 - 2 \times 3/5)/2 \). The method outlined above was applied and the times moved 2,000 times.
Discrete time Markov chains

\[ P_m = P_{m+1} \]

(2.176)

\[ \mathbb{E}_m = f_{m+1} \mathbb{E}_m = f \]

(2.177)

that \( P_m \) is a reversible Markov chain

(2.178)

distribution \( f \). By positivity it is irreducible.

\( \square \)

The Gibbs sampler draws, in this very simple case the \( \Lambda_i \) are iid, gives iid \( \Lambda_i^{(j)} \), \( i=1, \ldots, N \), and then estimates \( \theta \) from the stationary distribution \( \pi \) of the resulting value from \( \pi \). But if we knew how the Gibbs sampler! As outlined in Exercise 14, any distribution is exponentially fast, so we have a burn-in period in order to get close enough to the burn-in period do we actually start. The length of the burn-in period is a subject of

(continued) Since the Gibbs sampler is possible. The reverse chain must have the forward chain. We use this to create an observed value \( X_0=x \), we run the Gibbs chain \( x=y \). The we simulate \( N-1 \) paths \( N \) in \( y \), yielding observations \( x_{(i)}^{(j+1)}=x_{(i+1)}^{(j+1)} \). Newly (early) is an observation from \( \pi \), the same \( x_{(j)}^{(1)} \), \( x_{(j)}^{(2)} \), \ldots, \( x_{(j)}^{(N)} \) 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 \( x_i \) and \( x_j \). We need to construct a Markov chain having the uniform distribution on \( S \) as its stationary distribution. Let \( y=(y_{ij}) \) be a configuration in \( S \). Consider any sub-rectangle of \( y \) 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 \( z \) in \( S \). We call this procedure a switch, and the sub-rectangle switchable. Figure 2.3 shows this concept.

![Figure 2.3. Two switchable sub-rectangles in a 5x5 table.](image)

Any table \( z \in S \) can be reached from \( y \) 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 \( y \) to \( z \) in one step, we must do that by switching a single rectangle. Thus \( P_{y,z}=P_{z,y} \), 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 decreasing 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 \( \chi^2 \) statistic for independence in the 2x2 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 \( (N_{11}+c_{11}/N)^2/(c_{11}+c_{11}/N) \), which for this table is \( (1-2/3/5)^2/(2/3/5)=0.033 \). 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
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 \( q_{xy} \) (having \( q_{xy} = q_{yx} \)). 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 \( q_{xy} \).
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 \( f(x) \) satisfies the positivity condition (1.14), the Metropolis algorithm generates an ergodic Markov chain with stationary distribution \( f = f(x); x \in S \).

### 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

\[
N_i(n) = \sum_{t=1}^{n} I(x_t = i)
\]

the likelihood of the parameter becomes

\[
L_i(P) = \prod_{n=1}^{N} f_{x_t}^{N_i(n)} (1 - f_{x_t})^{N-x_t}
\]

where \( L_i(P) \) is the likelihood for the parameter \( i \).

In other words, we can treat the likelihood of the parameter of ones separately using the independence assumptions.

We want to maximize the likelihood of the parameter of ones separately using the independence assumptions.

Setting the derivative with respect to \( P_i \) to zero,

We can think of \( P_i \) as a step function with \( P_i \) as the height of the step. Thus, the estimates are the jumps in the step function, which represent the transitions at each step. The likelihood of the parameter becomes

\[
L_i(P) = \prod_{n=1}^{N} f_{x_t}^{N_i(n)} (1 - f_{x_t})^{N-x_t}
\]

where \( L_i(P) \) is the likelihood for the parameter \( i \).

Let \( S \) be the state space, which is always finite. We will assume that the process is in state \( i \) at each time, and that transitions occur at times \( \tau \). Then the likelihood for the parameter of ones separately using the independence assumptions is

\[
L_i(P) = \prod_{n=1}^{N} f_{x_t}^{N_i(n)} (1 - f_{x_t})^{N-x_t}
\]

where \( L_i(P) \) is the likelihood for the parameter \( i \).

**Proposition 2.11** Let \( S \) be the state space, which is always finite. We will assume that the process is in state \( i \) at each time, and that transitions occur at times \( \tau \). Then the likelihood for the parameter of ones separately using the independence assumptions is

\[
L_i(P) = \prod_{n=1}^{N} f_{x_t}^{N_i(n)} (1 - f_{x_t})^{N-x_t}
\]

where \( L_i(P) \) is the likelihood for the parameter \( i \).

**Proof**

We can think of \( P_i \) as a step function with \( P_i \) as the height of the step. Thus, the estimates are the jumps in the step function, which represent the transitions at each step. The likelihood of the parameter becomes

\[
L_i(P) = \prod_{n=1}^{N} f_{x_t}^{N_i(n)} (1 - f_{x_t})^{N-x_t}
\]

where \( L_i(P) \) is the likelihood for the parameter \( i \).

Let \( S \) be the state space, which is always finite. We will assume that the process is in state \( i \) at each time, and that transitions occur at times \( \tau \). Then the likelihood for the parameter of ones separately using the independence assumptions is

\[
L_i(P) = \prod_{n=1}^{N} f_{x_t}^{N_i(n)} (1 - f_{x_t})^{N-x_t}
\]

where \( L_i(P) \) is the likelihood for the parameter \( i \).

**Proposition 2.11** Let \( S \) be the state space, which is always finite. We will assume that the process is in state \( i \) at each time, and that transitions occur at times \( \tau \). Then the likelihood for the parameter of ones separately using the independence assumptions is

\[
L_i(P) = \prod_{n=1}^{N} f_{x_t}^{N_i(n)} (1 - f_{x_t})^{N-x_t}
\]

where \( L_i(P) \) is the likelihood for the parameter \( i \).
Discrete time Markov chains

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

Parameter of interest is a point in the space of all transition matrices. Let $N_{ij}(n)=\sum_{t=1}^{n} 1(\xi_{t-1}=i, \xi_{t}=j)$ count the number of $i,j$-transitions. If $N_{ij}(n)=n_{ij}$, the likelihood (2.9) takes the form

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

where $L_i(\mathbf{P}_l) = \prod_{j \in S} p_{ij}^{n_{ij}}$ depends only on the elements in the $i$th row $\mathbf{P}_l$ of $\mathbf{P}$. In other words, we are estimating $|S|$ independent probability distributions. Let $l(\mathbf{P}_0, \mathbf{P}) = \log L(\mathbf{P}_0, \mathbf{P})$. Then (2.80) corresponds, with obvious notation, to

$$l(\mathbf{P}_0, \mathbf{P}) = l_0(\mathbf{P}_0) + \sum_{i \in S} l_i(\mathbf{P}_i).$$

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

$$l_i(\mathbf{P}_l) + \lambda(\mathbf{P}_i, 1^T - 1) = \sum_{j \in S} n_{ij} \log p_{ij} + \lambda(\sum_{j \in S} p_{ij} - 1).$$

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

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

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 $\hat{T}_i = n_i$, i.e., that state $i$ is reached for the first time at time $n_i$, 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 $\mathbf{P}$.

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

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

□

---

The Metropolis algorithm, employs an auxiliary algorithm, the Hastings (1971). The chain transitions are determined by the following update method:

1. Choose any $i \in S$ and $x_0$.
2. Choose $x_1$ according to $p_{x_0x_1}$.
3. Choose $x_2$ according to $p_{x_1x_2}$.
4. Continue.

We define a new state $x$ to be a new state $x'$ with probability $1-r$. The new state $x'$ depends on the previous state, so that $x'$ is still a Gibbs sampler, the simple way to this satisfies detailed balance. We order the values so that $f(i) \leq f(j)$ implies $q_{ij} f(i) f(j) = q_{ji} f(j) f(i)$, using the same matrix. A generalization of both the and order is due to Hastings (1971), and we need to find the following result.

**Proposition 1.14** The Metropolis-Hastings algorithm is a Markov chain with stationary distribution $\pi$. Any function $f$ is a positive probability.

---

The discrete Markov chain, we saw in section 2.6, has a transition matrix, and thus any function of the likelihood. In this section we study the likelihood theory for both the parametric and the nonparametric case, where the
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

$$L_c(\hat{\mathbb{P}}) = \prod_{i \in S} L_q(\hat{\mathbb{P}}).$$

(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 $L(p_0, \mathbb{P})$ subject to the additional constraint that $p_0=p_0\mathbb{P}$. 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

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

(2.185)

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

$$\frac{n_{10} + 1}{(1-p_{11})} + \frac{n_{11}}{p_{11}} = -\frac{1}{1-(p_{11}-p_{01})}$$

$$\frac{n_{00} + 1}{(1-p_{01})} + \frac{n_{01}}{p_{01}} = \frac{1}{1-(p_{11}-p_{01})}$$

(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

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

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 $Y_n=(x_n, x_{n+1})$, let

$$\bar{p}_{kl} = \frac{P(x_{n+1}=y|x_n=x)}{P(x_{n+1}=y|x_n=x)}.$$  

(b) If $(X_n)$ is ergodic, there exists an invariant distribution $\bar{\pi}$ given by

$$P(Y_n=(x_n, x_{n+1}) = \bar{\pi}(x) \bar{p}_{kl}.  

The remaining parts of the proof follow.

We now use the step chain to estimate the number of transitions of the estimator $\bar{\mathbb{P}}$. Let $N_{ij}(n)$ denote the number of $(i,j)$-transitions to time $n$.

Theorem 2.14. As $n \to \infty$, $N_{ij}(n) \to \infty$ for all $i$, $j$.

Proof. If $p_{ij} = 0$, then $N_{ij}(n) = 0$ obviously. If $p_{ij} > 0$, then

$$N_{ij}(n) = n \left( \frac{1}{n} N_{ij}(n) \right)$$

and since the step chain is ergodic,

(a) that

$$\frac{1}{n} N_{ij}(n) \to \bar{p}_{ij}$$

Using Theorem 2.10, we have

$$\bar{p}_{ij} = \frac{P(x_{n+1}=y|x_n=x)}{P(x_{n+1}=y|x_n=x)}.$$
Likelihood theory for Markov chains

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 \( S = \{(i, j) \in S^2 : p_{ij} > 0\} \), initial distribution \( \tilde{\pi}_0 \) given by \( \tilde{\pi}_0(i, j) = p_0(i)p_{ij} \) and transition matrix \( \tilde{P} = (\tilde{P}_{ij}) \) given by

\[
\tilde{P}_{ij} = 1(i = l)p_{ij}. \tag{2.187}
\]

(b) If \((X_n)\) is ergodic and \( p_0 = \pi \), then \((Y_n)\) is also ergodic with stationary initial distribution \( \tilde{\pi} \) given by \( \tilde{\pi}(i, j) = \pi(i)p_{ij} \).

**Proof**

\[
P(Y_n = (i, j) \mid Y_{n-1} = (k_1, l_1), \ldots, Y_0 = (k_n, l_n)) = P(X_{n+1} = j, X_n = i \mid X_{n-1} = k_1, \ldots, X_0 = k_n) = P(X_{n+1} = j, X_n = i \mid X_{n-1} = k_1) \tag{2.188}
\]

\[
= P(Y_n = (i, j) \mid Y_{n-1} = (k_1, l_1))
\]

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

\[
P(Y_n = (i, j) \mid Y_{n-1} = (k, l)) = P(X_{n+1} = j \mid X_n = i, X_{n-1} = k) \times P(X_n = i \mid X_{n-1} = k) \tag{2.189}
\]

\[
= \begin{cases} 
P(X_{n+1} = j \mid X_n = i) & \text{if } i = l \\ 0 & \text{otherwise.} \end{cases}
\]

The remaining parts are proved using similar computations. \(\Box\)

We now use the step chain and the ergodic theorem to show strong consistency of the estimator \( \hat{\phi}_{ij} \). Let \( \hat{\phi}_{ij}(n) \) 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 \( \hat{\phi}_{ij}(n) \to p_{ij} \) with probability \( 1 \) as \( n \to \infty \) for all \( i, j \in S \), regardless of the initial distribution.

**Proof**

If \( p_{ij} = 0 \) then \( \hat{\phi}_{ij}(n) = 0 \), so we only need to consider \((i, j) \in \hat{S}) \). But

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

and since the step chain is ergodic we have from Theorem 2.10 and Lemma 2.5 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
Discrete time Markov chains

\[
\frac{1}{n}N_i(n) \to \pi_i \quad \text{with probability 1},
\]  
(2.192)

whence the result follows. \(\square\)

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

**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 \(P_{ij}\).

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

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,

\[
\left[ N_i(n)^{\frac{1}{2}}(p_{ij} - p_{ii}), i, j \in S \right] \overset{d}{\to} \mathcal{N}(0, \Sigma) \tag{2.193}
\]

where

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

**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. \(\square\)

**Proof** Since \(n\pi_i/N_i(n) \to 1\) we need only show that

\[
\left[ \frac{N_{ij}(n) - N_i(n)p_{ij}}{(n\pi_i)^{\frac{1}{2}}}, i, j \in S \right] \overset{d}{\to} \mathcal{N}(0, \Sigma). \tag{2.195}
\]

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

\[
\left[ \frac{Q_{ij}(n) - [n\pi_i]p_{ij}}{[n\pi_i]^{\frac{1}{2}}}, i, j \in S \right] \overset{d}{\to} \mathcal{N}(0, \Sigma). \tag{2.196}
\]
Likelihood theory for Markov chains

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

\[ D_n = (n\pi_i)^{-\frac{1}{2}}(N_i(n) - n\pi_i)p_{ij} - Q_{ij}(n) + n\pi_i p_{ij} \to 0. \quad (2.197) \]

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

\[ D_n = (n\pi_i)^{-\frac{1}{2}}(S_n - n\pi_i). \quad (2.198) \]

Then

\[
\begin{align*}
P( | D_n | > \varepsilon) & \leq P( | D_n | > \varepsilon, | N_i(n) - n\pi_i | \leq x\varepsilon^\frac{1}{2} ) \\
& + P( | N_i(n) - n\pi_i | > x\varepsilon^\frac{1}{2} )
\end{align*}
\]

(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 \( | m - n\pi_i | \leq x\varepsilon^\frac{1}{2} \). Using Chebyshev's inequality twice yields an upper bound of

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

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

\[ \mathbb{E}(S_m - n\pi_i)^4 \leq (x\varepsilon^\frac{1}{2} + 1)\kappa + 3(x\varepsilon^\frac{1}{2} + 1)^2 \sigma^4. \quad (2.201) \]

The sum in (2.200) has at most \( 2x\varepsilon + 1 \) terms, so we get

\[
\begin{align*}
P( | D_n | > \varepsilon) & \leq \frac{2x\varepsilon^\frac{1}{2} + 1}{n^2 \pi_i^2 \varepsilon^4} \left( (x\varepsilon^\frac{1}{2} + 1)\kappa + 3(x\varepsilon^\frac{1}{2} + 1)^2 \sigma^4 \right) \\
& + P( | N_i(n) - n\pi_i | > x\varepsilon^\frac{1}{2} )
\end{align*}
\]

(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 \( p_{11}(1-p_{11}) \pi_i \). We estimate that variance using \( \hat{p}_{11} = n_{11}/n_1 \) and \( \hat{p}_{11} = n_{11}/n_1 \) where \( n_{11} = \Sigma_{i=3}^{15} n_{i1} \), etc. Since \( n_{11} = 843, n_1 = 771 \) 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, $0.95^2 = 0.9025$. 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>
<thead>
<tr>
<th></th>
<th>Vowel next</th>
<th>Consonant next</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>1106</td>
<td>7532</td>
<td>8638</td>
</tr>
<tr>
<td>Consonant</td>
<td>7533</td>
<td>3829</td>
<td>11362</td>
</tr>
<tr>
<td>Total</td>
<td>8639</td>
<td>11361</td>
<td>20000</td>
</tr>
</tbody>
</table>

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 $\rho$ of switching from one type of character to another. The transition matrix for this hypothesis is

$$P = \begin{pmatrix} 1-\rho & \rho \\ \rho & 1-\rho \end{pmatrix},$$  \hspace{1cm} (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 $p_{ij} = p_{ij}(\theta)$ depend on an unknown parameter $\theta$, taking values in $\Theta$, an open subset of $R^d$. We will need some regularity conditions:

**Conditions A:**
(i) $D = \{i,j:p_{ij}(\theta) > 0\}$ is independent of $\theta$.
(ii) Each $p_{ij}(\theta) \in C^3$, i.e., each $p_{ij}(\theta)$ is three times continuously differentiable.
(iii) The $dx \times r$-matrix $\partial p_{ij}(\theta) / \partial \theta_k$, $i,j \in D$, $k = 1, \ldots, r$, where $d$ is the cardinality of $D$, has rank $r$.
(iv) For each $\theta$ there is only one ergodic class and no transient states.

### Likelihood theory

We now argue conditionally that

$$l(\tilde{\theta}) = -\frac{1}{T} \ln L,$$

where as before $\tilde{\theta}$ is the vector $\theta$ of the log-likelihood equal to

$$\frac{\partial}{\partial \theta}.$$

Let $\theta_0$ be the maximum likelihood estimate (MLE) (1961). We will

**Theorem 2.1**

(i) There is a consistent estimate

(ii) $\sqrt{T} (\hat{\theta} - \theta_0) \approx N(0, \Sigma)$

(iii) $\text{Var} \sqrt{T} (\hat{\theta} - \theta_0) \approx I^{-1}(\theta_0) \Sigma I^{-1}(\theta_0)$

The quantity is

**Application**

minimizing the log-likelihood

$$l(\theta) = \ln L(\theta),$$

where $L(\theta)$ is the likelihood is

$$l(\theta) = \ln \left( \frac{1}{T} \prod_{t=1}^{T} \frac{1}{P(Y_t|X_{t-1})} \right),$$

so from Theorem 2.1

$$(-l''(\tilde{\theta}))^{-1/2} \approx \Sigma,$$

where confidence intervals are adequate.