Brazilian Journal of Probability and Statistics
2017, Vol. 31, No. 3, 569–617
DOI: 10.1214/16-BJPS325

A brief tutorial on transformation based Markov Chain Monte Carlo and optimal scaling of the additive transformation

Kushal Kr. Dey[a] **and Sourabh Bhattacharya**[b]

[a]*University of Chicago*
[b]*Indian Statistical Institute*

Abstract. We consider the recently introduced Transformation-based Markov Chain Monte Carlo (TMCMC) (*Stat. Methodol.* **16** (2014) 100–116), a methodology that is designed to update all the parameters simultaneously using some simple deterministic transformation of a one-dimensional random variable drawn from some arbitrary distribution on a relevant support. The additive transformation based TMCMC is similar in spirit to random walk Metropolis, except the fact that unlike the latter, additive TMCMC uses a single draw from a one-dimensional proposal distribution to update the high-dimensional parameter. In this paper, we first provide a brief tutorial on TMCMC, exploring its connections and contrasts with various available MCMC methods.

Then we study the diffusion limits of additive TMCMC under various set-ups ranging from the product structure of the target density to the case where the target is absolutely continuous with respect to a Gaussian measure; we also consider the additive TMCMC within Gibbs approach for all the above set-ups. These investigations lead to appropriate scaling of the one-dimensional proposal density. We also show that the optimal acceptance rate of additive TMCMC is 0.439 under all the aforementioned set-ups, in contrast with the well-established 0.234 acceptance rate associated with optimal random walk Metropolis algorithms under the same set-ups. We also elucidate the ramifications of our results and clear advantages of additive TMCMC over random walk Metropolis with ample simulation studies and Bayesian analysis of a real, spatial dataset with which 160 unknowns are associated.

1 Introduction

Markov Chain Monte Carlo (MCMC), particularly, the Metropolis–Hastings (MH) methods, have revolutionized Bayesian computation—this pleasing truth, however, is often hard to appreciate in the face of the challenges posed by computational complexities and convergence issues of traditional MCMC. Indeed, exploration of very high-dimensional posterior distributions using MCMC can be both computationally very expensive and troublesome convergence-wise. Thus, there seems

Key words and phrases. Additive transformation, diffusion limit, high dimension, optimal scaling, random walk, transformation-based Markov Chain Monte Carlo.
Received March 2015; accepted June 2016.

to be trade-off between the great flexibility of MCMC algorithms [see, for example, Storvik (2011), Martino and Read (2013)] and choice of the right MCMC algorithm that ensures good convergence properties and reasonable computational complexity. Investigation of connections between varieties of available MCMC methods, as provided in the aforementioned papers, seems to be important to decide upon a suitable MCMC algorithm, given any paricular problem at hand.

The random walk Metropolis (RWM) algorithm is a popular MH algorithm because of its simplicity and ease in implementation, but unless great care is taken to properly scale the proposal distribution the algorithm can have poor convergence properties. For instance, if the variance of the proposal density is small, then the jumps will be small in magnitude, implying that the Markov chain will require a large number of iterations to explore the entire state-space. On the other hand, large variance of the proposal density causes too many rejections of the proposed moves, again considerably slowing down convergence of the underlying Markov chain. The need for an optimal choice of the proposal variance is thus inherent in the RWM algorithms. The pioneering approach towards obtaining an optimal scaling of the RWM proposal is due to Roberts, Gelman and Gilks (1997) in the case of target densities associated with independent and identical (*iid*) random variables; generalization of this work to more general set-ups are provided by Bedard (2007) (target density associated with independent but non-identical random variables) and Mattingly, Pillai and Stuart (2011) (target density absolutely continuous with respect to a Gaussian measure). The approach used in all these works is to study the diffusion approximation of the high-dimensional RWM algorithm, and maximization of the speed of convergence of the limiting diffusion. The optimal scaling, the optimal acceptance rate and the optimal speed of convergence of the limiting diffusion, along with the complexity of the algorithm are all obtained from this powerful approach.

In practice, a serious drawback of the RWM algorithm in high dimensions is that there is always a positive probability that a particular co-ordinate of the high-dimensional random variable is ill-proposed; in that case the acceptance ratio will tend to be extremely small, prompting rejection of the entire high-dimensional move. In general, unless the high-dimensional proposal distribution, which need not necessarily be a random walk proposal distribution, is designed with extreme care, such problem usually persists. Unfortunately, such carefully designed proposal density is rare in high dimensions. To combat these difficulties Dutta and Bhattacharya (2014) proposed an approach where the entire block of parameters can be updated simultaneously using some simple deterministic transformation of a scalar random variable sampled from some arbitrary distribution defined on some suitable support. The strategy effectively reduces the high-dimensional proposal distribution to a one-dimensional proposal, greatly improving the acceptance rate and computational speed in the process. This methodology is no longer Metropolis–Hastings for dimensions greater than one; the proposal density in more than one dimension becomes singular because it is induced by a

one-dimensional random variable. However, in one-dimensional cases this coincides with Metropolis–Hastings with a specialized mixture proposal density; in particular, the additive transformation based TMCMC coincides with RWM in one-dimensional situations. Dutta and Bhattacharya (2014) refer to this new general methodology as Transformation-based MCMC (TMCMC). In their work the authors point out several advantages of the additive transformation in comparison with the other valid transformations. For instance, they show that additive TMCMC requires less number of 'move-types' compared to other valid transformations; moreover, the acceptance rate has a simple form for additive transformations since the Jacobian of additive transformations is 1.

The contribution of this paper is two-fold. First, we provide a brief tutorial on TMCMC, attempting to convey the key ideas and the properties in simple terms and from several perspectives. We explore various connections and contrasts with existing MCMC algorithms.

Second, we investigate the diffusion limits of additive TMCMC in high-dimensional situations under various forms of the target density when the one-dimensional random variable used for the additive transformation is drawn from a left truncated zero-mean normal density. In particular, we consider situations when the target density corresponds to *iid* random variables, independent but non-identically distributed random variables; we also study the diffusion limit of additive TMCMC when the target is absolutely continuous with respect to a Gaussian measure. Since all these forms are considered in the MCMC literature related to diffusion limits and optimal scaling of RWM, comparisons of our additive TMCMC-based approaches can be made with the respective RWM-based approaches. Furthermore, in each of the aforementioned set-ups, we also consider additive TMCMC within Gibbs approach, where one or multiple components of the high-dimensional random variable are updated by additive TMCMC, conditioning on the remaining components. This we compare with the corresponding RWM within Gibbs approach under the same settings of the target densities.

Briefly, our scaling investigations show that the optimal additive TMCMC acceptance rate in all the set-ups is 0.439, as opposed to 0.234 associated with RWM. Moreover, we point out that even though the optimal diffusion speed of RWM is slightly greater than that of additive TMCMC, the diffusion speed associated with additive TMCMC is more robust with respect to the choice of the scaling constant. In other words, if the optimal scaling constant for RWM is somewhat altered, this triggers a sharp fall in the diffusion speed, but in the case of additive TMCMC the rate of decrease of diffusion speed is much slower. Investigation of the consequences of this phenomenon with simulation studies reveal severe decline in the performance of RWM in comparison with additive TMCMC.

This non-robustness of RWM with respect to scale choices other than the optimal, presents quite important consequences for applied MCMC practitioners, which we elaborate with a real, spatial data analysis problem. In a nutshell, in the context of the spatial problem, we have provided a method, which appears to

be generally applicable, for approximately achieving 44% and 23% acceptance rates for additive TMCMC and RWM; however, achieving the desired acceptance rates in general problems where optimal scaling theories are yet lacking, does not guarantee that the achieved acceptance rates correspond to optimal scales, as there are usually very many scale choices corresponding to the same acceptance rate. Because of such sub-optimality, in the real spatial problem, RWM faces very serious performance problems. On the other hand, additive TMCMC, because of its robustness with respect to the scales, performs quite reasonably.

Our paper is structured as follows. In Section 2, we provide a brief tutorial on TMCMC. We develop the theory for optimal additive TMCMC scaling in the *iid* set-up in Section 3; in the same section (Section 3.1) we also develop the corresponding theory for additive TMCMC within Gibbs in the *iid* situation. In Section 4, we extend the additive TMCMC-based optimal scaling theory to the independent but non-identical set-up; in Section 4.1 we outline the corresponding TMCMC within Gibbs case. We then further extend our additive TMCMC based optimal scaling theory to the aforementioned dependent set-up in Section 5, presenting the formal result in Section 5.2; the corresponding TMCMC within Gibbs case is considered in Section 5.3. In Section 6, we provide numerical comparisons between additive TMCMC and RWM in terms of optimal acceptance rates and diffusion speeds; in Section 7, we illustrate our theoretical results and compare the performances of additive TMCMC and RWM using simulation studies, illustrating that the former is a far more effective algorithm in comparison with the latter. In Section 8, we compare additive TMCMC with RWM with respect to a 160-dimensional posterior density associated with a real, spatial dataset, vividly demonstrating the clear superiority of additive TMCMC over RWM. Finally, we make concluding remarks in Section 9.

Apart from the main developments provided in this article, we provide additional details in our supplementary material [Dey and Bhattacharya (2016b)], whose sections and figures have the prefix "S-" when referred to in this article. Briefly, in Section S-1, we provide details on computational efficiency of TMCMC. Specifically, we demonstrate with an experiment the superior computational speed of additive TMCMC in comparison with RWM, particularly in high dimensions. In Section S-2 we discuss, with appropriate experiments, the necessity of optimal scaling in additive TMCMC, while in Sections S-3 and S-4 we delve into the robustness issues associated with the scale choices of additive TMCMC and RWM. In Section S-5, we include brief discussions of adaptive versions of RWM and TMCMC. Moreover, the proofs of all our technical results are provided in Sections S-6 and S-7 of the supplement.

2 A brief overview of TMCMC

Suppose that we are simulating from a d dimensional space (usually $\mathbb{R}^d$, where $\mathbb{R}$ is the real line), and suppose we are currently at a point $x = (x_1, \ldots, x_d)$. Let

us define the d-dimensional random vector $b = (b_1, \ldots, b_d)$, such that, for $i = 1, \ldots, d$,

$$b_i = \begin{cases} +1 & \text{with probability } p_i; \\ 0 & \text{with probability } 1 - p_i - q_i; \\ -1 & \text{with probability } q_i, \end{cases} \tag{1}$$

where, for each i, $0 < p_i, q_i < 1$ such that $p_i + q_i \leq 1$. Let $\varepsilon \sim \varrho(\varepsilon) = \tilde{\varrho}(\varepsilon) I_{\mathbb{S}}(\varepsilon)$, where $\tilde{\varrho}(\cdot)$ is any arbitrary density supported on some suitable space $\mathbb{S}$; here $I_{\mathbb{S}}(\cdot)$ denotes the indicator function of $\mathbb{S}$.

TMCMC uses moves of the following type:

$$(x_1, \ldots, x_d) \to \big(T^{b_1}(x_1, \varepsilon), \ldots, T^{b_d}(x_d, \varepsilon)\big), \tag{2}$$

where $T^{+1}(x_i, \varepsilon)$, the forward transformation to co-ordinate x_i, and $T^{-1}(x_i, \varepsilon)$, the backward transformation to x_i, are bijective for fixed ε and injective for fixed x_i, satisfying

$$T^{+1}\big(T^{-1}(x_i, \varepsilon), \varepsilon\big) = T^{-1}\big(T^{+1}(x_i, \varepsilon), \varepsilon\big) = x_i. \tag{3}$$

The transformation

$$T^0(x_i, \varepsilon) \equiv x_i, \qquad \forall \varepsilon \in \mathbb{S}, \tag{4}$$

indicates no change to the co-ordinate x_i while updating the vector $x = (x_1, \ldots, x_d)$ to $x^* = \mathcal{T}_b(x, \varepsilon)$, where $\mathcal{T}_b(x, \varepsilon)$ denotes the updated vector $(T^{b_1}(x_1, \varepsilon), \ldots, T^{b_d}(x_d, \varepsilon))$. Assuming for simplicity of illustration that $p_i = q_i$ for $i = 1, \ldots, d$, move (2) is to be accepted with probability

$$\alpha = \min\left\{1, \frac{\pi(x^*)}{\pi(x)} J^b(x, \varepsilon)\right\}, \tag{5}$$

where $J^b(x, \varepsilon) = |\frac{\partial(\mathcal{T}^b(x,\varepsilon),\varepsilon)}{\partial(x,\varepsilon)}|$ is the Jacobian of the transformation associated with $\mathcal{T}^b$. For general $(p_1, \ldots, p_d)$ and $(q_1, \ldots, q_d)$, the acceptance ratio depends upon these probabilities; see Dutta and Bhattacharya (2014).

2.1 Detailed balance

In the supplement to Dutta and Bhattacharya (2014) the proof of detailed balance has been provided, but here we refurnish the proof with more details and with more intuitive discussion. For the purpose of detailed balance, we need the following definition of "conjugate" $b^c = (b_1^c, \ldots, b_d^c)$ of the random vector d:

$$b_i^c = \begin{cases} +1 & \text{with probability } q_i; \\ 0 & \text{with probability } 1 - p_i - q_i; \\ -1 & \text{with probability } p_i. \end{cases} \tag{6}$$

This definition is needed for returning to x from x^*, so that moving from x to x^* using the transformation $\mathcal{T}^b$ has, in essence, the same probability as returning from

x^* to x using the transformation $\mathcal{T}^{b^c}$. Details are provided below; at this point we note that for the ith co-ordinate x_i, the probability of making a forward move to x_i^* using $T^{+1}(x_i, \varepsilon)$ is p_i, which is also the probability of returning from x_i^* to x_i using the backward move $T^{-1}(x_i^*, \varepsilon) = T^{(+1)^c}(x_i^*, \varepsilon)$.

Letting K denote the Markov transition kernel associated with TMCMC, note that for moving from x to x^*, the kernel satisfies

$$\begin{aligned}\pi(x)K(x \to x^*) &= \pi(x)P(b)\varrho(\varepsilon)\min\left\{1, \frac{\pi(x^*)}{\pi(x)}J^b(x,\varepsilon)\right\} \\ &= \min\{\pi(x)P(b)\varrho(\varepsilon), P(b)\varrho(\varepsilon)\pi(x^*)J^b(x,\varepsilon)\},\end{aligned} \tag{7}$$

where $P(b)$ is the probability of b responsible for the movement of the underlying Markov chain from x to x^*. For returning from x^* to x, the kernel satisfies

$$\begin{aligned}&\pi(x^*)K(x^* \to x) \\ &\quad= \pi(x^*)P(b^c)\varrho(\varepsilon)J^b(x,\varepsilon)\min\left\{1, \frac{\pi(x)}{\pi(x^*)}J^{b^c}(x^*,\varepsilon)\right\} \\ &\quad= \min\{\pi(x^*)P(b^c)\varrho(\varepsilon)J^b(x,\varepsilon), \\ &\qquad \pi(x)P(b^c)\varrho(\varepsilon)J^b(x,\varepsilon) \times J^{b^c}(x^*,\varepsilon)\} \\ &\quad= \min\{\pi(x^*)P(b^c)\varrho(\varepsilon)J^b(x,\varepsilon), \\ &\qquad \pi(x)P(b^c)\varrho(\varepsilon)J^b(x,\varepsilon) \times J^{b^c}(x^*,\varepsilon)\}.\end{aligned} \tag{8}$$

It follows from (3) and (4) that $\mathcal{T}^{b^c}(\mathcal{T}^b(x,\varepsilon)) = x$, so that

$$\begin{aligned}&J^b(x,\varepsilon) \times J^{b^c}(x^*,\varepsilon) \\ &\quad= \left|\frac{\partial(\mathcal{T}^b(x,\varepsilon),\varepsilon)}{\partial(x,\varepsilon)}\right| \times \left|\frac{\partial(\mathcal{T}^{b^c}(\mathcal{T}^b(x,\varepsilon)),\varepsilon)}{\partial(\mathcal{T}^b(x,\varepsilon),\varepsilon)}\right| \\ &\quad= \left|\frac{\partial(\mathcal{T}^b(x,\varepsilon),\varepsilon)}{\partial(x,\varepsilon)}\right| \times \left|\frac{\partial(x,\varepsilon)}{\partial(\mathcal{T}^b(x,\varepsilon),\varepsilon)}\right| \\ &\quad= 1.\end{aligned} \tag{9}$$

Substituting (9) in (8), we obtain

$$\pi(x^*)K(x^* \to x) = \min\{\pi(x^*)P(b^c)g(\varepsilon)J^b(x,\varepsilon), \pi(x)P(b^c)g(\varepsilon)\}. \tag{10}$$

If, for simplicity of illustration we assume $p_i = q_i$ for $i = 1, \ldots, d$, it follows that $P(b) = P(b^c)$, so that (10) is equal to (7), proving detailed balance. Detailed balance of course holds for general probabilities $(p_1, \ldots, p_d)$ and $(q_1, \ldots, q_d)$; see the supplement of Dutta and Bhattacharya (2014).

2.2 Discussion on the independence of the acceptance probability of the proposal density ϱ

An important feature of the TMCMC acceptance probability distinguishing it from the acceptance probability of the Metropolis–Hastings algorithms is its independence of the proposal density ϱ, irrespective of whether or not it is symmetric, and for all valid transformations T^b. The reason for this is implicit in the above detailed balance arguments—the same ε is generated from the proposal density ϱ while moving forward from x to x^* as well as while returning to x from x^*. This is exactly the reason why $\varrho(\varepsilon)$ features in both (7) and (8). Consequently, detailed balance is satisfied only if the acceptance ratio of TMCMC is independent of ϱ.

2.3 Relationship of TMCMC with the MH methodology

Note that, although in Section 2 we indicated a single ranodom variable ε to be used in the transformations, it is permissible to use k random variables $\{\varepsilon_1, \ldots, \varepsilon_k\}$, where $k \in \{1, \ldots, d\}$. Here it is also important to remark that general TMCMC with $k = d$ does not reduce to general MH method for the very reason that the acceptance ratio of TMCMC is always independent of the proposal density, while in MH it is not. For dimension $d = 1$, however, TMCMC boils down to an MH algorithm with a specialized two-component mixture proposal density, where the mixture components correspond to the two available move types, forward and backward. See Dutta and Bhattacharya (2014) for the complete technical details.

As pointed out by a reviewer, TMCMC can be viewed as a MH within Gibbs methodology, which updates the variables one at a time using general MH [Geyer (2011) criticizes this terminology since Gibbs is a special case of MH]. This can be seen as follows. Suppose that we assign positive probabilities to the sets $\tilde{b}_i = (b_1 = 0, \ldots, b_{i-1} = 0, b_i, b_{i+1} = 0, \ldots, b_d = 0)$, for $i = 1, \ldots, d$, where $b_i \in \{-1, 0, 1\}$. For TMCMC, at each iteration, we can select one of the sets $\tilde{b}_{i^*}$ by choosing $i^* \in \{1, \ldots, d\}$ at random, and generate $b_{i^*} \in \{-1, 0, 1\}$ with probabilities q_{i^*}, $1 - p_{i^*} - q_{i^*}$ and p_{i^*}. The corresponding x_{i^*}, corresponding to b_{i^*}, is then updated according to the TMCMC mechanism. This being a single-dimensional TMCMC step, coincides with an MH within Gibbs procedure with a specialized proposal mechanism.

2.4 Additive and multiplicative transformations in TMCMC

2.4.1 *Additive TMCMC.* The additive TMCMC, which is of our interest in this work, uses moves of the following type:

$$(x_1, \ldots, x_d) \to (x_1 + b_1\varepsilon, \ldots, x_d + b_d\varepsilon),$$

where $\varepsilon \sim \varrho(\varepsilon)$. Here $\varrho(\cdot)$ is an arbitrary density with support $\mathbb{R}_+$, the positive part of the real line. In other words, we assume that $T^{b_i}(x_i, \varepsilon) = x_i + b_i\varepsilon$. In this work, we shall assume that $p_i = 1/2$ and $q_i = 1/2$ for $i = 1, \ldots, d$, so that the

probabilities of choosing $b_i = 0$ and $b_i^c = 0$ are zero. Indeed, as proved in Dutta and Bhattacharya (2014), this is a completely valid and efficient choice for additive transformations. For this work, we set $\varrho(\varepsilon) I_{\{\varepsilon>0\}} \equiv N(0, \frac{\ell^2}{d}) I_{\{\varepsilon>0\}}$.

Note that, for each i, $b_i\varepsilon \sim N(0, \frac{\ell^2}{d})$, but even though $b_i\varepsilon$ are pairwise uncorrelated ($E(b_i\varepsilon \times b_j\varepsilon) = 0$ for $i \neq j$), they are not independent since all of them involve the same ε. Also observe that $b_i\varepsilon + b_j\varepsilon = 0$ with probability $1/2$ for $i \neq j$, showing that the linear combinations of $b_i\varepsilon$ need not be normal. In other words, the joint distribution of $(b_1\varepsilon, \ldots, b_d\varepsilon)$ is not normal, even though the marginal distributions are normal and the components are pairwise uncorrelated. This also shows that $b_i\varepsilon$ are not independent, because independence would imply joint normality of the components.

Thus, a single ε is simulated from a truncated normal distribution, which is then either added to, or subtracted from each of the d coordinates of x with probability $1/2$. Assuming that the target distribution is proportional to π, the new move $x^* = (x_1 + b_1\varepsilon, \ldots, x_d + b_d\varepsilon)$ is accepted with probability

$$\alpha = \min\left\{1, \frac{\pi(x^*)}{\pi(x)}\right\}. \tag{11}$$

The RWM algorithm, unlike additive TMCMC, proceeds by simulating $\varepsilon_1, \ldots, \varepsilon_d$ independently from $N(0, \frac{\ell^2}{d})$, and then adding ε_i to the co-ordinate x_i, for each i. The new move is accepted with probability having the same form as (11). As discussed in Section 2.3 for TMCMC it is permissible to use k random variables $\{\varepsilon_1, \ldots, \varepsilon_k\}$, where $k \in \{1, \ldots, d\}$. Thus, with such a proposal, if $k = d$, additive TMCMC reduces to RWM, showing that the latter is a special case of the former.

Discussion of computational efficiency of TMCMC is already provided in Dutta and Bhattacharya (2014). In this article, we supplement the discussion by demonstrating with an experiment the substantial computational gains of additive TMCMC over RWM, particularly in high dimensions; see Section S-1.

2.4.2 *Multiplicative TMCMC.* In contrast with additive TMCMC, multiplicative TMCMC proceeds by generating ε from some relevant distribution $\varrho(\varepsilon)$ supported on $[-1, 1] \setminus \{0\}$ and then either multiplying or dividing the current states by ε. In this case, it is necessary to assign positive probabilities to $b_i = 0$ and $b_i^c = 0$ to ensure irreducibility. For $i = 1, \ldots, d$, the general multiplicative TMCMC algorithm applies the transformation $x_i\varepsilon^{b_i}$ to x_i, which necessitates multiplication of the Jacobian $\varepsilon^{\sum_{i=1}^d b_i}$ to the ratio of the target distribution evaluated at the new and current states, for computation of the acceptance ratio; see Dutta (2012), Dutta and Bhattacharya (2014) and Dey and Bhattacharya (2016a).

There is an alternative version of multiplicative TMCMC that proceeds by generating ε from $\varrho(\varepsilon)$ that is supported on $(0, 1]$, then making the transformation $x_i \mapsto c_i x_i \varepsilon^{b_i}$, where $c_i = 1$ with probability r_i and -1 with probability $1 - r_i$, where $0 < r_i < 1$. Here, it is again permissible to set the probabilities of $b_i = 0$

and $b_i^c = 0$ to zero. The Jacobian remains the same as in the previous multiplicative version.

Note that the multiplicative proposal allows relatively large jumps, while the additive moves are local in nature. Hence, judicious combination of the two proposals may allow local moves as well as large jumps, which would result in a more efficient algorithm compared to the individual proposals. Indeed, Dey and Bhattacharya (2016a) demonstrate that a mixture of the additive and multiplicative transformations outperforms individual additive TMCMC and multiplicative TMCMC.

2.4.3 *Additive-multiplicative TMCMC.* Apart from these TMCMC mechanisms where all the coordinates are given either additive transformation or multiplicative transformation, it is possible to give additive transformation to some co-ordinates and multiplicative to the rest; this TMCMC mechanism has been referred to as additive-multiplicative TMCMC in Dutta and Bhattacharya (2014) and Dey and Bhattacharya (2016a).

2.5 A manifold interpretation of the TMCMC proposal mechanism

A reviewer observed that for a d-dimensional target distribution, the TMCMC proposal can be viewed as generating points in a manifold $\mathbb{M}$ such that $\mathbb{M} \subseteq \mathbb{R}^d$. For instance, for additive TMCMC this manifold $\mathbb{M}$ is formed by hyperplanes within $\mathbb{R}^d$. When $d = 2$, Figure 1 depicts the manifold as intersecting straight lines with the current state (x_1, x_2) being the point of intersection. For the two versions of multiplicative TMCMC, the manifolds are shown in Figures 2 and 3, respectively.

The TMCMC proposal implicitly describes the manifold parametrically, where $\varepsilon \in \mathbb{R}^k$ plays the role of the parameter. If $k = 1$, the proposed values must belong to appropriate curves embedded in $\mathbb{R}^d$. It is important to clarify, as the reviewer also observed, that $\varrho(\varepsilon)$ does not completely define the TMCMC proposal; the complete TMCMC proposal consists of first generating ε from ϱ and then transforming the current state deterministically with the help of ε in a way that ensures the transformed state falls within the appropriate manifold.

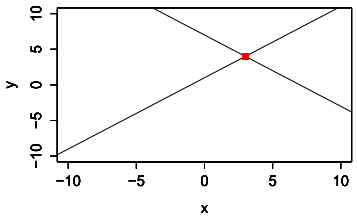


Figure 1 *Relevant manifold representing the one-step proposal mechanism for additive TMCMC for $d = 2$ given the current state, denoted by the dark patch at the intersecting point of the two lines.*

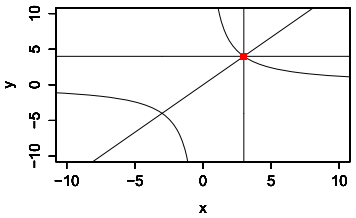


Figure 2 *Relevant manifold representing the one-step proposal mechanism for the first version of multiplicative TMCMC for $d = 2$ given the current state, denoted by the dark patch at the intersecting point of the straight lines and the curve towards the upper right portion of the diagram.*

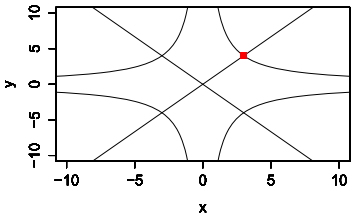


Figure 3 *Relevant manifold representing the one-step proposal mechanism for the second version of multiplicative TMCMC for $d = 2$ given the current state, denoted by the dark patch at the intersecting point of the curve and the straight line towards the upper right portion of the diagram.*

Since, for $d > 1$, $\mathbb{M} \subset \mathbb{R}^d$ even for $k = d$, the above manifold viewpoint also succeeds in explaining why, even for $k = d$, the TMCMC proposal does not reduce, in general, to the standard MH methodology. This perspective also clearly shows the singularity of the proposal, which also explains the association of the Jacobian with the acceptance probability of TMCMC. Note that for $d = 1$ (so that $k = 1$), however, we must have $\mathbb{M} = \mathbb{R}$.

2.6 Contrast of TMCMC with deterministic transformation based generalized Gibbs/MH approaches

TMCMC uses deterministic transformations to update the variables; however, deterministic transformations have also been considered by Liu and Yu (1999), Liu and Sabatti (2000), Kou, Xie and Liu (2005) in an attempt to improve mixing behaviour of the underlying usual Gibbs or MH algorithm. In a nutshell, these authors apply suitable deterministic transformations, usually additive, as $(x_1, \ldots, x_d) \to (x_1 + \varepsilon, \ldots, x_d + \varepsilon)$, or multiplicative, $(x_1, \ldots, x_d) \to (\eta x_1, \ldots, \eta x_d)$, to the samples generated at each iteration of Gibbs or MH in a way that the stationarity of the target distribution is preserved, while mixing properties of chain after the transformation may perhaps be improved. To maintain stationarity, ε and η must be simulated from some appropriate distribution which is often impossible to generate from [see Liu and Yu (1999)]. To alleviate the problem Liu and Sabatti (2000) [see also Kou, Xie and Liu (2005)] suggest MH methods.

As pointed out in the supplement of Dutta and Bhattacharya (2014), these generalized Gibbs/MH methods are mere attempts towards improving mixing of the underlying MCMC. These are not stand-alone methodologies which can converge to the target by themselves, unlike TMCMC. In fact, as shown in the supplement of Dutta and Bhattacharya (2014) the aforementioned additive and multiplicative transformations of generalized Gibbs/MH approaches are themselves not even irreducible, and are hence generally non-convergent. See the supplement of Dutta and Bhattacharya (2014) for detailed discussions regarding various issues pertaining to these approaches.

It is important to point out that the aforementioned generalized Gibbs/MH approaches are not the first to consider deterministic transformations, the earlier methods being the hit-and-run algorithm [see, for example, Berbee et al. (1987), Bélisle, Romeijn and Smith (1993), Romeijn and Smith (1994), Smith (1996)] and the adaptive direction sampling algorithm [Gilks, Roberts and George (1994), Roberts and Gilks (1994)]. The hit-and-run algorithm proceeds by generating random directions from an available set of directions and then considers addition of a scalar times the sampled direction to the current state as the updated state, where the scalar is drawn from an appropriate, but non-trivial distribution associated with the target density. The adaptive direction sampler generalizes the hit-and-run algorithm by selecting the directions and the current state adaptively, based on all the previous iterations. Since sampling the scalar directly is usually difficult or impossible, this step can be replaced with an appropriate MH step as in the case of generalized Gibbs/MH methods discussed above [see, for example, Romeijn and Smith (1994)].

The adaptive version of additive TMCMC [Dey and Bhattacharya (2015)], where the scales are chosen adaptively, comes close to the Metropolized versions of hit-and-run and adaptive direction sampling methods. The differences are that, in the latter algorithms, the directions (and also the current state in adaptive direction method) are chosen randomly and require a Jacobian of transformations to be evaluated at both the numerator and denominator of the acceptance ratio apart from the proposal distribution (if not symmetric), while in adaptive additive TMCMC the directions (and the current state) are chosen deterministically, and the acceptance ratio depends neither on any Jacobian, nor on proposal densities.

2.7 Contrast of the TMCMC idea with reversible jump Markov chain Monte Carlo (RJMCMC)

Interestingly, although both TMCMC and RJMCMC are based on deterministic transformations, the philosophy of TMCMC is in sharp contrast with that of RJMCMC. First, TMCMC is designed for simulation from fixed-dimensional distributions while RJMCMC is meant for generating from variable-dimensional distributions. Second, the deterministic transformations in RJMCMC are not required for move-types that are not meant for changing dimensions, while in TMCMC

none of the moves change dimensions, yet all of them are based on deterministic transformations. Third, the acceptance rates of dimension changing moves of RJMCMC depend upon the proposal density because for moving from lower to higher dimension simulation of random variables is required but for returning to lower dimension from higher dimension no simulation is necessary. For instance, for moving from $x_1 \in \mathbb{R}$ to $(x_1^*, x_2^*) \in \mathbb{R}^2$, one may simulate $u \sim \varphi$, where φ is some density supported on $\mathbb{R}$, and then make the transformation $x_1^* = x_1 - u$ and $x_2^* = x_1 + u$. However, for returning from (x_1^*, x_2^*) to x_1, one simply sets $x_1 = \frac{x_1^* + x_2^*}{2}$, without simulating any random variable. On the other hand, as discussed in Section 2.2, since ε is simulated from ϱ while moving forward and also for moving back, the detailed balance condition dictates that the acceptance ratio of TMCMC must always be independent of ϱ.

2.8 Discussion of extension of TMCMC to variable dimensional problems

Das and Bhattacharya (2016) extend TMCMC to accommodate variable-dimensional problems; they refer to the new variable dimensional methodology as Transdimensional Transformation based Markov chain Monte Carlo (TTMCMC). TTMCMC is designed to update the entire set of parameters, both fixed and variable dimensional, as well as the number of parameters, in a single block using simple deterministic transformations of some low-dimensional (often one-dimensional) random variable drawn from some fixed, but arbitrary distribution defined on some relevant support. Again, the acceptance probability is independent of the proposal density. The advantages of TMCMC over Metropolis–Hastings algorithms are clearly carried over to the advantages of TTMCMC over RJMCMC. In fact, since it is well-known that efficient implementation of RJMCMC is generally infeasible, TTMCMC offers huge advantages in this regard; see Das and Bhattacharya (2016) for details.

3 Optimal scaling of additive TMCMC when the target density is a product based on *iid* random variables

It must be emphasized that the proposal density for ε in TMCMC can be any distribution on the positive support. Similarly, the RWM algorithm also does not require the proposal to be normal. However, the optimal scaling results for RWM inherently assume normality, and for the sake of comparison, we have also restricted our focus on $\varepsilon \sim N(0, \frac{\ell^2}{d}) I_{\{\varepsilon > 0\}}$ in the subsequent sections.

In this paper, we are primarily interested in choosing the parameters of the process judiciously so as to enhance the performance of the chain. Our method as stated above involves only a single parameter—the proposal variance, or to be more precise, the scaling factor ℓ. Details on the need for optimal scaling of additive TMCMC with regard to the scaling factor ℓ are provided in Section S-2 of the supplement.

In this section, we consider the problem of optimal scaling in the simplest case where the target density π is a product of *iid* marginals, given by

$$\pi(x) = \prod_{i=1}^{d} f(x_i). \tag{12}$$

Assuming that the TMCMC chain is started at stationarity, we shall show that for each component of X, the corresponding one-dimensional process converges to a diffusion process which is analytically tractable and whose diffusion and drift speeds may be numerically evaluated. It is important to remark that it is possible to relax the assumption of stationarity; see Jourdain, Lelièvre and Miasojedow (2013) in the context of RWM, although we do not pursue this in our current work.

Let $X_t^d = (X_{t,1}, \ldots, X_{t,d})$. We define ${U_t}^d = X_{[dt],1}$ ($[\cdot]$ denotes the integer part), the sped up first component of the actual additive TMCMC-induced Markov chain. Note that this process proposes a jump every $\frac{1}{d}$ time units. As $d \to \infty$, that is, as the dimension grows to ∞, the process essentially becomes a continuous time diffusion process.

Before proceeding first let us introduce the notion of Skorohod topology [Skorohod (1956)]. It is a topology generated by a class of functions from $[0, 1] \to \mathbb{R}$ for which the right-hand side and the left-hand side limits are well defined at each point (even though they may not be the same). It is an important tool for formulating Poisson process, Levy process and other stochastic point processes. As considered in Roberts, Gelman and Gilks (1997) here we also consider the metric separable topology on the above class of functions as defined in Skorohod (1956). In other words, whenever we mention convergence of discrete time stochastic processes to diffusion process in this paper, we mean convergence with respect to this topology.

In what follows, we assume the following:

$$E_f\left(\frac{f'(X)}{f(X)}\right)^4 < \infty, \tag{13}$$

$$E_f\left(\frac{f''(X)}{f(X)}\right)^4 < \infty, \tag{14}$$

$$E_f\left(\frac{f'''(X)}{f(X)}\right)^4 < \infty. \tag{15}$$

$$E_f\left|\frac{f''''(X)}{f(X)}\right| < \infty. \tag{16}$$

These assumptions can also be somewhat relaxed, depending upon the order of the Taylor's series expansions used in the proofs. Following Roberts, Gelman and Gilks (1997), let us denote weak convergence of processes in the Skorohod topology by $\Rightarrow$.

We next present our formal result in the *iid* situation, the proof of which is presented in Section S-6.1. Our proof differs from the previous approaches associated with RWM particularly because as already shown in Section 2, in additive TMCMC, the terms $b_i\varepsilon$ are not jointly normally distributed unlike the RWM-based approaches. Thus, unlike the RWM-based approaches, in our case obtaining appropriate normal approximation to relevant quantities are not assured. To handle the difficulty, we had to apply Lyapunov's central limit theorem on sums associated with the discrete random variables $\{b_i; i = 2, \ldots, d\}$, *conditional on* ε (and b_1). This required us to verify Lyapunov's condition [see, for example, Koralov and Sinai (2007)] before applying the central limit theorem. We then integrated over ε and b_1. These issues make our proof substantially different from the previous approaches associated with RWM. It is important to remark that, not only in this $i.i.d.$ scenario, but in all the set-ups that we consider in this paper, application of Lyapunov's central limit theorem, conditionally on ε (and often b_1), is crucial, before finally integrating over the conditioned variables to obtain our results.

Theorem 3.1. *Assume that f is positive with at least three continuous derivatives and that the fourth derivative exists almost everywhere. Also assume that $(\log f)'$ is Lipschitz continuous, and that (13)–(16) hold. Let $X_0^d \sim \pi$, that is, the d-dimensional additive TMCMC chain is started at stationarity, and let the transition be given by $(x_1, \ldots, x_d) \to (x_1 + b_1\varepsilon, \ldots, b_d\varepsilon)$, where for $i = 1, \ldots, d$, $b_i = \pm 1$ with equal probability and $\varepsilon \equiv \frac{\ell}{\sqrt{d}}\varepsilon^*$, where $\varepsilon^* \sim N(0,1)I_{\{\varepsilon^*>0\}}$. We then have*

$$\{U_t^d; t \geq 0\} \quad \Rightarrow \quad \{U_t; t \geq 0\},$$

where $U_0 \sim f$ and $\{U_t; t \geq 0\}$ satisfies the Langevin stochastic differential equation (SDE)

$$dU_t = g(\ell)^{1/2}\, dB_t + \frac{1}{2}g(\ell)\big(\log f(U_t)\big)'\, dt, \tag{17}$$

with B_t denoting standard Brownian motion at time t,

$$g(\ell) = 4\ell^2 \int_0^\infty u^2 \Phi\left(-\frac{u\ell\sqrt{\mathbb{I}}}{2}\right)\phi(u)\, du; \tag{18}$$

$\Phi(\cdot)$ and $\phi(\cdot)$ being the standard normal cumulative distribution function (cdf) and density, respectively, and

$$\mathbb{I} = E_f\left(\frac{f'(X)}{f(X)}\right)^2. \tag{19}$$

In connection with our diffusion equation (see Equation (18) in connection with the proof of Theorem 3.1 in Section S-6.1), we note that our SDE is also Langevin like the usual RWM approach. But, we have a different *speed* and it is interesting

to compare how the two *speed* functions of our method is related to that of RWM and also, how it alters the optimal expected acceptance rate of the process. In what follows, we use the terms *speed* and *diffusion speed* of the process, given by $g(\ell)$ as in (18) interchangeably.

Corollary 3.1. *The diffusion speed $g(\ell)$ is maximized by*

$$\ell_{\text{opt}} = \frac{2.426}{\sqrt{\mathbb{I}}}, \tag{20}$$

and the optimal acceptance rate is given by

$$\begin{aligned} \alpha_{\text{opt}} &= 4\int_0^\infty \Phi\left(-\frac{u\ell_{\text{opt}}\sqrt{\mathbb{I}}}{2}\right)\phi(u)\,du \\ &= 0.439 \qquad (\textit{up to three decimal places}). \end{aligned} \tag{21}$$

3.1 TMCMC within Gibbs for *iid* product densities

The main notion of Gibbs sampling is to update one or multiple components of a multidimensional random vector conditional on the remaining components. In TMCMC within Gibbs, we update only a fixed proportion c_d of the d co-ordinates, where c_d is a function of d and we assume that as $d \to \infty$, then $c_d \to c$, for some $0 < c \leq 1$. In order to explain the transitions in this process analytically, we define an indicator function χ_i for $i = 1, \ldots, d$. For fixed d,

$$\begin{aligned} \chi_i &= 1 \qquad \text{if transition takes place in the } i\text{th co-ordinate} \\ &= 0 \qquad \text{if no transition takes place in the } i\text{th co-ordinate.} \end{aligned} \tag{22}$$

Our assumptions imply that

$$P(\chi_i = 1) = c_d; \qquad i = 1, \ldots, d. \tag{23}$$

Then a feasible transition with respect to additive TMCMC can be analytically expressed as

$$(x_1, \ldots, x_d) \to (x_1 + \chi_1 b_1 \varepsilon, \ldots, x_d + \chi_d b_d \varepsilon), \tag{24}$$

where $\varepsilon \equiv \frac{\ell}{\sqrt{d}}\varepsilon^*$, where $\varepsilon^* \sim N(0,1) I_{\{\varepsilon^* > 0\}}$. We then have the following theorem, the proof of which is presented in Section S-6.2 of the supplement.

Theorem 3.2. *Assume that f is positive with at least three continuous derivatives and that the fourth derivative exists almost everywhere. Also assume that* $(\log f)'$ *is Lipschitz continuous, and that* (13)–(16) *hold. Suppose also that the transition is given by* (24) *and that as $d \to \infty$, $c_d \to c$, for some $0 < c \leq 1$. Let $X_0^d \sim \pi$, that is, the d-dimensional additive TMCMC chain is started at stationarity. We then have*

$$\left\{U_t^d; t \geq 0\right\} \quad \Rightarrow \quad \{U_t; t \geq 0\},$$

where $U_0 \sim f$ *and* $\{U_t; t \geq 0\}$ *satisfies the Langevin SDE*

$$dU_t = g_c(\ell)^{1/2}\, dB_t + \frac{1}{2} g_c(\ell)\big(\log f(U_t)\big)' \, dt, \tag{25}$$

where

$$g_c(\ell) = 4c\ell^2 \int_0^\infty u^2 \Phi\left(-\frac{u\ell\sqrt{c\mathbb{I}}}{2}\right)\phi(u)\, du, \tag{26}$$

and $\mathbb{I}$ *is given by* (19).

Corollary 3.2. *The diffusion speed* $g_c(\ell)$ *is maximized by*

$$\ell_{\text{opt}} = \frac{2.426}{\sqrt{c\mathbb{I}}}, \tag{27}$$

and the optimal acceptance rate is given by

$$\begin{aligned} \alpha_{\text{opt}} &= 4 \int_0^\infty \Phi\left(-\frac{u\ell_{\text{opt}}\sqrt{c\mathbb{I}}}{2}\right)\phi(u)\, du \\ &= 0.439 \qquad (\textit{up to three decimal places}). \end{aligned} \tag{28}$$

4 Diffusion approximation for independent but non-identical random variables

So far we have considered only those target densities π which correspond to *iid* components of x. Now, we extend our investigation to those target densities that are associated with independent but not identically distributed random variables. That is, we now consider

$$\pi(x) = \prod_{i=1}^{d} f_i(x_i). \tag{29}$$

We concentrate on a particular form of the target density involving some scaling constant parameters, as considered in Bedard (2008b), Bédard and Rosenthal (2008).

$$\pi(x) = \prod_{j=1}^{d} \theta_j(d) f\big(\theta_j(d) x_j\big). \tag{30}$$

As before, we assume that f is twice continuously differentiable with existence of third derivative almost everywhere, and that $(\log f)'$ is Lipschitz continuous. We define $\Theta(d) = \{\theta_1(d), \theta_2(d), \ldots, \theta_d(d)\}$ and we shall focus on the case where $d \to \infty$. Some of the scaling terms are allowed to appear multiple times. We assume that the first k terms of the parameter vector may or may not be identical, but the

remaining $d-k$ terms can be split into m subgroups of independent scaling terms. In other words,

$$\begin{aligned}\Theta(d) = \big(&\theta_1(d), \theta_2(d), \ldots, \theta_k(d), \theta_{k+1}(d), \ldots, \theta_{k+m}(d),\\ &\underbrace{\theta_{k+1}(d), \ldots, \theta_{k+1}(d)}_{r(1,d)-1}, \underbrace{\theta_{k+2}(d), \ldots, \theta_{k+2}(d)}_{r(2,d)-1}, \ldots,\\ &\underbrace{\theta_{k+m}(d), \ldots, \theta_{k+m}(d)}_{r(m,d)-1}\big),\end{aligned} \tag{31}$$

where $r(1,d), r(2,d), \ldots, r(m,d)$ are the number of occurrences of the parameters in each of the m distinct classes. We assume that for any i,

$$\lim_{d\to\infty} r(i,d) = \infty. \tag{32}$$

Also, we assume a particular form of each scaling parameter $\theta_i(d)$:

$$\begin{aligned}\frac{1}{\{\theta_i(d)\}^2} &= \frac{K_i}{d^{\lambda_i}}; \qquad i = 1, \ldots, k, \quad \text{and}\\ \frac{1}{\{\theta_i(d)\}^2} &= \frac{K_i}{d^{\gamma_i}}; \qquad i = k+1, \ldots, k+m.\end{aligned} \tag{33}$$

Assume that ${\theta_i}^{-2}(d)$ are so arranged that γ_i are in a decreasing sequence for $i = k+1, \ldots, k+m$ and also let λ_i form a decreasing sequence from $i = 1, \ldots, k$. According to Bedard (2007), the optimal form of the scaling variance $\sigma^2(d)$ should be of the form $\sigma^2(d) = \frac{\ell^2}{d^\alpha}$, where ℓ^2 is some constant and α satisfies

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{d^\alpha} < \infty, \quad \text{and} \quad \lim_{d\to\infty} \frac{d^{\gamma_i} r(i,d)}{d^\alpha} < \infty; \qquad i = 1, \ldots, m. \tag{34}$$

Here, let U_t^d be the process at time t sped up by a factor of d^α. That is, $U_t^d = (X_1([d^\alpha t]), \ldots, X_d([d^\alpha t]))$. We then have the following theorem, the proof of which is provided in Section S-6.3 of the supplement.

Theorem 4.1. *Assume that the target distribution is of the form* (30), *where* f *is positive with at least three continuous derivatives and that the fourth derivative exists almost everywhere. Also assume that* $(\log f)'$ *is Lipschitz continuous, and that* (13)–(16), (31), (32), (33) *and* (34) *hold. Let* $X_0^d \sim \pi$, *that is, the* d*-dimensional additive TMCMC chain is started at stationarity. Let the transition be given by* $(x_1, \ldots, x_d) \to (x_1 + b_1\varepsilon, \ldots, x_d + b_d\varepsilon)$, *where, for* $i = 1, \ldots, d$, $b_i = \pm 1$ *with equal probability and* $\varepsilon \equiv \frac{\ell}{d^{\frac{\alpha}{2}}}\varepsilon^*$, *with* $\varepsilon^* \sim N(0,1) I_{\{\varepsilon^* > 0\}}$. *We then have*

$$\big\{U_t^d; t \geq 0\big\} \quad \Rightarrow \quad \{U_t; t \geq 0\},$$

where $U_0 \sim f$ *and* $\{U_t; t \geq 0\}$ *satisfies the Langevin SDE*

$$dU_t = g_\xi(\ell)^{1/2}\, dB_t + \frac{1}{2} g_\xi(\ell) \big(\log f(U_t)\big)'\, dt, \tag{35}$$

where

$$g_\xi(\ell) = 4\ell^2 \int_0^\infty u^2 \Phi\left(-\frac{u\ell\xi\sqrt{\mathbb{I}}}{2}\right)\phi(u)\,du. \tag{36}$$

Corollary 4.1. *The diffusion speed $g_c(\ell)$ is maximized by*

$$\ell_{\text{opt}} = \frac{2.426}{\xi\sqrt{\mathbb{I}}}, \tag{37}$$

and the optimal acceptance rate is given by

$$\begin{aligned}\alpha_{\text{opt}} &= 4\int_0^\infty \Phi\left(-\frac{u\ell_{\text{opt}}\xi\sqrt{\mathbb{I}}}{2}\right)\phi(u)\,du \\ &= 0.439 \qquad (\textit{up to three decimal places}).\end{aligned} \tag{38}$$

4.1 TMCMC within Gibbs for independent but non-identical random variables

As in Section 3.1, here also we define transitions of the form (24), where χ_i, having the same definitions as (22) and (23), indicates whether or not the ith co-ordinate x_i will be updated.

The rest of the proof is a simple modification of the proof for independent but non-identical random variables provided in Section S-6.3. There we must replace

$$\frac{1}{r(i,d)}\sum_{j=1}^{r(i,d)}\left(\frac{f'(u_j)}{f(u_j)}\right)^2 \to E\left[\left\{\frac{f'(U)}{f(U)}\right\}^2\right] = \mathbb{I}$$

with

$$\frac{c_d}{c_d r(i,d)}\sum_{j=1}^{c_d r(i,d)}\left(\frac{f'(u_j)}{f(u_j)}\right)^2 \to cE\left[\left\{\frac{f'(U)}{f(U)}\right\}^2\right] = c\mathbb{I}. \tag{39}$$

With the above modification the diffusion speed can be calculated as

$$g_{c,\xi}(\ell) = 4c\ell^2 \int_0^\infty \left\{u^2 \Phi\left(-\frac{u\ell\xi\sqrt{c\mathbb{I}}}{2}\right)\right\}\phi(u)\,du. \tag{40}$$

Formally, we have the following theorem.

Theorem 4.2. *Assume that the target distribution π is of the form (30), where f is positive with at least three continuous derivatives and that the fourth derivative exists almost everywhere. Also assume that $(\log f)'$ is Lipschitz continuous, and that (13)–(16), (31), (32), (33) and (34) hold. Let $X_0^d \sim \pi$, that is, the d-dimensional additive TMCMC chain is started at stationarity. Let the transition be $(x_1, \ldots, x_d) \to (x_1 + \chi_1 b_1 \varepsilon, \ldots, \chi_d b_d \varepsilon)$, where for $i = 1, \ldots, d$, $P(\chi_i = 1) = c_d$,*

$b_i = \pm 1$ with equal probability, and $\varepsilon \equiv \frac{\ell}{d^{\frac{\alpha}{2}}}\varepsilon^*$, with $\varepsilon^* \sim N(0,1)I_{\{\varepsilon^*>0\}}$. We then have

$$\{U_t^d; t \geq 0\} \quad \Rightarrow \quad \{U_t; t \geq 0\},$$

where $U_0 \sim f$ and $\{U_t; t \geq 0\}$ satisfies the Langevin SDE

$$dU_t = g_{c,\xi}(\ell)^{1/2}\, dB_t + \frac{1}{2} g_{c,\xi}(\ell)\big(\log f(U_t)\big)' dt, \tag{41}$$

where $g_{x,\xi}(\ell)$ is given by (40).

Corollary 4.2. *The diffusion speed $g_{c,\xi}(\ell)$ is maximized by*

$$\ell_{\text{opt}} = \frac{2.426}{\xi\sqrt{c\mathbb{I}}}, \tag{42}$$

and the optimal acceptance rate is given by

$$\begin{aligned} \alpha_{\text{opt}} &= 4\int_0^\infty \Phi\left(-\frac{u\ell_{\text{opt}}\xi\sqrt{c\mathbb{I}}}{2}\right)\phi(u)\, du \\ &= 0.439 \qquad (\textit{up to three decimal places}). \end{aligned} \tag{43}$$

5 Diffusion approximation for a more general dependent family of distributions

So far, we assumed that the target density π is associated with either *iid* or mutually independent random variables, with a special structure. Now, we extend our notion to a much wider class of distributions where there is a particular form of dependence structure between the components of the distribution. In determining these non-product measures, we adopted the framework of Mattingly, Pillai and Stuart (2011), Beskos, Roberts and Stuart (2009), Beskos and Stuart (2009). For clarity, we first discuss this in the case of finite dimension d, and then discuss the generalization in infinite dimensions.

Let $x^d \in \mathbb{R}^d$ denote the first d co-ordinates of $x \in \mathbb{R}^\infty$. Let us assume that the d-dimensional target density π^d satisfies

$$\frac{d\pi^d}{d\pi_0^d}(x^d) = M_{\Psi^d}\exp(-\Psi^d(x^d)), \tag{44}$$

where Ψ^d is measurable with respect to the Borel σ-field on $\mathbb{R}^d$, M_{Ψ^d} is an appropriate normalizing constant depending upon Ψ^d, and π^d has the density

$$\pi_0^d(x^d) = \prod_{j=1}^d \frac{1}{\lambda_j}\phi\left(\frac{x_j}{\lambda_j}\right) \tag{45}$$

with respect to the Lebesgue measure. In other words, under π_0^d, $x_j \sim N(0, \lambda_j^2)$; $j = 1, 2, \ldots, d$.

Then, with respect to Lebesgue measure, π^d has the following density:

$$\pi^d(x^d) = M_{\Psi^d} \exp\bigl(-\Psi^d(x^d)\bigr) \prod_{i=1}^{d} \frac{1}{\lambda_i} \phi\left(\frac{x_i}{\lambda_i}\right). \tag{46}$$

The above finite dimensional structure can be represented in terms of projection onto the first d eigenfunctions of an appropriate covariance operator associated with a Hilbert space. Indeed, let $(\mathbb{H}, \langle\cdot\rangle, \|\cdot\|)$ denote a real, separable Hilbert space. Consider a covariance operator $\Sigma : \mathbb{H} \to \mathbb{H}$, which is self-adjoint, positive, and trace class operator on $\mathbb{H}$ with a complete orthonormal eigenbasis $\{\lambda_j^2, \phi_j\}_{j=1}^{\infty}$ such that

$$\Sigma\phi_j = \lambda_j^2 \phi_j; \qquad j = 1, 2, \ldots. \tag{47}$$

As in Mattingly, Pillai and Stuart (2011), we assume that the eigenvalues are arranged in decreasing order and $\lambda_j > 0$.

Now note that any function x in $\mathbb{R}^{\infty}$ can be uniquely represented as

$$x = \sum_{j=1}^{\infty} x_j \phi_j \qquad \text{where } x_j = \langle x, \phi_j \rangle. \tag{48}$$

The function x can be identified with its co-ordinates $\{x_j\}_{j=1}^{\infty}$ which belongs to the space of square-summable sequences. Note that Σ is diagonal with respect to the co-ordinates of this eigenbasis, and if $x_j \sim N(0, \lambda_j^2)$; $j = 1, 2, \ldots$ independently, then by the Karhunen-Loéve expansion [see, for example, Prato and Zabczyk (1992)], x follows the Gaussian measure π_0, which is an infinite dimensional generalization of (45). In particular, we assume that π_0 is a Gaussian measure with mean 0 and covariance Σ.

Now, let $\Psi^d(\cdot) = \Psi(P^d\cdot)$, where P^d denotes projection (in $\mathbb{H}$) onto the first d eigen functions of Σ, and Ψ is a real π_0-measurable function on $\mathbb{R}^{\infty}$. Then $\pi^d(x^d)$ given by (46) can be represented as

$$\pi^d(x) = M_{\Psi^d} \exp\left(-\Psi^d(x) - \frac{1}{2}\langle x, (\Sigma^d)^{-1} x\rangle\right), \tag{49}$$

where $\Sigma^d = P^d \Sigma P^d$. As $d \to \infty$, (49) approximates the target density $\pi(x)$, where the Radon Nikodym derivative of the target π with respect to the Gaussian measure π_0 is given by

$$\frac{d\pi}{d\pi_0}(x) = M_{\Psi} \exp\bigl(-\Psi(x)\bigr). \tag{50}$$

Hence, for our purpose we shall work with the finite-dimensional approximation (49); as $d \to \infty$, the appropriate piecewise linear, continuous interpolant (to be defined subsequently in Section 5.2) that is described by our additive TMCMC algorithm and associated with π^d will converge to the correct diffusion equation associated with the infinite dimensional distribution π represented by (50).

5.1 Representation of the additive TMCMC algorithm in the dependent set-up

Under the TMCMC set up, the move at the $(k+1)$th time point can be explicitly stated in terms of the position at kth time point as follows

$$x^{k+1} = \gamma^{k+1} y^{k+1} + (1 - \gamma^{k+1}) x^k, \tag{51}$$

where

$$\gamma^{k+1} \sim \text{Bernoulli}\left(\min\left\{1, \frac{\pi^d(y^{k+1})}{\pi^d(x^k)}\right\}\right).$$

We define the move y^{k+1} as

$$y^{k+1} = x^k + \sqrt{\frac{2\ell^2}{d}}\, \Sigma^{\frac{1}{2}} \xi^{k+1}, \tag{52}$$

where $\xi^{k+1} = ({b_1}^{k+1}\varepsilon^{k+1}, \ldots, {b_d}^{k+1}\varepsilon^{k+1})$ with $b_i = \pm 1$ with probability $1/2$ each, and $\varepsilon \sim N(0,1) I_{\{\varepsilon > 0\}}$. From (49), it follows that $\min\{1, \frac{\pi^d(y^{k+1})}{\pi^d(x^k)}\}$ can be written as $\min\{1, e^{\mathbb{Q}(x^k, \xi^{k+1})}\}$ where $\mathbb{Q}(x, \xi)$ is given by

$$\mathbb{Q}(x,\xi) = \frac{1}{2}\left\|\Sigma^{-\frac{1}{2}}\left(P^d x\right)\right\|^2 - \frac{1}{2}\left\|\Sigma^{-\frac{1}{2}}\left(P^d y\right)\right\|^2 + \Psi^d(x) - \Psi^d(y). \tag{53}$$

Using (52), one obtains

$$\mathbb{Q}(x,\xi) = -\sqrt{\frac{2\ell^2}{d}}\langle \eta, \xi\rangle - \frac{\ell^2}{d}\|\xi\|^2 - r(x,\xi), \tag{54}$$

where

$$\eta = \Sigma^{-\frac{1}{2}}\left(P^d x\right) + \Sigma^{\frac{1}{2}} \nabla \Psi^d(x), \tag{55}$$

and

$$r(x,\xi) = \Psi^d(y) - \Psi^d(x) - \left\langle \nabla \Psi^d(x),\, P^d y - P^d x \right\rangle. \tag{56}$$

We further define

$$R(x,\xi) = -\sqrt{\frac{2\ell^2}{d}} \sum_{j=1}^{d} \eta_j \xi_j - \frac{\ell^2}{d} \sum_{j=1}^{d} {\xi_j}^2, \tag{57}$$

and

$$R_i(x,\xi) = -\sqrt{\frac{2\ell^2}{d}} \sum_{j=1, j\neq i}^{d} \eta_j \xi_j - \frac{\ell^2}{d} \sum_{j=1, j\neq i}^{d} {\xi_j}^2. \tag{58}$$

Using Lemma 5.5 of Mattingly, Pillai and Stuart (2011), for large d one can show that

$$\mathbb{Q}(x,\xi) = R(x,\xi) - r(x,\xi) \approx R_i(x,\xi) - \sqrt{\frac{2\ell^2}{d}}\eta_i\xi_i. \tag{59}$$

Using (57) and (59), it can be seen that $\mathbb{Q}(x,\xi)$ is approximately equal to $R(x,\xi)$ as d goes to ∞, where $R(x,\xi)$ in our case is given by

$$R(x,\xi) = -\varepsilon\sqrt{\frac{2\ell^2}{d}}\sum_{j=1}^{d}\eta_j b_j - \ell^2\varepsilon^2. \tag{60}$$

Note that in the case of Mattingly, Pillai and Stuart (2011), conditional on x, $R_i(x,\xi)$ was independent of ξ_i, which enabled them to compute $E_0(\min\{1, e^{\mathbb{Q}(x,\xi)}\}\xi_i)$ by first computing it over ξ_i and then over $\xi \setminus \xi_i$. However, such independence does not hold in our case since all the components of ξ involve ε.

To obtain $E_0(\min\{1, e^{\mathbb{Q}(x,\xi)}\}\xi_i)$ in our case, we need to obtain the asymptotic distribution of $\mathbb{Q}(x,\xi)$ for large d. Since our TMCMC based proposal is not *iid*, we verify Lyapunov's central limit theorem; see Section S-7.1. For obtaining the diffusion approximation in this dependent set-up we need to obtain the expected drift and the expected diffusion coefficient. In Section S-7.2, we calculate the expected drift and in Section S-7.3, we obtain the expected diffusion coefficient.

5.2 Formal statement of our main result in the general dependent set-up

Before formally stating our result in the dependent set-up, we need to provide the explicit form of a continuous interpolant which converges to the solution of the appropriate SDE.

Note that we can construct, following Mattingly, Pillai and Stuart (2011), the following continuous interpolant

$$z^d(t) = (dt - k)x^{k+1} + (k + 1 - dt)x^k, \qquad k \le dt < k + 1. \tag{61}$$

Observe that $z^d(t)$ admits the following representation

$$z^d(t) = z^0 + \int_0^t \vartheta^d\big(\bar{z}^d(s)\big)\,ds + \sqrt{2g(\ell)}W^d(t), \tag{62}$$

where $z^0 \sim \pi$, $g(\ell) = \ell^2\beta$, $\vartheta^d(x) = dE_0(x^1 - x)$, $\bar{z}^d(t) = x^k$; $t \in [t^k, t^{k+1}]$ is a piecewise constant interpolant of x^k, where

$$t^k = k\Delta t, \qquad \eta^{k,d} = \sqrt{\Delta t}\sum_{j=1}^{k}\Gamma^{j,d}, \tag{63}$$

$$W^d(t) = \eta^{[dt],d} + \frac{dt - [dt]}{\sqrt{d}}\Gamma^{[dt]+1,d}; \qquad t \in [0,T], \tag{64}$$

where $T > 0$ is fixed.

In fact as $d \to \infty$, there exists $\widehat{W}^d \Rightarrow W$ such that $z^d(t)$ admits the following representation:

$$z^d(t) = z^0 - g(\ell) \int_0^t \big(z^d(s) + \Sigma \nabla \Psi\big(z^d(s)\big)\big)\, ds + \sqrt{2g(\ell)}\widehat{W}^d(t). \tag{65}$$

It can be shown, proceeding in the same way, and using the same assumptions on the covariance operator and Ψ as Mattingly, Pillai and Stuart (2011), that $z^d(t)$ converges weakly to z [see Mattingly, Pillai and Stuart (2011) for the rigorous definition], where z satisfies the SDE given by

$$\frac{dz}{dt} = -g(\ell)\big(z + \Sigma \nabla \Psi(z)\big) + \sqrt{2g(\ell)}\frac{dW}{dt}, \qquad z(0) = z^0, \tag{66}$$

where $z^0 \sim \pi$, W is a Brownian motion in a relevant Hilbert space with covariance operator Σ, and

$$g(\ell) = \ell^2 \beta, \tag{67}$$

is the diffusion speed.

Our result, which we state as Theorem 5.1, requires the same assumptions on the decay of eigenvalues λ_j^2 of Σ and properties of Ψ that were also required by Mattingly, Pillai and Stuart (2011). For the sake of completeness we present these assumptions below. But before that we need to define some new notation, as follows.

Using the expansion (48), following Mattingly, Pillai and Stuart (2011) we define the Sobolev spaces $\mathbb{H}^r; r \in \mathbb{R}$, where the inner products and norms are defined by

$$\langle x, y\rangle_r = \sum_{j=1}^{\infty} j^{2r} x_j y_j, \qquad \|x\|_r^2 = \sum_{j=1}^{\infty} j^{2r} x_j^2.$$

For an operator $L : \mathbb{H}^r \to \mathbb{H}^l$, we denote, following Mattingly, Pillai and Stuart (2011), the operator norm on $\mathbb{H}$ by $\|L\|_{\mathcal{L}(\mathbb{H}^r, \mathbb{H}^l)}$ defined by

$$\|L\|_{\mathcal{L}(\mathbb{H}^r, \mathbb{H}^l)} = \sup_{\|x\|_r = 1} \|Lx\|_l.$$

5.2.1 *Assumptions.*

(1) *Decay of eigenvalues* λ_j^2 *of* Σ: There exist $M_-, M_+ \in (0, \infty)$ and $\kappa > \frac{1}{2}$ such that

$$M_- \le j^{\kappa} \lambda_j \le M_+ \qquad \forall j \in \mathbb{Z}_+ = \{1, 2, 3, \ldots\}. \tag{68}$$

(2) *Assumptions on* Ψ: There exist constants $M_i \in \mathbb{R}$, $i \le 4$ and $s \in [0, \kappa - \frac{1}{2})$ such that

$$M_1 \le \Psi(x) \le M_2\big(1 + \|x\|_s^2\big) \qquad \forall x \in \mathbb{H}^s \tag{69}$$

$$\big\|\nabla \Psi(x)\big\|_{-s} \le M_3\big(1 + \|x\|_s\big) \qquad \forall x \in \mathbb{H}^s \tag{70}$$

$$\big\|\partial^2 \Psi(x)\big\|_{\mathcal{L}(\mathbb{H}^r, \mathbb{H}^l)} \le M_4 \qquad \forall x \in \mathbb{H}^s. \tag{71}$$

(3) *Assumptions on* Ψ^d: The functions Ψ^d satisfy the same conditions imposed on Ψ given by (69), (70) and (71) with the same constants uniformly across d.

Theorem 5.1. *Let assumptions* (1)–(3), *as stated in Section* 5.2.1, *hold. Let* $x^0 \sim \pi^d$, *where* π^d *is given by* (49) *and let* $z^d(t)$ *be given by* (61). *Then* z^d *converges weakly to the diffusion process* z *given by* (66) *with* $z(0) \sim \pi$.

Corollary 5.1. *The diffusion speed* $g(\ell)$ *is maximized by*

$$\ell_{\text{opt}} = \frac{2.426}{\sqrt{2}} = 1.715, \tag{72}$$

and the optimal acceptance rate is given by

$$\begin{aligned} \alpha_{\text{opt}} &= 4\int_0^\infty \Phi\Big(-\frac{\ell_{\text{opt}} u}{\sqrt{2}}\Big)\phi(u)\,du \\ &= 0.439 \qquad (\textit{up to three decimal places}). \end{aligned} \tag{73}$$

5.3 TMCMC within Gibbs for the dependent family of distributions

As before, here we define transitions of the form (24), where the random variable χ_i; $i = 1, \ldots, d$ indicates whether or not the ith coordinate of x will be updated. Formally,

$$x^{k+1} = \gamma^{k+1} y^{k+1} + (1 - \gamma^{k+1}) x^k, \tag{74}$$

where

$$\gamma^{k+1} \sim \text{Bernoulli}\Big(\min\Big\{1, \frac{\pi^d(y^{k+1})}{\pi^d(x^k)}\Big\}\Big).$$

We define the new move y^{k+1} of the same form as (52), but with the indicator variables χ_i incorporated appropriately. In other words,

$$y^{k+1} = x^k + \sqrt{\frac{2\ell^2}{d}}\,\Sigma^{\frac{1}{2}}\xi^{k+1}, \tag{75}$$

where $\xi^{k+1} = (\chi_1^{k+1} {b_1}^{k+1} \varepsilon^{k+1}, \ldots, \chi_d^{k+1} {b_d}^{k+1} \varepsilon^{k+1})$; $b_i = \pm 1$ with probability $1/2$ each, $\varepsilon \sim N(0,1) I_{\{\varepsilon > 0\}}$, and for any $k > 0$ and for $i = 1, \ldots, d$, $P(\chi_i^{k+1} = 1) = c_d$. As before, we assume that $c_d \to c$ as $d \to \infty$, where $0 < c \leq 1$.

The proof again required only minor modification to the above proof provided in the case of this dependent family of distributions. Here, additionally, we only need to take expectations with respect to χ_i^{k+1}; $i = 1, \ldots, d$, so that we now have

$$[\mathbb{Q}(x, \xi) | b_i, \varepsilon] \approx d\Big(-\ell^2\varepsilon^2 - c\varepsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i, 2\ell^2\varepsilon^2 c^2\Big).$$

Proceeding in the same manner as in the above proof, we obtain a stochastic differential equation of the same form as (66), but with $g(\ell)$ replaced with

$$g_c(\ell) = c\ell^2\beta_c, \tag{76}$$

where

$$\beta_c = 4\int_0^\infty u^2\Phi\left(-\frac{\ell u}{c\sqrt{2}}\right)\phi(u)\,du.$$

The result can be stated formally as follows:

Theorem 5.2. *Let assumptions* (1)–(3), *as stated in Section* 5.2.1, *hold. Let* $x^0 \sim \pi^d$, *where* π^d *is given by* (49) *and let* $z^d(t)$ *be given by* (61), *where* $z^d(t)$ *depends upon* x^k *and* x^{k+1} *through* $\xi^{k+1} = (\chi_1^{k+1}{b_1}^{k+1}\varepsilon^{k+1}, \ldots, \chi_d^{k+1}{b_d}^{k+1}\varepsilon^{k+1})$, *where for any* $k > 0$ *and for* $i = 1, \ldots, d$, $P(\chi_i^{k+1} = 1) = c_d$, *other definitions remaining the same as before. Then* z^d *converges weakly to the diffusion process* z *having the same form as* (66), *but* $g(\ell)$ *replaced with* $g_c(\ell)$ *given by* (76), *and as before,* $z(0) \sim \pi$.

Corollary 5.2. *The diffusion speed* $g_c(\ell)$ *is maximized by*

$$\ell_{\text{opt}} = \frac{2.426c}{\sqrt{2}} = 1.715c, \tag{77}$$

and the optimal acceptance rate is given by

$$\begin{aligned}\alpha_{\text{opt}} &= 4\int_0^\infty \Phi\left(-\frac{\ell_{\text{opt}}u}{c\sqrt{2}}\right)\phi(u)\,du \\ &= 0.439 \qquad (\textit{up to three decimal places}).\end{aligned} \tag{78}$$

6 Comparison with RWM

6.1 Comparison in the *iid* set-up

Note that for both the standard RWM algorithm and our additive TMCMC algorithm, the diffusion process reduces to the Langevin diffusion having the same form, but different diffusion speeds. For the RWM algorithm, the diffusion speed $h(\ell)$ is given by $h(\ell) = 2\ell^2\Phi(-\frac{\ell\sqrt{\mathbb{I}}}{2})$, and the optimal acceptance rate is $2\Phi(-\frac{\ell_{\text{opt}}\sqrt{\mathbb{I}}}{2})$, where ℓ_{opt} maximizes $h(\ell)$. A comparison between (18) and the above diffusion speed reveals that if, instead of the standard normal distribution, z_1^* associated with equation (13) of the supplement, corresponding to the proof of Theorem 3.1, had a distribution that assigned probability $1/2$ to each of $+1$ and -1, then the additive TMCMC-based diffusion speed would reduce to the RWM-based diffusion speed.

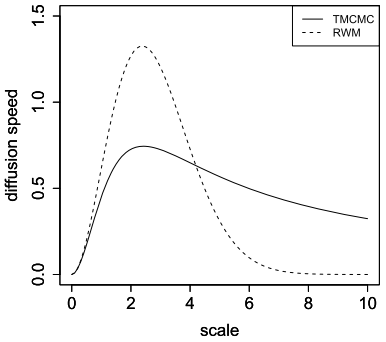


Figure 4 *Comparison of diffusion speeds of TMCMC and RWM in the iid case.*

Note that the optimum value of ℓ in RWM is $\ell_{\text{opt}} = \frac{2.381}{\sqrt{\mathbb{I}}}$ and the corresponding expected acceptance rate is 0.234. However, in TMCMC it is observed on maximizing (18) that $\ell_{\text{opt}} = \frac{2.426}{\sqrt{\mathbb{I}}}$ and the corresponding expected acceptance rate is 0.439; see Corollary 3.1. Hence, although the values of the optimizer ℓ_{opt} are close for RWMH and additive TMCMC, the optimal acceptance rate of the latter is significantly higher. This much higher acceptance rate for TMCMC is to be expected because effectively just a one-dimensional proposal distribution is used to update the entire high-dimensional random vector x.

Figure 4 compares the diffusion speeds of TMCMC and RWM in the *iid* case. Observe that the maximum diffusion speed for RWM is greater than that of TMCMC. However, the graph for RWM falls much more steeply compared to TMCMC for large ℓ, showing that the diffusion speed is quite sensitive towards misspecification of the scaling constant, and that scaling constants other than the maximizer can substantially decrease the diffusion speed. On the other hand, the graph for TMCMC is much more flat, indicating relatively more robustness with respect to the choice of ℓ.

As we will see, the same phenomenon holds for all the other set-ups, such as the target distributions with non-identical and dependent components. This is an important issue in practice for general high-dimensional target distributions, particularly with non-identical and dependent components since, as discussed in Sections S-3 and S-4, in practice, tuning the scaling constants of the proposal distributions to approximately achieve the optimal acceptance rate is generally infeasible in high dimensions, which in turn makes the maximum diffusion speed infeasible to achieve. For the RWM algorithm any such misspecification entails a sharp fall in the diffusion speed. Since in high dimensions misspecifications are very much likely, RWM is quite generally prone to sub-optimal performances. From the discussion presented in Section S-5, it can be anticipated that in very high dimensions, it may not be practically feasible to achieve the optimal acceptance rate using adaptive algorithms based on RWM. On the other hand, additive TMCMC remains far

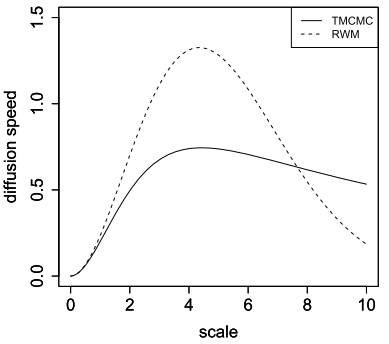


Figure 5 *Comparison of diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the iid case, with* $c = 0.3$.

more robust even in the face of such mis-specifications, thus significantly cutting down the risk of poor performance in high dimensions. Adaptive algorithms based on additive TMCMC are also demonstrated by Dey and Bhattacharya (2015) to be much more efficient compared to the adaptive RWM algorithms.

6.1.1 *Within Gibbs comparison in the iid set-up.* Now we compare TMCMC within Gibbs based diffusion speed and optimal acceptance rate given by

$$g_c(\ell) = 4c\ell^2 \int_0^\infty u^2 \Phi\left(-\frac{u\ell\sqrt{c\mathbb{I}}}{2}\right)\phi(u)\,du \tag{79}$$

(see (26) of the supplement, Section S-6.2) and (28) with those of RWM within Gibbs. The diffusion speed for the RWM within Gibbs algorithm is $h_c(\ell) = 2c\ell^2\Phi(-\frac{\ell\sqrt{c\mathbb{I}}}{2})$, and the optimal acceptance rate is $2\Phi(-\frac{\ell_{\text{opt}}\sqrt{c\mathbb{I}}}{2})$, where ℓ_{opt} maximizes $h_c(\ell)$; see Neal and Roberts (2006). It turns out that ℓ_{opt} for RWM within Gibbs is given by $\frac{2.381}{\sqrt{c\mathbb{I}}}$, and the optimal acceptance rate is 0.234, as before. Figure 5 compares the diffusion speeds associated with TMCMC within Gibbs and RWM within Gibbs, with $c = 0.3$. Once again, we observe that the diffusion speed of TMCMC within Gibbs is more robust with respect to misspecification of the scale.

6.2 Comparison in the independent but non-identical set-up

The equations

$$g_\xi(\ell) = 4\ell^2 \int_0^\infty \left\{u^2 \Phi\left(-\frac{u\ell\xi\sqrt{\mathbb{I}}}{2}\right)\right\}\phi(u)\,du \tag{80}$$

(see also (39) of the supplement) and (38) provide the diffusion speed and the optimal acceptance rate for TMCMC in the independent but non-identical set-up. The corresponding quantities for RWM are given by $2\ell^2\Phi(-\frac{\ell\xi\sqrt{\mathbb{I}}}{2})$ and $2\Phi(-\frac{\ell_{\text{opt}}\xi\sqrt{\mathbb{I}}}{2})$,

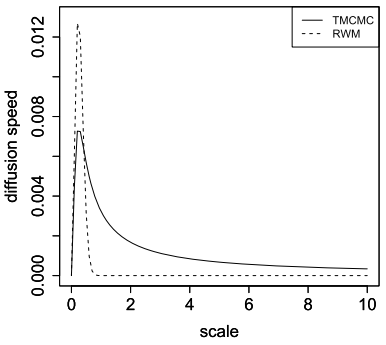


Figure 6 *Comparison of diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the independent but non-identical case, with* $\xi = 10$.

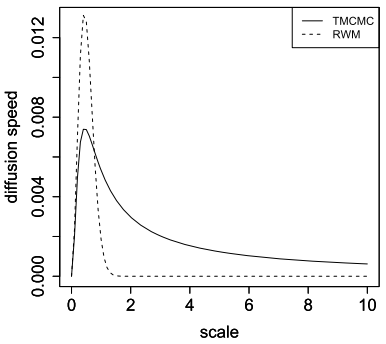


Figure 7 *Comparison of diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the independent but non-identical case, with* $\xi = 10$, $c = 0.3$.

respectively. As before, the optimal acceptance rates remain 0.234 and 0.439 for RWM and TMCMC, respectively. Figure 6 compares the diffusion speeds associated with TMCMC and RWM, with $\xi = 10$. Here both the graphs are steep, but that for RWM is much more steeper, leading to the same observations regarding robustness with respect to misspecification of scale.

6.2.1 *Within Gibbs comparison in the independent but non-identical set-up.* It can be easily shown that the RWM-based diffusion speed and the acceptance rate in the independent but non-identical set-up are $2c\ell^2\Phi(-\frac{\ell\xi\sqrt{c\mathbb{I}}}{2})$ and $2\Phi(-\frac{\ell_{\text{opt}}\xi\sqrt{c\mathbb{I}}}{2})$, respectively. These are to be compared with the TMCMC-based quantities given by (40) and (43). The optimal acceptance rates for TMCMC and RWM, as before, are 0.234 and 0.439. Conclusions similar as before are reached on observing Figure 7 that compares the diffusion speeds of TMCMC and RWM in this case.

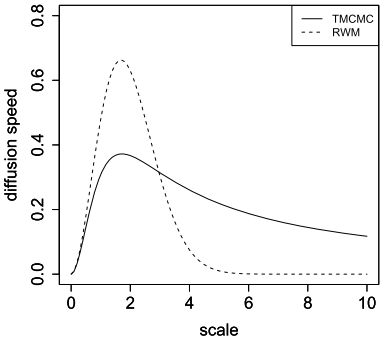


Figure 8 *Comparison of diffusion speeds of TMCMC and RWM in the dependent case.*

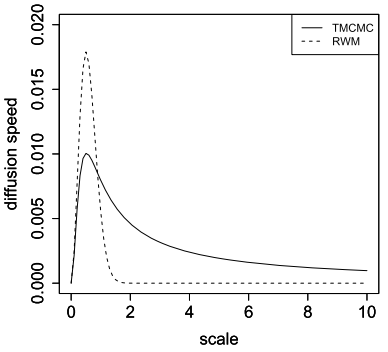


Figure 9 *Comparison of diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the dependent case, with* $c = 0.3$.

6.3 Dependent case

In the dependent case, the diffusion speed and the optimal acceptance rate of additive TMCMC are of the forms (67) and (73), respectively. As usual, the TMCMC-based optimal acceptance rate turns out to be 0.439. The corresponding RWM-based optimal acceptance rate, having the form $2\Phi(-\frac{\ell_{\text{opt}}}{\sqrt{2}})$, turns out to be 0.234 as before, where ℓ_{opt} maximizes the corresponding diffusion speed $2\ell^2\Phi(-\frac{\ell}{\sqrt{2}})$. Similar information as before are provided by Figure 8.

6.3.1 *Within Gibbs comparison in the dependent set-up.* In the dependent case, it is easily shown that the RWM-based diffusion speed and the acceptance rate are, respectively, $2c\ell^2\Phi(-\frac{\ell}{c\sqrt{2}})$ and $2\Phi(-\frac{\ell_{\text{opt}}}{c\sqrt{2}})$. The corresponding TMCMC-based quantities are (76) and (78). The optimal acceptance rates remain 0.234 and 0.439 for RWM and TMCMC. Figure 9, comparing the diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the dependent set-up, lead to similar observations as before.

7 Simulation experiments

So far, we have invested most of our efforts in the theoretical development of optimal scaling mechanism in the additive TMCMC case. Now, we shall consider some simulation experiments to illustrate the performance of our method with respect to the standard RWM methodology, under the *iid*, independent but non-identical, and dependent set-ups.

7.1 Comparison of additive TMCMC and RWM in the *iid* case

We compare the performance of RWM and TMCMC corresponding to three different choices of the proposal variance, with scalings ℓ being 2.4 (approximately optimal for both RWM and additive TMCMC) and 6 (sub-optimal for both RWM and additive TMCMC) respectively. We consider target densities of dimensions ranging from 2 to 200. For our purpose, we consider the target density π to be the multivariate normal distribution with mean vector zero and covariance matrix I, the identity matrix. The starting point x_0 is randomly generated from $U(-2, 2)$, the uniform distribution on $(-2, 2)$. The univariate density of ε for TMCMC was taken to be a left-truncated normal having mean 0 and variance $\frac{\ell^2}{d}$ for each coordinate, where ℓ is the value of the scaling constant. For RWM, each coordinate of the d dimensional proposal density was assumed to have the above distribution, but without the truncation.

In each run, the chain was observed up to 100,000 trials (including the rejected moves). The choice of burn-in was made somewhat subjectively, removing one fourth of the total number of iterates initially. This choice was actually a bit conservative as both RWM and TMCMC were found to be sufficiently close to the target density well ahead of the chosen point. We measured the efficiency of the TMCMC chain with respect to the RWM chain using certain performance evaluation measures—*Acceptance rate*, *Average Jump Size* (AJS), *Integrated Auto-Correlation Time* (IACT) and *Integrated Partial Auto-Correlation Time* (IPACT) [see Roberts and Rosenthal (2009)]. All calculations of AJS, IACT, IPACT were done corresponding to the process after burn-in in order to ensure stationarity. In calculating the integrated autocorrelation time, we considered 25 lags of ACF. IPACT was similarly computed. The first eight columns of Table 1 compare the performances of TMCMC and RWM with respect to these measures.

7.1.1 *Average Kolmogorov–Smirnov distance for comparing convergence of TMCMC and RWM.* The measures acceptance rate, IACT, IPACT and AJS do not explicitly measure how close the MCMC-based empirical distribution is to the target distribution. For this, we also considered the Kolmogorov-Smirnov (K-S) distance to evaluate the performances of the MCMC algorithms. We ran 100 copies of the RWM and TMCMC chains starting from the same initial point and with the same target density π and observed how well the empirical distribution

Table 1 *The performance evaluation of RWM and TMCMC chains for different dimensions. It is assumed that proposal has independent normal components for RWM with same proposal variance along all co-ordinates. The proposal scales are* 2.4 (*optimal*) *and* 6 (*sub-optimal*). *All calculations done after burn in*

		Test									
		Acceptance Rate (%)		IACT		IPACT		AJS		Average K-S distance	
Dimension	Scaling	RWM	TMCMC	RWM	TMCMC	RWM	TMCMC	RWM	TMCMC	RWM	TMCMC
2	2.4 (opt)	34.9	44.6	6.08	7.04	2.46	2.55	0.93	0.74	0.1651	0.1657
	6 (sub-opt)	18.66	29.15	7.08	8.08	2.52	2.56	0.79	0.62	0.1659	0.1655
5	2.4 (opt)	28.6	44.12	9.98	12.45	2.67	2.77	1.15	0.79	0.1659	0.1664
	6 (sub-opt)	2.77	20.20	15.6	14.11	2.77	2.81	0.39	0.48	0.1693	0.1674
10	2.4 (opt)	25.6	44.18	15.16	18.26	2.77	2.88	1.22	0.73	0.1667	0.1677
	6 (sub-opt)	1.37	20.34	17.55	16.31	2.91	2.86	0.25	0.49	0.1800	0.1688
100	2.4 (opt)	23.3	44.1	18.14	18.46	2.88	2.89	1.34	0.73	0.1794	0.1671
	6 (sub-opt)	0.32	20.6	18.62	18.25	2.89	2.88	0.26	0.69	0.1787	0.1684
200	2.4 (opt)	23.4	44.2	18.4	18.67	2.88	2.89	1.3	0.92	0.1813	0.1735
	6 (sub-opt)	0.33	20.7	18.86	18.74	2.89	2.89	0.09	0.54	0.1832	0.1755

corresponding to these 100 copies, after discarding the burn-in period, fits the true density by evaluating the K-S distance [Smirnov (1948)] at each time point for both the chains. As an overall measure we take the average of the K-S distances over all the time points. This averaging over the time points makes sense since the chains are assumed to be in stationarity after the burn-in period, and hence every time point must yield the same (stationary) distribution. Our average K-S distance can be viewed as quantifying how well the MCMC algorithm explores the stationary distribution after convergence is attained. The average K-S distances for RWM and TMCMC are shown in the last two columns of Table 1.

7.1.2 *Observations regarding the results presented in Table* 1. As evident from Table 1, TMCMC seems to have a uniformly better acceptance rate than RWM for all dimensions and all choices of proposal variances. There is sufficient gain in acceptance rate over RWM even for 2 dimensions and the difference increases once we move to higher dimensions or consider larger proposal variances. That large proposal variance would affect the performance of RWM is intuitively clear, because in this case getting an outlying observation in any of the d co-ordinates becomes more likely.

An interesting observation from Table 1 is that even for 2 dimensions, our acceptance ratio corresponding to the optimal scaling of 2.4 is very close to 0.44 and it remains close to the optimal value for all the dimensions considered. It is interesting to note that 0.44 is also the (non-asymptotic) optimal acceptance rate of RWM for one-dimensional proposals in certain settings obtained by minimizing the first order auto-correlation of the RWM algorithm; see Roberts and Rosenthal (2001), Roberts and Rosenthal (2009). Since in one dimension additive TMCMC is equivalent to RWM and because the former is effectively a one-dimensional algorithm irrespective of the actual dimensionality, this perhaps intuitively suggests that for TMCMC, the optimal acceptance rate will remain very close to 0.44 irrespective of dimensionality. For RWM however, the optimal acceptance rate is quite far from 0.234 for smaller dimensions. From the asymptotics perspective (setting aside the above argument regarding TMCMC being effectively one-dimensional for any actual dimension), this demonstrates that convergence to the diffusion equation occurs at a much faster rate in TMCMC as compared to RWM. Hence, even in smaller dimensions a TMCMC user can tune the proposal to achieve approximately 44% acceptance rate. Indeed, in low dimensions the tuning exercise is far more easier than in higher dimensions.

When the scale is changed from the optimum value 2.4 to the sub-optimal value 6, we witness very significant drop in the acceptance rates of RWM. Particularly for dimensions $d = 100$ and $d = 200$ the acceptance rate of RWM falls off very sharply and becomes almost negligible. In keeping with the discussion presented in Sections S-3 and S-4 of the supplement this indicates how difficult it can be in the case of general, high-dimensional target distributions, to adjust the RWM proposal to achieve the acceptance rates between 15% and 50%, as suggested by Roberts

and Rosenthal (2001). On the other hand, for any dimension, the acceptance rate of TMCMC remains more than 20%, indicating it is a lot more easier and safer to tune the TMCMC proposal.

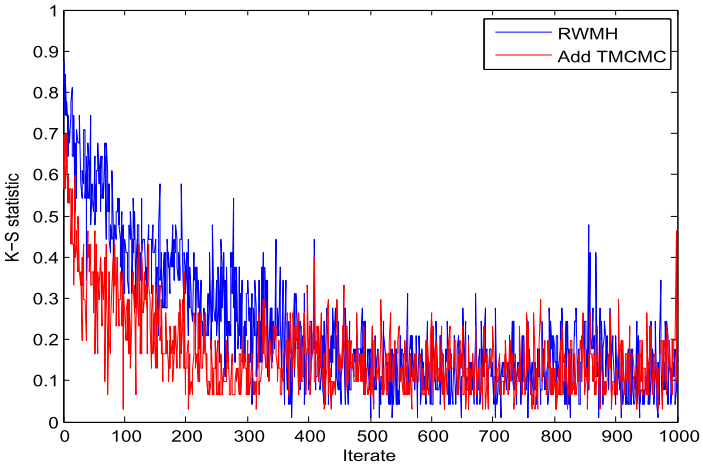
The measure IACT is uniformly higher for TMCMC for all dimensions when the optimal scale is considered. This is to be expected since the maximum diffusion speed is higher for RWM, and IACT decreases as diffusion speed increases. However, when the scale is sub-optimal, IACT of TMCMC is uniformly lower than that of RWM in all dimensions. This is in accordance with the discussion on the lack of robustness of RWM and the relative robust behaviour of the diffusion speed of TMCMC with respect to scale changes, presented in Sections S-3 and S-4 of the supplement. Indeed, the sub-optimal scale choice causes the diffusion speed of RWM to drop sharply, increasing the integrated autocorrelation in the process. On the other hand, the diffusion speed of TMCMC remains relatively more stable, thus not allowing IACT to increase significantly.

Although in the lower dimensions IPACT is slightly higher for TMCMC than for RWM, in dimensions 10, 100 and 200, it is slightly lesser for TMCMC when the scale is suboptimal (for $d = 200$ IPACT is almost the same for both the algorithms in the sub-optimal case).

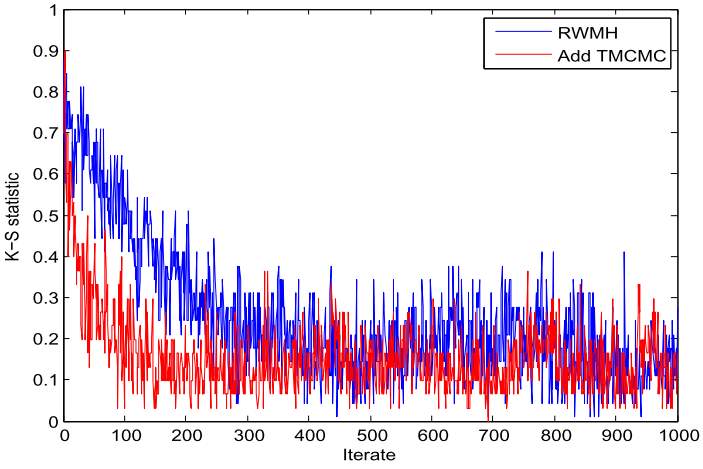
The average jump size, AJS, is uniformly somewhat larger for RWM compared to TMCMC when the scale is optimally chosen. However, for the sub-optimal scaling, AJS for TMCMC is significantly larger than those for RWM for dimensions $d = 5, 10, 100, 200$. Since in general sub-optimal scaling is to be expected, as per the discussions in Sections S-3 and S-4 of the supplement, one can expect better exploration (in terms of AJS) of the general, high-dimensional target density, by additive TMCMC.

For dimensions $d = 100$ and $d = 200$, the average K-S distance is smaller for TMCMC with respect to both optimal and sub-optimal scales. Moreover, for the sub-optimal scale, the K-S distance is uniformly smaller for TMCMC for all the dimensions considered. Furthermore, note that for the sub-optimal scale, as the dimension increases, the difference between the average K-S distances of RWM and TMCMC also increases. This suggests that at least when the scale is sub-optimal, TMCMC performs increasingly better than RWM in terms of better exploration of the target density, as dimension increases.

7.1.3 *Visualizing the rate of convergence of TMCMC and RWM to the stationary distribution using Kolmogorov–Smirnov distance.* Apart from measuring the performance of the chains after stationarity, one might be interested in visualizing how fast the chains converge to the target density starting from an initial value. In other words, it is of interest to know which of these chains have a steeper downward trend with respect to the other, when the respective optimal scales are used for both the algorithms. To investigate this empirically, we again use the K-S distance, plotting the distances with respect to the iteration number (time). Thus, while the average K-S distance, calculated after the burn-in, provides an overall measure of



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

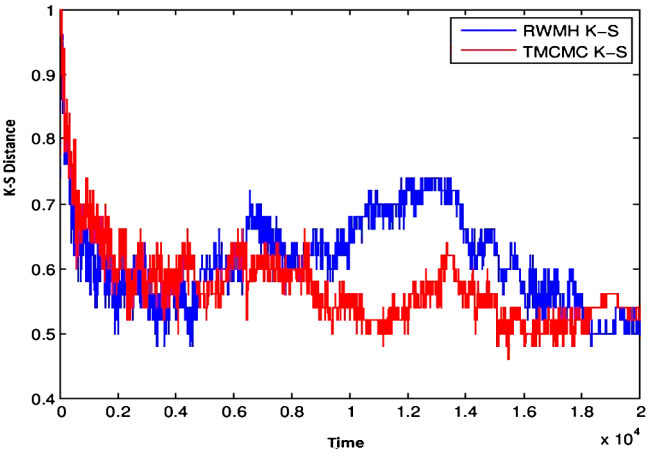


(b) $d = 30,\ \ell = 6$.

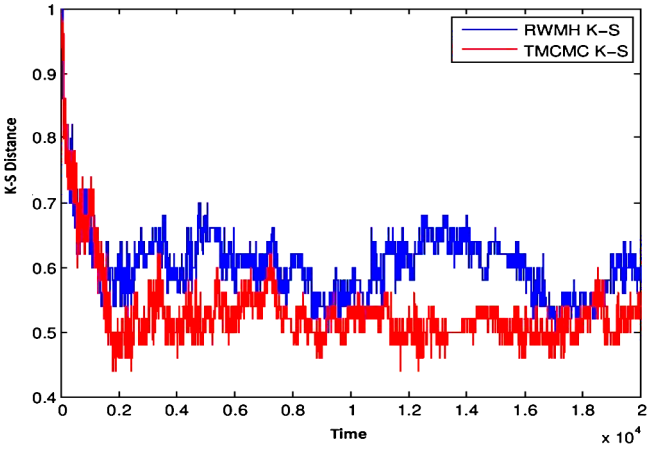
Figure 10 *K-S distance comparison before burn-in between the RWM and the TMCMC chains for dimension* $d = 30$.

how well an MCMC algorithm explores the stationary distribution after convergence, a simple plot of the K-S distances with respect to time can help visualize the rate of convergence of the MCMC algorithm to stationarity.

For smaller dimensions like 2 and 10, we did not perceive much difference between the two chains in terms of the plots of the K-S distance. But for higher dimensions, we observed a significant improvement in convergence for our TMCMC method in comparison with that of RWM. Two instances, for dimensions $d = 30$ and $d = 100$, are presented in Figures 10 and 11, respectively.



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



(b) $d = 100,\ \ell = 6$.

Figure 11 *K-S distance comparison before burn-in between the RWM and the TMCMC chains for dimension* $d = 100$.

7.2 Comparisons between additive TMCMC and RWM in the indepdendent, but non-identical set-up

We now compare additive TMCMC with RWM under an instance of independent, but non-identical situation provided in Bedard (2008a). In particular, we assume the target distribution to have independent normal components with all the means zero, and variances given by $\theta^{-2}(d) = (d^{-1/5}, d^{-1/5}, 3, d^{-0.5}, 3, d^{-0.5}, \ldots, 3, d^{-0.5})$. For our purpose, we set $d = 50$ and implement 30 chains each for additive TMCMC and RWM, each chain run for 10,000 iterations. For any given iteration,

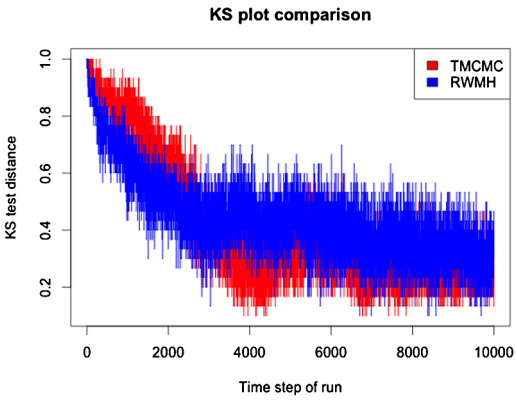


Figure 12 *Independent but non-identical case*: *K-S comparisons between additive TMCMC and RWM for* $\theta^{-2}(d) = 3$.

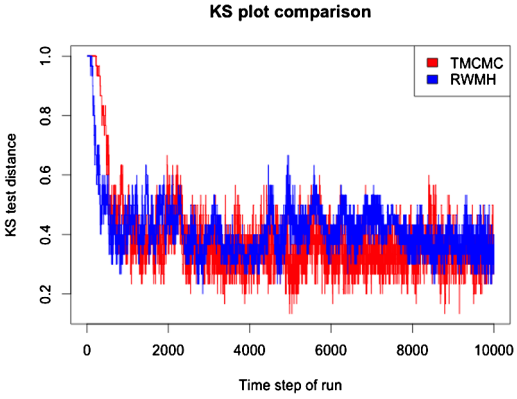


Figure 13 *Independent but non-identical case*: *K-S comparisons between additive TMCMC and RWM for* $\theta^{-2}(d) = d^{-0.5}$.

and for both TMCMC and RWM, we then compute the K-S distances based on the 30 chains, which are then compared.

Note that $\theta^{-2}(d)$ consists of three forms of co-ordinates, namely, $\theta^{-2}(d) = 3$, $\theta^{-2}(d) = d^{-0.5}$, and $\theta^{-2}(d) = d^{-\frac{1}{5}}$. These are associated with three distinct marginal target densities. In the figures below, for these three distinct marginals, we separately compare TMCMC and RWM using K-S distances. That is, Figures 12, 13 and 14 compare TMCMC and RWM when $\theta^{-2}(d) = 3$, $\theta^{-2}(d) = d^{-0.5}$, and $\theta^{-2}(d) = d^{-\frac{1}{5}}$, respectively.

All the three instances, Figures 12, 13 and 14, clearly demonstrate the superiority of additive TMCMC over RWM.

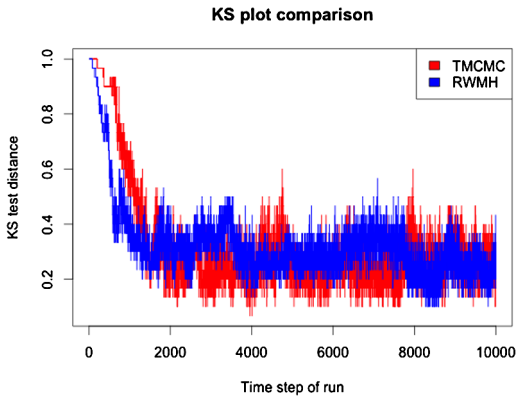


Figure 14 *Independent but non-identical case*: *K-S comparisons between additive TMCMC and RWM for* $\theta^{-2}(d) = d^{-\frac{1}{5}}$.

7.3 Comparisons between additive TMCMC and RWM in the depdendent set-up

We now compare additive TMCMC and RWM in the dependent set-up, as (46). That is, here we consider target densities of the type

$$\pi_d(x_d) = \exp(-x_d^T M_d x_d) \prod_{i=1}^{d} \frac{1}{\lambda_i} \phi\left(\frac{x_i}{\lambda_i}\right).$$

For our purpose, we consider the following forms of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ and M_d:

$$\boldsymbol{\lambda} = \alpha\left(1, \frac{1}{d}, \frac{1}{d^2}, \ldots, \frac{1}{d^d}\right),$$

and

$$M_d = \gamma(1-\rho)\mathbf{I}_d + \rho \mathbf{1}_d \mathbf{1}_d^T,$$

where $\mathbf{I}_d$ is the identity matrix of order d and $\mathbf{1}_d$ is the d-component vector with all elements 1. We report the results of our experiments with three set-ups: (a) $\rho = 0.3$, $\alpha = 0.1$, $\gamma = 100$; (b) $\rho = 0.3$, $\alpha = 0.01$, $\gamma = 100$, and (c) $\rho = 0.3$, $\alpha = 0.01$, $\gamma = 1000$. The corresponding K-S based comparisons are provided in Figures 15, 16 and 17. In all the three experiments, TMCMC very convincingly outperforms RWM. With various other choices of ρ, α and γ we observed similar results (not reported due to lack of space).

7.4 Discussion on simulation studies with multivariate Cauchy and multivariate t as target densities

We have so far restricted ourselves to comparisons between TMCMC and RWM when the target distribution is Gaussian (*iid*, independent but non-identical, and

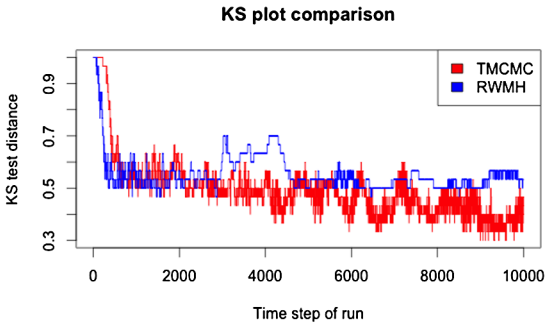


Figure 15 *Dependent case*: *K-S comparisons between additive TMCMC and RWM for* $\rho = 0.3$, $\alpha = 0.1$, $\gamma = 100$.

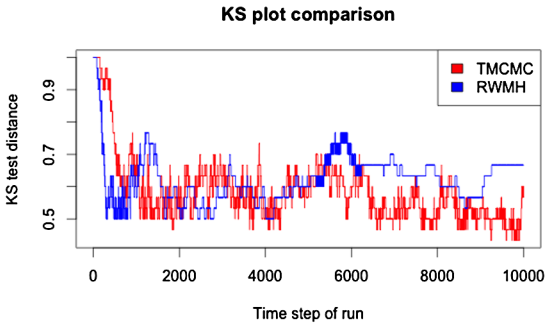


Figure 16 *Dependent case*: *K-S comparisons between additive TMCMC and RWM for* $\rho = 0.3$, $\alpha = 0.01$, $\gamma = 100$.

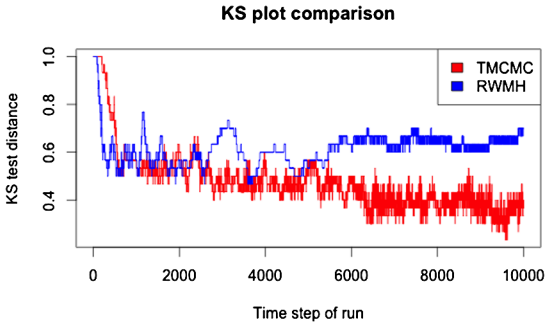


Figure 17 *Dependent case*: *K-S comparisons between additive TMCMC and RWM for* $\rho = 0.3$, $\alpha = 0.01$, $\gamma = 1000$.

dependent). However, Dey and Bhattacharya (2016a) conduct comparative studies between the two algorithms when the targets are multivariate Cauchy and multivariate t. Briefly, they compare the algorithms with respect to K-S distance,

when the location vectors and scale matrices are $\boldsymbol{\mu} = \mathbf{0}_d$ and $\boldsymbol{\Sigma} = diag\{0.7\mathbf{1}_d'\} + 0.3\mathbf{1}_d\mathbf{1}_d'$, respectively, where $\mathbf{0}_d$ is a d-dimensional vector with all elements 0. For multivariate t, they choose $\nu = 10$ degrees of freedom. For both RWMH and additive TMCMC they consider the scale of the proposal distribution to be 2.4, and illustrate the methods for dimension $d = 50$.

In fact, they compare the performances of RWM and additive TMCMC by two methods. In one method, since the above-mentioned target densities are not in the super-exponential family [see Dey and Bhattacharya (2016a) and the references therein], they transform them to superexponential distributions using a diffeomorphism proposed by Johnson and Geyer (2012), obtain samples from the transformed target densities using RWM and TMCMC, and then give inverse transformations to the simulated values so that they finally represent the original multivariate Cauchy and multivariate t. The other method is direct application of the algorithms to the original targets. In other words, Dey and Bhattacharya (2016a) also apply RWM and TMCMC directly to multivariate Cauchy and multivariate t, without resorting to diffeomorphism, and compare their performances. However, they also note that there are substantial gains with respect to mixing properties in the diffeomorphism based approach.

In either case, Dey and Bhattacharya (2016a) demonstrate that additive TMCMC outperforms RWM quite significantly in the case of the dependent, high-dimensional target densities. They even compare their performances in the case of *iid* Cauchy and t (with $\nu = 10$ degrees of freedom) distributions and reach the same conclusions.

8 Comparison of additive TMCMC and RWM in the case of a real, spatial data set

We now compare additive TMCMC and RWM with respect to a real, spatial dataset on radionuclide count data on Rongelap Island, analysed by Diggle, Tawn and Moyeed (1998) using a Bayesian hierarchical spatial model. This dataset and the model has been used subsequently by Christensen (2006) and Dutta and Bhattacharya (2014), to evaluate performances of Metropolis–Hastings and TMCMC algorithms, respectively.

8.1 Model and prior specification

For $i = 1, \ldots, 157$, Diggle, Tawn and Moyeed (1998) model the radionuclide count data as

$$Y_i \sim \text{Poisson}(M_i),$$

where

$$M_i = t_i \exp\{\beta + S(\mathbf{x}_i)\};$$

t_i is the duration of observation at location $\mathbf{x}_i$, β is an unknown parameter and $S(\cdot)$ is a zero-mean Gaussian process with isotropic covariance function of the form

$$\operatorname{Cov}\big(S(\mathbf{z}_1), S(\mathbf{z}_2)\big) = \sigma^2 \exp\{-(\alpha\|\mathbf{z}_1 - \mathbf{z}_2\|)^{\delta}\}$$

for any two locations $\mathbf{z}_1$, $\mathbf{z}_2$. In the above, $\|\cdot\|$ denotes the Euclidean distance between two locations, and $(\sigma^2, \alpha, \delta)$ are unknown parameters. Following Christensen (2006), we set $\delta = 1$, and assume uniform priors on the entire parameter space corresponding to $(\beta, \log(\sigma^2), \log(\alpha))$. Thus, there are 160 parameters to be updated in each iteration of additive TMCMC and RWM.

8.2 Optimal scaling

Note that the likelihood times the prior in this case can be approximately expressed as (46), that is, although the Poisson likelihood is expressible in the form $\exp(-\Psi^d(x^d))$, the Gaussian process prior for $S(\cdot)$ does not of course admit the form $\prod_{i=1}^{d} \frac{1}{\lambda_i}\phi(\frac{x_i}{\lambda_i})$ because of its dependence structure. Hence, this is an instance of a target density which does not fall within the class of densities for which optimal scaling theory has been developed. As is recommended in general cases, one may attempt tuning the parameters to approximately achieve the optimal acceptance rate. But this is a difficult task because of the large dimensionality, as already discussed. A far more important cause for concern is that, even if one succeeds in approximating the optimal acceptance rate, the corresponding scales will generally still be sub-optimal, because of the existence of many solutions such that the optimal acceptance rate holds. Indeed, we devise a method for approximately obtaining the optimal acceptance rates, but show that the corresponding scales lead to really poor performance of RWM, while thanks to the robustness property of additive TMCMC, the latter yields much reasonable performance. Details follow.

8.2.1 *Pilot TMCMC for facilitating approximate optimal scaling.* We first consider a pilot TMCMC run consisting of 11×10^6 iterations, with the same TMCMC algorithm used by Dutta and Bhattacharya (2014). In other words, for the pilot run we draw $\varepsilon \sim N(0, 1)\mathbb{I}(\varepsilon > 0)$, and consider the following additive transformations:

$$\begin{aligned}
T(\beta, \varepsilon) &= \beta \pm 2\varepsilon,\\
T\big(\log(\sigma^2), \varepsilon\big) &= \log(\sigma^2) \pm 5\varepsilon,\\
T\big(\log(\alpha), \varepsilon\big) &= \log(\alpha) \pm 5\varepsilon,\\
T\big(S(\mathbf{x}_i), \varepsilon\big) &= S(\mathbf{x}_i) \pm 2\varepsilon \qquad \text{for } i = 1, \ldots, 157,
\end{aligned}$$

where "+" and "−" occur with probability 1/2 each.

After discarding the first 10^6 iterations as burn-in, we then store one TMCMC sample in every 100 iterations to yield 10^5 thinned TMCMC realizations. With these stored realizations, we then obtain the empirical variance-covariance matrix of the 160 unknowns and store the 160 eigenvalues of the matrix.

8.2.2 *Approximately optimal acceptance rates of additive TMCMC and RWM using the stored eigenvalues.* For $i = 1, \ldots, 160$, letting θ_i denote the unknowns, and λ_i, the corresponding eigenvalues, we then consider the following additive transformations for TMCMC:

$$T(\theta_i, \varepsilon) = \theta_i \pm c_{\text{TMCMC}} \sqrt{\frac{2\lambda_i \ell^2_{\text{opt,TMCMC}}}{d}} \varepsilon,$$

where $\varepsilon \sim N(0,1)\mathbb{I}(\varepsilon > 0)$, "+" and "−" occur with probability 1/2 each, $\ell_{\text{opt,TMCMC}} = 1.715$ (see (72)), and c_{TMCMC} is a tuning parameter for adjusting the acceptance rate to about 44%. It turned out, after setting $c_{\text{TMCMC}} = 0.95$, that the empirical acceptance rate obtained from a TMCMC run of length 1.01×10^8, after discarding the first 1.7×10^7 iterations as burn-in, is very accurately approximated as 0.439. Implementing this TMCMC algorithm, we store one in every 100 realizations after the burn-in to obtain 8.5×10^5 thinned additive TMCMC realizations from the posterior distribution. The total time for the implementation took 46 hours and 51 minutes on a 64 bit machine with CPU MHz 1600 and about 8 GB memory.

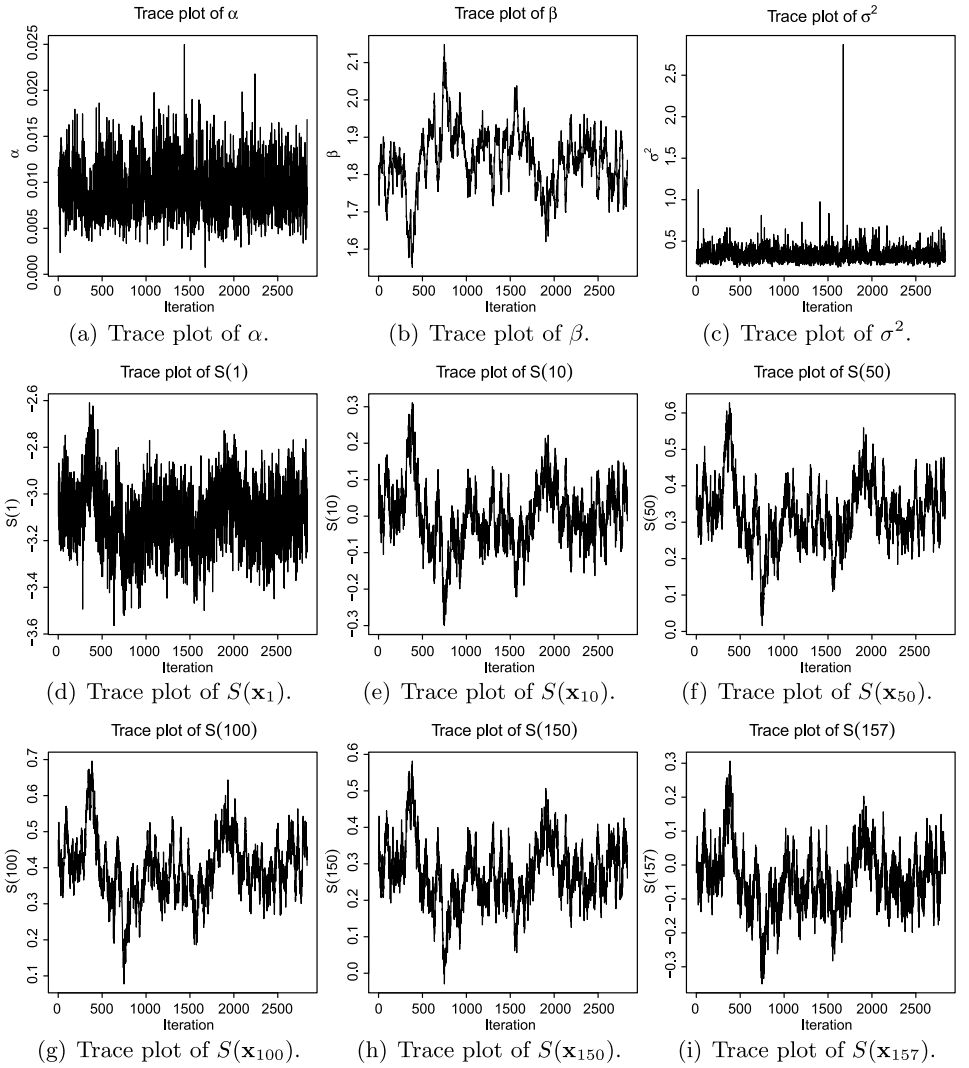
For RWM, we consider the following proposal:

$$T(\theta_i, \varepsilon_i) = \theta_i + c_{\text{RWM}} \sqrt{\frac{2\lambda_i \ell^2_{\text{opt,RWM}}}{d}} \varepsilon_i,$$

where $\varepsilon \overset{iid}{\sim} N(0,1)$, $\ell_{\text{opt,RWM}} = 1.715$, and $c_{\text{RWM}} = 0.95 = c_{\text{TMCMC}}$. We obtain the empirical acceptance rate from a RWM run of length 1.01×10^8, after discarding the frst 1.7×10^7 iterations as burn-in, as approximately 0.228. We implement this RWM algorithm, storing one in every 100 realizations after the burn-in to obtain 8.5×10^5 thinned RWM realizations from the posterior. The total time for the implementation took 46 hours and 53 minutes on the same machine on which TMCMC was implemented.

8.3 Results of comparison

Figures 18 and 19 show the trace plots of the stored realizations obtained by TMCMC and RWM, respectively, after a further thinning of size 300. Such substantial further thinning is required to facilitate effortless visual comparison of the autocorrelation plots for TMCMC and RWM shown in Figure 20. In Figures 18 and 19, it is worth observing that, RWM, composed of 160 ε_i's in this example, is prone to require a large number of iterations to return to any given set with positive posterior probability, once it leaves it. On the other hand, TMCMC marches off to convergence much faster than RWM, exploiting its more localized move types thanks to a single ε. This insight is more formalized by the comparison of



(a) Trace plot of α. (b) Trace plot of β. (c) Trace plot of σ^2.

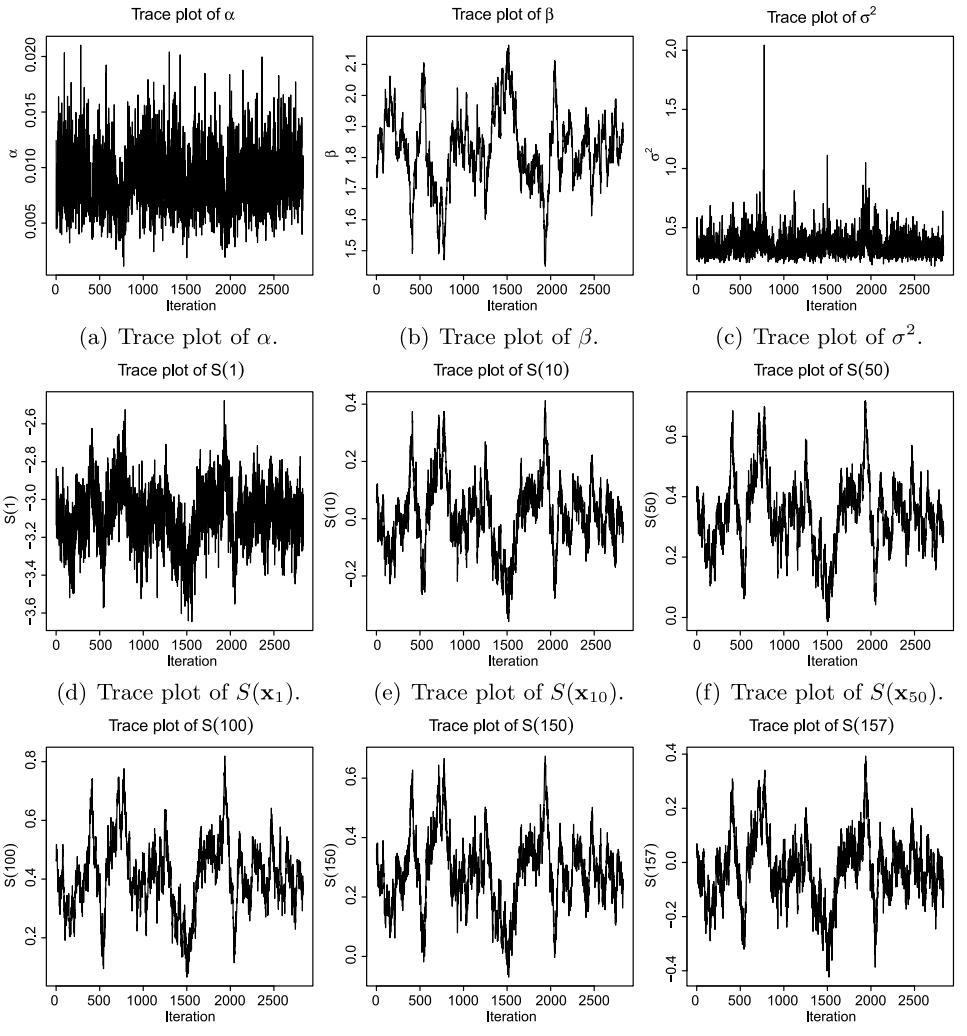
(d) Trace plot of $S(\mathbf{x}_1)$. (e) Trace plot of $S(\mathbf{x}_{10})$. (f) Trace plot of $S(\mathbf{x}_{50})$.

(g) Trace plot of $S(\mathbf{x}_{100})$. (h) Trace plot of $S(\mathbf{x}_{150})$. (i) Trace plot of $S(\mathbf{x}_{157})$.

Figure 18 *Rongelap island data*: *TMCMC based trace plots.*

the associated autocorrelation plots shown in Figure 20. It is obvious that TMCMC significantly outperforms RWM in terms of autocorrelations in all the cases.

Since scaling is directly related to autocorrelation (see Sections S-3 and S-4 of the supplement), it is clear that poor scaling of RWM in comparison with additive TMCMC is the reason for the relatively poor performance of the former. Indeed, even though we could approximately achieve the desired acceptance rates, there are many solutions for the scales, given the same acceptance rate; in fact, selecting reasonable scales gets increasing difficult with increasing dimensions. Thus, it is



(a) Trace plot of α. (b) Trace plot of β. (c) Trace plot of σ^2.

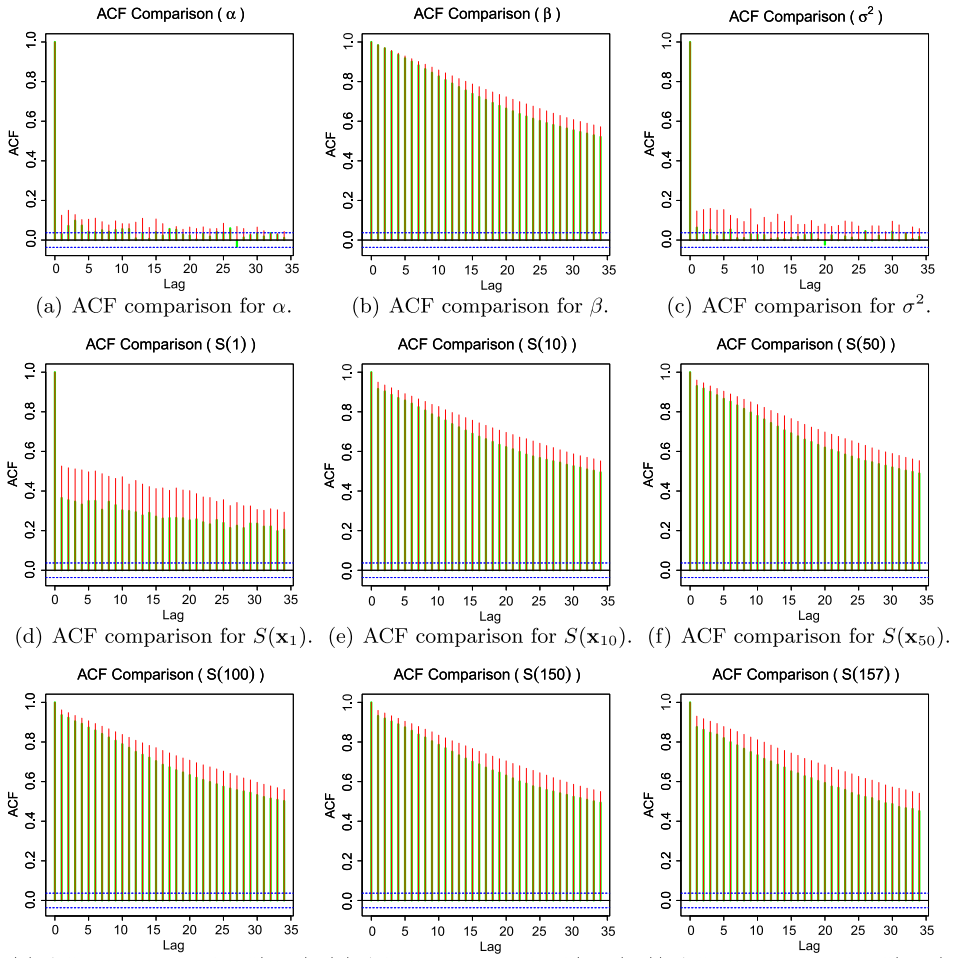
(d) Trace plot of $S(\mathbf{x}_1)$. (e) Trace plot of $S(\mathbf{x}_{10})$. (f) Trace plot of $S(\mathbf{x}_{50})$.

(g) Trace plot of $S(\mathbf{x}_{100})$. (h) Trace plot of $S(\mathbf{x}_{150})$. (i) Trace plot of $S(\mathbf{x}_{157})$.

Figure 19 *Rongelap island data*: *RWM based trace plots*.

highly unlikely that the chosen scales are even reasonable in this high-dimensional example, for either TMCMC or RWM. As a result it makes sense to conclude with respect to the autocorrelations that, in this real data study, sensitivity of RWM with respect to optimal scales is the reason for its relatively poor performance, while robustness of TMCMC in this regard is the reason for its quite reasonable performance.

Thus, in this real data example, additive TMCMC very clearly and very convincingly outperforms RWM.



(a) ACF comparison for α. (b) ACF comparison for β. (c) ACF comparison for σ^2.

(d) ACF comparison for $S(\mathbf{x}_1)$. (e) ACF comparison for $S(\mathbf{x}_{10})$. (f) ACF comparison for $S(\mathbf{x}_{50})$.

(g) ACF comparison for $S(\mathbf{x}_{100})$. (h) ACF comparison for $S(\mathbf{x}_{150})$. (i) ACF comparison for $S(\mathbf{x}_{157})$.

Figure 20 *Rongelap island data*: *Comparisons of the ACF's based on TMCMC and RWM. The darker hue depicts the TMCMC-based ACF's and the lighter hue corresponds to the RWM-based ACF's.*

9 Conclusion

Overall, our assessment is that TMCMC is clearly advantageous compared to RWM from various perspectives. It has significantly less computational complexity and the acceptance rate corresponding to the optimal scaling for TMCMC (0.439) is almost twice that of RWM (0.234). Although the maximum diffusion speed of RWM is somewhat higher than that of additive TMCMC, the latter is much more robust with respect to misspecifications of the scales. The advantages of such robustness are spelt out in the discussions in Sections S-3 and S-4 of the

supplement. Our simulation studies reported in Section 7 and Table 1, and the real data analysis in Section 8, clearly vindicate these discussions.

Related to the discussions on robustness and the difficulty of choosing proper scalings in high dimensions is also the issue of increasing computational complexity, particularly in the Bayesian paradigm. Note that complex, high-dimensional posteriors routinely arise in Bayesian applications. It is extremely uncommon among MCMC practitioners to use the RWM algorithm for updating all the parameters in a single block associated with any significantly high-dimensional posterior arising from any complex Bayesian application. We presume that the extreme difficulty of determining proper scalings in practice prevent the researchers from using the RWM as an algorithm for updating all the parameters in a single block. Indeed, as we demonstrated with our simulation study reported in Table 1, misspecification even in the case of the simple target distribution being a product of *iid* normal densities, leads to acceptance rates that are almost zero. In the context of the real data study in Section 8, we proposed a method for approximately obtaining the presumed optimal acceptance rates, and which appears to be generally applicable, but as we demonstrated, RWM failed to exhibit adequate mixing properties. Adaptive strategies may be thought of as alternative methods, but these are yet to gain enough popularity among applied MCMC practitioners; moreover, as we mention in Section S-5 of the supplement, extremely long runs may be necessary to reach adequate acceptance rates for adaptive RWM, which may be prohibitive in very high dimensions, for example, when the acceptance ratio involves high-dimensional matrix inversions at every iteration, such as in our spatial example.

The aforementioned difficulties force the researchers to use RWM to *sequentially* update the parameters, either singly, or in small blocks. Since one (or just a few) parameters are updated at a time by RWM, the acceptance rate can be controlled at each stage of the sequential updating procedure. However, this sequential procedure also requires computation of the acceptance ratio as many times as every small block is updated in a sequence. If each parameter is updated singly (that is, each small block consists of only one element), then the computational complexity increases d-folds compared to the procedure where all the d parameters are updated in a single block. Thus, when d is large, the computation can become prohibitively slow.

On the other hand, TMCMC is designed to update all the parameters in a single block in such a way that the acceptance rate remains reasonable in spite of the high dimensionality and complexity of the target distribution. Our simulation studies and real data example in Section 8 show that mis-specification of the scales do not have drastic effect on the efficiency of additive TMCMC, thanks to its robustness property. As a result, with much less effort compared to that required for RWM, we can achieve reasonable scalings that ensure adequate performance of additive TMCMC, so that resorting to sequential updating will not be necessary.

This also implies that unlike RWM, additive TMCMC can save enormous computational effort when the dimension d is large. Finally, adaptive TMCMC may

be of much value in very high dimensions because of its quick convergence to the correct optimal acceptance rate, and for ensuring good performance. The details will be covered in Dey and Bhattacharya (2015).

Our empirical findings reported in this article with respect to the simulation studies pertaining to *iid*, independent but non-identical, as well as dependent cases clearly point towards supremacy of TMCMC over RWM. Quite importantly, in the real, spatial data example, TMCMC outperformed RWM very significantly. Indeed, even though we could tune the scales so as to achieve approximately the respective optimal acceptance rates, the chosen scales need not be actually optimal, for either of TMCMC and RWM. Here RWM is convincingly outperformed by TMCMC thanks to its remarkably robust nature with respect to the choice of scales. Thus, all our experiments, particularly, the challenging real data example, lead us to clearly recommend TMCMC in general situations.

Given the importance of the general TMCMC idea, we have decided to create a software for its general usage. In this regard, we have now made available an R package **tmcmcR** for implementing TMCMC along with its adaptive versions at the Github page https://github.com/kkdey/tmcmcR. The software will be continuously updated in accordance with further developments of TMCMC; moreover, TTMCMC, the variable-dimensional version of TMCMC, will also be incorporated, and kept updated.

As part of our future work, we plan to extend TMCMC to multiple-try TMCMC, and investigate the corresponding optimal scaling theory. By multiple-try TMCMC we mean the TMCMC algorithm that selects the next proposal from a set of available, perhaps dependent, proposals. For MH-adapted versions of such an idea, see, for example, Liu, Liang and Wong (2000) and Liang, Liu and Caroll (2010), Martino and Read (2013). Bédard, Douc and Moulines (2012) investigated scaling analysis of such methods in the MH context. The advantages of TMCMC over MH quite reasonably lead us to expect substantial gains of multiple-try TMCMC over multiple-try MH.

Acknowledgments

We are sincerely grateful to the two reviewers for very encouraging and constructive comments which helped substantially improve the quality and presentation of our manuscript.

Supplementary Material

Supplement to "A brief tutorial on transformation based Markov Chain Monte Carlo and optimal scaling of the additive transformation" (DOI: 10.1214/16-BJPS325SUPP; .pdf). Additional details are provided in this supplementary material, whose sections and figures have the prefix "S-" when referred

to in this article. Briefly, in Section S-1, we provide details on computational efficiency of TMCMC. Specifically, we demonstrate with an experiment the superior computational speed of additive TMCMC in comparison with RWM, particularly in high dimensions. In Section S-2 we discuss, with appropriate experiments, the necessity of optimal scaling in additive TMCMC, while in Sections S-3 and S-4 we delve into the robustness issues associated with the scale choices of additive TMCMC and RWM. In Section S-5, we include brief discussions of adaptive versions of RWM and TMCMC. Moreover, the proofs of all our technical results are provided in Sections S-6 and S-7 of the supplement.

References

Bédard, M. (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *The Annals of Applied Probability* **17**, 1222–1244. MR2344305

Bédard, M. (2008a). Efficient sampling using Metropolis algorithms: Applications of optimal scaling results. *Journal of Computational and Graphical Statistics* **17**, 312–332. MR2439962

Bédard, M. (2008b). Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. *Stochastic Processes and their Applications* **118**, 2198–2222. MR2474348

Bédard, M., Douc, R. and Moulines, E. (2012). Scaling analysis of multiple-try MCMC methods. *Stochastic Processes and their Applications* **122**, 758–786. MR2891436

Bédard, M. and Rosenthal, J. S. (2008). Optimal scaling of Metropolis algorithms: Heading toward general target distributions. *Canadian Journal of Statistics* **36**, 483–503. MR2532248

Bélisle, C. J. P., Romeijn, H. E. and Smith, R. L. (1993). Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research* **18**, 255–266. MR1250117

Berbee, H. C. P., Boender, C. G. E., Rinnooy Kan, A. H. G., Scheffer, C. L., Smith, R. L. and Telgen, J. (1987). Hit-and-run algorithms for the identification of nonredundant linear inequalities. *Mathematical Programming* **37**, 184–207. MR0883020

Beskos, A., Roberts, G. O. and Stuart, A. M. (2009). Optimal scalings for local Metropolis–Hastings chains on non-product targets in high dimensions. *The Annals of Applied Probability* **19**, 863–898. MR2537193

Beskos, A. and Stuart, A. M. (2009). MCMC methods for sampling function space. In *ICIAM07: 6th International Congress on Industrial and Applied Mathematics* (R. Jeltsch and G. Wanner, eds.) 337–364. Zürich: European Mathematical Society. MR2588600

Christensen, O. F. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics* **15**, 1–17. MR2269360

Das, M. and Bhattacharya, S. (2016). Transdimensional transformation based Markov chain Monte Carlo. Preprint. Available at https://arxiv.org/abs/1403.5207.

Dey, K. K. and Bhattacharya, S. (2016). Adaptive transformation based Markov chain Monte Carlo. Manuscript under preparation.

Dey, K. K. and Bhattacharya, S. (2016a). On geometric ergodicity of additive and multiplicative transformation based Markov chain Monte Carlo in high dimensions. *Brazilian Journal of Probability and Statistics*. To appear. Available at https://arxiv.org/abs/1312.0915.

Dey, K. K. and Bhattacharya, S. (2016b). Supplement to "A brief tutorial on transformation based Markov Chain Monte Carlo and optimal scaling of the additive transformation." DOI:10.1214/16-BJPS325SUPP.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299–350. MR1626544

Dutta, S. (2012). Multiplicative random walk Metropolis–Hastings on the real line. *Sankhya B* **74**, 315–342. MR3046902

Dutta, S. and Bhattacharya, S. (2014). Markov chain Monte Carlo based on deterministic transformations. *Statistical Methodology* **16**, 100–116. Also available at arXiv:1306.6684. Supplement available at arXiv:1106.5850. MR3110892

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.) 3–48. New York: Chapman & Hall/CRC. MR2858443

Gilks, W. R., Roberts, G. O. and George, E. I. (1994). Adaptive direction sampling. *The Statstician* **43**, 179–189.

Johnson, L. T. and Geyer, C. J. (2012). Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm. *The Annals of Statistics* **40**, 3050–3076. MR3097969

Jourdain, B., Lelièvre, T. and Miasojedow, B. (2013). Optimal scaling for the transient phase of the random walk Metropolis algorithm: The mean-field limit. Preprint. Available at arXiv:1210.7639v2. MR3349007

Koralov, L. B. and Sinai, Y. G. (2007). *Theory of Probability and Random Processes*. New York: Springer. MR2343262

Kou, S. C., Xie, X. S. and Liu, J. S. (2005). Bayesian analysis of single-molecule experimental data. *Applied Statistics* **54**, 469–506. MR2137252

Liang, F., Liu, C. and Caroll, R. (2010). *Advanced Markov chain Monte Carlo methods: Learning from past samples*. New York: Wiley. MR2828488

Liu, J. S., Liang, F. and Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association* **95**, 121–134. MR1803145

Liu, J. S. and Sabatti, S. (2000). Generalized Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* **87**, 353–369. MR1782484

Liu, J. S. and Yu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274. MR1731488

Martino, L. and Read, J. (2013). On the flexibility of the design of multiple try Metropolis schemes. *Computational Statistics* **28**, 2797–2823. MR3141364

Mattingly, J. C., Pillai, N. S. and Stuart, A. M. (2011). Diffusion limits of the random walk Metropolis algorithm in high dimensions. *The Annals of Applied Probability* **22**, 881–930. MR2977981

Neal, P. and Roberts, G. O. (2006). Optimal scaling for partially updating MCMC. *Algorithms. The Annals of Applied Probability* **16**, 475–515. MR2244423

Prato, G. D. and Zabczyk, J. (1992). *Stochastic Equations in Infinite Dimensions*. *Encylopedia of Mathematics and Its Applications* **44**. Cambridge: Cambridge University Press. MR1207136

Roberts, G., Gelman, A. and Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability* **7**, 110–120. MR1428751

Roberts, G. O. and Gilks, W. R. (1994). Convergence of adaptive direction sampling. *Journal of Multivariate Analysis* **49**, 287–298. MR1276441

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science* **16**, 351–367. MR1888450

Roberts, G. O. and Rosenthal, R. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* **18**, 349–367. MR2749836

Romeijn, H. E. and Smith, R. L. (1994). Simulated annealing for constrained global optimization. *Journal of Global Optimization* **5**, 101–126. MR1291094

Skorohod, A. V. (1956). Limit theorems for stochastic processes. *Theory of Probability and its Applications* **1**, 261–290. MR0084897

Smirnov, N. V. (1948). Tables for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **19**, 279–281. MR0025109

Smith, R. L. (1996). The hit-and-run sampler: A globally reaching Markov sampler for generating arbitrary multivariate distributions. In *Proceedings of the 1996 Winter Simulation Conference* (J. M. Charnes, D. J. Morrice, D. T. Brunner and J. J. Swain, eds.), 260–264.

Storvik, G. (2011). On the flexibility of Metropolis–Hastings acceptance probabilities in auxiliary variable proposal generation. *Scandinavian Journal of Statistics* **38**, 342–358. MR2829604

Department of Statistics
University of Chicago
Chicago, IL 60637
USA

Interdisciplinary Statistical Research Unit
Indian Statistical Institute
Kolkata, 700108
India
E-mail: bhsourabh@gmail.com