On Single Variable Transformation Approach to Markov Chain Monte Carlo

Kushal K. Dey$^{\dagger+}$, Sourabh Bhattacharya*

† University of Chicago, IL
* Indian Statistical Institute, Kolkata
$+$ Corresponding author: kkdey@uchicago.edu

1 Introduction

In today's times, Monte Carlo methods have everyday use in Statistics and other disciplines like Computer Science, Systems Biology and Astronomy. This technique of generating random samples even from very high dimensional spaces involving very complicated data likelihoods and posterior distributions has simplified many pressing real life problems in recent times. In particular, Bayesian computation, simulation from complex posterior distribution and asymptotics of Bayesian algorithms have benefited a lot from this mechanism (see Gelfand and Smith [GS90],Tierney [Tie94], Gilks *et al* [GS96]). A very standard approach of simulating from multivariate distributions is to use the Random Walk Metropolis-Hastings (MH) algorithm [Has70][MRR53]. The convergence and optimal scaling of this algorithm has been extensively studied [RGG97]. However, there are obvious scopes for improving upon this algorithm, pertaining mainly to the choice of proper proposal distribution and the time-complexity associated with the process. However, there are certain glaring problems that one may encounter while using RWMH. For very high dimensional datasets, convergence of RWMH to the target density is pretty slow and one requires too many iterations, largely due to the fact that in RWMH, we need to update each co-ordinate at a time and this may lead to very small acceptance probability for high dimensions. The TMCMC algorithm proposed in [DB11] tries to address this problem. It uses simple deterministic transformations using a single random variable and a single proposal density chosen appropriately. In this paper, we primarily study one version, termed as the Additive TMCMC method and typically deal with the ergodic behavior of the chain in high dimensions. Our aim is to present a comparative study of the Additive TMCMC and the standard RWMH algorithms with respect to the ergodic behavior of the two.

This paper is organized as follows. In **Section 2**, we shall present the Additive TMCMC algorithm and discuss the intuition behind this algorithm. In **Section 3**, we discuss some theoretical results regarding the ergodic behavior of the chain. **Section 4** focuses on how to optimally select the proposal density for the chain when the target density has a product structure. In **Section 5**, we present the comparative simulation study of the Additive TMCMC and the RWMH chains and analyze the results.

2 Algorithm

We first briefly describe how additive TMCMC works. We explain it for the bivariate case – the multivariate extension would analogously follow. Suppose we start at a point (x_1, x_2). We generate an $\varepsilon > 0$ from some pre-specified proposal distribution q defined on $\mathbb{R}^+$. Then in additive TMCMC we have the following four possible transitions

$$\begin{aligned}
(x_1, x_2) &\rightarrow (x_1 + \varepsilon, x_2 + \varepsilon) \\
(x_1, x_2) &\rightarrow (x_1 + \varepsilon, x_2 + \varepsilon) \\
(x_1, x_2) &\rightarrow (x_1 + \varepsilon, x_2 + \varepsilon) \\
(x_1, x_2) &\rightarrow (x_1 + \varepsilon, x_2 + \varepsilon)
\end{aligned} \tag{1}$$

This means we are moving along two lines in each transition from the point (x_1, x_2) , one parallel to the line $y = x$ and the other parallel to the direction $y = -x$. Each of the four transitions described above are indexed as I_k for kth transition, where k may vary from 1 to 4 in the bivariate case, and in general from 1 to 2^d in $\mathbb{R}^d$. We choose with equal probability. As with the standard RWMH case, we do attach some probabilities with accepting/rejecting the proposed move such that the reversibility condition is satisfied thereby guaranteeing convergence. Formally, the algorithm may be presented as follows.

Algorithm 2.1. *Suppose we are at $\boldsymbol{x}_n = (x_1, x_2, \cdots, x_d)$ at the nth iteration.*

1. *Generate $\varepsilon \sim g(.)$ on $\mathbb{R}^+$.*

2. *Select randomly one move type and define*

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

$$\boldsymbol{y} = (x_1 + b_1\varepsilon, x_2 + b_2\varepsilon, \cdots, b_d\varepsilon) \tag{2}$$

$$\alpha(\boldsymbol{x}, \varepsilon) = min\left\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}_n)}\right\} \tag{3}$$

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

Now we discuss why intuitively we feel this algorithm is a better option compared to the RWMH algorithm. We tested using simulation experiments that our algorithm requires less computational time to run in MATLAB R2013b compared to RWMH (see **Fig 1**).

But this is not the major issue. In a standard RWMH algorithm in d dimensions, we need to generate d many ε_i's, for $i \in \{1, 2, \cdots, d\}$. 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

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

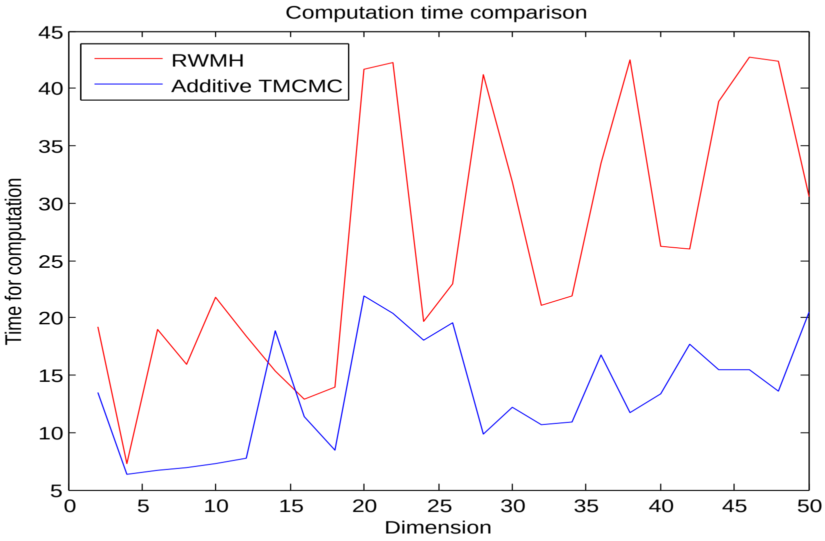


Figure 1: *Computation time (in MATLAB R2013b) of one run of 1,00,000 iterations with RWM and TMCMC algorithms corresponding to dimensions varying from 2 to 50 with target density being product of* $N(0,5)$ *and the proposal density for additive TMCMC being* $TN_{>0}(0,1)$ *(truncated* $N(0,1)$ *left truncated at* 0*) and for RWMH proposal, every component has* $N(0,1)$ *distribution . It is observed that TMCMC has consistently less computation time compared to RWM specially for higher dimensions.*

If d is very large, then by chance, we may get some very small or large values of $\varepsilon_i \sim q(.)$ (note that 5% observations are expected to lie outside the 95% confidence region and these are the points that are problematic). This would result in certain very small values of $f(x_i + \varepsilon_i)$ 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. Note that the additive TMCMC uses only one ε to update all the co-ordinates using sign change and this counters the above problem. So, we can expect a much higher acceptance rate for the additive TMCMC over the RWMH algorithm. But there are two pertinent questions here. Firstly, how much can we improve on the RWMH algorithm in terms of the acceptance rate? Secondly, how would the sample we get using the TMCMC method compare to the RWMH algorithm in terms of the convergence of the iterates to the target density and the mixing among the iterates once the target is attained. We address the first issue in **Section 4** and the second in **Section 5**.

3 Ergodic Properties of the Additive TMCMC

In case of Markov chains on discrete spaces, there is a well-established notion of irreducibility. However, on general state spaces, such a notion no longer works. This is why we define ψ irreducibility. A Markov chain is said to be ψ*-irreducible* if there exists a measure ψ such that

$$\psi(A) > 0 \implies \exists n \quad with \quad P^n(x,A) > 0 \qquad \forall x \in \chi \tag{4}$$

where χ is the state space of the Markov chain (in our case, it would most often be $\mathbb{R}^d$ for some d). To talk about the convergence of the process, we must ensure that it is λ-irreducible, where λ is the Lebesgue measure. We also need additional concepts of aperiodicity and *small* sets. A set E is said to be *small* if there exists a $n > 0$, $\delta > 0$ and some measure ν such that

$$P^n(x,.) > \delta\nu(.) \qquad x \in E \tag{5}$$

A chain is called *aperiodic* if the gcd of all such n for **Eqn 5** holds is 1. All these concepts of λ-irreducibility, aperiodicity and small sets are very important for laying the basic foundations of stability. The following theorem due to Dutta and Bhattacharya [DB11] establishes these properties for the additive TMCMC chain

Result 3.1. *Let π be a continuous target density which is bounded away from 0 on $\mathbb{R}^d$. Also, let the proposal density q be positive on all compact sets on $\mathbb{R}^+$. Then, the every non-empty bounded set in $\mathbb{R}^d$ is small, and this can be used to show that the chain is both λ-irreducible and also aperiodic.*

A proof of this result can be found in Dutta and Bhattacharya [DB11] along with a graphical interpretation. In fact in that paper, a stronger result has been shown that for any $n > d$ (d represents the dimensionality of the state space), minorization condition is satisfied.

$$P^n(x,A) \geq \delta\lambda(.) \qquad x \in E\ (bdd.\ Borel\ set)$$

where λ is the Lebesgue measure and he explicit form of the δ depends on the bounds on the proposal density. The λ irreducibility follows trivially from this minorization condition. The aperiodicity follows because the above result is true for all $n > d$ and the gcd of such n would be 1.

Let P be the transition kernel of a ψ-irreducible, aperiodic Markov chain with the stationary distribution π, then the chain is geometrically ergodic if $\exists$ a function $V \geq 1$ and finite at least at one point, and also constants ρ and M, so that

$$||P^n(x,.) - \pi(.)||_{TV} \leq M.V(x)\rho^n \quad \forall n \geq 1 \tag{6}$$

where $||\nu||_{TV}$ denotes the *total variation norm.*

$$||\nu||_{TV} = \sup_{g:|g|\leq V} \nu(g)$$

The reason for preferring geometric ergodicity is that under this condition, one can apply Central Limit Theorem to a wide class of functions of the Markov Chain, and hence, one can also speak about the stability of these ergodic estimates (see Roberts, Gelman and Gilks [RGG97]). A very standard way of checking geometric ergodicity is a result that involves the Foster-Lyapunov drift criteria. P is said to have a geometric drift to a set E if there is a function $V \geq 1$, finite for at least one point and constants $\lambda < 1$ and $b < \infty$ so that

$$PV(x) \leq \lambda.V(x) + b1_E(x) \tag{7}$$

where $PV(x) = \int V(y).P(x,y)dy$ is basically the expectation of V after one transition given that one starts at the point x. Theorems 14.0.1 and 15.0.1 in Meyn and Tweedie [MT93] establish the fact that if P has a geometric drift to a small set E, then under certain regularity conditions, P is π almost everywhere geometric ergodic and the converse is also true.

The first result we present is basically adaptation of a result due to (Mengerson and Tweedie, 1996). We now show a sufficient condition that would ensure that **Eqn 7** holds.

Lemma 3.1. *If $\exists\, V$ such that $V \geq 1$ and finite on bounded support, such that the following holds*

$$limsup_{|x|\to\infty} \frac{PV(x)}{V(x)} < 1 \tag{8}$$

$$\frac{PV(x)}{V(x)} < \infty \qquad \forall x \tag{9}$$

*Then this V satisfies the geometric drift condition in **Eqn 7** and hence the chain must be geometric ergodic. Also, if for some V finite , the geometric drift condition is satisfied, then the above condition must also hold true.*

Result 3.2. *If π the target density is sub-exponential and has contours that are nowhere piecewise parallel to $\{x : |x_1| = |x_2| = \cdots = |x_d|\}$, then the additive TMCMC chain satisfies geometric drift if and only if*

$$\liminf_{\|x\|\to\infty} Q^{(1)}(x, A^{(1)}(x)) > 0. \tag{10}$$

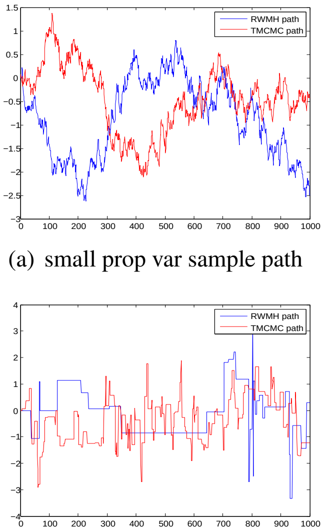
This implies that this condition is all we require to show geometric ergodicity for the additive TMCMC chain.

A proof of this result is given in Dey and Bhattacharya [DB13a]. A similar result holds true for the RWMH algorithm as well (see Jarner and Hansen [JH00] and Roberts and Tweedie [RT96]) except that there we do not need the constraint that the contours are not piecewise parallel to $\{x : |x_1| = |x_2| = \cdots = |x_d|\}$, but this is true for most densities we commonly encounter. Even if this condition is not satisfied, we can still show geometric ergodicity for a modified TMCMC chain with moves from $(x_1, x_2, \cdots, x_d)$ to $(x_1 + b_1c_1\varepsilon_1, x_2 + b_2c_2\varepsilon_2, \cdots, x_d + b_dc_d\varepsilon_d)$ where c_i's are some positive scalars not all equal.

4 Optimal Scaling of Additive TMCMC

In this section, we shall restrict our focus on target densities that are product of iid components $\pi = \prod_{i=1}^{d} f$ and the proposal density for ε is given by $TN_{>0}(0, \frac{l^2}{d})$, where l is called the scaling term of the proposal. This section would be dedicated to finding out the optimal value of this scaling l and determining the limiting expected acceptance rate of the additive TMCMC under the optimal scaling scenario. If the variance of the proposal density is very small, then jumps will be of smaller magnitude and this would mean the Markov chain would take lot more time to traverse the entire space and in the process, the convergence rate would be pretty low. On the other hand, if the variance is very large, then our algorithm will reject too many of the moves. An instance of this argument is depicted in **Fig 2**.

There is an extensive theory on optimal scaling of RWMH chains (see Beskos, Roberts and Stuart [BRS09], Bedard [Bed09] [Bed07], Neal and Roberts [NR06], Roberts, Gelman and Gilks [RGG97]). The magic number for RWMH has been the optimal acceptance rate value of 0.234, which has been achieved through maximization of speed of the process for a wide range of distributions- i.i.d set up and some special class of independent but non-i.i.d. component set up as well. We have employed an analogous speed where we have optimized the diffusion speed of our process to get optimal acceptance rate. We present



(a) small prop var sample path

(b) large prop variance sample path

Figure 2: *The graphical representation of a co-ordinate for a 5-dimensional chain with target density being product of $N(0,1)$ and the values of the scaling factor l for the two cases taken to be $l = 0.8$ and $l = 8$ respectively for the two scenarios a) and b) depicted in the graph*

a rough sketch of our approach here, for detailed analysis we refer the reader to Dey and Bhattacharya [DB13b].

We assume that f is Lipschitz continuous and satisfies the following conditions

$$(C1) \quad E\left[\left\{\frac{f^{'}(X)}{f(X)}\right\}^{8}\right] = M_1 < \infty \tag{11}$$

$$(C2) \quad E\left[\left\{\frac{f^{''}(X)}{f(X)}\right\}^{4}\right] = M_2 < \infty \tag{12}$$

We define ${U_t}^d = X^d_{[dt],1}$, the sped up first component of the actual Markov chain. Note that this process makes a transition at an interval of $\frac{1}{d}$. As we set $d \to \infty$, meaning that as the dimension of the space blows to ∞, the sped up additive TMCMC process essentially converges to a continuous time Diffusion process.

For our purpose, we define the discrete time generator of the TMCMC approach, as

$$\begin{aligned} G_d V(x) &= \frac{d}{2^d} \sum_{\left\{\begin{array}{l} b_i \in \{-1,+1\} \\ \forall i = 1,\ldots,d \end{array}\right\}} \int_0^\infty \left[\left(V\left(x_1 + b_1\varepsilon, \ldots, x_d + b_d\varepsilon\right) - V\left(x_1,\ldots,x_d\right)\right)\right. \\ &\qquad \left. \times \left(\min\left\{1, \frac{\pi(x_1 + b_1\varepsilon,\ldots,x_d + b_d\varepsilon)}{\pi(x_1,x_2,\ldots,x_d)}\right\}\right)\right] q(\varepsilon) d\varepsilon. \end{aligned} \tag{13}$$

In the above equation, we may assume that V belongs to the space of inifinitely differentiable functions on compact support (see, for example, [Bed07]) for further details).

Note that this function is measurable with respect to the Skorokhod topology and we can treat G_d as a continuous time generator that has jumps at the rate d^{-1}. Given our restricted focus on a one dimensional component of the actual process, we assume V to be a function of the first co-ordinate only. Under this assumption, the generator defined in (13) is a function of only ε and b_1, and can be rephrased as

$$\begin{aligned} G_d V(x) &= \frac{d}{2}\int_0^\infty \sum_{b_1\in\{-1,+1\}} \Bigg[\Big(V(x_1+b_1\varepsilon)-V(x_1)\Big) \\ &\qquad \times E_{b_2,\ldots,b_d}\left(\min\left\{1,\frac{\pi(x_1+b_1\varepsilon,\ldots,x_d+b_d\varepsilon)}{\pi(x_1,\ldots,x_d)}\right\}\right)\Bigg] q(\varepsilon)d\varepsilon, \end{aligned} \tag{14}$$

where $E_{b_2,\ldots,b_d}$ is the expectation taken conditional on b_1 and ε.

First we show that the quantity $G_dV(x)$ is a bounded quantity.

$$\begin{aligned} G_dV(x) &\leq dE_{\{b_1,\varepsilon\}}\left[V(x_1+b_1\varepsilon)-V(x_1)\right] \\ &= dV^{'}(x_1)E_{\{b_1,\varepsilon\}}(b_1\varepsilon)+\frac{d}{2}V^{''}(x_1^*)E_{\{b_1,\varepsilon\}}(\varepsilon^2) \\ &\leq \ell^2 M, \end{aligned} \tag{15}$$

where x_1^* lies between x_1 and $x_1+b_1\varepsilon$ and M is the maximum value of $V^{''}$.

We derive the limit of $G_dV(x)$ as $d\to\infty$ that will give us the infinitesimal generator of the associated diffusion process for the additive TMCMC chain. It can be shown that

Proposition 4.1. *If $X\sim N(\mu,\sigma^2)$, then*

$$E\left[\min\left\{1,e^X\right\}\right]=\Phi\left(\frac{\mu}{\sigma}\right)+e^{\left\{\mu+\frac{\sigma^2}{2}\right\}}\Phi\left(-\sigma-\frac{\mu}{\sigma}\right), \tag{16}$$

where Φ is the standard Gaussian cdf.

Using this proposition, we can write

$$\begin{aligned} & E\Bigg|_{b_1\varepsilon}\left[\min\left\{1,\frac{\pi(x_1+b_1\varepsilon,\ldots,x_d+b_d\varepsilon)}{\pi(x_1,\ldots,x_d)}\right\}\right] \\ =\quad & \Phi\left(\frac{\eta(x_1,b_1,\varepsilon)-\frac{(d-1)\varepsilon^2}{2}\mathbb{I}}{\sqrt{(d-1)\varepsilon^2\mathbb{I}}}\right)+e^{\eta(x_1,b_1,\varepsilon)}\Phi\left(-\sqrt{(d-1)\varepsilon^2\mathbb{I}}-\frac{\eta(x_1,b_1,\varepsilon)-\frac{(d-1)\varepsilon^2}{2}\mathbb{I}}{\sqrt{(d-1)\varepsilon^2\mathbb{I}}}\right) \\ =\quad & \mathbb{W}(b_1\varepsilon,x_1). \end{aligned} \tag{17}$$

Note that using Taylor series expansion around x_1, we can write (**??**) as

$$\eta(x_1,b_1,\varepsilon)=b_1\varepsilon\left[\log f(x_1)\right]^{'}+\frac{\varepsilon^2}{2}\left[\log f(x_1)\right]^{''}+b_1\frac{\varepsilon^3}{3!}\left[\log f(\xi_1)\right]^{'''}, \tag{18}$$

where ξ_1 lies between x_1 and $x_1 + b_1\varepsilon$. Again re-writing $b_1\varepsilon$ as $\frac{\ell}{\sqrt{d}} z_1^*$, where z_1^* follows a $N(0,1)$ distribution, η and $\mathbb{W}$ can be expressed in terms of ℓ and z_1^* as

$$\eta(x_1, z_1^*, d) = \frac{\ell z_1^*}{\sqrt{d}} [\log f(x_1)]^{'} + \frac{\ell^2 z_1^{*2}}{2!d} [\log f(x_1)]^{''} + \frac{\ell^3 z_1^{*3}}{3! d^{\frac{3}{2}}} [\log f(\xi_1)]^{'''} \tag{19}$$

and

$$\mathbb{W}(z_1^*, x_1, d) = \Phi\left(\frac{\eta(x_1, z_1^*, d) - \frac{z_1^{*2}\ell^2}{2}\mathbb{I}}{\sqrt{z_1^{*2}\ell^2\mathbb{I}}} \right) + e^{\eta(x_1, z_1^*, d)} \Phi\left(\frac{-\frac{z_1^{*2}\ell^2\mathbb{I}}{2} - \eta(x_1, z_1^*, d)}{\sqrt{z_1^{*2}\ell^2\mathbb{I}}} \right). \tag{20}$$

The last line follows as the expression $\eta(x_1, b_1, \varepsilon)$ depends on b_1 and ε only through the product $b_1\varepsilon$.

Now we consider the Taylor series expansion around x_1 of the term

$$\begin{aligned} & dE_{z_1^*}\left[\left(V\left(x_1 + \frac{z_1^*\ell}{\sqrt{d}} \right) - V(x_1) \right) \mathbb{W}(z_1^*, x_1, d) \right] \\ = \; & dE_{z_1^*}\left[\left\{ V^{'}(x_1)\frac{z_1^*\ell}{\sqrt{d}} + \frac{1}{2}V^{''}(x_1)\frac{z_1^{*2}\ell^2}{d} + \frac{1}{6}V^{'''}(\xi_1)\frac{z_1^{*3}\ell^3}{d^{\frac{3}{2}}} \right\} \mathbb{W}(z_1^*, x_1, d) \right]. \end{aligned} \tag{21}$$

From (20) it is clear that $\mathbb{W}(z_1^*, x_1, d)$ is continuous but not differentiable at the point 0. Using Taylor series expansion of the terms $\Phi\left(\frac{\eta(x_1, z_1^*, d) - \frac{z_1^{*2}\ell^2}{2}\mathbb{I}}{\sqrt{z_1^{*2}\ell^2\mathbb{I}}} \right)$, $e^{\eta(x_1, z_1^*, d)}$ and $\Phi\left(\frac{-\frac{z_1^{*2}\ell^2\mathbb{I}}{2} - \eta(x_1, z_1^*, d)}{\sqrt{z_1^{*2}\ell^2\mathbb{I}}} \right)$ about $\eta = 0$, we get the expression of $G_d(V(x)$ to be

$$G_d V(x) \approx V^{'}(x_1)\frac{\ell^2}{2}[logf(x_1)]^{'} E_{z_1^*}\left[z_1^{*2}\mathscr{V}(z_1^*) \right] + \frac{1}{2}V^{''}(x_1)\ell^2 E_{z_1^*}\left[z_1^{*2}\mathscr{V}(z_1^*) + \mathscr{O}(d^{-\frac{1}{2}}) \right]. \tag{22}$$

where

$$\mathscr{V}(z_1^*) \to 2\Phi\left(-\frac{|z_1^*|\ell\sqrt{\mathbb{I}}}{2} \right) = 2\left[1 - \Phi\left(\frac{|z_1^*|\ell\sqrt{\mathbb{I}}}{2} \right) \right]. \tag{23}$$

The infinitesimal generator $GV(x)$ obtained as the limit of the $GV_d(x)$ has therefore a simpler form

$$GV(x) = h(l)\left[\frac{1}{2}(logf)^{'}(x_1)V^{'}(x_1) + \frac{1}{2}V^{''}(x_1) \right] \tag{24}$$

This is the form of the generator for a Langevin diffusion process with

$$h(l) = 4l^2 \int_0^\infty z^2 \Phi\left(-\frac{\sqrt{{z_1}^2 l^2 \mathbb{I}}}{2} \right) \tag{25}$$

The function h is called the diffusion speed and we maximize this quantity with respect to l to derive the optimal scaling. For our case, $l_{opt} = \frac{2.4}{\sqrt{I}}$ and we put this value in the formula for asymptotic expected acceptance rate to get

$$\alpha_{opt} = 4\int_0^\infty \Phi\left(-\frac{|u|\ell_{opt}\sqrt{\mathbb{I}}}{2}\right)\phi(u)du. \tag{26}$$

For RWMH too, the diffuion process is Langevin but the form of the diffusion speed is slightly different (see Roberts, Gelman and Gilks [RGG97]).

$$h_{RWMH}(l) = 2l^2\Phi\left(\frac{-l\sqrt{I}}{2}\right) \tag{27}$$

It was noted in [RGG97] that the limiting expected acceptance rate corresponding to optimal scaling in RWMH (optimal scaling is similar to that of additive TMCMC actually $l_{opt} = \frac{2.4}{\sqrt{I}}$) is 0.234, while for that for the optimal scaling in additive TMCMC is 0.439 which is almost twice as that of RWMH. This shows that indeed additive TMCMC has much higher expected acceptance rate compared to RWMH. The graphs of the diffusion speeds over different l for the additive TMCMC and for standard RWMH is presented in **Fig 3**.

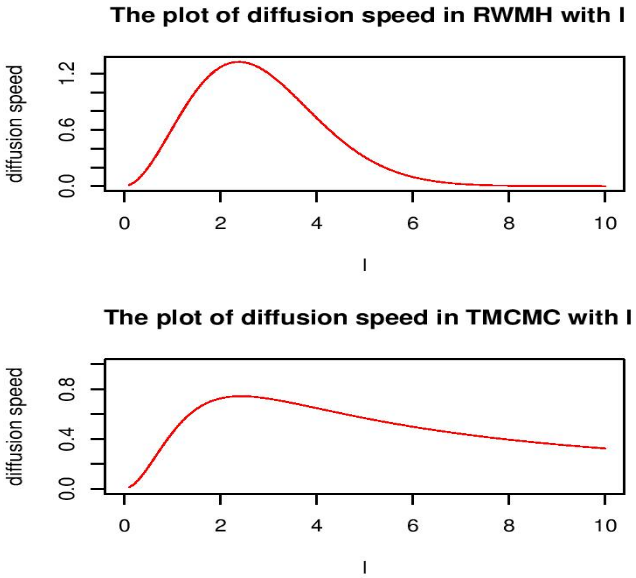


Figure 3: *The plot of the diffusion speed with respect to the scaling factor l for the RWMH and the additive TMCMC chains.*

Note that the diffusion speed at l_{opt} is higher for RWMH compared to additive TMCMC implying that once stationarity is reached, there will be faster mixing among the iterates in RWMH compared to additive TMCMC, although an interesting observation is that even if l deviates slightly from l_{opt}, the diffusion speed in additive TMCMC case remains much more stable while that for RWMH drops very fast. However, in all the calculations we have done so far and in the consideration of the diffusion speed and its implications, we must keep in mind an inherent assumption, we assume that the process is in stationarity. The major question to address now is that which chain has faster convergence to stationarity and we address this in the next section by running some simulations for both the chains and measuring the correspondence of the empirical distributions at each time point for RWMH and additive TMCMC with respect to that of the target density.

5 Simulation study comparison

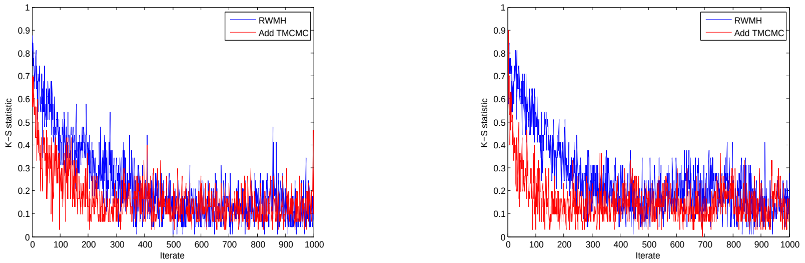
In this section, we compare the RWMH and the additive TMCMC methods using two parameters, one being the acceptance rate and the other, the Kolmogorov-Smirnov distance at each time point, of the empirical distribution with respect the target density. For the first measure, we observed the acceptance rates of the two algorithms for varying dimesnions and scaling factors l. The results are reported in **Table 1**.

Dim	Test Scaling	$Acceptance$ $rate(\%)$	
		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

Table 1: *A table representing the acceptance rates of the RWMH and additive TMCMC approaches for varying dimensions and varying scaling factors l, with the target density given by a iid product of* $N(0,1)$ *densities.*

Table 1 validates the fact for higher dimensions under optimal scaling, the acceptance rate of RWMH and additive TMCMC are indeed 0.234 and 0.439 respectively as the observed values are very close. Also, we see that for a fixed dimensions, as scaling increases, the acceptance rate falls drastically for RWMH and this worsens with increase in dimensionality. For dimensions 100 and 200, we skipped giving the acceptance rate for scaling $l = 10$ as it was understandably very small for RWMH. Comparatively, additive TMCMC is much more stable with change of scaling even for high dimensions. This also validates the robustness of the diffusion speed with respect to scaling l in **Fig 3**.

For the second measure of KS distance comparison, we run a number of chains, say L, starting from one fixed point for both MCMC and TMCMC adaptations. Corresponding to each time point t, we shall thus get L many iterates. The notion is that as time t increases (specially after burn-in), these L many iterates should be close to an independently drawn random sample from the target distribution π. So, if we look at the KS statistic for the empirical distribution of these iterates along any particular dimension with respect to the

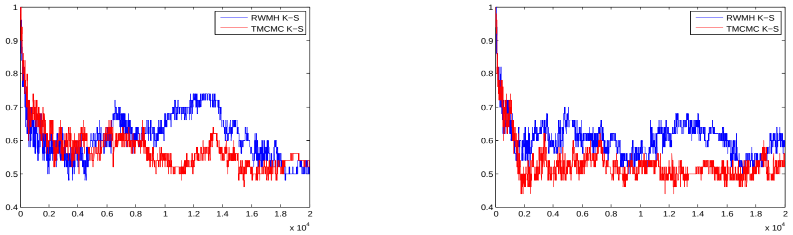


(a) $d = 30,\ \ell = 2.4$.

(b) $d = 30,\ \ell = 4$.

Figure 4: *The KS distance graph for the RWMH and the additive TMCMC chains for* 30 *dimensional target density being the product of iid* $N(0,1)$ *components and the scalings for the two graphs being* $l = 2.4$ *and* $l = 4$. *Notice that the KS graph for additive TMCMC seems to be lower compared to that of RWMH implying faster rate of convergence for additive TMCMC*

marginal of π along that dimension, we should find the test statistic decreasing with time and finally being very close to 0 after a certain time point. Now the question of interest is of the two approaches, additive TMCMC and RWMH, for which method the graph decays faster to 0. Corresponding to two different dimensions $d = 10$ and $d = 100$, and two scalings $\ell = 2.4$ (optimal given that $\mathbb{I} = 1$ for the target density product of $N(0,1)$ components) and $\ell = 4$, we present the two graphs of additive TMCMC and RWMH simultaneously in **Fig** 4 and **Fig** 5.



(a) $d = 100,\ \ell = 2.4$.

(b) $d = 100,\ \ell = 4$.

Figure 5: *The KS distance graph for the RWMH and the additive TMCMC chains for* 100 *dimensional target density being the product of iid* $N(0,1)$ *components and the scalings for the two graphs being* $l = 2.4$ *and* $l = 4$. *Notice that the KS graph for additive TMCMC seems to be lower compared to that of RWMH implying faster rate of convergence for additive TMCMC*

Therefore in conclusion it can be stated that

- 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 constraints on the target density, 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 meaning that once stationarity is attained RWMH will provide better samples than additive TMCMC.

- The KS test comparison in the simulation study shows that for high dimensions , TMCMC has lower KS statistic value compared to the RWMH when the chain is not stationary. This also suggests that additive TMCMC reaches burn-in faster than RWMH for higher dimensions. But once burn-in is reached, ideally the two methods should both give KS values close to 0 and that is why we see the KS graphs o stabilize with time for both the approaches.

Bibliography

[Bed07] M. Bedard. Weak Convergence OF Metropolis Algorithms For Non-i.i.d. Target Distributions. *The Annals of Applied Probability*, pages 1222–1244, 2007.

[Bed09] M. Bedard. On the optimal scaling problem of metropolis algorithms for hierarchical target distributions. *preprint*, 2009.

[BRS09] A. Beskos, G.O. Roberts, and A.M Stuart. Optimal scalings for local Metropolis-Hastings chains on non-product targets in high dimensions. *The Annals of Applied Probability*, pages 863–898, 2009.

[DB11] S Dutta and S Bhattacharya. Markov Chain Monte Carlo Based on Deterministic Transformations. *Statistical Methodology*, pages 100–116, 2011.

[DB13a] K.K. Dey and S Bhattacharya. On Geometric ergodicity of additive Transformation-based Markov Chain Monte Carlo Algorithm. *arXiv:1312.0915*, 2013.

[DB13b] K.K. Dey and S Bhattacharya. On Optimal scaling of Non-adaptive Additive Transformation based Markov Chain Monte Carlo. *arXiv:1307.1446*, 2013.

[GS90] A.E. Gelfand and A.F.M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, pages 398–409, 1990.

[GS96] Richardson S. Gilks, W. R. and D. J. Spiegelhalter. Markov chain Monte Carlo in practice. *Interdisciplinary Statistics, Chapman & Hall, London.*, 1996.

[Has70] W.K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, pages 97–109, 1970.

[JH00] S.F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process.Appl.*, pages 341–361, 2000.

[MRR53] N Metropolis, A.W. Rosenbluth, and A.H. Rosenbluth, M.N.and Teller. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, pages 1087–1092, 1953.

[MT93] S.P. Meyn and R.L. Tweedie. Markov chains and stochastic stability. 1993.

[NR06] P. Neal and G.O. Roberts. Optimal Scaling for Partially Updating MCMC Algorithms. *The Annals of Applied Probability*, pages 475–515, 2006.

[RGG97] G.O. Roberts, A Gelman, and W.R Gilks. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The Annals of Applied Probability*, pages 110–120, 1997.

[RT96] G.O. Roberts and R.L. Tweedie. Geometric convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms. *Biometrika*, pages 95–110, 1996.

[Tie94] L Tierney. Markov chains for exploring posterior distributions. *Ann. Statist*, pages 1701–1762, 1994.