On Single Variable Transformation Approach to Markov Chain Monte Carlo

Kushal K. Dey
Sourabh Bhattacharya

Indian Statistical Institute, Kolkata

Random Walk Metropolis Hastings (RWMH) - Algorithm

Algorithm

Suppose we are at $\mathbf{x}_n = (x_1, x_2, \cdots, x_d)$ *at the* n*th iteration.*

1. *Generate* $\epsilon_1, \epsilon_2, \cdots, \epsilon_d \sim q(.)$ *independently [* q *symmetric around 0 and of the form* $q(x, y) = g(x - y)$*].*
2. *Define*

$$y = (x_1 + \epsilon_1, x_2 + \epsilon_2, \cdots, x_d + \epsilon_d) \quad \alpha(\mathbf{x}_n, \epsilon) = \min\left\{1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x}_n)}\right\}$$

3. *Set* $\mathbf{x}_{n+1} = \left\{ \begin{array}{llll} \mathbf{y} & \text{with} & \text{prob.} & \alpha(\mathbf{x}_n, \epsilon) \\ \mathbf{x}_n & \text{with} & \text{prob.} & 1 - \alpha(\mathbf{x}_n, \epsilon) \end{array} \right\}$

A problem with RWMH- An Instance

- Assume that the target density π is the product density, $\pi = \prod_{i=1}^{d} f()$ of iid components f. Then the acceptance rule for RWMH comprises of the ratio

$$\frac{\pi(\mathbf{x} + \epsilon)}{\pi(x)} = \prod_{i=1}^{d} \frac{f(x_i + \epsilon_i)}{f(x_i)}$$

- If d is very large, then by chance, we may get some very small or large values of ϵ_i.
- This would result in certain very small values of $f(x_i + \epsilon_i)$ for some i and thereby drastically reduce the above ratio. So, the chain has the problem of remaining stuck at a point for a long time.

Additive Transformation based MCMC (TMCMC) - Algorithm (Dutta and Bhattacharya, 2013)

Algorithm

Suppose we are at $\mathbf{x}_n = (x_1, x_2, \cdots, x_d)$ *at the* n*th iteration.*

1. *Generate* $\epsilon \sim g(.)$ *on* $\mathbb{R}^+$.
2. *Select randomly one move type and define*

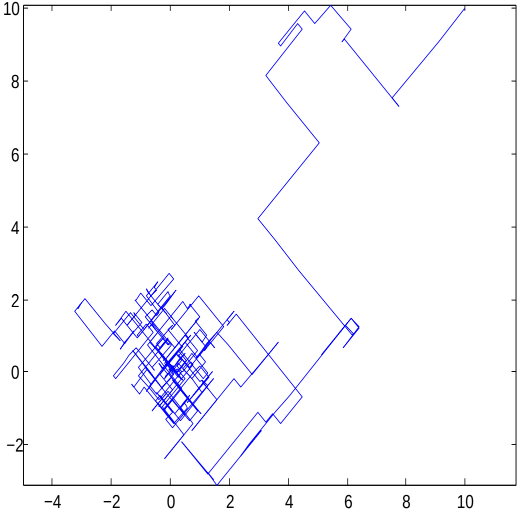
$$b_1, b_2, \cdots, b_d \overset{iid}{\sim} \text{DiscrUnif}\{-1, 1\}$$

$$\mathbf{y} = (x_1 + b_1\epsilon, x_2 + b_2\epsilon, \cdots, b_d\epsilon) \tag{1}$$

$$\alpha(\mathbf{x}, \epsilon) = \min\left\{1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x}_n)}\right\} \tag{2}$$

3. *Set* $\mathbf{x}_{n+1} = \left\{ \begin{array}{llll} \mathbf{y} & \text{with} & \text{prob.} & \alpha(\mathbf{x}_n, \epsilon) \\ \mathbf{x}_n & \text{with} & \text{prob.} & 1 - \alpha(\mathbf{x}_n, \epsilon) \end{array} \right\}$

A TMCMC chain for $\mathrm{N}(0, 1) \times \mathrm{N}(0, 1)$ chain



General State Space Markov Chain - Concepts

We state some notions of importance for general state space Markov chains.

- ψ irreducibility : A Markov chain on χ is said to be ψ-*irreducible* if there exists a measure ψ such that

$$\psi(\mathrm{A}) > 0 \implies \exists \mathrm{n} \quad \text{with} \quad \mathrm{P^n(x, A)} > 0 \qquad \forall \mathrm{x} \in \chi \tag{3}$$

- Small set: A set E is said to be *small* if there exists a $\mathrm{n} > 0$, $\delta > 0$ and some measure ν it satisfies the *Minorization condition*.

$$\mathrm{P^n(x, .)} > \delta\nu(.) \qquad \mathrm{x} \in \mathrm{E} \tag{4}$$

- Aperiodicity: A chain is called *aperiodic* if the gcd of all such n for **Eqn 4** holds is 1.

We state some notions of importance for general state space Markov chains.

- ψ irreducibility : A Markov chain on χ is said to be ψ-*irreducible* if there exists a measure ψ such that

$$\psi(\mathrm{A}) > 0 \implies \exists \mathrm{n} \quad \text{with} \quad \mathrm{P}^{\mathrm{n}}(\mathrm{x}, \mathrm{A}) > 0 \qquad \forall \mathrm{x} \in \chi \tag{3}$$

- Small set: A set E is said to be *small* if there exists a $\mathrm{n} > 0$, $\delta > 0$ and some measure ν it satisfies the *Minorization condition*.

$$\mathrm{P}^{\mathrm{n}}(\mathrm{x}, .) > \delta\nu(.) \qquad \mathrm{x} \in \mathrm{E} \tag{4}$$

- Aperiodicity: A chain is called *aperiodic* if the gcd of all such n for **Eqn 4** holds is 1.

General State Space Markov Chain - Concepts

We state some notions of importance for general state space Markov chains.

- ψ irreducibility : A Markov chain on χ is said to be ψ-*irreducible* if there exists a measure ψ such that

$$\psi(\mathrm{A}) > 0 \implies \exists \mathrm{n} \quad \text{with} \quad \mathrm{P}^{\mathrm{n}}(\mathrm{x}, \mathrm{A}) > 0 \qquad \forall \mathrm{x} \in \chi \tag{3}$$

- Small set: A set E is said to be *small* if there exists a $\mathrm{n} > 0$, $\delta > 0$ and some measure ν it satisfies the *Minorization condition*.

$$\mathrm{P}^{\mathrm{n}}(\mathrm{x}, .) > \delta\nu(.) \qquad \mathrm{x} \in \mathrm{E} \tag{4}$$

- Aperiodicity: A chain is called *aperiodic* if the gcd of all such n for **Eqn 4** holds is 1.

Ergodic properties of the Additive TMCMC chain

- By definition of the acceptance probability α, the TMCMC chain satisfies the detailed balance or the reversibility condition.
- The additive TMCMC chain is λ- irreducible and aperiodic. Also the minorization condition holds for dimension d,

$$\mathrm{P}^{2\mathrm{d}}(\mathrm{x}, \mathrm{A}) \geq \delta\nu(.) \qquad \mathrm{x} \in \mathrm{E}(\mathrm{Borelset})$$

- For a sub-exponential family, the geometric ergodicity holds for the Additive TMCMC chain

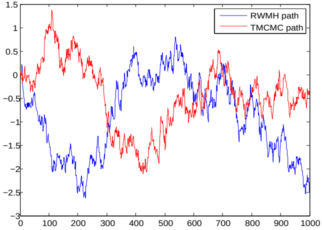
$$||\mathrm{P}^{\mathrm{n}}(\mathrm{x}, .) - \pi(.)||_{\mathrm{TV}} \leq \mathrm{M}(\mathrm{x})\rho^{\mathrm{n}} \qquad \forall \text{ large n} \tag{5}$$

- It ensures that CLT holds for the ergodic averages $\mathrm{S}_{\mathrm{n}}(\mathrm{g}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{g}(\mathrm{X}_{\mathrm{i}})$ of a function g evaluated at the visited states of the Markov Chain [Jarner and Hansen, 2000].

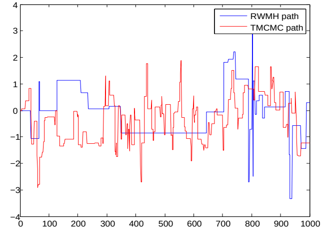
Optimal scaling of the TMCMC chain

- Tuning of associated parameters such as proposal variances is crucial for achieving efficient mixing but also quite difficult.
- Higher values of proposal variances lead to poor acceptance rates while lower values lead to higher ACF and slow movement of the chain thereby rendering the estimates to be locally concentrated.
- Adaptive MCMC algorithm uses techniques to learn better parameter values as it moves with little user intervention.
- A non-adaptive mechanism to deal with this problem is to find a suitable optimizing criterion whcih would lead to an optimal value of proposal variance.

Need for optimizing proposal variance



(a) small prop var sample path



(b) large prop variance sample path

Optimal scaling results in i.i.d case- TMCMC

- The discrete time generator function corresponding to V, a function of the first component only, underlying the TMCMC process for normal proposals $\epsilon \sim \mathrm{N}(0, \frac{\mathrm{l}^2}{\mathrm{d}})$ converges to (as $\mathrm{d} \to \infty$) .

$$\mathrm{GV}(\mathrm{x}) = \mathrm{h}(\mathrm{l}) \left[\frac{1}{2} (\log \mathrm{f})^{'}(\mathrm{x}_1) \mathrm{V}^{'}(\mathrm{x}_1) + \frac{1}{2} \mathrm{V}^{''}(\mathrm{x}_1) \right] \tag{6}$$

where

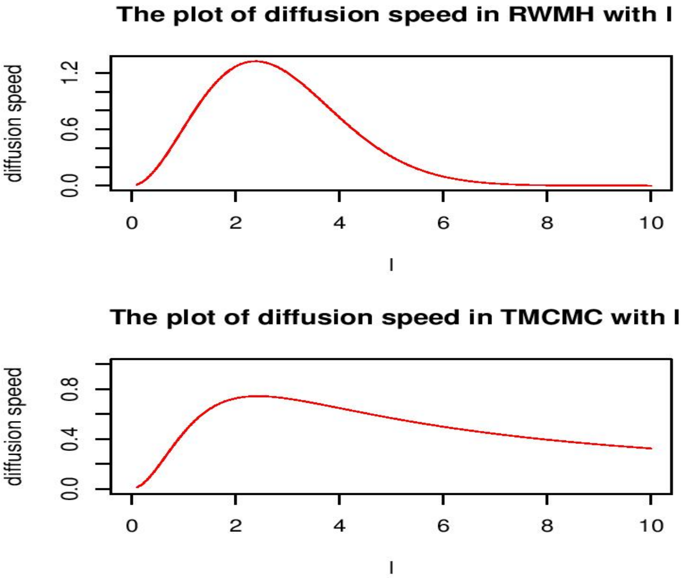
$$\mathrm{h}(\mathrm{l}) = 4\mathrm{l}^2 \int_0^\infty \mathrm{z}^2 \Phi \left(- \frac{\sqrt{\mathrm{z}_1{}^2 \mathrm{l}^2 \mathbb{I}}}{2} \right) \tag{7}$$

- The function $\mathrm{h}(\mathrm{l})$ corresponds to the limiting speed of the process is optimized with respect to l to derive the best scaling.

Optimal scaling results

- We optimize the diffusion speed to get a value of optimal scaling and calculate optimal acceptance rate.
- For RWMH, the diffusion process is Langevin and the optimal acceptance rate corresponding to maximizing the diffusion speed is 0.234.
- In our case, the optimal acceptance rate has been found out to be 0.439 with the optimizing value l_{opt} being very close to that of RWMH (around $\frac{2.4}{\sqrt{I}}$).

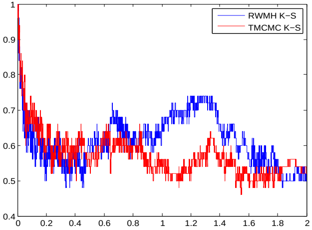
Graphical interpretation of diffusion speed (General)



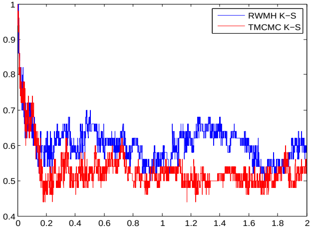
Simulation analysis

Dim	Scaling \ Test	Acceptance rate(%)	
		RWMH	TMCMC
2	2.4	34.9	44.6
	6	18.66	29.15
	10	3.83	12.36
5	2.4 (opt)	28.6	44.12
	6	2.77	20.20
	10	0.45	12.44
10	2.4 (opt)	25.6	44.18
	6	1.37	20.34
	10	0.03	7.94
100	2.4 (opt)	23.3	44.1
	6	0.32	20.6
200	2.4 (opt)	23.4	44.2
	6	0.33	20.7

K-S test plot between RWMH and TMCMC - 100D with scale 2.4 and 4



(c) K-S comparison 1



(d) K-S comparison 2

Summary and Final Remarks

- TMCMC is simple to interpret and does not depend heavily on the target density, and additionally has much lesser computational burden and time complexity.
- Under sub-exponential target density with some regularity conditions on the proposal distribution structure, the TMCMC algorithm is geometrically ergodic.
- TMCMC has a higher acceptance rate of 0.439 corresponding to 0.234 for the RWMH algorithm. As observed, our algorithm is more robust to change of scale and across dimensions. But the mixing or diffusion speed of RWMH is higher.
- The KS test comparison with the target density shows that for high dimension, TMCMC has almost uniformly lower KS statistic value compared to the RWMH.

THANK YOU