On ergodic behaviour of Additive Transformation based Markov Chain Monte Carlo

By

Kushal Kumar Dey

M.Stat 2nd year

Indian Statistical Institute, Kolkata

Dissertation

Dr. Sourabh Bhattacharya
(Advidor)

Dr. Ayan Basu

Dr. Subir K. Bhandari

On ergodic behavior of Additive Transformation based Markov Chain Monte Carlo

Kushal Kumar Dey

M.Stat 2nd year
Indian Statistical Institute, Kolkata

ABSTRACT

The advent of Markov Chain Monte Carlo and its widespread applications has revolutionized statistical literature since 1990. Despite its tremendous success, it has certain limitations that have formed the basis of modern day research aimed at improving upon the existing algorithm in terms of convergence, computational complexity etc. Random Walk Metropolis Hastings (RWMH) algorithm is the most widely used version of MCMC in higher dimensions. But, besides having huge computational complexity for high dimensions, it has been found to have pretty low acceptance rate for even moderately high dimensions like 10-20, and this becomes almost close to 0 for very high dimensions. In current times, bulk of the data we encounter, be it in Microarray analysis, Systems Biology, Machine learning, have huge dimensions - ranging from a few hundreds to even millions. For such processes, RWMH algorithm will scarcely move and so the entire purpose of using such techniques in high dimensions will fail completely.

In this article, we have presented a new technique- Additive Transformation based Markov Chain Monte Carlo (TMCMC)- that works on the abstraction of using a single variable to update all the co-ordinates of the process instead of separately updating each co-ordinate as is done in RWMH. This method not only has less time complexity but also leads to much better acceptance rate especially for higher dimensions. The main goal of our research has been to establish the stochastic stability and the ergodicity properties of

TMCMC. In the following chapters, we shall discuss the geometric ergodicity behavior, optimal scaling (for i.i.d samples and samples with a special dependence structure) and finally we shall focus on adaptive versions of this approach. Both from a theoretical viewpoint and from practical perspective (through an extensive simulation study), the performance of TMCMC has been evaluated and compared with that of RWMH. We shall finally discuss about the ongoing and future works of ours and the strength and limitations of TMCMC.

Acknowledgments

Foremost, I would like to express my gratitude towards my advisor Dr. Sourabh Bhattacharya for his support motivation and patience, which has helped me all the time in the time of my research and even in writing this Masters' thesis. I could not have imagined a better supervisor and mentor for my Masters' dissertation.

Besides my advisor, I would like to thank my thesis readers Dr. Anup Dewanji, Dr. Ayan Basu and Dr. Subir K. Bhandari for making out time from their busy schedule for my thesis presentation and for the insightful comments and support.

I would like to thank Indian Statistical Institute, for giving me the opportunity to work on such an interesting research field in my final year of Masters' study, and providing me the much needed research experience, thereby preparing myself for my upcoming 5 years as a PhD graduate student.

Finally, I would use this opportunity to thank my Mom and Dad for everything , and all my friends at ISI, Kolkata for being a part of my journey throughout these 5 years at this prestigious Institute.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Monte Carlo methods have everyday use in Statistics and other disciplines like Computer Science, Systems Biology and Astronomy in today's times. This technique of generating random samples from very high dimensional spaces involving very complicated data likelihoods and posterior distributions has simplified many pressing real life problems in the recent times. In particular, Bayesian computation, simulation from complex posterior distribution and asymptotics of Bayesian algorithms have benefited a lot from this mechanism (see Gelfand and Smith [GS90],Tierney [Tie94], Gilks *et al* [GS96]). A very standard approach of simulating from complicated distributions is to use the Metropolis-Hastings (MH) algorithm [Has70][MRR53]. However, there are obvious scopes for improving upon this algorithm, pertaining mainly to the choice of proper proposal distribution and the time-complexity associated with the process. The computational efficiency is of utmost significance when one is dealing with very high dimensional datasets. Random Walk Metropolis Hastings (RWMH) algorithm, proposed by Metropolis *et al* [MRR53], is the standard procedure for dealing with such multidimensional datasets. The convergence and optimal scaling of such algorithms have been extensively studied [RGG97]. However, there are certain glaring problems that one may encounter while using RWMH. For very high dimensional datasets, convergence to the target density is pretty slow and one requires too many iterations, a difficulty further aggravated by the fact that in RWMH, we need to update each co-ordinate at a time and this may lead to very small acceptance probability if the dimension of the dataset

is too high. The TMCMC algorithm proposed in [DB11] uses simple deterministic transformations using a single random variable and a single proposal density chosen appropriately. Primarily one version, termed as the additive TMCMC method, has been researched, applied to some real life datasets with satisfactory performance and its efficiency over the usual RWMH in terms of better acceptance rate has also been established.

1.1 Basic Concepts and Algorithm

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

$$\begin{aligned}
(x_1, x_2) &\to (x_1 + \epsilon, x_2 + \epsilon) \\
(x_1, x_2) &\to (x_1 + \epsilon, x_2 + \epsilon) \\
(x_1, x_2) &\to (x_1 + \epsilon, x_2 + \epsilon) \\
(x_1, x_2) &\to (x_1 + \epsilon, x_2 + \epsilon)
\end{aligned} \tag{1.1.1}$$

This means we are moving along two lines in each transition from the point (x_1, x_2) , one parallel to the line $y = x$ and the other parallel to the direction $y = -x$. Each of the four transitions described above are indexed as I_k for kth transition, where k may vary from 1 to 4 in the bivariate case, and in general from 1 to 2^d in $\mathbb{R}^d$. We choose a direction with probability $p(I_k)$ for the I_kth move. In this paper, we have mainly focused on the case with all $p(I_k)$ equal. As with the standard MCMC case,

we do attach some probabilities with accepting/rejecting the proposed move such that the reversibility condition is satisfied thereby guaranteeing convergence. Formally, the algorithm may be presented as follows.

Crucially, a single ϵ is used to update all the co-ordinates which would ensure a significant computational gain over RWMH or any other MCMC related methods. Indeed as we shall show in **Chapter 6**, for dimensions of the order of 100-120, we may need to simulate a huge number of variables from the proposal density in case of RWMH, and time complexity is immense, while TMCMC on the other hand is much faster and the chain will also move faster compared to that of RWMH due to higher acceptance rate. The singleton ϵ also shows that there is no mixture proposal density corresponding to TMCMC, implying that TMCMC cannot be derived from RWMH for any standard choice of proposal density.

It must be stated very clearly at the onset that the way we have defined the moves for the additive model is just one instance of defining a move, keeping in mind some related issues like irreducibility and reversibility as we shall soon discuss. But an experimenter can take complete freedom in defining other move types and they may go on to yield better results than our approach as well. However, to us, this definition is simple and has both visual and analytical interpretation and that is why we stick with it. However, among other possible choices of moves, one particular case - termed Random Dive Metropolis Hastings- has already been investigated to great detail by Dutta and Bhattacharya [Dut10]. It uses the four move types starting from (x_1, x_2) given by $(x_1\epsilon, x_2\epsilon)$, $(x_1\epsilon, \frac{x_2}{\epsilon})$, $(\frac{x_1}{\epsilon}, x_2\epsilon)$ and $(\frac{x_1}{\epsilon}, \frac{x_2}{\epsilon})$. Among other possible moves that may interest readers would be $(x_1 + \epsilon, x_2\epsilon)$, $(x_1 - \epsilon, x_2\epsilon)$, $(x_1 + \epsilon, \frac{x_2}{\epsilon})$ and $(x_1 - \epsilon, \frac{x_2}{\epsilon})$ which enforces additive move along one co-ordinate and multiplicative move along other. One may improve on these moves by making the moves symmetrical about the co-ordinates - with probability p, we choose from one of the above move types and with probability

$(1-p)$, we consider the reverse of these move types with respect to co-ordinates (taking additive step along second co-ordinate and multiplicative step along 1st co-ordinate). What we want to highlight here is that the main abstraction behind our methodology is using a single ϵ to update all the co-ordinates. Now, we state the main algorithm for additive TMCMC corresponding to 2 dimensions (**Algorithm 1.1.1**).

Algorithm 1.1.1 *Additive TMCMC on* R^2

- Input: Initial value x_0, and number of iterations N.
- For $t = 0, \ldots, N-1$
  1. Generate $\epsilon \sim q(.)$ and $u \sim U(0,1)$ independently so that
  2. If $0 < u < \frac{1}{4}$, then set

$$x^\star = (x_1 + \epsilon, x_2 + \epsilon) \quad and \quad \alpha(x^\star, \epsilon) = min\left\{1, \frac{\pi(x^\star)}{\pi(x)}\right\}$$

  3. If $\frac{1}{4} < u < \frac{1}{2}$, then set

$$x^\star = (x_1 + \epsilon, x_2 - \epsilon) \quad and \quad \alpha(x^\star, \epsilon) = min\left\{1, \frac{\pi(x^\star)}{\pi(x)}\right\}$$

  4. If $\frac{1}{2} < u < \frac{3}{4}$, then set

$$x^\star = (x_1 - \epsilon, x_2 + \epsilon) \quad and \quad \alpha(x^\star, \epsilon) = min\left\{1, \frac{\pi(x^\star)}{\pi(x)}\right\}$$

  5. If $\frac{3}{4} < u < 1$, then set

$$x^\star = (x_1 - \epsilon, x_2 - \epsilon) \quad and \quad \alpha(x^\star, \epsilon) = min\left\{1, \frac{\pi(x^\star)}{\pi(x)}\right\}$$

6. Set $x_{n+1} = \left\{ \begin{array}{llll} x^\star & with & prob. & \alpha(x^\star, \epsilon) \\ x & with & prob. & 1 - \alpha(x^\star, \epsilon) \end{array} \right\}$

- End for

For higher dimensions, say $\mathbb{R}^d$, we have 2^d many moves as each co-ordinate is increased or decreased by the single ϵ. So, in such a case, we split the interval $[0, 1]$ into 2^d many equal parts. We order the various move types and if the uniform random variable u falls in the kth part, then we apply the move type of order k ($k = 1(|)2^d$).

The algorithm stated here is a much simplified one than the ones (**Algorithm 2.1** and **Algorithm 2.2**) stated in Dutta and Bhattacharya [Dut10], which presents the general mechanism irrespective of move types. For instance, we take the move probability for each of these moves to be same and our concentration lies on additive moves only for which the Jacobian is 1 and the acceptance rate has a nicer looking expression.

That the reversibility condition holds for this algorithm has been established in [DB11]. Since we are working on a general state space, so the general notions of irreducibility and aperiodicity of Markov chains do not hold in these spaces. This forces one to introduce analogous concepts for general state spaces (see Meyn and Tweedie [MT93]). We shall briefly describe these concepts as they will form the groundwork for much of our studies related to TMCMC.

1.2 Irreducibility criteria

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

measure ψ such that

$$\psi(A) > 0 \implies \exists n \quad with \quad P^n(x, A) > 0 \qquad \forall x \in \chi \tag{1.2.1}$$

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

$$P^n(x, .) > \delta\nu(.) \qquad x \in E \tag{1.2.2}$$

A chain is called *aperiodic* if the gcd of all such n for **Eqn 1.2.2** holds is 1. All these concepts of λ-irreducibility, aperiodicity and small sets are very important for laying the basic foundations of stability. The following theorem due to Dutta and Bhattacharya [DB11] establishes these properties for the additive TMCMC chain

Theorem 1 *Let π be the target density which is bounded away from 0 on $\mathbb{R}^d$. Also, let the proposal density q be positive on all compact sets on $\mathbb{R}^+$. Then, the every non-empty bounded set in $\mathbb{R}^d$ is small, and this can be used to show that the chain is both λ-irreducible and also aperiodic.*

Proof 1 *We prove this result for d=2. For higher values of d, the proof is analogous.*

We shall show that for any bounded non-empty set E,

$$P^2(x, A) > \delta\nu(A) \qquad x \in E \quad A \in \mathbb{B}(\mathbb{R}^2) \tag{1.2.3}$$

The above result implies that the additive TMCMC method is 2-step *small. It has been shown already that the usual RWMH is* 1-step *small [RT96]. Since E is bounded, it can be enclosed by a compact rectangle C with $\lambda(C) > 0$, whose sides are parallel to the $\{y = x\}$ and $\{y = -x\}$ directions. Note that in one step starting from $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, one can move to only a specific set of points, namely*

$$\begin{pmatrix} x_1 \pm \epsilon \\ x_2 \pm \epsilon \end{pmatrix}$$

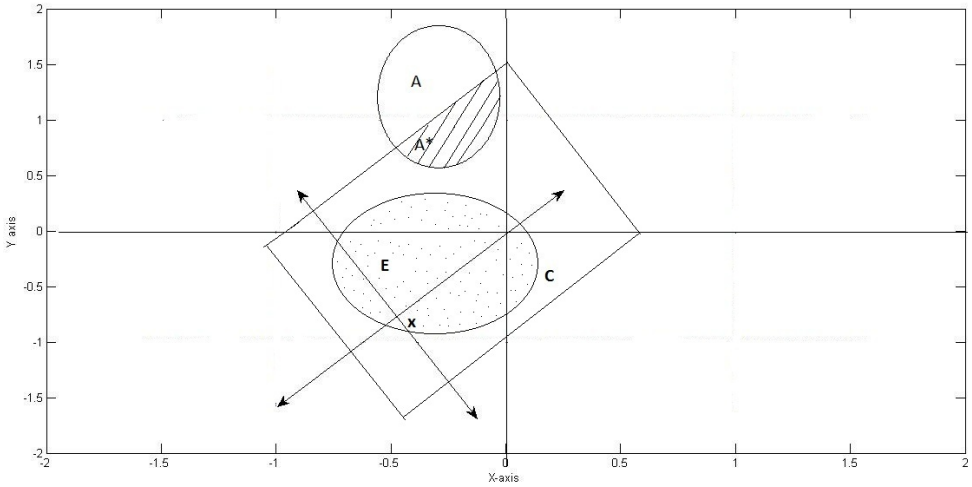


Fig. 1.2.1: The basic set up for checking irreducibility has been presented. E is the bounded set under consideration for "small" ness property. x is a particular point in E. We enclose E by a large compact rectangle C with sides parallel to the lines $\{y = x\}$ and $\{y = -x\}$. Let A be any arbitrary Borel set. $A^\star$ is the intersection of A with C. The possible directions of transitions from x have been depicted by bold arrows. It is clear that no transition from x can enter $A^\star$. This shows that TMCMC cannot be *1-step* small.

It can be easily seen from ***Fig 6.4.1****, that it may very well happen that in the first step, for no choice of ϵ does the particle reach the set A i.e. $P(x, A) = 0$, and if $\nu(A) > 0$, then by no means can the* 1-step *small property hold. So, we must resort to checking*

for 2-step *smallness. In two steps, the points that may be traversed from* $\boldsymbol{x}$ *are of the for*

$$\begin{pmatrix} x_1 \pm \epsilon_1 \pm \epsilon_2 \\ x_2 \pm \epsilon_1 \pm \epsilon_2 \end{pmatrix}$$

Check that in 2 two steps, varying over ϵ*, any state can be reached from any starting position. Altogether for a fixed* $\epsilon = (\epsilon_1, \epsilon_2)$*, there are* 16 *moves and we index them as* $I_1, I_2, \ldots, I_{16}$*, where* $x_{I_k}(\epsilon)$ *denotes the movement from* $\boldsymbol{x}$ *by the path indexed by* I_k *for some* ϵ*. Let* $x^1_{I_k}(\epsilon)$ *be the the intermediate step in the path attained after the first move. We define*

$$A^{I_k} = \{\epsilon : x_{I_k}(\epsilon) \in A^\star \quad and \quad x^1_{I_k}(\epsilon) \in A^\star\} \tag{1.2.4}$$

Let $p(I_k)$ *be the move probability for the* I_k *path and let* p_{min} *and* p_{max} *be the minimum and the maximum move probabilities. Also define*

$$m = \inf_{z \in C} \ \pi(z) \quad M = \sup_{z \in C} \ \pi(z) \quad h = \inf_{\delta \in C} \ q(\delta) \tag{1.2.5}$$

The above three quantities are finite and not equal to 0 because of the assumptions stated in the Theorem. Also, given a starting point x*, if we consider any point* y *in* $a^\star$*, there is a unique choice of* $\epsilon = (\epsilon_1, \epsilon_2)$ *for which we can move from* x *to* y*. Using this notion and the fact that* λ *is invariant of change of location or the axes, we have for bivariate data*

$$\sum_{k=1}^{16} \lambda(A^{I_k}) = \sum_{k=1}^{16} \lambda(x + A^{I_k}) = \lambda(A^\star) \tag{1.2.6}$$

Check that the fact that the sides of the compact rectangle C *are parallel to* $\{y = x\}$ *and* $\{y = -x\}$*, in the transition from the starting point* x *to the end point* y *belonging to* $A^\star$*, the intermediate step must also belong to* $A^\star$*. It can be easily checked that*

$$
\begin{aligned}
P^2(x,A) &\geq P^2(x,A^\star) \\
&\geq p_{min}^2 \sum_{k=1}^{16} \int_{A^{I_k}} q(\epsilon_1)q(\epsilon_2) \left(min \left\{ \frac{p_{min}\pi(x^1_{I_k}(\epsilon))}{p_{max}\pi(x)}, 1 \right\} \right) . \\
&\qquad \left(min \left\{ \frac{p_{min}\pi(x_{I_k}(\epsilon))}{p_{max}\pi(x^1_{I_k}(\epsilon))}, 1 \right\} \right) d\epsilon_1 d\epsilon_2 \\
&\geq p_{min}^2 h^2 (min \left\{ \frac{p_{min}.m}{p_{max}M}, 1 \right\})^2 \sum_{k=1}^{16} \lambda(A^{I_k}) \qquad \textit{by } \textbf{\textit{Eqn 1.2.5}} \\
&= p_{min}^2 h^2 (min \left\{ \frac{p_{min}.m}{p_{max}M}, 1 \right\})^2 \lambda(A^\star) \qquad \textit{by } \textbf{\textit{Eqn 1.2.6}} \\
&= p_{min}^2 h^2 (min \left\{ \frac{p_{min}.m}{p_{max}M}, 1 \right\})^2 \lambda_C(A) \qquad (1.2.7) \\
& \qquad (1.2.8)
\end{aligned}
$$

This explains analytically that 2-step *smallness is indeed satisfied. In this case* λ_C *plays the role of the measure* ν *and* $\delta = p_{min}^2 h^2 (min \left\{ \frac{p_{min}.m}{p_{max}M}, 1 \right\})^2$*, which is essentially a positive quantity. This automatically implies that the chain is* λ_C *measurable. But note that given any set* A *with* $\lambda(A) > 0$*, we can choose* C *large enough so that* $\lambda_C(A) > 0$*. This implies that the chain is* λ*-irreducible. Aperiodicity follows from the fact that by similar mechanism one can show that for any* k *greater than or equal to 2, the chain is* k-step *small. Hence gcd of all such* n *for which* ***Eqn 1.2.2*** *holds must be 1.*

1.3 Concluding remarks

In this introductory section, we described the TMCMC algorithm and discussed the various interesting properties of this algorithm- the reversibility, λ- irreducibility and aperiodicity of our chain. For any other definition of move types, one must ensure these conditions are satisfied as they form they are necessary to ensure convergence of

the chain. It would be our great pleasure if readers can come up with better move types that satisfy all the desirable criteria and perform better than our algorithm and the standard MCMC algorithms. Also, an interesting revelation was the fact that the minorization condition in **Eqn 1.2.7** was satisfied for P^2 the two step kernel or for any other higher order kernel instead of P, the one step kernel, as in standard RWMH algorithm. However, given that λ-irreducibility and aperiodicity have been proved for the TMCMC chain and that the reversibility condition holds imply that our chain indeed converges to the stationary distribution π.

Now that convergence has been guaranteed, our next focal point will be the rate of convergence. The most desirable rate would be the geometric or the exponential rate that would ensure very fast convergence. There is a rich theory on geometric ergodicity of the RWMH chains for a wide range of distributions (see Roberts and Tweedie [RT96], Jarner and Hansen [JH00], Mengersen and Tweedie [MT96]). Even a weaker form of ergodicity, polynomial ergodicity of RWMH chains has been extensively studied (see Jarner and Roberts [JR02], Mengersen and Tweedie [MT96]). We shall concentrate on establishing geometric ergodicity properties for our TMCMC chain and through simulation experiments, try to compare the rates of decay with that of the RWMH chain.

Chapter 2

Geometric ergodicity of TMCMC chain

2.1 Motivation

The reason for preferring geometric ergodicity is that under this condition, one can apply Central Limit Theorem to a wide class of functions of the Markov Chain, and hence, one can also speak about the stability of these ergodic estimates. We first define what is meant by geometric ergodicity.

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

$$||P^n(x,.) - \pi(.)||_{TV} \leq M.V(x)\rho^n \quad \forall n \geq 1 \tag{2.1.1}$$

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

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

A very standard way of checking geometric ergodicity is a result that involves the Foster-Lyapunov drift criteria. P is said to have a geometric drift to a set E if there is a function

$V \geq 1$, finite for at least one point and constants $\lambda < 1$ and $b < \infty$ so that

$$PV(x) \leq \lambda.V(x) + b1_E(x) \tag{2.1.2}$$

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

The first result we present is basically adaptation of a result due to (Mengerson and Tweedie, 1996). We now show a sufficient condition that would ensure that **Eqn 2.1.2** holds.

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

$$limsup_{|x|\to\infty} \frac{PV(x)}{V(x)} < 1 \tag{2.1.3}$$

$$\frac{PV(x)}{V(x)} < \infty \qquad \forall x \tag{2.1.4}$$

*Then this V satisfies the geometric drift condition in **Eqn 2.1.2** and hence the chain must be geometric ergodic. Also, if for some V finite , the geometric drift condition is satisfied, then the above condition must also hold true.*

Proof 2 *Assume that for some V finite and $V \geq 1$, the geometric drift condition in **Eqn 2.1.2** is satisfied. Now, by dividing both sides by $V(x)$, we shall get*

$$\frac{PV(x)}{V(x)} \leq \lambda + b\frac{1_E(x)}{V(x)}$$

Since V is finite then given that $V > 1$, we have

$$\frac{PV(x)}{V(x)} \leq \lambda + b \quad < \infty$$

Also if $|x| \to \infty$ then as E is a bounded small set, then $1_E(x) \to 0$, and hence

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

For the converse, let us fix a value $\gamma < 1$. Let R be a particular large value so that if $|x| > R$, then

$$\frac{PV(x)}{V(x)} < \gamma \quad |x| > R \quad \implies PV(x)\gamma.V(x) \quad |x| > R$$

Also $E = \{x : |x| \leq R\}$ is a compact set and by ***Eqn 2.1.4****,*

$$PV(x) \leq \frac{PV(x)}{V(x)} V(x)$$

Note here that $\frac{PV(x)}{V(x)}$ is finite by the hypothesis and the function V is also finite on any bounded set and E is one such set. This implies that $PV(x)$ is finite.

Take b to be the maximum value (finite) that $PV(x)$ can attain on the set E. and we know via the proof of irreducibility that such a E is small, then it is easy to see that $\forall x$

$$PV(x) \leq \gamma.V(x) + b1_E(x)$$

This proves the lemma.

So, in order to check geometric ergodicity, it is enough to prove **Eqn 2.1.3** and **Eqn 2.1.4** for the given chain. But unfortunately, it is not easy to establish for most distributions.

So, we need some additional assumptions. We start by showing that if the ϵ be sufficiently small so that ϵ^k is negligible for all $k > 2$, then we can derive simpler sufficient conditions for geometric ergodicity.

2.2 Sufficient conditions for geometric ergodicity

We restrict ourselves to $d = 2$, since the results will analogously follow for higher dimensions as well. We first state a lemma.

We index the transitions as

$$\begin{aligned} I_1 &\rightsquigarrow (x_1, x_2) \to (x_1 + \epsilon, x_2 + \epsilon) \\ I_2 &\rightsquigarrow (x_1, x_2) \to (x_1 + \epsilon, x_2 - \epsilon) \\ I_3 &\rightsquigarrow (x_1, x_2) \to (x_1 - \epsilon, x_2 + \epsilon) \\ I_4 &\rightsquigarrow (x_1, x_2) \to (x_1 - \epsilon, x_2 - \epsilon) \end{aligned} \tag{2.2.1}$$

For each k=1:4,

$$A^{I_k}(x) = \{\epsilon : \pi(x_{I_k}(\epsilon)) \geq \pi(x)\}$$

$$R^{I_k}(x) = \{\epsilon : \pi(x_{I_k}(\epsilon)) < \pi(x)\}$$

Here, corresponding to each transition I_k, $A^{I_k}(x)$ and $R^{I_k}(x)$ denote the acceptance region and the potential rejection rejection respectively for that transition at the point

x. We choose V function of the form

$$V(x) = \frac{M}{\pi(x)^s} \qquad 0 < s < 1 \tag{2.2.2}$$

In this case, we have the freedom to choose M and c. Note that as $|x| \to \infty$, it is expected that $V(x)$ goes to ∞, as $\pi(x)$ would decrease to 0. So, there exists a set $\{x : |x| > R\}$ over which V is larger than 1. We choose our target density π such that it is bounded on any bounded support. This would imply that we can always choose M large enough so that $V > 1$ even on $\{x : |x| > R\}$. This satisfies that V is uniformly greater than 1. That it is finite would follow if we take a target density that does not drop to 0 on any bounded region. Then

$$\begin{aligned}
\frac{PV(x)}{V(x)} &= \frac{1}{4}\sum_{k=1}^{4}\Big[\int_{A^{I_k}} q(\epsilon).\frac{V(x_{I_k}(\epsilon))}{V(x)}d\epsilon + \int_{R^{I_k}} q(\epsilon)\Big\{1 - \frac{\pi(x_{I_k}(\epsilon))}{\pi(x)}\Big\}d\epsilon \\
&\qquad + \int_{R^{I_k}} q(\epsilon)\frac{\pi(x_{I_k}(\epsilon))}{\pi(x)}.\frac{V(x_{I_k}(\epsilon))}{V(x)}d\epsilon\Big] \\
&= \frac{1}{4}\sum_{k=1}^{4}\Big[\int_{A^{I_k}} q(\epsilon)\Big\{\frac{\pi(x)}{\pi(x_{I_k}(\epsilon))}\Big\}^s d\epsilon + \int_{R^{I_k}} q(\epsilon)\Big\{1 - \Big\{\frac{\pi(x_{I_k}(\epsilon))}{\pi(x)}\Big\}\Big\}d\epsilon \\
&\qquad + \int_{R^{I_k}} q(\epsilon)\Big\{\frac{\pi(x_{I_k}(\epsilon))}{\pi(x)}\Big\}^{1-s} d\epsilon\Big] \\
&= \frac{1}{4}\sum_{k=1}^{4}\Big[\int_{A^{I_k}} q(\epsilon).\Big\{\frac{\pi(x)}{\pi(x_{I_k}(\epsilon))}\Big\}^s d\epsilon \\
&\qquad + \int_{R^{I_k}} q(\epsilon)\Big\{1 - \Big\{\frac{\pi(x_{I_k}(\epsilon))}{\pi(x)}\Big\} + \Big\{\frac{\pi(x_{I_k}(\epsilon))}{\pi(x)}\Big\}^{1-s}\Big\}d\epsilon\Big]
\end{aligned}$$

We consider those distributions that have well-defined probability contours in $\mathbb{R}^2$. In such cases, since the distribution must decay at the tails, so if $|x| \to \infty$, then corre-

sponding to two transitions I_1 and I_4, we get with probability going to 1,

$$\begin{aligned}\pi(x_1+\epsilon, x_2+\epsilon) &< \pi(x_1, x_2)\\ \pi(x_1-\epsilon, x_2-\epsilon) &> \pi(x_1, x_2)\end{aligned} \tag{2.2.3}$$

Now we state a result from Dutta and Bhattacharya [Dut10].

Lemma 2 *For each $1 > \lambda > 0$ and for any fixed s lying strictly between 0 and 1,*

$$\lambda^s + \lambda^{1-s} - \lambda < 1 \tag{2.2.4}$$

Also, we make the following assumptions

$$\begin{aligned}(A1)\ \underset{|x|\to\infty}{limsup} \int_\epsilon \left\{ \frac{1}{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}} \mathbb{I}_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}>1} \right\} &= \underset{|x|\to\infty}{limsup} \int_\epsilon \left\{ \frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} \mathbb{I}_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} \right\}\\ (A2)\ \underset{|x|\to\infty}{limsup} \int_\epsilon \left\{ \frac{1}{\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}} \mathbb{I}_{\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}>1} \right\} &= \underset{|x|\to\infty}{limsup} \int_\epsilon \left\{ \frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)} \mathbb{I}_{\frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}<1} \right\}\end{aligned} \tag{2.2.5}$$

Now armed with these assumptions and results,

$$
\begin{aligned}
4 \underset{|x|\to\infty}{limsup}\ & \frac{PV(x)}{V(x)} \\
= \quad & \underset{|x|\to\infty}{limsup} \Bigg[\int_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} \left\{ 1 - \frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} + \left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} \right]^{1-s} \right\} q(\epsilon) d\epsilon \\
& + \int_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}<1} \left\{ 1 - \frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)} + \left[\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)} \right]^{1-s} \right\} q(\epsilon) d\epsilon \\
& + \int_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}>1} \left\{ \frac{1}{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}} \right\}^s q(\epsilon) d\epsilon + \int_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}>1} \left\{ \frac{1}{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}} \right\}^s q(\epsilon) d\epsilon \\
& + \int_{\frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}<1} \left\{ 1 - \frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)} + \left[\frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)} \right]^{1-s} \right\} q(\epsilon) d\epsilon \\
& + \int_{\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} \left\{ 1 - \frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} + \left[\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} \right]^{1-s} \right\} q(\epsilon) d\epsilon \\
& + \int_{\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}>1} \left\{ \frac{1}{\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}} \right\}^s q(\epsilon) d\epsilon \\
& \quad + \int_{\frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}>1} \left\{ \frac{1}{\frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}} \right\}^s q(\epsilon) d\epsilon \Bigg]
\end{aligned}
\tag{2.2.6}
$$

Define i.i.d variables b_1 and b_2 such that b_i takes value $+1$ and -1 with probability $\frac{1}{2}$.

We define I_{b_1,b_2} as follows

$$
\begin{aligned}
I_{b_1,b_2} \quad = \quad & \underset{|x|\to\infty}{limsup} \left[\int_{\frac{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon)}{\pi(x_1,x_2)}<1} \left\{ 1 - \frac{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon)}{\pi(x_1,x_2)} + \left[\frac{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon)}{\pi(x_1,x_2)} \right]^{1-s} \right\} q(\epsilon) d\epsilon \right. \\
& + \quad \underset{|x|\to\infty}{limsup} \int \left\{ \frac{1}{\frac{\pi(x_1-b_1\epsilon,x_2-b_2\epsilon)}{\pi(x_1,x_2)}} \right\}^s I_{\frac{\pi(x_1-b_1\epsilon,x_2-b_2\epsilon)}{\pi(x_1,x_2)}>1} q(\epsilon) d\epsilon
\end{aligned}
\tag{2.2.7}
$$

It is easy to see that

$$4 \underset{|x|\to\infty}{limsup} \frac{PV(x)}{V(x)} \leq I_{+1,+1} + I_{+1,-1} + I_{-1,+1} + I_{-1,-1} \tag{2.2.8}$$

We abbreviate $I_{+1,+1}$ as I_1 .

$$\begin{aligned}
I_1 \quad &= \quad \underset{|x|\to\infty}{limsup} \Bigg[\int_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} \left\{ 1 - \frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} + \left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} \right]^{1-s} \right\} q(\epsilon) d\epsilon \\
&\qquad\qquad + \int_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}>1} \left\{ \frac{1}{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}} \right\}^s q(\epsilon) d\epsilon \Bigg] \\
&= \quad \underset{|x|\to\infty}{limsup} \int \Bigg[\left\{ 1 - \frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} + \left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} \right]^{1-s} \right\} I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} \\
&\qquad\qquad + \quad \left\{ \frac{1}{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}} \right\}^s I_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}>1} \Bigg] q(\epsilon) d\epsilon \\
&\leq \quad \underset{|x|\to\infty}{limsup} \int \left[\left\{ 1 - \frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} + \left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} \right]^{1-s} \right\} I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} \right] \\
&\qquad\qquad + \quad \underset{|x|\to\infty}{limsup} \int \left\{ \frac{1}{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}} \right\}^s I_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}>1} q(\epsilon) d\epsilon
\end{aligned} \tag{2.2.9}$$

Note that the integrand in this case is bounded as

$$
\begin{aligned}
&\left|\left\{1-\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}+\left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}\right]^{1-s}\right\}I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1}\right| \\
&\qquad\qquad +\left|\left\{\frac{1}{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}}\right\}^{s}I_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}>1}\right| \\
&\leq\left\{1+\left|\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}\right|+\left|\left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}\right]^{1-s}\right|\right\}I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} \\
&\qquad\qquad +\left|\left\{\frac{1}{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}}\right\}^{s}I_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}>1}\right| \\
&\leq 4
\end{aligned}
$$

Using Assumption $(A1)$, it can be checked that

$$
\begin{aligned}
I_1 \quad &\leq \quad \underset{|x|\to\infty}{limsup}\int\left[\left\{1-\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}+\left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}\right]^{1-s}\right\}I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1}\right] \\
&\qquad + \quad \underset{|x|\to\infty}{limsup}\int\left\{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}\right\}^{s}I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1}q(\epsilon)d\epsilon \\
&= \quad \underset{|x|\to\infty}{limsup}\int\left[\left\{1-\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}+\left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}\right]^{1-s}\right.\right. \\
&\qquad \left.\left.+\left[\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}\right]^{s}\right\}I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1}\right] \\
&< \quad 2\ \underset{|x|\to\infty}{limsup}\int I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} \qquad\qquad \textit{by Lemma 4.2}
\end{aligned}
\tag{2.2.10}
$$

Using similar approach the assumption $(A2)$, we have

$$
\begin{aligned}
4\,\underset{|x|\to\infty}{limsup}\,\frac{PV(x)}{V(x)} &< 2\left[\underset{|x|\to\infty}{limsup}\int I_{\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} + \underset{|x|\to\infty}{limsup}\int I_{\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}<1}\right] \\
&< \int\left\{2\,I_{\underset{|x|\to\infty}{limsup}\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} + 2\,I_{\underset{|x|\to\infty}{limsup}\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}<1}\right. \\
&\quad \left. + \quad 2\,I_{\underset{|x|\to\infty}{limsup}\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1} + 2\,I_{\underset{|x|\to\infty}{limsup}\frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}<1}\right\} \\
&= 2 + 2\int\left\{I_{\underset{|x|\to\infty}{limsup}\frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)}<1} + I_{\underset{|x|\to\infty}{limsup}\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)}<1}\right\}
\end{aligned}
\tag{2.2.11}
$$

Note that for two sequences a_n and b_n,

$$
limsup\,(a_n) - limsup\,(b_n) \leq limsup\,(a_n - b_n)
$$

Using this, one can infer from $(A1)$ to $(A3)$ that

$$
\{\epsilon : \underset{|x|\to\infty}{limsup}\frac{\pi(x_1+\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} < 1\} = \{\epsilon : \underset{|x|\to\infty}{limsup}\frac{\pi(x_1-\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)} < 1\}^c
$$

$$
\{\epsilon : \underset{|x|\to\infty}{limsup}\frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} < 1\} = \{\epsilon : \underset{|x|\to\infty}{limsup}\frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)} < 1\}^c
$$

And then, we can say that

$$
4\underset{|x|\to\infty}{limsup}\,\frac{PV(x)}{V(x)} < \quad 2+2 \quad = 4 \tag{2.2.12}
$$

This establishes the first sufficient condition stated in Lemma 4.1. The condition that $\frac{PV(x)}{V(x)}$ is finite follows from the fact that our choice of V is finite for any $0 < s < 1$ given that that the density function π is bounded.

An assumption made here is that $\left\{\epsilon : \frac{\pi(x_1-\epsilon,x_2+\epsilon)}{\pi(x_1,x_2)} = 1\right\}$ and $\left\{\epsilon : \frac{\pi(x_1+\epsilon,x_2-\epsilon)}{\pi(x_1,x_2)} = 1\right\}$ are measure zero sets. But, check that this assumption can be relaxed and the same procedure can be carried out even if these sets are not of 0 measure.

Example 1 *Consider now the bivariate uniform distribution*

$$\pi(x,y) \propto \frac{1}{\lambda(C)} \qquad x,y \in C \quad C \text{ bounded in } \mathbb{R}^2 \tag{2.2.13}$$

Note that for this distribution, outside the compact set C $\pi = 0$ *and in such cases, if transitions occur outside* C,(A1)-(A3) *will not make much sense. But luckily, our algorithm and also standard MCMC are so designed that such transitions are always rejected and one would be restricted to* C *all the time. Under such a case* (A1)-(A3) *trivially hold since all the* π *values are identical. Also this distribution is uniformly ergodic in the sense that the rate of convergence in* (7) *is independent of* x.

However, the sufficient conditions derived above fail for the case of Gaussian distributions, which propels us to search for stronger sufficient conditions that would include Gaussian distributions, if at all possible. We cite the next theorem that does suffice the need.

We shall now give a sufficient condition for geometric ergodicity in TMCMC for a broad class of distributions. This proof follows on the lines of Jarner and Hansen [JH00] and has been suitably modified for our TMCMC case. First we define the notion of subexponential densities.

A density π is said to be sub-exponential if it is positive with continuous first derivative

and satisfies

$$\lim_{|x|\to\infty} n(x).\nabla log\pi(x) = -\infty \tag{2.2.14}$$

where $n(x)$ denotes the unit vector $\frac{x}{|x|}$. This would imply that for any $K > 0$, $\exists\ R > 0$ such that

$$\frac{\pi(x + cn(x))}{\pi(x)} \leq e^{-cK} \qquad\qquad x \geq R, c \geq 0 \tag{2.2.15}$$

This means that π is decaying at a rate better than exponential along any direction. It is very easy to check that the Gaussian (univariate as well as multivariate for any variance covariance matrix) or the Gamma distributions (univariate or independent multivariate) indeed satisfy these conditions.

Theorem 2 *If π the target density is sub-exponential and has contours that are nowhere piecewise parallel to $y = -x$, then the TMCMC chain satisfies geometric drift if*

$$\liminf_{|x|\to\infty} Q(x, A(x)) > 0 \tag{2.2.16}$$

For this particular scenario, one can check that the function V that satisfies the geometric drift condition is $V(x) = \frac{c}{\sqrt{\pi(x)}}$ corresponding to some $c > 0$.

Proof 3 *Following the notation of [JH00], let $C_{\pi(x)}$ be the contour of the density π that contains the value $\pi(x)$. We define the radial cone $C_{\pi(x)}(\delta)$ around $C_{\pi(x)}$ to be*

$$C_{\pi(x)}(\delta) = \left\{y + an(y) : y \in C_{\pi(x)}, -\delta < s < \delta\right\} \tag{2.2.17}$$

By the hypothesis we can say there exists a $\epsilon > 0$ such that

$$\underset{|x|\to\infty}{limsup}\ Q(x, R(x)) \leq 1 - 2\epsilon^{\frac{1}{2}} \tag{2.2.18}$$

Take the belt length δ such that the probability that a move from x, the starting point, falls within this δ belt is 0. That it is possible can be seen as follows. Note that $\exists$ a compact set E such that $Q(x, E^c) < \frac{\epsilon}{2}$. So, if we can get a δ so that $Q(C_{\pi(x)}(\delta) \cap E) < \frac{\epsilon}{2}$, then we are done. Note that of the two outward and inward moves given by $(+\epsilon, +\epsilon)$ and $(-\epsilon, -\epsilon)$ for any point in the 1st co-ordinate, the probability that such a move results in a value within $C_{\pi(x)}(\delta)$ is proportional to the width δ and thus can be made sufficiently small. For the other two moves, note that since the contours ($\cap$ E) are nowhere piecewise parallel to $y = -x$, then we cut the contour at finitely many points. Infinitely many points of intersection can be ruled out because of the intersection with E which is compact. If that was the case, this this infinite collection of interesting points would have a limit point in E, which is not possible as the points are isolated.

Now, in this situation, $\exists$ R_ϵ so that any point y outside the δ bound around x,

$$\frac{\pi(y)}{\pi(x)} < \epsilon \qquad\qquad |x| > R_\epsilon \tag{2.2.19}$$

This done by taking the shortest line from y to the origin and suppose it (extended if needed) cuts the contour $C_{\pi(x)}$ at z. There will be two such values of z, we choose the one that is nearest to x. Then by the fact that $\pi(x)$ is same as $\pi(z)$ and ***Eqn 2.2.15****, we get the result in* ***Eqn 2.2.19****. This still requires a small argument left. To ensure that this z indeed satisfies $|z| > R_\epsilon$. Now look at he set E, the effective move set. Join each point in E to the origin by a straight line and extend it if needed to cut the contour. We get a segment of contour $D(x)$ bounded and closed that contains x. Now since this*

set is a bounded one, we can always choose x large enough so that all these points have norm greater than R_ϵ. Armed with these results, we look for dimension d and $s = \frac{1}{2}$,

$$\begin{aligned}\frac{PV(x)}{V(x)} &= \frac{1}{2^d}\sum_{b_1,\cdots,b_d}\int_{A(x)}\left[\frac{\pi(x_1,x_2)}{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon)}\right]^{\frac{1}{2}} g(\epsilon)d\epsilon \\ &\quad + \frac{1}{2^d}\sum_{b_1,\cdots,b_d}\int_{R(x)}\left[1-\frac{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon)}{\pi(x_1,x_2)}+\left\{\frac{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon)}{\pi(x_1,x_2)}\right\}^{\frac{1}{2}}\right] g(\epsilon)d\epsilon\end{aligned} \tag{2.2.20}$$

Split the integral over $R(x)$ and that over $A(x)$ into parts- within $C_{\pi(x)}(\delta)$ and outside $C_{\pi(x)}(\delta)$, and then we get for x such that $|x| > R_\epsilon$ and all the bounded region $D(x)$ has all points with norm greater than R_ϵ,

$$\begin{aligned}\frac{PV(x)}{V(x)} &< \epsilon + \epsilon^{\frac{1}{2}}Q(x,A(x)) + \left(1+\epsilon^{\frac{1}{2}}\right)Q(x,R(x)) \\ &= \epsilon + \epsilon^{\frac{1}{2}} + Q(x,R(x)) \\ &= 1-\epsilon^{\frac{1}{2}}+\epsilon \qquad\qquad < 1\end{aligned} \tag{2.2.21}$$

Note that for spherically symmetric sub-exponential distributions (for example standard Gaussian), these conditions naturally hold. For instance, the fact that no part of the contour is parallel to $y = -x$ is quite obvious. To check that $\underset{|x|\to\infty}{liminf}\ Q(x, A(x)) > 0$, it is enough to perceive that at any point in the 1st co-ordinate, the inward direction first stays in the acceptance region and then moves to rejection region after some time. Now, perceive that the minimum distance to be traversed to reach the acceptance region from any point in the first co-ordinate through the inward move is proportional to the norm

value of the point. Now, let M_ϵ be the that value such that the $\int_{-M}^{M} g(\epsilon)d\epsilon > 1-\epsilon$. Now choose x such that $|x| > 3M_\epsilon$ (radius of $C_{\pi(x)}$ is greater than $3M_\epsilon$), then $Q(x, A(x)) > \frac{1-\epsilon}{4} > 0$. Now take the truncated sequence x_n with $|x_n| \to 0$ and x_n has radius greater than $3M_\epsilon$, then along this sequence, the limit of $Q(x, A(x))$ is greater than $\frac{1-\epsilon}{4}$. Thus $\underset{|x|\to\infty}{liminf}\ Q(x, A(x)) > 0$ condition is satisfied.

Note that the constraint that no part of the contour can be piecewise parallel to $y = -x$ does not really cause too much of a problem because the only common distribution that satisfies this property is the Laplace distribution and it is not sub-exponential. So, in a sense we are not losing much over the RWMH algorithm.

Since the minorization condition is satisfied for P^2, from all these derivations, we can at most say that P^2 is geometrically ergodic. However, since the total variation norm is a decreasing function of n, we can say

$$||P^n(x,.) - \pi(.)||_{TV} < \frac{M}{\sqrt{\pi(x)}}\rho^{2\lfloor \frac{n}{2} \rfloor} \tag{2.2.22}$$

which is almost as good as geometric ergodicity.

2.3 Concluding remarks

In this chapter, we made some basic progress as far as convergence of TMCMC is concerned. We derived sufficient conditions for the geometric drift condition and although not exact, the convergence rate for the TMCMC chain in such cases is almost as good as geometric ergodicity. However, geometric ergodicity is just a theoretical notion that has certain desirable properties. It does not tell us how fast or slow the convergence rate really is. Note that the rate depends heavily on π, the target density and ρ. What we would like to do is to find out how the parameter ρ varies, what are the guiding

factor and what could be the optimal value of ρ. We shall start analyzing these issues next chapter onwards.

for different choices of proposal variances). Corresponding to each time point t, we shall thus get L many iterates. The notion is that as time t increases (specially after burn-in), these L many iterates should be close to an independently drawn random sample from the target distribution π. So, if we look at the KST statistic for the empirical distribution of these iterates along a particular dimension with respect to the marginal of π along that dimension, we should find the test statistic decreasing with time and finally being very close to 0 after a certain time point. burn in.

Chapter 3

Optimal scaling of TMCMC algorithm (independent components)

Having established the geometric ergodicity properties of the TMCMC method, we now focus on the optimal scaling of the proposal distribution so as to ensure convergence to stationarity in optimal time. This is particularly essential from an algorithmic point of view in order to reduce the time complexity of the simulation algorithm. If the variance of the proposal density is very small, then jumps will be of smaller magnitude and this would mean the Markov chain would take lot more time to traverse the entire space and in the process, the convergence rate would be pretty low (see **Fig 6.4.2**). On the other hand, if the variance is very large, then our algorithm will reject too many of the moves. An instance of that is presented in **Fig 6.4.1** where we depict the sample paths of two processes in RWMH and TMCMC processes, one with dimension 5 and proposal distribution is normal (independently so along each co-ordinate for RWMH) and variance 10. In another case, we interchanged the dimension and the variance. In both the cases we found there were lots of rejections and the sample path remained static for quite some time (especially for RWMH). Also, it can be deduced that there is an interplay between the proposal variance and the dimension, increase in dimension and increase in variance both contribute to lower acceptance rates. Therefore, it is most intuitive that the optimal scaling for both the RWMH and TMCMC chains will depend on the dimension of the underlying process as well.

There is an extensive theory on optimal scaling of RWMH chains (see Beskos, Roberts and Stuart [BRS09], Bedard [Bed09] [Bed07], Neal and Roberts [NR06], Roberts, Gelman and Gilks [RGG97]). The magic number for RWMH has been the optimal acceptance rate value of 0.234, which has been achieved through maximization of speed of the process for a wide range of distributions- i.i.d set up and some special class of independent but non-i.i.d. component set up as well. We have employed an analogous speed where we have optimized the diffusion speed of our process to get optimal acceptance rate. For the sake of readers' interests, I am splitting this topic into two separate chapters. In this chapter, we shall concentrate on optimal scaling of i.i.d families and independent families. In the next chapter, we shall proceed to the more involved general dependent case scenario.

Note that our moves at each step were symmetric and the magnitude of the jump would depend on the choice of the proposal density q. Again, q must have its support as $\mathbb{R}^+$, and from now onwards, we shall assume that it is a $N(0,1)$ distribution truncated at 0. Note that at each step , we sample only one ϵ from this proposal distribution and updates all the co-ordinates at one go. The notion of symmetrical transitions at each step can be expressed from a mathematical point of view as follows. Suppose that we are simulating from a d dimensional space (usually $\mathbb{R}^d$), and suppose we are at a point $x = (x_1, x_2, \ldots, x_d)$, and we use a single ϵ updating. Now, in order to analytically present the transitions, we define d random variables $b_1, b_2, \ldots, b_d$, such that

$$\begin{aligned} b_i &= +1 \ \textit{with probability} \ \frac{1}{2} \\ &= -1 \ \textit{with probability} \ \frac{1}{2} \end{aligned} \tag{3.0.1}$$

for each i, $i = 1, 2, \ldots, d$. So, we can write the transition from x given ϵ as

$$(x_1, x_2, \ldots, x_d) \to (x_1 + b_1\epsilon, x_2 + b_2\epsilon, \ldots, x_d + b_d\epsilon) \tag{3.0.2}$$

This obviously means that we have to define $d+1$ many random variables corresponding to d many random updates ϵ_1, ϵ_2, $\ldots$, ϵ_d as in the standard RWMH case, but simulating b_i is far easier an exercise (equivalent to a single toss of a fair coin) than simulating ϵ_i's, from a computational point of view. We shall show in this paper that from the viewpoint of optimal convergence also, our algorithm performs way better than the standard RWMH.

3.1 I.I.D. components case study

Assume that simplest case in which the target density π is a product of iid marginals.

$$\pi_d(x) = \prod_{i=1}^{d} f(x_i) \tag{3.1.1}$$

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

$$(C1) \quad E\left[\left\{\frac{f'(X)}{f(X)}\right\}^8\right] \;=\; M_1 \;< \infty \tag{3.1.2}$$

$$(C2) \quad E\left[\left\{\frac{f''(X)}{f(X)}\right\}^4\right] \;=\; M_2 \;< \infty \tag{3.1.3}$$

We shall show that for each fixed one-dimensional component of X, the one-dimensional process converges to a Diffusion process which is analytically tractable and its diffusion and drift speeds may be numerically evaluated.

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

The following theorem due to Roberts, Gelman and Gilks (1997) [RGG97] establishes the Diffusion process approximation of the standard RWMH process. In this case the proposal density variance has been chosen to be proportional to $\frac{1}{d}$. Before stating the theorem, first let us introduce the notion of Skorohod topology [Sko56]. It is a topology generated by a class of functions from $[0,1] \to \mathbb{R}$ for which Right hand limit and Left hand limit are well defined at each point (may not be same). It is an important tool for formulating Poisson process, Levy process and other stochastic point processes. We consider the metric separable topology $\mathbb{J}_1$ on the above class of functions as defined in [Sko56]. When we speak of convergence of discrete time stochastic processes to diffusion process in this paper, we imply convergence with respect to this $\mathbb{J}_1$ topology.

Theorem 3 *Let f be a positive Lipschitz continuous twice differentiable function that satisfies **Eqn 3.1.2** and **Eqn 3.1.3**, and let us define ${X_0}^n = (X^1_{0,1}, X^2_{0,2}, \ldots, ...X^n_{0,n})$ and $X^i_{0,j} = X^j_{0,j} \forall i \leq j$. Then as $n \to \infty$, U^n converges to U with respect to the Skorokhod topology and U satisfies the Langevin Stochastic Differential Equation.*

$$dU_t = (h(l))^{\frac{1}{2}} dB_t + h(l) \frac{f'(U_t)}{2f(U_t)} dt \tag{3.1.4}$$

where

$$h(l) = 2l^2 \Phi\left(\frac{-l\sqrt{I}}{2}\right) \tag{3.1.5}$$

where Φ *is the Gaussian cumulative distribution function and* I *is the information matrix of* f*.* $h(l)$ *is called the speed measure of the diffusion process. It can be shown that* $h(l)$ *is maximized for*

$$\hat{l} = \frac{2.38}{\sqrt{I}} \tag{3.1.6}$$

$$\alpha(\hat{l}) = 0.23 \quad h(\hat{l}) = \frac{1.3}{I} \tag{3.1.7}$$

We shall formulate an analogous diffusion process for the TMCMC approach and then either analytically or numerically try to compare the diffusion speeds and the acceptance rates of the two competing mechanisms.

First, we define the discrete time generator of the TMCMC approach, given by

$$\begin{aligned} G_d V(x) \quad = \quad & \frac{d}{2^d} \sum_{\left\{ \begin{array}{l} b_i \in \{-1,+1\} \\ \forall i = 1(|)d \end{array} \right\}} \int_0^\infty \left[\left(V\left(x_1 + b_1\epsilon, \ldots, x_d + b_d\epsilon\right) - V\left(x_1, \ldots, x_d\right) \right) \right. \\ & \left. \times \left(min\left\{ 1, \frac{\pi(x_1 + b_1\epsilon, x_2 + b_2\epsilon, \ldots, x_d + b_d\epsilon)}{\pi(x_1, x_2, \ldots, x_d)} \right\} \right) \right] q(\epsilon) d\epsilon \end{aligned} \tag{3.1.8}$$

Note that this function is measurable with respect to the Skorokhod topology and we can treat G_n as a continuous time generator that has jumps at the rate d. Given our restricted focus on a one dimensional component of the actual process, we assume V to be a function of the 1^{st} co-ordinate only. Under this assumption, the generator defined in **Eqn 3.1.8** is a function of only ϵ and b_1. and can be rephrased as

$$G_dV(x) = \frac{d}{2}\int_0^{\infty} \sum_{b_1 \in -1,+1} \left[\left(V(x_1 + b_1\epsilon) - V(x_1) \right) \times E_{b_2,b_3,\ldots,b_d} \left(min \left\{ 1, \frac{\pi(x_1 + b_1\epsilon, x_2 + b_2\epsilon, \ldots, x_d + b_d\epsilon)}{\pi(x_1, x_2, \ldots, x_d)} \right\} \right) \right] q(\epsilon) d\epsilon \tag{3.1.9}$$

where $E_{b_2,b_3,\ldots,b_d}$ is the expectation taken conditional on b_1 and ϵ. Note that if ϵ is taken to be truncated Normal $(0, 1)$, with left truncation at 0, then $b_i\epsilon$ for each $i = 1, 2, \ldots, d$ follows $N(0, 1)$ and since $E(b_i b_j \epsilon)$ for $i \neq j$ is also 0, we can say that $b_i\epsilon$ are uncorrelated. However, $b_i\epsilon$ are not independent because if they were, the pairwise sum would have been normal. However, since $b_i\epsilon + b_j\epsilon$ is equal to 0 with probability $\frac{1}{2}$ if $i \neq j$, the normality assumption for the sum is contradicted. Also, if we are working on a d dimensional space, we assume that $\epsilon \sim TN_0^L(0, \frac{l^2}{d})$ where we use the notation TN_0^L means truncated normal with left truncation at 0.

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

$$\begin{aligned} G_dV(x) &\leq dE\left[V(x + b\epsilon) - V(x)\right] \\ &= dV^{'}(x)E(b\epsilon) + d\frac{1}{2}V^{''}(x)E(\epsilon^2) \\ &\leq l^2\frac{d}{d-1}M \end{aligned} \tag{3.1.10}$$

where M is the maximum value of $V^{''}$, provided it is bounded. Most of the common choices of V are bounded and we can talk about its maximum value.

We state **Prop 1** that will prove handy.

Proposition 1 *If* $X \sim N(\mu, \sigma^2)$, *then*

$$E\left[min\left\{1, e^X\right\}\right] = \Phi\left(\frac{\mu}{\sigma}\right) + e^{\{\mu + \frac{\sigma^2}{2}\}}\Phi\left(-\sigma - \frac{\mu}{\sigma}\right) \tag{3.1.11}$$

where Φ *is the standard Gaussian cdf.*

Note that

$$\begin{aligned} E\Big|_{b_1\epsilon} & \left[min\left\{1, \frac{\pi(x_1 + b_1\epsilon, x_2 + b_2\epsilon, \ldots, x_d + b_d\epsilon)}{\pi(x_1, x_2, \ldots, x_d)}\right\}\right] \\ = \quad E\Big|_{b_1\epsilon} & \left[min\left\{1, exp\left(log(f(x_1 + b_1\epsilon)) - log(f(x_1))\right.\right.\right. \\ & \left.\left.\left. + \sum_{j=2}^{d}\left\{b_j\epsilon\left\{log(f(x_j))\right\}' + \frac{\epsilon^2}{2!}\left\{log(f(x_j))\right\}'' + \frac{b_j\epsilon^3}{3!}\left\{log(f(z_j))\right\}'''\right\}\right)\right\}\right] \end{aligned} \tag{3.1.12}$$

Since $b_j \forall j = 2(|)d$ are iid, as $d \to \infty$, one can apply Lyapunov's Central Limit Theorem [the Lyapunov condition holds for δ =2] , and we have that

$$\frac{\sum_{j=2}^{d} b_j\left[\epsilon\left\{log(f(x_j))\right\}' + \epsilon^3\left\{log(f(z_j))\right\}'''\right]}{\sqrt{\sum_{j=2}^{d}\left[\epsilon\{log(f(x_j))\}' + \epsilon^3\left\{log(f(z_j))\right\}'''\right]^2}} \xrightarrow{d} N(0,1) \tag{3.1.13}$$

Call

$$\eta(x_1, b_1, \epsilon) = log(f(x_1 + b_1\epsilon)) - log(f(x_1)) \tag{3.1.14}$$

Given this, we can say that as $d \to \infty$, **Eqn 3.1.12** reduces to $E(min\left\{1, e^X\right\})$ where

X is given as

$$X \sim N\left(\eta(x_1, b_1, \epsilon) + \frac{\epsilon^2}{2!}\sum_{j=2}^{d}\left\{log(f(x_j))\right\}'', \epsilon^2\sum_{j=2}^{d}\left[\left\{log(f(x_j))\right\}'\right]^2 + \Delta\right) \tag{3.1.15}$$

where

$$\Delta = \epsilon^4\sum_{j=2}^{d} 2\left\{log(f(x_j))\right\}'\left\{log(f(z_j))\right\}''' + \epsilon^6\sum_{j=2}^{d}\left[\left\{log(f(z_j))\right\}'''\right]^2 \tag{3.1.16}$$

It can be shown using Slutsky's theorem that this Δ term can be neglected as $d \to \infty$. For the sake of technicality, we have to justify that under the convergence in distribution as in **Eqn 3.1.13**, the expectation in **Eqn 3.1.12** also converges to $E(min\left\{1, e^X\right\})$. But this is obvious using the Skorohod representation theorem and using the fact that the expectation is taken over a function which is bounded and hence uniformly integrable.

Using the fact that

$$-\frac{1}{d-1}\sum_{j=2}^{d}\left\{log(f(x_j))\right\}'' = \frac{1}{d-1}\sum_{j=2}^{d}\left[\left\{log(f(x_j))\right\}'\right]^2 \overset{d\to\infty}{\longrightarrow} \mathbb{I} \tag{3.1.17}$$

where $\mathbb{I}$ is the information matrix corresponding to the density f. Also, by definition of ϵ, we may assume that $(d-1)\epsilon^2$ is a finite positive quantity. If that is the case, then the normal parameters in **Eqn 3.1.15** can be reformulated (using Slutsky's theorem) as

$$X \sim N\left(\eta(x_1, b_1, \epsilon) - \frac{(d-1)\epsilon^2}{2}\mathbb{I},\ (d-1)\epsilon^2\mathbb{I}\right) \tag{3.1.18}$$

Then using **Prop 1**, we can reduce the expression in **Eqn 3.1.12** to the following

$$
\begin{aligned}
E\Big|_{b_1\epsilon} & \left[min\left\{1, \frac{\pi(x_1+b_1\epsilon, x_2+b_2\epsilon, \ldots, x_d+b_d\epsilon)}{\pi(x_1, x_2, \ldots, x_d)}\right\}\right] \\
= & \quad \Phi\left(\frac{\eta(x_1, b_1, \epsilon) - \frac{(d-1)\epsilon^2}{2}\mathbb{I}}{\sqrt{(d-1)\epsilon^2\mathbb{I}}}\right) + e^{\eta(x_1,b_1,\epsilon)}\Phi\left(-\sqrt{(d-1)\epsilon^2\mathbb{I}} - \frac{\eta(x_1, b_1, \epsilon) - \frac{(d-1)\epsilon^2}{2}\mathbb{I}}{\sqrt{(d-1)\epsilon^2\mathbb{I}}}\right) \\
= & \quad \mathbb{W}(b_1\epsilon, x_1)
\end{aligned}
\tag{3.1.19}
$$

Note that using Taylor series expansion around x_1, we can write **Eqn 3.1.14** as

$$
\eta(x_1, b_1, \epsilon) = b_1\epsilon\left[logf(x_1)\right]' + \frac{\epsilon^2}{2}\left[logf(x_1)\right]'' + b_1\frac{\epsilon^3}{3!}\left[logf(\xi_1)\right]''' \tag{3.1.20}
$$

Also, we can rewrite $b_1\epsilon$ as $\frac{l}{\sqrt{(d-1)}}z_1$, where z_1 follows a $N(0,1)$ distribution and then we can write η in terms of l and z_1 as

$$
\eta(x_1, z_1) = b_1\frac{lz_1}{\sqrt{d}}\left[logf(x_1)\right]' + \frac{l^2{z_1}^2}{2!d}\left[logf(x_1)\right]'' + b_1\frac{l^3{z_1}^3}{3!d^{\frac{3}{2}}}\left[logf(\xi_1)\right]''' \tag{3.1.21}
$$

$$
\mathbb{W}(z_1, x_1, d) = \Phi\left(\frac{\eta(x_1, z_1) - \frac{{z_1}^2l^2}{2}\mathbb{I}}{\sqrt{{z_1}^2l^2\mathbb{I}}}\right) + e^{\eta(x_1,z_1)}\Phi\left(\frac{-\frac{{z_1}^2l^2\mathbb{I}}{2} - \eta(x_1, z_1)}{\sqrt{{z_1}^2l^2\mathbb{I}}}\right) \tag{3.1.22}
$$

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

product $b_1\epsilon$. Now we consider the Taylor series expansion around x_1 of the term

$$\begin{aligned} & dE_{z_1}\left[\left(V\left(x_1+\frac{z_1 l}{\sqrt{d}}\right)-V\left(x_1\right)\right)\mathbb{W}\left(z_1,x_1,d\right)\right] \\ = \; & dE_{z_1}\left[\left\{V^{'}(x_1)\frac{z_1 l}{\sqrt{d}}+\frac{1}{2}V^{''}(x_1)\frac{{z_1}^2 l^2}{d}+\frac{1}{6}V^{'''}(\xi_1)\frac{{z_1}^3 l^3}{d^{\frac{3}{2}}}\right\}\mathbb{W}\left(z_1,x_1,d\right)\right] \end{aligned} \tag{3.1.23}$$

From **Eqn 3.1.22**, it is quite clear that $\mathbb{W}(z_1, x_1)$ is continuous but is not differentiable at the point 0. So, we cannot split it into a Taylor series around 0. Also, note that $\mathbb{W}$ is a bounded function with respect to d. This follows from the fact that Φ is a bounded function and that $\eta(x_1, z_1)$ is a function that goes to 0 as $d \to \infty$. We assume that $log(f)$ is twice continuously differentiable and $[log(f)]^{(k)}$ is bounded for $k = 1(|)3$. Under this assumption, we can neglect the cubic terms of z_1, and hence essentially our generator reduces to

$$G_d V(x) = V^{'}(x_1)\sqrt{d}lE_{z_1}\left[z_1\mathbb{W}\left(z_1,x_1,d\right)\right]+\frac{1}{2}V^{''}(x_1)l^2E_{z_1}\left[{z_1}^2\mathbb{W}\left(z_1,x_1,d\right)\right] \tag{3.1.24}$$

In evaluating the second integral, we can simplify the expression using Dominated Convergence Theorem, so that as $d \to \infty$,

$$\mathbb{W}(z_1, x_1, d) \to 2\Phi\left(-\frac{\sqrt{{z_1}^2 l^2 \mathbb{I}}}{2}\right) \tag{3.1.25}$$

Note that the limiting distribution in that case is independent of x_1. But, the first integral is a cause of concern. In order to manage it, we use an approximation for

Normal cumulative distribution given by Aludaat and Alodat [AA08].

$$\Phi(x) \approx 0.5\left[1+\sqrt{1-e^{-\sqrt{\frac{\pi}{8}}x^2}}\right] \quad \forall x \geq 0 \tag{3.1.26}$$

Note that since the limiting form of $\mathbb{W}(z_1, x_1, d)$ is an even function, hence $E_{z_1}\left[z_1 \mathbb{W}\left(z_1, x_1, d\right)\right]$ must go to 0 as $d \to \infty$. The fact that the limiting value of $\sqrt{d}E_{z_1}\left[z_1 \mathbb{W}\left(z_1, x_1, d\right)\right]$ is finite follows from **Eqn 3.1.10** which proves that $G_d V(x)$ is bounded.

All integrals can be numerically computed using the approximation in [AA08]. Note that for standard RWMH, the Diffusion process reduces to the Langevin diffusion where the limiting form of $G_d V(x)$, say L is given by

$$L = h(l)\left[\frac{1}{2}V^{''}(x) + \frac{1}{2}[logf(x)]^{'}V^{'}(x)\right] \tag{3.1.27}$$

The coefficient of $V^{''}(x)$ is called the diffusion speed and it is in this case $h(l)$ given by **Eqn 3.1.5**. Note that if a priori, we had not chosen the distribution of z_1 to be $N(0,1)$ but were flexible in the choice of its distribution, say q_Z, such that it has 0 mean and variance 1, then note that our speed would have been equivalent to the speed of the standard RWMH if the mass were concentrated only around $+1$ and -1.
We choose that l that maximizes the quantity $2l^2$. $\int \left\{ {z_1}^2 \Phi\left(-\frac{\sqrt{{z_1}^2 l^2 \mathbb{I}}}{2}\right)\right\} q_Z(z)dz$, and use it to compute the acceptance rate.

Questions:

- Can the expectations be analytically derived? The second expectation is possible to derive analytically but the first one is difficult to handle.

- The analytical value of l_{max} in our case. Again for the first expectation, difficult to handle. Since the equation is not Langevin, the concept of speed is a bit blurred.

- Generalization of this method to independent but not identically distributed, and to dependent families of distributions as well.

3.2 TMCMC within Gibbs for iid product densities

The main notion of Gibbs' sampling is to update one or multiple components of a multidimensional stochastic process 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 $c_d \to c$. In order to explain the transitions in this process analytically, we define an indicator function $\mathbb{I}_i$ for $i = 1(|)d$. For fixed d,

$$\begin{aligned} \chi_i &= 1 \quad if\ transition\ in\ i^{th}\ co-ordinate \\ &= 0 \quad if\ no\ transition\ in\ i^{th}\ co-ordinate \end{aligned} \tag{3.2.1}$$

$$P(\chi_i = 1) = c_d \ \ \forall i = 1(|)d \tag{3.2.2}$$

Then a feasible transition can be analytically written as

$$(x_1, x_2, \ldots, x_d) \to (x_1 + \chi_1 b_1 \epsilon, x_2 + \chi_2 b_2 \epsilon, \ldots, x_d + \chi_d b_d \epsilon) \tag{3.2.3}$$

We can write down the generator $G_dV(x)$ as follows

$$G_dV(x) = \frac{d}{2}P(\chi_1 = 1)\int_0^\infty \sum_{b_1 \in -1,+1} \left[\left(V(x_1 + b_1\epsilon) - V(x_1) \right) \times E_{\{b_2,\ldots,b_d,\chi_2,\ldots,\chi_d\}} \left(min \left\{ 1, \frac{\pi(x_1 + b_1\epsilon, x_2 + \chi_2 b_2\epsilon, \ldots, x_d + \chi_d b_d\epsilon)}{\pi(x_1, x_2, \ldots, x_d)} \right\} \right) \right] q(\epsilon)d\epsilon \tag{3.2.4}$$

Note that since V is a function of x_1 only, if χ_1 is equal to 0, then no transition takes place and the value of the generator is 0. So, we need to consider only the feasible moves into account. Since b_j and χ_j always occur as products,

$$E\left\{ \begin{array}{l} b_2, b_3, \ldots, b_d \\ \chi_2, \chi_3, \ldots, \chi_d \end{array} \right\} = E_{\{b_2\chi_2, b_3\chi_3, \ldots, b_d\chi_d\}} \tag{3.2.5}$$

We apply similar form of approach as in **Eqn 3.1.12**. We have to leave out $(1-c_d)(d-1)$ many terms at each step and we sum over $c_d d$ many terms inside the exponential. We make a very vital assumption that $c_d \to c$, which forces $c_d(d-1)$ to go to ∞ as $d \to \infty$. We apply Lyapunov's Central Limit Theorem (again the Lyapunov assumption holds good for δ=2) as before to obtain

$$X \sim N\left(\eta(x_1, b_1, \epsilon) + \frac{\epsilon^2}{2!}\sum_{j=2}^{c_d d} \{log(f(x_j))\}'', \epsilon^2 \sum_{j=2}^{c_d d} \left[\{log(f(x_j))\}' \right]^2 + \Delta \right) \tag{3.2.6}$$

$$X \sim N\left(\eta(x_1, b_1, \epsilon) - \frac{c_d(d-1)\epsilon^2}{2}\mathbb{I},\ c_d(d-1)\epsilon^2\mathbb{I} \right) \tag{3.2.7}$$

Analogously, we define $\mathbb{W}(z_1, x_1, c_d, d)$ as the following

$$\mathbb{W}(z_1, x_1, c_d, d) = \Phi\left(\frac{\eta(x_1, z_1) - \frac{{z_1}^2 l^2 c_d \mathbb{I}}{2}}{\sqrt{{z_1}^2 l^2 c_d \mathbb{I}}}\right) + e^{\eta(x_1, z_1)}\Phi\left(\frac{-\frac{{z_1}^2 l^2 c_d \mathbb{I}}{2} - \eta(x_1, z_1)}{\sqrt{{z_1}^2 l^2 c_d \mathbb{I}}}\right) \quad (3.2.8)$$

So, finally the limiting form of the generator would also be quite analogous to what we acquired in the previous section

$$G_d V(x) = V^{'}(x_1) c_d \sqrt{d} l E_{z_1}\left[z_1 \mathbb{W}\left(z_1, x_1, c_d, d\right)\right] + \frac{1}{2} c_d V^{''}(x_1) l^2 E_{z_1}\left[{z_1}^2 \mathbb{W}\left(z_1, x_1, d\right)\right] \quad (3.2.9)$$

$$\mathbb{W}(z_1, x_1, c_d, d) \to 2\Phi\left(-\frac{\sqrt{{z_1}^2 l^2 c \mathbb{I}}}{2}\right) \quad (3.2.10)$$

Our diffusion speed is then $2cl^2$. $\int \left\{ {z_1}^2 \Phi\left(-\frac{\sqrt{{z_1}^2 l^2 c \mathbb{I}}}{2}\right)\right\} q_Z(z) dz$. This is corresponding to the diffusion speed $h_c(l) = 2ci^2\Phi\left(-\frac{\sqrt{l^2 c \mathbb{I}}}{2}\right)$ for the standard RWMH within Gibbs approach as given by Roberts and Neal [NR06].

3.3 Diffusion approximation for independent but not *iid* random variables

So far we considered only those target densities π which comprises of *iid* components. Now, we extend our investigation to those target densities that comprise of independnet but not identically distributed random variables. So,

$$\pi_d(x) = \prod_{i=1}^{d} f_i(x_i) \quad (3.3.1)$$

We concentrate on a particular form of the target density involving some scaling constant parameters, as considered in [Bed08][BR08].

$$\pi(x_d) = \prod_{j=1}^{d} \theta_j(d) f(\theta_j(d) x_j) \tag{3.3.2}$$

We assume the Lipschitz continuity and C^3 properties of f together with **Eqn 3.1.2** and **Eqn 3.1.3**. 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.

So,

$$\Theta(d) = \Bigg(\theta_1(d), \theta_2(d), \ldots, \theta_k(d), \theta_{k+1}(d), \ldots, \theta_{k+m}(d), \underbrace{\theta_{k+1}(d), \ldots, \theta_{k+1}(d)}_{c(1,d)-1}, \underbrace{\theta_{k+2}(d), \ldots, \theta_{k+2}(d)}_{c(2,d)-1}, \ldots, \underbrace{\theta_{k+m}(d), \ldots, \theta_{k+m}(d)}_{c(m,d)-1}, \Bigg) \tag{3.3.3}$$

where $c(1,d), c(2,d), \ldots, c(m,d)$ are the number of occurrences of the parameters in each of the m distinct classes. We assume that for any i, $\lim_{d\to\infty} c(i,d) = \infty$. Also, we assume a particular form of each scaling parameter $\theta_i(d)$.

$$\frac{1}{\{\theta_i(d)\}^2} = \frac{K_i}{d^{\lambda_i}} \qquad \forall i = 1, 2, \ldots, k \qquad\qquad \frac{1}{\{\theta_i(d)\}^2} = \frac{K_i}{d^{\gamma_i}} \qquad \forall i = k{+}1, 2, \ldots, k{+}m$$

Assume that the ${\theta_i}^{-2}(d)$ be so arranged that γ_i are in a decreasing sequence for $i = 1(|)m$ and also λ_i form a decreasing sequence from $j = 1(|)k$. According to [Bed07], the optimal form of the scaling variance $\sigma^2(d)$ would be of the form $\sigma^2(d) = \frac{l^2}{d^\alpha}$, where l^2 is some

constant and α satisfies

$$\lim_{d\to\infty}\frac{d^{\lambda_1}}{d^{\alpha}}<\infty \qquad\qquad \lim_{d\to\infty}\frac{d^{\gamma_i}c(i,d)}{d^{\alpha}}<\infty \qquad \forall i=1,2,\ldots,m \tag{3.3.4}$$

Let $\mathbb{Z}_t$ be the process at time t sped up by a factor of d^{α}, that means $\mathbb{Z}_t=(X_1(d^{\alpha}t),X_2(d^{\alpha}t),$ $\ldots,X_d(d^{\alpha}t))$. So, the generator function of the process can be written as

$$\begin{aligned} G_dV(x) &= \frac{d^{\alpha}}{2}\int_0^{\infty}\sum_{b_1\in-1,+1}\left[\Big(V(x_1+b_1\epsilon)-V(x_1)\Big)\right. \\ &\qquad \left.\times E_{b_2,b_3,\ldots,b_d}\left(min\left\{1,\frac{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon,\ldots,x_d+b_d\epsilon)}{\pi(x_1,x_2,\ldots,x_d)}\right\}\right)\right]q(\epsilon)d\epsilon \end{aligned} \tag{3.3.5}$$

Once again, we look at the term $min\left\{1,\frac{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon,\ldots,x_d+b_d\epsilon)}{\pi(x_1,x_2,\ldots,x_d)}\right\}$ and to write it as $min1,e^X$ where $X\sim N(0,1)$. Assume $\theta_1(d)=1$.

$$\begin{aligned} &E\Big|_{b_1\epsilon}\left[min\left\{1,\frac{\pi(x_1+b_1\epsilon,x_2+b_2\epsilon,\ldots,x_d+b_d\epsilon)}{\pi(x_1,x_2,\ldots,x_d)}\right\}\right] \\ &= E\Big|_{b_1\epsilon}\left[min\left\{1,exp\left(log(f(x_1+b_1\epsilon))-log(f(x_1))\right.\right.\right. \\ &\quad +\sum_{j=2}^{k}\left\{b_j\epsilon\left\{log(f(\theta_j(d)x_j))\right\}'+\frac{\epsilon^2}{2!}\left\{log(f(\theta_j(d)x_j))\right\}''+\frac{b_j\epsilon^3}{3!}\left\{log(f(\theta_j(d)z_j))\right\}'''\right\} \\ &\quad \left.\left.\left.+\sum_{j=k+1}^{d}\left\{b_j\epsilon\left\{log(f(\theta_j(d)x_j))\right\}'+\frac{\epsilon^2}{2!}\left\{log(f(\theta_j(d)x_j))\right\}''+\frac{b_j\epsilon^3}{3!}\left\{log(f(\theta_j(d)z_j))\right\}'''\right\}\right)\right\}\right] \end{aligned} \tag{3.3.6}$$

Note that since ϵ can be written as $\frac{lz_1}{d^{\frac{\alpha}{2}}}$ where we assume that $\alpha>0$, hence, since k is

finite, the first sum in the expression in **Eqn 3.3.6** goes in probability to 0 because . Then, we apply Lyapunov CLT on b_j for $j = k+1, \ldots, d$, which comprises infinitely many random variables as $d \to \infty$ and finally applying Slutsky's theorem which states that

$$Z_n \xrightarrow{d} Z \qquad Y_n \xrightarrow{P} 0 \qquad Z_n + Y_n \xrightarrow{d} Z$$

we get

$$X \sim N\left(\eta(x_1, b_1, \epsilon) + \frac{\epsilon^2}{2!}\sum_{j=k+1}^{d}\left\{log(f(\theta_j(d)x_j))\right\}^{''}, \epsilon^2\sum_{j=2}^{d}\left[\left\{log(f(\theta_j(d)x_j))\right\}^{'}\right]^2 + \Delta\right) \tag{3.3.7}$$

where

$$\Delta = \epsilon^4\sum_{j=2}^{d} 2\left\{log(f(\theta_j(d)x_j))\right\}^{'}\left\{log(f(\theta_j(d)z_j))\right\}^{'''} + \epsilon^6\sum_{j=2}^{d}\left[\left\{log(f(\theta_j(d)z_j))\right\}^{'''}\right]^2 \tag{3.3.8}$$

As $d \to \infty$, then

$$\epsilon^2\sum_{j=2}^{d}\left[\left\{log(f(\theta_j(d)x_j))\right\}^{'}\right]^2 \to \sum_{i=1}^{m}\frac{l^2{z_1}^2 d^{\gamma_i}}{K_i d^{\alpha}} E\left[\left\{\frac{f^{'}(Y)}{f(Y)}\right\}^2\right] \;=\; l^2{z_1}^2\xi^2\mathbb{I} \tag{3.3.9}$$

We then follow a similar approach as in the previous two cases to obtain

$$G_d V(x) = V^{'}(x_1) d^{\frac{\alpha}{2}} l E_{z_1}\left[z_1 \mathbb{W}\left(z_1, x_1, d, \xi\right)\right] + \frac{1}{2}V^{''}(x_1) l^2 E_{z_1}\left[{z_1}^2 \mathbb{W}\left(z_1, x_1, d, \xi\right)\right] \tag{3.3.10}$$

where

$$\mathbb{W}(z_1, x_1, d, \xi) == \Phi\left(\frac{\eta(x_1, z_1) - \frac{{z_1}^2 l^2 \xi^2}{2}\mathbb{I}}{\sqrt{{z_1}^2 l^2 \xi^2 \mathbb{I}}}\right) + e^{\eta(x_1, z_1)}\Phi\left(\frac{-\frac{{z_1}^2 l^2 \xi^2 \mathbb{I}}{2} - \eta(x_1, z_1)}{\sqrt{{z_1}^2 l^2 \xi^2 \mathbb{I}}}\right) \tag{3.3.11}$$

By an argument similar to **Eqn 3.1.10**, we can claim that $G_d V(x)$ is bounded under this set up as well and that forces $d^{\frac{\alpha}{2}} l E_{z_1}\left[z_1 \mathbb{W}(z_1, x_1, d, \xi)\right]$ to be a bounded function. The diffusion speed can be calculated analogously as

$$h_B(l) = 2l^2. \int \left\{ {z_1}^2 \Phi\left(-\frac{\sqrt{{z_1}^2 l^2 \xi^2 \mathbb{I}}}{2}\right)\right\} q_Z(z) dz$$

3.4 Observations and Concluding remarks

So far we have theoretically derived the diffusion process approximation of our TMCMC approach and also found the diffusion speed guiding the process in each case. Note that standard RWMH had resulted in Langevin diffusion equation for the *iid*, independent components with appropriate scaling (see Roberts, Gelman and Gilks [RGG97], M. Bedard [Bed08]). As a result the concept of *speed* of the Langevin SDE under standard approach was pretty clear. Note that in TMCMC however, the notion of *speed* is not very clear as the diffusion equation has varying drift and diffusion speeds unlike in the case of RWMH. One may call the coefficient of V'' to be the diffusion speed of the process. But, still one cannot compare in terms of *speed* between the two approaches. However, there is one common ground on which comparisons can be done and that is in terms of the expected acceptance rate given by $E_y\left[min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}\right]$ where y varies over all proposed moves from x. We know that for all the standard RWMH scenarios, the acceptance rate has been found to be 0.234. We shall see how much better or worse does our method perform compared to the standard RWMH in terms of expected acceptance

rate.

First, we considered the *iid* set up where the diffusion speed corresponding to our approach is given by

$$h_I(l) = 2l^2. \int \left\{ {z_1}^2 \Phi\left(-\frac{\sqrt{{z_1}^2 l^2 \mathbb{I}}}{2} \right) \right\} q_Z(z) dz \tag{3.4.1}$$

and it is easily checked that the expected acceptance ratio, as defined by us, in this case is

$$AR_{TMCMC} = 2. \int \left\{ \Phi\left(-\frac{\sqrt{{z_1}^2 l^2 \mathbb{I}}}{2} \right) \right\} q_Z(z) dz \tag{3.4.2}$$

Note from **Thm 3** that the optimum value of l in RWMH is $\frac{2.38}{\sqrt{\mathbb{I}}}$ and corresponding expected acceptance rate is 0.23. However, TMCMC it was observed on maximizing **Eqn 3.4.1** that the optimal value of l is $\frac{2.42}{\sqrt{\mathbb{I}}}$ and the corresponding expected acceptance rate is 0.438, which is a huge improvement on the standard RWMH process.

We shall have the same acceptance rate for TMCMC within GIbbs and its standard counterpart as well because in that case we can just define $\mathbb{I}_c = c\mathbb{I}$ and notice that we shall have the optimal l value in RWMH and TMCMC to be $\frac{2.38}{\sqrt{\mathbb{I}_c}}$ and $\frac{2.42}{\sqrt{\mathbb{I}_c}}$ respectively and the resulting acceptance rates would remain the same. The same is true for the target density with properly scaled independent components considered in **Section 3** where instead of c in previous argument, we consider ξ^2.

We present the plot of the diffusion speed with respect to scaling factor l for i.i.d. set up **Fig 6.4.3**, and the that corresponding to the TMCMC within Gibbs approach for various parameters in **Fig 6.4.4**, **Fig 6.4.5** and **Fig 6.4.6**.

This chapter shows that for a wide range of distributions (i.i.d and independent components with different scalings) indeed TMCMC is better than RWMH as far as acceptance rate goes. Later on, in **Chapter 6**, we shall present some simulation experiments with optimal scalings and compare the practical acceptance rates and investigate the convergence with respect to some performance evaluation measures. Now, we shall over to a more general class of distributions where we relax independence and impose a weaker dependence criterion instead. We shall then compare the optimal acceptance rates of TMCMC and RWMH for that class of distributions.

Chapter 4

Optimal scaling of TMCMC algorithm (dependent components)

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 [MPS11], [BRS09], [BS08], [Bed09]. In particular, we assume that π_0 is a Gaussian measure with mean 0 and covariance Σ and the Radon Nikodym derivative of the target density π with respect to π_0 is given by

$$\frac{d\pi}{d\pi_0} = M_\Psi \exp(-\Psi(x)), \tag{4.0.1}$$

for a real π_0-measurable function Ψ on $\mathbb{R}^\infty$, and M_Ψ is a normalizing constant. Since Σ is assumed to be positive definite, we can always find eigenvalues ${\lambda_j}^2$ and orthonormal eigenvectors ϕ_j such that

$$\Sigma\phi_j = {\lambda_j}^2\phi_j; \quad j = 1, 2, \ldots. \tag{4.0.2}$$

Any vector x in $\mathbb{R}^\infty$ can be uniquely represented as

$$x = \sum_{j=1}^{\infty} x_j\phi_j, \quad \text{where} \quad x_j = \langle x, \phi_j \rangle \tag{4.0.3}$$

4.1 TMCMC diffusion approximation for Gaussian dominated dependent family

Following [MPS11] we approximate the inifinite-dimensional measure using a finite dimensional target measure π^d involving only the first d co-ordinates of x. As $N \to \infty$, π^d would eventually converge to π almost everywhere. We define

$$\Psi^d(x) = \Psi(P^d x), \quad \text{and} \quad \frac{d\pi}{d\pi_0} \propto \exp(-\Psi^d(x)). \tag{4.1.1}$$

As in [MPS11], we represent $\pi^d(x)$ as

$$\pi^d(x) = M_{\Psi^d} \exp\left(-\Psi^d(x) - \frac{1}{2}\langle x, (\Sigma^d)^{-1} x\rangle \right) \tag{4.1.2}$$

where $\Sigma^d = P^d \Sigma P^d$. Under the TMCMC model 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

$$x^{k+1} = \gamma^{k+1} y^{k+1} + \left(1 - \gamma^{k+1}\right) x^k, \tag{4.1.3}$$

where

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

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

$$y^{k+1} = x^k + \sqrt{\frac{2\ell^2}{d}} \Sigma^{\frac{1}{2}} \xi^{k+1}, \tag{4.1.4}$$

where $\xi^{k+1} = \left({b_1}^{k+1} \epsilon^{k+1}, \ldots, {b_d}^{k+1} \epsilon^{k+1} \right)$ where $b_i = \pm 1$ with probability $1/2$ each, and $\epsilon \sim N(0,1) I_{\{\epsilon > 0\}}$. From (4.1.2) it follows that $\min\left\{ 1, \frac{\pi(y^{k+1})}{\pi(x^k)} \right\}$ can be written

as $\min\left\{1, e^{\mathbb{Q}}\right\}$ where $\mathbb{Q}$ is given by

$$\mathbb{Q} = \frac{1}{2}\left\|\Sigma^{-\frac{1}{2}}\left(P^d x^k\right)\right\|^2 - \frac{1}{2}\left\|\Sigma^{-\frac{1}{2}}\left(P^d y^{k+1}\right)\right\|^2 + \Psi^d(x^k) - \Psi^d(y^{k+1}). \tag{4.1.5}$$

Using (4.1.4), one obtains

$$\mathbb{Q} = -\sqrt{\frac{2\ell^2}{d}}\langle \eta, \xi\rangle - \frac{\ell^2}{d}\|\xi\|^2 - r(x,\xi), \tag{4.1.6}$$

where

$$\eta = \Sigma^{-\frac{1}{2}}(P^d x^k) + \Sigma^{\frac{1}{2}}\nabla\Psi^d(x^k), \tag{4.1.7}$$

and

$$r(x,\xi) = \Psi^d(y^{k+1}) - \Psi^d(x^k) - \langle \nabla\Psi^d(x^k), P^d y^{k+1} - P^d x^k\rangle. \tag{4.1.8}$$

We further define

$$R(x,\xi) = -\sqrt{\frac{2\ell^2}{d}}\sum_{j=1}^{d}\eta_j\xi_j - \frac{\ell^2}{d}\sum_{j=1}^{d}{\xi_j}^2, \tag{4.1.9}$$

and

$$R_i(x,\xi) = -\sqrt{\frac{2\ell^2}{d}}\sum_{j=1,j\neq i}^{d}\eta_j\xi_j - \frac{\ell^2}{d}\sum_{j=1,j\neq i}^{d}{\xi_j}^2 \tag{4.1.10}$$

Using Lemma 5.5 of [MPS11], for large d one can show that

$$\mathbb{Q} \;=\; R(x,\xi) - r(x,\xi) \;\approx\; R_i(x,\xi) - \sqrt{\frac{2\ell^2}{d}}\eta_i\xi_i. \tag{4.1.11}$$

Using (4.1.9) and (4.1.11) it can be seen that $\mathbb{Q}$ is approximately equal to $R(x,\xi)$ as d

goes to ∞, where $R(x,\xi)$ in our case is given by

$$R(x,\xi) = -\epsilon\sqrt{\frac{2\ell^2}{d}}\sum_{j=1}^{d}\eta_j b_j - \ell^2\epsilon^2. \tag{4.1.12}$$

To apply Lyapunov's central limit theorem we need to show the following: with probability 1 with respect to π,

$$\frac{\sum_{j=1}^{d} E\left(\frac{b_j\eta_j}{\sqrt{d}}\right)^4}{\left(\sqrt{\sum_{j=1}^{d}\frac{\eta_j^2}{d}}\right)^4} = \frac{\sum_{j=1}^{d}\frac{\eta_j^4}{d^2}}{\left(\frac{\|\eta\|^2}{d}\right)^2} \to 0, \quad \text{as} \quad d\to\infty. \tag{4.1.13}$$

By Lemma 5.2 of [MPS11], $\frac{\|\eta\|^2}{d} \to 1$ π-almost surely as $d\to\infty$. This implies that the denominator of the left hand side of (4.1.13) goes to 1 π-almost surely, as $d\to\infty$. Now, $\left(\frac{\|\eta\|^2}{d}\right)^2 = \sum_{j=1}^{d}\frac{\eta_j^4}{d^2} + \sum_{i=1}^{d}\frac{\eta_i^2}{d}\left(\sum_{j\neq i}\frac{\eta_j^2}{d}\right)$. Except on a π-null set $\mathcal{N}$, where $\sum_{j=1}^{d}\frac{\eta_j^2}{d}$ need not converge to 1, we have, for given $\epsilon>0$ and d_0 depending upon ϵ, $1-\epsilon < \sum_{j\neq i}\frac{\eta_j^2}{d} < 1+\epsilon$ and $1-\epsilon < \sum_{i=1}^{d}\frac{\eta_i^2}{d} < 1+\epsilon$, for $d\geq d_0$. Hence, for $d\geq d_0$, $-\epsilon^2-2\epsilon < \epsilon^2-2\epsilon = (1-\epsilon)^2-1 < \sum_{i=1}^{d}\frac{\eta_i^2}{d}\left(\sum_{j\neq i}\frac{\eta_j^2}{d}\right) - 1 < (1+\epsilon)^2-1 = \epsilon^2+2\epsilon$, so that $\left|\sum_{i=1}^{d}\frac{\eta_i^2}{d}\left(\sum_{j\neq i}\frac{\eta_j^2}{d}\right)-1\right| < \epsilon^2+2\epsilon$, showing that $\sum_{i=1}^{d}\frac{\eta_i^2}{d}\left(\sum_{j\neq i}\frac{\eta_j^2}{d}\right)\to 1$ on $\mathcal{N}^c$, the complement of $\mathcal{N}$. Since on $\mathcal{N}^c$, $\left(\frac{\|\eta\|^2}{d}\right)^2\to 1$, we must have $\sum_{j=1}^{d}\frac{\eta_j^4}{d^2}\to 0$ on $\mathcal{N}^c$, showing that Lyapunov's condition (4.1.13) holds almost surely with respect to π.

Using Lyapunov's central limit theorem on b_j, and using the result that $\frac{\|\eta\|^2}{d}\to 1$ π-almost surely as $d\to\infty$, we obtain, for sufficiently large d,

$$R(x,\xi) \approx \mathbb{Z} \approx N(-\ell^2\epsilon^2, 2\ell^2\epsilon^2). \tag{4.1.14}$$

Now, (4.1.9) and the fact that for large d, $\mathbb{Q} \approx R(x, \xi)$, imply

$$\mathbb{Q} \approx -\epsilon\sqrt{\frac{2\ell^2}{d}}\left(\eta_i b_i + \sum_{j=1, j\neq i}^{d} \eta_j b_j\right) - \ell^2\epsilon^2, \tag{4.1.15}$$

so that

$$[\mathbb{Q}|b_i, \epsilon] \approx N\left(-\ell^2\epsilon^2 - \epsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i, 2\ell^2\epsilon^2\right) \tag{4.1.16}$$

4.1.1 Expected drift

In order to obtain the diffusion approximation, we first obtain the expected drift conditions. In order to do that, we first define $\mathcal{F}_k$ to be the sigma algebra generated by $\{x^n, \xi^n, \gamma^k, n \leq k\}$, and denote the conditional expectations $E(\cdot|\mathcal{F}_k)$ by $E_k(\cdot)$. We then note that under stationarity, $E_k\left(x^{k+1} - x^k\right) = E_0\left(x^1 - x^0\right)$, and using (4.1.3) we can write

$$\begin{aligned} dE_0\left({x_i}^1 - {x_i}^0\right) &= dE_0\left[\gamma^1\left({y_i}^1 - {x_i}^1\right)\right] \\ &= dE_0\left[\alpha(x, \xi)\sqrt{\frac{2\ell^2}{d}}\left(\Sigma^{\frac{1}{2}}\xi\right)_i\right] \\ &= \frac{1}{\eta_i}\lambda_i\sqrt{2\ell^2 d}E_0\left[\min\left\{1, e^{\mathbb{Q}}\right\}\xi_i\right]\eta_i, \end{aligned} \tag{4.1.17}$$

where $\alpha(x,\xi) = \min\left\{1, \frac{\pi(y_i{}^1)}{\pi(x_i{}^1)}\right\}$. The last step follows from (4.0.2); noting that $\lambda_i \Sigma^{-\frac{1}{2}} \phi_i = \phi_i$, (4.1.7) yields

$$\begin{aligned}
\lambda_i \eta_i &= \lambda_i \left\langle \Sigma^{-\frac{1}{2}}(P^d x^0) + \Sigma^{-\frac{1}{2}} \nabla \Psi^d(x^0), \phi_i \right\rangle \\
&= \left\langle P^d x^0 + \nabla \Psi^d(x^0), \phi_i \right\rangle \\
&= \left(P^d x^0 + \nabla \Psi^d(x^0)\right)_i .
\end{aligned} \tag{4.1.18}$$

Thus, we can write

$$dE_0\left(x_i{}^1 - x_i{}^0\right) = \frac{1}{\eta_i}\left(P^d x^0 + \nabla \Psi^d(x^0)\right)_i \sqrt{2\ell^2 d} E_0\left[\min\left\{1, e^{\mathbb{Q}}\right\} \xi_i\right]. \tag{4.1.19}$$

Now, writing $\mu = -\ell^2\epsilon^2 - \epsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i$, $\sigma = \sqrt{2}\ell\epsilon$, and using (4.1.16), it follows that

$$\begin{aligned}
&\sqrt{d} E_0\left[\min\left\{1, e^{\mathbb{Q}}\right\} \xi_i\right] \\
&= \sqrt{d} E_{b_i\epsilon}\left[b_i \epsilon E_0\left\{\min\left\{1, e^{\mathbb{Q}}\right\} \middle| b_i, \epsilon\right\}\right] \\
&\approx \sqrt{d} E_{b_i\epsilon}\left[b_i \epsilon\left\{\Phi\left(\frac{\mu}{\sigma}\right) + e^{\mu + \frac{\sigma^2}{2}} \Phi\left(-\sigma - \frac{\mu}{\sigma}\right)\right\}\right] \\
&= \sqrt{d} E_{b_i\epsilon}\left[b_i \epsilon\left\{\Phi\left(-\frac{\ell\epsilon}{\sqrt{2}} - \frac{\eta_i b_i}{\sqrt{d}}\right)\right.\right. \\
&\qquad \left.\left. + e^{-\epsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i} \Phi\left(-\frac{\ell\epsilon}{\sqrt{2}} + \frac{\eta_i b_i}{\sqrt{d}}\right)\right\}\right].
\end{aligned} \tag{4.1.20}$$

Using the following Taylor's series expansions

$$\begin{aligned}
\Phi\left(-\frac{\ell\epsilon}{\sqrt{2}}-\frac{\eta_i b_i}{\sqrt{d}}\right) &= \Phi\left(-\frac{\ell\epsilon}{\sqrt{2}}\right)-\frac{\eta_i b_i}{\sqrt{d}}\phi\left(-\frac{\ell\epsilon}{\sqrt{2}}\right)+\frac{\eta_i^2}{d}\phi'(w_1),\\
\Phi\left(-\frac{\ell\epsilon}{\sqrt{2}}+\frac{\eta_i b_i}{\sqrt{d}}\right) &= \Phi\left(-\frac{\ell\epsilon}{\sqrt{2}}\right)+\frac{\eta_i b_i}{\sqrt{d}}\phi\left(-\frac{\ell\epsilon}{\sqrt{2}}\right)+\frac{\eta_i^2}{d}\phi'(w_2),\\
e^{-\epsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i} &= 1-\epsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i+\frac{2\ell^2\epsilon^2\eta_i^2}{d}e^{-w_3},
\end{aligned} \tag{4.1.21}$$

where w_1 lies between $-\frac{\ell\epsilon}{\sqrt{2}}$ and $-\frac{\ell\epsilon}{\sqrt{2}}-\frac{\eta_i bi}{\sqrt{d}}$; w_2 lies between $-\frac{\ell\epsilon}{\sqrt{2}}$ and $-\frac{\ell\epsilon}{\sqrt{2}}+\frac{\eta_i bi}{\sqrt{d}}$, and w_3 lies between 0 and $\epsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i$, and noting that $E_{b_i\epsilon}\left[\Phi\left(-\frac{\ell\epsilon}{\sqrt{2}}\right)\right]=E_{b_i\epsilon}\left[b_i\phi\left(-\frac{\ell\epsilon}{\sqrt{2}}\right)\right]=0$, $\frac{1}{\sqrt{d}}E_{b_i\epsilon}\left[\epsilon\phi\left(-\frac{\ell\epsilon}{\sqrt{2}}\right)\right]\to 0$ as $d\to\infty$, (4.1.20) can be easily seen to be of the form

$$\begin{aligned}
\sqrt{d}E_0\left[\min\left\{1,e^{\mathbb{Q}}\right\}\xi_i\right] &\approx \sqrt{d}E_{b_i\epsilon}\left[b_i\epsilon\left\{\Phi\left(-\frac{\ell\epsilon}{\sqrt{2}}-\frac{\eta_i b_i}{\sqrt{d}}\right)\right.\right.\\
&\qquad \left.\left.+e^{-\epsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i}\Phi\left(-\frac{\ell\epsilon}{\sqrt{2}}+\frac{\eta_i b_i}{\sqrt{d}}\right)\right\}\right]\\
&\approx -\sqrt{2\ell^2}\eta_i\times 2\int_0^\infty u^2\Phi\left(-\frac{\ell u}{\sqrt{2}}\right)\phi(u)du\\
&= -\sqrt{\frac{\ell^2}{2}}\eta_i\beta,
\end{aligned} \tag{4.1.22}$$

where

$$\beta=4\int_0^\infty u^2\Phi\left(-\frac{\ell u}{\sqrt{2}}\right)\phi(u)du. \tag{4.1.23}$$

Hence, we can re-write (4.1.19) as

$$\begin{aligned}
dE_0\left({x_i}^1-{x_i}^0\right) &= \frac{1}{\eta_i}\left(P^d x^0+\nabla\Psi^d(x^0)\right)_i\sqrt{2\ell^2 d}E_0\left[\min\left\{1,e^{\mathbb{Q}}\right\}\xi_i\right]\\
&= -\ell^2\beta\left(P^d x^0+\nabla\Psi^d(x^0)\right)_i.
\end{aligned} \tag{4.1.24}$$

4.1.2 Expected diffusion coefficient

Now we evaluate the expected diffusion coefficients involving the cross product terms. For $1 \leq i \neq j \leq d$, we have

$$dE_0\left[\left({x_i}^1 - {x_i}^0\right)\left({x_j}^1 - {x_j}^0\right)\right] = dE_0\left[\left\{\gamma^1\left({y_i}^1 - {x_i}^0\right)\right\}\left\{\gamma^1\left({y_j}^1 - {x_j}^0\right)\right\}\right]$$

Check that if $i \neq j$, then the above expectation is 0 using the fact that $b_i b_j \epsilon$ has 0 mean for $i \neq j$. However for $i = j$, using (4.1.18) again, we can reduce the above expectation to

$$\begin{aligned} dE_0\left[\left({x_i}^1 - {x_i}^0\right)\left({x_j}^1 - {x_j}^0\right)\right] &= dE_0\left[\left({x_i}^1 - {x_i}^0\right)^2\right] \\ &= dE_0\left[\alpha(x^0, \xi)\left({y_i}^1 - {x_i}^0\right)^2\right] \\ &= 2\ell^2 {\lambda_i}^2 E_0\left[{\xi_i}^2 \min\left\{1, e^{\mathbb{Q}}\right\}\right]. \end{aligned} \tag{4.1.25}$$

Using the same Taylor's series expansions (4.1.21) it is easily seen that

$$\begin{aligned} E_0\left[{\xi_i}^2 \min\left\{1, e^{\mathbb{Q}}\right\}\right] &\approx 4\int_0^\infty u^2 \Phi\left(-\frac{\ell u}{\sqrt{2}}\right)\phi(u)du \\ &= \beta. \end{aligned} \tag{4.1.26}$$

Hence,

$$\begin{aligned} dE_0\left[\left(x_i{}^1 - x_i{}^0\right)\left(x_j{}^1 - x_j{}^0\right)\right] &= 2\ell^2\lambda_i{}^2 E_0\left[\xi_i{}^2 \min\left\{1, e^{\mathbb{Q}}\right\}\right] \\ &\approx 2\ell^2\lambda_i{}^2\beta \\ &= 2\ell^2\beta\langle\phi_i, \Sigma\phi_i\rangle. \end{aligned} \tag{4.1.27}$$

It follows that

$$dE_0\left[\left(x^1 - x^0\right)\otimes\left(x^1 - x^0\right)\right] \approx 2\ell^2\beta\Sigma^d. \tag{4.1.28}$$

Now defining the piecewise constant interpolant of x^k given by

$$z^d(t) = x^k; \quad t \in [t^k, t^{k+1}], \tag{4.1.29}$$

where $t^k = k/d$, it can be shown, proceeding in the same way, and using the same assumptions on the covariance operator and Ψ as [MPS11], that $z^d(t)$ converges weakly to z (see [MPS11] for the rigorous definition), where z satisfies the following stochastic differential equation:

$$\frac{dz}{dt} = -g(\ell)\left(z + \Sigma\nabla\Psi(z)\right) + \sqrt{2g(\ell)}\frac{dW}{dt}, \quad z(0) = z^0, \tag{4.1.30}$$

where $z_0 \sim \pi$, W is a Brownian motion in a relevant Hilbert space with covariance operator Σ, and

$$g(\ell) = \ell^2\beta, \tag{4.1.31}$$

is the diffusion speed. If ℓ_{opt} maximizes the diffusion speed, then the optimal acceptance

rate is given by

$$\alpha_{opt} = 4 \int_0^\infty \Phi\left(-\frac{\ell_{opt} u}{\sqrt{2}}\right) \phi(u) du. \tag{4.1.32}$$

4.2 TMCMC within Gibbs for this dependent family of distributions

As before, here we define transitions of the form (**??**), where the random variable χ_i indicates whether or not the i-th co-ordinate of x will be updated. The proof again required only minor modification to the above proof provided in the case of this dependent family of distributions. Here we only need to take expectations with respect to $\chi_i; i = 1, \ldots, d$, so that we now have

$$[\mathbb{Q}|b_i, \epsilon] \approx N\left(-\ell^2\epsilon^2 - c\epsilon\sqrt{\frac{2\ell^2}{d}}\eta_i b_i, 2\ell^2\epsilon^2 c^2\right).$$

Proceeding in the same manner as in the above proof, we obtain a stochastic differential equation of the same form as (4.1.30), but with $g(\ell)$ replaced with

$$g_c(\ell) = c\ell^2\beta_c, \tag{4.2.1}$$

where

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

The optimal acceptance rate is given by

$$\alpha_{opt} = 4 \int_0^\infty \Phi\left(-\frac{\ell_{opt} u}{c\sqrt{2}}\right) \phi(u) du, \tag{4.2.2}$$

where ℓ_{opt} maximizes $g_c(\ell)$.

4.3 Observations and Concluding remarks

The optimal acceptance rate for similar distributions corresponding to standard RWMH and its resultant Langevin diffusion has been found to be around 0.234 (see Mattingly, Pillai and Stuart [MPS11]). While speeds cannot be compared, acceptance rates can be. In the dependent case, the diffusion speed is and the optimal acceptance rate are of the forms (4.1.31) and (4.1.32), 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 $2\Phi\left(-\frac{\ell_{opt}}{\sqrt{2}}\right)$, turn out to be 0.234 as before, where ℓ_{opt} maximizes the corresponding diffusion speed $2\ell^2\Phi\left(-\frac{\ell}{\sqrt{2}}\right)$. Similar information as before are provided by Figure 6.4.7.

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, $2c\ell^2\Phi\left(-\frac{\ell}{c\sqrt{2}}\right)$, and $2\Phi\left(-\frac{\ell_{opt}}{c\sqrt{2}}\right)$. The corresponding TMCMC-based quantities are (4.2.1) and (4.2.2), respectively. The optimal acceptance rates remain 0.234 and 0.439 for TMCMC and RWM, respectively. Figure 6.4.8, comparing the diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the dependent set-up, lead to similar observations as before.

Chapter 5

Adaptive versions of TMCMC

Some highly desirable properties of the non-adaptive TMCMC mechanism like geometric ergodicity, optimal scaling have been studied and the relative advantage of this method over MCMC has been established in **Chapter 2**, **Chapter 3** and **Chapter 4**. This paper is a follow up study of the adaptive versions of the TMCMC mechanism and the comparative analysis of its performance with respect to adaptive MCMC methods. First, we discuss the ergodicity properties of adaptations on the TMCMC chain where we present the arguments and theoretical background for adaptive MCMC based on coupling as suggested in Roberts and Rosenthal [RR07] and its natural extension to the TMCMC case as well. Next, we present some examples of adaptive TMCMC, which are primarily derived from MCMC adaptations as proposed in Haario *et al* [HT01], Haario *et al*[HHT05], Roberts *et al*[RR09].

5.1 Preliminaries

Let π be the target density from which we intend to simulate a random sample. Since the chain we run in MCMC depends the choice of the proposal variance (assuming the proposal distribution is symmetric about 0 and follows a Gaussian distribution), therefore each RWMH or TMCMC chain has an index λ that takes into account all the parameters associated with the prescribed algorithm. We now state some preliminary notions of adaptive Markov chains without restricting our focus to either RWMH or

TMCMC. Let $\{P_\lambda\}$ be a collection of Markov chain kernels with an associated parameter λ, and a stationary distribution π such that

$$\pi P_\lambda = \pi$$

As shown in Meyn and Tweedie [MT93], if the family of Markov kernels $\{P_\lambda\}$ can be shown to be ϕ-irreducible and aperiodic (which is the case for RWMH [RT96] and TMCMC [DB11]), then the total variation distance of the Markov kernel, for a fixed λ, with respect to π goes to zero for any starting point x_0. This means if we run a Markov chain for a fixed choice of the parameter λ, irrespective of our starting value, we converge to a process with distribution π- the target distribution.

While convergence is guaranteed, the rate of this convergence however depends on various choices of the parameter λ. It has been shown that for distributions that are spherically symmetric, the RWMH chain and the TMCMC chain both are geometrically ergodic (see **Chapter 2** and Roberts and Tweedie [RT96]). This means there $\exists$ a function $V \geq 1$ and finite at least at one point, and also constants $0 < \rho < 1$ and M, so that for each λ,

$$||{P_\lambda}^n(x, .) - \pi(.)||_{TV} \leq M.V(x)\rho^n \quad \forall n \geq 1 \tag{5.1.1}$$

where $||\nu||_{TV}$ denotes the *total variation norm*. This condition implies convergence at a geometric rate for each λ (which can represent the proposal variance), but the function V and M will depend on the choice of λ. Under certain regularity conditions, such an optimal choice of λ has been suggested for both RWMH (see Roberts, Gelman and Gilks [RGG97]) and TMCMC (see **Chapter 3** and **Chapter 4**). However, from a practical point of view, it would be desirable to have an algorithm that starts from a

random value of the proposal variance and then sequentially updates it at each iteration such that in the long run, this adaptive algorithm converges to the target density π, thereby eliminating the problem of choice of the parameter λ.

In formal terms, let X_n be a $\mathcal{X}$ valued random variables and represents the state of the system at time n. We denote Λ_n as a $\mathcal{Z}$ valued random quantity that specifies the parameter. We define a filtration $\mathcal{F}_n$ on (X_n, Λ_n) as follows

$$\mathcal{F}_n = \sigma\left(X_0, X_1, \cdots, X_n, \Lambda_0, \Lambda_1, \cdots, \Lambda_n\right) \tag{5.1.2}$$

Therefore we have for any set A belonging to the projection of $\mathcal{F}_n$ on the $\mathcal{X}$ space,

$$P\left[X_{n+1} \in A | X_n = x, \Lambda_n = \lambda, \mathcal{F}_{n-1}\right] = P_{lambda}(x, A) \quad x \in \mathcal{X}\ \lambda \in \mathcal{Z} \tag{5.1.3}$$

Typically we define then

$$W^n((x, \lambda), A) = P\left[X_n \in A | X_0 = x, \Lambda_0 = \lambda\right] \quad x \in \mathcal{X}\ \lambda \in \mathcal{Z} \tag{5.1.4}$$

W^n represents the unconditional distribution obtained by integrating over Λ_j's for $j < n$. We define the total variation distance between the adaptive chain and the target density π

$$U(x, \lambda, n) = ||W^n((x, \lambda), .) - \pi(.)||_{TV} \tag{5.1.5}$$

The chain will be ergodic if this distance goes to 0 ans $n \to \infty$ whatever be my choice of x and λ. It can be shown that the adaptive chain may not be ergodic [RR09]. However Roberts and Rosenthal have provided some sufficient conditions to check for the

ergodicity of the adaptive chain under two scenarios- the uniformly converging case and the non-uniformly converging case, the second being a relaxation of the first condition [RR07].

5.2 Ergodicity of the Adaptive TMCMC chain

We start with the following theorem due to Roberts and Rosenthal [RR07].

Theorem 4 *Consider an adaptive Markov Chain algorithm, on a state space $\mathcal{X}$ and adaptation space $\mathcal{Y}$ denoted by $\{P_\lambda\}$ having a stationary distribution π. Then under the following conditions, the algorithm is ergodic.*

1. *(* Simultaneous uniform ergodicity*) For any arbitrary $\epsilon > 0$, there exists a N_ϵ such that $|| \{P_\lambda\}^N (x, .) - \pi(.)||_{TV} < \epsilon$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.*

2. *(* Diminishing adaptation*) Define*

$$Dn = \sup_{x \in \mathcal{X}} || \left\{P_{\Lambda_{n+1}}\right\}^N (x, .) - \left\{P_{\Lambda_n}\right\}^N (x, .)||_{TV} \tag{5.2.1}$$

which is a $\mathcal{F}_{n+1}$ measurable random variable. Then

$$D_n \xrightarrow{P} 0 \quad as\ n \to \infty \tag{5.2.2}$$

The condition of Simultaneous uniform ergodicity can be relaxed and ergodicity will hold if

$$M_\epsilon(x, \lambda) = inf \left\{ n \geq 1 : || \{P_\lambda\}^N (x, .) - \pi(.)||_{TV} < \epsilon \right\} \tag{5.2.3}$$

is finite for any choice of x, λ and ϵ. Thus for both the RWMH and TMCMC chains, if the adaptations are done so as to satisfy the conditions stated in **Thm 4** or the relaxation

of the conditions for the non uniformly converging case as discussed above. For jointly compact sample space ($\mathcal{X}$) and adaptation space ($\mathcal{Y}$), there is a simpler analog of the sufficient conditions in **Thm** 4, as discussed by Roberts and Rosenthal(2007). A similar result but with subtle modifications that takes into account the restriction of moves for TMCMC is established in a simplified form in textbfCor 1.

Note that the TMCMC chain is defined only over Euclidean spaces and so both $\mathcal{X}$ and $\mathcal{Y}$ take values either from $\mathbb{R}^d$ or some subspace of it. Under such assumptions, the following corollary to **Thm** 4 holds for adaptive TMCMC chains.

Corollary 1 *Suppose* $\{P_\lambda\}$ *be a class of TMCMC chains with a Gaussian proposal distribution having 0 mean and* λ *being the variance of the proposal density. Assume that the proposal density is uniformly bounded and the space* $\mathcal{X} \times \mathcal{Y}$ *is compact with respect to a metric topology. With respect to some reference measure* ν *(usually the Lebesgue measure), the n-step Markov proposal kernel has continuous density in* $\mathcal{X} \times \mathcal{Y}$ *space for* $n \geq 2$*. Also assume that the Radon Nikodym derivative of* π *with respect to this reference measure is also continuous.Then if an adaptation is performed over the parameter space* Λ *for this chain and if such a version of adaptive TMCMC satisfies the condition in Theorem 1 (2), then the adaptive chain is ergodic.*

Proof 4 *Suppose for a fixed* λ*, we start a TMCMC chain from a point* x *(a d-dimensional point). Let* $\alpha_\lambda(x)$ *be the acceptance rate of the TMCMC chain (probability that we move from from our current position). Explicitly it is of the form*

$$\alpha_\lambda(x) = \frac{d}{2^d} \sum_{\left\{ \begin{array}{l} b_i \in \{-1,+1\} \\ \forall i = 1(|)d \end{array} \right\}} \int_0^\infty min \left\{ 1, \frac{\pi(x_1 + b_1\epsilon, x_2 + b_2\epsilon, \ldots, x_d + b_d\epsilon)}{\pi(x_1, x_2, \ldots, x_d)} \right\} q_\gamma(\epsilon) d\epsilon \tag{5.2.4}$$

Now, we can write $\{P_\lambda\}(x, dz)$ *as a mixture of two components as in* ***Eqn refeq:cor1****.*

$$\{P_\lambda\}(x, dz) = (1 - \alpha_\lambda(x))\, \delta_x(z) + r_\lambda(x, z)\nu(dz) \tag{5.2.5}$$

Note that for the first step transition in TMCMC, $r_\lambda(x, z)$ *is positive only for those values of* z *that can be reached from* x *in the first step, meaning that* z *is either on the line passing through* x *and parallel to the line* $y = x$ *or on the line orthogonal to it at and intersecting at* x*. For this first step transition, the proposal Markov kernel therefore does not possess a continuous Radon Nikodym derivative. However for n-th step transition probability, where* $n \geq 2$*,*

$$\{P_\lambda\}^n(x, dz) = (1 - \alpha_\lambda(x))^n\, \delta_x(z) + {r_\lambda}^n(x, z)\nu(dz) \tag{5.2.6}$$

where z *can take any value from* $\mathcal{X}$*. In the above context* ν *is any dominating measure, but we consider it to be the Lebesgue measure on* $\mathbb{R}^d$ *without loss of generality and thus it can be easily checked that the measures* δ_x *and* ν *are mutually singular. We can then say*

$$|\{P_\lambda\}^N(x, .) - \pi(.)||_{TV} = (1 - \alpha_\lambda(x))^n + \frac{1}{2}\int_{\mathcal{X}} |{r_\lambda}^n(x, z) - s(z)|\nu(dz) \tag{5.2.7}$$

where $\pi(dz) = s(z)\nu(dz)$ *and the integral is basically the Lebesgue-Stieltjes integral. By bounded convergence theorem this quantity is jointly continuous in* $\mathcal{X} \times \mathcal{Y}$*. By the hypothesis this quantity converges to 0 for each* λ *and using compactness of* $\mathcal{X} \times \mathcal{Y}$*, this convergence is uniform. This automatically implies that the second integral on the right hand side of* ***Eqn 5.2.7*** *converges to 0 uniformly for all* λ *as* $n \to \infty$ *and this establishes the condition (1) of simultaneous uniform ergodicity in* ***Thm 4****.*

textbfCor 1 gives an easy technique of determining for a class of proposal distributions, the ergodic behavior of the adaptive TMCMC chain. However in most situations that we deal with in real life, the spaces $\mathcal{X}$ or $\mathcal{Y}$ are not compact. However the adaptation used by Haario *et al* [HT01] assumes the sample space to be compact and proposal kernels to be Multivariate normal centered at x, the current point and having variance covariance matrix λ. The way the algorithm is designed, guarantees that this parameter set is also closed and bounded and this implies by **Cor 1** that Haario *et al*'s adaptive chain for both RWMH and TMCMC are ergodic.

5.3 Some methods of adaptation in TMCMC

In this section, we introduce to some methods of adaptation in TMCMC, which are obtained as natural analogs to the adaptive methods in Random Walk Metropolis Hastings algorithms.The three methods we have used are

- Haario *et al*'s method (MCMC analog: see [HT01])
- SCAM algorithm (MCMC analog: see [HHT05])
- RAMA algorithm (MCMC analog: see [RR09])
- Adaptation by Stochastic approximation

Now we shall briefly describe the modifications to these methods introduced for application in case of the TMCMC chain.

5.3.1 Haario *et al*'s method (2001)

In this method, the sample space $\mathcal{X}$ is assumed to be compact. At each iteration, we update the proposal variance (assuming the proposal distribution is Gaussian with mean 0) such that the nth stage iterate of the parameter, say λ_n converges and condition (2) of Diminishing adaptation in **Thm** 4 holds. The adaptation rule in this case is

$$\begin{aligned} {\eta_n}^2 &= {\eta_0}^2 && n \leq N_0 && (5.3.1) \\ &= s_d var(\epsilon_1, \epsilon_2, \cdots, \epsilon_{n-1}) + s_d \delta I_d && n > N_0 && (5.3.2) \end{aligned}$$

where ${\eta_n}^2$ is the proposal variance at the nth stage and η_0 is a constant. Note that for $n > N_0$, the actual adaptation is taking place. s_d is a parameter depending on the dimension d. As a basic choice, we adopt the optimal value of scaling in TMCMC [DB13]. This optimal value is around $\frac{(2.42)^2}{d}$ which is very close to the value obtained for RWMH in [RGG97]. The choice of N_0 is flexible, but the bigger it is, the lesser will be the impact of adaptation on the chain. While computing the variance of the ϵ_i's in the 2nd term, we take only those values for which the move is accepted, or in other words, the ones that causes the chain to move.

As already discussed, if one takes the sample space to be compact, then Haario *et al*'s method satisfies the first condition of **Thm** 4. That the second condition holds is quite easily seen as the empirical variance term is a consistent estimator for the true proposal covariance.

Intuitively the above choice of adaptation can be explained in the following way. Suppose

our proposal density is a mixture of two normal distributions, one component of which has very high mixing proportion and converges in distribution to the Normal distribution with the optimal proposal variance and the other part is normal with constant variance (for a fixed dimension d), and with very low mixing proportion. Formally,

$$\begin{aligned} q_n(\epsilon) &= N\left(0, \frac{(0.1)^2}{d}\right) && n \leq 2d \quad (5.3.3) \\ &= \beta N\left(0, \frac{(0.1)^2}{d}\right) && (5.3.4) \\ &\quad +(1-\beta)N\left(0, s_d var\left(\epsilon_1, \epsilon_2, \cdots, \epsilon_{n-1}\right)\right) && n > 2d \quad (5.3.5) \end{aligned}$$

Note that the proposal variance for the above Mixture adaptation has a form analogous to the Haario *et al*'s method discussed above but with a particular choice of N_0 depending on the dimension d and of δ which is equal to $(0.1)^2$.Another method very similar in lines to Haario *et al*'s method is the SCAM algorithm.

5.3.2 SCAM Algorithm

SCAM algorithm for TMCMC is primarily again a modification of the basic approach suggested by Haario *et al*.In this case, we choose

$$\begin{aligned} {\eta_n}^2 &= 5^2 && n \leq 10 && (5.3.6) \\ &= s_d\left[var\left(\epsilon_1, \epsilon_2, \cdots, \epsilon_{n-1}\right) + 0.05\right] && n \geq 11 && (5.3.7) \end{aligned}$$

In the MCMC case, however the SCAM algorithm is quite different from that of the

TMCMC method, where for $n \geq 11$, when we update the variance of the jth co-ordinate, we use the variance of the jth co-ordinate of each of the sample points in the 2nd expression.So SCAM for MCMC updates the proposal variance for each co-ordinate separately. In TMCMC, since we are generating a single ϵ at each stage, we are not required to deal with different dimensions separately and thus our method looks very similar to that of the Haario *et al*'s method.

5.3.3 Regional Adaptive Metropolis Algorithm (RAMA)

In this method, we split the space $\mathcal{X}$ into several partitions $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_s$. The main abstraction underlying this approach is that we run different proposal densities over different regions, but aim to make the acceptance probability close to the optimal value of 0.439 for non-adaptive TMCMC (see**Chapter 3** and **Chapter 4**). For each of the s regions, we assign a particular value ${a_n}^k$ for the kth region and nth iteration, such that if our current location x belongs to the region k, then we generate ϵ from a proposal density $N(0, exp({a_n}^k))$ for the nth iterate. As an example, for a 2-partition RAMA, our usual choice of the proposal density can be

$$q_n(\epsilon) = N\left(0,\ exp(a_n)\ \mathbb{I}_{||x||<d} + exp(b_n)\ \mathbb{I}_{||x||>d}\right) \tag{5.3.8}$$

We do not update the coefficients ${a_n}^k$'s every iterate. But we fix a number of iterates M, such that every M iterations onwards, these coefficients are updated by a small amount δ_n. After each batch of M iterations, we check the number of acceptances in that particular batch and from each region. If for region k, it is found to be less than 0.439, then we decrease the coefficient by δ_n, otherwise we increase it by the same amount. We use Roberts and Rosenthal's choice of δ_n [RR09] given by $\delta_n = min(0.01, \frac{1}{\sqrt{n}})$. The initial choice of coefficients is done by randomly generating from a $\mathcal{U}(-2, 2)$ distribution

so as to cover a broad spectrum of starting values of the coefficients in constructing this adaptive chain.

5.3.4 Adaptation by Stochastic Approximation

The final method of adaptation we discuss for TMCMC uses the concept of Stochastic approximation, introduced by Robbins and Monro [RM51], where we recursively update a parameter sequence $\{\theta_n\}$ in the following way

$$\theta_{n+1} = \theta_n + c_n(h(\theta_n) + \epsilon_{n+1}) \tag{5.3.9}$$

where h represents a function that is estimated with the noise $\{\epsilon_n\}$. If the noise process dies out, then $\{\theta_n\}$ will converge to the solution of $h(\theta) = 0$ under appropriate conditions. For our adaptation, we take c_n to be a sequence of real numbers lying between $\frac{1}{n}$ and $\frac{1}{\sqrt{n}}$.Such a choice ensures the result in **Eqn 5.3.10**.

$$\sum_{n=0}^{\infty} c_n = \infty \quad \sum_{n=0}^{\infty} {c_n}^2 < \infty \tag{5.3.10}$$

We proceed as per the TMCMC algorithm, but at each iterate, we update the proposal standard deviation η_n as in **Eqn 5.3.11**.

$$\eta_{n+1} = max\left\{\eta_n + c_n\left(\alpha(X_n, \epsilon_n) - u\right), 0\right\} \tag{5.3.11}$$

where $\alpha(X_n, \epsilon_n)$ represents the acceptance rate value at the nth iterate and u is the optimal scaling value (0.439) for TMCMC algorithm. We update the proposal variance irrespective of whether the move has been accepted or rejected. Note that if the expected acceptance rate $M(\eta_n) = E\left[\alpha(X_n, \epsilon_n)\right]$ is a decreasing function of the choice of η_n (which

is quite intuitively justified) and $M^{'}(\eta_{opt})$ for the optimal proposal variance η_{opt} exists and positive, then the algorithm will converge to a chain having optimal acceptance rate of 0.439.

5.4 Concluding remarks

In this chapter, we developed the theory for adaptive versions of TMCMC and then defined several adaptation mechanisms. It would be curious to investigate the performance of these approaches to the RWMH analogs. For this purpose, we need to define some measure of performance that will aid us in this process. In the next chapter, we shall consider Simulation experiments and much of our focus there would be on these adaptive measures and their relative performance with respect to RWMH adaptations.

Chapter 6

Simulation experiments – Case Studies

So far, we have mostly indulged in developing the theory of our additive TMCMC model. However, the theoretical results must be validated through applications and experiments. This is what we seek to achieve now. In this chapter, we shall consider some simulation experiments over varying dimensions and proposal densities with the aim of gaining new insight on our methods and to compare our method with the RWMH technique purely from an experimental point of view. We first determine some performance evaluation measures that will aid us in our comparison of the two processes, RWMH and TMCMC for non-adaptive as well as the adaptive versions.

6.1 Performance evaluation of Adaptive TMCMC with respect to Adaptive MCMC

In this section, we present a broad comparative analysis of the various adaptive versions of RWMH and TMCMC. We use three types of performance evaluation techniques.

- **Acceptance Rate**: A principal reason of using adaptive versions and seeking optimal scaling is to ensure that the acceptance rate is maximized meaning that our chain moves fast in space with respect to time. This is crucial because

otherwise if one selects a sample out of this process, most of the observations may be identical and this would reduce the effective sample size of the simulation.

- **Average Jump Size (AJS)**: We record each jump length, the distance between between X_{n+1} and X_n at each $n \geq N_B$, where N_B is the burn-in time, and average over all these jumps. A higher value of jump size will mean higher variation in the simulated observations after burn-in.

- **Integrated Autocorrelation Time (IACT) / Integrated partial autocorrelation time (IPACT)**: After burn-in, we compute the autocorrelation function of the chain upto a sufficiently high order. Then we define IACT as in **Eqn 6.1.1**.

$$IACT_N = \left[1 + 2\sum_{t=1}^{N-1}\left(1 - \frac{t}{N}\right)\hat{\rho}(t)\right] \tag{6.1.1}$$

where $\hat{\rho}(t)$ is the sample autocorrelation function of order t and N is a predetermined maximum order selected for integration.

As $N \to \infty$, then this the expression in **Eqn 6.1.1** reduces to the following.

$$IACT = 1 + 2\sum_{t=1}^{\infty}\hat{\rho}(t) \tag{6.1.2}$$

IACT actually manages to pool the effect of autocorrelation of various orders together to look in some sense at the aggregate dependence of the series on the past observations. A higher value of IACT would mean the chain would show less amount of variation and the moves may be restricted.

In the same way, we computed also $IPACT$- the integrated partial autocorrelation time where we replace the ACF's in the expression in **Eqn 6.1.1** by the PACFs. The rest is analogous to $IACT$.

- **Kolmogorov Smirnov Test Statistic (KST)**: Kolmogorov-Smirnov test is a very standard test for equality of two distributions. In this case we try to observe how closely the Adaptive MCMC and Adaptive TMCMC chains approximate the actual target distribution π. We run a number of chains, say L, starting from one fixed point for both MCMC and TMCMC adaptations. Corresponding to each time point t, we shall thus get L many iterates. The notion is that as time t increases (specially after burn-in), these L many iterates should be close to an independently drawn random sample from the target distribution π. So, if we look at the KST statistic for the empirical distribution of these iterates along a particular dimension with respect to the marginal of π along that dimension, we should find the test statistic decreasing with time and finally being very close to 0 after a certain time point. What we try to observe is how for a particular adaptation, the KST statistic for MCMC and TMCMC analogs behave and how closely it approximates π after burn-in. The main interest would lie in observing if one method has a uniformly better KST statistic value compared to the other.

All these measures highlight different aspects of performance (the acceptance rate, the variation of state in the chain after burn-in, convergence to π etc.). In **Tab 6.2** and **Tab 6.3**, we present the results of a comprehensive simulation study for different dimensions and change of starting proposal variance, that would help compare the performance of various adaptive procedures in both MCMC and TMCMC set up, corresponding to each of the aforementioned evaluation measures.

Now we test for simulation experiments on a wide range of dimensions and proposal densities.

6.2 Basic simulation of non-adaptive methods

First we compare the performance of RWMH and TMCMC for corresponding to 3 different choices of proposal variance, with scalings l being the optimal one 2.42 (TMCMC), 6 and 10 respectively. We considered data of varying dimensions ranging from 2 to 200 in order to get hold of a broader picture. For our data analysis we considered the target density π to be a $MVN(0, I)$ distribution and the starting point x_0 to be randonmly generated from $U(-2, 2)$ distribution. The proposal density was also taken to be Gaussian (with independent co-ordinates as for RWMH) having mean 0 and variance $\frac{l^2}{d}$ for each co-ordinate, where l is the value of the scaling constant. In each run, the chain was observed upto 1,00,000 trials (accepted or rejected). The choice of burn-in was made a bit subjectively, removing one fourth of the total number of iterates initially and it was found via simulation study that was quite sufficient- both RWMH and TMCMC seemed to reach burn-in well before that time, but still we were more conservative. 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 and then summed over as in **Eqn 6.1.1**. $IPACT$ was similarly computed. For computing the KST, we repeated the experiment with the same starting point and keeping all other factors same, 100 times and then averaged over all such iterates along all the co-ordinates. **Table 6.2** presents the detailed results of the simulation experiment. The sample path plots and the decreasing trend in KST over time for both RWMH and TMCMcorresponding to different scalings and idmensions are also presented.

6.3 Basic simulation of adaptive methods

For the adaptive methods, it was found that the choice of the initial value of the proposal density (corresponding to the ones where it is not specified) did not significantly change the results if the proposal variance is taken more or less close to the target density variance. The basic mechanism of simulation for these adaptive methods is the same as in the non-adaptive set up. In RAMA 2, we considered the two partition of the data space as $||x|| < 2$ and $||x|| > 2$, while in RAMA 3 , we considered the 3-fold partition as $||x|| < 2$, $2 < ||x|| < 5$ and $||x|| > 5$. For the RAMA procedures, we needed to update the parameters after running the process for some time. For the simulation experiments, for each of coefficients, we ran the process for 100 times, before updating the coefficients. Overall, we used 1000 many updates implying that the total length of the chain was of 1,00,000 runs. We used Haario *et al*'s method proposed in 2001 [HT01] with $N_0 = 2$ and $\delta = 0.10$ in textbfEqn 5.3.1 and $\eta_0 = 1$. The results were compared to the more refined approach suggested in Haario *et al* in 2005 [HHT05]. For the stochastic approximation procedure, we took c_n to be $\frac{1}{n}$ as it satisfies the desired properties of convergence. **Table 6.3** presents the detailed results of the simulation experiment. The sample path plots and the decreasing trend in KST over time for both RWMH and TMCMC corresponding to different scalings and dimensions are also presented.

6.4 Observations and Concluding remarks

We first summarize the main observations from the graphs and the tables corresponding to these simulation experiment.

- As anticipated, TMCMC seems to have a uniformly better acceptance rate than RWMH for all dimensions and all choices of proposal variances. There is sufficient

gain in acceptance rate over RWMH even for 2 dimensions and the difference broadens once we move to higher dimensions or consider higher proposal variances. That high proposal variance would affect the performance of RWMH is quite intuitive as getting an outlying observation in any of the d co-ordinates becomes more likely and that may affect the acceptance rate. Since we update by a single variable in TMCMC, the chances are less compared to RWMH.

- For optimal scaling as theoretically proved, TMCMC has a higher acceptance rate of 0.439 corresponding to 0.234- the optimal acceptance rate for the RWMH algorithm. An interesting observation from **Table 6.2** is that even from dimension 2, our acceptance ratio corresponding to the optimal scaling of 2.4 is very close to 0.44. This is not the case for RWMH. For very high dimensions (100 and 200), the optimal acceptance rate is attained more or less for both TMCMC and RWMH but our optimal acceptance ratio seems to be far more robust across dimensions (specially the lower dimensions) compared to RWMH.

- Another form of robustness is achieved over RWMH by changing the scale from 2.4 by a small amount, say 6. We witness much less significant drop in the acceptance rate under this change of scale than RWMH which often falls pretty badly and becomes almost negligible.

- Despite this inherent lack of randomness due to deterministic transfprmation in TMCMC, we find that in terms of convergence, we are almost as good as RWMH for smaller proposal variances and lower dimensions an dseem to have a slight edge over these methods for higher dimensions ans higher proposals. This is mainly because, for RWMH in higher dimensions, the effective sample size reduces greatly and so, the chain remains static for a good length of time and this compromises with the convergence.

- The measures $IACT$, $IPACT$ are extremely close, RWMH has a little less value compared to TMCMC in lower dimensions but in higher dimensions, TMCMC performs better than RWMH. The same is true for the average jump size (AJS) as well. For TMCMC AJS value is not very high but seems to be quite consistent. For RWMH, the AJS value is more in lower dimensions, implying that it has better variation than TMCMC after burn-in, but for high dimensions, the AJS value drops significantly as the chain does not move adequately and very few jumps are significant.

- The adaptive versions of TMCMC perform really well and unlike in case of RAMA 2 or RAMA 3, where RWMH does not quite reach the optimal value of acceptance rate for 10 dimenions even after 1,00,000 runs, TMCMC is quite close to the optimal value of 0.439 for dimension 10. In fact for dimension 100 in RAMA 3, we witnessed a significant drop in acceptance, but TMCMC stays robust. However, Haario's method does not seem to be very consistent both for RWMH and TMCMC. Atchade's method of stochastic approximation stands out as the best for both RWMH and TMCMC.

The overall assessment from our point of view is that TMCMC has less computational and time complexity (which is very significant for higher dimensions) than RWMH and also a much better acceptance rate for any dimension and proposal over RWMH. In addition, it has nice convergence properties- it is almost geometrically ergodic for a class of sub-exponential distributions (**Chapter 2**) and as observed from simulation experiments, the convergence is at par with MCMC despite the deterministic approach for lower dimensions, and actually better for higher dimensions. Also, the optimal acceptance rate is way higher than RWMH and is much more stable across dimensions and robust across scale changes. Finally, we find that the adaptive versions of TMCMC show better and more stable convergence to optimal acceptance rate than RWMH. All

these points very strongly prescribe that one should prefer to use TMCMC over RWMH in any dimension and for any proposal.

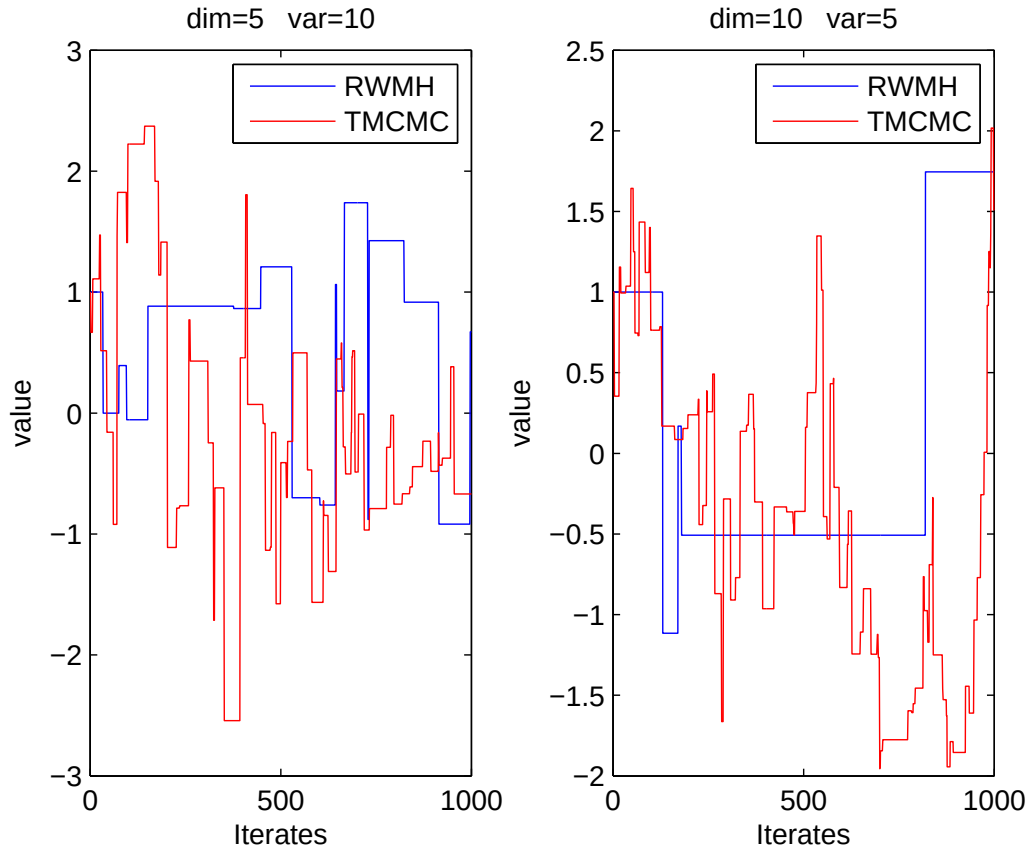


Fig. 6.4.1: Sample paths of RWMH and TMCMC paths for two cases: (left) dimension=5 and proposal variance=10 (aong each co-ordinate for RWMH) (right) dimension=10 proposal variance=5 (along each co-ordinate for RWMH), proposal distribution normal for TMCMC and independent normal in each co-ordinate for RWMH and centered at 0.

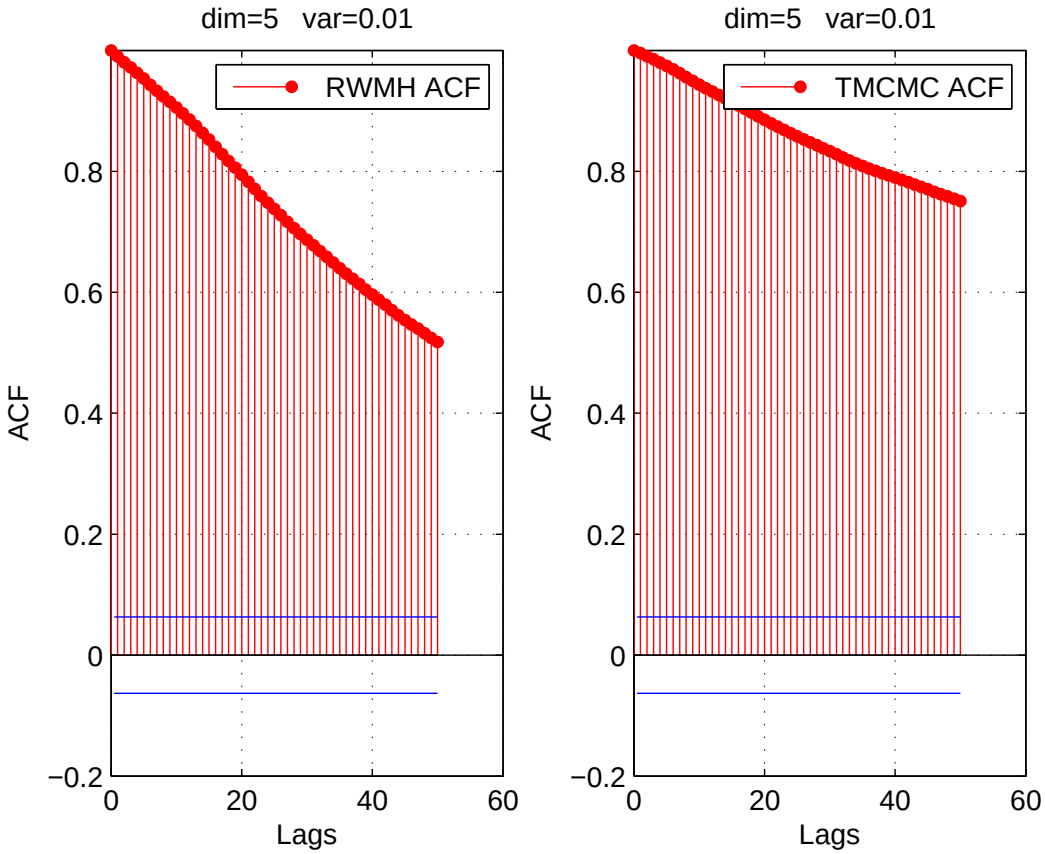


Fig. 6.4.2: acf plot of sample paths for RWMH and TMCMC for dimension=5 and proposal distribution normal for TMCMC and independent normal for RWMH along each co-ordinate, with center 0 and variance 0.01 (along each co-ordinate for RWMH).

Table 6.2.1: The performance evaluation of RWMH and TMCMC chains for varying dimensions. It is assumed that proposal has independent normal components for RWMH with same proposal variance along all co-ordinates. The proposal scales range from optimal (2.4) to 10. All calculations done after burn in

Dimension	Scaling \ Test	*Acceptance rate*(%)		*IACT*		*IPACT*		*AJS*		*Avg. K-S test*	
		RWMH	TMCMC	RWMH	TMCMC	RWMH	TMCMC	RWMH	TMCMC	RWMH	TMCMC
2	2.4	34.9	44.6	6.08	7.04	2.46	2.55	0.93	0.74	0.1651	0.1657
	6	18.66	29.15	7.08	8.08	2.52	2.56	0.79	0.62	0.1659	0.1655
	10	3.83	12.36	13.74	11.91	2.83	2.77	0.22	0.32	0.1676	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	2.77	20.20	15.6	14.11	2.77	2.81	0.39	0.48	0.1693	0.1674
	10	0.45	12.44	18.26	15.61	2.88	2.83	0.61	0.42	0.1697	0.1678
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	1.37	20.34	17.55	16.31	2.91	2.86	0.25	0.49	0.1800	0.1688
	10	0.03	7.94	18.79	15.25	2.83	2.71	0.08	0.14	0.1748	0.1674
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	0.32	20.6	18.62	18.25	2.89	2.88	1.05	0.69	0.1787	0.1684
200	2.4 (opt)	23.4	44.2	18.4	18.67	2.88	2.89	1.3	9.22	0.1813	0.1735
	6	0.33	20.7	18.86	18.74	2.89	2.89	0.09	0.54	0.1832	0.1755

Table 6.3.1: The performance evaluation of various adaptive versions of RWMH and TMCMC chains for varying dimensions. It is assumed that proposal has independent normal components for RWMH with same proposal variance along all co-ordinates.

Dimension	Test / Method	*Acceptance rate*(%)		*IACT*		*IPACT*		*AJS*		*Avg. K-S test*	
		RWMH	TMCMC	RWMH	TMCMC	RWMH	TMCMC	RWMH	TMCMC	RWMH	TMCMC
2	Haario	77.3	63.9	11.94	9.35	2.74	2.70	0.29	0.55	0.1684	0.1774
	SCAM	21.8	41.2	6.21	6.59	2.46	2.37	0.77	0.77	0.1772	0.1752
	RAMA-2	40.2	49.42	7.03	7.27	2.54	2.57	0.87	0.69	0.2104	0.2111
	RAMA-3	36.6	50.6	6.54	7.68	2.44	2.57	0.81	0.72	0.2136	0.2147
	Atchade	36.5	48.03	6.01	8.56	2.45	2.65	0.93	0.61	0.2106	0.2119
10	Haario	45.37	74.9	14.06	16.92	2.81	2.86	1.02	0.45	0.1667	0.1788
	SCAM	4	22.3	18.37	16.01	2.94	2.84	0.09	0.54	0.1795	0.1782
	RAMA-2	34.7	40.06	13.99	15.23	2.81	2.82	1.00	0.67	0.2104	0.2114
	RAMA-3	17.3	39.4	17.00	16.16	3.11	2.90	0.57	0.61	0.2125	0.2138
	Atchade	22.4	42.2	13.21	15.25	2.76	2.82	1.21	0.77	0.2105	0.2137
100	Haario	17.93	76.94	17.01	18.68	2.89	2.88	0.07	0.33	0.185	0.1779
	SCAM	0.62	11.32	17.00	16.24	2.88	2.86	0.06	0.48	0.1912	0.1976
	RAMA 2	20.19	42.1	18.28	18.43	2.88	2.89	0.69	0.75	0.2135	0.2122
	RAMA 3	6	40.09	18.83	18.44	2.89	2.88	0.03	0.84	0.2154	0.2141
	Atchade	23.1	43.9	18.13	18.44	2.88	2.88	1.27	0.75	0.2120	0.2119

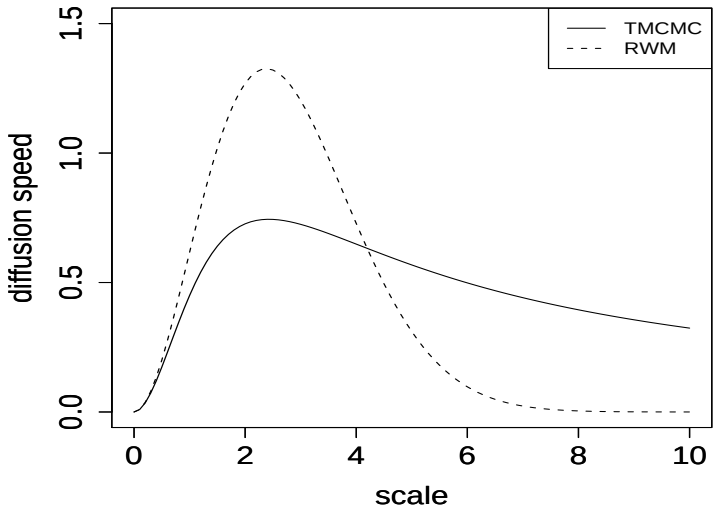


Fig. 6.4.3: Comparison of diffusion speeds of TMCMC and RWM in the *iid* case.

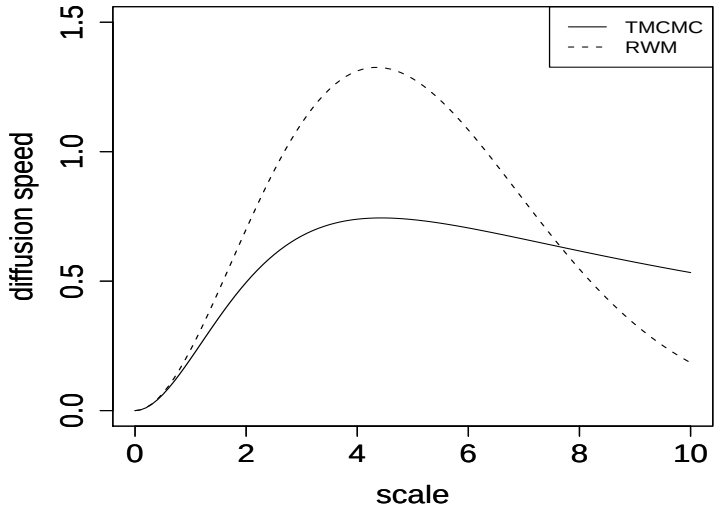


Fig. 6.4.4: Comparison of diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the *iid* case, with $c = 0.3$.

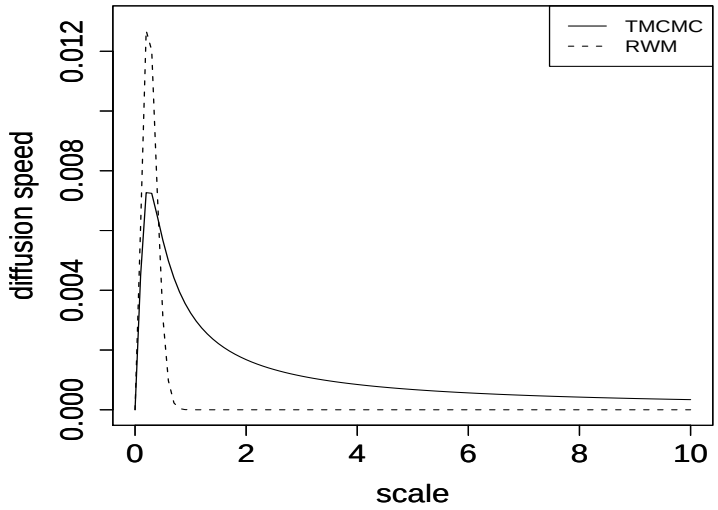


Fig. 6.4.5: Comparison of diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the independent but non-identical case, with $\xi = 10$.

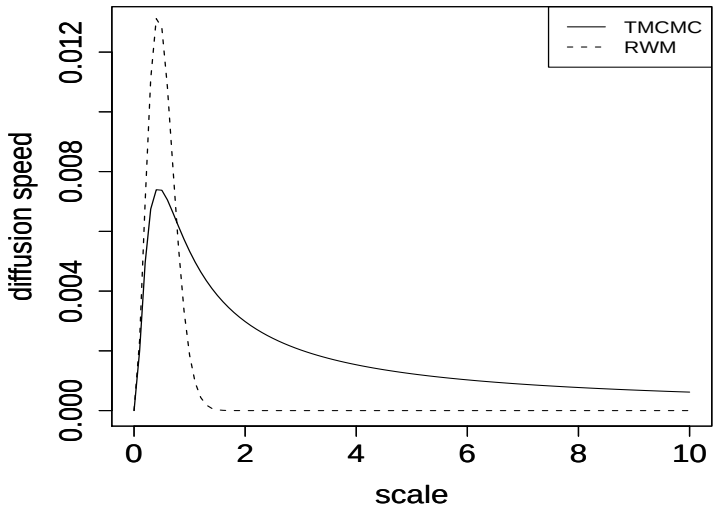


Fig. 6.4.6: Comparison of diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the independent but non-identical case, with $\xi = 10$, $c = 0.3$.

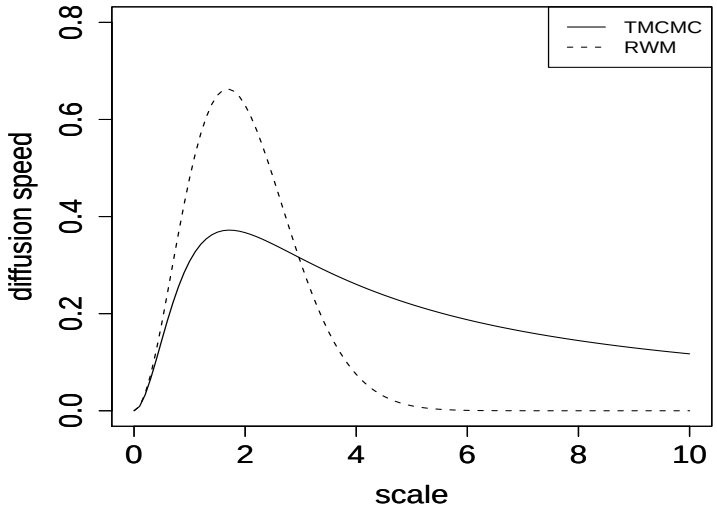


Fig. 6.4.7: Comparison of diffusion speeds of TMCMC and RWM in the dependent case.

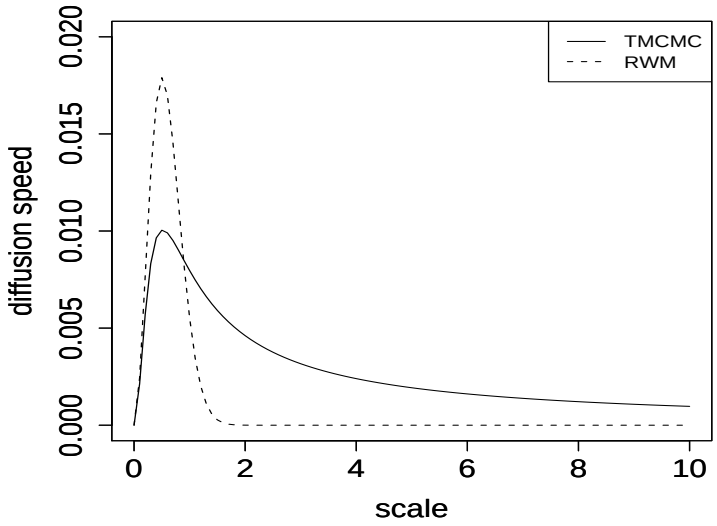
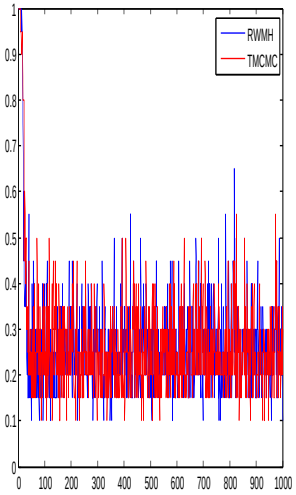
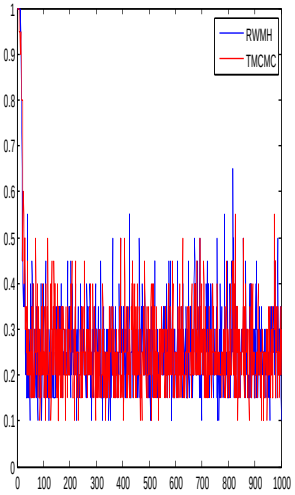


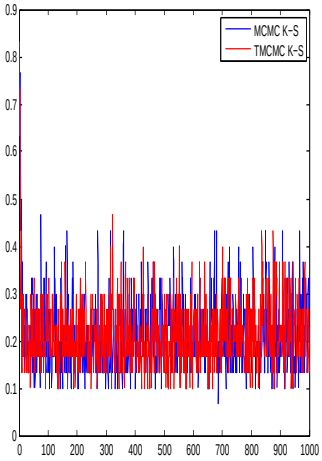
Fig. 6.4.8: Comparison of diffusion speeds of TMCMC within Gibbs and RWM within Gibbs in the dependent case, with $c = 0.3$.



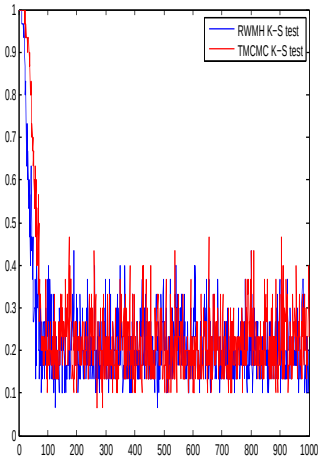
(a) GENERAL



(b) SCAM



(c) RAMA 2



(d) ATCHADE

Fig. 6.4.9: K-S test comparison across the various time points for four methods starting from top left to bottom right a) General Non-adaptive method b) SCAM algorithm by Haario *et al* [HHT05] c) RAMA algorithm with 2 partitions and d) Atchade's method. The data is two dimensional and starting point is $(1, 1)$

:

Bibliography

[AA08] K.M. Aludaat and M.T. Alodat. A Note on Approximating the Normal Distribution Function. *Applied Mathematical Sciences*, pages 425–429, 2008. 37

[Bed07] M. Bedard. Weak Convergence OF Metropolis Algorithms For Non-i.i.d. Target Distributions. *The Annals of Applied Probability*, pages 1222–1244, 2007. 28, 41

[Bed08] M. Bedard. Optimal Acceptance Rates for Metropolis Algorithms: Moving Beyond 0.234. *Stochastic Process. Appl.*, pages 2198–222, 2008. 41, 44

[Bed09] M. Bedard. On the optimal scaling problem of metropolis algorithms for hierarchical target distributions. *preprint*, 2009. 28, 47

[BR08] M. Bedard and J.S. Rosenthal. Optimal Scaling of Metropolis Algorithms: Heading Toward General Target Distributions. *Canad. J. Statist.*, pages 483–503, 2008. 41

[BRS09] A. Beskos, G.O. Roberts, and A.M Stuart. Optimal scalings for local Metropolis-Hastings chains on non-product targets in high dimensions. *The Annals of Applied Probability*, pages 863–898, 2009. 28, 47

[BS08] A. Beskos and A.M Stuart. Mcmc methods for sampling function space. *R. Jeltsch and G. Wanner, editors, ICIAM Invited Lecture 2007. European Mathematical Society*, 2008. 47

[DB11] S Dutta and S Bhattacharya. Markov Chain Monte Carlo Based on Deterministic Transformations. *arXiv:1106.5850*, 2011. 2, 5, 6, 59

[DB13] K.K. Dey and S Bhattacharya. On Optimal scaling of Non-adaptive Additive Transformation based Markov Chain Monte Carlo. 2013. 65

[Dut10] S Dutta. Multiplicative random walk Metropolis-Hastings on the real line. *arXiv:1008.5227*, 2010. 3, 5, 16

[GS90] A.E. Gelfand and A.F.M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, pages 398–409, 1990. 1

[GS96] Richardson S. Gilks, W. R. and D. J. Spiegelhalter. Markov chain Monte Carlo in practice. *Interdisciplinary Statistics, Chapman & Hall, London.*, 1996. 1

[Has70] W.K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, pages 97–109, 1970. 1

[HHT05] E. Saksman H. Haario and J. Tamminen. Componentwise adaptation for high dimensional MCMC. *Comput. Stat.*, pages 265–274, 2005. viii, 58, 64, 74, 86

[HT01] Saksman E. Haario, H. and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, pages 223–242, 2001. 58, 64, 74

[JH00] S.F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process.Appl.*, pages 341–361, 2000. 10, 21, 22

[JR02] S.F. Jarner and G.O. Roberts. Polynomial Convergence rates of Markov Chains. *The Annals of Applied Probability*, page 224â247, 2002. 10

[MPS11] J.C. Mattingly, N.S. Pillai, and A.M Stuart. Diffusion Limits of the Random Walk Metropolis Algorithm in High Dimensions. *The Annals of Applied Probability*, 2011. 47, 48, 49, 50, 55, 57

[MRR53] N Metropolis, A.W. Rosenbluth, and A.H. Rosenbluth, M.N.and Teller. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, pages 1087–1092, 1953. 1

[MT93] S.P. Meyn and R.L. Tweedie. Markov chains and stochastic stability. 1993. 5, 12, 59

[MT96] K.L. Mengersen and R.L. Tweedie. Rates of Convergence of the Hastings and Metropolis Algorithms. *The Annals of Statistics*, pages 101–121, 1996. 10

[NR06] P. Neal and G.O. Roberts. Optimal Scaling for Partially Updating MCMC Algorithms. *The Annals of Applied Probability*, pages 475–515, 2006. 28, 40

[RGG97] G.O. Roberts, A Gelman, and W.R Gilks. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The Annals of Applied Probability*, pages 110–120, 1997. 1, 28, 30, 44, 59, 65

[RM51] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat*, pages 400–407, 1951. 68

[RR07] G.O. Roberts and J.S. Rosenthal. Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms . *Journal of Applied Probability*, pages 454–478, 2007. 58, 61

[RR09] G.O. Roberts and J.S. Rosenthal. Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, pages 349–367, 2009. 58, 60, 64, 67

[RT96] G.O. Roberts and R.L. Tweedie. Geometric convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms. *Biometrika*, pages 95–110, 1996. 7, 10, 59

[Sko56] A.V. Skorohod. Limit theorems for stochastic processes. *Theor. Probability Appl.*, pages 261–290, 1956. 30

[Tie94] L Tierney. Markov chains for exploring posterior distributions. *Ann. Statist*, pages 1701–1762, 1994. 1