MARKOV CHAIN MONTE CARLO: EXPOSITION AND APPLICATIONS

FINAL PROJECT REPORT FOR MATH 697AM

Long Le
Department of Mathematics and Statistics
university of massachusetts Amherst

November 28, 2021

ABSTRACT

This is a final project report for *Math 697AM: Applied Math and Modeling* at UMass Amherst with Prof. *Yao Li*. In this report, we introduce the necessary mathematical background needed for and the property of Markov Chain Monte Carlo (MCMC). We will also explain two MCMC methods: **Metropolis-Hastings** algorithm and **Gibbs Sampler**. We then focus on the **Ising problem** and its two applications in physics and image reconstruction.

***K*eywords** MCMC · Metropolis-Hastings · Gibbs Sampler · Ising model · Image denoising.

1 Preliminary

In this section, we will introduce the theory of a generic Markov chain in discrete time and discrete space.

Definition 1 (Markov property)**.** *A sequence of r.vs $\{X_t\}_{t\in\mathbb{N}}$ taking values in a common state space E is said to have Markov property if and only if*

$$\mathbb{P}[X_{n+1}|X_1,...,X_n] = \mathbb{P}[X_{n+1}|X_n]$$

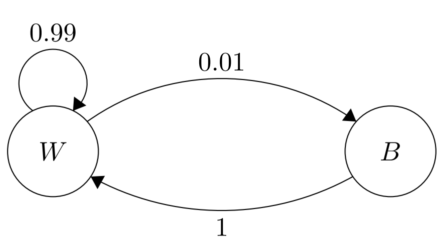
for all $n \in \mathbb{N}$. [1] *In addition, if we have*

$$\mathbb{P}[X_{n+1} = j|X_n = i] = p_{ij}$$

*for all $n \in \mathbb{N}$ then we say that the chain is **time-homogeneous**.*

If the Markov chain $\{X_t\}$ is time homogeneous, then we usually put the probability $\mathbb{P}[X_{n+1} = j|X_n = i] = p_{ij}$ into a transition matrix P and write $X_t \sim Markov(\lambda, P)$ with some initial distribution vector λ.

Example 1 (Machine failure,[Bertsekas and Tsitsiklis(2002)])**.** *A machine can either be working or broken down in any given day. If it is working, it can break down with probability* 0.01*. If it is broken down, it will be repaired and work on the next day with probability* 1*.*



[1]Here we use $\mathbb{P}[X_n]$ to denote $\mathbb{P}[X_n = x]$ for some $x \in E$, whose value we are not particularly interested in.

Assume that we start with a working machine, then the initial distribution is $\lambda = (1, 0)$. The transition matrix is

$$P = \begin{array}{c c} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} 0.99 & 0.01 \\ 1 & 0 \end{pmatrix} \end{array}$$

Definition 2 (Stationary distribution)**.** *A probability distribution π (row vector) is stationary with respect to P if and only if*

$$\pi P = \pi$$

The readers should note these two important fact

1. If our chain starts in stationarity (i.e $\lambda \sim \pi$) then $X_t \sim \pi$ for all $t \in \mathbb{N}$.
2. Under some technical conditions (positively recurrent and aperiodic for our discrete case), the chain has unique stationary π and will converge to π.

Finally, we will introduce the last property that would be important for our MCMC construction.

Definition 3 (Detailed balance)**.** *$X_t \sim Markov(\lambda, P)$ is said to be in detailed balance with distribution μ iff*

$$\mu_i P_{i,j} = \mu_j P_{j,i}.$$

If X_t starts in μ, then we can " run time backwards".

$$\mathbb{P}[X_0 = x_0, ..., X_n = x_n] = \mathbb{P}[X_0 = x_n, ..., X_n = x_0] \quad \textit{(time reversibility)}$$

Lemma 1 (Reversibility implies stationarity)**.** *If P is in detailed balance with μ then μ is the stationary distribution for $X_t \sim Markov(\lambda, P)$.*

Proof.

$$(\lambda P)_i := \sum_{j\in\Omega} \lambda_j P_{ji} = \sum_{j\in\Omega} \lambda_i P_{ij} = \lambda_i (\sum_{j\in\Omega} P_{ij}) = \lambda_i.$$

□

For a more detail treatment of Markov chain, we refer the reader to [Norris(1997)].

2 Markov Chain Monte Carlo (MCMC)

2.1 Problem

Suppose that we want to estimate

$$\int_{\mathbb{R}^d} f(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x}$$

for some function f and density π. If the integral is too hard to solve analytically, which is often the case, then the two general approaches to tackle the problem are

1. Quadrature
2. Monte-Carlo

Unfortunately, numerical integration does not tend to work on high dimensions and it also requires f, π to be smooth. The alternative, Monte Carlo, is the main topic of this report.

In Monte Carlo simulation, we would use the empirical distribution to estimate the underlying density π. Specifically, we Sample $X_1, ..., X_N \overset{i.i.d}{\sim} \pi$ and take the estimate

$$\overline{f}_N = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

Although this estimator is not necessarily unbiased, it is consistent. By the law of large number, we have the following almost surely convergence

$$\mathbb{P}[\lim_{N\to\infty} \overline{f}_N = \mathbb{E}[f(X_i)] = \int f(x)p(x)\,dx] = 1$$

One caveat is that we will often sample from some distribution other than π as we usually only know π up to some normalizing constant Z that is expensive to compute. Indeed, the two algorithms that will be introduced shortly do not compute Z.

There are many Monte Carlo algorithms, including rejection sampling and importance sampling. Markov chain Monte Carlo is another class of algorithms. The idea is to devise a chain with stationary distribution π and then take $X_1, ..., X_n$ samples obtained from simulating the chain. We have two requirement

1. **Correctness**: Chain has to be ergodic with π. That is π has to be the stationary distribution and the chain has to converge to π.
2. **Efficient**: Compute the next state of the chain has to be efficient.

2.2 Error bound of MCMC

In this section, we quickly estimate the variance of the estimator $\overline{f}_N$ obtained from $X_1, ..., X_N$ samples from the chain $X_t \sim Markov(\pi, P)$ starting in stationarity. Recall from basic probability that

$$Cov(X_1 + X_2 + ... + X_n) = \sum_{i,j} Cov(X_i, X_j).$$

where $Cov(X_i, X_i) = X_i$. So

$$Cov(X_1 + X_2 + ... + X_n) = \sum_i Var(X_i) + 2\sum_{i\neq j} Cov(X_i, X_j).$$

We can greatly simplify this expression. But before we proceed, let us introduce some more terminology.

Definition 4 (Weakly (covariance) stationary process)**.** *X_t is weakly stationary (WWS) process if and only if*

1. $\mathbb{E}(x_t) = \mu$ *for all* t
2. $Cov(x_t, x_s) = \sigma(t - s)$
3. $var(x_t) = \sigma(0) < \infty$

Because $\{X_t\} \sim Markov(\pi, P)$, we know that $X_n \sim \pi$ for all $n \in \mathbb{N}$. This is known as strict stationarity.

Lemma 2. $\{X_t\} \sim Markov(\pi, P)$ *is weakly stationary.*

Proof. □

2.3 Mixing time

References

[Bertsekas and Tsitsiklis(2002)] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability*. Athena Scientific books. Athena Scientific, 2002. ISBN 9781886529403. URL `https://books.google.com/books?id=bcHaAAAAMAAJ`.

[Norris(1997)] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, UK, 1997.