On Single Variable Transformation Approach to Markov Chain Monte Carlo

Kushal K. Dey[†+] , Sourabh Bhattacharya[*]

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

1 Introduction

In today's times, Markov Chain Monte Carlo (MCMC) 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 Metropolis-Hastings (MH) algorithm [Has70][MRR53] using the random walk proposal. We refer to such algorithm as the Random Walk Metropolis Hastings (RWMH) algorithm. The convergence and optimal scaling of this algorithm has been extensively studied [RGG97]. However, despite the advances, there are certain glaring problems that one may encounter while using RWMH. For very high dimensional, non-standard target distributions, choosing the scales optimally is not feasible in practice, and hence, attempts of jointly updating the parameters using RWMH face serious drop in the acceptance rate, which, in turn, leads to poor convergence. Methods of adpatively selecting the scales usually take very large number of iterations to even converge to the optimal scales; particularly in complex and very high-dimensional situations, this exercise is computationally burdensome in the extreme. The alternative method of updating the parameters sequentially is not only computationally burdensome in high-dimensional problems, high posterior correlation among the parameters usually cause very slow convergence. These issues are discussed in much detail in [DB13b].

The TMCMC methodology proposed in Dutta and Bhattacharya [DB11] tries to address these problems. The methodology uses simple deterministic transformations using (typically) a single random variable having an appropriately chosen proposal density. In this paper, we primarily study one version, termed as the Additive TMCMC (ATMCMC) method, and deal with the ergodic behavior of the chain in high dimensions. Our aim is to present a comparative study of ATMCMC and the standard RWMH algorithm with respect to their ergodic behaviors.

This paper is organized as follows. In **Section 2**, we present the ATMCMC 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 ATMCMC and RWMH and analyze the results.

2 Algorithm

We first briefly describe how additive TMCMC (ATMCMC) 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 "move-types":

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

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 the kth transition, where k varies from 1 to 4 in the bivariate case, and in general from 1 to 2^d in $\mathbb{R}^d$. For simplicity we assume that the move-types are chosen with equal probability; see Dutta and Bhattacharya [DB11] for the general case. 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 $\varepsilon \sim g(\cdot)$ 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* $\boldsymbol{x}_{n+1} = \left\{ \begin{array}{lll} \boldsymbol{y} & with \;\; prob. & \alpha(\boldsymbol{x}_n,\varepsilon) \\ \boldsymbol{x}_n & with \;\; prob. & 1-\alpha(\boldsymbol{x}_n,\varepsilon) \end{array} \right\}$

Now we intuitively discuss why ATMCMC is a better option compared to the RWMH algorithm. Firstly, we tested using simulation experiments (all conducted in MATLAB

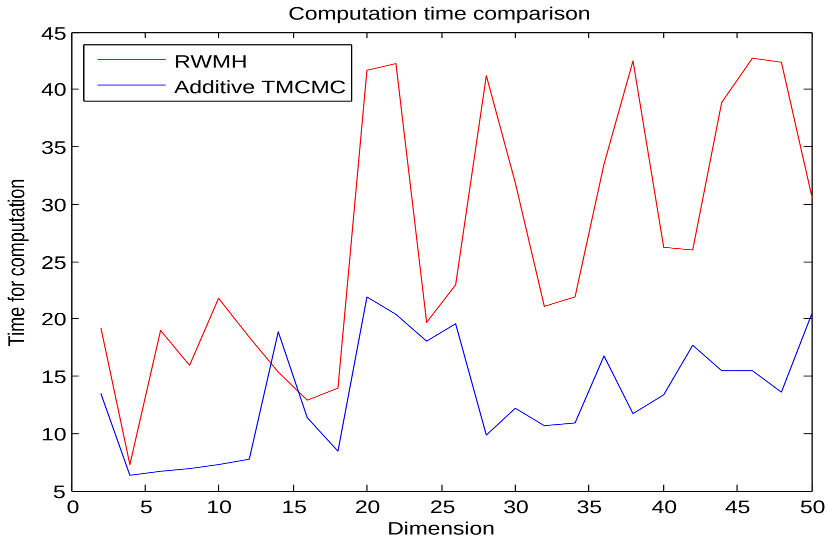


Figure 1: *Computation time (in MATLAB R2013b) of one run of 100,000 iterations with RWM and TMCMC algorithms corresponding to dimensions varying from 2 to 50 with target density being product of $N(0,5)$ and the proposal density for additive TMCMC being $TN_{>0}(0,1)$ (truncated $N(0,1)$ left truncated 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.*

R2013b) that our algorithm requires less computational time to run compared to RWMH (see **Fig 1**).

Secondly, and more importantly, ATMCMC is expected to have much higher acceptance rate than RWMH. We discuss this as follows.

In a standard RWMH algorithm in d dimensions, we need to generate d many ε_i's, for $i \in \{1,2,\cdots,d\}$. For simpliicty of illustration, 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 the ratio

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

If d is very large, then, by chance, we may obtain some very small or large values of $\varepsilon_i \sim q(\cdot)$ (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 ATMCMC 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 ATMCMC 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 ATMCMC 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 ATMCMC

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). For convergence of the process, we must ensure that it is μ-irreducible, where μ 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 $n > 0$, $\delta > 0$ and some measure ν such that

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

A chain is called *aperiodic* if the $g.c.d$ of all such n for **Eqn 5** holds, is 1. All these concepts of μ-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 ATMCMC 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, 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 aperiodic.*

A proof of this result can be found in Dutta and Bhattacharya [DB11], along with a graphical interpretation; see also Dey and Bhattacharya [DB13a]. In fact, in Dutta and Bhattacharya [DB11], a stronger result has been proved that for any $n > d$ (d represents the dimensionality of the state space), the minorization condition is satisfied. From the monorization condition, λ irreducibility follows trivially. Aperiodicity follows because the above result is true for all $n > d$ and the $g.c.d$ of such n is 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 $V \geq 1$, which is finite at least one point, and also constants $\rho \in (0,1)$ and M $(< \infty)$, such that

$$||P^n(x,\cdot) - \pi(\cdot)||_{TV} \leq MV(x)\rho^n \qquad \forall n \geq 1, \tag{6}$$

where $||\nu||_{TV}$ denotes the *total variation norm*, defined as

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

Apart from ensuring geometric rate of convergence of the Markov chain, another utility of geometric ergodicity is that one can apply Central Limit Theorem to a wide class of functions of the Markov chain, and hence, one can also investigate 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 $c < \infty$ such that

$$PV(x) \leq \lambda V(x) + c1_E(x), \tag{7}$$

where $PV(x) = \int V(y)P(x,y)dy$ is 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 Mengersen and Tweedie [MT96]. 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 hold:*

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

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

*Then this V satisfies the geometric drift condition in **Eqn 7**, and hence the chain must be geometrically 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*

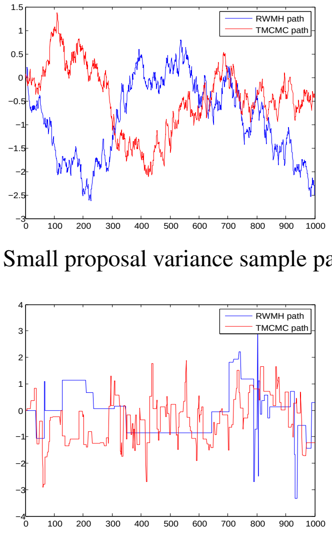
$$\liminf_{\|x\|\to\infty} Q(x, A(x)) > 0, \tag{10}$$

where $A(x)$ denotes the acceptance region when x is updated, and $Q(x,A(x))$ denotes the probability of the acceptance region under the ATMCMC proposal distribution associated with the density $q(\cdot)$ of ε.

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 products 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 will be dedicated to obtaining the optimal value of this scaling l and determining the limiting expected acceptance rate of ATMCMC under the optimal scaling scenario. If the variance of the proposal density is very small, then the jumps will be of smaller magnitude and this would mean the Markov chain would take very many iterations to traverse the entire state space, and in the process, the convergence



(a) Small proposal variance sample path

(b) Large proposal variance sample path

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

rate would be very small. 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 - iid set up, some special class of independent but non-identical set up, as well as a dependent set-up. For our purpose, we have developed an optimal scaling theory for ATMCMC where we have optimized the diffusion speed of our process to obtain optimal acceptance rate for ATMCMC. We present 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 $d \rightarrow \infty$, meaning that as the dimension of the space blows to ∞, the sped up ATMCMC process essentially converges to a continuous time diffusion process.

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

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

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

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

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

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

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

where x_1^* lies between x_1 and $x_1+b_1\varepsilon$ and M_V 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 ATMCMC chain. It can be shown that

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

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

where Φ is the standard Gaussian cdf.

Using this proposition, we can write

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

Note that using Taylor series expansion around x_1, we can represent $\eta(x_1,b_1,\varepsilon)$ 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 $\frac{l}{\sqrt{d}}z_1^*$, where z_1^* follows a $N(0,1)$ distribution, η and $\mathbb{W}$ can be expressed in terms of l and z_1^* as

$$\eta(x_1,z_1^*,d)=\frac{lz_1^*}{\sqrt{d}}\left[\log f(x_1)\right]^{'}+\frac{l^2z_1^{*2}}{2!d}\left[\log f(x_1)\right]^{''}+\frac{l^3z_1^{*3}}{3!d^{\frac{3}{2}}}\left[\log f(\xi_1)\right]^{'''} \tag{19}$$

and

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

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

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

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 $\Phi\left(\frac{\eta(x_1,z_1^*,d)-\frac{z_1^{*2}l^2}{2}\mathbb{I}}{\sqrt{z_1^{*2}l^2\mathbb{I}}}\right)$, $e^{\eta(x_1,z_1^*,d)}$ and $\Phi\left(\frac{-\frac{z_1^{*2}l^2\mathbb{I}}{2}-\eta(x_1,z_1^*,d)}{\sqrt{z_1^{*2}l^2\mathbb{I}}}\right)$ about $\eta=0$, we obtain the expression of $G_d(V(x)$ as

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

where

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

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

$$GV(x) = h(l)\left[\frac{1}{2}(\log f)'(x_1)V'(x_1) + \frac{1}{2}V''(x_1)\right]. \qquad (24)$$

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

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

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 plug this value in the formula for asymptotic expected acceptance rate to obtain

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

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

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

It was noted in [RGG97] that the limiting expected acceptance rate corresponding to optimal scaling in RWMH is 0.234, while for that for the optimal scaling in additive TMCMC is 0.439 which is almost twice as that of RWMH. It is to be noted that the optimal scale of RWMH is $l_{opt} = \frac{2.4}{\sqrt{I}}$, which, up to the first decimal place, is the same as that of ATMCMC. The graphs of the diffusion speeds over different l for ATMCMC and for standard RWMH are presented in **Fig 3**.

Note that the diffusion speed at l_{opt} is higher for RWMH compared to additive TMCMC (ATMCMC) implying that once stationarity is reached, there will be faster mixing among the iterates in RWMH compared to ATMCMC. However, an interesting observation is that if l deviates slightly from l_{opt}, the diffusion speed of RWMH drops much faster compared to that of ATMCMC. Thus, ATMCMC is much more robust compared to RWMH with respect to the scaling. This is very important in complex and high-dimensional practical situations where achieving the optimal scaling usually turns out to be infeasible; recall the discussion regarding this in Section 1. Although our above analysis holds true only for the case when all the components of the product density are *iid*, however, this condition can be relaxed to include independent components with appropriate scaling and inherent regularization properties as in Bedard (2009) [Bed09] and Dey and Bhattacharya (2013) [DB13a] and also to non-regular component densities in Dey and Bhattacharya [DB14].

Also, 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 our inherent assumption that the process is in stationarity. The major question to address now is that which chain has faster convergence to stationarity. We address this in the next section via simulation studies.

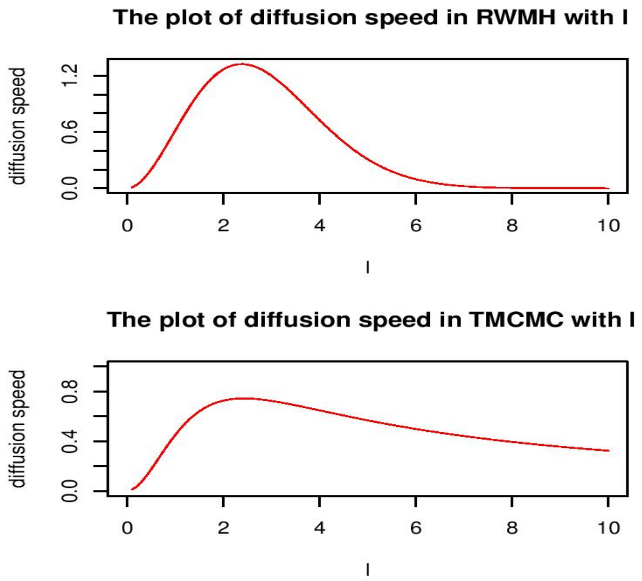


Figure 3: *The plot of the diffusion speed with respect to the scaling factor l for RWMH and ATMCMC chains.*

5 Simulation study comparison

In this section, we compare RWMH and additive TMCMC methods using two parameters, one being the acceptance rate and the other, the Kolmogorov-Smirnov (KS) distance between the empirical distribution at each time point and 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**.

Table 1 validates that for higher dimensions, under optimal scaling, the acceptance rates of RWMH and additive TMCMC are indeed 0.234 and 0.439 respectively, as the observed values are very close to the theoretical ones. Also, we see that for fixed dimensions, as scaling increases away from the optimal value, the acceptance rate falls drastically for RWMH and this worsens with increase in dimensionality. For dimensions 100 and 200, we skipped providing the acceptance rates 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 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 RWMH and ATMCMC adaptations. Corresponding to each time point t, we 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 observe the KS statistic for the empirical distribution of these iterates along any particular dimension with respect to the marginal of π along that dimension, we expect the test statistic to be 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, ATMCMC and RWMH, for which method the graph decays faster

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

to 0? Corresponding to two different dimensions $d = 10$ and $d = 100$, and two scalings $l = 2.4$ (optimal given that $\mathbb{I} = 1$ for the target density product of $N(0,1)$ components) and $l = 4$, we present the two graphs of additive TMCMC and RWMH simultaneously in **Fig 4** and **Fig 5**. Both the figures, but particularly the latter, clearly indicate faster convergence of ATMCMC to the stationary distribution.

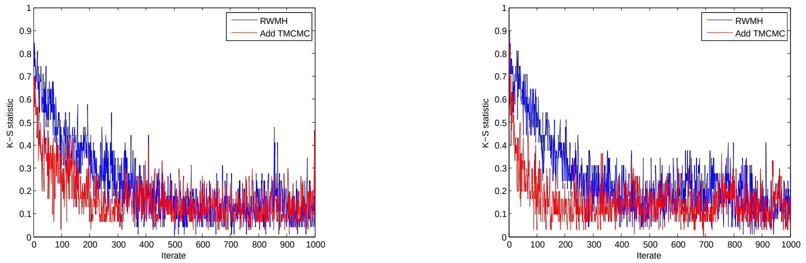
Therefore in conclusion it can be stated that

- ATMCMC 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 ATMCMC algorithm is geometrically ergodic.

- ATMCMC 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 ATMCMC.

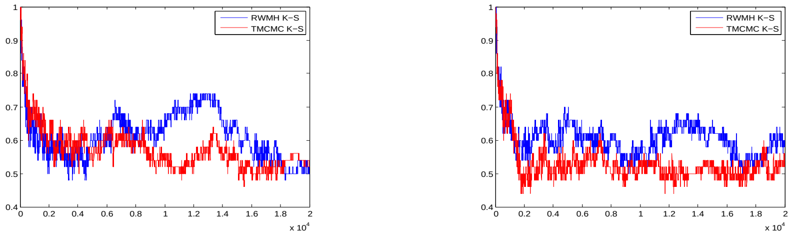
- The KS test comparison in the simulation study shows that for high dimensions, ATMCMC has lower KS statistic value compared to RWMH when the chain is not stationary. This also suggests that ATMCMC reaches burn-in faster than RWMH for higher dimensions. But once burn-in is reached, ideally the two methods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



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

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

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



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

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

Figure 5: *The KS distance graph for RWMH and ATMCMC chains for a* 100 *dimensional target density, which is the product of iid* $N(0,1)$ *components. The scalings for the two graphs are* $l = 2.4$ *and* $l = 4$. *Here the KS graph for ATMCMC is clearly lower compared to that of RWMH implying faster rate of convergence for ATMCMC.*

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.

[DB14] K.K. Dey and S Bhattacharya. On Optimal Scaling of Additive Transformation Based Monte Carlo Under Non-Regular Cases. *arXiv:1405.0913*, 2014.

[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.

[MT96] K.L. Mengersen and R.L. Tweedie. Rates of Convergence of the Hastings and Metropolis Algorithms. *The Annals of Statistics*, pages 101–121, 1996.

[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.