MARKOV CHAIN MONTE CARLO

RYAN WANG

Abstract. This paper gives a brief introduction to Markov Chain Monte Carlo methods, which offer a general framework for calculating difficult integrals. We start with the basic theory of Markov chains and build up to a theorem that characterizes convergent chains. We then discuss the Metropolis-Hastings algorithm.

1. Introduction

The Monte Carlo method refers to a large class of algorithmic methods that rely on random sampling. The term itself was coined by physicists at Los Alamos Laboratory during World War II. In this paper we focus on Markov Chain Monte Carlo (MCMC), which involves performing a random walk on a system of interest. The first MCMC algorithm was published by a group of Los Alamos researchers, Metropolis et al. (1953), with the purpose of computing difficult integrals. Hastings (1970) later generalized their original work, but it was not until the late 1980s that computers became powerful enough to support practical applications. Since then MCMC has become a ripe area for research and has found applications in diverse fields such as genetics, finance, and cryptography, among others. It offers a comparatively simple framework for solving otherwise difficult computational problems which perhaps explains its popularity. In section 2 we define a Markov chain and discuss some important properties for convergence. In section 3 we discuss the Metropolis-Hastings algorithm for generating a random sample and an application in Bayesian inference.

2. Finite Markov Chains

Intuitively, a Markov chain is a random process with the property that behavior of the system in one period depends only on its behavior in the previous period. The system transitions, or changes states, with certain transition probabilities. In this paper we discuss only finite, time-homogeneous Markov chains.

Example 2.1. Consider a random walk on the number line $\{1, ..., N\}$. Starting at an initial state the random walker moves right to the next higher integer with probability p and left to the next lower integer with probability $1-p$. If the walker reaches one of the boundary points then he moves back inside the interval with probability 1. Denote $p(i, j)$ as the transition probability of moving from point i to point j. We have that

$$p(i, i+1) = p,\ p(i, i-1) = 1-p,\ 0 < i < N,$$

$$p(1, 2) = 1,\ p(N, N-1) = 1,$$

and $p(i,j) = 0$ for all remaining i, j. We note that these probabilities depend only on the current state of the system and ignore how the system may have arrived at that state.

Definition 2.2. Consider a sequence of random variables $\{X_i, i \geq 0\}$ which take values in the finite set $\Omega = \{1, ..., N\}$. We refer to the sequence as a finite Markov chain if it satisfies

$$P\{X_t = x_t | X_0 = x_0, ..., X_{t-1} = x_{t-1}\} = P\{X_t = x_t | X_{t-1} = x_{t-1}\}$$

for all t and $x_0, ..., x_t \in \Omega$. We think of t as an index for time.

Definition 2.3. Now consider the probability of moving from a state i to another state j, where only the length of the time interval matters

$$p(i,j) = Pr\{X_t = j | X_{t-1} = i).$$

We refer to $p(i,j)$ as a time-homogeneous one-step transition probability.

Definition 2.4. We now form an $N \times N$ matrix from the one-step transition probabilities

$$\mathbf{P} = ||p(i,j)||_{i,j \in \Omega}$$

called the transition matrix. Note that each row in the matrix gives a distribution so that $\mathbf{P}$ is a stochastic matrix. In other words all entries are nonnegative and

$$\sum_{j \in \Omega} p(i,j) = 1 \ \forall i \in \Omega.$$

Definition 2.5. Similarly, the probability of moving from i to j in n steps is given by the n-step transition probability

$$p_n(i,j) = Pr\{X_{n+k} = j | X_k = i\}.$$

Lemma 2.6. *The n-step transition probability is given by the ij entry in the matrix $\boldsymbol{P}^n$.*

Proof. The statement is clearly true for $n = 1$. Assuming it to be true for a given n we apply the law of total probability to get

$$\begin{aligned} p_{n+1}(i,j) &= Pr\{X_{n+1} = j | X_0 = i\} \\ &= \sum_k Pr\{X_n = k | X_0 = i\} Pr\{X_{n+1} = j | X_n = k\} \\ &= \sum_k p_n(i,k) p(k,j). \end{aligned}$$

We know that $p(k,j)$ is the kj entry of $\mathbf{P}$ and by induction hypothesis we know that $p_n(i,k)$ is the ik entry of $\mathbf{P}^n$. Thus $p_{n+1}(i,j)$ is the ij entry in $\mathbf{P}^n\mathbf{P} = \mathbf{P}^{n+1}$. □

Corollary 2.7. $p_{n+m}(i,j) = \sum_{k \in \Omega} p_n(i,k) p_m(k,j)$.

As a result, we can fully describe a Markov chain with only its transition matrix and a probability distribution for the starting value, $\bar{\phi}_0$. Here $\bar{\phi}_0$ is a row vector. In other words we can write down the probability the chain is in a given state at a given time,

$$Pr\{X_n = j\} = \sum_{i \in \Omega} \bar{\phi}_0(i) p_n(i,j). \tag{2.8}$$

In fact for any transition matrix there exists a Markov chain with the given transitions. We now consider the convergence of Markov chains.

Definition 2.9. $\bar{\pi}$ is a stationary distribution if

$$\bar{\pi} = \bar{\pi}P.$$

Definition 2.10. The Markov chain given by $\mathbf{P}$ and $\bar{\phi}_0$ converges to $\bar{\pi}$ if

$$\lim_{n\to\infty} \bar{\phi}_0 \mathbf{P}^n = \bar{\pi}.$$

Definition 2.11. A chain is irreducible if for all $i, j \in \Omega$, there exists an integer n such that $\mathbf{P}^n(i,j) > 0$. In other words, every state can eventually be reached starting from any other state.

Definition 2.12. Let $\mathcal{T}_i = \{n \geq 1 | \mathbf{P}^n(i,i) > 0\}$ be the set of times for which it is possible to return to i, starting from i. We define the period of a state i to be the greatest common divisor of $\mathcal{T}_i$. We say that state i is aperiodic if $d_i = 1$.

Lemma 2.13. *If a Markov chain is irreducible, then $d_i = d_j$ for all $i, j \in \Omega$.*

Proof. For arbitrary states i and j, there exist integers $n, m > 0$ such that $P_n(i,j) > 0$ and $P_m(j,i) > 0$. Applying (2.7) we can rewrite

$$p_{n+m}(i,i) = \sum_{k\in\Omega} p_n(i,k)p_m(k,i) \geq p_n(i,j)p_m(j,i) > 0.$$

Thus we have that $n + m \in \mathcal{T}_i$ which implies $d_i | (n+m)$. Now suppose $l \in \mathcal{T}_j$ so that $p_l(j,j) > 0$. We can similarly rewrite

$$p_{n+m+l}(i,i) \geq p_n(i,j)p_l(j,j)p_m(j,i) > 0$$

so that $n + m + l \in \mathcal{T}_i$ and $d_i | (n+m+l)$, implying $d_i | l$. By definition $d_j = \gcd\{l | p_l(j,j) > 0\}$, hence $d_i \leq d_j$. Repeating the argument we have $d_j \leq d_i$. □

From this we see that all states of an irreducible chain have the same period, so we can refer to the period d of the chain itself. We call an irreducible chain aperiodic if $d = 1$.

Lemma 2.14. *If $\boldsymbol{P}$ is the transition matrix for an irreducible, aperiodic Markov chain, then there exists an integer $M > 0$ such that for all $n > M$, $\boldsymbol{P}^n$ has strictly positive entries.*

Proof. Recall $\mathcal{T}_i = \{n \geq 1 | \mathbf{P}^n(i,i) > 0\}$ and note that $\mathcal{T}_i$ is closed under addition since

$$p_{s+t}(i,i) \geq p_s(i,i)p_t(i,i).$$

We use the fact from number theory that any nonempty subset of the nonnegative integers which is closed under addition and has greatest common divisor d contains all but finitely many of the elements $\{0, d, 2d, ...\}$. In particular, since the chain is irreducible, there exists t_i such that $t > t_i$ implies $t \in \mathcal{T}_i$ and thus $p_t(i,i) > 0$. Again by irreducibility, there exists $m(i,j)$ such that $p_{m(i,j)}(i,j) > 0$. So for all $t > t_i$

$$p_{t+m(i,j)}(i,j) \geq p_t(i,i)p_{m(i,j)}(i,j) > 0.$$

Take $M = \max\{t_i + m(i,j)\}$ over all pairs (i,j). For all $n \geq M$, $i, j \in \Omega$ we have $p_n(i,j) > 0$.

□

Definition 2.15. For a state $i \in \Omega$, we denote the first hitting time for i to be

$$T_i = \min\{t \geq 1 | X_t = i\}.$$

We now prove the existence and uniqueness of a stationary distribution for irreducible, aperiodic Markov chains.

Theorem 2.16. *If $\boldsymbol{P}$ is the transition matrix for an irreducible Markov chain then there exists a stationary distribution π of $\boldsymbol{P}$ and $\pi(i) > 0$ for all $i \in \Omega$.*

Proof. Take an arbitrary state $k \in \Omega$ and set $X_0 = k$. Define the expected number of visits to j before returning to k by

$$\bar{\pi}(j) = \sum_{t=0}^{\infty} P\{X_t = j, T_k > t\}.$$

We know $\{T_k \geq t+1\} = \{T_k > t\}$ and that this event is fully determined by the values of $X_0, ..., X_t$. Thus,

$$P\{X_t = i,\ X_{t+1} = j,\ T_k \geq t+1\} = P\{X_t = i,\ T_k \geq t+1\}p(i,j).$$

We use these facts to verify that $\bar{\pi}$ is indeed a stationary distribution for $\mathbf{P}$. For any $j \in \Omega$ we have

$$\begin{aligned}
\bar{\pi}(j)\mathbf{P} &= \sum_{i\in\Omega} \bar{\pi}(i)p(i,k) \\
&= \sum_{i\in\Omega}\sum_{t=0}^{\infty} P\{X_t = i,\ T_k > t\}p(i,j) \\
&= \sum_{t=0}^{\infty}\sum_{i\in\Omega} P\{X_t = i,\ T_k \geq t+1\}p(i,j) \\
&= \sum_{t=0}^{\infty}\sum_{i\in\Omega} P\{X_t = i,\ X_{t+1} = j,\ T_k \geq t+1\} \\
&= \sum_{t=0}^{\infty} P\{X_{t+1} = j,\ T_k \geq t+1\} \\
&= \sum_{t=1}^{\infty} P\{X_t = j,\ T_k \geq t\} \\
&= \sum_{t=1}^{\infty} P\{X_t = j,\ T_k > t\} + \sum_{t=1}^{\infty} P\{X_t = j,\ T_k = t\} \\
&= \bar{\pi}(j) - P\{X_0 = j,\ T_k > 0\} + \sum_{t=1}^{\infty} P\{X_t = j,\ T_k = t\}. \\
&= \bar{\pi}(j) - P\{X_0 = j\} + P\{X_{T_k} = j\}.
\end{aligned}$$

Now consider two cases. If $j = k$ then $P\{X_0 = j = k\} = P\{X_{T_k} = j = k\} = 1$. If $j \neq k$ then $P\{X_0 = j = k\} = P\{X_{T_k} = j = k\} = 0$. Thus $\bar{\pi} = \bar{\pi}\mathbf{P}$.

□

Definition 2.17. A function $h : \Omega \to \mathbb{R}$ is harmonic at i if

$$h(i) = \sum_{j\in\Omega} p(i,j)h(j).$$

If we represent h as a column vector, we see that a function is harmonic at every point in Ω if it satisfies $h = \mathbf{P}h$.

Lemma 2.18. *Suppose $\boldsymbol{P}$ is irreducible. If h satisfies $h = \boldsymbol{P}h$ then it is a constant function.*

Proof. Since Ω is a finite set, there exists $i_0 \in \Omega$ such that $h(i_0) = M$ is a maximum. Now suppose there exists $j \in \Omega$ with $p(i_0, j) > 0$ such that $h(j) < M$. Then we have the following contradiction

$$h(i_0) = p(i_0, j)h(j) + \sum_{k \neq j} p(i_0, k)h(k) < M.$$

Thus $h(j) = M$ for all j with $p(i_0, j) > 0$. Since $\mathbf{P}$ is irreducible we know that any state can be reached from any other by transitions of positive probability. That is for any $k \in \Omega$ there exists a path $i_0, i_1, ..., i_n$ from i_0 to k with $p(i_m, i_{m+1}) > 0$ for $m = \{0, ..., n-1\}$. So $h(k) = M$ for all $k \in \Omega$.

□

Theorem 2.19. *If $\boldsymbol{P}$ is the transition matrix for an irreducible Markov chain, then the stationary distribution π is unique.*

Proof. By Lemma (2.18), any h satisfying $(\mathbf{P} - \mathbf{I})h = \mathbf{0}$ must be constant. In other words the kernel of $\mathbf{P} - \mathbf{I}$ must have dimension 1. Applying the rank-nullity theorem, the rank and thus the row rank of $\mathbf{P} - \mathbf{I}$ is $|\Omega| - 1$. Now the space of row vectors satisfying $\pi = \pi P$ must be one-dimensional. Since this space contains only one vector representing a distribution, that is a vector whose entries sum to one, π is unique. □

Definition 2.20. Consider two distributions p and q on Ω. We define their total variation distance to be the maximum difference in probability they assign to a given event. That is

$$||p - q||_{TV} = \max_{A \subset \Omega} |p(A) - q(A)|.$$

We now state the basic convergence theorem involving total variation distance. See [2] for a proof.

Theorem 2.21. *Suppose $\boldsymbol{P}$ is the transition matrix for an irreducible, aperiodic Markov chain with stationary distribution π. Then there exist $\alpha \in (0, 1)$ and $C > 0$ such that*

$$\max_{i \in \Omega} ||P^t(i, \cdot) - \pi||_{TV} \leq C\alpha^t.$$

3. Markov Chain Monte Carlo

Suppose we want to generate samples from a complicated distribution $\pi(\theta)$. MCMC methods offer a solution to an inverted version of the problem we previously considered. In particular we take our target distribution to be the stationary distribution and construct a Markov chain that converges to it. Realizations from this process are kept as dependent samples. The following example illustrates a situation in which such a procedure might be desirable.

Example 3.1 (Bayesian Inference)**.** Suppose we have a vector of observed data $\boldsymbol{x}$ and a model of the process generating that data. In standard statistical inference

we wish to estimate the vector of parameters $\boldsymbol{\theta}$ for the model. In Bayesian inference we assume that $\boldsymbol{\theta}$ follows some distribution, called the 'posterior' distribution, rather than having a fixed value. Using Bayes Theorem we obtain the posterior distribution by combining the 'likelihood' of the data and a 'prior' distribution for $\boldsymbol{\theta}$. In other words,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto L(\boldsymbol{x}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}). \tag{3.2}$$

Here the likelihood is the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{\theta}$. The prior distribution $p(\boldsymbol{\theta})$ expresses uncertainty regarding $\boldsymbol{\theta}$ before observing data. In principle we can obtain the posterior distribution for any given $\boldsymbol{\theta}$. In practice this requires knowledge of the normalizing constant $Z = 1/\int L(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, which may be difficult or impossible to compute. Using MCMC methods we can generate samples from the $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ and easily calculate moments of the distribution.

We now discuss a general method known as the Metropolis-Hastings (M-H) algorithm which requires minimal knowledge of the target distribution. Our goal is to construct transitions $p(i,j)$ such that the Markov chain given by $\mathbf{P} = ||p(i,j)||_{i,j\in\Omega}$ converges to the target distribution $\bar{\pi}$. In M-H, we construct transitions by choosing an appropriate 'proposal distribution' $q(i,j)$ and 'acceptance probability' $\alpha(i,j)$. Sampling from the proposal-generating distribution, we propose a transition from the current state to the next. The acceptance probability governs whether we actually change states or remain in the current state. We will use the following result to derive transitions such that $\mathbf{P}$ converges to the correct distribution.

Lemma 3.3 (Detailed Balance)**.** *If a distribution $\bar{\pi}$ satisfies*

$$\bar{\pi}(i)p(i,j) = \bar{\pi}(j)p(j,i) \forall i,j \in \Omega$$

then it is a stationary distribution for the Markov chain given by $\boldsymbol{P}$.

Proof. Sum both sides over j. Since $\mathbf{P}$ is stochastic, we have

$$\sum_{j\in\Omega} \bar{\pi}(i)p(i,j) = \sum_{j\in\Omega} \bar{\pi}(j)p(j,i) = \bar{\pi}(x).$$

□

By (3.3) we know that if our proposal $q(i,j)$ satisfies the detailed balance condition then π is the stationary distribution for the constructed Markov chain. Accept proposals with probability one and we have the correct transitions.

Now suppose that for some i,j we have $\pi(i)q(i,j) > \pi(j)q_(j,i)$. Roughly speaking the proposal distribution $q(i,j)$ causes the chain to move from i to j too often and from j to i too rarely. Rather than accepting every proposed transition, accept proposals with probability $\alpha(i,j)$. If the proposal is rejected then the chain remains at i. We derive the following transition probabilities

$$P_{MH}(i,j) = \begin{cases} q(i,j)\alpha(i,j) & \text{if } j \neq i \\ 1 - \sum_{k:k\neq i} q(i,k)\alpha(i,k) & \text{if } j = i. \end{cases}$$

Because the chain moves from j to i too rarely, we choose $\alpha(j,i)$ to be as large as possible, one. Now plug into the detailed balance condition

$$\pi(i)q(i,j)\alpha(i,j) = \pi(j)q(j,i)$$

and solve for the acceptance probability

$$\alpha(i,j) = \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}.$$

Reversing the inequality above, we set $\alpha(i,j) = 1$ and derive $\alpha(j,i)$ in the same way.

Theorem 3.4 (Metropolis-Hastings)**.** *Let $\boldsymbol{P} = ||p(i,j)||_{i,j\in\Omega}$ and $\boldsymbol{Q} = ||q(i,j)||_{i,j\in\Omega}$ denote transition matrices such that*

$$p(i,j) = \begin{cases} q(i,j) & \text{if } i \neq j,\ \alpha(i,j) \geq 1 \\ q(i,j)\alpha(i,j) & \text{if } i \neq j,\ \alpha(i,j) < 1 \\ 1 - \sum_{k:k\neq i} q(i,k)\alpha(i,k) & \text{if } x = y \end{cases}$$

where

$$\alpha(i,j) = \min\{\frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}, 1\}.$$

Then $\boldsymbol{P}$ satisfies the detailed balance condition.

Proof. This is true by construction. □

We implement M-H by choosing an initial point x_0 and proposal density $q(x,y)$. We then proceed in the following way

- For $j = 1, ..., N$ generate y from $q(x_{j-1}, \cdot)$ and u from $U(0,1)$.
- If $u \leq \alpha(x_j, y)$ put $x_j = y$.
- Else put $x_j = x_{j-1}$.
- Discard the first b samples $\{x_1, ..., x_b\}$, letting the Markov chain reach its stationary distribution.
- Return $\{x_{b+1}, ..., x_N\}$ and treat as a dependent sample from π.

Thus far we have not put any restrictions on the proposal distribution q. In principle we only require knowledge of the ratio $\pi(y)q(y,x)/p(x)q(x,y)$ up to a constant and that the chain sufficiently explores the state space. The original Metropolis procedure called for a symmetric proposal density, where

$$q(x,y) = q(y,x).$$

In this case the acceptance probability reduces to $\alpha(x,y) = \pi(y)/\pi(x)$. A generalization of the Metropolis proposal is the random walk proposal. Here, we generate a new value y by adding the current value x to a random variable z. That is

$$y = x + z,\ z \sim f$$

where z follows some distribution f. We have $q(x,y) = f(z)$. Oftentimes f is chosen to be a uniform or multivariate normal distribution. Another popular proposal is the independent proposal. In this case we generate a new value independently of the current value. Thus

$$q(x,y) = f(y).$$

Example 3.5 (Gibbs Sampler)**.** The following algorithm, known as the Gibbs Sampler, is a special case of M-H where proposals are generated from posterior conditional distributions and accepted with probability one. Consider a random vector $\boldsymbol{\theta} = (\theta^{(1)}, ...\theta^{(p)})$ with distribution π and let $\pi(\theta^{(i)}|\boldsymbol{\theta}^{(-i)})$ denote the conditional distribution of $\theta^{(i)}$ given $\theta^{(1)}, ..., \theta^{(i-1)}, \theta^{(i+1)}, ..., \theta^{(p)}$. We modify the M-H

algorithm by updating the vector componentwise to generate samples. That is given the current position, $\boldsymbol{\theta}_t = (\theta_t^{(1)}, ...\theta_t^{(p)})$, draw sequentially from the conditionals

$$\theta_{t+1}^{(1)} \sim \pi(\theta^{(1)}|\boldsymbol{\theta}_t^{(-1)})$$

$$\vdots$$

$$\theta_{t+1}^{(p)} \sim \pi(\theta^{(p)}|\boldsymbol{\theta}_t^{(-p)})$$

to generate $\boldsymbol{\theta}_{t+1}$.

In practice there are a number of issues left to address when implementing MCMC. Naturally we want to determine how many samples to discard, or in other words the rate of convergence of the chain. We may also want to determine an optimal initial value and whether choice of initial value impacts convergence. We also want to know how well the chain mixes, or how often the algorithm accepts proposals. A poorly mixing chain in which proposal values are frequently rejected will not sample the space very efficiently. For example, suppose we sample from a normal distribution using a random walk proposal where f is uniform with width a. Choosing a too large will yield a poorly mixing chain. The first 1000 samples from such a chain with $a = 25$ are plotted in Figure 1. Compare that to a well mixing chain with $a = 1$ plotted in Figure 2. Furthermore we may want to know how many samples are needed to efficiently summarize the target distribution. Since generated samples are positively correlated with each other, we expect that the effective sample size will be smaller than N. These are all areas of current research.

Acknowledgments. It is a pleasure to thank my mentors Al and Shawn for their guidance and insight.

References

[1] Robert Gramacy. Notes on Monte Carlo Inference.
[2] David Levin, Yuval Peres, and Elizabeth Wilmer. Markov Chains and Mixing Times.

FIGURE 1. Random Walk M-H with U[-25,25] Proposal

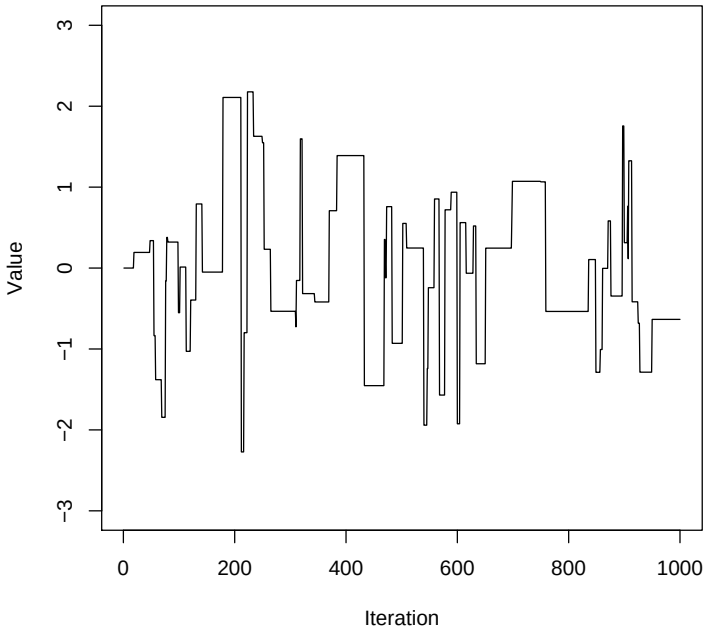

FIGURE 2. Random Walk M-H with U[-1,1] Proposal