

Lecture 34: Markov Chains and Their Estimation

Charles A. Bouman

Purdue University

Friday, November 13, 2026

Outline

- 1 What Markov chains model
- 2 Definition of a Markov chain; the state space Ω
- 3 Marginal distribution $\pi_j^{(n)}$ and transition probabilities $P_{ij}^{(n)}$; homogeneous chains
- 4 Homogeneous does not mean stationary
- 5 The sequence probability $p(x) = \tau_{x_0} \prod_{n=1}^N P_{x_{n-1}, x_n}$
- 6 Example 13.1: the binary chain with flip probability ρ
- 7 Parameter estimation: the statistics N_j and $K_{i,j}$
- 8 Markov chains are an exponential family; the ML estimates

Core idea

A Markov chain is built on the simple idea that **each new state depends only on the previous state**. That assumption gives a closed-form sequence probability, and because the distribution is an exponential family, ML parameter estimation reduces to counting transitions.

Today's Big Ideas

- A Markov chain is a sequence of discrete states in which the distribution of each new state depends only on the previous state.
- A chain is homogeneous when its transition probabilities do not change with time, but a homogeneous chain can still have a marginal distribution that changes with time.
- The Markov property makes the probability of a whole state sequence the product of the probability of the initial state with one transition probability for each step.
- In the binary example, the probability of the sequence depends on the data only through the number of times the chain changes state.
- The distribution of a Markov chain is an exponential family, so maximum likelihood estimation reduces to counting transitions and normalizing each row of the transition matrix.

What Markov chains model

It is often useful to model the discrete state of a real-world system as it evolves in time. So the discrete-time random process $\{X_n\}_{n=0}^N$, which takes values on a countable set Ω , might be used to represent many things:

- for a **speech signal**, each state could represent the particular phoneme being voiced;
- for **music**, it could represent the specific note of a discrete musical scale;
- for a **digitally modulated signal**, it could represent the discrete binary values zero and one.

Markov chains represent data as a sequence of states in time, and they are based on the simple idea that each new state depends only on the previous state. This simple assumption makes them easy to analyze but still allows them to be very powerful tools for modeling a wide variety of physical processes.

At the end of the chapter we will ask what happens as time tends toward infinity: when the transient behavior decays and the chain reaches steady state, we call the Markov chain **ergodic**.

Definition of a Markov chain

While the behavior of physical systems can never be fully characterized by a stochastic model, it is often useful to approximate the distribution of each new state, X_n , as depending only on the previous state, X_{n-1} . In this case, we call the random process a Markov chain.

Definition: Markov Chain

Let $\{X_n\}_{n=0}^{\infty}$ be a discrete-time random process taking values in the countable set Ω . Then we say that X_n is a **Markov chain** if for all $j \in \Omega$ and for all $n > 0$,

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

where Ω is the **state space** of the Markov chain.

Conventions for the state space

- Since the state space is discrete and countable, we can assume, without loss of generality, that $\Omega = \{0, \dots, M-1\}$, where $M = \infty$ is allowed.
- We also allow for the possibility that the Markov chain is of finite length, so that it consists of the sequence $X_0, \dots, X_N$.
- In the case where Ω is continuously valued, then we refer to X_n as a **discrete-time Markov process**.
- However, we will focus primarily on Markov chains because they are easy to analyze and of great practical value. Nonetheless, most of the results we derive are also true for discrete-time Markov processes.

Notation: marginals, transitions, and homogeneity

In order to analyze a Markov chain, we need to first define some basic notation. The marginal distribution, $\pi_j^{(n)}$, and **transition probabilities**, $P_{i,j}^{(n)}$, of the Markov chain are defined 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.

Homogeneous does not mean stationary

One might initially guess that a homogeneous Markov chain will have a stationary distribution, so that, for example, its marginal distribution will not vary with time.

Why that guess is wrong

This is not true, since the Markov chain can have transient behavior depending on the initial distribution of states at $n = 0$.

Nonetheless, it is reasonable to expect that after a sufficient time, any transient behavior would die out, and the Markov chain might reach some type of steady-state behavior.

In fact, Section 13.4 will treat this issue in detail.

The probability of a state sequence

From the parameters of the Markov process, we can derive an expression for the probability of the sequence $\{X_n\}_{n=0}^N$. In order to simplify notation, we denote the distribution of the initial state by

$$\tau_j = \pi_j^{(0)} = P\{X_0 = j\} .$$

Then the probability of a particular state sequence for the Markov chain is given by the product of the probability of the initial state, τ_{x_0} , with the probability for 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 13.1: a binary Markov chain

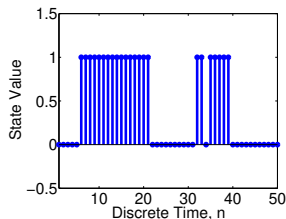
Let $\{X_n\}_{n=0}^N$ be a Markov chain with state space $\Omega = \{0, 1\}$ and parameters given by

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

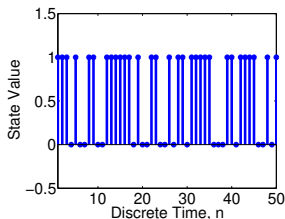
This Markov chain starts with an equal chance of being zero or one. Then with each new state, it has probability ρ of changing states and probability $1 - \rho$ of remaining in the same state.

- If ρ is small, then the probability of changing states 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.
- When ρ is approximately one, then the state is likely to change with each new value of n .

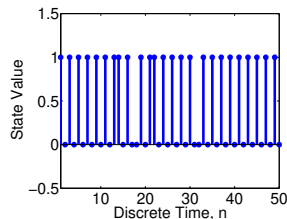
Example 13.1: three distinct behaviors



$$\rho = 0.1$$



$$\rho = 0.5$$



$$\rho = 0.9$$

Illustrations of 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 when $\rho = 0.9$, the state changes value 90% of the time.

Example 13.1: counting the state changes

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

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

So for this chain, the probability of the whole sequence depends on the data only through the single number K .

Section 13.2: why parameter estimation is the challenge

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 that Markov chains are homogeneous.
- 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 of estimating these parameters.

¹The 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 from the fact that there are M rows in the transition matrix and each row has $M - 1$ degrees of freedom since it must sum to one.

The statistics N_j and $K_{i,j}$

Fortunately, we will see that Markov chains are exponential distributions, so parameters are easily estimated from natural sufficient statistics (Section 12.7.1).

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 state j .
- The statistic N_j counts the number of times the initial state has value j .

The sequence probability and the log likelihood

Using these statistics, we can express the probability of the Markov chain sequence as

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

which means that the log likelihood has the form²

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

²This expression assumes that the parameters are nonzero, but the results for the ML parameter estimates still hold in the general case.

Markov chains are an exponential family

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 Lecture 33 (Section 12.7.1)

An exponential family has $\log p(y|\theta) = \langle \eta(\theta), T(y) \rangle + d(\theta) + s(y)$, with $T(y)$ the natural sufficient statistic. Here the log likelihood is a sum of $\log \tau_j$ and $\log P_{i,j}$ weighted by N_j and $K_{i,j}$, which is exactly that form.

From the log likelihood, we can derive the ML estimates of the parameter θ in much the same manner as is done for the ML estimates of the parameters of a Bernoulli sequence.

The ML parameter estimates

This results in the ML estimates

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

What the answer says

So the ML estimate of transition parameters is quite reasonable. It simply counts the rate at which a particular transition occurs from state i to state j and then normalizes this quantity so it sums to one for each initial state i .

Summary

- A Markov chain satisfies $P\{X_n = j | X_k \text{ for all } k < n\} = P\{X_n = j | X_{n-1}\}$ on a countable state space $\Omega = \{0, \dots, M-1\}$.
- Notation: $\pi_j^{(n)} = P\{X_n = j\}$, $P_{i,j}^{(n)} = P\{X_n = j | X_{n-1} = i\}$, and $\tau_j = \pi_j^{(0)}$.
Homogeneous means $P_{i,j}^{(n)}$ does not depend on n ; it does not imply a stationary marginal.
- Sequence probability: $p(x) = \tau_{x_0} \prod_{n=1}^N P_{x_{n-1}, x_n}$, so the chain is parameterized by τ_j and $P_{i,j}$.
- Example 13.1: binary chain with flip probability ρ , and $p(x) = (1/2)(1 - \rho)^{N-K} \rho^K$ where K counts state changes.
- Estimation: M^2 parameters for a homogeneous chain; natural sufficient statistics N_j and $K_{i,j}$; log likelihood $\sum_j \{N_j \log \tau_j + \sum_i K_{i,j} \log P_{i,j}\}$.
- ML estimates: $\hat{\tau}_j = N_j$ and $\hat{P}_{i,j} = K_{i,j} / \sum_j K_{i,j}$ — count transitions and normalize each row.

Reading: Sections 13.1–13.2. Next lecture: Section 13.3, Hidden Markov Models.