Optimal Acceptance Rates for Metropolis Algorithms: Moving Beyond 0.234

Mylène Bédard [*]

Abstract

Recent optimal scaling theory has produced a condition for the asymptotically optimal acceptance rate of Metropolis algorithms to be the well-known 0.234 when applied to certain multidimensional target distributions. These d-dimensional target distributions are formed of independent components, each of which is scaled according to its own function of d. We show that when the 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. Specifically, 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 new speed measure v, except in particular cases where it converges to a one-dimensional Metropolis algorithm with acceptance rule α^*. We also discuss the use of inhomogeneous proposals, which might reveal essential in specific cases.

Keywords: Weak convergence, optimal scaling, Langevin diffusion, generator, Markov chain Monte Carlo

1 Introduction

This paper contains theoretical results aiming to optimize the efficiency of Metropolis algorithms ([9], [8]), which are widely used in various domains to generate data from highly complex probability distributions. Although versatile and easy to implement, Metropolis algorithms often suffer from slow convergence; achieving the optimal convergence speed for the algorithm thus becomes an important, and sometimes even a crucial matter. A way to reach some level of optimality in the performance of the algorithm is through the selection of the scaling parameter of the proposal distribution (see [3] and [4]).

Roberts, Gelman, and Gilks (1997) have been the first authors to publish theoretical results about the optimal scaling problem for Metropolis algorithms with Gaussian proposals (see [12]). They established that for high-dimensional target distributions formed of independent and identically distributed (*iid*) components, the acceptance rate optimizing the efficiency of the process approaches 0.234. This provides an exceptionally simple guideline to be applied

[*] Département de mathématiques et de statistique, Université de Montréal, Montréal, Canada, H3C 3J7. Email: bedard@dms.umontreal.ca   January 2006, revised September 2007

MSC 2000 subject classifications. Primary 60F05; secondary 65C40.

by practitioners. Despite the *iid* assumption, this result was believed to hold under various perturbations of the target distribution; other authors then studied diverse extensions of the *iid* target model, and all obtained results corroborating those of [12] (see, for instance, [5], [13], [6], and [10]).

In [1], we recently studied d-dimensional target distributions with independent components, but where each component possesses its own scaling term which might be a function of the dimension of the target distribution, d. This setting significantly differs from the *iid* case since it produces target distributions for which some components might have their probability mass spread out over the real line, or concentrated within a narrow interval of the state space. We provided a necessary and sufficient condition on the scaling terms under which the algorithm admits the same limiting process as in [12], and thus produces the same asymptotically optimal acceptance rate (AOAR), 0.234. This condition roughly ensures that the d target scaling terms are similar to each other. There is however a nontrivial question that remains unanswered: what happens when this condition fails to hold? In particular, is 0.234 still the AOAR and if not, then how can the speed of convergence of the algorithm be optimized?

The present paper aims at answering these questions. Specifically, we consider the exact same target model as in [1], but we now focus on the case where the previous condition is violated i.e., where there is at least one scaling term differing significantly from the others. Consequently, there is now some target components whose density is concentrated in much narrower intervals of the state space than others, and those components inevitably converge faster to their stationary distribution. In [1], the presence of such components was prevented by the condition on the scaling terms since we considered components converging at (more or less) the same rate only. We now prove that the existence of significantly different target components affects the limiting behavior of the algorithm (as $d \to \infty$). In particular, we prove that the acceptance rate 0.234 is not asymptotically optimal and most importantly, we present methods for determining the correct AOAR. This addresses the issue raised in [14]'s Open Problem #3. We shall also see that if our target model includes scaling terms that are extremely small compared to the others, then the optimization problem for solving for the AOAR is ill-posed; this thus calls for the use of inhomogeneous proposal scalings. In general, the well-known 0.234 constitutes an upper bound for the AOAR, which might take any value in [0,0.234] depending on the scaling terms of our target model. This work is the first instance in the literature of limiting processes and AOARs differing from those of [12] (and of [1]) for Metropolis algorithms.

In order to prove the theorems in this paper, it is necessary to determine the limiting behavior (as $d \to \infty$) of every "one-dimensional path" of the generated multidimensional Markov chain. This approach is similar to traditional treatments of optimal scaling issues and for the target model considered here and in [1], it involves a $\mathcal{L}^1$-convergence of the processes' generators argument. There is however one important variation with previous methods: due to the presence of target components converging to their stationary distribution much faster than others, it becomes necessary to study every "one-dimensional path" of the algorithm marginally. In other words, we study each one-dimensional process without possessing any information about the past of the other target components. Although this difference is more of a conceptual notion, it also affects the mathematical approach to a certain extent as we take an extra expectation of the generators involved (this is explained in details at the beginning of Appendix A). The conclusions obtained when dealing with significantly different scaling terms (i.e., when the condition in [1] is violated) and appearing in Sections 3 and 4 greatly differ from the limiting Langevin diffusion obtained in [12] and [1]. In particular, the limiting distribution of the fastest converging components is a Metropolis algorithm with a special acceptance rule; the other components still converge to a Langevin diffusion process, but the speed measure of the process is different and this is what yields smaller AOARs for these targets.

In Section 2, we shall first introduce the target distribution and the Metropolis algorithm, along with a method for determining the optimal form for the proposal scaling as a function of d. The main optimal scaling results shall be presented in Section 3, followed by some extensions in Section 4, and concluding with a discussion in Section 5. The proofs of the various results are presented in Appendix A, and make use of the lemmas of Appendix B.

2 Sampling Using the Metropolis Algorithm

2.1 The Target Distribution

Suppose that we want to generate data from the following d-dimensional product density

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

We impose some regularity conditions on the density f: f is a positive C^2 function, $(\log f\left(x\right))'$ is Lipschitz continuous, $\mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^4\right] = \int_{\mathbf{R}} \left(\frac{f'(x)}{f(x)}\right)^4 f(x)\ dx < \infty$, and similarly $\mathrm{E}\left[\left(\frac{f''(X)}{f(X)}\right)^2\right] < \infty$.

We define the scaling vector $\mathbf{\Theta}\left(d\right) = \left(\theta_1\left(d\right), \ldots, \theta_d\left(d\right)\right)$ i.e., the vector containing the d target scaling terms; although independent, the d target components are thus not identically distributed. Since we are interested in studying the limiting distribution of the path of every individual component forming the Metropolis algorithm as $d \to \infty$, we shall ultimately become interested in an infinite version of the vector $\mathbf{\Theta}\left(d\right)$. For this to make sense, we however have to make some assumptions about the form of this scaling vector.

We shall assume that $\mathbf{\Theta}\left(d\right)$ contains at least $1+m$, but at most $n+m$ different terms, where $n, m \geq 1$; this means that some scaling terms will appear more than once in $\mathbf{\Theta}\left(d\right)$ as d gets larger. Specifically, we suppose that the scaling terms $\theta_1\left(d\right), \ldots, \theta_n\left(d\right)$ are scaling terms that might or might not be all identical, and that are distinct of any other scaling term $\theta_{n+1}, \ldots, \theta_d\left(d\right)$ in $\mathbf{\Theta}\left(d\right)$ for $d \geq n$. The last $d-n$ scaling terms can be divided into m different groups of identical scaling terms; furthermore the larger is d, the greater is the number of scaling terms in each of the m different groups. Accordingly, we can express the scaling vector as follows:

$$\boldsymbol{\Theta}(d)=\Bigg(\theta_1(d),\ldots,\theta_n(d),\theta_{n+1}(d),\ldots,\theta_{n+m}(d),\underbrace{\theta_{n+1}(d),\ldots,\theta_{n+1}(d)}_{c(\mathcal{J}(1,d))-1},\ldots,\underbrace{\theta_{n+m}(d),\ldots,\theta_{n+m}(d)}_{c(\mathcal{J}(m,d))-1}\Bigg),$$

where $c(\mathcal{J}(1,d)),\ldots,c(\mathcal{J}(m,d))$ denote, in a given dimension, the number of scaling terms in each of these m groups.

In fact, more can be said about these functions. A simple way to identify the positions of components possessing a common scaling term is to introduce the mutually exclusive sets $\mathcal{J}(i,d)=\{j\in\{1,\ldots,d\};\theta_j(d)=\theta_{n+i}(d)\}$, $i=1,\ldots,m$. Hence, $\dot{\bigcup}_{i=1}^{m}\mathcal{J}(i,d)=\{n+1,\ldots,d\}$ and the vector $\mathbf{X}^{(d)}_{\mathcal{J}(1,d)}$ (say) contains the target components with a scaling term equal to $\theta_{n+1}(d)$. Because the $\mathcal{J}(i,d)$s do not necessarily possess the same number of elements for finite values of d, we introduce corresponding cardinality functions:

$$c(\mathcal{J}(i,d)) \;=\; \#\{j\in\{1,\ldots,d\};\theta_j(d)=\theta_{n+i}(d)\},\quad i=1,\ldots,m. \tag{2}$$

Here, $c(\mathcal{J}(i,d))$ is some polynomial function of d satisfying $\lim_{d\to\infty}c(\mathcal{J}(i,d))=\infty$.

Having defined the form of the scaling vector $\boldsymbol{\Theta}(d)$, we now need to introduce some assumptions about the form of each scaling term as a function of d. For this task, it shall be convenient to work in terms of $\theta_j^{-2}(d)$s rather than in terms of $\theta_j(d)$s. We let

$$\boldsymbol{\Theta}^{-2}(d)=\left(\frac{K_1}{d^{\lambda_1}},\ldots,\frac{K_n}{d^{\lambda_n}},\frac{K_{n+1}}{d^{\gamma_1}},\ldots,\frac{K_{n+m}}{d^{\gamma_m}},\underbrace{\frac{K_{n+1}}{d^{\gamma_1}},\ldots,\frac{K_{n+1}}{d^{\gamma_1}}}_{c(\mathcal{J}(1,d))-1},\ldots,\underbrace{\frac{K_{n+m}}{d^{\gamma_m}},\ldots,\frac{K_{n+m}}{d^{\gamma_m}}}_{c(\mathcal{J}(m,d))-1}\right), \tag{3}$$

where $\{K_j, j=1,\ldots,n+m\}$ are some positive and finite constant terms. Finally, without loss of generality, we assume $\theta_1^{-2}(d),\ldots,\theta_n^{-2}(d)$ and $\theta_{n+1}^{-2}(d),\ldots,\theta_{n+m}^{-2}(d)$ to be respectively arranged according to an asymptotic increasing order. Equivalently, we have $-\infty<\lambda_n\leq\lambda_{n-1}\leq\ldots\leq\lambda_1<\infty$ and $-\infty<\gamma_m\leq\gamma_{m-1}\leq\ldots\leq\gamma_1<\infty$. We shall refer to both $\theta_j(d)$ and $\theta_j^{-2}(d)$ as target scaling terms.

In [1], we provided a condition on the scaling terms of this same target distribution that ensures an AOAR of 0.234 for the Metropolis algorithm. We now consider the case where there exists at least one scaling term in (3) that is significantly smaller than the others, i.e.

$$\lim_{d\to\infty}\frac{\theta_1^2(d)}{\sum_{j=1}^{d}\theta_j^2(d)}>0, \tag{4}$$

which is the complement of Condition 5, Theorem 1 in [1]. From (3), it is clear that the asymptotically smallest scaling term is either $\theta_1^{-2}(d)$ or $\theta_{n+1}^{-2}(d)$. It is interesting to notice that under the fulfilment of (4) this uncertainty is resolved and K_1/d^{λ_1} is smallest for large d. There might however be more than just one target component possessing a $O(d^{-\lambda_1})$ scaling term, so we let $b=\max(j\in\{1,\ldots,n\};\lambda_j=\lambda_1)$ be the number of such components. This was not the case in [1], where both $\theta_1^{-2}(d)$ and $\theta_{n+1}^{-2}(d)$ were allowed to be the asymptotically smallest scaling term, as long as $\theta_1^{-2}(d)$ was not too small compared to $\theta_{n+1}^{-2}(d)$.

When studying a particular component of the d-dimensional algorithm, it is necessary for the corresponding scaling term to be independent of d to avoid trivial limiting processes. Consequently, if interested in the i^*th component, we set $\theta_{i^*}(d) = 1$; this can be done without loss of generality by applying a linear transformation to the target distribution. This also means that we shall deal with different vectors $\mathbf{\Theta}^{-2}(d)$ when studying different components of the algorithm.

2.2 The Algorithm and Proposal Scaling

To obtain a sample from the target distribution described in Section 2.1 we apply a Metropolis algorithm, which generates a Markov chain $\mathbf{X}^{(d)}(0), \mathbf{X}^{(d)}(1), \ldots$ whose invariant distribution is $\pi\left(d, \mathbf{x}^{(d)}\right)$. The method works as follows. Given the time-t state of the chain, $\mathbf{X}^{(d)}(t)$, we propose a value $\mathbf{Y}^{(d)}(t+1)$; this value is generated from the proposal distribution, taken to be $N\left(\mathbf{X}^{(d)}(t), \sigma^2(d) I_d\right)$, where $\sigma^2(d)$ is a scalar and I_d the d-dimensional identity matrix. We then accept $\mathbf{Y}^{(d)}(t+1)$ as the new value for the chain i.e., $\mathbf{X}^{(d)}(t+1) = \mathbf{Y}^{(d)}(t+1)$, with probability $\alpha\left(d, \mathbf{X}^{(d)}(t), \mathbf{Y}^{(d)}(t+1)\right) = \min\left\{1, \frac{\pi\left(d, \mathbf{Y}^{(d)}(t+1)\right)}{\pi\left(d, \mathbf{X}^{(d)}(t)\right)}\right\}$; otherwise, we set $\mathbf{X}^{(d)}(t+1) = \mathbf{X}^{(d)}(t)$.

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 favor 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 the best value for the proposal scaling, we need to determine its form as a function of d. Clearly, $\theta_1^{-2}(d)$ must be taken into account in this choice to prevent the algorithm from proposing too large jumps for components with smaller scaling terms. Furthermore, we should bear in mind that the larger is d, the greater are the odds of proposing an improbable move in one of the directions. To circumvent such a situation, it is reasonable to opt for a proposal scaling that is a decreasing function of d. We can show that the optimal proposal scaling form for the target in Section 2.1 (i.e., no matter if Condition (4) holds or not) is given by $\sigma^2(d) = \ell^2/d^\eta$, where ℓ^2 is some constant and η is the smallest number satisfying

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{d^\eta} < \infty \qquad \text{and} \qquad \lim_{d\to\infty} \frac{d^{\gamma_i} c\left(\mathcal{J}(i,d)\right)}{d^\eta} < \infty, \quad \text{for } i = 1, \ldots, m. \tag{5}$$

Under the fulfilment of Condition (4), we can find the exact form of $\sigma^2(d)$. For Condition (4) to be met, its reciprocal has to satisfy

$$\begin{aligned} &\lim_{d\to\infty} \sum_{j=1}^{d} \theta_1^{-2}(d)\, \theta_j^2(d) \\ &\quad = \lim_{d\to\infty} \frac{K_1}{d^{\lambda_1}} \left(\frac{d^{\lambda_1}}{K_1} + \ldots + \frac{d^{\lambda_n}}{K_n} + c\left(\mathcal{J}(1,d)\right) \frac{d^{\gamma_1}}{K_{n+1}} + \ldots + c\left(\mathcal{J}(m,d)\right) \frac{d^{\gamma_m}}{K_{n+m}} \right) < \infty. \end{aligned}$$

It must then be true that $\lim_{d\rightarrow\infty} c\left(\mathcal{J}\left(i,d\right)\right) d^{\gamma_i}/d^{\lambda_1} < \infty$, $\forall i \in \{1,\ldots,m\}$, in which case it is clear from (5) that $\sigma^2\left(d\right) = \ell^2/d^{\lambda_1}$. In other words, not only $\theta_1^{-2}\left(d\right)$ is the asymptotically smallest scaling term, but it also is small enough so as to "act of proposal scaling" for the algorithm. This conclusion is the opposite to that achieved in [1], where we required $\eta > \lambda_1$ for 0.234 to be the AOAR. Note that the requirement $\theta_{i^*}\left(d\right) = 1$, as mentioned at the end of Section 2.1, implies that $\sigma^2\left(d\right)$ is no greater than ℓ^2; hence, it remains finite for any d. In particular, the proposal scaling takes its largest form when $i^* \in \{1,\ldots,b\}$.

In Section 3, we shall differentiate two situations. In the first case, which happens when there is 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 $O(d^{\lambda_1})$, we say that the b small scaling terms dictate in part only the behavior of the algorithm. Indeed, everything else being held constant, ignoring these scaling terms would not affect the proposal distribution. Failing this, we say that the b components govern the accept/reject ratio of the algorithm as ignoring these terms would result in a larger value for η. We shall study these two situations separately. Note that this is a crucial distinction with [1], in which $\theta_1\left(d\right),\ldots,\theta_n\left(d\right)$ played an insignificant role as far as the acceptance of the proposed moves was concerned.

The proposal scaling (space) being a function of d, an appropriate rescaling of the elapsed time between each step is now required to guarantee a nontrivial limiting process as $d \rightarrow \infty$. Let $\mathbf{Z}^{(d)}\left(t\right)$ be the time-t value of the generated Markov chain sped up by a factor of $d^{\eta} = d^{\lambda_1}$; specifically, $\mathbf{Z}^{(d)}\left(t\right) = \mathbf{X}^{(d)}\left(\left[d^{\lambda_1}t\right]\right) = \left(X_1^{(d)}\left(\left[d^{\lambda_1}t\right]\right),\ldots,X_d^{(d)}\left(\left[d^{\lambda_1}t\right]\right)\right)$, where $[\cdot]$ is the integer part function. This (continuous-time) sped up process proposes on average d^{λ_1} moves during each unit interval, instead of only one. Subsequent sections shall be devoted to study the limiting distribution of each component, or each one-dimensional path, of the process $\left\{\mathbf{Z}^{(d)}\left(t\right), t \geq 0\right\}$ as $d \rightarrow \infty$.

Before presenting optimal scaling results we introduce the notion of π-average acceptance rate, defined in [12] as

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

where q is the normal proposal density. The asymptotic efficiency of the algorithm is closely connected to this concept, as shall be seen next section.

3 Optimal Scaling Results

This section introduces weak convergence results for the sped up algorithm, which shall later be used to establish the value of ℓ^2 leading to the chain that converges fastest to its invariant distribution. As mentioned previously, it is necessary to distinguish the case where the b small scaling terms are the only ones to govern the proposal scaling from that where they share this responsibility with at least one of the m groups having an infinite number of scaling terms in the limit, $\mathbf{\Theta}_{\mathcal{J}(i,d)}^{-2}\left(d\right)$. We begin by the latter.

3.1 New AOARs

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 $\Phi(\cdot)$. In order to ease notation, we adopt the following convention for defining vectors: $\mathbf{X}^{(b-a)} = (X_{a+1}, \ldots, X_b)$; furthermore, $\mathbf{X}^{(b-a)-}$ means that the component of interest, X_{i^*}, is excluded. We also use the following notation for conditional expectations: $\mathrm{E}[f(X,Y)|X] = \mathrm{E}_Y[f(X,Y)]$.

Theorem 1. *Consider a Metropolis algorithm with proposal distribution $\mathbf{Y}^{(d)} \sim N\left(\mathbf{x}^{(d)}, \frac{\ell^2}{d^{\lambda_1}} I_d\right)$, applied to a target density as in (1) satisfying the specified conditions on f, with $\mathbf{\Theta}^{-2}(d)$ as in (3), and $\theta_{i^*}(d) = 1$. Consider the i^*th component of the process $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$ i.e., $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\} = \left\{X_{i^*}^{(d)}\left(\left[d^{\lambda_1} t\right]\right), t \geq 0\right\}$, 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 Condition (4) is satisfied and

$$\exists i \in \{1, \ldots, m\} \text{ such that } \lim_{d \to \infty} \frac{c(\mathcal{J}(i,d))\, d^{\gamma_i}}{d^{\lambda_1}} > 0, \tag{6}$$

with $c(\mathcal{J}(i,d))$ as in (2).

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

$$\begin{aligned} \alpha^*\left(\ell^2, X_{i^*}, Y_{i^*}\right) &= \mathrm{E}_{\mathbf{Y}^{(b)-}, \mathbf{X}^{(b)-}}\left[\Phi\left(\frac{\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right. \\ &\qquad \left. + \prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)} \Phi\left(\frac{-\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right]. \end{aligned} \tag{7}$$

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

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

where

$$\upsilon(\ell) = 2\ell^2 \mathrm{E}_{\mathbf{Y}^{(b)}, \mathbf{X}^{(b)}}\left[\Phi\left(\frac{\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right]. \tag{8}$$

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

$$E_R = \lim_{d \to \infty} \sum_{i=1}^{m} \frac{c(\mathcal{J}(i,d))}{d^{\lambda_1}} \frac{d^{\gamma_i}}{K_{n+i}} \mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]. \tag{9}$$

Remark 1 For $j \in \{1, \ldots, b\}$, the term $\theta_j(d) X_j$ is distributed according to f and is $O(1)$ for any d. We use $\theta_j X_j = \theta_j(d) X_j$, and similarly $\theta_j Y_j = \theta_j(d) Y_j$ (which is normally distributed with variance ℓ^2), to emphasize this fact (and because it does not make sense to write the limiting process as a function of d).

Remark 2 Contrarily as in [1], these processes are with respect to the filtration $\mathcal{F}^{Z_{i^*}^{(d)}}(t)$, and are thus marginal limiting processes. In other words, they are based on past information about the component of interest only. Note that marginally, these processes are Markovian; based on $\mathcal{F}^{\mathbf{Z}^{(d)}}(t)$ however i.e., given the past moves of the d-dimensional process, we do not average over $\mathbf{X}^{(b)-}$ and the limiting one-dimensional processes are not Markovian anymore, but are part of higher dimensional Markov processes (see the last paragraph of Section 3.1).

Interestingly, the one-dimensional processes associated with the b components of smallest order, $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\}$ for $i^* = 1, \ldots, b$, each possess a discrete-time limiting process. Indeed, they explore the space faster since $\sigma^2(d)$ perfectly suits them, so a speed-up time factor is not required. They converge to stationarity in $O(1)$ iterations. The acceptance rule α^*, which satisfies the detailed balance condition, is affected by specific components only; for finite values of d, they constitute the components that are more likely to cause the rejection of the proposed moves. Intuitively, we know that the more components rule the algorithm, the harder it is to accept moves. Everything else being held constant, we thus expect α^* to decrease as b and/or E_R increase.

Let's now examine α^* in the simple role of an acceptance function rather than as part of some limiting process. To do this, we assume that we are sampling from some finite-dimensional target density $f(x_1, \ldots, x_k)$ using a $N\left(\mathbf{X}^{(k)}(t), \tau^2\right)$ proposal for some $\tau > 0$. Instead of accepting the proposed moves according to the usual rule α, we instead use the new rule α^*. The quantities ℓ^2 and E_R being independent of both the proposal and target densities, we can simply see them as parameters; since they always appear together as $\ell^2 E_R$, we conveniently introduce $\xi = \ell^2 E_R$. We thus expect the algorithm to perform differently depending on which value for ξ is selected. In the one-dimensional case where we would like to sample from $f(x)$ for instance, we have

$$\alpha^*(\xi, x, y) = \Phi\left(\frac{\log \frac{f(y)}{f(x)} - \xi/2}{\sqrt{\xi}}\right) + \frac{f(y)}{f(x)} \Phi\left(\frac{\log \frac{f(x)}{f(y)} - \xi/2}{\sqrt{\xi}}\right);$$

if we let $\xi \to \infty$, then $\alpha^*(\xi, x, y) \to 0$ and the chain never moves; when $\xi = 0$, then $\alpha^* = \alpha$, the usual rule as introduced in Section 2.2. In [11], the optimal Metropolis acceptance rule (in terms of asymptotic variance) was shown to be α because it favors the mixing of the chain by improving the sampling of all possible states. The efficiency of α^* is thus inversely proportional to its parameter ξ (see Figure 1). Note however that when α^* steps back into its original role, where it is part of the limiting process described in Theorem 1, then E_R is not a parameter anymore (its value is fixed by the target density); furthermore since the proposal distribution depends on ℓ, setting $\ell = 0$ is obviously not optimal as this would yield a static chain.

For $i^* = b+1, \ldots, d$, we find that $\sigma^2(d) = \ell^2/d^{\lambda_1}$ is of smaller order than $\theta_{i^*}^{-2}(d) \equiv 1$; consequently, the processes $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\}$ do not explore their space as efficiently as when $i^* = 1, \ldots, b$ and now converge in $O(d^{\lambda_1})$ iterations. We thus obtain continuous-time limiting Langevin diffusion processes whose speed measure is different from that in [1] and [12]; it now depends on $\mathbf{X}^{(b)}, \mathbf{Y}^{(b)}$ and this alters the value of the AOAR.

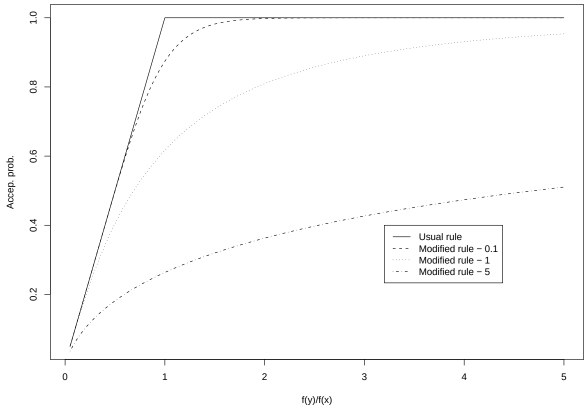

Figure 1: Graph of α^* versus the density ratio $f\left(y\right)/f\left(x\right)$ for $\xi = 0, 0.1, 1, 5$ (from top to bottom) and when $b = 1$.

Since there are two different types of limiting processes (we only had the diffusive limit in [1]), we now face the dilemma as to which should be chosen to determine the AOAR. The algorithm either accepts or rejects all d individual moves in a given step so we must have a common acceptance rate in all directions. For $i^* = b+1, \ldots, d$, the only quantity depending on the proposal variance (i.e. on ℓ) in the limit is $\upsilon\left(\ell\right)$, so optimizing the mixing rate of any of these one-dimensional processes is equivalent to choosing the diffusion process that goes the fastest (so the value of ℓ for which the speed measure $\upsilon\left(\ell\right)$ is optimized). For $i^* = 1, \ldots, b$, efficiency criteria are not all equivalent as we are dealing with discrete-time limiting processes; attempting to select the "best" ℓ would then depend on our choice of efficiency measure. To be confident that $\left\{\mathbf{Z}^{(d)}\left(t\right), t \geq 0\right\}$ has reached stationarity, we must be certain that $\left\{Z_{i^*}^{(d)}\left(t\right), t \geq 0\right\}$ have all reached stationarity. Since components possessing a diffusion limit mix in $O(d^{\lambda_1})$ iterations while those converging to a discrete-time process mix according to $O\left(1\right)$, it thus makes sense to rely on $\upsilon\left(\ell\right)$ to achieve the fastest convergence of the algorithm to its stationary distribution.

The following corollary provides an equation for the asymptotic acceptance rate of the algorithm as a function of $\upsilon\left(\ell\right)$ in (8).

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

An analytical solution for the value $\hat{\ell}$ that maximizes the function $\upsilon\left(\ell\right)$ cannot be obtained. However, this maximization problem can be easily resolved with numerical methods. For densities f satisfying the regularity conditions mentioned in Section 2.1, $\hat{\ell}$ is finite and unique. This thus yields an AOAR $a(\hat{\ell})$ and although an explicit form is not available for this quantity, we can still draw some conclusions about $\hat{\ell}$ and the AOAR. First, Condition (4) ensures the existence of $1 \leq b \leq \infty$ components with a scaling term substantially 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 $\left\{ Z_{i^*}^{(d)}(t), t \geq 0 \right\}$ for $i^* \in \{b+1, \ldots, d\}$, 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 $\theta_{i^*}(d) \equiv 1$. Hence, the first b components obviously provoke a reduction of $\hat{\ell}$ and the AOAR, which is now necessarily smaller than 0.234. In particular, both quantities get smaller as b increases.

Traditionally, optimal scaling results have been derived using joint limiting processes (i.e., processes based on $\mathcal{F}^{\mathbf{Z}^{(d)}}(t)$); this method has worked well in the literature (including in [1]) since every component of the algorithms considered was Markovian in the limit, and thus independent of all the other components. In the present case, the joint limiting process is itself Markovian, but the individual components from the joint limiting process are not; $\left\{ \mathbf{Z}^{(b)}(t), t \geq 0 \right\}$ converges to some b-dimensional Metropolis algorithm, while the last $d-b$ components become *conditionally iid* given $\mathbf{X}^{(b)}$ and converge to a *common* limiting diffusion, whose drift and volatility depend on $\mathbf{X}^{(b)}$. Since $\mathbf{X}^{(b)}$ affects the accept/reject ratio of the algorithm for any d, it becomes difficult to rely on joint limiting processes, as it is not clear how diffusion processes can be optimized (apart from optimizing the speed measure, which is not available anymore). If we knew how to optimize the diffusion, we would obtain $\hat{\ell}\left(\mathbf{X}^{(b)}\right)$, and a global value $\hat{\ell}$ could then be found by averaging $\hat{\ell}\left(\mathbf{X}^{(b)}\right)$ with respect to $\pi\left(\mathbf{X}^{(b)}\right)$; this is equivalent here to optimizing the marginal limiting processes, as suggested in the previous paragraphs. Note, however, that the similarity of the marginal processes for the last $d-b$ components of the algorithm is crucial to the success of this method, as we need all components possessing a continuous-time limit to agree about a same AOAR.

3.2 Homogeneous Proposals: An Impasse

We now consider the remaining case, where the b scaling terms of smallest order entirely determine the proposal scaling. This is the case where $\left\{ Z_{i^*}^{(d)}(t), t \geq 0 \right\}$, $i^* = 1, \ldots, b$ converge "too fast" compared to the overall convergence speed of the algorithm.

Theorem 3. *In the settings of Theorem 1 but with Condition (6) replaced by*

$$\lim_{d \to \infty} \frac{c(\mathcal{J}(i,d))\, d^{\gamma_i}}{d^{\lambda_1}} = 0 \quad \forall\, i \in \{1, \ldots, m\}, \tag{10}$$

the conclusions of Theorem 1 are preserved, but the acceptance rule of the limiting Metropolis algorithm is now $\alpha^*(X_{i^*}, Y_{i^*}) = \mathrm{E}_{\mathbf{Y}^{(b)-}, \mathbf{X}^{(b)-}}\left[1 \wedge \prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)}\right]$ *and the speed measure for the limiting Langevin diffusion is* $\upsilon(\ell) = 2\ell^2 \mathrm{P}_{\mathbf{Y}^{(b)}, \mathbf{X}^{(b)}}\left(\sum_{j=1}^{b} \varepsilon(X_j, Y_j) > 0\right)$. *Furthermore,* $\lim_{d \to \infty} a(d, \ell) = a(\ell) \equiv \upsilon(\ell)/\ell^2$.

Since $\sigma^2(d)$ is now entirely ruled by the first b components, it follows that E_R in (9) is equal to 0. When $b = 1$, the acceptance rule of the limiting Metropolis algorithm reduces to the usual rule, α. In that case, the first component not only becomes independent of the others as $d \to \infty$, but is totally unaffected by these $d-1$ components in the limit, which move too slowly compared to the pace of X_1. When trying to optimize the speed measure $\upsilon(\ell)$, we realize that it is unbounded as a function of ℓ for virtually any density f satisfying the regularity conditions of Section 2.1. This indicates that ℓ should be chosen as large as possible; however, we also have $a(\ell) \to 0$ as $\ell \to \infty$. We conclude that when Condition (10) is satisfied, the proposal distribution with variance $\sigma^2(d) = \ell^2/d^{\lambda_1}$ generates moves that are way too small for the last $d-b$ components, forcing large values of ℓ to compensate. In theory, we thus obtain a well-defined limiting process, but in practice we lie in an impasse when it comes to optimizing the efficiency of the algorithm. We shall see in Section 4 that for such cases, inhomogeneous proposal scalings constitute a wiser option.

One particularity of the results introduced in Section 3 is their ability to optimize Metropolis algorithms when sampling from any multivariate normal target, regardless of the correlation structure amongst its components. Thanks to the invariance property of such targets under orthogonal transformations, it suffices to transform the covariance matrix Σ into a diagonal matrix, whose diagonal elements are the eigenvalues of Σ. Under this transformation, the target components are independent and the eigenvalues can be used to verify if Condition (4) is satisfied. In the case it is, we can determine the AOAR with Theorem 1 or 3, depending on which of Condition (6) or (10) is satisfied. Otherwise, the AOAR is the well-known 0.234 as demonstrated in [1].

For instance, let $X_1 \sim N(0,1)$ and $X_i \sim N(X_1, 1)$, $i = 2, \ldots, d$; $(X_1, \ldots, X_d)$ is normally distributed with mean vector $\mathbf{0}$ and Σ such that $\sigma_{i,j} = 2$ for $2 \leq i = j \leq d$, $\sigma_{i,j} = 1$ otherwise. All but two eigenvalues of Σ are equal to 1, and the other two are $O(d)$ and $O(1/d)$ respectively. Conditions (4) and (6) are thus satisfied, so based on Theorem 1, (8) may be optimized to find $\hat{\ell}^2 = 3.8$ along with an AOAR lying about 0.2. Detailed examples about the application of the theorems in this section are presented in [2].

4 Inhomogeneous Proposal Scalings and Extensions

Homogeneous proposal scalings are generally suboptimal to heterogeneous ones. Hence, inhomogeneous proposal scalings for our particular target model might be used to further improve the efficiency of the algorithm, as mentioned in Section 3.2 of [1]. However, as realized in Section 3.2 of the current paper, 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. In such a case inhomogeneous proposal scalings are not just an improvement, they are a necessity; this approach shall ensure a decent speed of convergence for each of the target components when facing models as those in Section 3.2.

In particular, consider $\mathbf{\Theta}^{-2}(d)$ in (3) and set the proposal scaling of the component X_j as follows: $\sigma_j^2(d) = \ell^2/d^{\eta_j}$, where $\eta_j = \lambda_1$ for $j = 1, \ldots, n$ and η_j is the smallest value such that $\lim_{d\to\infty} c(\mathcal{J}(i,d))\, d^{\gamma_i}/d^{\eta_j} < \infty$ for $j = n+1, \ldots, d$, $j \in \mathcal{J}(i,d)$. When studying $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\}$, we still assume $\theta_{i^*}(d) = 1$, but we now let $\mathbf{Z}^{(d)}(t) = \mathbf{X}^{(d)}([d^{\eta_{i^*}} t])$. We have the following result.

Theorem 4. *In the settings of Theorems 1 and 3 (i.e. no matter if Condition (6) or (10) is satisfied) but with proposal scalings as just described, the conclusions of Theorem 1 and*

Corollary 2 are preserved and E_R is now expressed as

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

Since E_R is larger than before, the optimal scaling value $\hat{\ell}$ is now smaller than with homogeneous proposal scalings. Indeed, when $\sigma^2\left(d\right)$ was only 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 $\hat{\ell}^2$. In [1], it was easily verified that inhomogeneous proposals do not affect the value of the AOAR. The same statement does not hold anymore, although we can affirm that the AOAR cannot be greater than 0.234; since ℓ is assumed to be the same in each direction, the algorithm can hardly do better than for *iid* targets even though the proposal has been personalized.

For instance, consider a d-dimensional normal target with mean vector $\mathbf{0}$, variances equal to $\left(1/d^2, 1, \ldots, 1\right)$, and null covariances. In that case, $\sigma^2\left(d\right) = \ell^2/d^2$, which yields an AOAR of 0 by Theorem 3. Switching to the inhomogeneous proposal scalings $\left(\ell^2/d^2, \ell^2/d, \ldots, \ell^2/d\right)$ instead yields an AOAR of about 0.16; this is obviously far more efficient than the former (see [2] for more details).

The results of Section 3 may be extended to target distributions with scaling terms admitting broader forms. We start by relaxing the form of the scaling terms as a function of d, and also 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 as $d \to \infty$, we allow the constant terms to be randomly distributed according to a distribution satisfying specific moment conditions; we also allow the scaling terms of a common group to be different functions of the dimension, provided they are of the same order. In addition, we let the cardinality functions $c\left(\mathcal{J}\left(i,d\right)\right)$ be general functions of d, with $c\left(\mathcal{J}\left(i,d\right)\right) \to \infty$ as $d \to \infty$. Specifically, we have

$$\mathbf{\Theta}^{-2}\left(d\right) = \left(K_1\left(\theta_1^*\left(d\right)\right)^{-2}, \ldots, K_n\left(\theta_n^*\left(d\right)\right)^{-2}, K_{n+1}\left(\theta_{n+1}^*\left(d\right)\right)^{-2}, \ldots, K_d\left(\theta_d^*\left(d\right)\right)^{-2}\right). \quad (11)$$

We assume that $\left\{K_j, j \in \mathcal{J}\left(i,d\right)\right\}$ are *iid* and chosen randomly from some distribution with $\mathrm{E}\left[K_j^{-2}\right] < \infty$. Without loss of generality, we also take $\mathrm{E}\left[K_j^{-1}\right] = \beta_i$ for $j \in \mathcal{J}\left(i,d\right)$. Recall that the scaling term of the component of interest does not depend on d, and we therefore have $\theta_{i^*}^{-2}\left(d\right) = K_{i^*}$. To have sensible limiting theory, we restrict our attention to functions such that $\lim_{d\to\infty} \theta_j^*\left(d\right)$ exists.

The m groups of scaling terms appearing infinitely often in the limit are now defined by

$$\mathcal{J}\left(i,d\right) = \left\{j \in \left\{1,\ldots,d\right\}; 0 < \lim_{d\to\infty} \frac{\theta_i'\left(d\right)}{\theta_j\left(d\right)} < \infty\right\}, \quad i = 1,\ldots,m. \quad (12)$$

Here, $\theta_i'\left(d\right)$ is a baseline function such that $\left\{\theta_j^*\left(d\right), j \in \mathcal{J}\left(i,d\right)\right\}$ is $O\left(\theta_i'\left(d\right)\right)$. We again assume that the first n and the next m scaling terms are respectively classified according to an asymptotically increasing order. To support the previous modifications, we restrict $\theta_i'\left(d\right)$, $i = 1,\ldots,m$ to be of different orders, otherwise we could just group the scaling terms of a same order together. In addition, we suppose that there does not exist a $\theta_j^*\left(d\right)$, $j = 1,\ldots,n$

of the same order as one of the $\theta'_i(d)$, $i = 1, \ldots, m$. Hence, $\mathbf{\Theta}^{-2}(d)$ contains at least m and at most $n + m$ scaling terms of different orders; if there is infinitely many scaling terms of a same order in the limit, they thus necessarily belong to the same of the m groups.

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

$$\lim_{d\to\infty} \theta_1^2(d)\, \sigma_\eta^2(d) < \infty \quad \text{and} \quad \lim_{d\to\infty} c\left(\mathcal{J}(i,d)\right) \theta_i'^2(d)\, \sigma_\eta^2(d) < \infty \;\text{ for } i = 1, \ldots, m. \qquad (13)$$

We then have the following result.

Theorem 5. *Consider the settings of Theorem 1 with $\mathbf{\Theta}^{-2}(d)$ as in (11), $\theta_{i^*} = \theta_{i^*}(d) = K_{i^*}^{-1/2}$, and proposal scaling $\sigma^2(d) = \ell^2\sigma_\eta^2(d)$ where $\sigma_\eta^2(d)$ satisfies (13). Suppose that*

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

holds instead of Condition (4), and that either

$$\exists i \in \{1, \ldots, m\} \;\textit{ such that }\; \lim_{d\to\infty} \frac{c\left(\mathcal{J}(i,d)\right)\theta_i'^2(d)}{\theta_1^2(d)} > 0 \qquad (14)$$

holds instead of Condition (6), or

$$\lim_{d\to\infty} \frac{c\left(\mathcal{J}(i,d)\right)\theta_i'^2(d)}{\theta_1^2(d)} = 0 \quad \forall i \in \{1, \ldots, m\} \qquad (15)$$

holds instead of Condition (10).

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 if (14) is met (or Theorem 3 if (15) is met) for the first b components, but where it satisfies the Langevin SDE

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

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

Under this setting, the quantity E_R is now given by

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

where $c\left(\mathcal{J}(i,d)\right)$ is the cardinality function of (12). In addition, the conclusion of Corollary 2 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 $d-n$ constant terms affects the quantity E_R through the terms $\beta_1, \ldots, \beta_m$; therefore, it only has an impact on the limiting processes when Condition (14) is satisfied.

This theorem assumes a somewhat general form for the target distribution. Contrarily to what was proven in [1], the AOARs in this paper are not independent of the target distribution and scaling terms anymore, and finding the exact AOAR involves maximizing the corresponding speed measure $\upsilon(\ell)$. Although 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 rather than blindly tune the algorithm according to the 0.234 rule.

5 Discussion

This paper studies the asymptotic behavior of the Metropolis algorithm applied to a particular type of target distribution with non-*iid* components. It provides a condition under which the limiting behavior of the algorithm differs from that in [12] for the *iid* case (and thus also from [1]). The cause of this discrepancy is the existence of a finite number of components converging significantly faster than the others to their stationary distribution, yielding AOARs that are no greater than the usual 0.234 and which often significantly differ from this value. These results are the first to admit limiting processes and AOARs that are different from those found by [12] for Metropolis algorithms with Gaussian proposal. 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.

The existence of components converging significantly faster than others is the direct cause of the divergence between our results and previously published work on optimal scaling. This is also the reason why dealing with marginal rather than joint limiting processes now becomes necessary. In cases where the discrepancy between the speed of convergence of some components is too large, the algorithm considered becomes inefficient for large d. Since the proposal scaling of the algorithm is governed by a finite number of target components only, the resulting overall convergence speed of the algorithm is extremely slow; we then concluded that using inhomogeneous proposal scalings would be infinitely more efficient in such situations.

It is worth mentioning that although asymptotic, the results presented in this paper work well in relatively small dimensions. In addition, the method provided to determine the optimal form for the proposal variance as a function of d turns out to be a useful guideline in practice. The results of this paper are then relatively easy to apply for practitioners, as it suffices to verify which of Conditions (4), (6), and (10) are satisfied, and then numerically optimize the appropriate speed measure to find the optimal value $\hat{\ell}$.

A particular case where 0.234 was shown not to be optimal was when considering widely used normal hierarchical models. This might seem surprising, given that multivariate normal distributions were believed to adopt a conventional limiting behavior. Consequently, it would be interesting to see if similar results could be derived for target distributions with nontrivial correlation structures, or for other types of MCMC algorithms (the MALA, for instance). Such questions are presently under investigation. In particular, we presently study the behavior of the Metropolis algorithm applied to hierarchical target models, which are very popular in the Bayesian community.

Appendix A. Theorems Proofs

We now prove the results introduced in Sections 3 and 4. Because of their similarity, we shall present a detailed proof of Theorem 1, and outline the main differences for the other theorems. The proofs are based on Theorem 8.2 and Corollary 8.6 of Chapter 4 in [7], which say that it suffices to verify $\mathcal{L}^1$-convergence of the processes' generator to assess weak convergence of the sequence of processes considered (see [1] for more details). Contrarily to [1], we are now dealing with one-dimensional marginal processes. Since these processes do not satisfy the Markov property, it thus becomes necessary to work with generator-like expressions. This variation does not cause any technical difficulty as Theorem 8.2 in [7] is still valid for such approximately Markovian settings. This however implies computing an extra expectation with respect to $\mathbf{X}^{(d)-}$ (compared to [1], in which we were only taking an expectation with respect to $\mathbf{Y}^{(d)}$).

Generators of processes are usually expressed as a function of some arbitrary (smooth) test function h; in the case where the sequence of processes converges to a Langevin diffusion process, the space of infinitely differentiable functions with compact support on $\mathbf{R}$, C_c^∞, is a core for the limiting generator and we are thus allowed to verify $\mathcal{L}^1$-convergence of generators for $h \in C_c^\infty$ only (again, see [1]). In the case of discrete-time limiting processes, we need to verify $\mathcal{L}^1$-convergence for $h \in \overline{C}$, the space of bounded continuous functions on $\mathbf{R}$.

Proof of Theorem 1

For $i^* \in \{1, \ldots, b\}$ and an arbitrary test function $h \in \overline{C}$, we first show that

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

where

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

is the generator-like of the i^*th component of the sped up Metropolis algorithm and $G_{MH}h\left(X_{i^*}\right) = \mathrm{E}_{Y_{i^*}}\left[\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right)\alpha^*\left(\ell^2, X_{i^*}, Y_{i^*}\right)\right]$ is the generator of a Metropolis algorithm with acceptance rule α^* as in (7). Given that $\theta_{i^*}(d) \equiv 1$, it follows that $\lambda_j = \lambda_1 = 0$ for $j = 1, \ldots, b$ and thus $\sigma^2(d) = \ell^2$. Since $\sigma^2(d)$ does not depend on the dimension of the target, there is no need to speed up the d-dimensional Metropolis algorithm, justifying the discrete-time nature of the limiting process.

From Lemma B.1, we have $\mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - \widehat{G}h\left(d, X_{i^*}\right)\right|\right] \to 0$ as $d \to \infty$, where

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

with $z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ as in (B.1). We are then left to show $\mathcal{L}^1$-convergence of the generator-like $\widehat{G}h\left(d, X_{i^*}\right)$ to the generator of the Metropolis algorithm with acceptance rule α^*. Substituting explicit expressions for the generators and using the triangle's inequality along with the fact that h is bounded gives

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

for some constant K. By Lemma B.2 the RHS converges to 0, proving the first part of Theorem 1.

To complete the proof, we must now show that for $i^* \in \{b+1, \ldots, d\}$ and an arbitrary $h \in C_c^\infty$, $\lim_{d\to\infty} \mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - G_L h\left(X_{i^*}\right)\right|\right] = 0$, where

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

is the generator of a Langevin diffusion process with speed measure $\upsilon\left(\ell\right)$ as in (8).

Combining Lemma 7 in [1] and the triangle's inequality, we find $\mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - \widetilde{G}h\left(d, X_{i^*}\right)\right|\right] \to 0$ as $d \to \infty$, where

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

We conclude the proof by showing that $\widetilde{G}h\left(d, X_{i^*}\right)$ is asymptotically equivalent to the generator of the Langevin diffusion, $G_L h\left(d, X_{i^*}\right)$. Substituting explicit expressions for the generators, grouping some terms and using the triangle's inequality yield

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

Since the function h has compact support, it implies that h itself and its derivatives are bounded in absolute value by some constant. Combining the Bounded Convergence Theorem (see [15], for instance) with Lemma 8 in [1], and then applying Lemma B.3 of the following section, we 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 first combining Lemma 10 in [1] with the Bounded Convergence Theorem, and then applying Lemma B.4.

Proof of Theorem 3

The distinction with the proof of Theorem 1 is that $z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right) \to_p \sum_{j=1}^{b}\varepsilon\left(X_j,Y_j\right)$. To realize this, we first notice that $\sum_{j=b+1}^{n}\varepsilon\left(d,X_j,Y_j\right) \to_p 0$ (Proposition A.2 in [1]) and that $\sum_{i=1}^{m} R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right) \to 0$ (Proposition A.3 in [1] along with Condition (10)). From (B.1), this implies that $\lim_{d\to\infty}\mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)\right] = \sum_{j=1}^{b}\varepsilon\left(X_j,Y_j\right)$. Therefore,

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

where the inequality has been obtained by applying Chebychev's inequality and the last equality has been obtained from the conditional distribution in (B.7).

For the first part of Theorem 3, we can conclude that $\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge e^{z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] \to_p \mathrm{E}_{\mathbf{Y}^{(b)-},\mathbf{X}^{(b)-}}\left[1\wedge\prod_{j=1}^{b}\frac{f(\theta_jY_j)}{f(\theta_jX_j)}\right]$ as $d\to\infty$, since $1\wedge e^{z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}$ is a continuous and bounded function. For the second part of the theorem, it suffices to use the fact that $\mathrm{E}\left[1\wedge e^{z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)}\right] \to \mathrm{E}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left[1\wedge\prod_{j=1}^{b}\frac{f(\theta_jY_j)}{f(\theta_jX_j)}\right]$ as $d\to\infty$ along with the following decomposition

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

and Proposition C.3 to conclude that

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

Proof of Theorem 5

Although an elaborate notation is necessary due to the general form of the functions $c\left(\mathcal{J}\left(i,d\right)\right)$, $i=1,\ldots,m$ and $\theta_j\left(d\right)$, $j=1,\ldots,d$, 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 be different, provided they are of the same order. To deal with this characteristic of the model, we write $\theta_j\left(d\right)=K_j^{-1/2}\theta_i'\left(d\right)\left[\theta_j^*\left(d\right)/\theta_i'\left(d\right)\right]$. This representation is used to allow factoring a common function of d for all components of a same group. Instead of factoring the term $\theta_{n+i}^2\left(d\right)$ as in Theorem 1, we factor $\beta_i\left(\theta_i'\left(d\right)\right)^2$. We are then left with the ratio $\left(\theta_j^*\left(d\right)/\theta_i'\left(d\right)\right)^2$ but since this converges to 1 as $d\to\infty$, the rest of the proof can be repeated with minor adjustments. A typical quantity we deal with is

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

Appendix B. Lemmas

Discrete-time Generator $\widehat{G}h(d, X_{i^*})$

Lemma B.1. *If $\lambda_1 = 0$, then $\lim_{d\to\infty}\mathrm{E}\left[\left|Gh(d,X_{i^*}) - \widehat{G}h(d,X_{i^*})\right|\right] = 0\ \forall h\in\overline{C}$, where*

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

with

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

Also, $\varepsilon(d,X_j,Y_j) = \log\{f(\theta_j(d)\,Y_j)/f(\theta_j(d)\,X_j)\}$ and for $i = 1,\ldots,m$,

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

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

Proof. The generator-like of the i^*th component of the sped up Metropolis algorithm in (A.1) may be reexpressed as

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

We first concentrate on the inner expectation. Using properties of the log function and a three-term Taylor expansion, we obtain

$$\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right] \tag{B.3}$$

$$= \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\exp\left\{\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right)+\sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}\left[\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)\right.\right.\right.$$

$$\left.\left.\left.+\frac{1}{2}\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)^2+\frac{1}{6}\frac{d^3}{dU_j^3}\log f\left(\theta_j\left(d\right)U_j\right)\left(Y_j-X_j\right)^3\right]\right\}\right]. \tag{B.4}$$

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

Before verifying if $\widehat{G}h\left(d,X_{i^*}\right)$ 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 $z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$. By the triangle inequality, the Lipschitz property of the function $1\wedge e^x$ (see Proposition 2.2 in [12]), and noticing that the first two terms of the function $z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$ cancel out with the first two terms of the exponential term in (B.4), we obtain

$$\left|\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right]-\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge e^{z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\right|\leq$$

$$\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{1}{2}\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)^2+\frac{\ell^2}{2d^{\lambda_1}}\left(\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\right)^2\right)\right.\right.$$

$$\left.\left.+\frac{1}{6}\sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}\frac{d^3}{dU_j^3}\log f\left(\theta_j\left(d\right)U_j\right)\left(Y_j-X_j\right)^3\right|\right].$$

Noticing that the first double sum forms the random variables $W_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)},\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)$'s introduced in Lemma C.1 we find

$$\left|\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right]-\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge e^{z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\right|$$

$$\leq \sum_{i=1}^{m}\mathrm{E}\left[\left|W_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)},\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right]+\sum_{i=1}^{m}c\left(\mathcal{J}\left(i,d\right)\right)\ell^3K\frac{d^{3\gamma_i/2}}{d^{3\lambda_1/2}} \tag{B.5}$$

for some constant K. We are now ready to verify $\mathcal{L}^1$-convergence of the generators. Using the triangle's inequality and (B.5), we find

$$\mathrm{E}\left[\left|Gh\left(d,X_{i^*}\right)-\widehat{G}h\left(d,X_{i^*}\right)\right|\right] \leq \sum_{i=1}^{m}\mathrm{E}\left[\left|W_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)},\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right]\mathrm{E}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\right]$$

$$+\sum_{i=1}^{m}c\left(\mathcal{J}\left(i,d\right)\right)\ell^3K\frac{d^{3\gamma_i/2}}{d^{3\lambda_1/2}}\mathrm{E}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\right].$$

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

Convergence to the Acceptance Rule α^*

Lemma B.2. *If $\lambda_1 = 0$ and Conditions (4) and (6) are satisfied, then*

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

with $z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$ as in (B.1) and $\alpha^*\left(\ell^2, X_{i^*}, Y_{i^*}\right)$ as in (7).

Proof. We first use conditional expectations to obtain

$$\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] = \mathrm{E}_{\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}}\left[\mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[1 \wedge e^{z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\right]. \quad \text{(B.6)}$$

Since $\left(Y_j - X_j\right)|X_j \sim iid\ N\left(0,\ell^2\right)$ for $j = n+1,\ldots,d$, then

$$z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)\Big|\,\mathbf{Y}^{(n)},\mathbf{X}^{(d)} \sim N\left(\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right),\ \ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)\right). \quad \text{(B.7)}$$

Applying Proposition 2.4 in [12], we can express the inner expectation in (B.6) in terms of $\Phi(\cdot)$, the *cdf* of a standard normal random variable

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

We need to study the convergence of every term included in previous expression. Condition (4) 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, defined in the present case as $b = \max\left(j \in \{1,\ldots,n\}; \lambda_j = 0\right)$. It is thus pointless to study the convergence of these b variables as they are independent of d. However, we can study the convergence of the other $n-b$ components and from Proposition A.2 in [1] we know that $\varepsilon\left(d,X_j,Y_j\right) \to_p 0$, since $\lambda_j < 0$ for $j = b+1,\ldots,n$. Similarly, we can use Proposition A.3 in [1] and Condition (6) to conclude that $\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right) \to_p E_R > 0$, with E_R as in (9).

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

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

From the triangle's inequality, we then obtain

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

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 to conclude that the previous expression converges to 0 as $d \to \infty$. □

Volatility of the Langevin Diffusion

Lemma B.3. *If Conditions (4) and (6) are satisfied, then*

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

where

$$\begin{aligned}
z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right) &= \sum_{j=1,j\neq i^*}^{n}\varepsilon\left(d,X_j,Y_j\right)+\sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d),j\neq i^*}\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right) \\
&\quad -\frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)
\end{aligned} \tag{B.8}$$

and E_R is as in (9).

Proof. First note that (B.8) is similar to (B.1), except that the i^*th component is now excluded from the expression. Therefore, it is easily seen from the proof of Lemma B.2 that

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

From Proposition C.2, both terms of the sum have the same expectation and the previous expression thus simplifies to

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

By Proposition A.2 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 A.3 in [1], we also know that $\sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right) \to_p E_R$, where E_R is as in (9) and is strictly positive by Condition (6). Applying Slutsky's and the Continuous Mapping Theorems thus yield

$$\Phi\left(\frac{\sum_{j=1,j\neq i^*}^{n} \varepsilon\left(d, X_j, Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)}{\sqrt{\ell^2 \sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)}}\right) \to_p \Phi\left(\frac{\sum_{j=1,j\neq i^*}^{b} \varepsilon\left(X_j, Y_j\right) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right) \text{(B.9)}$$

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

Drift of the Langevin Diffusion

Lemma B.4. *If Conditions (4) and (6) are satisfied, then*

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

where $z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)$ *and* E_R *are as in (B.8) and (9) respectively.*

Proof. The proof of this result is similar to that of Lemma B.3 and for this reason, we shall not repeat every detail. Since we know the conditional distribution of $z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)\Big|\mathbf{Y}^{(n)-}, \mathbf{X}^{(d)-}$, we can use Proposition 2.4 in [12] to conclude that

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

From Proposition C.2, the unconditional expectation simplifies to

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

Using (B.9) along with the Bounded Convergence Theorem completes the proof of the lemma. □

Appendix C. Auxiliary Results

Lemma C.1. *For* $i = 1, \ldots, m$, *let*

$$W_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)}, \mathbf{Y}_{\mathcal{J}(i,d)}^{(d)}\right) = \frac{1}{2}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right) X_j\right)\right)\left(Y_j - X_j\right)^2$$

$$+\frac{\ell^2}{2d^{\lambda_1}}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\right)^2,$$

where $Y_j\left|X_j\right.\sim N\left(X_j,\ell^2/d^{\lambda_1}\right)$ *and* X_j *is distributed according to the density* $\theta_j\left(d\right)f\left(\theta_j\left(d\right)x_j\right)$*, independently for all* $j=1,...,d$*. Then, for* $i=1,\ldots,m$*,* $\mathrm{E}\left[\left|W_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)},\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right]\to 0$ *as* $d\to\infty$*.*

Proof. From the proof of Lemma 6 in [1], we have

$$\mathrm{E}_{\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}}\left[\left|W_i\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)},\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right]\le\frac{\ell^2}{d^{\lambda_1}}\theta^2_{n+i}\left(d\right)\sqrt{c\left(\mathcal{J}\left(i,d\right)\right)}K$$
$$+\frac{\ell^2}{2d^{\lambda_1}}\theta^2_{n+i}\left(d\right)c\left(\mathcal{J}\left(i,d\right)\right)\left|\frac{1}{c\left(\mathcal{J}\left(i,d\right)\right)}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f\left(X_j\right)+\left(\frac{d}{dX_j}\log f\left(X_j\right)\right)^2\right)\right|;$$

we also have

$$\left|S_i\left(d\right)\right|\equiv\left|\frac{1}{c\left(\mathcal{J}\left(i,d\right)\right)}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f\left(X_j\right)+\left(\frac{d}{dX_j}\log f\left(X_j\right)\right)^2\right)\right|\to_p 0.$$

By independence between the X_j's, $\mathrm{E}\left[S_i^2\left(d\right)\right]=\mathrm{E}\left[\left(\frac{f''(X)}{f(X)}\right)^2\right]/c\left(\mathcal{J}\left(i,d\right)\right)<\infty$ for all d. Then, as $a\to\infty$,

$$\sup_d \mathrm{E}\left[\left|S_i\left(d\right)\right|\mathbf{1}_{\{|S_i(d)|\ge a\}}\right]\le\sup_d\frac{1}{a}\mathrm{E}\left[\left(S_i\left(d\right)\right)^2\mathbf{1}_{\{|S_i(d)|\ge a\}}\right]\le\frac{K}{a}\to 0.$$

Since the uniform integrability condition is satisfied (see e.g., [15]), we find $\lim_{d\to\infty}\mathrm{E}\left[\left|S_i\left(d\right)\right|\right]=\mathrm{E}\left[\lim_{d\to\infty}\left|S_i\left(d\right)\right|\right]=0$. □

Proposition C.2. *Let* $\mathbf{Y}^{(d)}\left|\mathbf{X}^{(d)}\right.\sim N\left(\mathbf{X}^{(d)},\sigma^2\left(d\right)I_d\right)$*, where* X_j *is distributed according to the density* $\theta_j\left(d\right)f\left(\theta_j\left(d\right)x_j\right)$ *for* $j=1,\ldots,d$*. If* $\varepsilon\left(d,X_j,Y_j\right)$ *is as in Lemma B.1, then we have*

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

Proof. Developing the expectation on the LHS and simplifying the integrand yield

$$\int\int\Phi\left(\frac{\log\prod_{j=1,j\neq i^*}^{n}\frac{f(\theta_j(d)x_j)}{f(\theta_j(d)y_j)}-\frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}{\sqrt{\ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)}}\right)$$
$$\prod_{j=1,j\neq i^*}^{n}\theta_j\left(d\right)f\left(\theta_j\left(d\right)y_j\right)C\exp\left(-\frac{1}{2\sigma^2\left(d\right)}\sum_{j=1,j\neq i^*}^{n}\left(x_j-y_j\right)^2\right)d\mathbf{y}^{(n)-}d\mathbf{x}^{(n)-},$$

where C is a normalizing constant. Using Fubini's Theorem and swapping $\mathbf{y}^{(n)-}$ and $\mathbf{x}^{(n)-}$ then yield the desired result. □

Proposition C.3. *Let X_j be distributed according to the density $\theta_j(d) f(\theta_j(d) x_j)$ for $j = 1, \ldots, d$ and $\mathbf{Y}^{(d)} \Big| \mathbf{X}^{(d)} \sim N\left(\mathbf{X}^{(d)}, \sigma^2(d) I_d\right)$. We have*

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

Proof. Developing the LHS leads to

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

where C is a normalizing constant. 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 completed. His expertise, guidance and encouragements have been precious throughout my studies. I also acknowledge useful conversations with Professor Gareth O. Roberts, and am grateful to the Associate Editor and anonymous referee whose comments led to an improved version of the paper.

References

[1] Bédard, M. (2006). Weak Convergence of Metropolis Algorithms for Non-*iid* Target Distributions. *Ann. Appl. Probab.* **17**, 1222-44.

[2] Bédard, M. (2006). Efficient Sampling using Metropolis Algorithms: Applications of Optimal Scaling Results. *To appear in J. Comput. Graph. Statist.*

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

[6] 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.

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

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

[9] 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.

[10] Neal, P., Roberts, G.O. (2007). Optimal Scaling for Partially Updating MCMC Algorithms. *Ann. Appl. Probab.* **16**, 475-515.

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

[12] 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.

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

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

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