MCMC AND THE ISING MODEL

ADAM ZWEIGER

ABSTRACT. In this expository paper, we discuss two common Markov chain Monte Carlo(MCMC) methods for sampling from a complicated distribution: the Metropolis-Hastings algorithm and Gibbs Sampling. We then apply Gibbs Sampling to the Ising model, a model from statistical mechanics that is very difficult to simulate without MCMC because there is often a very large number of states. Finally, we shift our focus to mixing times for the Ising model. Much of the material from the sections on the Ising model is from [LP17]. This paper assumes some basic knowledge of Markov chains and methods to bound mixing times.

1. The Metropolis-Hastings Algorithm

The Metropolis-Hastings Algorithm is a method to produce a sequence of random samples from a given density. First, given a probability distribution π on Ω, suppose we wish to find a Markov chain with transition matrix P such that P has a stationary distribution of π. If we had such a chain, we could run the Markov chain until it approximates the stationary distribution. Then, running the chain further would give samples from the stationary distribution. To find such a transition matrix P, we pick any initial transition matrix Ψ, and over many iterations use it to build another chain with stationary distribution close to π :

- Pick an irreducible transition matrix, Ψ, over Ω.
- Pick an initial state $x_0 \in \Omega$ at time $t = 0$.
- Generate a candidate state x' at random according to the Markov chain, $\Psi(x'|x_t)$.
- Calculate the acceptance probability

$$A(x', x_t) = \min\left(1, \frac{\pi_{x'}\Psi_{x',x_t}}{\pi_{x_t}\Psi_{x_t,x'}}\right).$$

- Accept the candidate state x' with probability $A(x', x_t)$. If the state is accepted, we have $x_{t+1} = x'$. Otherwise, reject the candidate and set $x_{t+1} = x_t$. Return to step 3.

More precisely,

Algorithm 1: Metropolis-Hastings algorithm

```
Initialize $x_0 \in \Omega$ at random
for $i \leftarrow 0$ to $N - 1$ do
    Sample $x' \sim \Psi(x'|x_t)$
    Sample $u \sim [0, 1]$
    if $u < A(x', x_t) = \min\left(1, \frac{\pi_{x'}\Psi_{x',x_t}}{\pi_{x_t}\Psi_{x_t,x'}}\right)$ then
        $x_{t+1} = x'$
    else
        $x_{t+1} = x_t$
    end
end
```

This process forms a Markov chain, $x_0, x_1, \ldots$ with transition matrix P given by the transition probabilities,

$$P_{ij} = \begin{cases} \Psi_{ij} \cdot \min\left(1, \frac{\pi_j \Psi_{ji}}{\pi_i \Psi_{ij}}\right) & \text{if } i \neq j. \\ 1 - \sum\limits_{k \in \Omega | k \neq i} \Psi_{ik} \cdot \min\left(1, \frac{\pi_k \Psi_{ki}}{\pi_i \Psi_{ik}}\right) & \text{if } i = j. \end{cases}$$

Theorem 1.1. *The described transition matrix P for the Metropolis-Hastings algorithm is a reversible Markov chain with a stationary distribution of π.*

Definition 1.2. An ergodic Markov chain over state space Ω with transition matrix P is said to be reversible if there exists a probability distribution, π that satisfies

$$\pi_i P_{ij} = \pi_j P_{ji}$$

for all $i, j \in \Omega$.

Lemma 1.3. *Let P be a reversible ergodic transition matrix. The probability distribution π that satisfies the condition*

$$\pi_i P_{ij} = \pi_j P_{ji}$$

for all $i, j \in \Omega$ is the stationary distribution of P.

Proof. Suppose that π satisfies the condition. We have

$$(\pi P)_j = \sum_{i \in \Omega} \pi_i P_{ij} = \sum_{i \in \Omega} \pi_j P_{ji} = \pi_j \sum_{i \in \Omega} P_{ji} = \pi_j$$

for all j. Thus $\pi P = \pi$, so π is the stationary distribution. ■

Proof of Theorem 1.1. Due to the previous lemma, we need only show that P is reversible. That is, show that

$$P_{ij}\pi_i = \Psi_{ij} A(i,j) \pi_i = \Psi_{ji} A(j,i) \pi_j = P_{ji} \pi_j.$$

We finish with casework:

Case 1. $\Psi_{ij}\pi_i = \Psi_{ji}\pi_j$. Here,

$$\frac{\pi_i \Psi_{ij}}{\pi_j \Psi_{ji}} = \frac{\pi_j \Psi_{ji}}{\pi_i \Psi_{ij}} = 1,$$

so $A(i,j) = A(j,i) = 1$. Therefore the chain is reversible.

Case 2. $\Psi_{ij}\pi_i > \Psi_{ji}\pi_j$. In this case, $A(j,i) = 1$, and

$$A(i,j) = \frac{\pi_j \Psi_{ji}}{\pi_i \Psi_{ij}}.$$

Thus

$$\Psi_{ij} A(i,j) \pi_i = \Psi_{ij} \frac{\pi_j \Psi_{ji}}{\pi_i \Psi_{ij}} \pi_i = \pi_j \Psi ji = \Psi_{ji} A(j,i) \pi_j.$$

Case 3. $\Psi_{ij}\pi_i < \Psi_{ji}\pi_j$. Similarly, $A(i,j) = 1$, and

$$A(j,i) = \frac{\pi_i \Psi_{ij}}{\pi_j \Psi_{ji}}.$$

Thus

$$\Psi_{ji} A(j,i) \pi_j = \Psi_{ji} \frac{\pi_i \Psi_{ij}}{\pi_j \Psi_{ji}} \pi_j = \pi_i \Psi ij = \Psi_{ij} A(i,j) \pi_i.$$

■

The Metropolis-Hastings algorithm allows us to draw random samples from a complicated function $f(x)$ without knowing the normalization factor. Say we have $f(x) = Cp(x)$, where $p(x)$ is a probability density function and C is an unknown proportionality factor. Notice that the Markov Chain generated by the Metropolis-Hastings algorithm depends only on the ratios $\frac{\pi_j}{\pi_i}$ rather than the values themselves. If we have $\pi_x = p(x) = f(x)/C$, we can sample from this complicated distribution without ever knowing the normalization factor since

$$\frac{\pi_j}{\pi_i} = \frac{p(j)}{p(i)} = \frac{f(j)}{f(i)}.$$

Often, one picks a proposal distribution which is symmetric, i.e., $\Psi_{ij} = \Psi_{ji}$. This reduces the acceptance probability to

$$A(x_t, x') = \min\left(1, \frac{\pi_{x'} \Psi_{x',x_t}}{\pi_{x_t} \Psi_{x_t,x'}}\right) = \min\left(1, \frac{\pi_{x'}}{\pi_{x_t}}\right).$$

This way, if the candidate sample is more probable, we accept it with probability 1.

One issue with the Metropolis-Hastings algorithm and most other MCMC methods is that the chain might take a long time to approach stationarity. So, typically we *burn-in* the sampler by throwing out the first n samples.

Another problem is that the samples are correlated. One way to reduce this is by thinning the output, storing only every m^{th} point after the burn-in period.

2. Gibbs Sampling

Gibbs Sampling is an important variation of the Metropolis-Hastings algorithm with conditional distributions as the proposal distribution Ψ and acceptance probability 1. It samples from a distribution over several random variables by fixing all but one random variable, sampling that one conditioned on the others. It does this for each random variable. So, all we need are the conditional distributions. Let $x = (x_1, x_2, \ldots x_n)$ and $x_{-i} = (x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots, x_n)$. In this section, we denote components of the state vector with subscripts and the time with superscripts.

Algorithm 2: Gibbs Sampling

```
Initialize $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)} \in \Omega$ at random
for $t \leftarrow 0$ to $N-1$ do
    Pick an index $1 \leq j \leq n$ at random
    Sample $x_j^{(t+1)} \sim \pi(x_j | x_1^{(t+1)}, \ldots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)}, \ldots, x_n^{(t)})$
end
```

This runs a Markov chain on each of the random variables x_i. The rule to update x is choose a random index j, and then choose a new state according to

$$P(x_j^{(t)}, x_j') = \begin{cases} \frac{\pi_{x'}}{\pi(z : z_{-j} = x_{-j})} & \text{if } x'_{-j} = x_{-j}^{(t)}, \\ 0 & \text{Otherwise.} \end{cases}$$

Theorem 2.1. *The described transition matrix P for the Gibbs Sampler is a reversible Markov chain with a stationary distribution of π.*

Proof. We just need to check that π satisfies the detailed balance equation. Suppose we have arbitrary states x and y. If $x_{-j} \neq y_{-j}$, then

$$\pi_x P_{xy} = 0 = \pi_y P_{xy}.$$

Otherwise, we have

$$\begin{aligned} \pi_x P_{xy} &= \pi_x \frac{\pi_y}{\pi_{z:z_{-j}=x_{-j}}} \\ &= \pi_y \frac{\pi_x}{\pi_{z:z_{-j}=y_{-j}}} \\ &= \pi_y P_{yx}. \end{aligned}$$

■

One issue with the Gibbs Sampler is that we always accept samples based on conditional distributions. This results in high correlation between consecutive samples. As with the Metropolis-Hastings algorithm, one typically starts saving samples after a burn-in period and thins the outputs.

3. The Ising Model

The Ising model is a very commonly studied model, originally used to study ferromagnetism. However, the model has numerous applications outside of magnetism including simulating neurons in the brain and simulating lattice gasses.

The most commonly studied spin system is the nearest-neighbor Ising model. Consider a set of magnets, each having one of two possible orientations, or spins, represented by $+1$ and -1 being placed on the vertices of a graph. Let the vertex set of this graph be V and the edge set be E. a A configuration σ represents the orientations of all the magnets. So, $\sigma(v)$ is the spin of vertex v under configuration σ. The state space is $\Omega = \{-1, 1\}^V$. The probability distribution is given by

$$\mu(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z(\beta)}$$

where $\beta = \frac{1}{k_B T}$ is the inverse of the temperature,

$$H(\sigma) = -\sum_{v \sim w} \sigma(v)\sigma(w)$$

is the energy of a configuration, obtained by summing the interactions between neighboring vertices, and

$$Z(\beta) = \sum_{\sigma \in \Omega} e^{-\beta H(\sigma)}$$

is the normalizing constant, called the partition function. The β value, the inverse of temperature, determines the importance of the energy function. At high values, lower energy configurations are more likely. At low values, H is less important and μ is closer to the uniform distribution.

Note that it is impractical to compute the normalizing constant which is a sum over all $2^{|V|}$ configurations. This high amount of states motivates us to use Markov chain Monte

Carlo methods to simulate the Ising model. We will use Gibbs sampling, sometimes called Glauber dynamics for the Ising model. This moves from a starting configuration σ by picking a uniformly random vertex $w \in V$ and generating a new configuration according to μ and the conditional probability that it agrees with σ on vertices other than w.

The new state σ' agrees with σ everywhere except possibly at w, where $\sigma'(w) = 1$ with probability

$$\frac{\mu(\sigma_w = 1|\sigma_{-w})}{\mu(\sigma_w = 1|\sigma_{-w}) + \mu(\sigma_w = -1|\sigma_{-w})} = \frac{e^{(1)\beta \sum_{u:u\sim w} \sigma_u}/Z(\beta)}{e^{(1)\beta \sum_{u:u\sim w} \sigma_u}/Z(\beta) + e^{(-1)\beta \sum_{u:u\sim w} \sigma_u}/Z(\beta)}.$$

This simplifies to

$$p(\sigma, w) = \frac{e^{\beta S(\sigma,w)}}{e^{\beta S(\sigma,w)} + e^{-\beta S(\sigma,w)}} = \frac{1 + \tanh(\beta S(\sigma, w))}{2}.$$

where $S(\sigma, w) = \sum_{u:u\sim w} \sigma(u)$. This only depends on the spins at the vertices connected to w. Thus the transition matrix from configurations σ to σ' is

$$P(\sigma, \sigma') = \frac{1}{|V|} \sum_{w\in V} \frac{e^{\beta\sigma'(w)S(\sigma,w)}}{e^{\beta\sigma'(w)S(\sigma,w)} + e^{-\beta\sigma'(w)S(\sigma,w)}} \cdot \mathbf{1}_{\{\sigma(v)=\sigma'(v) \text{ for } v\neq w\}}.$$

This is reversible with respect to and has a stationary distribution of μ. (Note that $\mathbf{1}_{\{\sigma(v)=\sigma'(v) \text{ for } v\neq w\}}$ is the *indicator function* and in this case is 1 if $\sigma(v) = \sigma'(v)$ for $v \neq w$ and 0 otherwise.)

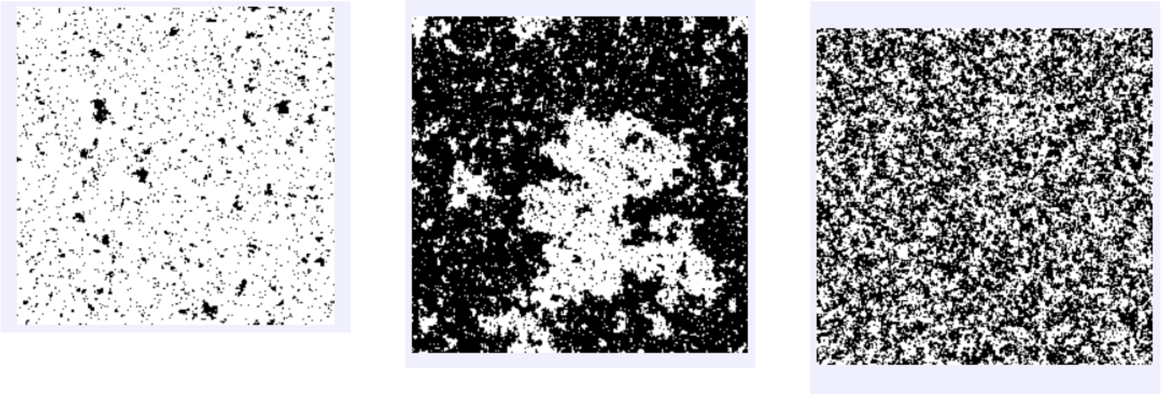

Figure 1. Glauber dynamics for the nearest-neighbor Ising model on a square lattice at low temperature $\beta > \beta_c$ (left), critical temperature $\beta = \beta_c$ (middle), and high temperature $\beta < \beta_c$ (right) after a sufficient burn-in period. At low temperatures, the model tends to be biased towards low energy configurations with more connected vertices having the same spin. As β increases, this goes away and at very high temperature with β close to 0, the energy function plays almost no effect and the distribution is approximately uniform. [LP17]

4. Mixing Times for the Ising Model

We now turn our focus to the mixing times for the Ising model and the Ising model on a complete graph. We begin by recalling Theorem 15.1 from Levin, Peres, and Wilmer's book.

Theorem 4.1. *[LP17] Consider the Ising model on a graph with n vertices and maximal degree Δ. Let $c(\beta) = 1 - \Delta \tanh(\beta)$. If $\Delta \cdot \tanh(\beta) < 1$, then*

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{n(\log n + \log(1/\varepsilon))}{c(\beta)} \right\rceil .$$

Proposition 4.2. *For $x \in [0, \infty)$,*

$$\tanh(x) \leq x.$$

Proof. Let $f(x) = x - \tanh(x)$. Taking the derivative,

$$f'(x) = 1 - (1 - \tanh^2(x)) = \tanh^2(x) \geq 0.$$

Since $f(x)$ is monotonically increasing and $\tanh(0) = 0$, $f(x)$ is always positive for $x > 0$. ■

By the proposition, at high temperatures when $\beta < \Delta^{-1}$, Theorem 4.1 holds and there is fast mixing on the order $O(n \log n)$.

Let K_n be the complete graph on n vertices. The interaction strength for the complete graph is $\sigma(v) \sum_{w:w\sim v} \sigma(w)$, which is of order n. So, let's replace β and take $\beta = \alpha/n$. The probability of updating to a $+1$ is now

$$p(\sigma, w) = \frac{e^{\alpha(S-\sigma(w))/n}}{e^{\alpha(S-\sigma(w))/n} + e^{-\alpha(S-\sigma(w))/n}}$$

where $S = \sum_{i=1}^{n} \sigma(i)$ is the total magnetization. We now have the following theorem about the mixing times for the complete graph Ising model, known as the Curie-Weiss model.

Theorem 4.3. *Let K_n be the complete graph on n vertices, and consider the Markov chain for the Ising model on K_n with $\beta = \alpha/n$.*

(i) If $\alpha < 1$, then

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{n(\log(n) + \log(1/\varepsilon))}{1 - \alpha} \right\rceil .$$

(ii) If $\alpha > 1$, there exists a constant $C_0 > 0$ such that

$$t_{\text{mix}}(\varepsilon) \geq C_0 e^{r(\alpha)n},$$

where $r(\alpha) > 0$.

Proof of (i). We have $\Delta = n - 1$ for the complete graph. Substituting,

$$\Delta \tanh(\beta) = (n-1)\tanh(\alpha/n).$$

Now, we prove that $\tanh(x) \leq x$ for nonnegative x. By Lemma 4.2,

$$\Delta \tanh(\beta) = (n-1)\tanh(\alpha/n) \leq \frac{n-1}{n}\alpha < \alpha.$$

Thus if $\alpha < 1$, then $\Delta \cdot \tanh(\beta) < 1$. Theorem 4.1 completes the proof. ■

Proof of (ii). We bound the mixing time using the bottleneck ratio. Let A_k be the set of configurations σ such that $|\{v : \sigma(v) = 1\}| = k$. We have $\pi(A_k) = \alpha_k/Z(\alpha)$ where

$$\alpha_k = \binom{n}{k} \exp\left(\frac{\alpha}{n}\left[\binom{k}{2} + \binom{n-k}{2} - k(n-k)\right]\right)$$

and $Z(\alpha)$ is a normalizing constant. Now, Stirling's formula tells us

$$\begin{aligned}\log\binom{n}{cn} &= \log\frac{n!}{(cn)!(n-cn)!}\\ &= \log(n!) - \log((cn)!) - \log((n-cn)!)\\ &\sim cn\log(\frac{n}{cn}) + (n-cn)\log(\frac{n}{n-nc})\\ &= -cn\log(c) - n(1-c)\log(1-c).\end{aligned}$$

We take logs to our expression for α_k and apply Stirling's formula:

$$\log(\alpha_{\lfloor cn\rfloor}) = n\varphi_\alpha(c)(1+o(1))$$

where

$$\varphi_\alpha(c) = -c\log(c) - (1-c)\log(1-c) + \alpha\frac{(1-2c)^2}{2}.$$

Define S to be the set of configurations σ with $\sum_{v\in V}\sigma(v) < 0$ and similarly define S' to be the set with $\sum_{v\in V}\sigma(v) > 0$. By symmetry, $2\pi(S) = \pi(S) + \pi(S') \leq 1$. Thus $\pi(S) \leq 1/2$. The only way to get from S to $S^c = \Omega\backslash S$ is through $A_{\lfloor n/2\rfloor}$. This is a bottleneck between states with positive magnetization and states with negative magnetization. Now,

$$Q(S, S^c) \leq \pi(A_{\lfloor n/2\rfloor})$$

and

$$\pi(S) = \sum_{j\leq\lfloor n/2\rfloor}\pi(A_j).$$

Let the maximum value of $\varphi_\alpha(c)$ on $[0, 1/2]$ be obtained at $c = c_\alpha$. We differentiate our expression for φ,

$$\varphi'_\alpha(1/2) = 0$$

and

$$\varphi''_\alpha(1/2) = -4(1-\alpha).$$

So for $\alpha > 1$, a local minimum is attained at $c = 1/2$. This means the maximum on the interval $[0, 1/2]$ must be at $c_\alpha < 1/2$. Therefore, we can bound the bottleneck ratio as

$$\Phi(S) = \frac{Q(S,S^c)}{\pi(S)} \leq \frac{\pi(A_{\lfloor n/2\rfloor})}{\pi(A_{\lfloor c_\alpha n\rfloor})} = \frac{a_{\lfloor n/2\rfloor}/Z(\alpha)}{a_{\lfloor c_\alpha n\rfloor}/Z(\alpha)} = \frac{\exp[\varphi_\alpha(1/2)n(1+o(1))]}{\exp[\varphi_\alpha(c_\alpha)n(1+o(1))]}.$$

Since $\varphi_\alpha(c_\alpha) > \varphi_\alpha(1/2)$, there is an $r(\alpha) > 0$ and a constant b such that $\Phi_\star \leq be^{-nr(\alpha)}$. Thus the mixing time is $\Omega(e^{nr(\alpha)})$. ■

At high temperatures($\alpha < 1$) the complete graph Ising model is fast mixing with order $n\log n$ and at low temperatures ($\alpha > 1$) the mixing time is exponential in n. In [LLP10], it is shown that for the case of $\alpha = 1$, t_{mix} is on the order of $n^{3/2}$. This interesting phenomenon of the mixing time transitioning from $\theta(n\log n)$ to $\theta(n^{3/2})$ to $\theta(e^n)$ is further studied in [DLP09]. Moreover, this critical slowdown is not unique to the complete graph Ising model. It turns out that other graphs, for example, the widely studied Ising model on a square lattice, exhibit this transition of mixing times at the critical temperature as well.

The mixing times of the Glauber dynamics for the Ising model is a broad and active topic of current research with many open problems . For more on the square lattice Ising model and its mixing times, we refer the reader to [LS10]. We also recommend [DLP09] and [LLP10] for further reading on the cutoff and mixing times of the complete graph Ising model.

References

[DLP09] Jian Ding, Eyal Lubetzky, and Yuval Peres. The mixing time evolution of glauber dynamics for the mean-field ising model. *Communications in Mathematical Physics*, 289(2):725–764, Apr 2009. URL: `http://dx.doi.org/10.1007/s00220-009-0781-9`, `doi:10.1007/s00220-009-0781-9`.

[LLP10] David A. Levin, Malwina J. Luczak, and Yuval Peres. Glauber dynamics for the mean-field ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, 146:223, Jan 2010. `doi:10.1007/s00440-008-0189-z`.

[LP17] D.A. Levin and Y. Peres. *Markov Chains and Mixing Times*. MBK. American Mathematical Society, 2017. URL: `https://books.google.com/books?id=f208DwAAQBAJ`.

[LS10] Eyal Lubetzky and Allan Sly. Critical ising on the square lattice mixes in polynomial time, 2010. `arXiv:1001.1613`.