Beyond Worst-Case Mixing Times for Markov Chains

Maxim Rabinovich, Aaditya Ramdas,
Michael I. Jordan, and Martin J. Wainwright
Department of Electrical Engineering and Computer Science, University of California—Berkeley

Convergence of MCMC

Definition (Total variation distance). Let X and Y be two random variables taking values in a set Ω. The total variation distance between them is defined by

$$d_{\mathrm{TV}}(X,\ Y) = \sup_{A \subset \Omega} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|.$$

Definition (Absolute spectral gap). If $P \in \mathbb{R}^{d \times d}$ is the transition matrix of an irreducible, aperiodic, and reversible Markov chain, with eigenvalues

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d > -1,$$

the absolute spectral gap is defined by

$$\gamma_* = 1 - \max\{\lambda_2,\ |\lambda_d|\}.$$

Theorem 1. *If (X_n) is an irreducible, aperiodic, and reversible Markov chain, and π_n denotes the distribution of X_n, then*

$$d_{\mathrm{TV}}(\pi_n,\ \pi) \leq \frac{(1-\gamma_*)^n}{\sqrt{\pi_{\min}}} \cdot d_{\mathrm{TV}}(\pi_0,\ \pi).$$

Why is MCMC so hard?

Total variation is a worst-case measure of distance.

Theorem 2. *If X and Y are random variables taking values in a set Ω, the total variation distance between them satisfies*

$$d_{\mathrm{TV}}(X,\ Y) = \sup_{f\colon \Omega \to [0,\ 1]} |\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]|.$$

Yet often we only care about very simple functions.

- Posterior mean corresponds to $f = x$.
- Posterior covariance corresponds to $f = xx^T - \mathbb{E}[X]\,\mathbb{E}[X]^T$.
- In a mixture model with cluster membership vector z, cluster co-membership probabilities correspond to $f = \mathbf{1}(z_i = z_{i'})$ for data indices $i \neq i'$.

Function mixing

Instead, convergence with respect to individual functions.

Definition (Function variation distance). Let X and Y be two random variables taking values in a set Ω and let $f\colon \Omega \to [0,\ 1]$. The function variation distance with respect to f is then defined by

$$d_f(X,\ Y) = |\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]|.$$

Definition (Function absolute spectral gap). Let q_j be the (left) eigenvectors of the transition matrix P and let $f\colon [d] \to [0,\ 1]$ be a function. Then the function absolute spectral gap is defined by

$$\gamma_f = 1 - \max_{j \neq 1\colon q_j^T f \neq 0} |\lambda_j|.$$

In words, it is the gap between 1 and the largest absolute value of an eigenvalue whose eigenspace f is not orthogonal to.

The function absolute spectral gap controls the rate of convergence in d_f.

Theorem 3. *If (X_n) is an irreducible, aperiodic, and reversible Markov chain with state space $[d]$, π_n denotes the distribution of X_n, and $f\colon [d] \to [0,\ 1]$, then*

$$d_f(\pi_n,\ \pi) \leq \frac{(1-\gamma_f)^n}{\sqrt{\pi_{\min}}} \cdot d_f(\pi_0,\ \pi).$$

Application: concentration of measure

Previous results give a single rate for all functions.

Theorem 4 (Uniform Hoeffding bound, Léon and Perron 2004). *Let (X_n) be an irreducible, aperiodic, and reversible Markov chain at equilibrium, and let $f\colon [d] \to [0,\ 1]$ be a function. If $\mu = \mathbb{E}_\pi[f]$ is the equilibrium expectation of f, then*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} f(X_n) - \mu\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\gamma_0}{2(2-\gamma_0)} \cdot \epsilon^2 N\right),$$

where $\gamma_0 = \min(1-\lambda_2,\ 1)$.

We prove adaptive rates.

Theorem 5 (Function-dependent Hoeffding bound). *With notation as above,*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} f(X_n) - \mu\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\gamma_f}{4\Lambda(\epsilon,\ \mu,\ \pi)} \cdot \epsilon^2 N\right),$$

where, letting $\nu = \min(\mu,\ 1-\mu)$,

$$\Lambda(\epsilon,\ \mu,\ \pi) = \log\left(\frac{4}{\nu\sqrt{\pi_{\min}}\epsilon^2}\right).$$

Furthermore, this holds even if the chain is not at equilibrium.

Examples and simulations

Definition (Lazy random walk on C_{2d})**.** The lazy random walk on the cycle graph with $2d$ vertices, C_{2d}, updates at each step according to the following rule:

- With probability $\frac{1}{2}$, stay at the current location.
- Otherwise, with probability $\frac{1}{2}$, move to the next node in clockwise order.
- Otherwise, move to the previous node in clockwise order.

We view the states in this Markov chain as indexed by integers in $\{0, \ldots, 2d-1\}$. For this chain, we have

time until $d_{\mathrm{TV}}(\pi_n,\ \pi) \leq \delta$ is on the order of $d^2 \log(1/\delta)$.

Example (Parity function)**.** Let f be the parity function defined by

$$f(i) = \begin{cases} 1 & \text{if } i \text{ is odd,} \\ 0 & \text{otherwise.} \end{cases}$$

Since both neighbors of any vertex have the opposite of its parity, it is easy to see that

$$\mathbb{E}[f(X_1) \mid X_0 = i] = \frac{1}{2},$$

so the function mixes in a single step.

Example (Trigonometric functions)**.** For $0 < j < 2d$, the trigonometric functions

$$g_j(i) = \frac{1 + \cos\left(\frac{\pi j i}{d}\right)}{2}$$

have

$$\gamma_{g_j} = \frac{1 - \cos\left(\frac{\pi j}{d}\right)}{2}.$$

Therefore, when $j = d \pm c$ for some constant $c > 0$, the function absolute spectral gap is on the order of a constant, and the chain mixes with respect to g_j in constant time.

Example (Random binary functions)**.** Consider a random binary function obtained by sampling $f(i) \sim \mathrm{Bern}(1/2)$ iid for each $i \in \{0, \ldots, 2d-1\}$. With probability $\geq 1 - \frac{\delta}{128\sqrt{d \log d}}$, we have that for any constant $0 < \delta < 1$,

time until $|\mathbb{E}[f(X_n)] - \mu| \leq \delta$ is at most on the order of $d \log^2 d$.