Chapter 1

Markov Chains and Hidden Markov Models

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.

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 x_k and all n
>
> $$P\{X_n = x_n | X_k = x_k \text{ for all } k < n\} = P\{X_n = x_n | X_{n-1} = x_{k-1}\} \ .$$

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

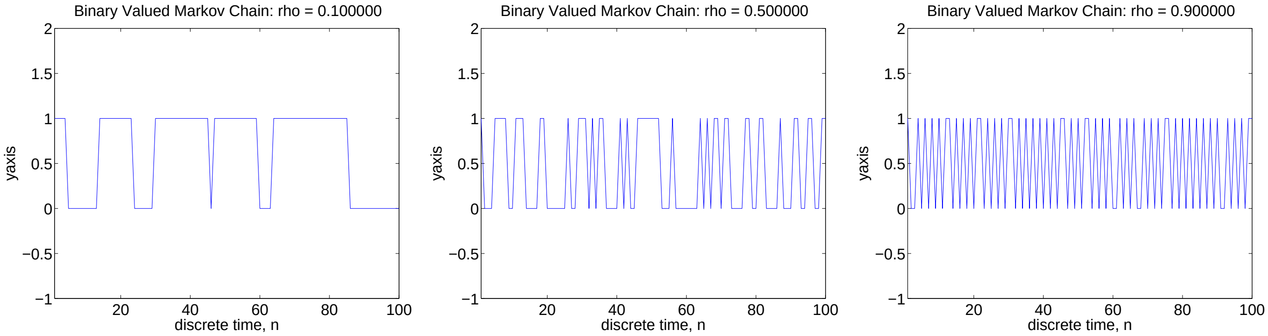

Figure 1.1: Three figures illustrating three distinct behaviors of the Markov chain example with $\rho = 0.1$, 0.5, and 0.9.

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

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

for $i, j \in \Omega$. If the transition probabilities, $P_{i,j}^{(n)}$, do not depend on time, n, then we say that the Markov chain is homogeneous. Note, that a homogeneous Markov chain may have time varying distribution because of the transients associated with an initial condition, but we might hope that a homogeneous Markov chain will reach a stationary distribution given a long enough time. We will discuss this later in more detail. For the remainder of this chapter, we will assume that Markov chains are homogeneous unless otherwise stated.

From the Markov process, we can derive an expression for the probability of the sequence $\{X_n\}_{n=0}^N$. The probability of the sequence is the product of the probability of the initial state, denoted by ρ_{x_0}, with each of the N transitions from state x_{n-1} to state x_n. This product is given by

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

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

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

$$\begin{aligned} \rho_j &= 1/2 \\ P_{i,j} &= \left\{ \begin{array}{cl} 1-\rho & \text{if } j = i \\ \rho & \text{if } j \neq i \end{array} \right. \end{aligned}$$

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 Markov chain is likely to stay in the same state for a long time. When $\rho = 1/2$ each new state will be independent of the previous state; and when ρ is approximate 1, then with almost every value of n, the state is likely to change. These three cases are illustrated in Figure 1.1. If we can 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, but even homogeneous Markov chains can have many parameters. In general, an $M = |\Omega|$ state Markov chain has a total of M^2 parameters. [1] When M is large, 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 $\theta = [\rho_i, P_{i,j} \ : \ \text{for } i, j \in \Omega]$, and define the statistics

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

[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 label X_n.

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

$$p(x|x_0) = \prod_{i\in\Omega}\prod_{j\in\Omega} [\rho_i]^{\tau_i}[P_{i,j}]^{K_{i,j}}$$

which means that the log likelihood has the form

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

Based on this, it is easy to see that X_n has an exponential distribution and that τ_i and $K_{i,j}$ are its nature 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

$$\begin{aligned}\hat{\rho}_i &= \tau_i \\ \hat{P}_{i,j} &= \frac{K_{i,j}}{\sum_{j\in\Omega} K_{i,j}} \ .\end{aligned}$$

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

1.3 Hidden Markov Models

One important application of Markov chains is in hidden Markov models (HMM). Figure 1.2 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 observations Y_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 change in behavior of Y_n over long time scales. The HMM model is sometimes referred to as *doubly-stochastic* because the unobserved stochastic process X_n controls the observed stochastic process Y_n.

Put some more discussion of applications here*

Let the density function for Y_n given X_n be given by

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

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

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

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

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

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 MAP State Sequence Estimation for HMMs

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

$$\hat{x} = \arg\max_{x \in \Omega^N} p(y, x|\theta) \ . \tag{1.2}$$

Interestingly, the optimization of (1.2) can be efficiently computed using *dynamic programming*. To do this, we first define the quantity $L(k, n)$ to be the log probability of the state sequence that results in the largest probability and starts with the value $X_n = k$. The values of $L(k, n)$ can be computed with the recursion

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

[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.

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

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

1.3.2 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 a multivariate Gaussian random vector with distribution $N(\mu_{x_n}, R_{x_n})$, so it is also necessary to estimate the parameters μ_k and R_k for each of the states $k\in\Omega$.

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

$$\begin{aligned}
t_{1,k} &= \sum_{n=0}^{N-1} Y_n\delta(x_n-k) \\
t_{2,k} &= \sum_{n=0}^{N-1} Y_nY_n^t\delta(x_n-k) \\
N_k &= \sum_{n=0}^{N-1} \delta(x_n-k) \\
\tau_k &= \delta(X_0-k) \\
K_{i,j} &= \sum_{n=1}^{N} \delta(X_n-j)\delta(X_{n-1}-i)\ .
\end{aligned}$$

The ML estimate of θ can then be calculated from the sufficient statistics as

$$\begin{aligned}
\hat{\mu}_k &= \frac{t_{1,k}}{N_k} \\
\hat{R}_k &= \frac{t_{2,k}}{N_k}-\frac{t_{1,k}\,t_{1,k}^t}{N_k^2} \\
\hat{\rho}_i &= \tau_i \\
\hat{P}_{i,j} &= \frac{K_{i,j}}{\sum_{j\in\Omega}K_{i,j}}\ .
\end{aligned}$$

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

$$\begin{aligned}
\bar{t}_{1,k} &= \sum_{n=0}^{N-1} Y_n P\{X_n = k|Y = y, \theta\} \\
\bar{t}_{2,k} &= \sum_{n=0}^{N-1} Y_n Y_n^t P\{X_n = k|Y = y, \theta\} \\
\bar{N}_k &= \sum_{n=0}^{N-1} P\{X_n = k|Y = y, \theta\} \\
\bar{\tau}_k &= P\{X_n = k|Y = y, \theta\} \\
\bar{K}_{i,j} &= \sum_{n=1}^{N} P\{X_n = j, X_{n-1} = i|Y = y, \theta\} \ ,
\end{aligned}$$

and the HMM model parameters are then updated using the M-step

$$\begin{aligned}
\hat{\mu}_k &\leftarrow \frac{t_{1,k}}{N_k} \\
\hat{R}_k &\leftarrow \frac{t_{2,k}}{N_k} - \frac{t_{1,k}\, t_{1,k}^t}{N_k^2} \\
\hat{\rho}_i &\leftarrow \tau_i \\
\hat{P}_{i,j} &\leftarrow \frac{K_{i,j}}{\sum_{j\in\Omega} K_{i,j}} \ .
\end{aligned}$$

While the M-step for the HMM is easily computed, the E-step requires the computation of the posterior probability of X_n given Y. However, this is not easily computed using Bayes rule due to the time dependencies of the Markov chain. Fortunately, there is a computationally efficient method for computing these posterior probabilities that exploits the 1D structure of the HMM. The algorithms for doing this are known as the *forward-backward* algorithms due to there forward and backward recursion structure in time.

The forward recursion is given by

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

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

$$\begin{aligned} P\{X_n = k | Y = y, \theta\} &= \frac{\alpha_n(i)\beta_n(i)}{p(y)} \\ P\{X_n = j, X_{n-1} = i | Y = y, \theta\} &= \frac{\alpha_{n-1}(i) p(y_n|j) P_{i,j} \beta_n(i)}{p(y)} \end{aligned}$$

where

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