Exploration of Markov Chain Monte Carlo Algorithms

Boyi Li

Supervisor: Professor Jeffrey Rosenthal

University of Toronto

boyi.li@mail.utoronto.ca

April 12, 2019

Motivation of MCMC [1]

- Suppose have **complicated** and **high-dimensional** unnormalized probability density $\pi = cg$
- Want get samples $X_1, X_2, ... \sim \pi$. (Hard to do Monte Carlo and Rejection sampler)
- Define a Markov chain (dependent random process) $X_0, X_1, X_2, ...$ in such a way that for large n, $X_n \approx \pi$.
- Then we can estimate $\mathbf{E}_\pi(h) \equiv \int h(x)\pi(x)dx$ by

$$\mathbf{E}_\pi(h) \approx \frac{1}{M-B} \sum_{i=B+1}^{M} h(X_i)$$

where B (burn-in) chosen large enough so $X_B \approx \pi$, and M chosen large enough to get good Monte Carlo estimates

Metroplis-Hasting Algorithm [1]

- Choose some initial value X_0.
- Then given X_{n-1}, choose a proposal $Y_n \sim q(X_{n-1}, \cdot)$
- Let $A_n = \frac{\pi(Y_n)q(Y_n, X_{n-1})}{\pi(X_{n-1})q(X_{n-1}, Y_n)}$ and $U_n \sim \text{Uniform}[0, 1]$.
- Then if $U_n < A_n$, set $X_n = Y_n$ ("accept"), otherwise set $X_n = X_{n-1}$ ("reject").
- Repeat, for $n = 1, 2, 3, ..., M$
- Proposal density could be not symmetric $q(x, y) \neq q(y, x)$
- If $q(x, y) >> q(y, x)$,then Metropolis chain would spend too much time at y and not enough at x, so need to accept fewer moves $x \to y$.
- Require $q(x, y) > 0$ iff $q(y, x) > 0$
- If proposal $Y_n \sim MVN(X_{n-1}, \sigma^2 I)$. "RWM". Choose σ such that the accept rate is 0.234. Best Performance. Optimal Scaling (Roberts and Rosenthal, Stat Sci 2001)

Independence Sampler [1]

- Choose some initial value X_0.
- Then given X_{n-1}, choose a proposal $Y_n \sim q(\cdot)$
- Let $A_n = \frac{\pi(Y_n)q(Y_n,X_{n-1})}{\pi(X_{n-1})q(X_{n-1},Y_n})$ and $U_n \sim \text{Uniform}[0,1]$.
- Then if $U_n < A_n$, set $X_n = Y_n$ ("accept"), otherwise set $X_n = X_{n-1}$ ("reject").
- Repeat, for $n = 1, 2, 3, ..., M$
- Special case for Metroplis-Hasting Algorithm. Proposal density independent of X_{n-1}.

Problem of interest [2]

Probability Kernel $P^n(x, A)$: The probability that tart from state x and move to set A in n step. $n \in \mathbb{N}, A \in \mathcal{X}$(state space). General State (Unaccountable) space (only cases about set A, $P^n(x, \{y\}) = 0$.)

Definition

The **total variation distance** between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ is:

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

Problem of Interest [2]

The concepts to total variance helps us to answer question: Is $\lim_{n\to\infty}\|P^n(x,\cdot)-\pi(\cdot)\|=0$? And, given $\epsilon>0$, how large must n be so that $\|P^n(x,\cdot)-\pi(\cdot)\|<\epsilon$? Can we get a qualitative bounds for n? Can we get a quantitative bounds for n?

Definition

A chain is ϕ-*irreducible* if there exists a non-zero σ-finite measure ϕ on $\mathcal{X}$ such that for all $A\subseteq\mathcal{X}$ with $\phi(A)>0$, and for all $x\in\mathcal{X}$, there exists a positive integer $n=n(x,A)$ such that $P^n(x,A)>0$.

Definition

A Markov chain with stationary distribution $\pi(\cdot)$ is *aperiodic* if there do not exist $d\geq 2$ and disjoint subsets $\mathcal{X}_1,\mathcal{X}_2,\mathcal{X}_3,...,\mathcal{X}_d\subseteq\mathcal{X}$ with $P(x,\mathcal{X}_{i+1})=1$ for all $x\in\mathcal{X}_i(i\leq i\leq d-1)$ and $P(x,\mathcal{X}_1)=1$ for all $x\in\mathcal{X}_d$, such that $\pi(\mathcal{X}_1)>0$(and hence $\pi(\mathcal{X}_i)>0$ for all i). Otherwise, the chain is *periodic*, with *period* d, and *periodic decomposition* $\mathcal{X}_1,...,\mathcal{X}_d$

Markov Chain Convergence Theorem [2]

Theorem

If a Markov chain on a state space with countably generated σ-algebra is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x \in \mathcal{X}$

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| = 0.$$

In particular, $\lim_{n\to\infty} P^n(x,A) = \pi(A)$ for all measurable $A \subseteq \mathcal{X}$.And If $\mathbf{E}_\pi(|h|) < \infty$, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n h(X_i) = \mathbf{E}_\pi(h)$. "LLN"

We can use this theorem to justify Metropolis-Hasting Algorithm.
We also call a chain is **Ergodic**, if it converges.

Ergodicity [2]

Answer questions: How fast does the chain converge?

Uniform Ergodic

A Markov chain having stationary distribution $\pi(\cdot)$ is *uniformly ergodic* if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M\rho^n, n = 1, 2, 3, ...$$

for some $\rho < 1$ and $M \leq \infty$.

Geometric Ergodicity

A Markov chain with stationary distribution $\pi(\cdot)$ is *geometrically ergodic* if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)\rho^n, \quad n = 1, 2, 3, ...$$

for some $\rho < 1$, where $M(x) < \infty$ for π-*a.e.* $x \in \mathcal{X}$

Useful facts [1], [3]

Fact

CLT holds for $\frac{1}{n}\sum_1^n h(X_i)$ if chain is geometrically ergodic and $E_\pi(|h|)^{2+\delta} < \infty$ for some > 0.

Can calculate confidence interval. Important.

Fact about Independence sampler

Independence sampler is geometric ergodic, if and only if theres is $\delta > 0$ such that $q(x) \geq \delta\pi(x)$ for π-a.e. $x \in \mathcal{X}$. If so, then $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\delta)^n$, for π-a.e. $x \in \mathcal{X}$.

Fact about RWM

RWM is geometrically ergodic essentially if and only if π has exponentially light tails, i.e. there are $a, b, c > 0$ such that $\pi(x) \leq ae^{b|x|}$ whenever $|x| > c$.

Adaptive MCMC [4]

"MCMC with learning"
$\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ be a collection of Markov chain kernels on $\mathcal{X}$, each of which has $\pi(\cdot)$ as a stationary distribution.
Assuming P_γ is ϕ-irreducible and aperiodic, this implies that P_γ be ergodic for $\pi(\cdot)$. The choice of γ is given by a $\mathcal{Y}$-valued random variable Γ_n,

$$\mathcal{G}_n = \sigma(X_0, ..., X_n, \Gamma_0, ..., \Gamma_n)$$

be the filtration generated by $\{(X_n, \Gamma_n\}$. Thus,

$$\mathbb{P}[X_{n+1} \in B | X_n = x, \Gamma_n = \gamma, \mathcal{G}_{n-1}] = P_\gamma(x, B), \quad x \in \mathcal{X}, \gamma \in \mathcal{Y},$$

while the conditional distribution of Γ_{n+1} given $\mathcal{G}_n$ is to be specified by the particular adaptive algorithm being used. We let

$$A^{(n)}((x,\gamma), B) = \mathbb{P}[X_n \in B | X_0 = x, \Gamma_0 = \gamma]$$

$$T(x,\gamma,n) = \|A^{(n)}((x,\gamma),\cdot) - \pi(\cdot)\| \equiv \sup_{B\in\mathcal{F}} |A^{(n)}((x,\gamma), B) - \pi(B)|$$

Is Adaptive MCMC ergodic?

Adaptive MCMC Convergence Theorem [4]

for $\epsilon > 0$, define the "ϵ convergence time function" $M_\epsilon : \mathcal{X} \times \mathcal{Y} \to \mathbf{N}$ by

$$M_\epsilon(x, \gamma) = \inf\{n \geq 1 : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \epsilon\}.$$

If each individual P_γ is ergodic, then $M_\epsilon(x, \gamma) < \infty$.
Let $x_* \in \mathcal{X}$ and $\gamma_* \in \mathcal{X}$., if

- Diminishing Adaptation: Adapt less and less as the algorithm proceeds.
  (i.e., $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| = 0$ in probability)
- Containment:For all $\epsilon > 0$, the sequence $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^\infty$ is bounded in probability: Given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$, i.e. for all $\delta > 0$, there is $N \in \mathbf{N}$ such that $\mathbb{P}[M_\epsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta$ for all $n \in N$.

Then $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$.

Adaptive MCMC [1]

Consider RWM on $X = R^d$.
Proposal Density. $Y_n \sim MVN(X_{n-1}, \Sigma)$ How to choose Σ?

- Previous σI_d, choose σ such that the accept rate is 0.234.
- Can do better. Choose $\Sigma = ((2.38)^2/d)\Sigma_0$. Σ_0 is the covariance matrix of the target distribution.
- Σ_0 usually unknown, but we can estimate it based on run so far. (Use generated variables). And for large n, hopefully we have $\Sigma_n = \Sigma_0$.(empirical covariance matrix)
- Usually also add ϵI_d to proposal covariance, to improve stability. $\epsilon = 0.05$.
- Can be justified by previous theorem.

Graph [1]

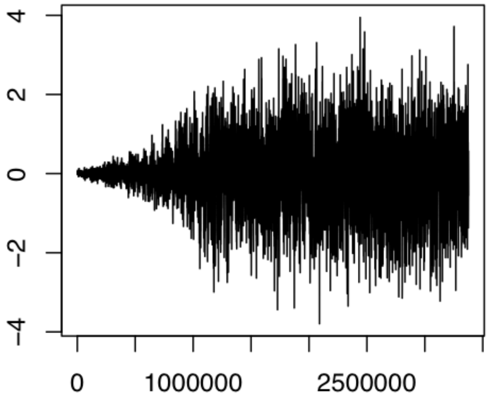

Figure: Trace plot for a Normal distribution in 200 dimensions

Adversarial Markov Chain [5]

- Assume we have probability kernel $P(x, \cdot)$, such that $\lim_{n\to\infty}\|P^n(x,\cdot) - \pi\| = 0, \quad x \in \mathcal{X}$. Ergodic.
- there is a constant $D < \infty$ such that P never moves more than a distance D, that is such that

$$P(x, \{y \in \mathcal{X} : \eta(x, y) \leq D\}) = 1, \quad x \in \mathcal{X}$$

- Let K be a bounded set.K_r is the set within distance r of K.
- It begins with $X_0 = x_0$ for some specific initial state x_0; assume that $x_0 \in K$. When ever the process is outside of K, it moves according to the Markov transition probabilities P
- When the process is inside of K, it can move arbitrarily, according to an adversary's wishes, in a nonanticipatory manner (i.e., adapted to $\{X_n\}$). And cannot move more than a distance D. i.e. Can only move to points within K_D.
- Theoretically, people are curious about the conditions to ensure X_n is bounded.

Application in Adaptive MCMC [5]

"Only Adapt in a bounded set"
Consider RWM on R_d. Let $K \subset R^d$ be a bounded set and D be a constant. Let Σ_* be a fix covariance matrix. Start from $x_0 \in K$. Proposal $Y_n \sim MVN(X_{n-1}, \Sigma)$.

- If $X_{n-1} \notin K$, let $\Sigma = \Sigma_*$
- If $X_{n-1} \in K$ with $dist(X_{n-1}, K) > 1$. Use Adaptive MCMC, i.e. $\Sigma = ((2.38)^2/d)\Sigma_{n-1}$
- If $X_{n-1} \in K$ with $dist(X_{n-1}, K) = u$ and $0 < u < 1$. Then combination. $Y_n \sim (1-u)N(X_{n-1}, \Sigma_*) + uN(X_{n-1}, ((2.38)^2/d)\Sigma_{n-1})$
- Reject if $|Y_n - X_{n-1}| > D$.
- Can add ϵI_d to proposal covariance when adapt.

Always work by related algorithm.

- J. Rosenthal, "Lecture notes of sta3431," , Sep. 2018.
- G. O. Roberts and J. Rosenthal, "General state space markov chains and mcmc algorithms," *Probability Surveys*, vol. 1, Apr. 2004. DOI: 10.1214/154957804100000024.
- S. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. Jan. 2009, ISBN: 978-0-521-73182-9.
- G. O. Roberts and J. Rosenthal, "Coupling and ergodicity of adaptive mcmc," , Apr. 2019.
- R. Craiu, L. Gray, K. Latuszynski, N. Madras, G. O. Roberts, and J. Rosenthal, "Stability of adversarial markov chains, with an application to adaptive mcmc algorithms," *The Annals of Applied Probability*, vol. 25, Mar. 2014. DOI: 10.1214/14-AAP1083.

Thank you!