Markov Chain Monte Carlo Notes

Jesse Young Lin

February 13, 2025

Contents

1 Preliminaries on Probability

Probability is extremely unintuitive for humans. Consequently, there is a variety of vocabulary and notation which is initially confusing but essential to understand. We will illustrate all the preliminaries with the example of a fair coin flip (which is formally known as a Bernoulli distributed random variable).

1.1 Probability distributions

A **random variable** is typically denoted with a capital letter, and its **realization** is often denoted with the lowercase letter. To model a series of coin flips, we consider a sequence of N random variables $\{X_i\}_{i=0}^N$. Say we flip the $i = 0$ coin and see heads. Then, we say that $x_0 = H$.

The **probability distribution** or **probability density function** of X_0 is given by the Bernoulli distribution, which is a function ρ

$$\rho(x) = \begin{cases} 1/2, & x = H \\ 1/2, & x = T \end{cases} .$$

This communicates the intuitive fact that if we flip a fair coin a large number of times, we expect it to land on heads half the time and tails the other half. We note that $\rho(x)$ is just a function over the set $\{H, T\}$, its probabilistic meaning is given by the notation

$$\mathbb{P}(X_0 = z) = \rho(z).$$

We use the $\mathbb{P}$ notation to describe the probability of particular realizations. For example, if $z = H$, then the above equation says that the probability the random variable X_0 has value H is given by the value of the function $\rho(z)$ at $z = H$, which is 1/2.

1.2 Conditional probability and (in)dependence of random variables

In our sequence of coin flips, we have considered the first coin X_0. What about the X_1 variable? Now, we need to introduce the assumption that X_1 is **independent** of X_0. Intuitively, this means that the realization (i.e., the result of the coin flip) x_1 does not depend on information from the realization x_0. Formally, we write

$$\mathbb{P}(X_1|X_0) = \mathbb{P}(X_1). \tag{1}$$

The left-hand side of this equation is the **conditional probability**. It is read as the conditional probability of X_1 conditioned on the realization of X_0. In words, what is the probability which describes X_1 given that we know the realization of X_0? As we are modelling a sequence of coin flips, we know that the probability distribution of the $i = 1$ should have nothing to do with whether the coin flip at $i = 0$ was heads or tails, which is what (1) tells us. Finally, we have been assuming that all the X_i are identically distributed, i.e., they are all Bernoulli distributed, which tells us that

$$\mathbb{P}(X_i = z) = \rho(z)$$

for any i. Therefore we have our final model of a sequence of coin flips, which is given by the sequence $\{X_i\}$ of independently and identically distributed random variables, all distributed according to the Bernoulli distribution.[1]

[1] We note also that independence is often defined via the following equation on the **joint probability**

$$\mathbb{P}(X_1 = x \text{ and } X_0 = y) = \mathbb{P}(X_1 = x)\mathbb{P}(X_0 = y).$$

This is equivalent to (1) if we use **Bayes' theorem**, which is the following statement which always holds

$$\mathbb{P}(X_1 = x \text{ and } X_0 = y) = \mathbb{P}(X_1 = x \mid X_0 = y)\mathbb{P}(X_0 = y).$$

If we denote using A and B the realizations

$$A = \{X_1 = x\}$$
$$B = \{X_0 = y\}$$

and use the set theoretic notation for "and", we get the common expression

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B).$$

This equation says that the probability of events A and B occurring simultaneously is equal to the probability that B occurs multiplied by the probability that A occurs given that we know event B occurs (the conditional probability).

A simple example of dependent random variables is the following. Imagine an urn filled with one red ball and one black ball. Let Y_i be the random variable corresponding to the i-th draw from the urn. Now,

$$\mathbb{P}(Y_1 = \text{red} \mid Y_0 = \text{red}) = 0$$
$$\mathbb{P}(Y_1 = \text{black} \mid Y_0 = \text{red}) = 1,$$

in other words if you first draw red then you know the next draw must be black. However

$$\mathbb{P}(Y_1 = \text{red}) = 1/2$$
$$\mathbb{P}(Y_1 = \text{black}) = 1/2.$$

which means if you just consider drawing from the urn twice, the second draw has a uniform probability of being either the red or the black one. This violates the equation (1).

1.3 Expectation value

The final concept is the most intuitive one. In the coin flip example we denoted the heads and tails by symbols $\{H, T\}$. Let's imagine a game where everytime you flip heads you gain 1 dollar and everytime you flip tails you lose 1 dollar. This is equivalent to assigning numbers $H \to 1$ and $T \to -1$. Let the sequence $\{Z_i\}$ of random variables correspond to the payoff at each step i of this game. The average payoff at any step i is evidently 0. Formally this is denoted with the **expectation value** defined as follows

$$\mathbb{E}(Z_i) = \sum_{z \in \{\pm 1\}} z \mathbb{P}(Z_i = z)$$

which is easy to compute:

$$\begin{aligned} \sum_{z \in \{\pm 1\}} z \mathbb{P}(Z_i = z) &= \sum_{z \in \{\pm 1\}} z \rho(z) \\ &= (1)(1/2) + (-1)(1/2) \\ &= 0. \end{aligned}$$

2 Markov Chains

2.1 Definition

A Markov chain is a sequence of random variables which is **memoryless**. In other words, for a sequence $\{X_i\}$

$$\mathbb{P}(X_j \mid X_{j-1}, X_{j-2}, \ldots, X_0) = \mathbb{P}(X_j \mid X_{j-1}). \tag{2}$$

Intuitively, the value of random variable X at time j depends only on its value at the previous time $j-1$ and it has no memory of the history before that. Equation (2) is called the **Markov property**. Often we define the **Markov transition matrix**

$$W(x, y) = \mathbb{P}(X_j = x \mid X_{j-1} = y).$$

The essential feature of Markovian systems is the ability to predict the future given an initial condition by repeated application of W, i.e.,

$$\mathbb{P}(X_n = x_n \mid X_0 = x_0) = \sum_{\{x_{n-1},\ldots,x_1\}} W(x_n, x_{n-1}) \cdots W(x_1, x_0)\mathbb{P}(X_0 = x_0). \tag{3}$$

A derivation is given in [2]. The above when applied to quantum mechanics is actually the celebrated Feynman path integral.

2.2 Equilibrium

If the values x_i in (3) are taken to assume only finitely many values, then we can write the equivalent matrix-vector equation

$$P_t = W^t P_0$$

where P_t is the vector of probabilities at time t, and W is a matrix. The superscript t represents the repeated multplication of the W.

The **equilibrium** or **invariant distribution** of a Markov chain is the probability distribution P which satisfies

$$P = WP. \tag{4}$$

In linear algebra, this condition (4) is known as an **eigenvalue equation**. One approach to solving the above is to consider the components

$$P_i = \sum_j W_{ij} P_j$$

[2] Bayes' theorem allows the following decomposition of the joint probabilities

$$\mathbb{P}(X_n, \ldots, X_0) = \sum_{\{x_{n-1},\ldots,x_0\}} \mathbb{P}(X_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)\mathbb{P}(X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)$$

where the sum is taken over all possible values of the $\{x_{n-1}, \ldots, x_0\}$, e.g., if each X_i models a coin flip then

$$\sum_{\{x_{n-1},\ldots,x_0\}} = \sum_{x_{n-1}\in\{H,T\}} \cdots \sum_{x_0\in\{H,T\}}.$$

Then,

$$\begin{aligned}
\mathbb{P}(X_n = x_n \mid X_0 = x_0) &= \sum \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)\mathbb{P}(X_{n-1} = x_{n-1}, \ldots, X_0 - x_0) \\
&= \sum \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1})\mathbb{P}(X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\
&= \sum W(x_n, x_{n-1})\mathbb{P}(X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\
&= \sum W(x_n, x_{n-1}) \cdots W(x_1, x_0)\mathbb{P}(X_0 = x_0) \\
&= \sum_{x_0} W^n(x_n, x_0)\mathbb{P}(X_0 = x_0)
\end{aligned}$$

then using the fact that $\sum_j W_{ji} = 1$ (i.e., the transition matrix must conserve probability), this is equivalent to

$$\sum_j W_{ji} P_i = \sum_j W_{ij} P_j$$

and one way to solve this is with a vector P that satisfies, for each component

$$W_{ji} P_i = W_{ij} P_j. \tag{5}$$

The condition (5) is called **detailed balance**. It indicates that, at all times, the rate of transitions between states $i \to j$ is exactly balanced by the rate of transitions $j \to i$. Intuitively, then, the probability P_k of being in any state k must be constant in time.

3 Markov Chain Monte Carlo

The essential idea is now immediate to state: to sample from a complex probability distribution P, it is often easier to design the transition matrix W of a Markov chain such that P is its invariant distribution. Then, independent simulations of the Markov chain can be done on a computer, and given sufficient time one expects that the simulated data obeys P.

Designing a Markov chain is often conceptually simple: for example, to sample from a chemical system in equilibrium the transition matrix is given directly by the kinetic rates and stochiometry of the reactants. Markov chains for most systems also benefit from an **exponential** convergence rate to equilibrium, which means the algorithm is usually quite efficient. [3]

The algorithm we use is the **Metropolis-Hastings** algorithm. It is essentially a specification of the transition matrix W. There are other choices for W, such as the **Gibbs sampler**, but the basic idea is the same.

[3] A notable exception occurs with systems which are at a critical point. This is an extremely rich subject, especially in study of the Ising model. It's out of scope of our project but I encourage looking it up!