Markov Chains

E.H.F.

May 11, 2004

1 Introduction

We are interested in Markov chains and the related problem of sampling an equilibrium distribution. We begin with some definitions and the mathematical properties of Markov chains, and then we discuss the famous Metropolis algorithm for sampling an equilibrium distribution. This is the algorithm used to perform Monte Carlo simulations.

2 The Markov Property

2.1 Definitions

Let Γ denote a state space, or the set of all states of a system. We use notation such as x and y to denote states in Γ. We will be interested in situations in which Γ is the set of all states of a statistical mechanical system, such as the 2^N states of an Ising model with N spins. The following discussion of Markov chains is rigorously true when Γ is a countable set, or when the set of all states can be put into one to one correspondence with the positive integers.[1] To begin, consider a more simple example where Γ represents the states of a dice:

$$\Gamma = \{1, 2, 3, 4, 5, 6\}. \tag{1}$$

It is possible to talk about the probability of each state, and it is reasonable to assign $P(x) = 1/6$ for all $x \in \Gamma$ to model the roll of a dice. We can also

[1] For example, the set of all real numbers in the interval $[0, 1]$ is not a countable set, but the set of all integers, not just positive integers, is a countable set.

talk about random variables on Γ that take on one of the six values. We will use this simple example to illustrate Markov chains.

A Markov chain $X_0, X_1, \ldots, X_n, \ldots$ is a sequence of random variables X_i on a state space Γ with the property that

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n) \quad (2)$$

where $P(X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n)$ is the condition probability of being in state x_{n+1} at step $n+1$ when the chain was in previously in states $x_n, x_{n-1}, \ldots, x_0$. Eqn. 2 is known as the Markov property, which implies that the $n+1$ state of the chain depends only on the n state of the chain. At any n state of the chain, the previous history is irrelevant in determining the next step in the chain. We define

$$p(x, y) \equiv P(X_{n+1} = x | X_n = y) \quad (3)$$

as the transition probability, and we assume that this relationship holds for all n. For this quantity to make sense as a transition *probability*, we must have that

$$\sum_{x \in \Gamma} p(x, y) = 1. \quad (4)$$

All sums over x or y are over states in Γ, so we will no longer write the Γ in the summation. For the simple example of the dice, $p(x, y)$ is a 6×6 matrix, and we can set $p(x, y) = 1/6$ for $x, y \in \Gamma$. For any state in this Markov chain, there is an equal probability to make a transition to each state of the system. We define

$$p^n(x, y) \equiv P(X_n = x | X_0 = y) \quad (5)$$

as the probability of being in state x n steps after being in state y. In our notation, the later state is to the left, and the earlier state is to the right.

2.2 Mathematical Properties

We introduce the notation that $b(x)$ is a vector indexed by the state space Γ; a vector $b(x)$ corresponds to a probability distribution if $\sum_x b(x) = 1$. Then Eqn. 4 is equivalent to saying that the vector $b(x) = 1$ for all $x \in \Gamma$ is a left eigenvector of $p(x, y)$ with eigenvalue 1,

$$\sum_x b(x) p(x, y) = b(y). \quad (6)$$

Note that $p(x, y)$ is not necessarily symmetric, so the left eigenvectors are not necessarily the same as the right eigenvectors. What can we say about the right eigenvector and eigenvalues? Consider the equation

$$\sum_y p(x, y)b(y) = \lambda b(x). \tag{7}$$

While the entries of $p(x, y)$ are real valued and non-negative, λ and $b(x)$ can in general be complex. Summing over x on both sides of the previous equation implies

$$\sum_y b(y) = \lambda \sum_x b(x) \tag{8}$$

so then either $\lambda = 1$ or $\sum_x b(x) = 0$. This implies that if there exists a probability distribution $b^*(x)$ that satisfies Eqn. 7, then the corresponding eigenvalue is 1. If there exists an eigenvector $b(x)$ such that the corresponding eigenvalue is not 1, then $\sum_x b(x) = 0$.

For an arbitrary Markov chain, it is possible to prove the following:

Theorem 1 *For a Markov chain,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} p^m(x, y) \tag{9}$$

always exists.

Since $p^m(x, y)$ is the probability of being in state x at the m step, then $n^{-1} \sum_{m=1}^{n} p^m(x, y)$ is the probability of being in state x averaged over n consecutive steps of the Markov chain. Theorem 1 implies that the *average* probability of being in a state converges as the Markov chain becomes arbitrarily long. This average number of visits is independent of the starting state y. Theorem 1 is a result of the law of large numbers.

Now we make a few more mild assumptions about the transition probability $p(x, y)$. First, we assume that for any $x, y \in \Gamma$ there exists sequence of states such that one can get from x to y with nonzero probability. A Markov chain with this property is said to be *irreducible*. Our Markov chain for the dice satisfies this property. Second, we assume that if we start in any state x, then the average number of steps it takes to return to x for the first time is finite. If this is true for all $x \in \Gamma$, then the Markov chain is said to be *positive recurrent*. For the dice Markov chain, one can show that the mean return time for any state is 6/25, so the Markov chain is positive recurrent.

Theorem 2 *For an irreducible, positive recurrent Markov chain*[2]*, there exists a unique right eigenvector* $b(x)$ *with eigenvalue 1 such that* $\sum_x b(x) \neq 0$.

This implies that

$$\sum_y p(x,y)b(y) = b(x). \tag{10}$$

This eigenvector equation remains unchanged if we multiple by a constant, so we divide by a normalization constant $\sum_x b(x)$. Then $b^*(x) \equiv b(x)/\sum_y b(y)$ is a probability distribution that is a right eigenvector of $p(x,y)$ with eigenvalue 1. We call $b^*(x)$ the stationary distribution, since the action of the transition probability on this distribution, or $\sum_y p(x,y)b^*(y)$, simply gives the same distribution.

With one addition assumption, it is possible prove a stronger result. For any state x, let $\mathcal{I}_x$ be the set of positive integers r such that there is a nonzero probability that $X_r = x$ if $X_0 = x$. Let d_x be the greatest common denominator of the set $\mathcal{I}_x$. If $d_x = 1$ for all $x \in \Gamma$, then the Markov chain is *aperiodic*. Note that $d_x = 1$ if $p(x,x) > 0$. It is possible to prove the following:

Theorem 3 *If a Markov chain is irreducible, positive recurrent, and aperiodic, then* $\lim_{n\to\infty} p^n(x,y) = b^*(x)$.

The probability of being in state x goes to $b^*(x)$ the stationary distribution in the limit of large n, not just the *average* probability of being in state x. This is an amazing result. For any initial probability distribution, the Markov chain always eventually reaches the stationary distribution. For our dice example, assume we start the Markov chain in state $X_0 = 1$. So then $P(X_0 = 1) = 1$ and $P(X_0 = y) = 0$ for all $y(\neq 1) \in \Gamma$. The theorem states that the Markov chain will eventually reach the stationary distribution, or $b^*(x) = 1/6$ for all $x \in \{1 \ldots 6\}$.

3 Sampling an Equilibrium Distribution

Up until this point, we have assumed that $p(x,y)$ is given and explored the properties of the resulting Markov chain. But in statistical mechanics, we are

[2]For an irreducible Markov chain, all the states are either positive recurrent or null recurrent, meaning that the mean return time to a state is infinite. The probability of a state $b^*(x)$ is the inverse of the mean return time, so then $b^*(x) = 0$ for all $x \in \Gamma$ for a null recurrent Markov chain so $b^*(x)$ can't be a distribution.

given a distribution $b^*(x)$ and would like to sample from this distribution. This is equivalent to choosing state $x \in \Gamma$ with probability $b^*(x)$. We can use a Markov chain to generate $b^*(x)$ in the limit of large n if we can construct the appropriate $p(x, y)$.

3.1 The Metropolis Algorithm

To do this, we construct a transition probability that satisfies $\sum_x p(x, y) = 1$ and the detailed balance condition:

$$p(x, y)b^*(y) = p(y, x)b^*(x). \tag{11}$$

If we can construct such a $p(x, y)$, then it immediately follows that

$$\sum_x p(y, x)b^*(x) = b^*(y) \tag{12}$$

by summing both sides of Eqn. 11 over x. A transition probability $p(x, y)$ that satisfies the detailed balance condition has $b^*(x)$ as its stationary distribution. Then a Markov chain governed by such a transition probability will sample the equilibrium distribution in the limit of large n.

In the 50's sometime, researchers at Los Alamos came up with a very clever way of constructing $p(x, y)$ that satisfies the detailed balance condition. Let $t(x, y)$ be a symmetric matrix such that $\sum_x t(x, y) = 1$; we call this the *trial probability*. Then

$$p(x, y) = t(x, y)\min\left(1, \frac{b^*(x)}{b^*(y)}\right) \tag{13}$$

where $\min(a, b)$ is the minimum of a and b satisfies the detailed balance condition. To prove this, note that for any constants a, b and c the relation $a \cdot \min(b, c) = \min(ab, ac)$ holds. Then inserting $p(x, y)$ in Eqn. 13 into the detailed balance condition in Eqn. 11 gives

$$t(x, y)\min\Big(b^*(y), b^*(x)\Big) = t(y, x)\min\Big(b^*(x), b^*(y)\Big) \tag{14}$$

which is true because of the symmetry of $t(y, x)$ and the min function. The transition probability in Eqn. 13 implies the following algorithm for generating a Markov chain:

- If we are in a state $y \in \Gamma$, then pick a state $x \in \Gamma$ from the trial probability $t(x, y)$.
- If $b^*(x) \geq b^*(y)$, accept the move.
- If $b^*(x) < b^*(y)$, accept the move with probability $b^*(x)/b^*(y)$.

This algorithm is particularly useful in sampling distributions in statistical mechanics, as we show in the next section.

3.2 Application to the Ising model

To illustrate the application of the Metropolis algorithm to statistical mechanics, consider a d dimensional nearest neighbor Ising model. If $s_i \in \{1, -1\}$ denotes the two states of a single spin, then $\Gamma = \{s_1, s_2, \ldots, s_N\}$ so there are 2^N possible states of the system. The Hamiltonian for such as system is

$$\mathcal{H}(\Gamma) = \sum_{i \neq j,} J_{ij} s_i s_j + H \sum_i s_i. \tag{15}$$

where $J_{ij} = J$ when i, j are nearest neighbor sites and $J_{ij} = 0$ otherwise. The probability of state Γ in the canonical ensemble is

$$P(\Gamma) = Q^{-1} \exp\Big(-\mathcal{H}(\Gamma)/kT\Big) \tag{16}$$

where Q is the canonical partition function

$$Q = \sum_\Gamma \exp\Big(-\mathcal{H}(\Gamma)/kT\Big). \tag{17}$$

We have some freedom in choosing the trial probability $t(x, y)$ as long as this matrix is symmetric. A simple trial probability considers changing the state of one spin at a time; this corresponds to a $t(x, y)$ that has a value of N^{-1} for states x, y that differ by the state of one spin and $t(x, y) = 0$ otherwise. For $N = 3$, we allow trial moves such as

$$\{1, -1, -1\} \longrightarrow \{1, -1, 1\} \tag{18}$$

and we pick each spin with equal probability. With this choice of $t(x, y)$, we obtain the following algorithm:

- For a state Γ, pick a spin at random to produce a new state Γ'.

- If $\mathcal{H}(\Gamma') \leq \mathcal{H}(\Gamma)$, accept the move.
- If $\mathcal{H}(\Gamma') > \mathcal{H}(\Gamma)$, pick a uniformly distributed random number Z from the interval $[0, 1]$. If $Z < \exp[-(\mathcal{H}(\Gamma') - \mathcal{H}(\Gamma))]$, then accept the move. Otherwise, reject the move.

If we make a trial move and the energy decreases, then accept the move. If the energy increases, accept the move with a certain probability that is related to the energy difference between the two states. For a larger energy increase from the old to the new state, there is a lower probability of accepting the move. For the nearest neighbor Ising model, this energy difference can be evaluated from the states of the $2d$ spins that are nearest neighbors to the spin that was flipped to create Γ'. The cleverness in this algorithm is that we never have to calculate the partition function Q.

4 Appendix

Don't read this. Haven't figured out why this might be important. Let's determine an upper bound on the magnitude of the eigenvalue λ. The relationship

$$\left| \sum_y p(x, y) b(y) \right| \leq \sum_y p(x, y) |b(y)| \tag{19}$$

implies that

$$\sum_y p(x, y) |b(y)| \geq |\lambda| |b(x)|. \tag{20}$$

We sum over x on both sides of the equation and obtain

$$\sum_y |b(y)| \geq |\lambda| \sum_x |b(x)| \tag{21}$$

by Eqn. 4. This implies that either $|\lambda| \leq 1$ or $\sum_x |b(x)| = 0$. The fact that $p(x, y)$ has an interpretation as a transition probability implies that the magnitude of all right eigenvalues must be less than or equal to unity or the corresponding eigenvector doesn't correspond to a probability distribution.