Weak Convergence of a Metropolis Algorithm for Bimodal Target Distributions

Michaël Lalancette[1] Mylène Bédard[2]

[1]Department of Statistical Sciences, University of Toronto,
lalancette@utstat.toronto.edu

[2]Département de Mathématiques et de Statistique, Université de Montréal,
bedard@dms.umontreal.ca

June 4, 2018

MCMC in a nutshell

- Probability measure Π
- Want to estimate (finite) expectation $\int h\, d\Pi$
- Idea: create a Markov chain $X_1, X_2, \ldots$ that has target distribution Π
- If the chain is well behaved (irreducible, aperiodic, Harris recurrent),

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \longrightarrow \int h\, d\Pi \quad \text{a.s.}$$

- In general, less efficient than iid sampling, but iid sampling is sometimes unrealizable

The Random Walk Metropolis algorithm

- RWM is a very simple MCMC algorithm
- π is a (possibly unnormalized) density on $\mathbb{R}^d$
- Initialize X_0
- Given X_t, independently generate Y *symmetric around 0*
- Set

$$X_{t+1} = \begin{cases} X_t + Y, & \text{w.p.} \quad \alpha_t \\ X_t, & \text{w.p.} \quad 1 - \alpha_t \end{cases}$$

- $\alpha_t = \min\left\{1, \frac{\pi(X_t+Y)}{\pi(X_t)}\right\}$
- Called *accept/reject step*

The Random Walk Metropolis algorithm

- Under mild conditions on instrumental distribution (distribution of increments Y),

$$\frac{1}{n}\sum_{t=1}^{n} h(X_t) \longrightarrow \frac{\int h(x)\pi(x)\,dx}{\int \pi(x)\,dx} \quad \text{a.s.}$$

- Usually $Y \sim N_d\left(0, \sigma^2 I_d\right)$
- In theory, works for any $\sigma > 0$...
- But small $\sigma \Rightarrow$ Small steps $\Rightarrow$ Slow exploration
- And large $\sigma \Rightarrow X_t + Y$ far from the mode $\Rightarrow \frac{\pi(X_t+Y)}{\pi(X_t)}$ small $\Rightarrow$ Most steps are rejected $\Rightarrow$ Slow exploration
- Must choose σ carefully

Optimal scaling

- [Roberts et al., 1997] proposed optimal scaling (in simplified framework)
- Assume $\pi(x) = \prod_{i=1}^{d} f(x_i)$
- Let $\sigma = \frac{\ell}{\sqrt{d}}$ (Large dimension $\Rightarrow$ Small steps)
- Let $\mathrm{X}^{(d)}(t)$ be Markov chain obtained (with step size depending on dimension d)
- And $X_j^{(d)}(t)$ its j-th component
- Smaller steps $\Rightarrow$ More steps needed, so accelerate chain: $Z_j^{(d)}(t) = X_j^{(d)}(\lfloor dt \rfloor)$
- Let $d \to \infty$

Accelerating the Markov chain

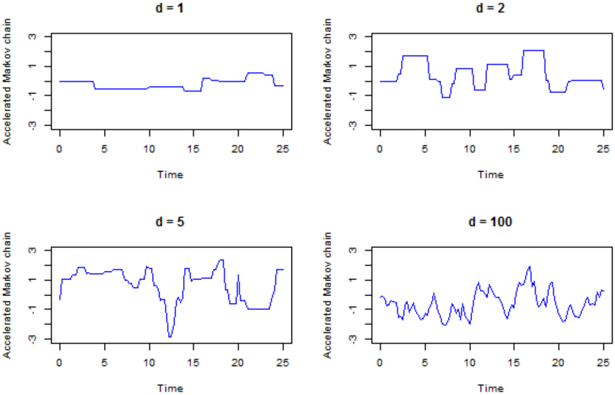

Figure: Trace of the accelerated first component of the Markov chain, $Z_1^{(d)}$, for different values of d. The target distribution is multivariate standard Gaussian.

Optimal scaling

- Find that $Z_j^{(d)}$ converges (weakly in Skorokhod topology) to Langevin diffusion with *speed measure* $v(\ell) = 2\ell^2\Phi\left(-\frac{\ell\sqrt{B}}{2}\right)$ and stationnary distribution f
- B is unknown parameter of f
- Find a value $\hat{\ell}$ that maximizes $v(\ell)$
- $\hat{\ell}$ is the only value that makes the asymptotic acceptance probability 0.234 (regardless of B)!
- So just simulate a few short runs and tune σ so that acceptance rate is roughly 0.234

The problem with bimodal distributions

- Suppose the target density π (on $\mathbb{R}^d$) has two distinct modes with a "hole" in between (low density region)
- Optimal scaling strategy tends to favor local exploration (Relatively small σ)
- Small steps $\Rightarrow$ Almost impossible to cross the hole in 1 step
- Almost impossible to accept steps into the hole ($\frac{\pi(X_t+Y)}{\pi(X_t)}$ very small)
- Chain gets stuck in a mode and never explores the other one

Solution: A new instrumental distribution

- Previously, $Y \sim N_d\left(0, \sigma^2 I_d\right)$
- Now, let

$$Y \sim \begin{cases} N_d\left(0, \sigma^2 I_d\right), & \text{w.p.} \quad 1-p \\ \mathcal{D}, \quad \text{w.p.} \quad p \end{cases},$$

  where $\mathcal{D}$ is any distribution on $\mathbb{R}^d$ *symmetric around 0*, $p \in (0,1)$
- In practice, choose $\mathcal{D}$ to favor switching modes
- Turns the "slowly mixing" chain into a "rapidly mixing" chain ([Guan and Krone, 2007])

1. Random Walk Metropolis algorithm

2. Weak convergence of our algorithm
    - Framework
    - The limiting processes

3. Summary and Future work

Framework

- Assume $\pi(x) = f_1(x_1) \prod_{i=2}^{d} f(x_i)$
- Instrumental distribution $\mathcal{D}$: $Y_1 \sim \mathcal{D}_1,\ Y_2, ..., Y_d \sim \mathrm{N}\left(0, \sigma^2\right)$, all independent
- Scale both $\sigma = \frac{\ell}{\sqrt{d}}$ and $p = 1 \wedge \frac{\beta}{d}$, but not $\mathcal{D}_1$
- Accelerate the chain by a factor of d (like before)
- The one-dimensional accelerated processes $Z_j^{(d)}$ weakly converge

The Langevin diffusion with Metropolis jumps

- Z_L and Z_{LM} are Langevin diffusions with speed measure $v(\ell)$ and stationnary distributions f and f_1
- At random times $T_1, T_2, \dots$ generated by a Poisson process of rate β, generate $Y(T_i) \stackrel{iid}{\sim} \mathcal{D}_1$
- At time T_i, Z_{LM} jumps by $Y(T_i)$ with probability $\alpha(\ell, Z_{LM}(T_i^-), Z_{LM}(T_i^-) + Y(T_i))$
- $$\alpha(\ell, x, y) = \Phi\left(\frac{\log\frac{f_1(y)}{f_1(x)} - \frac{\ell^2}{2}B}{\ell\sqrt{B}}\right) + \frac{f_1(y)}{f_1(x)}\Phi\left(\frac{-\log\frac{f_1(y)}{f_1(x)} - \frac{\ell^2}{2}B}{\ell\sqrt{B}}\right)$$

The limiting processes

Theorem

If the chain starts at stationnarity ($\mathbf{X}^{(d)}(0) \sim \pi$), then as $d \to \infty$,

$$Z_1^{(d)} \Rightarrow Z_{LM}, \quad \text{and for } j \geq 2, \quad Z_j^{(d)} \Rightarrow Z_L.$$

Here, $\Rightarrow$ represents *weak convergence in the Skorokhod topology*.

Optimizing the limiting processes

- Run classic RWM on π and tune σ so that acceptance rate is roughly 0.234
- Run RWM on f_1 with instrumental distribution $\mathcal{D}_1$ and acceptance probability $\alpha(\ell, x, y)$ to estimate acceptance probability of large steps, say λ
- Choose p, proportion of accepted large steps will be $\approx p\lambda$
- Number of times we switch modes $\approx np\lambda$

- Can work for much more general target distributions (align modes with rotation)
- Find objective way to choose p through non-asymptotics
- Can replace the large step distribution $\mathcal{D}$ by any *algorithm* ("something-inside-Metropolis")

Guan, Y. and Krone, S. (2007).
Small-world MCMC and convergence to multi-modal distributions: From slow mixing to fast mixing.
Ann. Appl. Probab., 17:284–304.

Roberts, G., Gelman, A., and Gilks, W. (1997).
Weak convergence and optimal scaling of random walk Metropolis algorithms.
Ann. Appl. Probab., 7:110–120.