Complexity Bounds for Markov Chain Monte Carlo

Jun Yang
(joint work with Jeffrey S. Rosenthal)

Department of Statistical Sciences
University of Toronto

Carleton University, August 10, 2017

Motivations

- Quantitative bounds for MCMC
  e.g. Drift and Minorization [Rosenthal1995].
- Convergence complexity of MCMC
  e.g. "little is known for MCMC complexity" by Yang, Wainwright, Jordan, AoS, 2016.
- Big data
  "Large p and large n" or "large p and small n".
- Directly translating the quantitative bounds is problematic
  e.g. Rajaratnam and Sparks, arXiv 2015:
  "We therefore hope that one consequence of our work will be to motivate the proposal and development of new ideas analogous to those of Rosenthal that are suitable for high-dimensional settings."

Basics

- Transition kernel:

$$P^k(x, A) := \Pr(X_k \in A \,|\, X_0 = x)$$

- Total variation distance:

$$\|P^k(x, \cdot) - \pi(\cdot)\| := \sup_{A \in \mathcal{B}} |P^k(x, A) - \pi(A)|$$

- Mixing Time:

$$t = \arg\min_k \|P^k(x, \cdot) - \pi(\cdot)\| \leq 1/4$$

- Geometrically Ergodic (usually proved by drift and minorization):

$$\|P^k(x, \cdot) - \pi(\cdot)\| \leq M(x)\gamma^k, \quad 0 < \gamma < 1$$

Splitting and Coupling

- Coupling inequality:

$$|P^k(x, A) - \pi(A)| \leq \Pr(T > k).$$

- If $P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in \mathcal{X}$ where $\epsilon > 0$, then

$$P(x, \cdot) = \epsilon Q(\cdot) + (1 - \epsilon)\frac{P(x, \cdot) - \epsilon Q(\cdot)}{1 - \epsilon}$$

- $|P^k(x, A) - \pi(A)| \leq (1 - \epsilon)^k$ (uniformly ergodic)
- Usually, there does not exists $Q(\cdot)$ and $\epsilon > 0$ such that $P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in \mathcal{X}$.

Drift and Minorization

Minorization Condition

For some set $R \subseteq \mathcal{X}$ (R is called the small set) for which $\pi(R) > 0$

$$P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in R$$

How long it takes between visits to R?

Drift Condition

For some function $f : \mathcal{X} \to \mathbb{R}^+$, some $0 < \lambda < 1$, and $b < \infty$

$$E[f(X_1) \,|\, X_0 = x] \leq \lambda f(x) + b, \quad \forall x \in \mathcal{X}.$$

Under the drift condition, choose small set $R = \{f(x) \leq d\}$, where $d > \frac{2b}{1-\lambda}$.

Issues in high-dimensional MCMC

Recall the drift condition: for $0 < \lambda < 1$ and $b < \infty$

$$E(f(X_1) \,|\, X_0 = x) \leq \lambda f(x) + b, \quad \forall x \in \mathcal{X}$$

When $p \to \infty$:

- $\lambda \to 1$ and b too large $\Rightarrow$ $d > \frac{2b}{1-\lambda}$ too large;
- The small set $R = \{f(x) \leq d\}$ is too large;
- Typical $\epsilon = O(e^{-p})$ $\Rightarrow$ upper bound on mixing time is exponentially increasing (Rajaratnam and Sparks, 2015).

It is unknown if [Rosenthal1995] can give tight convergence complexity order as $p \to \infty$.

Coupling using both "large set" and "small set"

Generalized Drift Condition

Let $R' \in \mathcal{X}$ be a large set, function $f(\cdot)$ satisfies

$$\begin{aligned} E(f(X_1)) \,|\, X_0 = x, X_1 \in R') &\leq E(f(X_1)) \,|\, X_0 = x) \\ &\leq \lambda f(x) + b, \quad \forall x \in R', \end{aligned}$$

New Quantitative Bounds (Y. and Rosenthal, 2017)

Under associated minorization condition, for any $0 < r < 1$

$$\begin{aligned} \|P^k(x_0, \cdot) - \pi(\cdot)\| \leq & (1-\epsilon)^{rk} + \frac{(\alpha A)^{rk}\left(1 + \frac{b}{1-\lambda} + f(x_0)\right) - \alpha^{rk}}{\alpha^k - \alpha^{rk}} \\ & + k\,\pi((R')^c) + \sum_{i=1}^{k} P^i(x_0, (R')^c). \end{aligned}$$

The large set R' should be chosen to balance the two parts.

Gibbs Sampler

Consider the following example:

$$\begin{aligned} Y_i \mid \theta_i \quad &\sim \mathcal{N}(\theta_i, 1), \quad 1 \le i \le n, \\ \theta_i \mid \mu, A &\sim \mathcal{N}(\mu, A), \quad 1 \le i \le n, \\ \mu &\sim \text{ flat prior on } \mathbb{R}, \\ A &\sim \mathbf{IG}(a, b). \end{aligned}$$

- Observed: $(Y_1, \ldots, Y_n)$;
- State: $x = (A, \mu, \theta_1, \ldots, \theta_n)$;
- $p = n + 2$
- Gibbs sampler for the posterior distribution of

$$\pi(\cdot) = \mathcal{L}(A, \mu, \theta_1, \ldots, \theta_n \mid Y_1, \ldots, Y_n).$$

Analysis of Gibbs Sampler

- This model was originally analyzed by Rosenthal in 1996.
- However, directly translating the bound gives $\epsilon = O(e^{-n})$.

We use the new quantitative bound to show

Theorem (Y. and Rosenthal 2017)

For large enough n, we have

$$\|P^k(x_0, \cdot) - \pi(\cdot)\| \leq C_1\gamma^k + C_2\frac{k(1+k)}{n} + C_3\frac{k}{\sqrt{n}},$$

which implies the mixing time is $O(1)$.

Proof

Consider the following order:

$$\mu^{(1)} \sim \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}\right),$$

$$\theta_i^{(1)} \sim \mathcal{N}\left(\frac{\mu^{(1)} + Y_i A^{(0)}}{1 + A^{(0)}}, \frac{A^{(0)}}{1 + A^{(0)}}\right),$$

$$A^{(1)} \sim \textbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right).$$

Suppose we use the initial state:

$$\bar{\theta}^{(0)} = \bar{Y}, \quad A^{(0)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} - 1.$$

Proof (cont.)

- Use the drift function
  $f(x) = n(\bar{\theta} - \bar{Y})^2 + n\left[\left(\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} - 1\right) - A\right]^2$
- We have

$$E[f(X_1) \,|\, X_0 = x] \leq \left(\frac{1+2A}{1+2A+A^2}\right)^2 f(x) + b,$$

  where $b = \mathcal{O}(1)$ for all $x \in \mathcal{X}$.
- This is not a valid drift function since $\left(\frac{1+2A}{1+2A+A^2}\right)^2$ depends on A.
- "bad" states if A is close to zero.

Proof (cont.)

- Let $\Delta = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, choose a threshold T and define the large set

$$R'_T = \left\{ x : \left(\frac{\Delta}{n-1} - 1 - A \right)^2 \leq \left(\frac{\Delta}{n-1} - 1 - T \right)^2 \right\}.$$

- We can verify the generalized drift condition:

$$E[f(X_1) \,|\, X_0 = x] \leq \left(\frac{1+2T}{1+2T+T^2} \right)^2 f(x) + b, \quad \forall x \in R'_T,$$

and

$$E[f(X_1) \,|\, X_0 = x, X_1 \in R'_T] \leq E[f(X_1) \,|\, X_0 = x], \quad \forall x \in R'_T.$$

Proof (cont.)

- First part: if $T = \mathcal{O}(1)$ then $\epsilon = \Omega(1)$, and

$$(1-\epsilon)^{rk} + \frac{(\alpha A)^{rk}\left(1 + \frac{b}{1-\lambda} + f(x_0)\right) - \alpha^{rk}}{\alpha^k - \alpha^{rk}} = C_1\gamma^k.$$

- Second part: for large enough n, we have

$$k\,\pi((R')^c) + \sum_{i=1}^{k} P^i(x_0, (R')^c) \leq C_2\frac{k(1+k)}{n} + C_3\frac{k}{\sqrt{n}}.$$

- Combing the above two parts to get

$$\|P^k(x_0, \cdot) - \pi(\cdot)\| \leq C_1\gamma^k + C_2\frac{k(1+k)}{n} + C_3\frac{k}{\sqrt{n}}.$$

References

- Rosenthal, *Minorization conditions and convergence rates for Markov chain Monte Carlo*, JASA, 1995.
- Rajaratnam and Sparks, *MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains*, arXiv:1508:00947, 2015.
- Yang, Wainwright, and Jordan, *On the computational complexity of high-dimensional Bayesian variable selection*, AoS, 2016.
- Y. and Rosenthal, *Complexity results for MCMC derived from quantitative bounds*, arXiv:1708.00829, 2017.