Markov Chain Monte Carlo

Heran Song

Introduction

Monte Carlo is a method to learn about probability models by simulating them. If we can simulate a random process, we can calculate it's expectation by averaging these simulations(sample mean). The strong law of large numbers tells us that the sample mean will converge almost surely to the theoretical mean as the number of simulations goes to infinity. Furthermore, if the random process has a finite variance, then the central limit theorem tells us the asymptotic behavior of our simulations. While ordinary Monte Carlo(OMC) can be used in many situations, one major draw back is that it becomes very difficult when the problem we encounter is multivariate. This is where Markov Chain Monte Carlo(MCMC) comes in. Instead of trying to simulate i.i.d random variables, we simulate a markov chain with specific invariant distribution and average over these simulations, hence the name Markov Chain Monte Carlo. In this paper, we will explain the theory behind MCMC, and introduce a specific algorithm called the Metropolis-Hastings algorithm, then give an example of it being applied to a problem.

Markov Chain Theory

Definition 1 Given a set $T \subset \mathbb{R}$ and a measurable space $(\Omega, \mathscr{F})$ a *filtration* $\mathbb{F}$ is a sequence of increasing σ-fields $\mathbb{F} = \{\mathscr{F}_t \ : \ t \in T\}$ such that $\mathscr{F}_s \subset \mathscr{F}_t \subset \mathscr{F} \ \forall s < t$ and $s, t \in T$.

Definition 2 Given a measureable space $(\Omega, \mathscr{F})$ and a filtration $\mathbb{F}$ on it, a *stochastic process* $\{X_t \ : \ t \in T\}$ is a collection of random variables on the common measurable space, where $\forall t \in T$, X_t is $\mathscr{F}_t$-measurable. The particular case where $\mathscr{F}_t = \sigma(X_t)$ is called the *natural filtration* of the process.

Definition 3 Given a discrete time stochastic process $X_0, X_1, X_2, ...$ taking values in an arbitrary state space S, X_t is a *Markov Chain* with respect to a filtration $\mathscr{F}_t$ if for $X_t \in \mathscr{F}_t \ \forall t \in T$ and $\forall B \subset S$, we have the *Markov Property*: $P\{X_{t+1} \in B | \mathscr{F}_t\} = P\{X_{t+1} \in B | X_t\}$.

If $P\{X_{t+1} \in B | X_t\} = P\{X_{s+1} \in B | X_s\} \ \forall s, t \in T$, then we say that the Markov Chain has stationary transition probabilities. From now on, we will only refer to Markov Chains with stationary transition probabilities unless stated otherwise.

In order to specify a Markov Chain model, we need both the initial distribution and the transition probabilities. The initial distribution is the marginal distribution of X_0. The transition probabilities specify the conditional distribution of $X_{t+1} | X_t$.

Note that a Markov Chain having stationary transition probabilities alone does not imply that the Markov Chain itself is stationary, the initial distribution also plays a factor.

Definition 4 A probability distribution is *invariant* for a specific transition probability if the Markov Chain that results from using that distribution as the initial distribution is stationary.

In general state space, transition probabilities are represented by operators operating on infinite dimensional spaces. Although the location of the operators represent different things.

Definition 5 Given a general state space S, the transition probabilities are specified by defining a *transition kernel* $P(x, B) = P\{X_t \in B | X_{t-1} = x\}$, $x \in S$, B is a measurable set in S satisfying:

- $\forall x$ the function $B \mapsto P(x, B)$ is a probability measure.
- $\forall$ fixed B the function $x \mapsto P(x, B)$ is a measurable function.

For left multiplication, if λ is a probability measure on the state space S, and X_{n-1} has distribution λ, then the distribution of X_n is given by $\lambda P(B) = \int \lambda(dx)P(x, B)$.

For right multiplication, if a kernel P has invariant distribution π and $f \in L^p(\pi)$ for some $p \geq 1$, then $Pf(x) = \int P(x, dy)f(y)$ is a well define element of $L^p(\pi)$.

Detailed Balance We say a kernel P is *reversible* with respect to π if $\int_A \pi(dx)P^\dagger(x, B) = \int_A \pi(dx)P(x, B) = \int_B \pi(dx)P(x, A)$, $A, B \in \mathscr{B}$, where $\mathscr{B}$ is the σ-field of S and $P^\dagger$ is the conjugate transpose of P. A condition known as *detailed balance*. Which is what makes MCMC work. When A=S, $\int \pi(dx)P(x, B) = \int_B \pi(dx) = \pi(B)$ or $\pi P = \pi$.

Claim P is reversible w.r.t $\pi \iff$ P is a self-adjoint operator on $L^2(\pi)$.
Proof: " $\longrightarrow$ "
Let $f, g \in L^2(\pi)$ and assume P is reversible,

$$\begin{aligned} < Px, y > &= \int\int \pi(dx)P(x, dy)f(x)g(y) \\ &= \int\int \pi(dx)P(x, dy)f(y)g(x) = \int\int \pi(dx)P^\dagger(x, dy)f(y)g(x) =< x, P^\dagger y > \end{aligned}$$

" $\longleftarrow$ "
Assume P is self-adjoint,

$$\begin{aligned} \implies \int\int \pi(dx)P^\dagger(x, dy)f(y)g(x) &= \int\int \pi(dx)P(x, dy)f(x)g(y) \\ &= \int\int \pi(dx)P(x, dy)g(x)f(y) \qquad \blacksquare \end{aligned}$$

Like how OMC has the SLLN and central limit theorem, there are also similar theorems for MCMC.

Birkhhoff Ergodic Theorem If $X_1, X_2, ...$ is a stationary real valued stochastic process that is ergodic, and $E(X_i) = \mu \, \forall i$, then $\overline{X_n} \to \mu$, as $n \to \infty$.

A stationary Markov Chain $X_1, X_2, ...$ is a stationary stochastic process, but not necessarily real valued. If g is a real valued function on the state space of the Markov Chain, then $g(X_1), g(X_2), ..$ is a stationary real valued stochastic process, known as a functional of the chain.

Functional CLT While expectation remains the same for any random variables, variance on the other hand is different for non i.i.d random variables.

Given a stationary Markov Chain $X_1, X_2, ...,$

$$\begin{aligned} Var(\sum_{i=1}^{n} X_i) &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j) \\ &= nVar(X_i) + 2\sum_{k=1}^{n-1}(n-k)Cov(X_j, X_{j+k}) \\ \implies \sigma_n^2 = nVar(\overline{X_n}) &= \gamma_0 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\gamma_k \quad where \ \gamma_k = Cov(X_j, X_{j+k}) \\ \implies \sigma_{clt}^2 &= \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \end{aligned}$$

It turns out that the last line is only true under the condition of uniform integrability and that $\sum_{k=1}^{\infty} |\gamma_k| < \infty$.

The Metropolis-Hastings Algorithm

With our theory behind Markov Chains, we are ready to introduce the Metropolis-Hastings algorithm.

The MH algorithm preserves any distribution π specified by an unnormalized density h with respect to a measure μ using an auxiliary transition probability specified by a density q(x,y) called the proposal distribution.

To use MH, we need the following:

- $\forall \ x$, we can evaluate $h(x)$.
- $\forall \ x$ and y, we can evaluate $q(x, y)$.
- $\forall \ x$, we can simulate a random variate with density $q(x, \cdot)$ w.r.t μ.

The MH algorithm:

1. Simulate a random variate y having density $q(x, \cdot)$.
2. Calculate $R = \frac{h(y)q(y,x)}{h(x)q(x,y)}$.
3. Set $x = y$ with probability $\min(1, R)$.

Introduce the MH kernel,

$$P(x, A) = r(x)I(x, A) + \int_A q(x, y)a(x, y)\mu(dy)$$

where $r(x) = 1 - \int q(x, y)a(x, y)\mu(dy)$, $a(x, y) = min(1, R)$, $I(x, A)$ is the identity kernel.

Claim A Markov Chain is reversible w.r.t the distribution π having unnormalized density h w.r.t μ under the Metroplis-Hastings algorithm.
Proof:

$$\begin{aligned}
a(x, y) = min(1, R) &= min(1, \frac{h(y)q(y, x)}{h(x)q(x, y)}) \\
R \leq 1 \implies a(x, y) &= \frac{h(y)q(y, x)}{h(x)q(x, y)} \\
R \geq 1 \implies a(x, y) = 1 \implies a(y, x) &= \frac{h(x)q(x, y)}{h(y)q(y, x)} \\
\implies a(x, y)h(x)q(x, y) &= a(y, x)h(y)q(y, x)
\end{aligned}$$

$$\begin{aligned}
\int\int f(x)g(y)\pi(dx)P(x, dy) &= \int\int f(x)g(y)\pi(dx)r(x)I(x, A) + \int\int f(x)g(y)\pi(dx)q(x, y)a(x, y)\mu(dy) \\
&= \int f(x)g(x)r(x)\pi(dx) + \int\int f(x)g(y)\pi(dx)q(x, y)a(x, y)\mu(dy) \\
&= \int f(x)g(x)r(x)h(x)\mu(dx) + \int\int f(x)g(y)q(x, y)a(x, y)h(x)\mu(dx)\mu(dy) \\
&= \int f(x)g(x)r(x)h(x)\mu(dx) + \int\int f(x)g(y)q(y, x)a(y, x)h(y)\mu(dx)\mu(dy) \\
&= \int f(x)g(x)r(x)h(x)\mu(dx) + \int\int f(y)g(x)q(x, y)a(x, y)h(x)\mu(dy)\mu(dx) \\
&= \int\int f(y)g(x)\pi(dx)P(x, dy) \qquad \blacksquare
\end{aligned}$$

where in the second to last line, the switch of measure is justified by Fubini's theorem.

Although general theory of MCMC is nice, the application of MCMC is a bit different since a computer has no sense of infinity or a measure zero. The theory still holds, but instead of kernels, we have matrices as operators on finite or countably infinite state space and integrals are replaced by summations.

The Ising Model

We now present an application of the MH algorithm to a physical model known as the Ising model. The Ising model tries to approximate the behavior of a ferromagnet under a change in temperature by simulate the spin of it's electron on a lattice. The case we will be using is a $N \times N$ square lattice with periodic boundary conditions or a mathematical torus.

The function of we are sampling from is the probability function for the model to be at a certain σ configuration following the Boltzman distribution given by,

$$P(\sigma; \beta) = \frac{\exp(-\beta H(\sigma))}{Z}$$

where:

- $H(\sigma) = -\sum_{<i\ j>} J_{ij}\sigma_i\sigma_j - \mu \sum_j h_j\sigma_j$ is known as the Hamiltonian in physics.
- $\sum_{<i\ j>}$, $i, j \in \{1, 2, ..., N \times N\}$ being sum of the nearest neighbors.
- J_{ij} being the coupling strength at site i,j.
- μ being the magnetic moment.
- h_j being the external magnetic field strength at site j.
- $\sigma = \{\sigma_1, \sigma_2, ..., \sigma_{N\times N}\} \in C$ being the current configuration of the system with $\sigma_j \in S = \{-1, 1\}$, and C is the configuration space of the system being $2^{N\times N}$ dimensions.
- $Z = \sum_\sigma \exp(-\beta H(\sigma))$ is the normalization constant known as the partition function in physics.
- $\beta = \frac{1}{k_B T}$ with k_B being the Boltzman's constant and T being the the temperature of the system in degrees Kelvin.

Usually the partition function Z is very hard to calculate, but one nice thing about MH is that we don't need to calculated it since they are canceled out. All we need is the unnormalized density.

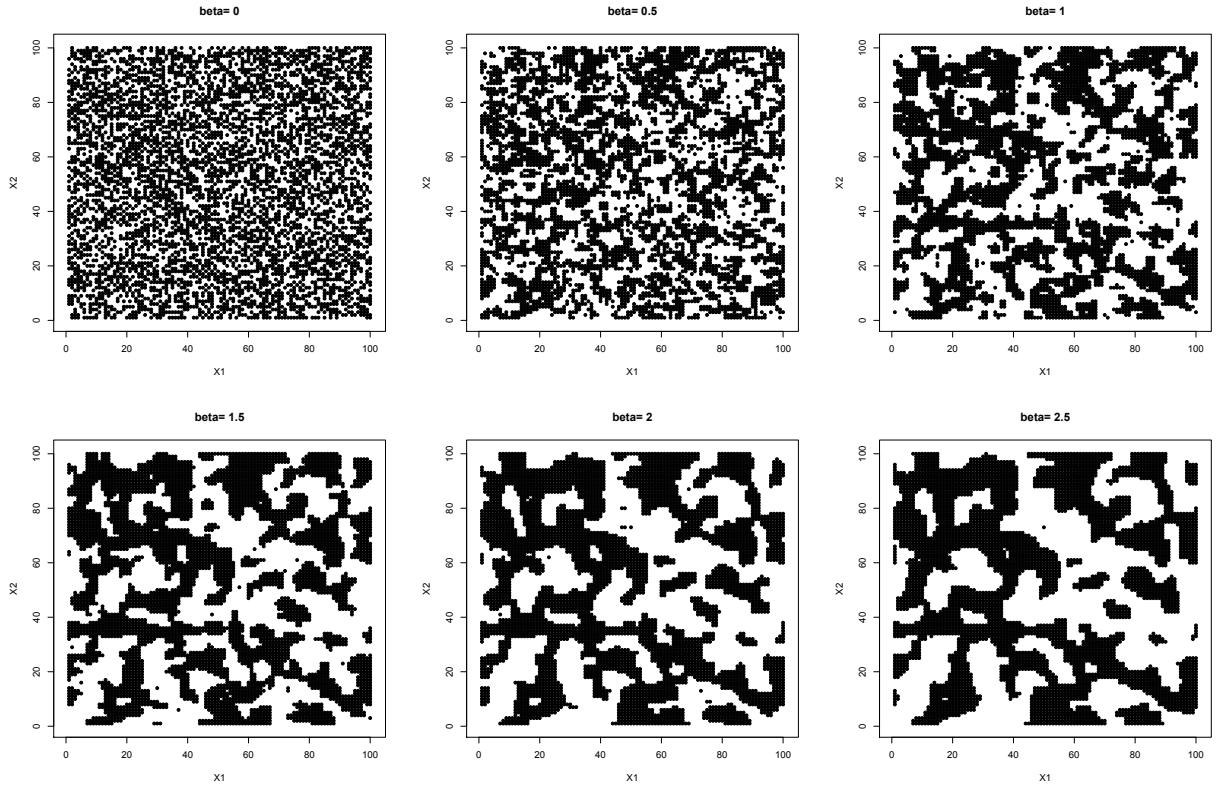
For simplicity, we will make the follow assumptions in our model.

1. $h_j = 0\ \forall j \in \{1, 2, ..., N \times N\}$. External field turned off.
2. $J_{ij} = J_{kl}\ \forall i, j, k, l \in \{1, 2, ..., N \times N\}$. Uniform coupling strength.
3. $J_{ij} = k_B \implies \beta = \frac{1}{T}$.
4. $P(i, j) = P(j, i)$ transition probability are symmetric. In our case, $P(i, j) = \frac{1}{N\times N}\ \forall i, j \in \{1, 2, ..., N \times N\}$.

Here is a simulations of $N = 100$, $T_0 = \infty(\beta = 0)$, $T_f = .4(\beta = 2.5)$ degrees kelvin, the code can be found in the appendix at the end. Our unnormalized density will be,

$$h(\sigma_{ij}) = \exp(-\beta H(\sigma_{ij}))$$
$$\implies R = \frac{h(\sigma'_{ij})}{h(\sigma_{ij})} = \exp(2\beta \sum_{<i\ j>} \sigma_i \sigma_j)$$

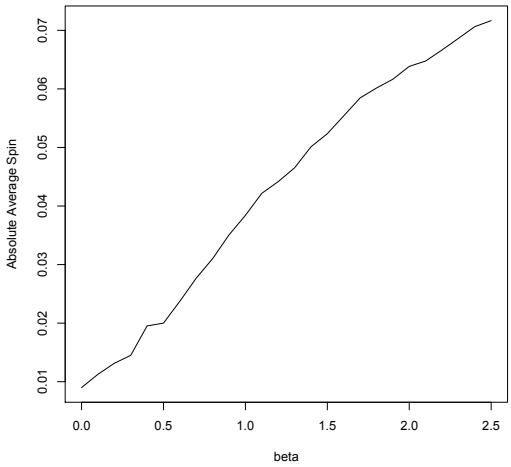
where σ_{ij} denotes the state at site i,j on the lattice. The proposal distribution is canceled out since it is symmetric.



As we sample a random site, the site is flipped with probability min(1,R). This is then repeated n times, then the process repeated for the next temperature. What we are doing is actually finding the global minimum of the function $P(\sigma; \beta)$. This method of minimizing the function is an optimization technique known as Simulated Annealing, but it's actually just a case of MH.

Looking at it from a physical view. As temperature lower, what started as a random lattice starts to form structure. We can see that islands of spins are beginning to form and the system is trying to reach a state of minimum energy. It turns out that there are actually 2 degenerate ground state (global minimum), and the system will pick one at random. In physics, this behavior is known as Spontaneous Symmetry Breaking.

Here we have the same simulation ran over 50 times. We took over an average and plot the absolute value of the average spin.



We can see that there is a sudden jump around $T = 2$ ($\beta = .5$) degrees kelvin. This is known as the critical temperature. It is the temperature where the system decided which one of the two degenerate ground state to head to (known as a phase transition in physics). The theoretical value is actually $T \approx 2.27$ ($\beta \approx 0.44$), which the founder was awarded a Nobel prize.

Appendix

```
n<-5000
N<-100 #lattice size

bc<-function(N,A){           #periodic bc
A[1,]<-A[N+1,]
A[,1]<-A[,N+1]
A[N+2,]<-A[2,]
A[,N+2]<-A[,2]
A}


M<-c()
Y<-rbinom((N+2)*(N+2),1,1/2)
for(i in 1:length(Y)){
if(Y[i]==0){Y[i]<--1}
}
X<-matrix(Y,N+2,N+2)  #IC

for(i in 0:25){
beta<-i*.1
for (k in 1:n){
X<-bc(N,X)
X1<-sample(2:(N+1),1); X2<-sample(2:(N+1),1) #picks random site from lattice
U<-runif(1)
if(U<exp(-2*beta*(X[X1,X2]*X[X1-1,X2]+X[X1,X2]*X[X1+1,X2]+X[X1,X2]*X[X1,X2-1]
+X[X1,X2]*X[X1,X2+1]))){
X[X1,X2]<--X[X1,X2]
}
}
M=c(M,sum(X[2:(N+1),2:(N+1)])/(N^2)) #average spin
}
plot(1,type="n",xlim=c(0,N+1),ylim=c(0,N+1),main=paste("beta=",beta),xlab="X1",ylab="X2")
for (i in 1:N) {
for (j in 1:N){
if (X[i,j]==1){
points(i,j,pch=20)
}}}
}

plot((0:25)*.1,M,type='l',ylab="Absolute Average Spin",xlab="beta")

################
M1<-matrix(0,50,26)
n<-5000
```

```
N<-100

for(l in 1:50){
Y<-rbinom((N+2)*(N+2),1,1/2)
for(i in 1:length(Y)){
if(Y[i]==0){Y[i]<--1}
}
X<-matrix(Y,N+2,N+2) #IC
M<-c()
for(i in 0:25){
beta<-i*.1
for (k in 1:n){
X<-bc(N,X)
X1<-sample(2:(N+1),1); X2<-sample(2:(N+1),1) #picks random site from lattice
U<-runif(1)
if(U<exp(-2*beta*(X[X1,X2]*X[X1-1,X2]+X[X1,X2]*X[X1+1,X2]+X[X1,X2]*X[X1,X2-1]
+X[X1,X2]*X[X1,X2+1]))){
X[X1,X2]<--X[X1,X2]
}
}
M=c(M,sum(X[2:(N+1),2:(N+1)])/(N^2)) #average spin
}
M1[l,]<-M
}

plot((0:25)*.1,apply(abs(M1),2,mean),type='l',ylab="Absolute Average Spin",xlab="beta")
```

Bibliography

1. Brooks, Steve. Gelman, Andrew. Jones, Galin. Meng, Xiao-Li. *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC/Taylor & Francis, 2011. Print.

2. Ross, Sheldon M., and Erol A. Pekoz. *A Second Course in Probability*. Boston: Probability-tore.com, 2007. Print.

3. Chandler, David. *Introduction to Modern Statistical Mechanics*. New York: Oxford UP, 1987. Print.