Contents

1 Markov Chains and Hidden Markov Models **3**

1.1 Markov Chains 4

1.2 Parameter Estimation for Markov Chains 6

1.3 Hidden Markov Models 8

1.3.1 State Sequence Estimation and Dynamic Programming 9

1.3.2 State Probability and the Foward-Backward Algorithm 9

1.3.3 Training HMMs with the EM Algorithm 11

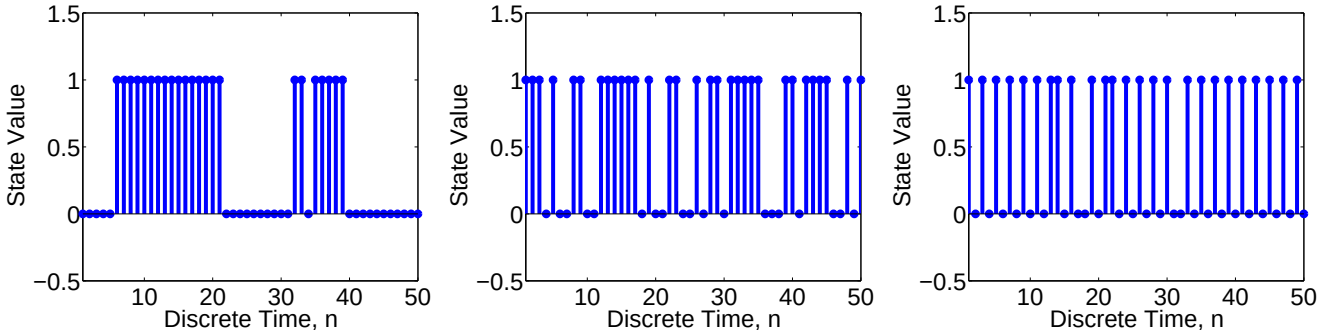
1.4 Stationary Distributions of Markov Chains 12

Chapter 1

Markov Chains and Hidden Markov Models

Notation:

- X_n - discrete time Markov chain
- $0 \leq n \leq N$ with X_0 being the initial state
- $\pi_j^{(n)} = P\{X_n = j\}$ - state distribution at time n
- $P_{i,j}^{(n)}$ - nonhomogeneous state transition probability
- $\tau_j = \pi_j^{(0)}$ - initial state distribution
- $N_j = \delta(X_0 - j)$ - initial state statistics
- $K_j = \sum_{n=1}^{N} \delta(X_n - j) = \sum_{i=0}^{M-1} K_{i,j}$ - state membership statistics
- $K_{i,j} = \sum_{n=1}^{N} \delta(X_n - j)\delta(X_{n-1} - i)$ - state transition statistics
- $b_j = \sum_{n=0}^{N-1} y_n \delta(x_n - j)$ - first order class statistics
- $S_j = \sum_{n=0}^{N-1} y_n y_n^t \delta(x_n - j)$ - second order class statistics
- $\Omega = [0, \cdots, M-1]$ - set of discrete states
- $\theta = [\tau_j, P_{i,j} : \text{for } i, j \in \Omega]$ - parameter vector
- π_j - ergodic distribution of Markov chain



$\rho = 0.02$ $\rho = 0.5$ $\rho = 0.8$

Figure 1.1: Three figures illustrating three distinct behaviors of the Markov chain example with $\rho = 0.1$, 0.5, and 0.9. In the case, of $\rho = 0.1$, the Markov Chain remains in its current state 90% of the time. When $\rho = 0.5$, the state is independent of the past with each new value, and with $\rho = 0.9$, the state changes value 90% of the time.

In this chapter, we will introduce the concept of Markov chains, and show how Markov chains can be used to model signals using structures such as hidden Markov models (HMM). Markov chains are based on the simple idea that each new sample is only dependent on the previous sample. This simple assumption makes them easy to analyze, but still allows them to be very powerful tools for modeling physical processes.

We finish the chapter by discussing what happens when a homogeneous Markov chain is run for a long time. Typically, such a Markov chain will reach a steady state, when it does, we call the Markov chair ergodic, and we give some relatively simple technical conditions to check for this situation. Ergodic analysis of Markov chains is very useful in a wide range of applications, but it will be of particular interest to us for applications such as Monte Carlo simulation, stochastic sampling from Markov Random Fields, and stochastic optimization.

1.1 Markov Chains

A Markov chain is a discrete-time and discrete-valued random process in which each new sample is only dependent on the previous sample. So let $\{X_n\}_{n=0}^{N}$ be a sequence of random variables taking values in the countable set Ω.

Definition: Then we say that X_n is a *Markov chain* if for all values of $j \in \Omega$ and all n

$$P\{X_n = j | X_k \text{ for all } k < n\} = P\{X_n = j | X_{n-1}\} \ .$$

Notice that a Markov chain is discrete in both time and value. This set of possible values for the Markov chain is known as the **state space**. Since the state space is discrete, we can assume, without lose of generality, that $\Omega = \{0, \cdots, M-1\}$, where $M = \infty$ is allowed.

Alternatively, a **discrete-time Markov process** is continuously valued and discrete in time. We will primarily focus on Markov chains because they are easy to analyze and of great practical value. However, most of the results we derive are also true for discrete-time Markov processes.

In order to analyze a Markov chain, we will first define notation to describe the marginal distribution of X_n and the probability of transitioning from state X_{n-1} to state X_n as

$$\begin{aligned} \pi_j^{(n)} &\triangleq P\{X_n = j\} \\ P_{i,j}^{(n)} &\triangleq P\{X_n = j | X_{n-1} = i\} \end{aligned}$$

for $i, j \in \Omega$. If the transition probabilities, $P_{i,j}^{(n)}$, do not depend on time, n, then we say that X_n is a **homogeneous Markov chain**. For the remainder of this chapter, we will assume that Markov chains are homogeneous unless otherwise stated. It is reasonable to expect that a homogeneous Markov chain may have a time varying distribution due to the transients associated with its initial condition. However, if the homogeneous Markov chain is given sufficient time to reach steady-state behavior, then we would expect that its distribution should become stable. Section 1.4 will treat this issue in detail.

From the Markov process, we can derive an expression for the probability of the sequence $\{X_n\}_{n=0}^N$. In order to simplify notation, let the distribution of the initial state be denoted by $\tau_j = P\{X_0 = j\}$. Then the probability of the Markov chain sequence is the product of the probability of the initial state, τ_{x_0}, with each of the N transitions from state x_{n-1} to state x_n. This product is given by

$$p(x) = \tau_{x_0} \prod_{n=1}^{N} P_{x_{n-1},x_n} \ .$$

From this expression, it is clear that the Markov chain is parameterized by its initial distribution, τ_j, and its transition probabilities, $P_{i,j}$.

Example 1.1.1: Let $\{X_n\}_{n=0}^{N}$ be a Markov chain with state-space $\Omega = \{0, 1\}$ and parameters given by

$$\begin{aligned} \tau_j &= 1/2 \\ P_{i,j} &= \begin{cases} 1-\rho & \text{if } j = i \\ \rho & \text{if } j \neq i \end{cases} . \end{aligned}$$

This Markov chain starts with an equal chance of being 0 or 1. Then with each new state, it has probability ρ of changing states, and probability $1 - \rho$ of remaining in the same state. If ρ is small, then the probability of changing state is small, and the Markov chain is likely to stay in the same state for an extended period of time. When $\rho = 1/2$ each new state is independent of the previous state; and when ρ is approximately 1, then the state is likely to change with each new value of n. These three cases are illustrated in Figure 1.1.

If we define the statistic

$$K = N - \sum_{n=1}^{N} \delta(X_n - X_{n-1}) \ ,$$

then K is the number of times that the Markov chain changes state. Using this definition, we can express the probability of the sequence as

$$p(x) = (1/2)(1-\rho)^{N-K}\rho^K \ .$$

1.2 Parameter Estimation for Markov Chains

Markov chains are very useful for modeling physical phenomena because, as with AR models, we simply predict the distribution of the next sample from the previous one. However, the challenge with Markov chains is often to accurately estimate the parameters of the model from real data. If the Markov chain is nonhomogeneous, then the number of parameters will grow with the length of the sequence. So it is often practically necessary to assume

a homogeneous Markov chain. However, even when the Markov chain is homogeneous, there are M^2 parameters to choose, where $M = |\Omega|$ is the number of states.[1] Therefore, it is important to have effective methods to estimate these parameters.

Fortunately, we will see that Markov chains are exponential distributions so parameters are easily estimated from natural sufficient statistics. Let $\{X_n\}_{n=0}^N$ be a Markov chain parameterized by $\theta = [\tau_j, P_{i,j} : \text{for } i, j \in \Omega]$, and define the statistics

$$\begin{aligned} N_j &= \delta(X_0 - j) \\ K_{i,j} &= \sum_{n=1}^{N} \delta(X_n - j)\delta(X_{n-1} - i) \ . \end{aligned}$$

The statistic $K_{i,j}$ essentially counts the number of times that the Markov chain transitions from state i to j, and the statistic N_j counts the number of times the initial state has value j. Using these statistics, we can express the probability of the Markov chain sequence as

$$p(x|x_0) = \left(\prod_{j \in \Omega} \tau_j^{N_j} \right) \left(\prod_{i \in \Omega} \prod_{j \in \Omega} P_{i,j}^{K_{i,j}} \right)$$

which means that the log likelihood has the form

$$\log p(x|x_0) = \sum_{i \in \Omega} \sum_{j \in \Omega} \left\{ N_j \log(\tau_j) + K_{i,j} \log(P_{i,j}) \right\} \ . \tag{1.1}$$

Based on this, it is easy to see that the distribution of the Markov chain X_n with parameter θ is from an exponential family, and that N_i and $K_{i,j}$ are its natural sufficient statistics. From (1.1), we can derive the maximum likelihood estimates of the parameter θ in much the same manner as is done for the ML estimates of the parameters of a Bernoulli sequence. This results in the ML estimates

$$\begin{aligned} \hat{\tau}_j &= N_j \\ \hat{P}_{i,j} &= \frac{K_{i,j}}{\Sigma_{j \in \Omega} K_{i,j}} \ . \end{aligned}$$

So the ML estimate of transition parameters is quite reasonable. It simply counts the rate at which a particular i to j transition occurs.

[1] The quantity M^2 results from the sum of M parameters for the initial state plus $M(M-1)$ parameters for the transition probabilities. The value $M(M-1)$ results form the fact that there are M rows to the transition matrix and each row has $M-1$ degrees of freedom since it must sum to 1.

Figure 1.2: This diagram illustrates the dependency of quantities in a hidden Markov model. Notice that the values X_n form a Markov chain in time, while the observed values, Y_n, are dependent on the corresponding labels X_n.

1.3 Hidden Markov Models

One important application of Markov chains is in **hidden Markov models (HMM)**. Figure 1.3 shows the structure of an HMM. The discrete values $\{X_n\}_{n=0}^N$ form a Markov chain in time, and their values determine the distribution of the corresponding observations $\{Y_n\}_{n=1}^N$. Much like in the case of the Gaussian mixture distribution, the labels X_n are typically not observed in the real application, but we can imagine that their existence explains the changes in behavior of Y_n over long time scales. As with the example of the mixture distribution from Section **??**, the HMM model is a doubly-stochastic model because the unobserved stochastic process X_n controls the observed stochastic process Y_n.

HMMs are very useful in applications such as audio or speech process. For these applications, Y_n are random feature vectors describing the local sound properties, and the discrete random variables X_n, are the underlying state of the audio signal. So for example in speech applications, X_n might be a label that assigns the specific phoneme or sound that is being voiced in the pronunciation of a word.

In order to analyze the HMM, we start by defining some notation. Let the density function for Y_n given X_n be given by

$$f(y|k) = P\{Y_n \in dy | X_n = k\} \ ,$$

and let the Markov chain X_n be parameterized by $\theta = [\tau_j, P_{i,j} : \text{for } i, j \in \Omega]$, then the density function for the sequences Y and X are given by [2]

$$p(y, x|\theta) = \tau_{x_0} \prod_{n=1}^{N} \left\{ f(y_n|x_n) P_{x_{n-1},x_n} \right\} \ .$$

and assuming that no quantities are zero, the log likelihood is given by

$$\log p(y, x|\theta) = \log \tau_{x_0} + \sum_{n=1}^{N} \left\{ \log f(y_n|x_n) + \log P_{x_{n-1},x_n} \right\} \ .$$

[2]This is actually a mixed probability density and probability mass function. This is fine as long as one remembers to integrate over y and sum over x.

There are two basic tasks which one typically needs to solve with HMMs. The first task is to estimate the unknown states X_n from the observed data, Y_n, and the second task is to estimate the parameters θ from the observations Y_n. The following two sections explain how these problems can be solved.

1.3.1 State Sequence Estimation and Dynamic Programming

One common estimate for the states, X_n, is the MAP estimate given by

$$\begin{aligned}\hat{x} &= \arg\max_{x\in\Omega^N} p(x|y,\theta) \\ &= \arg\max_{x\in\Omega^N} \log p(y,x|\theta)\ .\end{aligned}$$

When we expand this optimization problem out using the log likelihood of the HMM, it is not obvious that it can be solved in closed form.

$$\hat{x} = \arg\max_{x\in\Omega^N}\left\{\log\tau_{x_0} + \sum_{n=1}^{N}\left\{\log f(y_n|x_n) + \log P_{x_{n-1},x_n}\right\}\right\} \tag{1.2}$$

The difficulty appears to be in the coupling between terms n and $n-1$ cause by the transition probabilities of the Markov chain. Interestingly, the optimization of (1.2) can be efficiently computed using **dynamic programming**. To do this, we first define the quantity $L(j,n)$ to be the log probability of the state sequence $[x_n,\cdots,x_N]$ that results in the largest probability and starts with the value $x_n = j$. The values of $L(k,n)$ can be computed with the back-ward recursion

$$L(i,n-1) = \arg\max_{j\in\Omega}\left\{\log f(y_n|j) + \log P_{i,j} + L(j,n)\right\}$$

using the initial condition that $L(j,N) = 0$. Once these values are computed, the MAP sequence can be computed by

$$\begin{aligned}\hat{x}_0 &= \arg\max_{j\in\Omega}\left\{L(j,0) + \log\tau_k\right\} \\ \hat{x}_n &= \arg\max_{j\in\Omega}\left\{\log P_{\hat{x}_{n-1},j} + \log f(y_n|j) + L(j,n)\right\}\end{aligned}$$

1.3.2 State Probability and the Foward-Backward Algorithm

It some cases, it may be important to know the posterior distribution of the states, X_n, given the observations Y. For example, the estimator known

as the **maximizer of the posterior marginals (MPM)** minimizes the classification error of each state and is given by [**?**]

$$\begin{aligned} \hat{x} &= \arg\max_{x_n \in \Omega} p(x_n|y,\theta) && (1.3) \\ &= \arg\max_{x_n \in \Omega} \log p(x_n, y, \theta) \ . && (1.4) \end{aligned}$$

Calculation of the MPM estimator requires the either posterior distribution of X_n given Y. Unfortunately, a naive calculation of this joint distribution is given by a summation over N variables $X_0, \cdots, X_{n-1}, X_{n+1}, \cdots, X_N$, each of which can take on M values,

$$p(x_n, y|\theta) = \sum_{x_0=0}^{M-1} \sum_{x_{n-1}=0}^{M-1} \cdots \sum_{x_{n+1}=0}^{M-1} \sum_{x_N=0}^{M-1} p(x_n, x_{k\neq n}, y|\theta) \ ,$$

and evaluation of this sum is typically intractable since it requires M^N operations.

Fortunately, there is a computationally efficient method for computing these posterior probabilities that exploits the 1D structure of the HMM. The algorithm for doing this is known as the **forward-backward algorithm** due to its forward and backward recursion structure in time. The forward recursion is given by

$$\begin{aligned} \alpha_1(j) &= \tau_j \\ \alpha_{n+1}(j) &= \sum_{i=\Omega} \alpha_n(i) P_{i,j}\, p(y_{n+1}|j) \end{aligned}$$

The backward recursion is given by

$$\begin{aligned} \beta_N(i) &= 1 \\ \beta_{n+1}(i) &= \sum_{j=\Omega} P_{i,j} p(y_{n+1}|j) \beta_{n+1}(j) \end{aligned}$$

From this the required posterior probabilities can be calculated as

$$P\{X_n = k|Y = y, \theta\} = \frac{\alpha_n(i)\beta_n(i)}{p(y)} \qquad (1.5)$$

$$P\{X_n = j, X_{n-1} = i|Y = y, \theta\} = \frac{\alpha_{n-1}(i) p(y_n|j) P_{i,j} \beta_n(i)}{p(y)} \qquad (1.6)$$

where

$$p(y) = \sum_{i \in \Omega} \alpha_n(i)\beta_n(i) \ .$$

1.3.3 Training HMMs with the EM Algorithm

It order to Train the HMM, it is necessary to estimate the parameters for the HMM model. This will be done by employing the EM algorithm, so we will need to derive both the E and M-steps. In a typical application, the observe quantity Y_n is an L dimensional multivariate Gaussian random vector with distribution $N(\mu_{x_n}, R_{x_n})$, so it is also necessary to estimate the parameters μ_j and R_j for each of the states $j \in \Omega$.

In this case, the joint distribution $p(x, y|\theta)$ is an exponential distribution with parameter vector $\theta = [\mu_j, R_j, \tau_j, P_{i,j} : \text{for } i, j \in \Omega]$, and natural sufficient statistics given by

$$\begin{aligned} b_j &= \sum_{n=1}^{N-1} y_n \delta(x_n - j) \\ S_j &= \sum_{n=1}^{N-1} y_n y_n^t \delta(x_n - j) \\ N_j &= \delta(x_0 - j) \\ K_{i,j} &= \sum_{n=1}^{N} \delta(x_n - j)\delta(x_{n-1} - i) \ , \end{aligned}$$

and another useful statistic, K_j, can be computed from these as

$$K_j = \sum_{i \in \Omega} K_{i,j} = \sum_{n=1}^{N} \delta(x_n - j) \ .$$

Using these sufficient statistics, the ML estimate of θ is given by

$$\begin{aligned} \hat{\mu}_j &= \frac{b_j}{K_j} \\ \hat{R}_j &= \frac{S_j}{K_j} - \frac{\mu_j \, \mu_j^t}{K_j^2} \\ \hat{\tau}_j &= N_j \\ \hat{P}_{i,j} &= \frac{K_{i,j}}{\sum_{j \in \Omega} K_{i,j}} \ . \end{aligned}$$

From this and the results of the previous chapter, we know that we can compute the EM updates by simply substituting the natural sufficient statistics with their conditional expectation given Y. So to compute the EM

update of the parameter θ, we first compute the conditional expectation of the sufficient statistics in the E-step as

$$\begin{aligned}
\bar{b}_j &\leftarrow \sum_{n=0}^{N-1} y_n P\{X_n = j | Y = y, \theta\} \\
\bar{S}_j &\leftarrow \sum_{n=0}^{N-1} y_n y_n^t P\{X_n = j | Y = y, \theta\} \\
\bar{N}_j &\leftarrow P\{X_0 = j | Y = y, \theta\} \\
\bar{K}_{i,j} &\leftarrow \sum_{n=1}^{N} P\{X_n = j, X_{n-1} = i | Y = y, \theta\} \\
\bar{K}_k &\leftarrow \sum_{n=0}^{N-1} P\{X_n = k | Y = y, \theta\} \ ,
\end{aligned}$$

and the HMM model parameters, $\theta = [\mu_j, R_j, \tau_j, P_{i,j} : \text{for } i, j \in \Omega]$, are then updated using the M-step

$$\begin{aligned}
\hat{\mu}_j &\leftarrow \frac{b_j}{K_j} \\
\hat{R}_j &\leftarrow \frac{S_j}{K_j} - \frac{\mu_j \, \mu_j^t}{K_j^2} \\
\hat{\rho}_i &\leftarrow N_i \\
\hat{P}_{i,j} &\leftarrow \frac{K_{i,j}}{\sum_{j \in \Omega} K_{i,j}} \ .
\end{aligned}$$

Notice that the E-step requires the computation of the posterior probability of X_n given Y. Fortunately, this may be easily computed using the forward-backward algorithm given in equations (1.5) and (1.6) in the previous section.

1.4 Stationary Distributions of Markov Chains

In this section, we present the basic results of Markov chain theory that we will need to analyze simulation methods [1]. The fundamental concept is that a well behaved Markov chain will eventually reach a stable stationary distribution after the transient behavior dies away. The objective will be to define the technical conditions that ensure that this happens.

Let $\{X_n\}_{n=0}^{\infty}$ be a discrete valued and discrete state homogeneous Markov

chain taking values in the set Ω. Define the notation

$$\begin{aligned} \pi_j^{(n)} &\stackrel{\triangle}{=} P\{X_n = j\} \\ P_{i,j} &\stackrel{\triangle}{=} P\{X_n = j | X_{n-1} = i\} \ , \end{aligned}$$

and define $\pi^{(n)}$ to be the corresponding $1 \times |\Omega|$ row vector, and P to be the corresponding $|\Omega| \times |\Omega|$ matrix. A fundamental property of Markov chains is that the marginal probability density for time $n+1$ can be expressed as

$$\pi_j^{(n+1)} = \sum_{i \in \Omega} \pi_i^{(n)} P_{i,j} \ . \tag{1.7}$$

In matrix notation, this is equivalent to

$$\pi^{(n+1)} = \pi^{(n)} P \ . \tag{1.8}$$

Repeated application of (1.8) results in the equation

$$\pi^{(n+m)} = \pi^{(n)} P^m \ , \tag{1.9}$$

or equivalently

$$\pi_j^{(n+m)} = \sum_{i \in \Omega} \pi_i^{(n)} P_{i,j}^m \ . \tag{1.10}$$

where $P_{i,j}^m$ is defined by the recursion

$$P_{i,j}^{m+1} = \sum_{k \in \Omega} P_{i,k}^m P_{k,j} \ .$$

More generally, the Chapman-Kolmogorov relation states that

$$P_{i,j}^{m+k} = \sum_{k \in \Omega} P_{i,k}^m P_{k,j}^k \ .$$

We next define a number of properties for homogeneous Markov chains that we will need.

Definition 1 *Communicating states*
The states $i, j \in \Omega$ of a Markov chain are said to communicate *if there exists integers $n > 0$ and $m > 0$ such that $P_{i,j}^n > 0$ and $P_{j,i}^m > 0$.*

Intuitively, two states communicate if it is possible to transition between the two states. It is easily shown that communication is an equivalence property, so it partitions the set of states into disjoint sets that all communicate with each other. This leads to a natural definition for irreducible Markov chains.

Definition 2 *Irreducible Markov Chain*
A discrete time discrete state homogeneous Markov chain is said to be irreducible if for all $i, j \in \Omega$ i and j communicate.

So a Markov chain is irreducible if it is possible to change from any initial state to any other state in finite time.

In some cases, a state of a Markov chain my repeat periodically. This type of periodic repetition can last indefinitely.

Definition 3 *Periodic state*
We denote the period of a state $i \in \Omega$ by the value $d(i)$ where $d(i)$ is the largest integer so that $P_{i,i}^n = 0$ whenever n is not divisible by $d(i)$. If $d(i) > 1$, then we say that the state i is periodic.

It can be shown that states of a Markov chain that communicate must have the same period. Therefore, all the states of an irreducible Markov chain must have the same period. We say that the an irreducible Markov chain is aperiodic if all the states have period 1.

Using these definitions, we may now state a theorem which gives basic conditions for convergence of the distribution of the Markov chain.

Theorem 1 *Limit Theorem for Markov Chains*
Let $X_n \in \Omega$ be a discrete-state discrete-time homogeneous Markov chain with transition probabilities $P_{i,j}$ and the following additional properties

- *Ω is a finite set*
- *The Markov chain is irreducible*
- *The Markov chain is aperiodic*

There exists a unique stationary distribution π, which for all states i is given by

$$\pi_j = \lim_{n\to\infty} P_{i,j}^n > 0 \tag{1.11}$$

and which is the unique solution to the following set of equations.

$$1 = \sum_{i\in\Omega} \pi_i \tag{1.12}$$

$$\pi_j = \sum_{i\in\Omega} \pi_i P_{i,j} \ . \tag{1.13}$$

The relations of (1.13) are sometimes called the full balance equations (FBE). Any probability density which solves the FBE is guaranteed to be the stationary distribution of the Markov chain. Furthermore, in the limit as $n \to \infty$, the Markov chain is guaranteed to converged to this stationary distribution independently of the initial state. Markov chains that have this property of (1.11) are said to be *ergodic*. It can be shown that for ergodic Markov chains, expectations of state variables can be replace by time averages, which will be very useful in later sections.

Theorem 1 gives relatively simple conditions to establish that a Markov chain has a stationary distribution. However, while it may be known that a stationary distribution exists, it may be very difficult to compute the solution of the FBEs to determine the precise form of that distribution. It is often useful to use the property of reversibility as a method to solve this problem. First we must show that the time reverse of a Markov chain is itself a Markov chain.

Proposition 1 *Time Reverse of Markov Chains*
Let $\{X_n\}_{n=-\infty}^{\infty}$ be a Markov Chain. Then the time reversed process $Y_n = X_{-n}$ is also a Markov chain.

Since the time reversal of a Markov chain is also a Markov chain, it must also have a transition distribution which we denote at $Q_{i,j}$. Therefore, we know that

$$P\{X_n = i, X_{n+1} = j\} = \pi_i^n P_{i,j} = \pi_j^{n+1} Q_{j,i} \ .$$

If the Markov chain has a stationary distribution, then $\pi = \pi^n = \pi^{n+1}$, and we have that

$$\pi_i P_{i,j} = \pi_j Q_{j,i} \ .$$

Furthermore, if the Markov chain is reversible, then we know that $P_{i,j} = Q_{i,j}$. This yields the so-called detailed balance equations (DBE).

$$\pi_i P_{i,j} = \pi_j P_{j,i} \tag{1.14}$$

The DBE specify that the rate of transitions from state i to state j equals the rate of transitions from state j to i. This is always the case when a Markov chain is reversible.

Definition 4 *Reversible Markov Chain*
A homogeneous Markov chain with transition probabilities $P_{i,j}$ is said to be

reversible if there exists a stationary distribution which solves the detailed balance equations of (1.14).

Notice that if one finds a solution to the DBE, then this solution must also be a solution to the FBE, and is therefore the stationary distribution of an ergodic Markov chain. To see this

$$\begin{aligned}\sum_{i\in\Omega}\pi_i P_{i,j} &= \sum_{i\in\Omega}\pi_j P_{j,i} \\ &= \pi_j \sum_{i\in\Omega} P_{j,i} \\ &= \pi_j\end{aligned}$$

Finally, it is useful to study the convergence behavior of Markov chains. Let us consider the case when the Markov chains state Ω is finite. In this case, we may use a matrix representation for P. We know that any matrix P may be diagonalized using eigen decomposition and expressed in the form

$$P = E^{-1}\Lambda E$$

where the rows of E are the left hand eigenvectors of P, and Λ is a diagonal matrix of eigenvalues. Using this decomposition, we can see that

$$\begin{aligned}P^m &= P^{m-2}\,E^{-1}\Lambda E^{-1}\,E\Lambda E \\ &= P^{m-2}\,E^{-1}\Lambda^2 E \\ &= E^{-1}\Lambda^m E\end{aligned}$$

So the distribution at time n is given by

$$\pi^{(n)} = \pi^{(0)}E^{-1}\Lambda^m E$$

When P corresponds to a irreducible and aperiodic Markov chain, then it must have a stationary distribution. In this case, exactly one of the eigenvalues is 1, and the remaining eigenvalues have magnitude strictly less then 1. In this way, we can see that the distribution of $\pi^{(n)}$ converges geometrically to its stationary distribution π.

Chapter 1 Problems

1. Let $\{X_i\}_{i=0}^N$ be a Markov Chain with $X_i \in \{0, \cdots, M-1\}$ and transition probabilities given by $P\{X_n = j | X_{n-1} = i\} = P_{i,j}$ where $0 \leq i, j < M$, and initial probability $P\{X_0 = j\} = \tau_j$ and parameters $\theta = (\tau, P)$. Also define the statistics

$$\begin{aligned} N_j &= \delta(X_0 - j) \\ K_{i,j} &= \sum_{n=1}^{N} \delta(X_n - j)\delta(X_{n-1} - i) \ . \end{aligned}$$

   a) Use these statistics to derive an expression for the probability $p(x)$.

   b) Show that the Markov chain has an exponential distribution, and that N_j and $K_{i,j}$ are the natural sufficient statistics of the distribution.

   c) Derive the ML estimate of the parameters τ_j and $P_{i,j}$ in terms of the natural sufficient statistics.

Bibliography

[1] S. M. Ross. *Stochastic Processes*. John Wiley & Sons, New York, 1983.