ON SHARPER ESTIMATE OF THE CONVERGENT RATE OF METROPOLIS-HASTINGS ALGORITHM

CHENDONG SONG

1. Introduction and background

Markov Chain Monte Carlo (MCMC) algorithms have gained widespread popularity in the field of statistics as powerful tools for sampling from complex probability distributions. In simple terms, MCMC aims to construct a Markov chain with a stationary distribution π identical to the target distribution we wish to sample. One of the most commonly used MCMC algorithms is the Metropolis-Hastings algorithm.

Here's a breakdown of the Metropolis-Hastings algorithm: Start by choosing an initial state X_0. Then, given the current state X_n, generate a proposal Y_{n+1} from the proposal distribution $P(X_n, \cdot)$. Simultaneously, flip an independent coin with the probability of heads determined by $\alpha\left(X_n, Y_{n+1}\right)$, where:

$$\alpha(x, y) = \min\left[1, \frac{\pi(y)P(y, x)}{\pi(x)P(x, y)}\right].$$

Subsequently, in the event of obtaining heads, the proposal is "accepted" by setting $X_{n+1} = Y_{n+1}$; in the case of tails, the proposal is "rejected", and $X_{n+1} = X_n$. Update the iteration index from n to $n+1$ and repeat the process. The Markov chain generated by this algorithm satisfies the detailed balance condition, ensuring its stationary distribution is indeed π.

To ensure the algorithm's efficacy, it is crucial to determine whether the chain ultimately converges to the stationary distribution π and to explore the rate of convergence. Regarding the former, [2] demonstrates that if the chain is irreducible and aperiodic, it will ultimately converge to its stationary distribution. However, understanding the "speed of convergence" is a more complex matter. Initially, let's define the total variation distance between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ as:

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_A |\nu_1(A) - \nu_2(A)|.$$

For a Markov chain with a state space denoted by Ω and a transition kernel represented by P, the distance between X_t and the stationary distribution is quantified by

$$d(t) := \sup_{x \in \Omega} ||P^t(x, \cdot) - \pi(\cdot)||$$

This expression captures the evolution of the chain over time, reflecting the discrepancy between its distribution at time t and the stationary distribution π.

Given that every irreducible and aperiodic Markov chain eventually converges to its stationary distribution, $d(t)$ asymptotically approaches 0 as t tends to infinity. However, understanding the time required for this convergence is of great interest. To address this, we introduce the concept of *mixing time*.

$$t_{mix}(\epsilon) := \min\{t : d(t) \leq \epsilon\}. \tag{1}$$

In the context of MCMC, we define a subset $C \subseteq \Omega$ as small, or (n_0, ϵ, ν)-small, if there exists a positive integer n_0, a positive value ϵ, and a probability measure $\nu(\cdot)$ on Ω such that the following minorization condition is satisfied:

$$P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot) \quad x \in C,$$

Namely, for all $x \in C$ and measurable sets $A \subseteq \Omega$, the inequality $P^{n_0}(x, A) \geq \epsilon\nu(A)$ holds. Using the method of coupling, we can deduce that if a Markov chain satisfies minorization condition, then the total variance of a given Markov chain decreases exponentially over time.

Theorem 1.1 (Theorem 8 in [2])**.** *Consider a Markov chain with stationary distribution $\pi(\cdot)$. Suppose the minorisation condition is satisfied for some $n_0 \in \mathbf{N}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$, in the special case $C = \Omega$ (i.e., the entire state space is small). Then $||P^n(x, \cdot) - \pi(\cdot)||_{TV} \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for all $x \in \Omega$, where $\lfloor r \rfloor$ is the greatest integer not exceeding r.*

The theorem presented provides a non-asymptotic description of the convergence rate; however, it lacks a quantitative estimate for the parameters ϵ and n_0. In practical applications of the algorithm, understanding precisely when to terminate the iteration it is crucial. Consequently, establishing bounds for ϵ and n_0 becomes important.

In this article, my objective is to establish bounds for ϵ and n_0 within the framework of the Metropolis-Hastings algorithm. The first section focuses on scenarios where Ω is bounded, and the proposal distribution is uniform. Here, I precisely provide the expression $d(n) = (1 - 1/C)^n$, with C representing the maximum value of the target distribution. The second section extends the analysis to situations where $\Omega = \mathbb{R}$ and consider arbitrary proposal distribution. I give the bound that $d(n) \leq (1 - \inf_{x \in \mathbb{R}} \frac{f(x)}{\pi(x)})^n$, where π is the target distribution and f is the proposal function.

2. Finite Case

In the initial analysis, we examine a straightforward scenario. Let $\Omega = [0, 1]$ represent a one-dimensional finite interval, π denote the stationary distribution, and consider the proposal probability $Y_{n+1} = Q(X_n, \cdot) \sim U(0, 1)$. In other words, if the chain is at X_n at time n, we randomly sample Y_n from $[0, 1]$ and accept it with a probability of $\alpha = \min[1, \frac{\pi(Y_n)}{\pi(X_n)}]$. The subsequent proposition aims to attain the precise value of ϵ when $n = 1$.

Proposition 2.1. *The Markov chain of the Metropolis-Hastings theorem on $[0, 1]$ with uniform proposal transition is $(1, 1/C, \nu)$-small, where $C = \sup_{x \in [0,1]} \pi(x)$ and ν is a probability measure with pdf π.*

Proof. Firstly, consider the distribution of $P(x_0, \cdot)$, the distribution after 1 iteration with starting point x_0. Let E be the set $\{x \in [0, 1] | f(x) \geq f(x_0)\}$ and F be the set $\{x \in [0, 1] | f(x) < f(x_0)\}$. The uniform sampling Y_1 falls within E with probability $m(E)$ and within F with probability $m(F)$. Here, m denotes the Lebesgue measure on $\mathbb{R}$. Note that once the sample falls within E, it must be accepted. Consequently, the distribution of $P(x_0, \cdot)$ is also uniform on E. If the sample falls within F, it can be accepted with a probability of $f(x)/f(x_0)$. Denoting μ as the probability measure of $P(x_0, \cdot)$, for any measurable set M,

$$\mu(M) = \mu(M \cap \{x_0\}) + \frac{1}{\pi(x_0)}\nu(M \cap F) + m(M \cap E). \tag{2}$$

where $\mu(\{x_0\})$ is the probability of $X_1 = X_0 = x_0$, signifying the probability of rejection. Actually it is equal to

$$\int_F \frac{\pi(x_0) - \pi(x)}{\pi(x_0)} dx = m(F) - \frac{\nu(F)}{\pi(x_0)} \tag{3}$$

Therefore, for any measurable set M,

$$C\mu(M) \geq \frac{C}{\pi(x_0)}\pi(M \cap F) + \int_{M\cap E} C dx \geq \nu(M \cap F) + \nu(M \cap E) = \nu(M). \tag{4}$$

So the Markov chain is $(1, 1/C, \nu)$-small. □

It is evident that the convergence rate is influenced by the maximum value of the target pdf(π). When the maximum value is modest, the algorithm exhibits rapid convergence. For instance, with $C = 2$, just 10 iterations result in a total variation distance of approximately 1/1000 with the stationary distribution. However, in the presence of a local extreme maximum in the pdf, as in the case of $C = 1000$, around 2000 iterations are required for convergence.

Upon establishing that the chain is $(1, 1/C, \nu)$-small, we can deduce from Theorem 1.1 that $||P^n(x, \cdot) - \pi(\cdot)||_{TV} \leq (1 - 1/C)^n$ for all $x \in \Omega$. In other words, $d(n) \leq (1 - 1/C)^n$. Furthermore, it is noteworthy that this estimation is sharp. Specifically,

Proposition 2.2. *If X_t is a metropolis-hasting chain on [0,1] with uniform proposal with stationary distribution pdf π and maximal C, then $d(n) = (1 - 1/C)^n$ and $t_{mix}(\epsilon) = \log_{1-1/C} \epsilon$*

Proof. First, we establish that $d(n) \geq (1 - 1/C)^n$. Assume that π achieves its maximum at x_0. By the definition of d, we have $d(n) \geq ||P^n(x_0, \cdot) - \pi(\cdot)||TV$. Consider the distribution of $P^n(x_0, \cdot)$. At time t, assuming $X_t = x_0$, we initially propose $Yt + 1 \sim U(0, 1)$ and accept it with a probability of $\frac{\pi(Y_{t+1})}{\pi(X_t)}$, given that $\pi(X_t)$ is maximal. The probability of the proposal being accepted is then

$$\int_0^1 \frac{\pi(y)}{\pi(x_0)} dy = \frac{1}{C}. \tag{5}$$

This implies that at each time step, the Markov chain X_t has a probability of $1/C$ to remain at x_0 if the chain does not move before time t. Therefore, X_t has a probability of at least $(1-1/C)^t$ of consistently staying at x_0, i.e., $P^n(X_t, x_0) \geq (1 - 1/C)^n$. However, $\pi(x_0) = 0$ because any discrete point has a probability measure of 0 in a continuous distribution. Therefore,

$$d(n) \geq ||P^n(x_0, \cdot) - \pi(\cdot)||_{TV} \geq |P^n(x_0, x_0) - \pi(x_0)| \geq (1 - 1/C)^n \tag{6}$$

□

Remark 2.3. I have extensively studied discrete Markov chains, as detailed in [1], which covers various examples like random walks on cycles, tori, or hypercubes. The book offers numerous techniques to establish both upper and lower bounds of mixing times for Markov chains, yet the bounds often differ by a constant (e.g., $(1-c)^n \leq d(n) \leq 2(1-c)^n$). However, in the Metropolis-Hastings chain, it's remarkable that the upper and lower bounds of $d(n)$ are identical. This symmetry is an exceptional property of the Metropolis-Hastings Markov chain, distinguishing it from other Markov chains.

3. Infinite Case

In this section, we extend our analysis to a more general scenario where we aim to sample from a probability distribution with a probability density function (pdf) π on $\mathbb{R}$. We focus on the Metropolis-Hastings chain with a proposal $Q(X_n, \cdot)$ subject to a proposal distribution with

pdf f, where f can be readily sampled (e.g., a normal distribution). Remarkably, a similar estimate holds for this extended situation.

Theorem 3.1. *If (X_t) is a Metropolis-Hastings chain defined above with stationary distribution pdf π , then the total variance of this chain satisfies*

$$d(n) \leq (1 - \inf_{x\in\mathbb{R}} \frac{f(x)}{\pi(x)})^n \tag{7}$$

Remark 3.2. This scenario indeed constitutes a generalization of the case discussed in the first section. If we choose $f(x)$ to be uniformly distributed on [0,1] and the support of π is also on [0,1], then (7) aligns with Proposition 2.1. Additionally, it's important to note that if π and f coincide across the entire real line, then the acceptance probability $\alpha(x, y) = \min[1, \frac{\pi(y)f(x)}{\pi(x)f(y)}]$ is always equal to 1. This implies that every proposal is accepted, and the chain reaches its stationary distribution in a single transition.

Remark 3.3. In the theorem's setting, the proposal sample $Q(X_n, \cdot)$ is a fixed distribution, independent of the initial point X_n. This matter is very crucial . Indeed, if we were to allow the proposal sample to depend on X_n (e.g., using a normal distribution $N(X_n, \sigma)$), this estimation method would fail, and the convergence would no longer be guaranteed.

Proof. Denote $\epsilon = \inf_{x\in\mathbb{R}} \frac{f(x)}{\pi(x)}$, we prove that the chain is $(1, \epsilon, \pi) - small$. Thus by Theorem 1.1 we can deduce (7).

Here still consider the distribution of $P(x_0, \cdot)$. Let E be the set $\{x \in \mathbb{R}|\pi(x) \geq \pi(x_0)\}$ and F be the set $\{x \in \mathbb{R}|\pi(x) < \pi(x_0)\}$. Then the uniform sampling Y_1 drops at E with probability $\int_E f(x)dx$ and at F with probability $\int_F f(x)dx$. If the sample x drops at F, then it can be accepted with probability $\frac{\pi(x)f(x_0)}{\pi(x_0)f(x)}$ by the acceptance probability . Denote μ as the probability measure of $P_(x_0, \cdot)$ For any measurable set M,

$$\mu(M) = \mu(M \cap \{x_0\}) + \int_{M\cap F} f(x)\frac{\pi(x)f(x_0)}{\pi(x_0)f(x)}dx + \int_{M\cap E} f(x)dx \tag{8}$$

where $\mu(\{x_0\})$ is the probability of $X_1 = X_0 = x_0$,which means the proposal is rejected. Therefore, for any measurable set M,

$$\begin{aligned} \mu(M) &\geq \frac{f(x_0)}{\pi(x_0)} \int_{M\cap F} \pi(x)dx + \int_{M\cap E} \epsilon\pi(x)dx \\ &\geq \int_{M\cap F} \epsilon\pi(x)dx + \int_{M\cap E} \epsilon\pi dx = \epsilon \int_M \pi(x)dx. \end{aligned} \tag{9}$$

Therefore, the chain is $(1, \epsilon, \pi) - small$. □

MCMC is a method to sample an unknown pdf(π) from a given simple pdf(f). So If we already have a family of pdfs (f) that are easily accessible, such as the normal distribution with parameters of mean and variance $N(\mu, \sigma^2)$, the question arises: How do we choose an appropriate pdf from this family to minimize the convergence time of the Metropolis-Hastings algorithm? In fact, the goal is to maximize $\epsilon = \inf_{x\in\mathbb{R}} \frac{f(x)}{\pi(x)}$, where f represents the pdf of $N(\mu, \sigma^2)$ with undetermined parameters μ and σ, and π is an arbitrary pdf. To achieve this, optimization methods discussed in lectures, such as bisection or gradient descent, can be employed to determine the optimal parameters for f to find

$$\epsilon_\pi := \max_{\sigma,\mu} \inf_{x\in\mathbb{R}} \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\pi(x)} \tag{10}$$

Theoretically, if the pdf π possesses certain favorable properties, it is possible to establish an upper bound for ϵ. A direct question is whether $\epsilon_\pi > 0$, and the subsequent proposition provides an answer to this.

Proposition 3.4. *$\epsilon_\pi > 0$ if only if there exists $M > 0, C > 0, d \in \mathbb{R}$ such that $\pi(x) \leq Me^{-C(x-d)^2}$ for all $x \in \mathbb{R}$. If this condition holds, then $\epsilon_\pi \geq \frac{\sqrt{C}}{M\sqrt{\pi}}$.*

Proof. If $\pi(x) \leq Me^{-C(x-d)^2}$ for all $x \in \mathbb{R}$, then we can take $\mu = d$ and $\sigma = \sqrt{\frac{1}{2C}}$, bring it into (10) and deduce that $\epsilon_\pi \geq \frac{\sqrt{C}}{M\sqrt{\pi}}$. On the other hand, if not, then for any μ and σ and M, there exists some $x \in \mathbb{R}$ such that $\pi(x) > Mf_{\mu,\sigma}(x)$,so $\inf \frac{f_{\mu,\sigma}(x)}{\pi(x)} = 0$ for all μ and σ. Therefore $\epsilon_\pi = 0$. □

And a direct corollary is

Corollary 3.5. *If the support of π is compact, then $\epsilon_\pi > 0$.*

Remark 3.6. This section finds that in metropolis-hasting algorithm with fixed proposal distribution, the target distribution π should be somewhat "similar" to the proposal distribution, particularly they may have the similar convergence rate in the infinity. The calculation of normal distribution is only an example. In practice, we should choose the proposal distribution in accordance with the target distribution.

References

[1] Levin, David A., and Yuval Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[2] Roberts G O, Rosenthal J S. General state space Markov chains and MCMC algorithms[J]. 2004.