Optimal Acceptance Rates for Metropolis Algorithms: Moving Beyond 0.234

Mylène Bédard [*]

April 3, 2006

Abstract

Recent optimal scaling theory has produced a necessary and sufficient condition for the asymptotically optimal acceptance rate of random walk Metropolis algorithms to be the well-known 0.234 when applied to certain multidimensional target distributions with scaling terms possibly depending on the dimension. We show that when this condition is not met the limiting process of the algorithm is altered, yielding an asymptotically optimal acceptance rate which might drastically differ from the usual 0.234. In particular, we prove that as $d \to \infty$ the sequence of stochastic processes formed by say the i^*-th component of each Markov chain usually converges to a Langevin diffusion process with a distinct speed measure, except in particular cases where it converges to a one-dimensional Metropolis-Hastings algorithm with a singular acceptance rule. We also discuss the use of inhomogeneous proposals, which might reveal essential in specific cases.

1 Introduction

This paper contains theoretical results aiming to optimize the efficiency of random walk Metropolis (RWM) algorithms. This class of Metropolis-Hastings algorithms ([12], [11]) is easily implemented and widely used in various domains to generate data from highly complex probability distributions. The choice of a proposal density is however essential, and it is primordial to carefully choose the scaling parameter of this distribution in order to have some level of optimality in the performance of the algorithm (see [3], [4]).

[*] Department of Statistics, University of Toronto, Toronto, Ontario, Canada, M5S 3G3. Email: mylene@utstat.utoronto.ca

Roberts, Gelman and Gilks (1997) have been the first authors to publish theoretical results about the optimal scaling problem for RWM algorithms with Gaussian proposals (see [15]). They established that for d-dimensional target distributions with *iid* components, the asymptotic acceptance rate optimizing the efficiency of the process is 0.234 independently of the target density. These findings have had a deep impact in the MCMC literature as they provide an exceptionally simple guideline to be applied by all skilled-level practitioners. Afterwards, Roberts and Rosenthal (1998) carried out a similar study for Metropolis adapted Langevin algorithms (MALA), whose proposal distribution makes use of the gradient of the target density to generate wiser moves (see [16]). They obtained analogous conclusions but with an asymptotically optimal acceptance rate (AOAR) of 0.574. In spite of the *iid* assumption for the target components, in both cases the results are believed to be far more robust and to hold under various perturbations of the target distribution. Since their publication these papers have spawned much interest, and consequently other authors have studied diverse extensions of the *iid* model, which all corroborate the results found in [15] and [16] (see for instance [6], [7], [8], [13] and [17]).

We recently considered a model involving d-dimensional target distributions with independent components, where the scaling term of each component is allowed to depend on the dimension of the target distribution. This setting includes many popular statistical models, but it produces distributions having unstable scaling terms (since some of them might converge to 0 or ∞), which creates a significant distinction with *iid* target distributions. In [1], we provided a necessary and sufficient condition on the scaling terms under which the algorithm admits the same limiting diffusion process and the same AOAR as those found in [15].

In this paper, we prove that the acceptance rate 0.234 is not always asymptotically optimal and most importantly, we present methods for determining the correct AOAR when it is not. This addresses the issue raised in Open Problem #3 of [18]. In particular, we show that when there exists a finite number of scaling terms converging significantly faster than the others, then the asymptotic acceptance rate optimizing the efficiency of the algorithm depends on specific target components only and is smaller than 0.234. To this end, we first determine the limiting distribution of the sequence of stochastic processes formed by say the i^*-th component of each Markov chain, which is done by proving $\mathcal{L}^1$ convergence of the processes' generators. For the smallest order components we obtain a limiting RWM algorithm whose singular acceptance rule constitutes a new rule in the Metropolis-Hastings family. Since there exist multiple measures of efficiency for discrete-time processes, we choose not to use these components to determine the AOAR. We rather optimize the algorithm through the speed measure of the limiting Langevin diffusion process obtained from the other components. Speed measures being the sole measure of efficiency for diffusions, we realize that the AOAR is not 0.234 anymore since the speed measure now differs from that obtained in [15] and [1]. We shall also see that when the smallest order components are remote from the other components, the optimization problem is ill-posed which calls for the use of inhomogeneous proposal scalings.

The paper is divided as follows. We begin by introducing the target, that is the distribution we wish to sample. In Section 3, we then present the algorithm used to this end along with a method for determining the optimal form for the proposal scaling. The main optimal scaling results are presented in Section 4, while the particular case where the target is normally distributed is discussed in Section 5. Inhomogeneous proposals are then considered in Section 6, along with the optimal scaling problem under various extensions of the target distribution. The results are proved in Section 7, making use of the different lemmas proved in Sections 8 to 10. We conclude the paper with a discussion in Section 11.

2 The Target Distribution

Suppose we are interested in generating data from the following d-dimensional product density

$$\pi\left(d, \mathbf{x}^{(d)}\right) = \prod_{j=1}^{d} \theta_j\left(d\right) f\left(\theta_j\left(d\right) x_j\right), \tag{1}$$

where $\theta_j^{-2}\left(d\right)$, $j = 1, \ldots, d$ are the scaling terms of the target distribution. Although independent, the d components are not identically distributed as we let the scaling terms be such that

$$\boldsymbol{\Theta}^{-2}\left(d\right) = \left(\frac{K_1}{d^{\lambda_1}}, \ldots, \frac{K_n}{d^{\lambda_n}}, \underbrace{\frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+1}}{d^{\gamma_1}}}_{c(\mathcal{J}(1,d))}, \ldots, \underbrace{\frac{K_{n+m}}{d^{\gamma_m}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}}_{c(\mathcal{J}(m,d))}\right), \tag{2}$$

where $0 \leq n < \infty$, $1 \leq m < \infty$ and $\{K_j, j = 1, \ldots, n+m\}$ are some positive and finite constant terms. Note that the assumption of equality among the constants for a given γ_i, $i = 1, \ldots, m$ shall be relaxed in Section 6.

We finally assume that the density f satisfies some regularity conditions. In particular, we suppose that f is a positive C^2 function, $(\log f\left(X\right))'$ is Lipschitz continuous,

$$\mathrm{E}\left[\left(\frac{f'\left(X\right)}{f\left(X\right)}\right)^4\right] = \int_{\mathbf{R}} \left(\frac{f'\left(x\right)}{f\left(x\right)}\right)^4 f\left(x\right) dx < \infty$$

and

$$\mathrm{E}\left[\left(\frac{f''\left(X\right)}{f\left(X\right)}\right)^2\right] = \int_{\mathbf{R}} \left(\frac{f''\left(x\right)}{f\left(x\right)}\right)^2 f\left(x\right) dx < \infty.$$

Having defined the target distribution, we now describe in more details the form adopted by $\boldsymbol{\Theta}^{-2}\left(d\right)$ in (2). The first particularity to notice is that the first n terms appear only once each, while the balance is repeated according to some functions of d. That is, the last $d-n$ terms are separated into m different groups, in each of which the number of terms grows with the dimension. Ultimately, we shall be interested in studying the limiting distribution of every individual component forming the RWM algorithm as $d \to \infty$.

In order to study each of the $n+m$ different components, a rearrangement of the scaling terms in (2) is necessary. This will avoid referring to a component located at an infinite position as $d \to \infty$. We thus instead let

$$\mathbf{\Theta}^{-2}(d) = \left(\frac{K_1}{d^{\lambda_1}}, \ldots, \frac{K_n}{d^{\lambda_n}}, \frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}, \frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}, \ldots, \frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}\right). \quad (3)$$

That is, we first enumerate each one of the $n+m$ different scaling terms. Afterwards, we cycle through the remaining ones, i.e. the scaling terms that will appear an infinite number of times in the limit. The m groups to which they belong might however occupy different proportions of the vector $\mathbf{\Theta}^{-2}(d)$, and we should make sure to preserve the proportion in which they appear when cycling through them.

An simple way to identify the different groups of components whose scaling terms appear infinitely often in the limit is to introduce the mutually exclusive sets

$$\mathcal{J}(i,d) = \left\{ j \in \{1, \ldots, d\}; \theta_j^{-2}(d) = \frac{K_{i+n}}{d^{\gamma_i}} \right\}, \quad i = 1, \ldots, m.$$

All positions of components with a scaling term equal to K_{n+i}/d^{γ_i} thus belong to the i-th set $\mathcal{J}(i,d)$. The union of these m sets satisfies $\dot{\bigcup}_{i=1}^{m} \mathcal{J}(i,d) = \{n+1, \ldots, d\}$. They also provide an alternative way of expressing the d-dimensional product density in (1)

$$\pi\left(d, \mathbf{x}^{(d)}\right) = \prod_{j=1}^{n} \left(\frac{d^{\lambda_j}}{K_j}\right)^{1/2} f\left(\left(\frac{d^{\lambda_j}}{K_j}\right)^{1/2} x_j\right) \prod_{i=1}^{m} \prod_{j \in \mathcal{J}(i,d)} \left(\frac{d^{\gamma_i}}{K_{n+i}}\right)^{1/2} f\left(\left(\frac{d^{\gamma_i}}{K_{n+i}}\right)^{1/2} x_j\right).$$

Because the positions sets do not necessarily possess the same number of elements, we introduce corresponding cardinality functions. For $i = 1, \ldots, m$,

$$c(\mathcal{J}(i,d)) = \# \left\{ j \in \{1, \ldots, d\}; \theta_j^{-2}(d) = \frac{K_{n+i}}{d^{\gamma_i}} \right\}, \quad (4)$$

where $c(\mathcal{J}(i,d))$ is some polynomial function of the dimension satisfying $\lim_{d\to\infty} c(\mathcal{J}(i,d)) = \infty$ and subject to the constraint $\sum_{i=1}^{m} c(\mathcal{J}(i,d)) = d - n$.

Without loss of generality, we assume the scaling terms to be arranged according to an asymptotic increasing order within both groups formed by the first n and last $d-n$ scaling terms respectively. Letting $\preceq$ stand for "is asymptotically smaller than", we obtain $\theta_1^{-2}(d) \preceq \ldots \preceq \theta_n^{-2}(d)$ and similarly $\theta_{n+1}^{-2}(d) \preceq \ldots \preceq \theta_d^{-2}(d)$, which are equivalent to $-\infty < \lambda_n \le \lambda_{n-1} \le \ldots \le \lambda_1 < \infty$ and $-\infty < \gamma_m \le \gamma_{m-1} \le \ldots \le \gamma_1 < \infty$. For components with the same power of d, we must consequently refer to their constant to determine which scaling term is smaller. We finally point out that some of the first n components might have exactly the same scaling term. When this happens, we still refer to them as say K_j/d^{λ_j} and $K_{j+1}/d^{\lambda_{j+1}}$, with $K_j = K_{j+1}$ and $\lambda_j = \lambda_{j+1}$.

The objective of this paper is to study the limiting distribution of the first $n+m$ components of the RWM process applied to target distributions as just described. In [1], we provided a necessary and sufficient condition on the scaling terms of the target ensuring that the AOAR of the RWM algoritm is 0.234. We consider the same target distributions, but we now focus on the case where that condition is not satisfied. Specifically, we consider the case where there exists a finite number of target components associated with scaling terms significantly smaller than the others. This can be translated mathematically by

$$\lim_{d\to\infty} \frac{\theta_1^2(d)}{\sum_{j=1}^{d} \theta_j^2(d)} > 0, \tag{5}$$

which is the complement of Condition 8 in [1]. According to the form of the target, the asymtotically smallest scaling term in (3) would normally have to be either $\theta_1^{-2}(d)$ or $\theta_{n+1}^{-2}(d)$. However, it is interesting to notice that under the fulfilment of the previous condition this uncertainty is resolved and K_1/d^{λ_1} is smallest for large d. Furthermore, the existence of other target components having a $O\left(d^{\lambda_1}\right)$ scaling term is also possible. In particular, let $b = \max\left(j \in \{1,\dots,n\}; \lambda_j = \lambda_1\right)$; b is then the number of such components, which is finite and at most n.

To avoid getting a trivial limiting process when studying one particular component, it is necessary for its scaling term to be independent of d. In particular we set it equal to 1, which can be done without loss of generality by applying a linear transformation to the target distribution.

3 Algorithm and Proposal Scaling

To generate a sample from the target distribution described in Section 2 we apply a RWM algorithm, which consists in building a d-dimensional Markov chain $\mathbf{X}^{(d)}(0), \mathbf{X}^{(d)}(1), \dots$ having the target distribution as its stationary distribution. We choose the d-dimensional proposal to be normally distributed with independent components centered around the current state of the chain, i.e. $\mathbf{Y}^{(d)}(t+1) \sim N\left(\mathbf{X}^{(d)}(t), \sigma^2(d) I_{d\times d}\right)$ with density

$$q\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) = \left(2\pi\sigma^2(d)\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2(d)} \sum_{j=1}^{d} (y_i - x_i)^2\right).$$

The construction of the Markov chain is thus carried using the following steps. Given $\mathbf{X}^{(d)}(t)$, the state of the chain at time t, a value $\mathbf{Y}^{(d)}(t+1)$ is generated from $q\left(d, \mathbf{X}^{(d)}(t), \mathbf{y}^{(d)}\right)$. The probability of accepting the proposed value $\mathbf{Y}^{(d)}(t+1)$ as the new value for the chain is $\alpha\left(d, \mathbf{X}^{(d)}(t), \mathbf{Y}^{(d)}(t+1)\right)$, where

$$\alpha\left(d,\mathbf{x}^{(d)},\mathbf{y}^{(d)}\right)=\begin{cases}\min\left\{1,\frac{\pi\left(d,\mathbf{y}^{(d)}\right)}{\pi\left(d,\mathbf{x}^{(d)}\right)}\right\}, & \pi\left(d,\mathbf{x}^{(d)}\right)q\left(d,\mathbf{x}^{(d)},\mathbf{y}^{(d)}\right)>0\\ 1, & \pi\left(d,\mathbf{x}^{(d)}\right)q\left(d,\mathbf{x}^{(d)},\mathbf{y}^{(d)}\right)=0\end{cases}.$$

If the proposed move is accepted, the chain jumps to $\mathbf{X}^{(d)}(t+1)=\mathbf{Y}^{(d)}(t+1)$; otherwise, it stays where it is and $\mathbf{X}^{(d)}(t+1)=\mathbf{X}^{(d)}(t)$.

Note that the method just described refers to a Metropolis-Hastings algorithm. The Gaussian form of the proposal density however implies that the kernel driving the chain is a random walk, from where the appellation RWM algorithm. Furthermore, since the choice of proposal density supports the irreducibility and aperiodicity conditions of the transition kernel, and since the acceptance rule is chosen to ensure the reversibility of the chain with respect to the density $\pi(\cdot)$, it follows that the distribution of the generated chain converges to that of the target.

For the algorithm to be efficient, it is primordial to carefully choose the scaling parameter of the proposal distribution (the variance of the normal distribution, in our case). Large values for $\sigma^2(d)$ will generally favour jumps that are far away from the current state of the chain, often in regions where the target density is low. As a consequence, proposed moves will usually be rejected and the chain will linger on some states for long periods of time. On the other hand, small values for $\sigma^2(d)$ will generate short jumps, resulting in a dawdling exploration of the state space.

Prior to determine an exact value for the proposal scaling, it seems sensible to decide on its form as a function of d. Clearly, $\theta_1(d)$ must be taken into account in this choice to prevent from proposing jumps that would be too large for components with smaller scaling terms, and thus inefficiently amplifying the rejection rate. Furthermore, we should bear in mind that the number of components increases as $d\to\infty$, raising the odds of proposing a silly move in one direction. To circumvent such a situation, it is reasonable to opt for a proposal scaling that is decreasing as a function of d.

The optimal proposal scaling form is given by $\sigma^2(d)=\ell^2/d^\alpha$, where ℓ^2 is some constant and α is the smallest number satisfying

$$\lim_{d\to\infty}\frac{d^{\lambda_1}}{d^\alpha}<\infty \qquad \text{and} \qquad \lim_{d\to\infty}\frac{d^{\gamma_i}c\left(\mathcal{J}(i,d)\right)}{d^\alpha}<\infty, \quad \text{for } i=1,\ldots,m. \tag{6}$$

In fact, more can be said about the determination of the proposal scaling. Under the fulfilment of Condition (5), we show in Section 7.1 that we automatically obtain $\sigma^2(d)=\ell^2/d^{\lambda_1}$, that is a proposal scaling governed by the b asymptotically smallest scaling terms. This conclusion is opposite to that achieved in [1], where we concluded that for 0.234 to be the asymptotic acceptance rate optimizing the efficiency of the chain, the form of the proposal scaling have to be based on one of the m groups of scaling terms appearing infinitely often in the limit. This divergence is explained by the significant difference between λ_1 and any γ_i, $i = 1, \ldots, m$, introduced by Condition (5). Indeed, if the first b scaling terms were ignored in the determination of the parameter α, this could result in larger proposal scaling and rejection rate, compromising the convergence speed of the algorithm. Specifically, we can differentiate two situations: one where there exists at least one $i \in \{1, \ldots, m\}$ such that the term $c\left(\mathcal{J}\left(i, d\right)\right) d^{\gamma_i}$ is $O\left(d^{\lambda_1}\right)$ and the other one where there does not exist any such i. In the first case, we can say that the first b scaling terms dictate in part only the comportment of the algorithm since everything else being held constant, ignoring these scaling terms would not affect the proposal distribution. The second situation is however different in the sense that ignoring these terms would result in a larger value of α. We shall thus study these two situations separately.

The requirement that the scaling term of the component of interest be independent of the dimension implies that the proposal scaling is at most $\sigma^2\left(d\right) = \ell^2$; this then keeps it from diverging as the dimension grows. In particular, the proposal scaling will take its largest form when studying one of the first b components only. Having found the optimal form for the scaling of the proposal distribution, we can thus write

$$\mathbf{Y}^{(d)} - \mathbf{x}^{(d)} \sim N\left(\mathbf{0}, \frac{\ell^2}{d^{\lambda_1}}\ I_{d\times d}\right).$$

Initially, RWM algorithms are discrete-time processes and thus on a microscopic level, the chain evolves according to the transition kernel outlined earlier. The proposal scaling (space) being function of d, an appropriate re-scaling of the elapsed time between each step will guarantee the obtention of a nontrivial limiting process as $d \to \infty$. This corresponds to study the model from a macroscopic viewpoint and on this level, we shall see next section that the component of interest will usually behave as a Langevin diffusion process. The only exception to this will happen with the b smallest order components, i.e. when $\sigma^2\left(d\right) = \ell^2$, since a time-speeding factor should not be applied in this case.

Let $\mathbf{Z}^{(d)}\left(t\right)$ be the time-t value of the RWM process sped up by a factor of d^{α}. In particular,

$$\mathbf{Z}^{(d)}\left(t\right) = \mathbf{X}^{(d)}\left(\left[d^{\alpha}t\right]\right) = \left(X_1^{(d)}\left(\left[d^{\alpha}t\right]\right), \ldots, X_d^{(d)}\left(\left[d^{\alpha}t\right]\right)\right),$$

where $[\cdot]$ is the integer part function. The periods of time between each step are thus shorter and the sped up process will propose on average d^{α} moves during each time interval instead of just one. Subsequent sections shall be devoted to study the limiting distribution of each component of the process $\left\{\mathbf{Z}^{(d)}\left(t\right), t \geq 0\right\}$.

Finally, before attempting to find an optimal value for the constant ℓ^2, we should define a criterion by which measuring efficiency. The authors in [15] introduce the notion of π-average acceptance rate, which is

$$\begin{aligned} a\left(d, \ell\right) &= \mathrm{E}\left[1 \wedge \frac{\pi\left(d, \mathbf{Y}^{(d)}\right)}{\pi\left(d, \mathbf{X}^{(d)}\right)}\right] \\ &= \int\int \pi\left(d, \mathbf{x}^{(d)}\right) \alpha\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) q\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) d\mathbf{x}^{(d)} d\mathbf{y}^{(d)} \qquad (7) \end{aligned}$$

for our d-dimensional symmetric RWM algorithm. The asymptotic efficiency of the algorithm is closely connected to this concept, as shall be seen next section.

4 Optimal Scaling Results

This section introduces weak convergence results which shall later be used to establish an equation permitting numerically solving for the optimal ℓ^2 value. Among target distributions satisfying Condition (5), it will reveal necessary to separate out the case where the asymptotically smallest scaling terms are the only ones to govern the proposal scaling from the case where they share this responsibility with at least one of the m groups having an infinite number of scaling terms in the limit.

We denote weak convergence in the Skorokhod topology by $\Rightarrow$, standard Brownian motion at time t by $B(t)$, and the standard normal cumulative distribution function (*cdf*) by $\Phi(\cdot)$. Since the scaling term of the component of interest (θ_{i^*}) is taken equal to 1, a linear transformation on the target distribution might be required.

Theorem 1. *Consider a RWM algorithm with proposal distribution*

$$\mathbf{Y}^{(d)} \sim N\left(\mathbf{x}^{(d)}, \frac{\ell^2}{d^{\lambda_1}} I_{d\times d}\right),$$

applied to a target density as in (1) satisfying the specified conditions on f and with $\theta_j^{-2}(d)$, $j = 1, \ldots, d$ as in (3). Consider the i^*-th component of the process $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$, that is $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\} = \left\{X_{i^*}^{(d)}\left(\left[d^{\lambda_1} t\right]\right), t \geq 0\right\}$, and let $\mathbf{X}^{(d)}(0)$ be distributed according to the target density π in (1).

We have

$$\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\} \Rightarrow \{Z(t), t \geq 0\},$$

where $Z(0)$ is distributed according to the density f and $\{Z(t), t \geq 0\}$ is as below, if and only if

$$\lim_{d\to\infty} \frac{\theta_1^2(d)}{\sum_{j=1}^d \theta_j^2(d)} > 0 \tag{8}$$

and there is at least one $i \in \{1, \ldots, m\}$ satisfying

$$\lim_{d\to\infty} \frac{c(\mathcal{J}(i,d))\, d^{\gamma_i}}{d^{\lambda_1}} > 0, \tag{9}$$

with $c(\mathcal{J}(i,d))$ as in (4).

For $i^* = 1, \ldots, b$ with $b = \max\left(j \in \{1, \ldots, n\}; \lambda_j = \lambda_1\right)$, the limiting process $\{Z(t), t \geq 0\}$ is a Metropolis-Hastings algorithm with acceptance rule

$$\begin{aligned}\alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right) &= \mathrm{E}_{\mathbf{Y}^{(b)-}, \mathbf{X}^{(b)-}}\left[\Phi\left(\frac{\sum_{j=1}^{b} \varepsilon\left(X_j, Y_j\right) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right. \\ &\left. + \prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)} \Phi\left(\frac{-\sum_{j=1}^{b} \varepsilon\left(X_j, Y_j\right) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right]. \qquad (10)\end{aligned}$$

For $i^* = b+1, \ldots, d$, $\{Z(t), t \geq 0\}$ satisfies the Langevin stochastic differential equation (SDE)

$$dZ(t) = \upsilon(\ell)^{1/2} dB(t) + \frac{1}{2} \upsilon(\ell) \left(\log f\left(Z(t)\right)\right)' dt,$$

where

$$\upsilon(\ell) = 2\ell^2 \mathrm{E}_{\mathbf{Y}^{(b)}, \mathbf{X}^{(b)}}\left[\Phi\left(\frac{\sum_{j=1}^{b} \varepsilon\left(X_j, Y_j\right) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right].$$

In both cases, $\varepsilon\left(X_j, Y_j\right) = \log\left(f\left(\theta_j Y_j\right) / f\left(\theta_j X_j\right)\right)$ and

$$E_R = \lim_{d \to \infty} \sum_{i=1}^{m} \frac{c\left(\mathcal{J}(i, d)\right)}{d^{\lambda_1}} \frac{d^{\gamma_i}}{K_{n+i}} \mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]. \qquad (11)$$

Interestingly, the b components of smallest order each possess a discrete-time limiting process. In comparison with the other components, they already converge fast enough so a speed-up time factor is superfluous. Furthermore, the acceptance rule of the limiting Metropolis-Hastings algorithm is influenced by the components affecting the form of the proposal scaling only. These components are more likely to cause the rejection of the proposed moves, and in that sense they constitute the components having the deepest impact on the rejection rate of the algorithm, ultimately becoming the only ones having an impact as $d \to \infty$.

It is worth noticing the singular form of the acceptance rule, which verifies the detailed balance condition and can be shown to belong to the Metropolis-Hastings family (see [11]). In particular when $b = 1$ the expectation operator can be dropped and for a general asymmetric proposal density $q(x, y)$ we obtain

$$\alpha\left(\ell^2, x, y\right) = \Phi\left(\frac{\log \frac{f(y)q(y,x)}{f(x)q(x,y)} - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right) + \frac{f(y)\, q(y, x)}{f(x)\, q(x, y)} \Phi\left(\frac{\log \frac{f(x)q(x,y)}{f(y)q(y,x)} - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right).$$

The effectiveness of this new acceptance rule depends on the parameter ℓ^2. When $\ell^2 \to \infty$, the proposed moves become enormous and $\alpha\left(\ell^2, x, y\right) \to 0$, meaning that the chain never moves. At the other extreme, if $\ell^2 = 0$ then E_R is granted no importance, meaning that the components among $X_{b+1}, \ldots, X_d$ that affected the proposal scaling have no more impact on the acceptance probability. The resulting rule is thus the usual one, i.e. $1 \wedge \frac{f(y)q(y,x)}{f(x)q(x,y)}$.

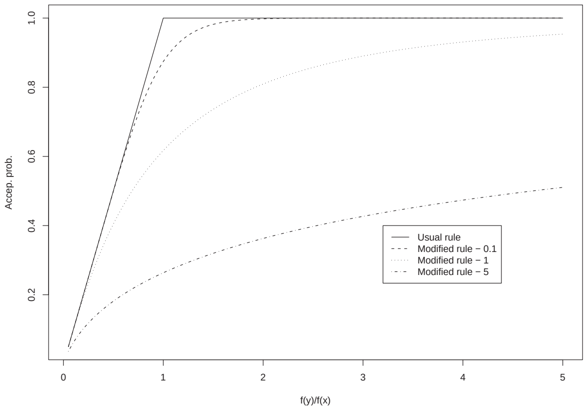

Figure 1: Modified acceptance rule in (10) as a function of the density ratio $f(y)/f(x)$ for different values of $\ell^2 E_R$ and when $b = 1$. From the top to the bottom, $\ell^2 E_R$ takes the values 0, 0.1, 1 and 5.

In [14], the optimal Metropolis-Hastings acceptance rule was shown to be $1 \wedge \frac{f(y)q(y,x)}{f(x)q(x,y)}$ because it favors the mixing of the chain by improving the sampling of all possible states. The efficiency of the modified acceptance rule is thus inversely proportional to its parameter ℓ^2, which is pictured in Figure 1. Coming back to the case where $b \geq 1$, we intuitively know that if many components are ruling the algorithm then it will be harder to accept moves. For $\ell^2 > 0$, we thus expect the probability of accepting the proposed move y_{i^*} given that we are at state x_{i^*} to get smaller as b and/or E_R get larger.

For $i^* = b+1, \ldots, d$, the scaling of the proposal distribution is function of d and a speed-up time factor is then required in order to get a sensible limit. Consequently, we obtain a continuous-time limiting process, and the speed measure of the limiting Langevin diffusion is now different from that found in [1] and [15]. It depends on exactly the same components as for the discrete-time limit and as we shall see, this alters the value of the AOAR.

Since there are two limiting processes for the same algorithm, we now face the dilemma as to which should be chosen to determine the AOAR. Indeed, the algorithm either accept or reject all d individual moves in a given step so it is important to have a common acceptance rate in all directions. The limiting distribution of the first b components being discrete, their AOAR is governed by a one-dimensional Metropolis-Hastings algorithm with a singular acceptance rule. This is however a source of ambiguities since for discrete-time processes, measures of efficiency are not unique and would yield different acceptance rates depending on which one is chosen. Fortunately, this issue does not exists for the limiting Langevin diffusion process obtained from the last $d-b$ components, as all measures of efficiency turn out to be equivalent. In our case, optimizing the efficiency corresponds to maximizing the speed measure of the diffusion $(\upsilon(\ell))$, which is justified by the fact that the speed measure is proportional to the mixing rate of the algorithm.

The following corollary provides an equation for the asymptotic acceptance rate of the algorithm as $d \to \infty$.

Corollary 2. *In the settings of Theorem 1 we have* $\lim_{d\to\infty} a(d,\ell) = a(\ell)$*, where*

$$a(\ell) = 2\mathrm{E}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left[\Phi\left(\frac{\sum_{j=1}^{b}\varepsilon(X_j,Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right].$$

An analytical solution for the value $\hat{\ell}$ that maximizes the function $\upsilon(\ell)$ cannot be obtained. However, this maximization problem can be easily resolved through the use of numerical methods. For virtually any density f satisfying the regularity conditions mentioned in Section 2, $\hat{\ell}$ will be finite and unique. This will thus yield an AOAR $a\left(\hat{\ell}\right)$ and although an explicit form is not available for this quantity, we can still draw some conclusions about the AOAR. First, Condition (5) ensures the existence of a finite number of components having a scaling term significantly smaller than the others. Since this constitutes the complement of the case treated in [1], we know that the variation in the speed measure is directly due to these components. When studying any component X_{i^*} with $i^* \in \{b+1,\ldots,d\}$, we also know that $\theta_j^{-2}(d) \to 0$ as $d \to \infty$ for $j = 1,\ldots,b$ since these scaling terms are of smaller order than $\theta_{i^*} = 1$. Hence, the first b components obviously provoke a reduction of the AOAR, which is now necessarily smaller than 0.234. In particular, the AOAR will get smaller as b increases.

We now consider the remaining case where there is again a finite number of components having scaling terms converging significantly faster than the others, but where these scaling terms are now entirely ruling the proposal scaling. This is the case where these special components are converging "too fast" compared with the overall convergence speed of the algorithm.

Theorem 3. *In the settings of Theorem 1 but with Condition (9) replaced by*

$$\lim_{d\to\infty}\frac{c(\mathcal{J}(i,d))\,d^{\gamma_i}}{d^{\lambda_1}} = 0 \quad \forall\, i \in \{1,\ldots,m\}, \tag{12}$$

the conclusions of Theorem 1 are preserved, but the acceptance rule is now

$$\alpha(X_{i^*},Y_{i^*}) \quad = \quad \mathrm{E}_{\mathbf{Y}^{(b)-},\mathbf{X}^{(b)-}}\left[1 \wedge \prod_{j=1}^{b}\frac{f(\theta_j Y_j)}{f(\theta_j X_j)}\right]$$

for the limiting Metropolis-Hastings algorithm and the speed measure is

$$\upsilon(\ell) = 2\ell^2 \mathrm{P}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left(\sum_{j=1}^{b}\varepsilon(X_j,Y_j) > 0\right)$$

for the limiting Langevin diffusion.

As in Theorem 1 there are two different limiting processes, depending on which component we focus. Since the proposal scaling is now entirely ruled by the first b components, it means that $E_R = 0$. When $b = 1$, the acceptance rule of the limiting RWM algorithm reduces to the usual rule. In that case, the first component not only becomes independent of the others as $d \to \infty$, but it is completely unaffected by these $d-1$ components in the limit, which move too slowly compared to the pace of X_1. For the last $d-b$ components, the limiting process is continuous and the speed measure of the diffusion is also affected by the first b components only. As mentioned previously, we use the speed measure of the diffusion to attempt optimizing the efficiency of the chain.

Corollary 4. *In the settings of Theorem 3, we have* $\lim_{d\to\infty} a(d,\ell) = a(\ell)$, *where*

$$a(\ell) = 2\mathrm{P}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left(\sum_{j=1}^{b} \varepsilon(X_j, Y_j) > 0\right).$$

When trying to optimize the speed measure, we realize that it is unbounded as a function of ℓ for virtually any choice of density f satisfying the regularity conditions of Section 2. This thus suggests that we should chose ℓ as large as possible, and hence indicates that bigger proposal scalings might be more appropriate. However, the scaling of the form $\sigma^2(d) = \ell^2/d^{\lambda_1}$ is the only one yielding a nontrivial limit. In addition, the asymptotic acceptance rate converges to 0 as $\ell \to \infty$. This thus confirms that our choice for the scaling of the proposal is right, but that the obtained AOAR is useless in practice. We conclude that when Condition (12) is satisfied, the first b components converge to 0 very rapidly, governing the choice for the parameter α in (6). However, such a proposal distribution generates moves that are way too small for the other components, from where the conclusion that we should chose ℓ as large as possible. Note that if we were choosing a smaller value for α, the proposed moves would be too big for the first b components, resulting in the divergence of the limiting process. In theory, we thus obtain a well-defined limiting process, but in practice we lie in an impasse when it comes to optimize the efficiency of the algorithm. We shall see in Section 6 that for such cases, inhomogeneous proposal scalings constitute a wiser option.

5 Normal Case

The results of the previous section can be somehow simplified when f is the standard normal density function, since it is then possible to compute expectations with respect to $\mathbf{X}^{(b)}$ and conditional on $\mathbf{Y}^{(b)}$. We obtain the following results.

Theorem 5. *In the settings of Theorem 1 with* $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, *the conclusions of Theorem 1 and Corollary 2 are preserved but with Metropolis-Hastings acceptance rule*

$$\alpha\left(\ell^2, x_{i^*}, y_{i^*}\right) = \mathrm{E}\left[\Phi\left(\frac{\varepsilon(x_{i^*}, y_{i^*}) - \frac{\ell^2}{2}\left(\sum_{j=1,j\neq i^*}^{b}\frac{\chi_j^2}{K_j} + E_R\right)}{\sqrt{\ell^2\left(\sum_{j=1,j\neq i^*}^{b}\frac{\chi_j^2}{K_j} + E_R\right)}}\right)\right. \tag{13}$$

$$+\frac{f\left(y_{i^*}\right)}{f\left(x_{i^*}\right)}\Phi\left(\frac{-\varepsilon\left(x_{i^*},y_{i^*}\right)-\frac{\ell^2}{2}\left(\sum_{j=1,j\neq i^*}^{b}\frac{\chi_j^2}{K_j}+E_R\right)}{\sqrt{\ell^2\left(\sum_{j=1,j\neq i^*}^{b}\frac{\chi_j^2}{K_j}+E_R\right)}}\right)\Bigg],$$

where χ_j^2*,* $j=1,\dots,b$ *are independent chi square random variables with 1 degree of freedom and* E_R *simplifies to*

$$E_R=\lim_{d\to\infty}\sum_{i=1}^{m}\frac{c\left(\mathcal{J}\left(i,d\right)\right)}{d^{\lambda_1}}\frac{d^{\gamma_i}}{K_{n+i}}.$$

In addition, the Langevin speed measure is now given by

$$\upsilon\left(\ell\right)=2\ell^2\mathrm{E}\left[\Phi\left(-\frac{\ell}{2}\sqrt{\sum_{j=1}^{b}\frac{\chi_j^2}{K_j}+E_R}\right)\right],$$

and the limiting average acceptance rate satisfies

$$a\left(\ell\right)=2\mathrm{E}\left[\Phi\left(-\frac{\ell}{2}\sqrt{\sum_{j=1}^{b}\frac{\chi_j^2}{K_j}+E_R}\right)\right].$$

Finally, $\upsilon\left(\ell\right)$ *is maximized at the unique value* $\hat{\ell}$ *and the AOAR is given by* $a\left(\hat{\ell}\right)$*.*

In practice, it then suffices to numerically maximize $\upsilon\left(\ell\right)$ and to run the algorithm with a proposal scaling equal to $\hat{\ell}^2/d^{\lambda_1}$ in order to optimize the efficiency of the technique. The proportion of accepted moves will then lie around $a\left(\hat{\ell}\right)$. Even though the previous result is asymptotic, it also works well in relatively small dimensions just as for *iid* target distributions (see [17]). For the second case where some components entirely rule the proposal scaling, we find the following result.

Theorem 6. *In the settings of Theorem 3 with* $f\left(x\right)=\left(2\pi\right)^{-1/2}\exp\left(-x^2/2\right)$*, the conclusions of Theorem 3 and Corollary 4 are preserved but with Langevin speed measure*

$$\upsilon\left(\ell\right)=2\ell^2\mathrm{E}\left[\Phi\left(-\frac{\ell}{2}\sqrt{\sum_{j=1}^{b}\frac{\chi_j^2}{K_j}}\right)\right],$$

where χ_j^2*,* $j=1,\dots,b$ *are independent chi square random variables with 1 degree of freedom.*

Furthermore, the limiting average acceptance rate now satisfies

$$a\left(\ell\right)=2\mathrm{E}\left[\Phi\left(-\frac{\ell}{2}\sqrt{\sum_{j=1}^{b}\frac{\chi_j^2}{K_j}}\right)\right].$$

When $b = 1$, the limiting process of X_1 is the usual one-dimensional RWM algorithm. As we said before, measures of efficiency are not unique in this case but to understand the situation, suppose we consider first-order efficiency. We then want to maximize the expected square jump distance, which will result in an better mixing of the chain. The acceptance rate maximizing this quantity is 0.45, as given from the finite-dimensional results for RWM algorithms in [17]. As b increases, more and more components affect the acceptance process and this results in a reduction of the AOAR towards 0.234. Indeed, when b is infinite then Condition (5) is not satisfied anymore and we find ourselves facing the complemental case introduced in [1]. It is however important to note that in such a case, the proposal scaling $\sigma^2(d) = \ell^2/d^{\lambda_1}$ would be inappropriate and we would have to determine a new value for α in (6) in order to handle this new case, using the cardinality function $c(\mathcal{J}(i,d))$ of these components. In the case of Theorem 5, the acceptance rule is more restrictive and accepting moves is thus harder. First-order efficiency is maximized when the acceptance rate is about 0.35 for $b = 1$, and decreases towards 0.234 as b increases. The difference between both rules thus becomes insignificant for large values of b.

The previous analysis allows to get some insight about the situation for discrete-time limits. Nonetheless in practice, continuous-time limits must be used to determine the AOAR that should be applied for optimal performance of the algorithm. In Theorem 5 the speed measure $\upsilon(\ell)$ is always smaller than $2\ell^2\Phi(-\ell E_R/2)$, the speed measure of the limiting Langevin diffusion when Condition (5) is not met (see [1]). Consequently, the optimal value $\hat{\ell}$ is bounded above by $2.38/\sqrt{E_R}$ and gets smaller as the number b of components increases. As expected, the diminution of the parameter ℓ is not important enough to outdo the factors χ_i^2/K_i and the AOAR is thus continually smaller than 0.234. This difference is intensified with the growth of b.

The speed measure in Theorem 6 is particular in the sense that its expectation term does not vanish fast enough to overturn the growth of ℓ^2. The optimal value $\hat{\ell}$ is thus infinite, yielding an AOAR converging to 0. This means that any acceptance rate will result in an algorithm that is inefficient in practice for large d. The best solution is to resort to inhomogeneous proposal distribution, which shall be discussed next section.

The particularity of the results introduced in this section resides in their ability to optimize RWM algorithms applied to any multivariate normal target with correlated components. Since multivariate normal distributions are invariant under orthogonal transformations, it suffices to transform the covariance matrix into a diagonal matrix, where the diagonal elements are the eigenvalues of the covariance matrix. Under this transformation, the target components are now independent and the eigenvalues can be used to verify if Condition (5) is satisfied. In the case where it is, we can use either Theorem 5 or 6 to determine the AOAR, depending on which of Conditions (9) or (12) is satisfied. Otherwise, the AOAR is the well-known 0.234 as demonstrated in [1].

Using a similar technique, these results can also be applied to normal hierarchical models since such models are jointly normal. For instance, in the case where $X_i \sim N(X_1, 1)$, $i = 2, \ldots, d$ with $X_1 \sim N(0, 1)$ the joint distribution has mean vector $\mathbf{0}$ and covariance matrix

$$\begin{pmatrix} 1 & 1 & \cdots & & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ & & \ddots & & \\ 1 & \cdots & 1 & 2 & 1 \\ 1 & & \cdots & 1 & 2 \end{pmatrix}.$$

Among the d eigenvalues of this matrix, $d-2$ are equal to 1 and the other two are $O\left(d\right)$ and $O\left(1/d\right)$. It is easily checked that both Conditions (5) and (9) are satisfied, so Theorem 5 can be applied to determine the AOAR, which turns out to be around 0.2. We note that in the case where the eigenvalues do not assume the form specified in Section 2, we can instead refer to the generalizations of Section 6 to reach the same conclusions. Detailed examples about the different applications of these results are presented in [2].

6 Inhomogeneous Proposal Scalings and Extensions

As realized previously, it is possible to face a situation where the efficiency of the algorithm cannot be optimized under homogeneous proposal scalings. This happens when a finite number of scaling terms request a proposal scaling of very small order, resulting in an excessively slow convergence of the other components. To overcome this problem, inhomogeneous proposal scalings will add a touch a personalization and ensure a decent speed of convergence in each direction.

The idea is to adjust the parameter α in (6) to better fit the different components. However, we should be careful as to preserve the stochastic property of the process $\left\{\mathbf{Z}^{(d)}\left(t\right), t \geq 0\right\}$. To possess this characteristic, it is mandatory that each component be sped up by the same time factor since all d components are dependent in finite dimensions. It turns out that this factor is d^{α}, the only factor yielding a nontrivial limiting process. The obtention of a reasonable limiting process is then also conditional on the proposal scaling of the component of interest being ℓ^2/d^{α}. Since we wish to study each of the first $n+m$ components, we then personalize the scaling of the last $d-n-m$ components only. That is, we adjust the proposal scaling of the components whose scaling term appears infinitely often in the limit, but making sure to put aside one component of each of these m groups.

In particular, consider the $\theta_j\left(d\right)$'s appearing in (3) and let $\mathbf{Z}^{(d)}\left(t\right) = \mathbf{X}^{(d)}\left(\left[d^{\alpha}t\right]\right)$ as before. We set the proposal scalings as follows: for $j = 1, \ldots, n+m$ let $\sigma^2\left(d\right) = \ell^2/d^{\lambda_1}$ and for $j = n+m+1, \ldots, d$, $j \in \mathcal{J}\left(i, d\right)$, let $\sigma^2\left(d\right) = \ell^2/\left(c\left(\mathcal{J}\left(i,d\right)\right)d^{\gamma_i}\right)$. We then have the following result.

Theorem 7. *In the settings of Theorems 1 and 3 (i.e. no matter if Condition (9) is satisfied or not) but with the proposal scaling as just described, the conclusions of Theorem 1 and Corollary 2 are preserved.*

Due to the increase of the term E_R, the optimal scaling value $\hat{\ell}$ will now be smaller than with homogeneous proposal scalings. This makes sense since when the proposal scaling of each component was based on λ_1, the algorithm had to compensate for the fact that the proposal scaling was maybe too small for certain groups of components with a larger value for $\hat{\ell}$. In [1], it was easily verified that the AOAR is unaffected by the use of inhomogeneous proposals. The same statement does not hold in the present case, although we can still affirm that the AOAR will not be greater than 0.234. Indeed, since ℓ is assumed to be fixed in each direction, the algorithm can hardly do better than for *iid* targets even though the proposal has been personalized. Recall that in the *iid* case, both homogeneous and inhomogeneous settings would yield the same proposal distribution, and talking about inhomogeneity is thus meaningless.

For a good example where the use of inhomogeneous proposal scalings is beneficial, consider a d-dimensional normal target with independent components, where the variances of the components are given by $(1/d^2, 1, \ldots, 1)$. The homogeneous proposal scaling is ℓ^2/d^2 and by Theorem 6, this yields an AOAR of 0. On the other hand, the inhomogeneous proposal scaling is $(\ell^2/d^2, \ell^2/d^2, \ell^2/d, \ldots, \ell^2/d)$, yielding an AOAR of about 0.16 which is obviously far more efficient than the former.

We now study how the results in previous sections extend to scaling terms admitting broader forms. We start by relaxing the assumption of equality among the scaling terms belonging to a common group. That is, within each of the m groups of scaling terms appearing infinitely often, we allow the constant terms to be randomly distributed according to a distribution satisfying specific moment conditions. Specifically, we have

$$\mathbf{\Theta}^{-2}(d) = \left(\frac{K_1}{d^{\lambda_1}}, \ldots, \frac{K_n}{d^{\lambda_n}}, \frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+c(\mathcal{J}(1,d))}}{d^{\gamma_1}}, \ldots, \frac{K_{n+\sum_{i=1}^{m-1} c(\mathcal{J}(i,d))+1}}{d^{\gamma_m}}, \ldots, \frac{K_d}{d^{\gamma_m}} \right). \quad (14)$$

We assume that $\{K_j, j \in \mathcal{J}(i,d)\}$ are *iid* and chosen randomly from some distribution with $\mathrm{E}\left[K_j^{-2}\right] < \infty$. Without loss of generality, we also take $\mathrm{E}\left[K_j^{-1/2}\right] = 1$ and denote $\mathrm{E}\left[K_j^{-1}\right] = b_i$ for $j \in \mathcal{J}(i,d)$. Recall that the scaling term of the component of interest does not depend on d, and we therefore have $\theta_{i^*}^{-2}(d) = K_{i^*}$.

To support the previous modifications, we now suppose that $-\infty < \gamma_m < \gamma_{m-1} < \ldots < \gamma_1 < \infty$, otherwise we could just group the scaling terms with a common power together. In addition, we also suppose that there does not exist a λ_j, $j = 1, \ldots, n$ equal to one of the γ_i, $i = 1, \ldots, m$ because if this was not true, we could just include the j-th scaling term in the i-th group.

Theorem 8. *Consider the settings of Theorem 1 (Theorem 3) with $\mathbf{\Theta}^{-2}(d)$ as in (14), $\theta_{i^*} = K_{i^*}^{-1/2}$ and replace Condition (8) by*

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{\sum_{j=1}^{n} d^{\lambda_j} + \sum_{i=1}^{m} c(\mathcal{J}(i,d))\, d^{\gamma_i}} > 0. \quad (15)$$

We have

$$\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\} \Rightarrow \{Z(t), t \geq 0\},$$

where $Z(0)$ is distributed according to the density $\theta_{i^*} f(\theta_{i^*} x)$ and $\{Z(t), t \geq 0\}$ is identical to the limit found in Theorem 1 (Theorem 3) for the first b components, but where it satisfies the Langevin SDE

$$dZ(t) = (\upsilon(\ell))^{1/2} dB(t) + \frac{1}{2}\upsilon(\ell)(\log f(\theta_{i^*} Z(t)))' dt$$

for the other $d-b$ components, with $\upsilon(\ell)$ as in Theorem 1 (Theorem 3).

For both limiting processes, we now use

$$E_R = \lim_{d\to\infty} \sum_{i=1}^{m} \frac{c(\mathcal{J}(i,d))\, d^{\gamma_i}}{d^{\alpha}}\; b_i \mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]$$

instead of (11) in Theorem 1, with

$$c(\mathcal{J}(i,d)) = \#\left\{j \in \{n+1,\ldots,d\}; \theta_j(d) \text{ is } O\left(d^{\gamma_i/2}\right)\right\}.$$

In addition, the conclusion of Corollary 2 (Corollary 4) is preserved.

Contrarily to what was done previously, this result does not assume a scaling term equal to 1 for the component of interest, but still presume that the scaling term in question is $O(1)$. The difference engendered by this modification consists in θ_{i^*} now appearing in the drift term of the Langevin diffusion. It is interesting to note that the randomness of the last $d-n$ scaling terms affects the quantity E_R through the factors $b_1,\ldots,b_m$. Therefore, it only has an impact on the limiting processes when Condition (9) is satisfied. Also note that Condition (15) is in fact the same as Condition (8) since constant terms are assumed to be finite. Consequently, both conditions could be use interchangeably in any of the previous results. For the present case however, Condition (15) is easier to compute due to the randomness of $K_{n+1},\ldots,K_d$.

The setting described in Section 2 for target distributions can be relaxed to allow for more generality. That is, we now permit functions other than polynomial for $c(\mathcal{J}(i,d))$, $i = 1,\ldots,m$, as long as they still satisfy $c(\mathcal{J}(i,d)) \to \infty$ as $d \to \infty$. We also broaden the form taken by $\theta_j(d)$, $j = 1,\ldots,d$. In order to have sensible limiting theory, we however restrict our attention to functions for which the limit exists as $d \to \infty$. We even allow the scaling terms $\left\{\theta_j^{-2}(d), j \in \mathcal{J}(i,d)\right\}$ to vary within each of the m groups, provided they are of the same order. That is, for $j \in \mathcal{J}(i,d)$ we suppose

$$\lim_{d\to\infty} \frac{\theta_j(d)}{\theta_i'(d)} = K_j^{-1/2},$$

for some reference function $\theta_i'(d)$ having no constant term (i.e. a constant term equal to 1) and some constant K_j coming from the distribution described for Theorem 8. Note that the reference functions are taken to be as simple as possible. As an example, if all components in a given group $i \in \{1, \ldots, m\}$ are $O(d)$, then $\theta'_i(d) = d$.

As just specified for Theorem 8, we assume that $\boldsymbol{\Theta}^{-2}(d)$ contains at least m and at most $n+m$ functions of different order. Hence, if there is infinitely many scaling terms of a same order in the limit, they necessarily belong to the same of the m groups, which are defined by

$$\mathcal{J}(i,d) = \left\{ j \in \{1, \ldots, d\} ; 0 < \lim_{d\to\infty} \theta_j^{-2}(d)\, \theta_i'^2(d) < \infty \right\}. \tag{16}$$

We again assume that the first n and the next m scaling terms are respectively classified according to an asymptotically increasing order. Specifically, let the first n terms of $\boldsymbol{\Theta}^{-2}(d)$ satisfy $\theta_1^{-2}(d) \preceq \ldots \preceq \theta_n^{-2}(d)$, and let the order of the terms $\theta_{n+1}^{-2}(d), \ldots, \theta_{n+m}^{-2}(d)$ be chosen to satisfy $\theta_1'^{-2}(d) \prec \ldots \prec \theta_m'^{-2}(d)$.

Due to the broader form for the target distribution, it is now necessary to modify the proposal scaling. In particular, let $\sigma^2(d) = \ell^2 \sigma_\alpha^2(d)$, with $\sigma_\alpha^2(d)$ the function of largest possible order such that

$$\lim_{d\to\infty} \theta_1^2(d)\, \sigma_\alpha^2(d) < \infty \quad \text{and} \quad \lim_{d\to\infty} c(\mathcal{J}(i,d))\, \theta_i'^2(d)\, \sigma_\alpha^2(d) < \infty \;\; \text{for } i = 1, \ldots, m. \tag{17}$$

We then have the following result.

Theorem 9. *Under the settings of Theorem 8, but with proposal scaling $\sigma^2(d) = \ell^2 \sigma_\alpha^2(d)$ where $\sigma_\alpha^2(d)$ satisfies (17) and with general functions for $c(\mathcal{J}(i,d))$ and $\theta_j(d)$ as defined previously, the conclusions of Theorem 8 are preserved, provided that*

$$\lim_{d\to\infty} \frac{\theta_1^2(d)}{\sum_{j=1}^n \theta_j^2(d) + \sum_{i=1}^m c(\mathcal{J}(i,d))\, \theta_i'^2(d)} > 0$$

holds instead of Condition (15),

$$\exists i \in \{1, \ldots, m\} \text{ such that } \lim_{d\to\infty} \frac{c(\mathcal{J}(i,d))\, \theta_i'^2(d)}{\theta_1^2(d)} > 0$$

holds instead of Condition (9), and

$$\lim_{d\to\infty} \frac{c(\mathcal{J}(i,d))\, \theta_i'^2(d)}{\theta_1^2(d)} = 0 \;\; \forall i \in \{1, \ldots, m\}$$

holds instead of Condition (12).

Under this setting, the quantity E_R is now given by

$$E_R = \lim_{d\to\infty} \sum_{i=1}^m c(\mathcal{J}(i,d))\, \sigma_\alpha^2(d)\, \theta_i'^2(d)\, b_i \mathrm{E}\left[\left(\frac{f'(X)}{f(X)} \right)^2 \right],$$

where $c(\mathcal{J}(i,d))$ is the cardinality function of (16).

This theorem assumes a somewhat general form for the target distribution. Contrarily to what was proven in [1], the AOAR is not independent of the target distribution and scaling terms anymore, and finding the exact AOAR involves solving the maximization problem of the speed measure $\upsilon(\ell)$ for every different case. Albeit the AOAR might turn out to be close to the usual 0.234 it is also possible to face a case where this rate is inefficient, from where the importance to determine the correct proposal scaling.

7 Theorems Proofs

This section aims to prove the optimal scaling theorems introduced in Sections 4 to 6. We shall present a detailed proof of Theorem 1, which makes use of the technical results introduced in Sections 8 and 9. However, due to the similarity between the proofs of the various results, we shall not include every detail when demonstrating the other theorems. We generally outline the main distinctions with the detailed demonstration only.

For the proofs presented subsequently to be complete, it is necessary to refer to various results appearing in [9]. The pillar of the results introduced in this paper is Corollary 8.1 of Chapter 4, which roughly says that the finite-dimensional distributions of a sequence of processes converge weakly to those of some Markov process provided that their generators converge in mean. According to Theorem 7.8 of Chapter 3, it is then sufficient to verify relative compactness of the sequence of stochastic processes to achieve weak convergence of the processes themselves. This property is easily assessed using Corollary 7.4 and Theorem 8.6 of Chapter 3, and by means of a continuity of probabilities argument along with the fact that the algorithm starts in stationarity.

The main objective here is thus to verify $\mathcal{L}^1$ convergence of generators, which are expressed as a function of some arbitrary test function h typically taken to be any smooth function. For our purpose however, the restriction to functions h belonging to the space of infinitely differentiable functions on compact support C_c^∞ is supported by Theorem 2.1 of Chapter 8 in [9]. This results says that for diffusions satisfying certain drift and volatility conditions, C_c^∞ is a core for the generator of the diffusion, roughly meaning that C_c^∞ depicts smooth functions well enough so that we can focus on this space of functions only. The approach used for the demonstrations is analogous to that for the RWM algorithm case in [13]. In that paper, the authors however base their proofs on a distinct result from [9] and instead prove uniform convergence of generators, which could not be use in the current situation.

In order to lighten the formulas, we characterize vectors as follows. The number in parentheses (say a) appearing at the exponent denotes the first a components of the d-dimensional vector. When a substraction of terms appears in the parentheses (say $b-a$), the vector is formed of the components $a+1,\ldots,b$. A minus sign appearing outside the brackets informs us that the i^*-th component, i.e. the component of interest, is excluded from the vector. For instance, the vector $\mathbf{X}^{(d-n)-}$ contains the last $d-n$ target random variables, i.e. the components having a scaling term appearing infinitely often, from which we excluded the i^*-th component.

We also adopt the following convention for conditional expectations. The expectation is computed with respect to the variables appearing as a subscript on the right of the expectation operator E. When there is no such subscript, this means that the expectation is taken with respect to all random variables included in the expression. For instance, we write

$$\mathrm{E}\left[f\left(X,Y\right)\right]=\mathrm{E}\left[\mathrm{E}\left[f\left(X,Y\right)|Y\right]\right]=\mathrm{E}_{Y}\left[\mathrm{E}_{X}\left[f\left(X,Y\right)\right]\right].$$

7.1 Restrictions on the Proposal Scaling

As stated in Section 2, Condition (5) ensures that there exists a finite number of scaling terms significantly smaller than the others. The impact of this condition is that it also determines the scaling of the proposal distribution, which we now demonstrate. For Condition (5) to be met, its reciprocal has to satisfy

$$\begin{aligned}&\lim_{d\to\infty}\sum_{j=1}^{d}\hat{\theta}^{-2}\left(d\right)\theta_j^2\left(d\right)\\&=\lim_{d\to\infty}\frac{K_1}{d^{\lambda_1}}\left(\frac{d^{\lambda_1}}{K_1}+\ldots+\frac{d^{\lambda_n}}{K_n}+c\left(\mathcal{J}\left(1,d\right)\right)\frac{d^{\gamma_1}}{K_{n+1}}+\ldots+c\left(\mathcal{J}\left(m,d\right)\right)\frac{d^{\gamma_m}}{K_{n+m}}\right)<\infty.\end{aligned}$$

It must then be true that

$$\lim_{d\to\infty}\frac{c\left(\mathcal{J}\left(i,d\right)\right)d^{\gamma_i}}{d^{\lambda_1}}<\infty,\quad\forall i\in\left\{1,\ldots,m\right\},$$

in which case

$$\lim_{d\to\infty}\sum_{j=1}^{d}\hat{\theta}^{-2}\left(d\right)\theta_j^2\left(d\right)\;=\;1+\sum_{j=2}^{b}K_1/K_j<\infty,$$

where $b=\max\left(j\in\left\{1,\ldots,n\right\};\lambda_j=\lambda_1\right)$, the number of components with a scaling term of the same order as that of X_1. In other words, $\theta_1^{-2}\left(d\right)$ is not only the asymptotically smallest scaling term, but it must also be small enough so as to act of proposal scaling for the algorithm, implying $\sigma^2\left(d\right)=\ell^2/d^{\lambda_1}$.

7.2 Proof of Theorem 1

We now show that the generator of the sped up RWM algorithm

$$Gh\left(d,X_{i^*}\right)\;=\;d^{\lambda_1}\mathrm{E}_{\mathbf{Y}^{(d)},\mathbf{X}^{(d)-}}\left[\left(h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right)\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right)\right]\tag{18}$$

converges in mean to the generator of a Metropolis-Hastings algorithm with particular acceptance rule in some cases, and to that of a Langevin diffusion in other cases. To this end, we shall use the results appearing in Sections 8 and 9.

Proof. We first need to show that for $i^* \in \{1, \ldots, b\}$ and an arbitrary test function $h \in C_c^\infty$,

$$\lim_{d\to\infty} \mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - G_{MH}h\left(X_{i^*}\right)\right|\right] = 0,$$

where

$$G_{MH}h\left(X_{i^*}\right) = \mathrm{E}_{Y_{i^*}}\left[\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right)\alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right)\right]$$

with acceptance rule

$$\begin{aligned}\alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right) &= \mathrm{E}_{\mathbf{Y}^{(b)-},\mathbf{X}^{(b)-}}\left[\Phi\left(\frac{\sum_{j=1}^b \varepsilon\left(X_j, Y_j\right) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right.\\ &\quad \left. + \prod_{j=1}^b \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)}\Phi\left(\frac{-\sum_{j=1}^b \varepsilon\left(X_j, Y_j\right) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right].\end{aligned}$$

Since the component of interest must have a scaling term of 1, we have $\lambda_j = \lambda_1 = 0$ for $j = 1, \ldots, b$ and the proposal scaling is $\sigma^2(d) = \ell^2$. Since $\sigma^2(d)$ does not depend on the dimension of the target, there is thus no need to speed up the RWM algorithm for the obtention of a nontrivial limit, justifying the discrete-time nature of the limiting process.

We first introduce a third generator asymptotically equivalent to the original generator $Gh\left(d, X_{i^*}\right)$. Specifically, let

$$\widetilde{G}h\left(d, X_{i^*}\right) = \mathrm{E}_{\mathbf{Y}^{(d)},\mathbf{X}^{(d)-}}\left[\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right)\left(1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right)\right],$$

with $v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ as in (20). By the triangle's inequality,

$$\begin{aligned}&\mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - G_{MH}h\left(X_{i^*}\right)\right|\right]\\ &\quad\le \mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - \widetilde{G}h\left(d, X_{i^*}\right)\right|\right] + \mathrm{E}\left[\left|\widetilde{G}h\left(d, X_{i^*}\right) - G_{MH}h\left(X_{i^*}\right)\right|\right],\end{aligned}$$

and the first expectation on the RHS converges to 0 as $d \to \infty$ by Lemma 10. To complete the proof, we are then left to show $\mathcal{L}^1$ convergence of the third generator $\widetilde{G}h\left(d, X_{i^*}\right)$ to that of the modified Metropolis-Hastings algorithm.

Substituting explicit expressions for the generators and using the triangle's inequality along with the fact that the function h has compact support gives

$$\begin{aligned}&\mathrm{E}\left[\left|\widetilde{G}h\left(d, X_{i^*}\right) - G_{MH}h\left(X_{i^*}\right)\right|\right]\\ &\quad\le K\mathrm{E}_{X_{i^*},Y_{i^*}}\left[\left|\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] - \alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right)\right|\right],\end{aligned}$$

where K is chosen such that $|h(Y_{i^*}) - h(X_{i^*})| < K$. Since the expectation on the RHS converges to 0 as ascertained by Lemma 11, this proves the first part of Theorem 1.

To complete the proof, we must show that for $i^* \in \{b+1, \ldots, d\}$ and an arbitrary test function $h \in C_c^\infty$,

$$\lim_{d\to\infty} \mathrm{E}\left[\left|Gh(d, X_{i^*}) - G_L h(X_{i^*})\right|\right] = 0,$$

where

$$G_L(X_{i^*}) = \upsilon(\ell)\left[\frac{1}{2}h''(X_{i^*}) + \frac{1}{2}h'(X_{i^*})(\log f(X_{i^*}))'\right]$$

is the generator of a Langevin diffusion process with $\upsilon(\ell)$ as in Theorem 1.

As in the first part of the proof, we introduce a third generator now given by

$$\begin{aligned} \widetilde{G}h(d, X_{i^*}) &= \frac{1}{2}\ell^2 h''(X_{i^*})\,\mathrm{E}\left[1 \wedge e^{\sum_{j=1,j\neq i^*}^{d} \varepsilon(d,X_j,Y_j)}\right] \qquad (19)\\ &\quad + \ell^2 h'(X_{i^*})(\log f(X_{i^*}))'\,\mathrm{E}\left[e^{\sum_{j=1,j\neq i^*}^{d} \varepsilon(d,X_j,Y_j)};\ \sum_{j=1,j\neq i^*}^{d} \varepsilon(d, X_j, Y_j) < 0\right]. \end{aligned}$$

From Lemma 7 in [1], this new expression is asymptotically equivalent to the original generator and hence $\mathrm{E}\left[\left|Gh(d, X_{i^*}) - \widetilde{G}h(d, X_{i^*})\right|\right] \to 0$ as $d \to \infty$. We then have to conclude the proof by showing that the third generator also converges in mean to the generator of the Langevin diffusion.

Substituting explicit expressions for the generators and the speed measure, grouping some terms and using the triangle's inequality yield

$$\begin{aligned} &\mathrm{E}\left[\left|\widetilde{G}h(d, X_{i^*}) - G_L h(X_{i^*})\right|\right]\\ &\leq \ell^2 \left|\frac{1}{2}\mathrm{E}\left[1 \wedge e^{\sum_{j=1,j\neq i^*}^{d} \varepsilon(d,X_j,Y_j)}\right] - \mathrm{E}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left[\Phi\left(\frac{\sum_{j=1}^{b}\varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right]\right| \mathrm{E}\left[\left|h''(X_{i^*})\right|\right]\\ &\quad + \ell^2 \left|\mathrm{E}\left[e^{\sum_{j=1,j\neq i^*}^{d} \varepsilon(d,X_j,Y_j)};\ \sum_{j=1,j\neq i^*}^{d} \varepsilon(d, X_j, Y_j) < 0\right]\right.\\ &\quad \left. -\mathrm{E}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left[\Phi\left(\frac{\sum_{j=1}^{b}\varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right]\right| \mathrm{E}\left[\left|h'(X_{i^*})(\log f(X_{i^*}))'\right|\right]. \end{aligned}$$

Since the function h has compact support, it implies that h itself and its derivatives are bounded by some constant. Therefore, $\mathrm{E}\left[\left|h''(X_{i^*})\right|\right]$ and $\mathrm{E}\left[\left|h'(X_{i^*})(\log f(X_{i^*}))'\right|\right]$ are both bounded by K, say. Using Lemma 8 in [1] and Lemma 12, we then conclude that the first term on the RHS goes to 0 as $d \to \infty$. We reach the same conclusion for the second term by applying Lemma 10 in [1] along with Lemma 13. □

7.3 Proof of Theorem 3

The proof of Theorem 3 is similar to that of Theorem 1. The key is to notice that $v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right) \to_p \sum_{j=1}^{b} \varepsilon\left(X_j, Y_j\right)$. This is done by first realizing that $\sum_{j=b+1}^{n} \varepsilon\left(d, X_j, Y_j\right) \to_p 0$ (Proposition 13 in [1]) and $\sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right) \to 0$ (Proposition 14 in [1], Condition (12)), implying that

$$\lim_{d\to\infty} \mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)\right] = \sum_{j=1}^{b} \varepsilon\left(X_j, Y_j\right).$$

Therefore,

$$\begin{aligned}
&\lim_{d\to\infty} \mathrm{P}\left(\left|v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right) - \sum_{j=1}^{b} \varepsilon\left(X_j, Y_j\right)\right| \geq \epsilon\right) \\
&= \lim_{d\to\infty} \mathrm{E}_{\mathbf{Y}^{(n)},\mathbf{X}^{(d)}}\left[\mathrm{P}_{\mathbf{Y}^{(d-n)}}\left(\left|v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right) - \sum_{j=1}^{b} \varepsilon\left(X_j, Y_j\right)\right| \geq \epsilon\right)\right] \\
&\leq \lim_{d\to\infty} \frac{1}{\epsilon^2} \mathrm{E}_{\mathbf{Y}^{(n)},\mathbf{X}^{(d)}}\left[\mathrm{Var}_{\mathbf{Y}^{(d-n)}}\left(v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)\right)\right] \\
&= \frac{1}{\epsilon^2} \lim_{d\to\infty} \sum_{i=1}^{m} \mathrm{E}\left[R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)\right] = 0,
\end{aligned}$$

where the inequality has been obtained by applying Chebychev's inequality and the last equality has been obtained from the conditional distribution in (25).

For the first part of the theorem, since $1 \wedge e^{v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)}$ is a continuous and bounded function, we then conclude that $\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)}\right] \to_p \mathrm{E}_{\mathbf{Y}^{(b)-},\mathbf{X}^{(b)-}}\left[1 \wedge \prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)}\right]$.

For the second part of the theorem, it suffices to use the fact that $\mathrm{E}\left[1 \wedge e^{v\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)}\right] \to \mathrm{E}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left[1 \wedge \prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)}\right]$ along with the following decomposition

$$\mathrm{E}\left[1 \wedge \prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)}\right] = \mathrm{P}\left(\prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)} > 1\right) + \mathrm{E}\left[\prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)}; \prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)} < 1\right]$$

and Proposition 15 to conclude that

$$\left|\mathrm{E}\left[1 \wedge e^{v\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)}\right] - 2\mathrm{P}\left(\prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)} > 1\right)\right| \to 0 \quad \text{as } d \to \infty.$$

7.4 Proof of Theorems 5 and 6

The proof of Theorem 5 is almost identical to the proof of Theorem 1. For the discrete-time limit, it suffice to use Lemma 14 instead of Lemma 11 to achieve the desired conclusion.

For the continuous-time limit, the proof also stays the same but a modification is needed in Lemmas 12 and 13. That is, the body of these proofs remains unchanged, but we use the conditional distribution for $v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ developed in (31) (Section 10) rather than that developed in Section 9.1 in (25) to find the appropriate speed measure. The results presented in Theorem 6 are then obtain by letting $E_R = 0$ in Theorem 5.

7.5 Proof of Theorem 8

The main distinction with the proofs of Theorems 1 and 3 will arise when facing an equation involving any one of the m different groups having an infinite number of components in the limit. Since the constant terms of the last $d-n$ scaling terms are now random, it is impossible to factorize the scaling terms of components belonging to the same group. To overcome this difficulty, we use changes of variables and conditional expectations. Instead of carrying the term $\theta_{n+i}^2(d) = d^{\gamma_i}/K_{n+i}$ which used to be common to all X_j, $j \in \mathcal{J}(i,d)$, we thus carry $\mathrm{E}\left[\theta_{n+i}^2(d)\right] = b_i d^{\gamma_i}$ from where the change in the formula for E_R. Apart from this adjustment, the proof can be carried as usual.

A good example of the necessary adjustment is

$$\begin{aligned}
&\mathrm{E}\left[\sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d}{dX_j}\log\theta_j(d)\, f\left(\theta_j(d)\, X_j\right)\right)^2\right] \\
&= \sum_{i=1}^{m}\mathrm{E}\left[\sum_{j\in\mathcal{J}(i,d)}\int\left(\frac{f'\left(\theta_j(d)\,x_j\right)}{f\left(\theta_j(d)\,x_j\right)}\right)^2\theta_j(d)\, f\left(\theta_j(d)\,x_j\right)dx_j\right] \\
&= \sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}\mathrm{E}\left[\theta_j^2(d)\right]\int\left(\frac{f'(x)}{f(x)}\right)^2 f(x)\,dx \\
&= \sum_{i=1}^{m} c\left(\mathcal{J}(i,d)\right) b_i d^{\gamma_i}\,\mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right].
\end{aligned}$$

7.6 Proof of Theorem 9

Although an elaborate notation is necessary due to the general form of the functions $c\left(\mathcal{J}(i,d)\right)$, $i = 1,\ldots,m$ and $\theta_j(d)$, $j = 1,\ldots,d$ the essence of the proof is preserved. The demonstration is however somehow altered by the possibility for the scaling terms within a given group $i \in \{1,\ldots,m\}$ to assume different functions of the dimension, provided they are of the same order. To deal with this characteristic of the model, we let

$$\theta_j(d) = K_j^{-1/2}\theta_i'(d)\,\frac{\theta_j^*(d)}{\theta_i'(d)},$$

where $\theta_j^*(d)$ is implicitly defined. This representation is used to allow factoring a common function of d for all components of a same group. Specifically, the proof is carried as before but instead of factoring the term $\theta_{n+i}^2(d)$ as in Theorem 1, we factor $b_i\theta_i'(d)$. We are then left with the ratio but since $\lim_{d\to\infty}\theta_j^*(d)/\theta_i'(d) = 1$, the rest of the proof can be repeated with minor adjustments.

8 Approximate Generator

We prove a result stating that the generator of the RWM algorithm is asymptotically equivalent to the generator of a Metropolis-Hastings algorithm with a distinct acceptance rule.

Lemma 10. *For any function $h \in C_c^\infty$, let*

$$\widetilde{G}h(d, X_{i^*}) = \mathrm{E}_{\mathbf{Y}^{(d)},\mathbf{X}^{(d)-}}\left[\left(h(Y_{i^*}) - h(X_{i^*})\right)\left(1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right)\right],$$

with

$$\begin{aligned} v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right) &= \sum_{j=1}^{n}\varepsilon(d, X_j, Y_j) + \sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}\frac{d}{dX_j}\log f(\theta_j(d)X_j)(Y_j - X_j) \\ &\quad - \frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right). \end{aligned} \tag{20}$$

Here,

$$\varepsilon(d, X_j, Y_j) = \log\frac{f(\theta_j(d)Y_j)}{f(\theta_j(d)X_j)}$$

and for $i = 1,\ldots,m$

$$R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right) = \frac{1}{d^{\lambda_1}}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d}{dX_j}\log\theta_j(d)f(\theta_j(d)X_j)\right)^2, \tag{21}$$

where $\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}$ is the vector containing the random variables $\{X_j, j\in\mathcal{J}(i,d)\}$.

Then if $\lambda_1 = 0$, we have

$$\lim_{d\to\infty}\mathrm{E}\left[\left|Gh(d, X_{i^*}) - \widetilde{G}h(d, X_{i^*})\right|\right] = 0.$$

Proof. The generator of the RWM algorithm is given by

$$\begin{aligned} Gh(d, X_{i^*}) &= \mathrm{E}_{\mathbf{Y}^{(d)},\mathbf{X}^{(d)-}}\left[\left(h(Y_{i^*}) - h(X_{i^*})\right)\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right)\right] \\ &= \mathrm{E}_{Y_{i^*}}\left[\left(h(Y_{i^*}) - h(X_{i^*})\right)\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right]\right]. \end{aligned}$$

We first concentrate on the inner expectation. Using properties of the log function, we get

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right]\\
&\quad=\ \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\exp\left\{\sum_{j=1}^{n}\log\frac{f\left(\theta_j\left(d\right)Y_j\right)}{f\left(\theta_j\left(d\right)X_j\right)}+\sum_{j=n+1}^{d}\log\frac{f\left(\theta_j\left(d\right)Y_j\right)}{f\left(\theta_j\left(d\right)X_j\right)}\right\}\right]\\
&\quad=\ \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\exp\left\{\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right)+\sum_{j=n+1}^{d}\left(\log f\left(\theta_j\left(d\right)Y_j\right)-\log f\left(\theta_j\left(d\right)X_j\right)\right)\right\}\right].
\end{aligned}$$

We now write the difference of the log functions as a Taylor expansion with three terms to obtain

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right]\\
&\quad=\ \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\exp\left\{\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right)+\sum_{j=n+1}^{d}\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)\right.\right.\\
&\qquad\left.\left.+\sum_{j=n+1}^{d}\frac{1}{2}\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)^2+\sum_{j=n+1}^{d}\frac{1}{6}\frac{d^3}{dU_j^3}\log f\left(\theta_j\left(d\right)U_j\right)\left(Y_j-X_j\right)^3\right\}\right],
\end{aligned}$$

for some $U_j\in\left(X_j,Y_j\right)$ or $\left(Y_j,X_j\right)$.

In order to compare the term in the exponential function with $v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$, it will be convenient to group the last $d-n$ components according to their scaling term. We have

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right]\\
&\quad=\ \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\exp\left\{\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right)+\sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}^{d}\left[\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)\right.\right.\right.\\
&\qquad\left.\left.\left.+\frac{1}{2}\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)^2+\frac{1}{6}\frac{d^3}{dU_j^3}\log f\left(\theta_j\left(d\right)U_j\right)\left(Y_j-X_j\right)^3\right]\right\}\right]. \qquad (22)
\end{aligned}$$

Before verifying if $\widetilde{G}h\left(d,X_{i^*}\right)$ is asymptotically equivalent to $Gh\left(d,X_{i^*}\right)$, we shall find an upper bound on the difference between the original inner expectation and the new acceptance rule involving $v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$. By the triangle inequality, we have

$$\begin{aligned}
&\left|\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right]-\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\right|\\
&\quad\leq\ \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\left\{1\wedge\exp\left(\log\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right)\right\}-\left\{1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right\}\right|\right].
\end{aligned}$$

By the Lipschitz property of the function $1 \wedge e^x$ and noticing that the first two terms of the function $v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ cancel out with the first two terms of the exponential term in (22), we obtain

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\left\{1 \wedge \exp\left(\log \frac{\pi\left(d, \mathbf{Y}^{(d)}\right)}{\pi\left(d, \mathbf{X}^{(d)}\right)}\right)\right\} - \left\{1 \wedge e^{v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)}\right\}\right|\right] \\
&\quad \leq \; \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\sum_{i=1}^{m} \sum_{j \in \mathcal{J}(i,d)} \left(\frac{1}{2}\frac{d^2}{dX_j^2} \log f\left(\theta_j\left(d\right) X_j\right)\left(Y_j - X_j\right)^2 + \frac{\ell^2}{2d^{\lambda_1}}\left(\frac{d}{dX_j} \log f\left(\theta_j\left(d\right) X_j\right)\right)^2\right)\right.\right. \\
&\qquad \left.\left. + \frac{1}{6}\sum_{i=1}^{m} \sum_{j \in \mathcal{J}(i,d)} \frac{d^3}{dU_j^3} \log f\left(\theta_j\left(d\right) U_j\right)\left(Y_j - X_j\right)^3\right|\right].
\end{aligned}$$

Noticing that the first double summation forms the random variables $W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)$'s of Lemma 6 in [1] we find

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\left\{1 \wedge \exp\left(\log \frac{\pi\left(d, \mathbf{Y}^{(d)}\right)}{\pi\left(d, \mathbf{X}^{(d)}\right)}\right)\right\} - \left\{1 \wedge e^{v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)}\right\}\right|\right] \\
&\quad \leq \; \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\sum_{i=1}^{m} W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right] \\
&\qquad + \frac{1}{6}\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\sum_{i=1}^{m} \sum_{j \in \mathcal{J}(i,d)} \frac{d^3}{dU_j^3} \log f\left(\theta_j\left(d\right) U_j\right)\left(Y_j - X_j\right)^3\right|\right] \\
&\quad \leq \; \sum_{i=1}^{m} \mathrm{E}\left[\left|W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right] + \sum_{i=1}^{m} c\left(\mathcal{J}\left(i,d\right)\right) \frac{K}{6}\sqrt{\frac{8}{\pi}} \frac{\ell^3}{d^{3\lambda_1/2}} \frac{d^{3\gamma_i/2}}{K_i^{3/2}}. \qquad (23)
\end{aligned}$$

We are now ready to verify the $\mathcal{L}^1$ convergence of the RWM generator to the approximate generator $\widetilde{G}h\left(d, X_{i^*}\right)$. By the triangle's inequality, we find

$$\begin{aligned}
&\mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - \widetilde{G}h\left(d, X_{i^*}\right)\right|\right] \\
&\quad \leq \; \mathrm{E}_{Y_{i^*},X_{i^*}}\left[\left|\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right) \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left(1 \wedge \frac{\pi\left(d, \mathbf{Y}^{(d)}\right)}{\pi\left(d, \mathbf{X}^{(d)}\right)}\right) - \left(1 \wedge e^{v\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)}\right)\right]\right|\right]
\end{aligned}$$

From (23), we observe

$$\begin{aligned}
\mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - \widetilde{G}h\left(d, X_{i^*}\right)\right|\right] \; \leq \; & \sum_{i=1}^{m} \mathrm{E}\left[\left|W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right] \mathrm{E}\left[\left|h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right|\right] \\
& + \sum_{i=1}^{m} \frac{K}{6} c\left(\mathcal{J}\left(i,d\right)\right) \sqrt{\frac{8}{\pi}} \frac{\ell^3}{d^{3\lambda_1/2}} \frac{d^{3\gamma_i/2}}{K_i^{3/2}} \mathrm{E}\left[\left|h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right|\right].
\end{aligned}$$

Because $h \in C_c^\infty$, there exists a constant such that $\left|h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right| \leq K$ and thus Lemma 6 in [1] implies the previous expectation converges to 0 as the dimension goes to infinity. □

9 Acceptance Rule, Volatility and Drift - Theorem 1

9.1 Convergence to the Modified Acceptance Rule

This section determines the limiting acceptance rule for the case where the RWM algorithm is not sped up by any time factor, and where the first b components of the target are not solely ruling the chain.

Lemma 11. *If $\lambda_1 = 0$ and if Conditions (5) and (9) are satisfied, then*

$$\mathrm{E}_{X_{i^*},Y_{i^*}}\left[\left|\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] - \alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right)\right|\right] \to 0 \quad \text{as} \quad d \to \infty,$$

with $v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$ as in (20) and $\alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right)$ as in (10).

Proof. We first use conditional expectations to obtain

$$\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] = \mathrm{E}_{\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}}\left[\mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\right]. \qquad (24)$$

In order to evaluate the inner expectation, we need to find the distribution of the function $v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$ conditional on $\mathbf{Y}^{(n)}$ and $\mathbf{X}^{(d)}$. Since $\left(Y_j - X_j\right)|X_j$ are independent and normally distributed with mean 0 and variance ℓ^2 for $j = n+1,\ldots,d$, we have

$$\begin{aligned} & v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)\Big|\,\mathbf{Y}^{(n)},\mathbf{X}^{(d)} \\ \sim\ & N\left(\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right),\ \ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)\right). \end{aligned} \qquad (25)$$

Applying Proposition 2.4 in [15], we can express the inner expectation in (24) in terms of $\Phi(\cdot)$, the *cdf* of a standard normal random variable

$$\begin{aligned} & \mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] \\ =\ & \Phi\left(\frac{\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)}}\right) \\ & + \exp\left(\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right)\right)\Phi\left(\frac{-\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)}}\right). \end{aligned}$$

We need to study the convergence of every term included in previous function. Condition (5) implies that $\theta_1^{-2}(d)$ is the asymptotically smallest scaling term and along with $\lambda_1 = 0$, this means that the fastest converging component has an $O(1)$ scaling term.

However there might be a finite number of other components also having an $O(1)$ scaling term. Recall that b is the number of such components and in the present case is defined as $b = \max(j \in \{1, \ldots, n\}; \lambda_j = 0)$. It is thus pointless to study the convergence of these b variables since they are independent of d. However, we can study the convergence of the other $n - b$ components and from Proposition 13 in [1] we know that $\varepsilon(d, X_j, Y_j) \rightarrow_p 0$ for $j = b+1, \ldots, n$ since $\lambda_j < 0$. Similarly, we can use Proposition 14 in [1] and Condition (9) to conclude that $\sum_{i=1}^{m} R_i \left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right) \rightarrow_p E_R > 0$, with E_R as in (11).

Using Slutsky's and the Continuous Mapping Theorems, we conclude that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] \rightarrow_p \\
&\quad \Phi\left(\frac{\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \frac{\ell^2}{2} E_R}{\sqrt{\ell^2 E_R}}\right) + \exp\left(\sum_{j=1}^{b} \varepsilon(X_j, Y_j)\right) \Phi\left(\frac{-\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \frac{\ell^2}{2} E_R}{\sqrt{\ell^2 E_R}}\right) \\
&\quad \equiv M\left(\ell^2, \mathbf{Y}^{(b)}, \mathbf{X}^{(b)}\right).
\end{aligned}$$

Using the triangle's inequality, we obtain

$$\begin{aligned}
&\mathrm{E}_{X_{i^*},Y_{i^*}}\left[\left|\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] - \alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right)\right|\right] \\
&\quad \leq \mathrm{E}_{\mathbf{Y}^{(n)},\mathbf{X}^{(d)}}\left[\left|\mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] - M\left(\ell^2, \mathbf{Y}^{(b)}, \mathbf{X}^{(b)}\right)\right|\right].
\end{aligned}$$

Since each term in the absolute value is positive and bounded by 1 and since the difference between them converges to 0 in probability, we can use the Bounded Convergence Theorem ([5], [10], [19]) to conclude that the previous expression converges to 0. □

9.2 Simplified Expression for the Approximate Volatility

Lemma 8 in [1] established that $\ell^2 \mathrm{E}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)}\right]$ is an asymptotically valid volatility term for the continuous-time generator $\widetilde{G}h(d, X_{i^*})$ in (19). We now wish to find a convenient expression for the volatility term of (19) as $d \to \infty$. This is achieved in the following lemma.

Lemma 12. *If Conditions (5) and (9) are satisfied, then*

$$\lim_{d\to\infty}\left|\mathrm{E}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)}\right] - 2\mathrm{E}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left[\Phi\left(\frac{\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right]\right| = 0,$$

where

$$\begin{aligned}
v\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right) &= \sum_{j=1, j\neq i^*}^{n} \varepsilon(d, X_j, Y_j) + \sum_{i=1}^{m} \sum_{j\in\mathcal{J}(i,d), j\neq i^*} \frac{d}{dX_j} \log f(\theta_j(d) X_j)(Y_j - X_j) \\
&\quad - \frac{\ell^2}{2} \sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right) \qquad (26)
\end{aligned}$$

and E_R is as in (11).

Proof. Note that (26) is similar to (20), excepted that the i^*-th component is now excluded from the definition. Therefore, it is easily seen from the proof of Lemma 11 that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d-n)-}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)}\right] \\
&= \Phi\left(\frac{\sum_{j=1,j\neq i^*}^{n} \varepsilon\left(d,X_j,Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2 \sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}}\right) \\
&\quad + \exp\left(\sum_{j=1,j\neq i^*}^{n} \varepsilon\left(d,X_j,Y_j\right)\right) \Phi\left(\frac{-\sum_{j=1,j\neq i^*}^{n} \varepsilon\left(d,X_j,Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2 \sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}}\right).
\end{aligned}$$

From Proposition 12 in [1], both terms of the sum have the same expectation and the previous expression thus simplifies to

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)}\right] \\
&= 2\mathrm{E}_{\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}}\left[\Phi\left(\frac{\sum_{j=1,j\neq i^*}^{n} \varepsilon\left(d,X_j,Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2 \sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}}\right)\right].
\end{aligned}$$

By Proposition 13 in [1], we have $\varepsilon\left(d,X_j,Y_j\right) \to_p 0$ since $\lambda_j < \lambda_1$ for $j = b+1,\ldots,n$. From Proposition 14 in [1], we also know that $\sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right) \to_p E_R$, where E_R is as in (11) and is strictly positive by Condition (9). Applying Slutsky's and the Continuous Mapping Theorems thus yield

$$\begin{aligned}
&\Phi\left(\frac{\sum_{j=1,j\neq i^*}^{n} \varepsilon\left(d,X_j,Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2 \sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}}\right) \to_p \\
&\qquad \Phi\left(\frac{\sum_{j=1,j\neq i^*}^{b} \varepsilon\left(X_j,Y_j\right) - \frac{\ell^2}{2}E_R}{\sqrt{\ell^2 E_R}}\right). \qquad (27)
\end{aligned}$$

Using the Bounded Convergence Theorem concludes the proof of the lemma. □

9.3 Simplified Expression for the Approximate Drift

Lemma 10 in [1] introduced a drift term that is asymptotically equivalent to the drift term of the continuous-time generator $\widetilde{G}h\left(d,X_{i^*}\right)$ in (19). The goal of the following lemma is to determine a simple expression for this new drift term as $d \to \infty$.

Lemma 13. *If Conditions (5) and (9) are satisfied, then*

$$\lim_{d\to\infty}\left|\mathrm{E}\left[e^{v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)};v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)<0\right]-\mathrm{E}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left[\Phi\left(\frac{\sum_{j=1}^{b}\varepsilon\left(X_j,Y_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\right]\right|=0,$$

where $v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)$ *and* E_R *are as in (26) and (11) respectively.*

Proof. The proof of this result is similar to that of Lemma 12 and for this reason, we shall not repeat every detail. Since we know the conditional distribution of $v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\big|\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}$, we can use Proposition 2.4 in [15] to obtain

$$\begin{aligned}&\mathrm{E}_{\mathbf{Y}^{(d-n)-}}\left[e^{v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)};v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)<0\right]\\ &=\exp\left(\sum_{j=1,j\neq i^*}^{n}\varepsilon\left(d,X_j,Y_j\right)\right)\Phi\left(\frac{-\sum_{j=1,j\neq i^*}^{n}\varepsilon\left(d,X_j,Y_j\right)-\frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}}\right).\end{aligned}$$

From Proposition 12 in [1], the unconditional expectation simplifies to

$$\begin{aligned}&\mathrm{E}\left[e^{v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)};v\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)<0\right]\\ &=\mathrm{E}\left[\Phi\left(\frac{\sum_{j=1,j\neq i^*}^{n}\varepsilon\left(d,X_j,Y_j\right)-\frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}}\right)\right].\end{aligned}$$

Using (27) along with the Bounded Convergence Theorem completes the proof of the lemma. □

10 Acceptance Rule - Theorem 5

We now consider the case where the target distribution is normally distributed. The aim of this section is to determine the acceptance rule for the limiting Metropolis-Hastings algorithm when one of the first b components is studied.

Lemma 14. *If* $f(x)=(2\pi)^{-1/2}\exp\left(x^2/2\right)$, $\lambda_1=0$ *and if Conditions (5) and (9) are satisfied, then*

$$\mathrm{E}_{X_{i^*},Y_{i^*}}\left[\left|\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]-\alpha\left(\ell^2,X_{i^*},Y_{i^*}\right)\right|\right]\to 0\quad \text{as } d\to\infty,$$

with $v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$ *as in (20) and* $\alpha\left(\ell^2,X_{i^*},Y_{i^*}\right)$ *as in (13).*

Proof. We first use conditional expectations to obtain

$$\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] = \mathrm{E}_{\mathbf{Y}^{(n)-},\mathbf{X}^{(d-n)}}\left[\mathrm{E}_{\mathbf{Y}^{(d-n)},\mathbf{X}^{(n)-}}\left[1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\right]. \quad (28)$$

In order to evaluate the inner expectation, we need to find the distribution of the function $v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$ conditional on $\mathbf{Y}^{(n)}$, $\mathbf{X}^{(d-n)}$ and X_{i^*}. First of all, we find

$$\left.\left(\sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)-\frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)\right)\right|\mathbf{X}^{(d-n)}$$
$$\sim N\left(-\frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right),\ \ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)\right). \quad (29)$$

Due to the normal form of the target components, it is possible to find the distribution of $\sum_{j=1,j\neq i^*}^{n}\varepsilon\left(d,X_j,Y_j\right)$. Since $Y_j\left|X_j\right.\sim N\left(X_j,\ell^2\right)$ and $X_j\sim N\left(0,K_j/d^{\lambda_j}\right)$, we obtain

$$Y_j=X_j+U_j,$$

where $U_j\sim N\left(0,\ell^2\right)$ is independent of X_j for $j=1,\ldots,n$. We then have

$$\begin{aligned}\varepsilon\left(d,X_j,Y_j\right) &= \frac{d^{\lambda_j}}{2K_j}\left(X_j^2-\left(X_j+U_j\right)^2\right)=-\frac{d^{\lambda_j}}{2K_j}\left(2X_jU_j+U_j^2\right)\\ &= -\left(\ell\widetilde{X}_j\widetilde{U}_j+\frac{\ell^2}{2}\widetilde{U}_j^2\right),\end{aligned}$$

where $\widetilde{X}_j\sim N\left(0,1\right)$ and $\widetilde{U}_j\sim N\left(0,d^{\lambda_j}/K_j\right)$.

By independence between $\widetilde{X}_j$ and $\widetilde{U}_j$ we have $\widetilde{X}_j\left|\widetilde{U}_j\right.\sim N\left(0,1\right)$, and hence

$$\left.\left(\frac{\ell^2}{2}\widetilde{U}_j^2+\ell\widetilde{U}_j\widetilde{X}_j\right)\right|\widetilde{U}_j\sim N\left(\frac{\ell^2}{2}\widetilde{U}_j^2,\ell^2\widetilde{U}_j^2\right). \quad (30)$$

Combining (29) and (30), we obtain the conditional distribution

$$\begin{aligned}&v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)\left|\mathbf{Y}^{(n)},X_{i^*},\mathbf{X}^{(d-n)}\right.\\ &\sim\ N\left(\varepsilon\left(X_{i^*},Y_{i^*}\right)-\frac{\ell^2}{2}\left(\sum_{j=1,j\neq i^*}^{n}\widetilde{U}_j^2+\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)\right),\right.\\ &\qquad\left.\ell^2\left(\sum_{j=1,j\neq i^*}^{n}\widetilde{U}_j^2+\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)\right)\right).\end{aligned} \quad (31)$$

Applying Proposition 2.4 in [15], we can express the inner expectation in (28) in terms of

$\Phi\left(\cdot\right)$

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d-n)},\mathbf{X}^{(n)-}}\left[1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\\
&\quad=\ \Phi\left(\frac{\varepsilon\left(X_{i^*},Y_{i^*}\right)-\frac{\ell^2}{2}\left(\sum_{j=1,j\neq i^*}^{n}\widetilde{U}_j^2+\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right)\right)}{\sqrt{\ell^2\left(\sum_{j=1,j\neq i^*}^{n}\widetilde{U}_j^2+\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right)\right)}}\right)\\
&\qquad\quad+\exp\left(\varepsilon\left(X_{i^*},Y_{i^*}\right)\right)\Phi\left(\frac{-\varepsilon\left(X_{i^*},Y_{i^*}\right)-\frac{\ell^2}{2}\left(\sum_{j=1,j\neq i^*}^{n}\widetilde{U}_j^2+\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right)\right)}{\sqrt{\ell^2\left(\sum_{j=1,j\neq i^*}^{n}\widetilde{U}_j^2+\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right)\right)}}\right).
\end{aligned}$$

We need to study the convergence of every term appearing in the preceding equation. From Proposition 14 in [1], we know that $\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right)\to_p E_R$. Because the variance of the components $\widetilde{U}_1,\ldots,\widetilde{U}_b$ does not vary with the dimension of the target, it is not relevant to talk about convergence for these variables. We can however study the convergence of the components $\widetilde{U}_{b+1},\ldots,\widetilde{U}_n$. Using Chebychev's inequality, we have for all $\epsilon>0$

$$\mathrm{P}\left(\left|\widetilde{U}_j\right|\geq\varepsilon\right)\ \leq\ \frac{\mathrm{Var}\left(\widetilde{U}_j\right)}{\epsilon^2}=\frac{1}{\epsilon^2}\frac{d^{\lambda_j}}{K_j}\to 0\ \text{ as } d\to\infty,$$

since $\lambda_j<\lambda_1=0$ for $j=b+1,\ldots,n$. Therefore, $\widetilde{U}_j\to_p 0$ for $j=b+1,\ldots,n$.

Using Slutsky's Theorem, the Continuous Mapping Theorem and the fact that $\widetilde{U}_j^2=\chi_j^2/K_j$ with χ_j^2, $j=1,\ldots,b$ distributed as independent chi square random variables with 1 degree of freedom, we conclude that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d-n)},\mathbf{X}^{(n)-}}\left[1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\to_p\Phi\left(\frac{\varepsilon\left(X_{i^*},Y_{i^*}\right)-\frac{\ell^2}{2}\left(\sum_{j=1,j\neq i^*}^{b}\frac{\chi_j^2}{K_j}+E_R\right)}{\sqrt{\ell^2\left(\sum_{j=1,j\neq i^*}^{b}\frac{\chi_j^2}{K_j}+E_R\right)}}\right)\\
&\quad+\exp\left(\varepsilon\left(X_{i^*},Y_{i^*}\right)\right)\Phi\left(\frac{-\varepsilon\left(X_{i^*},Y_{i^*}\right)-\frac{\ell^2}{2}\left(\sum_{j=1,j\neq i^*}^{b}\frac{\chi_j^2}{K_j}+E_R\right)}{\sqrt{\ell^2\left(\sum_{j=1,j\neq i^*}^{b}\frac{\chi_j^2}{K_j}+E_R\right)}}\right)\equiv M\left(\ell^2,X_{i^*},Y_{i^*},\left(\widetilde{\mathbf{U}}^2\right)^{(b)-}\right).
\end{aligned}$$

Using the triangle's inequality, we therefore obtain

$$\begin{aligned}
&\mathrm{E}_{X_{i^*},Y_{i^*}}\left[\left|\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]-\alpha\left(\ell^2,X_{i^*},Y_{i^*}\right)\right|\right]\\
&\quad\leq\mathrm{E}_{Y_{i^*},\left(\widetilde{\mathbf{U}}^2\right)^{(n)-},X_{i^*},\mathbf{X}^{(d-n)}}\left[\left|\mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[1\wedge e^{v\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]-M\left(\ell^2,X_{i^*},Y_{i^*},\left(\widetilde{\mathbf{U}}^2\right)^{(b)-}\right)\right|\right].
\end{aligned}$$

Since each term in the absolute value is positive and bounded by 1 and since the difference between them converges to 0 in probability, we can use the Bounded Convergence Theorem to conclude that the previous expression converges to 0. □

11 Discussion

In this paper, we have established a necessary and sufficient condition under which the AOAR developed for *iid* target distributions in [15] is not optimal, and even inefficient in some cases. The target density considered was an extension of the *iid* model where each component had the same density f, but where each of them was scaled according to $\theta_j^{-2}(d)$, $j = 1, \ldots, d$ possibly depending on the dimension of the target distribution. We also introduced an equation from which the appropriate AOAR can be numerically solved for optimal performance of the algorithm. These results are the first to admit limiting processes and AOARs that are different from those find by [15] for RWM algorithms. This work should then act as a warning for practitioners, who should be aware that the usual 0.234 might be inefficient even with seemingly regular targets.

An intuitive explanation for this phenomenon resides of course in the necessary and sufficient condition, which ensures that there exists a finite number of components converging significantly faster than the others. In particular, as the dimension of the target increases, the effect of these components on the performance of the algorithm remains significant, thus drawing down the AOAR. A surprising characteristic of RWM algorithms with Gaussian proposal distribution is thus that their AOAR is at most 0.234.

A particular case where the well-known 0.234 was shown not to be optimal was when considering widely used normal hierarchical models. This might seem surprising, given that multivariate normal distribution were believed to adopt a conventional limiting behavior. In some cases where the proposal scaling was governed by a finite number of target components, an infinite value of the proposal scaling was suggested. In such cases, the convergence speed of the algorithm was extremely slow no matter what proposal scaling was adopted, and we concluded that using inhomogeneous proposal distributions would be far more efficient.

Finally, it is worth mentioning that although asymptotic, the results presented in this paper work well in relatively small dimensions. In addition, the results provided about the optimal form for the proposal variance as a function of d turn out to be useful guidelines in practice. The results of this paper are then relatively easy to apply for practitioners, as it suffices to verify which conditions are satisfied, and then numerically optimize the appropriate equation to find the optimal value for ℓ.

Appendix

The following proposition demonstrates the equivalence between an expectation and a probability, and is useful to prove Theorem 3.

Proposition 15. *Let* $\mathbf{X}_j$ *be distributed according to the density* $\theta_j f(\theta_j x_j)$ *for* $j = 1, \ldots, d$

and also let $\mathbf{Y}^{(d)} \left| \mathbf{X}^{(d)} \right. \sim N\left(\mathbf{X}^{(d)}, \sigma^2(d)\, I_{d \times d}\right)$. *Then, we have*

$$\mathrm{E}\left[\prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)}; \prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)} < 1\right] \;=\; \mathrm{P}\left(\prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)} > 1\right).$$

Proof. Developing the LHS leads to

$$\begin{aligned}
&\mathrm{E}\left[\prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)}; \prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)} < 1\right] \\
&= \int\!\!\int \mathbf{1}_{\left(\prod_{j=1}^{b} \frac{f(\theta_j y_j)}{f(\theta_j x_j)} < 1\right)} \prod_{j=1}^{b} \frac{f(\theta_j y_j)}{f(\theta_j x_j)} C e^{-\frac{1}{2\sigma^2(d)} \sum_{j=1}^{b} (y_j - x_j)^2} \prod_{j=1}^{b} \theta_j f(\theta_j x_j)\, d\mathbf{y}^{(b)} d\mathbf{x}^{(b)} \\
&= \int\!\!\int \mathbf{1}_{\left(\prod_{j=1}^{b} \frac{f(\theta_j x_j)}{f(\theta_j y_j)} > 1\right)} C e^{-\frac{1}{2\sigma^2(d)} \sum_{j=1}^{b} (x_j - y_j)^2} \prod_{j=1}^{b} \theta_j f(\theta_j y_j)\, d\mathbf{y}^{(b)} d\mathbf{x}^{(b)},
\end{aligned}$$

where C is a constant term. Using Fubini's Theorem and swapping y_j and x_j yield the desired result. □

Acknowledgments

This work is part of my Ph.D. thesis and has been supported by NSERC of Canada. Special thanks are due to my supervisor, Professor Jeffrey S. Rosenthal, without who this work would not have been complete. His expertise, guidance and encouragements have been precious throughout my studies. I also acknowledge useful conversations with Professor Gareth O. Roberts.

References

[1] Bédard, M. (2006). Weak Convergence of Metropolis Algorithms for Non-*iid* Target Distributions. Submitted for publication.

[2] Bédard, M. (2006). Efficient Sampling using Metropolis Algorithms: Applications of Optimal Scaling Results. Submitted for publication.

[3] Besag, J., Green, P.J. (1993). Spatial statistics and Bayesian computation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **55**, 25-38.

[4] Besag, J., Green, P.J., Higdon, D., Mergensen, K. (1995). Bayesian computation adn stochastic systems. *Statist. Sci.* **10**, 3-66.

[5] Billingsley, P. (1995). *Probability and Measure, 3rd ed.* John Wiley & Sons, New York.

[6] Breyer, L.A., Piccioni, M., Scarlatti, S. (2002). Optimal Scaling of MALA for Nonlinear Regression. *Ann. Appl. Probab.* **14**, 1479-1505.

[7] Breyer, L.A., Roberts, G.O. (2000). From Metropolis to Diffusions: Gibbs States and Optimal Scaling. *Stochastic Process. Appl.* **90**, 181-206.

[8] Christensen, O.F., Roberts, G.O., Rosenthal, J.S. (2003). Scaling Limits for the Transient Phase of Local Metropolis-Hastings Algorithms. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 253-69.

[9] Ethier, S.N., Kurtz, T.G. (1986). *Markov Processes: Characterization and Convergence.* Wiley.

[10] Grimmett, G.R., Stirzaker, D.R. (1992). *Probability and Random Processes.* Oxford.

[11] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* **57**, 97-109.

[12] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-92.

[13] Neal, P., Roberts, G.O. (2004). Optimal Scaling for Partially Updating MCMC Algorithms. To appear in *Ann. Appl. Probab.*

[14] Peskun, P.H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika.* **60**, 607-12.

[15] Roberts, G.O., Gelman, A., Gilks, W.R. (1997). Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *Ann. Appl. Probab.* **7**, 110-20.

[16] Roberts, G.O., Rosenthal, J.S. (1998). Optimal Scaling of Discrete Approximations to Langevin Diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**, 255-68.

[17] Roberts, G.O., Rosenthal, J.S. (2001). Optimal Scaling for various Metropolis-Hastings algorithms. *Statist. Sci.* **16**, 351-67.

[18] Roberts, G.O., Rosenthal, J.S. (2004). General State Space Markov Chains and MCMC Algorithms. *Probab. Surveys* **1**, 20-71.

[19] Rosenthal, J.S. (2000). *A First Look at Rigorous Probability Theory.* World Scientific, Singapore.