Hierarchical models and the tuning of random walk Metropolis algorithms

Mylène Bédard
Département de mathématiques et de statistique
Université de Montréal
Montréal, Canada, H3C 3J7
bedard@dms.umontreal.ca

Abstract

We obtain weak convergence and optimal scaling results for the random walk Metropolis algorithm with a Gaussian proposal distribution. The sampler is applied to hierarchical target distributions, which form the building block of many Bayesian analyses. The global asymptotically optimal proposal variance derived may be computed as a function of the specific target distribution considered. We also introduce the concept of locally optimal tunings, *i.e.* tunings that depend on the current position of the Markov chain. The theorems are proved by studying the generator of the first and second components of the algorithm, and verifying their convergence to the generator of a modified RWM algorithm and a diffusion process, respectively. The rate at which the algorithm explores its state space is optimized by studying the speed measure of the limiting diffusion process. We illustrate the theory with two examples. Applications of these results on simulated and real data are also presented.

AMS 2000 subject classifications: Primary 60F05; secondary 65C40.

Keywords: acceptance probability; diffusion process; efficiency; position-dependent reversibility; RWM-within-Gibbs; weak convergence

1. Introduction

Randow walk Metropolis (RWM) algorithms are widely used to sample from complex or multidimensional probability distributions ([15], [12]). The simplicity and versatility of these samplers often make them the default option in the MCMC toolbox. Implementing a RWM algorithm involves a tuning step, to ensure that the process explores its state space as fast as possible, and that the sample produced be representative of the probability distribution of interest (the target distribution). In this paper, we solve an aspect of the tuning problem for a large class of target distributions with correlated components. This issue has mainly been studied for product target densities, but attention has recently turned towards more complex target models ([7], [14]). The specific type of target distribution considered here is formed of components which are related according to a hierarchical structure. These distributions are ubiquitous in several fields of research (finance, biostatistics, physics, to name a few), and constitute the basis of many Bayesian inferences.

Bayesian hierarchical models are comprised of a likelihood function $f(\mathbf{d}|\theta)$, which is the statistical model for the observed data $\mathbf{d}$. The parameters θ are then modeled using a prior distribution $\pi(\theta|\rho)$; since this prior might not be easy to determine, it is common practice to assume that the hyperparameters ρ are themselves distributed according to a non-informative prior distribution $\pi(\rho)$. The various models thus represent different levels of hierarchy and give rise to a posterior distribution $\pi(\theta, \rho|\mathbf{d})$, which is often quite complex. Most of the time, this distribution cannot be studied analytically or sampled directly, and thus simulation algorithms such as MCMC methods are required to perform a statistical analysis. Samplers such as the RWM, RWM-within-Gibbs, and Adaptive Metropolis (see [11]) are usually the default algorithms for such targets.

The idea behind RWM algorithms is to build a Markov chain having the Bayesian posterior (target) distribution as its stationary distribution. To implement this method, users must select a proposal distribution from which are generated candidates for the Markov chain. This distribution should ideally be similar to the target, while remaining accessible from a sampling viewpoint. A pragmatic choice is to let the proposed moves be normally distributed around the latest value of the sample. Tuning the variance of the normal proposal distribution (σ^2) has a significant impact on the speed at which the sampler explores its state space (hereafter referred to as "efficiency"), with extremal variances leading to slow-mixing algorithms. In particular, large variances seldom induce suitable candidates and result in lazy processes; small variances yield hyperactive processes whose tiny steps lead to a time-consuming exploration of the state space. Seeking for an intermediate value that optimizes the efficiency of the RWM algorithm, *i.e.* a proposal variance σ^2 offering sizable steps that are still accepted a reasonable proportion of the time, is called the optimal scaling problem.

The optimal scaling issue of the RWM algorithm with a Gaussian proposal has been addressed by many researchers over the last few decades. It has been determined in [17] that target densities formed of independent and identically distributed (i.i.d.) components correspond to an optimal proposal variance $\hat{\sigma}^2(n) \approx 5.66/\{n\mathbb{E}[(\log f(X))']\}$, where f is the density of one target component and n the number of target components. This optimal proposal variance has also be shown to correspond to an optimal expected acceptance rate of 23.4%, where the acceptance rate is defined as the proportion of candidates that are accepted by the algorithm. Generalizing this conclusion is an intricate task and further research on the subject has mainly been restricted to the case of target distributions formed of independent components (see [18], [16], [2], [3], [5], [6]). In the specific case of multivariate normal target distributions however, the optimal variance and acceptance rate may be easily determined (see [16], [1]). Lately, [7] and [14] have also performed scaling analyses of non-product target densities. These advances are important, as MCMC methods are mainly used when dealing with complex models, which only rarely satisfy the independence assumption among target components. These results however assume that the correlation structure among target components is known and used in generating candidates for the chain. This is a restrictive assumption that leads, as expected, to an optimal acceptance rate of 23.4% (see [18] for an explanation).

In this paper, we focus on solving the optimal scaling problem for a wide class of models that include a dependence relationship, the hierarchical distributions. Weak convergence results are derived without explicitly characterizing the dependency among target components, and thus rely on a Gaussian proposal distribution with diagonal covariance matrix. The optimal proposal variance may then be obtained from these results, *i.e.* by maximizing the speed measure of the limiting diffusion process. This constitutes significant advances in understanding the theoretical underpinnings of the RWM sampler. More importantly in practice, the results theoretically support the use of RWM-within-Gibbs over RWM samplers and provide a convenient approach for obtaining a new type of proposal variances. These proposal variances are a function of the current state of the Markov chain; they thus evolve with the chain and lead to more appropriate candidates in the RWM-within-Gibbs algorithm.

In the next section, we describe the target distribution and introduce some notation related to the RWM sampler. The theoretical optimal scaling results are stated in Section 3, and then illustrated with two examples using RWM samplers in Section 4. In Section 5, the potential of RWM-within-Gibbs with local scalings is illustrated in Bayesian contexts through a simulation study and an application on real data. Extensions are briefly discussed in Section 6, while appendices contain proofs.

2. Framework

Consider an n-dimensional target distribution consisting of a mixing component X_1 and of $n-1$ conditionally i.i.d. components X_i $(i = 2, \ldots, n)$ given X_1. Suppose that this distribution has a target density π with respect to Lebesgue measure, where

$$\pi\left(\mathbf{x}\right) = f_1\left(x_1\right) \prod_{i=2}^{n} f\left(x_i \left| x_1\right.\right) \ . \tag{1}$$

To obtain a sample from the target density in (1), we rely on a RWM algorithm with a Gaussian proposal distribution. This sampler builds an n-dimensional Markov chain $\{\mathbf{X}^{(n)}[j]; j \in \mathbb{N}\}$ having $\pi(\mathbf{x})$ as its stationary density. Given $\mathbf{X}^{(n)}[j] = \mathbf{x}$, the time-$j$ state of the Markov chain, one iteration is performed according to the following steps:

1. generate a candidate $\mathbf{Y}^{(n)}[j+1] = \mathbf{y}$ from a $\mathcal{N}(\mathbf{x}, D_n)$, where D_n is a diagonal variance matrix with elements $\left(\sigma_1^2(n), \sigma^2(n), \ldots, \sigma^2(n)\right)$. In particular, set $D_n = \ell^2 I_n/n$, where $\ell > 0$ is a tuning parameter and I_n the n-dimensional identity matrix;

2. compute the acceptance probability $\alpha(\mathbf{x},\mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right\}$;

3. generate $U[j+1] \sim \mathcal{U}(0,1)$;

4. if $U[j+1] \leq \alpha(\mathbf{x},\mathbf{y})$, accept the candidate and set $\mathbf{X}^{(n)}[j+1] = \mathbf{y}$; otherwise, the Markov chain remains at the same state for another time interval and $\mathbf{X}^{(n)}[j+1] = \mathbf{x}$.

Optimal scaling results widely rely on the use of Gaussian proposal distributions which, due to their symmetry, lead to a simplified form of the acceptance probability. Although generally not emphasized in the literature, we note that the proposal variance could also be a function of $\mathbf{x}$, which would result in a non-homogeneous random walk sampler. In that case, there would be no simplification in the Metropolis-Hastings acceptance probability and Step 2 would then replaced by

$$\alpha(\mathbf{x},\mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})q_n(\mathbf{x};\mathbf{y})}{\pi(\mathbf{x})q_n(\mathbf{y};\mathbf{x})}\right\} ,$$

where $q_n(\mathbf{y};\mathbf{x})$ is the density of a $\mathcal{N}(\mathbf{x}, D_n(\mathbf{x}))$.

In what follows we work towards finding the optimal value of ℓ, *i.e.* leading to an optimally mixing chain. The proofs of the theoretical results rely on CLTs and LLNs; as such, the results are obtained by letting $n \to \infty$. This is a common approach in MCMC theory and does not prevent users from applying the asymptotically optimal value of ℓ in lower dimensional contexts (as small as $n = 10$ or 15). Indeed, a particularity of optimal scaling results is that the asymptotic behaviour kicks in extremely rapidly, as shall be witnessed in the examples of Section 4.

The first thought of most MCMC users when facing a target density as in (1) would be to use a RWM-within-Gibbs algorithm, which consecutively updates subgroups of the n components in a given iteration. The tuning of RWM-within-Gibbs algorithms has been addressed in [16], but only for target distributions with i.i.d. components and Gaussian targets with correlation. Focusing on RWM algorithms is thus a good starting point to understand the behaviour of samplers applied to hierarchical target distributions. The results expounded in this paper lead to the concept of local tunings, which is particularly appealing in the context of RWM-within-Gibbs. Incidentally, the proofs in appendices provide a theoretical justification for the use of locally optimal scalings in RWM-within-Gibbs, see [4]. These findings are illustrated in the examples of Section 5.

In Sections 2.1, 2.2, and 3, we expound how to obtain asymptotically optimal variances D_n and $D_n(\mathbf{x})$ for RWM and RWM-within-Gibbs, respectively. Section 2.1 describes the regularity conditions imposed on $\pi(\mathbf{x})$, while Section 2.2 explains why the proposal matrix $D_n = \ell^2 I_n/n$ is the optimal choice for obtaining the theoretical results that shall be presented in Section 3.

2.1. Assumptions on the target density

To characterize the asymptotic behaviour of the conditionally i.i.d. components X_i ($i = 2, \ldots, n$), we impose some regularity conditions on the densities f_1 and f in (1). The density f_1 is assumed to be a continuous function on $\mathbb{R}$, with $\mathcal{X}_1 = \{x_1 : f_1(x_1) > 0\}$ forming an open interval.

For all fixed $x_1 \in \mathcal{X}_1$, $f(x|x_1)$ is a positive $\mathcal{C}^2$ density on $\mathbb{R}$ and $\frac{\partial}{\partial x}\log f(x|x_1)$ is Lipschitz continuous with constant $K(x_1)$ such that $\mathbb{E}[K^2(X_1)] < \infty$. Here, $\mathcal{C}^2$ denotes the space of real-valued functions with continuous second derivative. For all fixed $x \in \mathcal{X} = \mathbb{R}$, $f(x|x_1)$ is a $\mathcal{C}^2$ function on $\mathcal{X}_1$ and $\frac{\partial}{\partial x}\log f(x|x_1)$ is Lipschitz continuous with constant $L(x)$ such that $\mathbb{E}[L^4(X)] < \infty$. Furthermore,

$$\mathbb{E}_X\left[\left(\tfrac{\partial}{\partial X}\log f(X|x_1)\right)^4\right] < \infty \ \forall x_1 \in \mathcal{X}_1 \quad \text{with} \quad \mathbb{E}\left[\left(\tfrac{\partial}{\partial X}\log f(X|X_1)\right)^4\right] < \infty \ ; \tag{2}$$

hereafter, the notation $\mathbb{E}_X[\cdot]$ means that the expectation is computed with respect to X conditionally on the other variables in the expression; the first expectation in (2) is thus obtained according to the conditional distribution of X given X_1. Where there is no confusion possible, $\mathbb{E}[\cdot]$ shall be used to denote an expectation with respect to all random variables in the expression. The above regularity conditions constitute an extension of those stated in [3] for target distributions with independent components, and are weaker than would be a Lipschitz continuity assumption on the bivariate function $\frac{\partial}{\partial x}\log f(x|x_1)$. They also imply that the Lipschitz constants $K(x_1)$ and $L(x)$ themselves satisfy a Lipschitz condition.

We now impose further conditions on $f(x|x_1)$ to account for the movements of the coordinate X_1 when studying the asymptotic behaviour of a component X_i ($i = 2, \ldots, n$). These movements should not be too abrupt so for almost all fixed $x \in \mathcal{X}$, $\frac{\partial}{\partial x_1}\log f(x|x_1)$ is Lipschitz continuous with constant $M(x)$ such that $\mathbb{E}[M^2(X)] < \infty$ and

$$\mathbb{E}_X\left[\left(\tfrac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right] < \infty \ \forall x_1 \in \mathcal{X}_1 \quad \text{with} \quad \mathbb{E}\left[\left(\tfrac{\partial}{\partial X_1}\log f(X|X_1)\right)^2\right] < \infty \ . \tag{3}$$

Finally, in order to characterize the asymptotic behaviour of the mixing component X_1, we introduce assumptions that are closely related to the Bernstein von Mises Theorem. Let $\mathbf{X}_{2:n} = (X_2, \ldots, X_n)$, $\underline{\mathbf{X}} = (X_2, X_3, \ldots)$, and $\to_p$ denote convergence in probability. Assume that $\mathbb{V}(X_1|\mathbf{X}_{2:n}) \to_p 0$, and denote $\mu \equiv \mu(\underline{\mathbf{X}})$ such that $\mu_n \equiv \mu_n(\mathbf{X}_{2:n}) = \mathbb{E}[X_1|\mathbf{X}_{2:n}] \to_p \mu$ as $n \to \infty$, with $|\mu| < \infty$. Hereafter, we make a small abuse of notation by letting μ and μ_n sometimes denote the random variable or the realisation, depending on the context. Furthermore, define $\tilde{X}_1 = \sqrt{n}(X_1 - \mu_n)$; for almost all $\mathbf{x}_{2:n} \in \mathbb{R}^{n-1}$, the conditional density of $\tilde{X}_1$ given $\mathbf{x}_{2:n}$, $f_1(\mu_n + \tilde{x}_1/\sqrt{n}|\mathbf{x}_{2:n})/\sqrt{n}$, is assumed to converge almost surely to $g_1(\tilde{x}_1|\underline{\mathbf{x}})$, a continuous density on $\mathbb{R}$ with respect to Lebesgue measure. In fact, the information on X_1 increases linearly in n, meaning that the limiting density of $X_1|\mathbf{x}_{2:n}$ is degenerate, but that a standard rescaling leads to a non-trivial density on $\mathbb{R}$ (normal distribution).

2.2. Form of the proposal variance matrix D_n

In Section 3, we focus on deriving weak convergence and optimal scaling results for the RWM algorithm with a Gaussian proposal by letting n, the dimension of the target density in (1), approach ∞. Traditionally, asymptotically optimal scaling results have been obtained by studying the limiting path of a given component (X_2 say) as $n \to \infty$. In the case of target distributions with i.i.d. components (and some extensions), the components of the RWM algorithm are asymptotically independent of each other and their limiting behaviour is regimented by identical one-dimensional Markovian processes. In the current correlated framework, we expect the presence of an asymptotic dependence relationship among X_i

($i \in \{2, \ldots, n\}$) and X_1, in the spirit of (1). In the following section, we thus study the limiting behaviour of components X_1 and X_2 separately, on their respective conditional space. This approach allows us to quantify the mixing rate of each component X_i conditionally on the others, and to propose optimal scalings for the sampler.

To obtain non-trivial limiting processes describing the behaviour of the RWM sampler as $n \to \infty$, we need to fix the form of the proposal scalings $\sigma_1^2(n), \sigma^2(n)$. Whilst the proposals are independent, a single accept-reject step is used, which makes the paths of the components dependent. We aim to choose the maximal scalings that avoid a degenerate limit (of either 0 or 1) for this acceptance probability. Since the distribution of X_1 conditional on $\mathbf{X}_{2:n}$ contracts at a rate of $\sqrt{n}$, then if $\sigma_1(n)/\sqrt{n} \to \infty$ the proposed jumps in X_1 will be too large. If $\sigma_1(n)/\sqrt{n} \to 0$, then the change in X_1 makes no contribution to the acceptance probability in the limit; to maximise movements we, therefore, require $\sigma_1(n) \propto 1/\sqrt{n}$. Now, the conditional distribution of $\mathbf{X}_{2:n}$ given X_1 does not contract with n. Nonetheless, when proposing jumps in $\mathbf{X}_{2:n}$ using $\sigma^2(n) = \sigma^2$, the odds of rejecting an n-dimensional candidate increase with n and lead to a degenerate (null) acceptance probability. To overcome this problem we then let the proposal variance be a decreasing function of the dimension. In fact, since Lipschitz conditions control the contribution to the accept-reject ratio coming from the movements of X_1, a similar argument to that which leads to $\sigma(n) \propto 1/\sqrt{n}$ in the case of i.i.d. targets applies again here. We therefore set $D_n = \ell^2 I_n/n$, where $\ell > 0$ is a tuning parameter and I_n the n-dimensional identity matrix.

As $n \to \infty$, it becomes necessary to speed up time to compensate for the reduced movement along components $\mathbf{X}_{2:n}$. The time interval between each proposed candidate is thus set to $1/n$ and we study the continuous-time, sped up version of the initial Markov chain defined as $\{\mathbf{W}^{(n)}(t); t \geq 0\} = \{\mathbf{X}^{(n)}[\lfloor nt \rfloor]; t \geq 0\}$, where $\lfloor \cdot \rfloor$ is the floor function. Similarly to the i.i.d. case, a limiting diffusion is obtained for the rescaled one-dimensional process related to X_i ($i \geq 2$), but this time its behaviour is conditional on X_1.

Since the first coordinate X_1 converges to a point μ, a transformation $\tilde{X}_1 = \sqrt{n}(X_1 - \mu_n)$ is required to obtain the limiting behaviour of this component. We thus study the continuous-time process $\{\tilde{\mathbf{W}}^{(n)}(t); t \geq 0\} = \{(\tilde{X}_1^{(n)}[\lfloor t \rfloor], \mathbf{X}_{2:n}^{(n)}[\lfloor t \rfloor]); t \geq 0\}$; in other words, we are now looking at a magnified, centered version of the path associated to X_1. This transformation leads to proposal distributions $\tilde{Y}_1 = \sqrt{n}(Y_1 - \mu_n) \sim \mathcal{N}(\tilde{x}_1, \ell^2)$ and $Y_i \sim \mathcal{N}(x_i, \ell^2/n)$, $i = 2, \ldots, n$ with $\ell > 0$; it thus cancels the effect of n in $\sigma_1^2(n)$. Without the speed up of time, the limiting process for $\tilde{X}_1$ is then a propose-accept-reject on the conditional density for $\tilde{X}_1$, given the current values of $\mathbf{X}_{2:n}$; this is made precise in Theorem 1. When considering the diffusion limit for X_i ($i \geq 2$) with time sped-up, this effectively means that at every instant, X_1 is simply a sample from its conditional distribution given the current values of $\mathbf{X}_{2:n}$; this is made precise in Theorem 2.

We note that an alternative scaling of $\sigma_1(n) \propto 1/n$ could also be applied. The sped-up limiting process would then be a diffusion for all coordinates, and would be easier to describe. However, this would also be a deliberate handicapping of the algorithm since the change in X_1 would make no contribution to the acceptance probability in the limit. A suboptimal $\sigma_1^2(n)$, besides altering the movements of X_1, would thus also indirectly affect the efficiency according to which $\mathbf{X}_{2:n}$ explores its state space.

3. Asymptotics of the RWM algorithm

In this section we introduce results about the limiting behaviour (as $n \to \infty$) of the time- and scale-adjusted univariate processes $\{\tilde{W}_1^{(n)}(t); t \geq 0\}$ and $\{W_i^{(n)}(t); t \geq 0\}$ $(i = 2, \ldots, n)$. From these results we determine the asymptotically optimal scaling (AOS) values and acceptance rate (AOAR) that optimize the mixing of the algorithm.

Hereafter, we let $\Rightarrow$ denote weak convergence in the Skorokhod topology and $B(t)$ a Brownian motion at time t; the cumulative distribution function of a standard normal random variable is denoted by $\Phi(\cdot)$.

Theorem 1. *Consider a RWM algorithm with proposal distribution $\mathcal{N}(\mathbf{x}, \ell^2 I_n/n)$ used to sample from a target density π as in (1). Suppose that π satisfies the conditions on f_1 and f specified in Section 2.1, and that $\mathbf{X}^{(n)}(0)$ is distributed according to π in (1).*

If $\frac{1}{n}\sum_{i=2}^{n}\left(\frac{\partial}{\partial X_i}\log f(X_i|X_1 = \mu_n + \frac{\tilde{X}_1}{\sqrt{n}})\right)^2 \to_p \tilde{\gamma}(\mu)$ with

$$\tilde{\gamma}(\mu) \;=\; \mathbb{E}_X\left[\left(\frac{\partial}{\partial X}\log f(X|\mu(\underline{\mathbf{X}}))\right)^2\right] = \int_{\mathbb{R}}\left(\frac{\partial}{\partial x}\log f(x|\mu(\underline{\mathbf{X}}))\right)^2 f(x|\mu(\underline{\mathbf{X}}))\mathrm{d}x < \infty \;,$$

then the magnified process $\{\tilde{W}_1^{(n)}(t); t \geq 0\} \Rightarrow \{\tilde{W}_1(t); t \geq 0\}$. Here, $W_1(0)$ and $W_i(0)$ $(i = 2, 3, \ldots)$ are distributed according to the densities f_1 and f respectively, which implies that $\tilde{W}_1(0)$ is distributed according to the density g_1 in Section 2.1. Given the time-t state $\tilde{\mathbf{W}}(t) = (\tilde{x}_1, \underline{\mathbf{x}})$, the process $\{\tilde{W}_1(t); t > 0\}$ evolves as the continuous-time version of a special RWM algorithm applied to the target density $g_1(\tilde{x}_1|\underline{\mathbf{x}})$; the proposal distribution of this algorithm is a $\mathcal{N}(\tilde{x}_1, \ell^2)$ and the acceptance rule is defined as

$$\alpha^*(\tilde{x}_1, \tilde{y}_1|\underline{\mathbf{x}}) \;=\; \Phi\left(\frac{\log\frac{g_1(\tilde{y}_1|\underline{\mathbf{x}})}{g_1(\tilde{x}_1|\underline{\mathbf{x}})} - \frac{\ell^2}{2}\tilde{\gamma}(\mu)}{\ell\tilde{\gamma}^{1/2}(\mu)}\right) + \frac{g_1(\tilde{y}_1|\underline{\mathbf{x}})}{g_1(\tilde{x}_1|\underline{\mathbf{x}})}\Phi\left(\frac{-\log\frac{g_1(\tilde{y}_1|\underline{\mathbf{x}})}{g_1(\tilde{x}_1|\underline{\mathbf{x}})} - \frac{\ell^2}{2}\tilde{\gamma}(\mu)}{\ell\tilde{\gamma}^{1/2}(\mu)}\right) . \quad (4)$$

Proof. See Appendix A.1. □

This result describes the limiting path associated to the coordinate $\tilde{X}_1$ as $n \to \infty$, which is Markovian with respect to the history of the multidimensional chain $\mathcal{F}^{\tilde{\mathbf{W}}}(t)$. We recall that the conditional distribution of X_1 given $\mathbf{X}_{2:n}$ contracts at a rate of $\sqrt{n}$ and that $\sigma_1(n) \propto 1/\sqrt{n}$. Conditionally on $\underline{\mathbf{X}}$, the transformed $\tilde{X}_1$ thus mixes according to $\mathcal{O}(1)$ and explores its conditional state space much more efficiently than the other components, as shall be witnessed in Theorem 2. The asymptotic process found can be described as an atypical one-dimensional RWM algorithm, whose acceptance rule $\alpha^*(\tilde{x}_1, \tilde{y}_1|\underline{\mathbf{x}})$ and target density $g_1(\tilde{x}_1|\underline{\mathbf{x}})$ both vary according to $\underline{\mathbf{x}}$ at every iteration. The acceptance function α^* in (4) satisfies the reversibility condition with respect to $g_1(\tilde{x}_1|\underline{\mathbf{x}})$ (see [3] for more details about this acceptance function).

Theorem 1 is interesting from a theoretical perspective, but cannot be used to optimize the global mixing of the algorithm. Although we could try to determine the value of ℓ leading to the optimal mixing of X_1 on its conditional space, it will be wiser to focus instead on optimizing the mixing rate of $\mathbf{X}_{2:n}$ on its own conditional space given X_1. Since the distribution of X_1 contracts about μ_n, the position of this coordinate heavily depends on the current state of $\mathbf{X}_{2:n}$. We shall also see in Theorem 2 that given X_1, the coordinates X_i ($i \geq 2$) explore their conditional state space according to $\mathcal{O}(n)$. Since these coordinates take more time exploring their conditional distribution and heavily affect the position of X_1, then the global performance of the sampler is subjected to the mixing of $\mathbf{X}_{2:n}$ conditionally on X_1.

Theorem 2. *Consider a RWM algorithm with proposal distribution $\mathcal{N}(\mathbf{x}, \ell^2 I_n/n)$ used to sample from a target density π as in (1). Suppose that π satisfies the conditions on f_1 and f specified in Section 2.1, and that $\mathbf{X}^{(n)}(0)$ is distributed according to π in (1).*

For $i = 2, \ldots, n$, we have $\{W_i^{(n)}(t); t \geq 0\} \Rightarrow \{W_i(t); t \geq 0\}$, where $W_i(0)$ $(i \geq 2)$ is distributed according to f, and $W_1(0)$ according to f_1. Conditionally on $W_1(t)$, the evolution of $\{W_i(t); t > 0\}$ over an infinitesimal interval $\mathrm{d}t$ *satisfies*

$$\mathrm{d}W_i(t) = \upsilon^{1/2}(\ell, W_1(t))\mathrm{d}B(t) + \frac{1}{2}\upsilon(\ell, W_1(t))\tfrac{\partial}{\partial W_i(t)} \log f\left(W_i(t)|W_1(t)\right)\mathrm{d}t, \tag{5}$$

with

$$\upsilon(\ell, x_1) = 2\ell^2 \mathbb{E}_{Z_1}\left[\Phi\left(-\tfrac{\ell}{2}\gamma^{1/2}(x_1, Z_1)\right)\right], \tag{6}$$

$Z_1 = \sqrt{n}(Y_1 - x_1)/\ell \sim \mathcal{N}(0,1)$, and

$$\gamma(x_1, z_1) \;\;=\;\; z_1^2\ \mathbb{E}_X\left[\left(\tfrac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right] + \mathbb{E}_X\left[\left(\tfrac{\partial}{\partial X}\log f(X|x_1)\right)^2\right]. \tag{7}$$

Proof. See Appendix A.2. □

Equation (5) describes the behaviour of the process at the next instant, $(t + \mathrm{d}t)$, given its position at t. This expression should not come as a surprise: each rescaled component X_i ($i = 2, \ldots, n$) asymptotically behaves according to a diffusion process that is Markovian with respect to $\mathcal{F}^{(W_1, W_i)}(t)$. Examination of (5) also tells us that $f(W_i(t)|W_1(t))$ is invariant for this diffusion process (see [19], for instance). We finally recall that $\sigma(n) \propto 1/\sqrt{n}$ and therefore, conditionally on X_1, the rescaled X_i mixes according to $\mathcal{O}(n)$. Each coordinate X_i thus requires more iterations than were required by the coordinate X_1 to explore its conditional state space.

Since X_1 and X_i ($i \geq 2$) use different time rescaling factors, the asymptotic behaviour of these coordinates cannot be expressed as a bivariate diffusion process. To obtain such a diffusion, we would have to rely on inhomogeneous proposal variances to ensure that X_1 also mixes in $\mathcal{O}(n)$ iterations; as mentioned at the end of Section 2, this would require setting $\sigma_1(n) = \ell/n$, $\sigma(n) = \ell/\sqrt{n}$ for $\ell > 0$. This framework would of course be suboptimal as it would restrain the X_1 movements. Proposed jumps for X_1 would then become insignificant, and so the first term in (7) would be null.

Remark 3. *Studying the limiting behaviour of X_1 and X_i ($i = 2, \ldots, n$) separately does not cause information loss. In fact, studying the paths of X_1, X_2 simultaneously would require letting the test function h of the generator in* (A.3) *be a function of (X_1, X_2). Such a generator would however be developed as an expression in which cross-derivative terms (e.g. $\frac{\partial^2}{\partial x_1 \partial x_2} h(x_1, x_2)$) are null, which confirms that given the current state of the asymptotic process, one-dimensional moves are performed independently for each coordinate.*

The limiting processes in Theorems 1 and 2 indicate that the component X_1 explores its conditional state space at a different (higher) rate than $\mathbf{X}_{2:n}$ explores its own. Combined to the specific Markovian forms of the limiting processes obtained (with respect to $\mathcal{F}^{\tilde{\mathbf{W}}}(t)$ and $\mathcal{F}^{(W_1,W_i)}(t)$ respectively), this points towards the need for updating X_1 and $\mathbf{X}_{2:n}$ separately, assessing the superiority of RWM-within-Gibbs samplers for sampling from hierarchical targets. These algorithms update blocks of components successively, a design that allows fully exploiting the characteristics of the target considered. To our knowledge, this is the first time that asymptotic results are used to theoretically validate the superiority of RWM-within-Gibbs over RWM samplers for hierarchical target distributions. This theoretical superiority is obviously tempered in practice by an increased computational effort; the extent of this computational overhead is however difficult to quantify in full generality. To this end, Section 5 presents two examples that illustrate the performance of the RWM-within-Gibbs and compare it to RWM and Adaptive Metropolis samplers.

3.1. Optimal tuning of the RWM algorithm

To be confident that the n-dimensional chain has entirely explored its state space, we must be certain that every one-dimensional path has explored its own space. In the correlated framework considered, the overall mixing rate of the RWM sampler is only as fast as the slowest component. As explained in Section 3, optimal mixing of the algorithm shall be attained by optimizing the mixing of the coordinates X_i, $i = 2, \ldots, n$. In the limit, the only quantity that depends on the proposal variance (*i.e.* on ℓ) is $\upsilon(\ell, W_1(t))$ in (6). To optimize mixing, it thus suffices to find the diffusion process that goes the fastest, *i.e.* the value of ℓ for which the speed measure $\upsilon(\ell, W_1(t))$ is optimized.

The speed measure in (6) is quite intuitive; it is in fact similar to that obtained when studying i.i.d. target densities. The main difference lies in the form of $\gamma(x, z)$ which, in the i.i.d. case, is given by the constant term $\gamma = \mathbb{E}[(\frac{\partial}{\partial X} \log f(X))^2]$. The second term in (7) is thus equivalent to γ, and consists in a measure of roughness of the conditional density $f(x_i|x_1)$ under a variation of x_i $(i \geq 2)$. In the case of hierarchical target distributions, we find an extra term that might be viewed as a measure of roughness of $f(x_i|x_1)$ under a variation of x_1. This term is weighted by z_1^2, the square of the (standardized) candidate increment for the first component; in other words, the further the candidate y_1 is from the current x_1, the greater is the weight attributed to the associated measure of roughness. Of course, in optimizing the speed measure function, we do not need to know in advance the exact value of the proposed standardized increment z_1; the speed measure averages over this quantity.

It is interesting to note that optimizing the speed measure leads to local proposal variances of the form $\hat{\ell}^2(W_1(t))/n$. Such proposal variances would then be used for proposing a candidate at the next instant $t + \mathrm{d}t$, given the position of the mixing coordinate at time t. These local proposal variances thus vary from one iteration to another, by opposition to usual tunings in the literature that are fixed for the duration of the algorithm. Naturally, if both expectations in (7) are constant with respect to x_1, then the proposal variance obtained by maximizing the speed measure also is constant.

Remark 4. *It turns out that local proposal variances optimizing* (6) *are bounded above by* $2.38/\mathbb{E}_X^{1/2}[(\frac{\partial}{\partial X} \log f(X|x_1))^2]$*, the asymptotically optimal scaling (AOS) values for targets* *with i.i.d. components given a fixed $X_1 = x_1$. Indeed, if $X_1 = x_1$ is fixed across iterations, we find ourselves in an i.i.d. setting and the associated speed measure is expressed as $2\ell^2\Phi(-\ell\mathbb{E}_X^{1/2}[(\frac{\partial}{\partial X}\log f(X|x_1))^2]/2)$. The mentioned upper bounds then follow from the fact that the function $\Phi(\cdot)$ in (6) decreases faster in ℓ than $\Phi(\cdot)$ in the above expression.*

Relying on a local variance $\hat{\ell}(x_1)$ to propose a candidate for the next time interval is usually time-consuming, as it involves numerically solving for the appropriate local proposal variance at every iteration. Since the process is assumed to start in stationarity and X_1 explores its conditional state space faster than the other coordinates, we might determine a value $\hat{\ell}$ that is fixed across iterations by integrating the speed measure $\upsilon(\ell,\cdot)$ over $\mathcal{X}_1$ with respect to the marginal distribution f_1. Hence, the global (unconditional) asymptotically optimal scaling value $\hat{\ell}$ maximizes the function

$$\begin{aligned}\mathbb{E}_{X_1}\left[\upsilon\left(\ell, X_1\right)\right] &= 2\ell^2\mathbb{E}_{X_1,Z_1}\left[\Phi\left(-\frac{\ell}{2}\gamma^{1/2}(X_1,Z_1)\right)\right] \\ &= 2\ell^2\int_{\mathcal{X}_1}\int_{\mathbb{R}}\Phi\left(-\frac{\ell}{2}\gamma^{1/2}(x_1,z_1)\right)\phi(z_1)f_1(x_1)\,\mathrm{d}z_1\,\mathrm{d}x_1\ ,\end{aligned}$$

where $\phi(\cdot)$ is the probability density function of a standard normal random variable.

Remark 5. *The asymptotic process introduced in Theorem 2 naturally leads us to the concept of local proposal variances. It is however unclear whether the local tunings obtained by maximizing* (6) *really optimize the mixing rate of the algorithm. Indeed, the proof of Theorem 2 is carried out with ℓ^2 constant; this allows, among other things, relying on the simplified form for the acceptance probability. In order to claim that the local proposal variances obtained are optimal, a weak convergence result would need to be proven using a general proposal variance of the form $\sigma^2(n,x_1) = \ell^2(x_1)/n$. This extension is not trivial, as the ratio of proposal densities $q_n(\mathbf{x};\mathbf{y})/q_n(\mathbf{y};\mathbf{x})$ would then need to be included in the acceptance probability. Since the concept of locally optimal proposal variances is numerically demanding in the current framework, we choose to focus on ℓ^2 constant.*

In RWM-within-Gibbs, the blocks X_1 and $\mathbf{X}_{2:n}$ are updated consecutively and the situation is therefore different. In that case, local variances of the form $\sigma^2(n,x_1) = \ell^2(x_1)/n$ obtained by maximizing (6) *may be used to update the block $\mathbf{X}_{2:n}$. Since X_1 is updated separately, the first term in* (7) *is null, which makes local variances easier to compute. Furthermore, since local variances only depend on X_1 (which is updated separately), the ratio $q_n(\mathbf{x};\mathbf{y})/q_n(\mathbf{y};\mathbf{x})$ is equal to 1 and does not need to be included in the acceptance probability. Local variances are thus very appealing in that context and shall be studied in Section 5.*

Rather than tuning the sampler using the global AOS value, one may instead monitor the acceptance rate in order to work with an optimally mixing version of the RWM algorithm. To express optimal scaling results in terms of acceptance rates, we introduce the expected acceptance rate of the n-dimensional stationary RWM algorithm with a normal proposal:

$$a_n(\ell) = \int\int\alpha(\mathbf{x},\mathbf{y})\left(\frac{\ell}{\sqrt{n}}\right)^{-n}\phi_n\left(\frac{\mathbf{y}-\mathbf{x}}{\ell/\sqrt{n}}\right)\pi(\mathbf{x})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x}\ ,$$

where $\phi_n(\cdot)$ denotes the probability density function of an n-dimensional standard normal random variable. Optimal mixing results for the RWM sampler are summarized in the following corollary.

Corollary 6. *In the settings of Theorem 2, the global asymptotically optimal scaling value* $\hat{\ell}$ *maximizes*

$$2\ell^2 \int_{\mathcal{X}_1} \int_{\mathbb{R}} \Phi\left(-\frac{\ell}{2}\gamma^{1/2}(x,z)\right) \phi(z) f_1(x) \, \mathrm{d}z \, \mathrm{d}x \; .$$

Furthermore, we have that

$$\lim_{n\to\infty} a_n(\ell) = a(\ell) \equiv 2 \int_{\mathcal{X}_1} \int_{\mathbb{R}} \Phi\left(-\frac{\ell}{2}\gamma^{1/2}(x,z)\right) \phi(z) f_1(x) \, \mathrm{d}z \, \mathrm{d}x \; ,$$

and the corresponding asymptotically optimal acceptance rate is given by $a(\hat{\ell})$.

In contrast to the i.i.d. case, the AOAR found is not independent of the densities f_1 and f. Hence, there is not a huge advantage in choosing to tune the acceptance rate of the algorithm over the proposal variance; in fact, both approaches involve the same effort. Although it would also be possible to compute an overall acceptance rate associated to using local proposal variances, it could not be used to tune the algorithm. Building an optimal Markov chain based on local proposal variances would imply modifying the proposal variance at every iteration, which cannot be achieved by solely monitoring the acceptance rate.

For simplicity, the theoretical results expounded in this section attribute the same tuning constant ℓ to all n components. In practice, when a RWM algorithm is used to sample from a hierarchical target, users will likely want to use a different proposal variance for the mixing component X_1. In fact, the proofs of Theorems 1 and 2 easily generalize to the case of inhomogeneous proposal variances.

Corollary 7. *Let* $Y_1 \sim \mathcal{N}(x_1, \ell^2\kappa_1^2/n)$ *with* $0 < \kappa_1 < \infty$ *and* $\mathbf{Y}_{2:n} \sim \mathcal{N}(\mathbf{x}_{2:n}, \ell^2 I_{n-1}/n)$, *where* $Y_1, \mathbf{Y}_{2:n}$ *are independent. Then, Theorems 1 and 2 hold as stated, except that the limiting proposal distribution in Theorem 1 is* $\tilde{Y}_1 \sim \mathcal{N}(\tilde{x}_1, \ell^2\kappa_1^2)$ *and the random variable* Z_1 *in Theorem 2 is such that* $Z_1 \sim \mathcal{N}(0, \kappa_1^2)$.

In this paper, we consider the simple, yet useful hierarchical model described in (1) and featuring a single mixing component X_1. This is a natural starting point to study weak convergence of RWM algorithms for hierarchical targets, and even for correlated targets in general. There exist many generalizations of (1), just as there are many extensions of the proposal distribution considered. Some extensions of the hierarchical target are considered in the discussion, but we do not aim at presenting a detailed treatment of these cases.

4. Numerical studies

To illustrate the theoretical results of Section 3, we consider two toy examples: the first target distribution considered is a normal-normal hierarchical model in which the components $X_2, \ldots, X_n$ are related through their mean, while the second one is a gamma-normal hierarchical model in which $X_2, \ldots, X_n$ are related through their variance. In both cases, we show how to compute the optimal variance $\hat{\ell}$. We also study the performance of RWM samplers and conclude that even in relatively low-dimensional settings, the samplers behave according to the asymptotic results previously detailed.

4.1. Normal-normal hierarchical distribution

Consider an n-dimensional hierarchical target such that $X_1 \sim \mathcal{N}(0,1)$ and $X_i|X_1 \sim \mathcal{N}(X_1,1)$ for $i = 2,\dots,n$. To sample from this distribution, we use a RWM algorithm with a $\mathcal{N}(\mathbf{x}, \ell^2 I_n/n)$ proposal distribution. This simple target shall relate Theorem 2 to the theoretical results derived in [3].

Standard calculations lead to $X_1|\mathbf{X}_{2:n} \sim \mathcal{N}(\sum_{i=2}^n X_i/n, 1/n)$; as $n \to \infty$, $\mathbb{V}(X_1|\mathbf{X}_{2:n}) \to 0$ almost surely. If we let $\mu_n = \sum_{i=2}^n X_i/n$ and $\tilde{X}_1 = n^{1/2}(X_1 - \mu_n)$, then $\tilde{X}_1|\mathbf{X}_{2:n} \sim \mathcal{N}(0,1)$. Furthermore, the term $\sum_{i=2}^n (X_i - \mu_n - \tilde{X}_1/\sqrt{n})^2/n$ is reexpressed as $\sum_{i=2}^n (X_i - X_1)^2/n = \sum_{i=2}^n Z_i^2/n$ and thus converges in probability to $\mathbb{E}[Z^2] = \int (\frac{\partial}{\partial x} \log f(x|\mu))^2 f(x|\mu) \mathrm{d}x = 1$, where $Z_1,\dots,Z_n$ denote independent standard normal random variables. By Theorem 1, we can thus affirm that the component $\tilde{X}_1$ asymptotically behaves according to a one-dimensional RWM algorithm with a standard normal target and acceptance function as in (4); these do not, in the current case, depend on $\underline{\mathbf{x}}$.

Evaluating the function $\gamma(x_1, z_1)$ in (7) is a simple task and leads to $\gamma(x_1, z_1) = z_1^2 + 1$. The AOS value is then found by maximizing

$$\upsilon(\ell) = 2\ell^2 \mathbb{E}_{Z_1}\left[\Phi\left(-\frac{\ell}{2}\sqrt{Z_1^2+1}\right)\right]$$

with respect to ℓ, where $Z_1 \sim \mathcal{N}(0,1)$. This yields an AOS of $\hat{\ell}^2 = 4.00$ and a corresponding AOAR of $\upsilon(\hat{\ell})/\hat{\ell}^2 = 0.205$. These values are naturally smaller than those obtained for a target with i.i.d. components (5.66 and 0.234, respectively); indeed, the proposal distribution is formed of i.i.d. components and accordingly better suited for similar targets. Relying on a proposal with correlated components would however require a certain understanding of the target correlation structure, which goes against the general framework we wish to consider.

It is worth pointing out that the speed measure of the limiting diffusion process does not depend on X_1 in the present case. This holds for arbitrary densities f_1 and f satisfying the conditions in Section 2.1, provided that X_1 is a location parameter for X_i ($i \geq 2$). Since a variation in the location parameter does not perturb the roughness of the distribution, the AOS and AOAR found are valid both locally and globally. This means that $\hat{\ell}$, which remains fixed across iterations, is the best possible proposal scaling conditionally on the last position of the component X_1 (*i.e.* $\hat{\ell} = \hat{\ell}(x_1)$).

A second peculiarity of this example is that the target distribution is jointly normal with mean $\mathbf{0}$ and $n \times n$ covariance matrix Σ_n given by $\sigma_1^2 = 1$, $\sigma_j^2 = 2$ ($j = 2,\dots,n$), and $\sigma_{i,j} = 1$ $\forall i \neq j$ ($i,j = 1,\dots,n$). Normal distributions being invariant under orthogonal transformations, we can find a transformation under which the target components become mutually independent. The covariance matrix Σ_n is thus transformed into a diagonal matrix whose diagonal elements consist in the eigenvalues of Σ_n. In moderate to large dimensions, the eigenvalues can be approximated by $1/(n+1), (n+1), 1, \dots, 1$. It turns out that the optimal scaling problem for target distributions of this sort (*i.e.* formed of components that are i.i.d. up to a scaling term) has been studied in [1]. Solving for the AOS value and AOAR of the transformed target using Theorem 1 and Corollary 2 in [3] leads to values that are consistent with those obtained using Theorem 2 in Section 3.

To illustrate these theoretical results, we consider the 20-dimensional normal-normal target described above and run 50 RWM algorithms that differ by their proposal variance only.

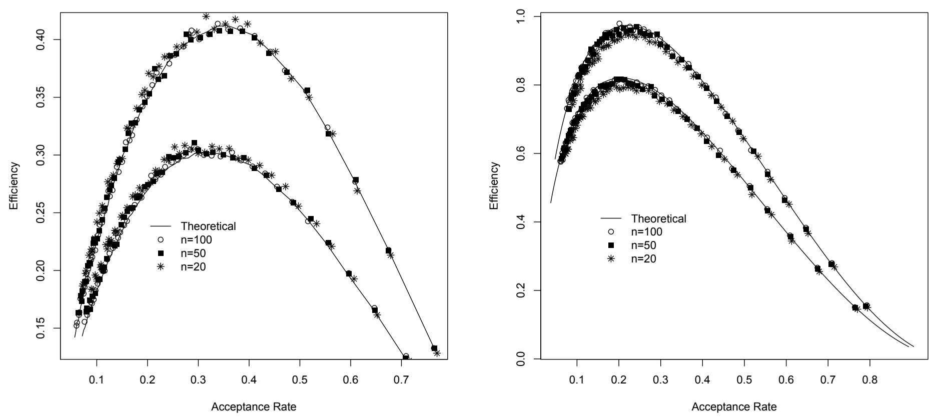

Figure 1: Efficiency of RWM algorithm against acceptance rate for the normal-normal hierarchical target. Left: efficiency of X_1 only; the top set of curves corresponds to homogeneous proposal variances. Right: efficiency of all n components; the top set of curves now corresponds to inhomogeneous proposal variances.

For each sampler, we perform 100,000 iterations (sufficient for convergence according to the autocorrelation function) and measure efficiency by recording the average squared jumping distance

$$\text{ASJD} \quad = \quad \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\left(x_i^{(n)}[j] - x_i^{(n)}[j-1]\right)^2; \tag{8}$$

here, N is the number of iterations and n is the dimension of the target distribution. We also record the average acceptance rate of each algorithm, expressed as

$$\text{AAR} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\mathbf{x}^{(n)}[j] \neq \mathbf{x}^{(n)}[j-1]\}.$$

We repeat these steps for 50- and 100-dimensional normal-normal targets, and combine all three curves of efficiency versus acceptance rate on a graph along with the theoretical efficiency curve of $\upsilon(\ell)$ versus the expected acceptance rate $\upsilon(\ell)/\ell^2$ (Figure 1, right graph, bottom set of curves). To assess the limiting behaviour of the coordinate X_1, we also plot the ASJD of this single component (for the 20-, 50-, and 100-dimensional cases) along with the ASJD for the limiting one-dimensional RWM sampler described in Theorem 1 (Figure 1, left graph, top set of curves).

We now repeat the numerical experiment by taking advantage of the available target variances in the tuning of the proposal distribution. Specifically, we let $Y_1 \sim \mathcal{N}(x_1, \ell^2/2n)$ be independent of $\mathbf{Y}_{2:n} \sim \mathcal{N}(\mathbf{x}_{2:n}, \ell^2/n)$ and run the RWM algorithm in dimensions 20, 50, and 100. The resulting simulated and theoretical efficiency curves are illustrated in Figure 1 (left graph, bottom set of curves; right graph, top set of curves). Although efficiency curves for X_1 are lower when using inhomogeneous proposal variances, this approach still results in a better overall performance (the curves in the right graph are higher than with homogeneous variances). The optimized theoretical efficiency is 0.974, which is related to an AOAR of

0.221. Despite the fact that Theorems 1 and 2 are valid asymptotically, the simulation study yields efficiency curves that are very close together; the theorems thus seem applicable in relatively low-dimensional settings.

Each set of curves on the right graph of Figure 1 agrees about the optimal acceptance rates 0.205 and 0.221, respectively. These optimal rates have been obtained by running an homogeneous sampler with optimal variance $\hat{\ell}^2/n = 4/n$ and an inhomogeneous sampler with optimal variance $4.4/n$, each optimizing (6). Any other proposal variance leads to a point that is lower on the efficiency curve.

According to the shape of these curves, tuning the acceptance rate anywhere between 0.15 and 0.3 would yield a loss of at most 10% in efficiency, and would still result in a Markov chain that rapidly explores its state space; in particular, using the usual 0.234 for this target would yield an almost optimal algorithm. Beyond finding the exact AOAR for a specific target distribution, there is thus a need for understanding when and why AOARs significantly differ from 0.234. At the present time, the only way to answer this question is by solving the optimal scaling problem for target distributions of interest.

4.2. Gamma-normal hierarchical distribution

As a second example, consider a gamma-normal hierarchical target such that $X_1 \sim \Gamma(\alpha, \lambda)$ and $X_i|X_1 \sim \mathcal{N}(0, 1/X_1)$, $i = 2, \ldots, n$. Although X_i ($i \geq 2$) are still normally distributed, the coordinate X_1 now acts through the variance of the normal variables. This results in a target that significantly differs from the distribution considered in the previous section, falling slightly outside the framework of Section 2 ($\frac{\partial}{\partial x_1} \log f(x|x_1)$ is now only locally Lipschitz continuous). We run the usual RWM algorithm to obtain a sample from this distribution.

Standard calculations lead to $X_1|\mathbf{X}_{2:n} \sim \Gamma(\alpha + (n-1)/2, \lambda + \sum_{i=2}^n X_i^2/2)$ and as $n \to \infty$, $\mathbb{V}(X_1|\mathbf{X}_{2:n}) \to_p 0$. The WLLN-type expression in Theorem 1 may be reexpressed as $\sum_{i=2}^n (\mu_n + \tilde{X}_1/\sqrt{n})^2 X_i^2/n = (\mu_n + \tilde{X}_1/\sqrt{n})(\sum_{i=2}^n Z_i^2/n)$, where $\mathbf{Z}_{1:n}$ are independent standard normal random variables. The condition is thus satisfied as it converges in probability to $\mu(\underline{\mathbf{X}}) = \mathbb{E}_X[(\frac{\partial}{\partial X} \log f(X|\mu))^2]$. Using Stirling's formula, it is not difficult to show that the density of $\tilde{X}_1|\mathbf{X}_{2:n}$ converges almost surely to that of a $\mathcal{N}(0, 2/\mu^2(\underline{\mathbf{X}}))$. By Theorem 1, the coordinate $\tilde{X}_1$ asymptotically behaves according to an atypical one-dimensional RWM algorithm with a normal target; the target variance however varies from one iteration to the next, and so does the acceptance function in (4).

To optimize the efficiency of the algorithm, we analyze the speed measure in (6); in the present case, it is expressed as

$$\upsilon(\ell, x_1) = 2\ell^2 \mathbb{E}_{Z_1}\left[\Phi\left(-\frac{\ell}{2}\sqrt{\frac{1}{2}\frac{Z_1^2}{x_1^2} + x_1}\right)\right],$$

where $Z_1 \sim \mathcal{N}(0,1)$. Maximizing the function $\mathbb{E}_{X_1}[\upsilon(\ell, X_1)]$ in Corollary 6 with respect to ℓ leads to the global AOS value, which is fixed across iterations; when $(\alpha, \lambda) = (3, 1)$ for instance, we find $\hat{\ell}^2 = 2.40$ and AOAR $= 0.204$.

The simulation study described in Section 4.1 has been performed for the gamma-normal target model with various α and λ. Specifically, for fixed α, λ, we consider a 10-dimensional

Table 1: Optimal efficiency and acceptance rate of chains in various dimension ($n = 10, 20, 50$), for different parameters α, λ of the gamma distribution for X_1. The theoretical optimal efficiency and acceptance rate are also included for comparison.

	Optimal efficiency				**Optimal acceptance rate**			
Parameters	Theoretical	$n = 10$	$n = 20$	$n = 50$	Theoretical	$n = 10$	$n = 20$	$n = 50$
$\alpha = 2, \lambda = 1$	0.6381	0.6036	0.6246	0.6456	0.1934	0.1984	0.1968	0.1857
$\alpha = 2, \lambda = 2$	0.8169	0.7430	0.7862	0.8239	0.1815	0.1888	0.1759	0.1886
$\alpha = 2, \lambda = 3$	0.8420	0.7623	0.8170	0.8608	0.1517	0.1682	0.1527	0.1593
$\alpha = 3, \lambda = 1$	0.4889	0.4503	0.4736	0.4926	0.2037	0.2370	0.2158	0.2001
$\alpha = 3, \lambda = 2$	0.7541	0.6739	0.7139	0.7405	0.2038	0.2265	0.2233	0.2040
$\alpha = 3, \lambda = 3$	0.8648	0.7554	0.8075	0.8497	0.1922	0.1931	0.1930	0.1882

gamma-normal target distribution and run 50 RWM algorithms possessing their own proposal variance. For each sampler, we perform 1,000,000 iterations (again sufficient for convergence according to the autocorrelation function) and measure efficiency by recording the ASJD of each chain. We then repeat these steps for 20- and 50-dimensional targets. Table 1 presents the optimal efficiency and acceptance rate for various α, λ. Those results are compared to the theoretical optimal values obtained by maximizing $\mathbb{E}_{X_1}[\upsilon(\ell, X_1)]$.

Although the corresponding graphs are omitted here, they yield curves similar to those obtained in Figure 1 for the normal-normal target. We note that even if the gamma-normal departs from a jointly normal distribution assumption and does not yield as nice a target distribution as in the previous example, the AOAR obtained is not too far from the 0.234 found for i.i.d. targets. The AOAR however tends to decrease as λ increases (*e.g.* 0.152 for $(\alpha, \lambda) = (2, 3)$).

In the current example, it also turns out that the agreement between theoretical and simulation results is altered for some values (α, λ). As mentioned above, one of the Lispchitz conditions is only valid locally and so the change in $\frac{\partial}{\partial x_1} \log f(x|x_1)$ becomes arbitrarily steep as $X_1 \to 0$. The amplitude of X_1 movements is, therefore, not adequately controlled for some choices of (α, λ) that yield a density f_1 assigning a significant probability close to 0. In cases where regularity assumptions are not all satisfied, the applicability of theoretical results may thus be affected by the choice of hyperparameters.

5. Applications in Bayesian contexts

The theoretical results presented in this paper have wide applicability and may be used to improve not only RWM algorithms, but other samplers as well (RWM-within-Gibbs, for instance). The examples below study the performance of optimally tuned samplers in the context of hierarchical Bayesian models. They show that the RWM-within-Gibbs sampler with local variances (*i.e.* variances that are a function of the current state of the chain) is superior to its counterpart with a fixed variance. It is also superior to traditional RWM algorithms and even Adaptive Metropolis (AM) samplers, which use the history of the chain to recursively update the covariance matrix of their proposal distribution (see [11]).

5.1. Scottish secondary school scores

The dataset `ScotsSec` in the package `mlmRev` in `R` contains the scores attained by 3,435 Scottish secondary school students on a standardized test taken at age 16. The primary schools attended by students are also recorded in this dataset; there are $n = 148$ different primary schools, and the number of students per primary school varies between 1 and 72. We use the following multilevel Bayesian framework to model these data

$$\begin{array}{ccccc} & & \mu & & \\ & \swarrow & \downarrow & \searrow & \\ \theta_1 & \ldots & & \ldots & \theta_n \\ \downarrow & & \downarrow & & \downarrow \\ Y_{1,1},\ldots,Y_{1,r_1} & & & & Y_{n,1},\ldots,Y_{n,r_n} \end{array} \qquad \begin{array}{l} \theta_i \sim t_\nu(\mu,1/\eta)\ ,\quad (1\le i\le n) \\ \\ Y_{ij}\sim \mathcal{N}(\theta_i,1/\tau)\ ,\quad (j=1,\ldots,r_i)\ . \end{array}$$

In this model, the variables $\mathbf{y}_{i,1:r_i}$ represent the observed scores obtained by the r_i students having attended primary school i, $i = 1, \ldots 148$. These observations are modeled according to a normal distribution with mean θ_i and variance $1/\tau$. The group sizes range from $r_{148} = 1$ to $r_{61} = 72$. The variables $\boldsymbol{\theta}_{1:148}$, which represent the mean scores of the standardized test for students having attended each of the 148 primary schools, are modeled using a Student distribution with $\nu = 4$ degrees of freedom. A translated and scaled Student distribution $t_\nu(\mu, 1/\eta)$ has a density proportional to $[1 + \eta(x - \mu)^2/\nu]^{-(\nu+1)/2}$. The mean and precision of the Student distribution, along with the precision of the normally distributed data, are attributed non-informative priors: $\pi(\mu) \propto 1$, $\pi(\eta) \propto \eta^{-1}$, and $\pi(\tau) \propto \tau^{-1}$.

This model leads to the $(n + 3)$-dimensional posterior density

$$\begin{aligned} \pi(\mu, \eta, \tau, \boldsymbol{\theta}_{1:n}|\{Y_{ij}\}) \quad \propto \quad & \eta^{-1}\ \tau^{-1} \prod_{i=1}^{n} \sqrt{\eta} \left[1 + \frac{\eta\,(\theta_i - \mu)^2}{\nu}\right]^{-(\nu+1)/2} \\ & \prod_{i=1}^{n}\prod_{j=1}^{r_i} \sqrt{\tau} \exp\left\{-\frac{\tau}{2}(y_{ij} - \theta_i)^2\right\}\ . \end{aligned} \tag{9}$$

The posterior density is too complex for analytic computation, and numerical integration must be ruled out due to the dimensionality of the problem. This distribution is best sampled with MCMC methods, although a classical Gibbs sampler must be ruled out, as the Student distribution destroys conjugacy. In the current setting, we propose to use a RWM-within-Gibbs with four blocks of variables: μ, η, τ, and $\boldsymbol{\theta}_{1:n}$. We are also interested in assessing the performance of full-dimensional RWM and AM algorithms in which μ, η, τ, and $\boldsymbol{\theta}_{1:n}$ are updated at once.

The RWM-within-Gibbs performs one-dimensional updates of μ, η, and τ using target densities $f(\mu|\eta, \tau, \boldsymbol{\theta}_{1:n}, \{Y_{ij}\})$, $f(\eta|\mu, \tau, \boldsymbol{\theta}_{1:n}, \{Y_{ij}\})$, and $f(\tau|\mu, \eta, \boldsymbol{\theta}_{1:n}, \{Y_{ij}\})$. It then performs an n-dimensional update of $\boldsymbol{\theta}_{1:n}$ with respect to the conditional density $f(\boldsymbol{\theta}_{1:n}|\mu, \eta, \tau, \{Y_{ij}\}) = \prod_{i=1}^{n} f(\theta_i|\mu, \eta, \tau, \mathbf{Y}_{i,1:r_i})$.

Since each block of variables is updated individually using a RWM sampler, we may compute local proposal variances for the fourth block using (6) and (7) in Theorem 2. The proposal variances maximizing (6) are adjusted according to the roughness of their corresponding target component's distribution, and should offer a better performance than a fixed proposal variance.

The target distribution of the fourth block satisfies

$$f(\boldsymbol{\theta}_{1:n}|\mu,\eta,\tau,\{Y_{ij}\}) \quad \propto \quad \prod_{i=1}^{n}\left[1+\frac{\eta\,(\theta_i-\mu)^2}{\nu}\right]^{-\frac{\nu+1}{2}} \exp\left\{-\frac{\tau}{2}\sum_{j=1}^{r_i}(y_{ij}-\theta_i)^2\right\} ,$$

hence the partial derivative of the one-dimensional log density with respect to θ_i is

$$\frac{\partial}{\partial\theta_i}\log f(\theta_i|\mu,\eta,\tau,\mathbf{Y}_{i,1:r}) = \tau\sum_{j=1}^{r_i}(y_{ij}-\theta_i) - \frac{\nu+1}{\nu}\sqrt{\eta}\left(\frac{T_i}{1+T_i^2/\nu}\right) , \tag{10}$$

where $T_i = \sqrt{\eta}(\theta_i-\mu) \sim t_\nu(0,1)$, $i = 1,\ldots,n$. Since the variables μ,η,τ are updated separately, then the first term in (7) is null, leading to

$$\gamma_i(\mu,\eta,\tau) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}f(\theta_i|\mu,\eta,\tau,\mathbf{Y}_{i,1:r_i})\right)^2\right] . \tag{11}$$

Optimizing (6) leads to local, inhomogeneous proposal variances of the form $2.38^2/\{n\gamma_i(\mu,\eta,\tau)\}$.

The terms $\gamma_i(\mu,\eta,\tau)$ in the proposal variances are not easy to obtain explicitly as the expectation in (11) must be computed with respect to the conditional distribution of θ_i given $(\mu,\eta,\tau,\mathbf{Y}_{i,1:r_i})$, which is not a Student distribution anymore. However, the terms $\gamma_i(\mu,\eta,\tau)$ may be averaged over the random variables $\mathbf{Y}_{i,1:r_i}$. Squaring (10) and computing the expectation first with respect to $\mathbf{Y}_{i,1:r_i}$ and then with respect to θ_i easily leads to

$$\mathbb{E}\left[\gamma_i(\mu,\eta,\tau)\right] = r_i\ \tau + \eta\ \frac{(\nu+1)^2}{\nu(\nu+2)}\ \frac{\Gamma((\nu+1)/2)\ \Gamma((\nu+4)/2)}{\Gamma(\nu/2)\ \Gamma((\nu+5)/2)} .$$

These terms yield local proposal variances that have been averaged over all possible datasets; these are the best local variances for the model under study when no information about the observations is available.

The RWM-within-Gibbs is then implemented using Gaussian proposal distributions with $\sigma_1 = 0.95$, $\sigma_2 = 0.025$, and $\sigma_3 = 0.0005$ for μ, η, and τ. This yields acceptance rates in the range 35%-50% for each sub-algorithm, as prescribed in the literature for one-dimensional target distributions (see [18]). We update $\boldsymbol{\theta}_{1:148}$ using a Gaussian proposal with local variances $2.38^2/\{n\mathbb{E}[\gamma_i(\mu,\eta,\tau)]\}$.

These steps are then repeated by running a RWM-within-Gibbs in which $\boldsymbol{\theta}_{1:148}$ is updated using a fixed proposal variance of 5^2. We also run a 151-dimensional RWM sampler with a $\mathcal{N}((\mu,\eta,\tau,\boldsymbol{\theta}_{1:148}), 4^2/151 * diag(1,0.01,0.001,1,\ldots,1))$ proposal distribution, and an AM algorithm in which the tuning factor of the proposal covariance matrix is 8.

The ASJD of the chain in (8) offers a reliable way of comparing the four samplers; it is reported in the first column of Table 2. A large value of this measure (relative to other samplers) is indicative of a process that rapidly explores its space, and is equivalent to ordering samplers according to their lag-1 autocorrelations. We also compare the relative efficiency of these samplers by calculating the effective sample size (ESS) of the variables μ, η, τ, and θ_2. The

Table 2: *Scottish dataset*: Efficiency and time-adjusted efficiency measures for the four samplers tested.

Sampler	Efficiency			Time-adjusted efficiency	
	Mean ASJD (a)	Min ESS (e)	Mean time (s)	a/s $(\times 100)$	e/s $(\times 100)$
RWM	2.9712	74.30	145.50	2.0421	51.07
Fixed RWM-w-G	6.4279	157.09	147.77	4.3499	106.31
Local RWM-w-G	8.4108	272.70	148.06	5.6807	184.18
Adapt. Met.	5.2476	473.83	1,081.52	0.4852	43.81

effective sample size represents the number of uncorrelated samples that are produced from the output of the sampler. It is also used as a convergence diagnostic: when its value is too small (< 100), we may have reasonable doubts that the chain really has converged. It is computed as

$$ESS = \frac{N}{1 + 2\sum_{k=1}^{\infty} \gamma(k)} ,$$

where N is the number of samples and $\sum_{k=1}^{\infty} \gamma(k)$ is the sum of lag-k sample autocorrelations. An ESS is produced for each variable; since we want to measure the number of samples that are uncorrelated over all variables, we report the minimum ESS (2nd column of Table 2). The ASJD and minimum ESS values are averaged over 10 runs of 100,000 iterations each, with a burn-in period of 1,000. These quantities are then normalized relative to the average running time of samplers (3rd column); this respectively yields the average square jumping distance per second (4th column), and the number of uncorrelated samples generated every second (5th column).

According to these results, the RWM-within-Gibbs with local variances is 1.3 times more efficient than the one with a fixed variance; the efficiency gain is even greater (1.7) if we consider the minimum ESS instead of the ASJD. Although the RWM sampler offers a slight improvement in terms of running time, it still results in efficiency measures that are significantly smaller than those of the RWM-within-Gibbs. The Adaptive Metropolis sampler could be an interesting alternative to the RWM-within-Gibbs, if it were not as expensive in terms of computational resources. Indeed, even if its ASJD is smaller than that of the RWM-within-Gibbs, its minimum ESS is greater. This sampler however requires significantly more time than the other samplers to complete its 100,000 iterations. When correcting for computational effort, it thus badly loses ground to its competitors.

The results in Table 2 thus illustrate that there is an important efficiency gain that is available from preferring a local RWM-within-Gibbs over its constant counterpart. Given that running times for both approaches are equivalent, we should clearly use local proposal variances whenever possible.

5.2. Stochastic volatility model

As a second example, we wish to study the performance of MCMC samplers in the context of a Bayesian hierarchical model that does not respect the regularity assumptions imposed by the theory of Section 3. We consider a stochastic volatility model in which the latent volatilities form an order-1 autoregressive process. The model, similar to those studied in [10] and [13], expresses the mean corrected returns d_i and log volatilites X_i, for $i \geq 1$, as

$$\begin{aligned} d_i &= \varepsilon_i \exp\{X_i/2\} , \\ X_{i+1} &= \phi X_i + \eta_{i+1} . \end{aligned}$$

The variables $\varepsilon_i \sim \mathcal{N}(0,1)$ and $\eta_i \sim \mathcal{N}(0,\tau^2)$ are uncorrelated white noises and we set $X_1 \sim \mathcal{N}(0,\tau^2/(1-\phi^2))$. Priors for the parameters τ^2 and ϕ are $\tau^2 \sim \mathrm{I}\Gamma(\delta,\lambda)$ and $(\phi+1)/2 \sim \beta(a,b)$, where $\mathrm{I}\Gamma(\delta,\lambda)$ is the inverse gamma distribution with density proportional to $x^{-(\delta+1)}\mathrm{e}^{-\lambda/x}$. This model leads to an $(n+2)$-dimensional posterior density $\pi(\tau^2,\phi,X_1,\ldots,X_n|\mathbf{d}_{1:n})$.

Before pursuing the analysis, we note that τ^2 and ϕ are constrained to subsets of $\mathbb{R}$; since the target density is rather sensitive to changes in these parameters, this will potentially affect the performance of MCMC approaches. To ensure fluidity in the samplers implemented, we apply the transformations $\tau^2 = \exp\{\kappa\}$ and $\phi = \tanh(\omega)$. The new variables κ, ω take values in $\mathbb{R}$ and the resulting $(n+2)$-dimensional posterior density is given by

$$\begin{aligned} \pi(\kappa,\omega,\mathbf{x}_{1:n}|\mathbf{d}_{1:n}) \quad &\propto \quad \exp\left\{-\kappa(\tfrac{n}{2}+\delta)\right\} \frac{\mathrm{e}^{-\omega(2b+1)}}{(1+\mathrm{e}^{-2\omega})^{a+b+1}} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}(x_i + d_i^2\mathrm{e}^{-x_i})\right\} \\ &\times \ \exp\left\{-\frac{\mathrm{e}^{-\kappa}}{2}\left[2\lambda + \frac{4\mathrm{e}^{-2\omega}}{(1+\mathrm{e}^{-2\omega})^2}x_1^2 + \sum_{i=2}^{n}\left(x_i - (\tfrac{1-\mathrm{e}^{-2\omega}}{1+\mathrm{e}^{-2\omega}})x_{i-1}\right)^2\right]\right\} \ . \end{aligned}$$

Using a 100-dimensional dataset $\mathbf{d}_{1:100}$ exhibiting low correlation (obtained from the stochastic volatility model with $\phi = 0.1$ and $\tau^2 = 0.75$), we sample this posterior density using RWM-within-Gibbs (local and fixed variances), traditional RWM, and AM algorithms. Hyperparameters are set to $\delta = 1$, $\lambda = 0.75$, $a = 10$, and $b = 6$.

For the RWM-within-Gibbs, we propose to divide the variables into 3 blocks: κ, ω, and $\mathbf{X}_{1:n}$. The proposal standard deviations associated to κ and ω are set to 0.2 and 0.27 respectively; each sub-algorithm thus accepts candidates according to a rate of $\approx 45\%$. The n-dimensional update of $\mathbf{X}_{1:n}$ is performed according to the conditional target density $\pi(\mathbf{x}_{1:n}|\kappa,\omega,\mathbf{d}_{1:n})$. In the case of the RWM-within-Gibbs with local variances, the terms

$$\gamma_i(\kappa,\omega) = \mathbb{E}[(\tfrac{\partial}{\partial X_i}\log\pi(\mathbf{X}_{1:n}|\kappa,\omega,\mathbf{d}_{1:n}))^2] \ , \quad i = 1,\ldots,n$$

in (7) are not easy to obtain as the full conditional distribution (given the data) is not normally distributed anymore. As before, we solve this problem by computing the expectation above with respect to $\mathbf{d}_{1:n}$ first, and then with respect to $\mathbf{X}_{1:n}$. The resulting proposal variances are thus averaged over all possible datasets; they are the best local proposal variances, independently of the specific dataset considered. Optimizing (6) for $i = 1,\ldots,n$ yields the n-dimensional vector

$$\frac{2.38^2}{n}\left(\tfrac{1}{2}+\mathrm{e}^{-\kappa}, \tfrac{1}{2}+\mathrm{e}^{-\kappa}\left(1+(\tfrac{1-\mathrm{e}^{-2\omega}}{1+\mathrm{e}^{-2\omega}})^2\right), \ldots, \tfrac{1}{2}+\mathrm{e}^{-\kappa}\left(1+(\tfrac{1-\mathrm{e}^{-2\omega}}{1+\mathrm{e}^{-2\omega}})^2\right), \tfrac{1}{2}+\mathrm{e}^{-\kappa}\right)^{-1} \ . \quad (12)$$

For the RWM-within-Gibbs with a fixed proposal variance, the proposal standard deviations associated to κ and ω are still 0.2 and 0.27. We then use the theory of Section 3 to obtain

Table 3: *Stochastic volatility* - Efficiency and time-adjusted efficiency measures for the four samplers tested.

Sampler	**Efficiency** Mean ASJD (a)	Min ESS (e)	Mean time (s)	**Time-adjusted efficiency** a/s ($\times$ 1,000)	e/s ($\times$ 1,000)
RWM	0.3994	103.37	367.34	1.0873	281.40
Fixed RWM-w-G	0.6420	116.58	371.93	1.7261	313.45
Local RWM-w-G	0.6740	132.40	371.38	1.8149	356.51
Adapt. Met.	0.6320	347.76	1,149.26	0.5499	302.59

an approximately optimal acceptance rate of 0.2 for the block $\mathbf{X}_{1:n}$. We reach a similar conclusion for the traditional RWM sampler. Naturally, we have to keep in mind that regularity assumptions are violated in the current context; the theoretical results might not be robust to a departure from those assumptions. In fact, given that the X_is are correlated, we expect the Adaptive Metropolis sampler to better capture this design and to outdo its competitors.

The initial covariance matrix of the Adaptive Metropolis algorithm is the $(n+2)$-dimensional identity matrix. We tune its acceptance rate as close as possible to 0.234, as suggested in the literature. For each sampler, we average the ASJD and minimum ESS over 10 runs of 200,000 iterations each, from which the first 10,000 iterations are discarded as burn-in. Time-adjusted ESJD and minimum ESS are again used a measures of efficiency; their values are reported in Table 3.

In terms of ASJD, the RWM-within-Gibbs with local variances is the best option, although its competitors also offer decent performances. The AM sampler does better, in absolute, for the minimum ESS; when accounting for computational effort however, the AM ends up outdone by the RWM-within-Gibbs (local and fixed). As before, we notice a net efficiency gain when preferring local variances to a fixed one in the RWM-within-Gibbs (net gain between 5% and 13%, depending on the efficiency measure). This modest gain is explained by the fact that, for the specific model studied, variations in κ and ω do not have a huge impact on the value of the local variances in (12). In spite of this, the impact of using local variances remains positive; generally, there does not seem to be a risk associated to using such local variances. Furthermore, the theoretical results seem applicable to contexts where regularity assumptions are violated (to some extent).

6. Discussion

In this paper, we have studied the tuning of RWM algorithms applied to single-level hierarchical target distributions. The optimal variance of the Gaussian proposal distribution has been found to depend on a measure of roughness of the density f with respect to x as before, but also with respect to the mixing coordinate x_1. This leads to local proposal variances that are a function of the mixing parameter x_1. It is however possible to average over the random variable X_1 to find a globally optimal proposal variance. In the case where X_1 is a location parameter, it does not affect the roughness of the density f and the optimal proposal scaling is valid both locally and globally.

Higher-level hierarchies could be studied using a similar approach. A target featuring p mixing components, expressed as

$$\pi\left(\mathbf{x}\right)=\prod_{j=1}^{p}f_j\left(x_j\right)\prod_{i=p+1}^{n}f\left(x_i\left|\mathbf{x}_{1:p}\right.\right)\ ,$$

with $\mathbf{x}_{1:p}=(x_1,\ldots,x_p)$ would lead to a result similar to Theorem 2, but with the function

$$\gamma(\mathbf{x}_{1:p},\mathbf{z}_{1:p})\ =\ \mathbb{E}_X\left[\left(\sum_{j=1}^{p}z_j\tfrac{\partial}{\partial x_j}\log f(X|\mathbf{x}_{1:p})\right)^2\right]+\ \mathbb{E}_X\left[\left(\tfrac{\partial}{\partial X}\log f(X|\mathbf{x}_{1:p})\right)^2\right],$$

where $\mathbf{z}_{1:p}=(z_1,\ldots,z_p)$ come from independent $\mathcal{N}(0,1)$ random variables. For a target whose mixing component (X_p say) depends itself on higher-level mixing components $X_1,\ldots,X_{p-1}$, expressed as $\pi\left(\mathbf{x}\right)=f_1\left(\mathbf{x}_{1:p}\right)\prod_{i=p+1}^{n}f\left(x_i\left|x_p\right.\right)$, the conclusions of Theorem 2 are still valid. These generalizations also hold for Corollary 7, with obvious adjustments $(Z_1\sim\mathcal{N}(0,\ell^2\kappa_1^2/n),\ldots,Z_p\sim\mathcal{N}(0,\ell^2\kappa_p^2/n))$. Similar extensions may be derived for other hierarchical models.

In the simulation study of Section 4, we found that the optimal acceptance rate most often lies around 0.2. In the gamma-normal example, there were some values of α,λ that led to significantly lower optimal acceptance rates (0.15 when $\alpha=2,\lambda=3$). The usual 0.234 is thus quite robust and, if preferred, should lead to an efficient version of the sampler. In the case of correlated targets, it would however be wiser to settle for an acceptance rate slightly below 0.234. Since we investigate correlated targets with a proposal distribution featuring a diagonal covariance matrix, it is not surprising to find an AOAR lower than 0.234; the latter is the AOAR for exploring a target distribution with independent components, which is an ideal situation when relying on a proposal distribution with independent components.

We conclude by outlining that the concept of locally optimal proposal variances reveals itself to be of interest with other types of samplers, such as RWM-within-Gibbs algorithms. Indeed, the asymptotic results of Section 3 are proof of the theoretical superiority of RWM-within-Gibbs over RWM when sampling from hierarchical targets. The examples of Section 5 illustrate the efficiency gain from using a RWM-within-Gibbs with local variances over some competitors, including an adaptive sampler. Similar ideas may also be applied to different samplers such as Metropolis-adjusted Langevin algorithms (MALA), but this goes beyond the scope of this paper.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Funding Statement

This work was supported by the National Sciences and Engineering Research Council of Canada (Grant number: RGPIN/03931-2014).

Acknowledgements

The author is grateful to P. Gagnon for useful comments. The author also wishes to thank anonymous referees for very constructive comments that greatly improved the manuscript.

References

[1] M. Bédard, On the Robustness of Optimal Scaling for Random Walk Metropolis Algorithms., Ph.D. thesis, University of Toronto, 2006.

[2] M. Bédard, Weak convergence of Metropolis algorithms for non-i.i.d. target distributions, Ann. Appl. Probab. 17 (2007) 1222–1244.

[3] M. Bédard, Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234, Stoch. Process. Appl. 118 (2008) 2198–2222.

[4] M. Bédard, Hierarchical models: Local proposal variances for RWM-within-Gibbs and MALA-within-Gibbs, Comput. Statist. Data Anal. 109 (2017) 231–246.

[5] M. Bédard, R. Douc, E. Moulines, Scaling analysis of multiple-try MCMC methods, Stoch. Process. Appl. 122 (2012) 758–786.

[6] A. Beskos, N. Pillai, G. Roberts, J. Sanz-Serna, A. Stuart, Optimal tuning of the hybrid Monte Carlo algorithm, Bernoulli 19 (2013) 1501–1534.

[7] A. Beskos, G. Roberts, A. Stuart, Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions, Ann. Appl. Probab. 19 (2009) 863–898.

[8] P. Billingsley, Convergence of probability measures, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, second edition, 1999.

[9] S. Ethier, T. Kurtz, Markov processes: Characterization and Convergence, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986.

[10] M. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, J. Roy. Statist. Soc. Ser. B 73 (2011) 123–214.

[11] H. Haario, E. Saksman, J. Tamminen, An adaptive Metropolis algorithm, Bernoulli 7 (2001) 223–242.

[12] W. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.

[13] S. Kim, N. Shepard, S. Chib, Stochastic volatility: likelihood inference and comparison with ARCH models, The Review of Economic Studies 65 (1998) 361–393.

[14] J. Mattingly, N. Pillai, A. Stuart, Diffusion limits of random walk Metropolis in high dimensions, Ann. Appl. Probab. 22 (2012) 881–930.

[15] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, Equation of state calculations by fast computing machines, The Journal of Chemical Physics 21 (1953) 1087–1092.

[16] P. Neal, G. Roberts, Optimal scaling for partially updating MCMC algorithms, Ann. Appl. Probab. 16 (2006) 475–515.

[17] G. Roberts, A. Gelman, W. Gilks, Weak convergence and optimal scaling of random walk Metropolis algorithms, Ann. Appl. Probab. 7 (1997) 110–120.

[18] G. Roberts, J. Rosenthal, Optimal scaling for various Metropolis-Hastings algorithms, Stat. Sci. 16 (2001) 351–367.

[19] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, M. Girolami, Langevin diffusions and the Metropolis-adjusted Langevin algorithm, Stat. Probab. Lett. 91 (2014) 14–19.

A. Appendix : Proofs of theorems

We now proceed to prove Theorems 1 and 2. To assess weak convergence of the processes $\{\tilde{W}_1^{(n)}(t); t \geq 0\}$ and $\{W_2^{(n)}(t); t \geq 0\}$ in the Skorokhod topology, we first verify weak convergence of finite-dimensional distributions. Whereas these processes are not themselves Markovian, they are $\mathcal{F}^{\tilde{\mathbf{W}}^{(n)}}(t)$-progressive and $\mathcal{F}^{(W_1^{(n)}, W_i^{(n)})}(t)$-progressive $\mathbb{R}$-valued processes respectively, and the aim of this section is to establish their convergence to some Markov processes. According to Theorem 8.2 of Chapter 4 in [9], we thus look at the pseudo generator of $\{\tilde{W}_1^{(n)}(t); t \geq 0\}$ (resp. $\{W_2^{(n)}(t); t \geq 0\}$), the univariate process associated to the component X_1 (resp. X_2) in the rescaled RWM algorithm introduced at the end of Section 2. We then verify $\mathcal{L}^1$-convergence to the generator of the special RWM sampler with acceptance rule (4) (resp. the generator of the diffusion in (5)).

To complete the proofs, Theorem 7.8 of Chapter 3 in [9] says that we must also assess the relative compactness of $\{\tilde{W}_1^{(n)}(t); t \geq 0\}$ and $\{W_2^{(n)}(t); t \geq 0\}$) for $n = 2, 3, \ldots$, as well as the existence of a countable dense set on which the finite-dimensional distributions weakly converge. This is achieved by using Corollary 8.6 of Chapter 4 in [9]; in the setting of Theorem 1, the satisfaction of applicability conditions is immediate; in the setting of Theorem 2, the satisfaction of the first condition is immediate, while the verification of the second condition is briefly discussed in Section A.2.

A.1. Proof of Theorem 1

In Theorem 1, it is assumed that $\{\tilde{W}_1^{(n)}(t); t \geq 0\}$ is the component of interest in $\{\tilde{\mathbf{W}}^{(n)}(t); t \geq 0\}$. Define the pseudo generator of $\{\tilde{W}_1^{(n)}(t); t \geq 0\}$ as

$$\tilde{G}_n h(\tilde{W}_1^{(n)}(t)) = \mathbb{E}\left[h(\tilde{W}_1^{(n)}(t+1)) - h(\tilde{W}_1^{(n)}(t)) \,\middle|\, \mathcal{F}^{\tilde{\mathbf{W}}^{(n)}}(t) \right] ,$$

where h is an arbitrary test function. By setting $\xi_n(t) = h(\tilde{W}_1^{(n)}(t))$ and $\varphi_n(t) = \tilde{G}_n h(\tilde{W}_1^{(n)}(t))$, conditions in part (c) of Theorem 8.2 (Chap. 4 in [9]) reduce to $\mathbb{E}\left[\left| \tilde{G}_n h(\tilde{W}_1{}^{(n)}(t)) - \tilde{G} h(\tilde{W}_1(t)) \right| \right]$

$\to 0$ as $n \to \infty$ for $h \in \overline{C}$ (the space of continuous and bounded functions on $\mathbb{R}$), where $\tilde{G}h(\tilde{W}_1(t))$ is the generator of the special RWM sampler described in Theorem 1.

The above may be reexpressed as $\mathbb{E}\left[\left|\tilde{G}_n h(\tilde{X}_1) - \tilde{G}h(\tilde{X}_1)\right|\right] \to 0$ as $n \to \infty$, where

$$\tilde{G}_n h(\tilde{x}_1) = \mathbb{E}_{\tilde{Y}_1}\left[\left(h(\tilde{Y}_1) - h(\tilde{x}_1)\right)\mathbb{E}_{\mathbf{Y}_{2:n}}\left[\alpha(\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\right]\right]$$

with $\tilde{\mathbf{x}}^{(n)} = (\tilde{x}_1, x_2, \ldots, x_n)$ and similarly for $\tilde{\mathbf{Y}}^{(n)}$. The density of $\tilde{\mathbf{x}}^{(n)}$ is $\frac{1}{\sqrt{n}}\pi(\mu_n + \frac{\tilde{x}_1}{\sqrt{n}}, \mathbf{x}_{2:n})$ with π as in (1), and thus $\alpha(\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) = 1 \wedge \frac{\pi(\mu_n + \tilde{Y}_1/\sqrt{n}, \mathbf{Y}_{2:n})}{\pi(\mu_n + \tilde{x}_1/\sqrt{n},\ \mathbf{x}_{2:n})}$; hereafter, $1 \wedge x = \min(1, x)$. Furthermore,

$$\tilde{G}h(\tilde{x}_1) = \mathbb{E}_{\tilde{Y}_1}\left[\left(h(\tilde{Y}_1) - h(\tilde{x}_1)\right)\alpha^*(\tilde{x}_1, \tilde{Y}_1|\underline{\mathbf{x}})\right]$$

with α^* as in (4). Note that there is a slight abuse of notation as, although h is a function of x_1 only, the generator $\tilde{G}_n h(\tilde{x}_1)$ is a function of $\tilde{\mathbf{x}}^{(n)}$; a similar remark holds for $\tilde{G}h(\tilde{x}_1)$. We now proceed to verify this condition. Hereafter, we use $\to_{a.s.}$, $\to_p$, and $\to_d$ to denote convergence almost surely, in probability, and in distribution.

In the current context where there is no time-rescaling factor, the limiting process shall remain a RWM algorithm. For $h \in \overline{C}$ and some $K > 0$, the triangle inequality implies

$$\begin{aligned}
\mathbb{E}\left[\left|\tilde{G}_n h(\tilde{X}_1) - \tilde{G}h(\tilde{X}_1)\right|\right] \quad \leq \quad & K\ \mathbb{E}\left[\left|\alpha(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) - \alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\right|\right] \\
& + K\ \mathbb{E}\left[\left|\alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) - \alpha_1(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\right|\right] \\
& + K\ \mathbb{E}\left[\left|\mathbb{E}_{\mathbf{Y}_{2:n}}\left[\alpha_1(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\right] - \alpha^*(\tilde{X}_1, \tilde{Y}_1|\underline{\mathbf{X}})\right|\right],
\end{aligned} \tag{A.1}$$

where the function $\alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})$ shall be defined in Lemma B.1 and $\alpha_1(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) = 1 \wedge \exp\left\{\varepsilon_1(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\right\}$. Here,

$$\begin{aligned}
\varepsilon_1(\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) \quad = \quad & \log\frac{f_1(\mu_n + \frac{\tilde{Y}_1}{\sqrt{n}}|\mathbf{x}_{2:n})}{f_1(\mu_n + \frac{\tilde{x}_1}{\sqrt{n}}|\mathbf{x}_{2:n})} + \sum_{i=2}^{n} \frac{\partial}{\partial x}\log f(x|\mu_n + \tfrac{\tilde{x}_1}{\sqrt{n}})\Big|_{x=x_i}(Y_i - x_i) \\
& - \frac{\ell^2}{2n}\sum_{i=2}^{n}\left(\frac{\partial}{\partial x}\log f(x|\mu_n + \tfrac{\tilde{x}_1}{\sqrt{n}})\Big|_{x=x_i}\right)^2,
\end{aligned} \tag{A.2}$$

with $\frac{1}{\sqrt{n}}f_1(\mu_n + \frac{\tilde{x}_1}{\sqrt{n}}|x_{2:n})$ representing the conditional density of $\tilde{X}_1$ given $\mathbf{x}_{2:n}$.

By Lemmas B.1 and B.2, the first and second terms in (A.1) respectively converge to 0 as $n \to \infty$; in the sequel, we thus study the last term. Since $\mathbf{Y}_{2:n} \sim \mathcal{N}(\mathbf{x}_{2:n}, \ell^2 I_{n-1}/n)$, the second and third terms on the right of (A.2) are normally distributed with mean M and variance V, where $V = -2M = \frac{\ell^2}{n}\sum_{i=2}^{n}\left(\frac{\partial}{\partial x}\log f(x|\mu_n + \frac{\tilde{x}_1}{\sqrt{n}})\Big|_{x=x_i}\right)^2$.

By assumption, this variance term converges in probability to $\ell^2\tilde{\gamma}(\mu)$; hence, the last two terms on the right of (A.2) converge in probability to a $\mathcal{N}(-\ell^2\tilde{\gamma}(\mu)/2, \ell^2\tilde{\gamma}(\mu))$. Regularity conditions allow us to invoke the (multivariate) Continuous Mapping Theorem, which implies

$$\alpha_1(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) \quad \to_p \quad 1 \wedge \exp\left\{\mathcal{N}\left(\log\frac{g_1(\tilde{Y}_1|\underline{\mathbf{X}})}{g_1(\tilde{X}_1|\underline{\mathbf{X}})} - \frac{\ell^2}{2}\tilde{\gamma}(\mu), \ell^2\tilde{\gamma}(\mu)\right)\right\}.$$

Proposition 2.4 in [17] then claims that the expectation of $1 \wedge \exp\{Z\}$, where Z is the normal random variable just introduced, is equal to $\alpha^*(\tilde{X}_1, \tilde{Y}_1|\underline{\mathbf{X}})$. The Bounded Convergence Theorem can then be used to conclude that the last term in (A.1) converges to 0 as $n \to \infty$.

A.2. Proof of Theorem 2

In Theorem 2, it is assumed that $\{W_i^{(n)}(t); t \geq 0\}$ $(i = 2, \ldots, n)$ is the component of interest in the rescaled process $\{\mathbf{W}^{(n)}(t); t \geq 0\}$. Without loss of generality, fix $i = 2$ and define the pseudo generator of $\{W_2^{(n)}(t); t \geq 0\}$ as

$$G_n h(W_2^{(n)}(t)) = n\mathbb{E}\left[h(W_2^{(n)}(t + \tfrac{1}{n})) - h(W_2^{(n)}(t)) \,\middle|\, \mathcal{F}^{(W_1^{(n)}, W_2^{(n)})}(t)\right] ,$$

where h is an arbitrary test function.

By setting $\xi_n(t) = h(W_2^{(n)}(t))$ and $\varphi_n(t) = G_n h(W_2^{(n)}(t))$, part (c) of Theorem 8.2 (Chapter 4 in [9]) reduces to the conditions $\sup_n \sup_{s \leq T} \mathbb{E}[|G_n h(W_2^{(n)}(s))|] < \infty$ for $T > 0$ and $h \in \overline{C}$, and $\mathbb{E}\left[\left|G_n h(W_2^{(n)}(t)) - Gh(W_2(t))\right|\right] \to 0$ as $n \to \infty$ for $h \in \overline{C}$, where $Gh(W_2(t))$ is the generator of the diffusion process described in Theorem 2.

Hereafter, we use the notation $\mathbf{Y}_{1,3:n} = (Y_1, Y_3, \ldots, Y_n)$. The latter condition may be reexpressed as $\mathbb{E}_{\mathbf{X}_{1:2}}\left[|G_n h(X_2) - Gh(X_2)|\right] \to 0$ as $n \to \infty$, where

$$G_n h(X_2) = n\mathbb{E}_{Y_2}\left[(h(Y_2) - h(X_2))\, \mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{1,3:n}}\left[\alpha(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})\right]\right] \tag{A.3}$$

with $\alpha(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = 1 \wedge \frac{\pi(\mathbf{Y}^{(n)})}{\pi(\mathbf{x}^{(n)})}$ and π as in (1), and

$$Gh(X_2) = \upsilon(\ell, X_1)\left\{\frac{1}{2}h''(X_2) + \frac{1}{2}\frac{\partial}{\partial X_2}\log f(X_2|X_1)h'(X_2)\right\} . \tag{A.4}$$

There is again a slight abuse of notation as, although h is a function of x_2 only, the generators $G_n h(x_2)$ and $Gh(x_2)$ are functions of x_1, x_2. Due to the form of (A.4), we can resort to Theorem 2.1 of Chapter 8 in [9] to assert that $\mathcal{C}_c^\infty$, the space of continuous and infinitely differentiable functions that are compactly supported on $\mathbb{R}$, forms a core for the generator of the diffusion in Theorem 2. The test function h in (A.3) and (A.4) might then be restricted to functions h belonging to $\mathcal{C}_c^\infty$.

We note that the condition $\sup_n \sup_{s \leq T} \mathbb{E}[|G_n h(W_2^{(n)}(s))|] < \infty$ for $T > 0$ and $h \in \overline{C}$ may be reexpressed as $\sup_n \mathbb{E}[|G_n h(X_2)|] < \infty$ for $h \in \mathcal{C}_c^\infty$. In fact, it is straight-forward to verify that $\mathbb{E}[(G_n h(X_2))^2] \leq K_h + \mathcal{O}(n^{-1})$ for some $K_h \in (0, \infty)$ which implies that the former is satisfied (this is achieved by considering a function similar to (B.6), in which the acceptance function is Taylor expanded to first order only, and by proceeding as in the proof of Lemma B.3). It also implies the satisfaction of the second applicability condition of Corollary 8.6 (Chapter 4 in [9]), which may be simplified as $\limsup_{n\to\infty} \mathbb{E}[(G_n h(X_2))^2] < \infty$ for $h \in \mathcal{C}_c^\infty$.

We now proceed to verify that $G_n h(X_2)$ converges in $\mathcal{L}^1$ to $Gh(x_2)$. To begin, we have from Lemma B.3 that $\mathbb{E}_{\mathbf{X}_{1:2}}\left[\left|G_n h(X_2) - G_n^{(1)} h(X_2)\right|\right] \to 0$ as $n \to \infty$, where $G_n^{(1)} h(x_2)$ is the generator of a diffusive process :

$$\begin{aligned} G_n^{(1)} h(X_2) &= \frac{\ell^2}{2} h''(X_2) \mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{1,3:n}} \left[\alpha(\mathbf{X}^{(n)}, \mathbf{Y}_{X_2}^{(n)}) \mathbb{1}_{\mathcal{X}_1}(Y_1) \right] \\ &\quad + \ell^2 h'(X_2) \tfrac{\partial}{\partial X_2} \log f(X_2|X_1) \mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{1,3:n}} \left[g(\mathbf{X}^{(n)}, \mathbf{Y}_{X_2}^{(n)}) \mathbb{1}_{\mathcal{X}_1}(Y_1) \right] . \end{aligned} \tag{A.5}$$

Note that it is not necessary to precise this expression further at this stage. Now, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{X}_{1:2}} \left[\left| G_n^{(1)} h(X_2) - G h(X_2) \right| \right] \leq \\ &\quad K_1 \; \mathbb{E}_{\mathbf{X}_{1:2}} \left[\left| \ell^2 \mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{1,3:n}} \left[\alpha(\mathbf{X}^{(n)}, \mathbf{Y}_{X_2}^{(n)}) \mathbb{1}_{\mathcal{X}_1}(Y_1) \right] - \upsilon(\ell, X_1) \right| \right] \\ &+ \quad K_2 \; \mathbb{E}_{\mathbf{X}_{1:2}} \left[\left| \tfrac{\partial}{\partial X_2} \log f(X_2|X_1) \right| \left| \ell^2 \mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{1,3:n}} \left[g(\mathbf{X}^{(n)}, \mathbf{Y}_{X_2}^{(n)}) \mathbb{1}_{\mathcal{X}_1}(Y_1) \right] - \frac{1}{2} \upsilon(\ell, X_1) \right| \right] \end{aligned} \tag{A.6}$$

for some $K_1, K_2 > 0$, since $h \in C_c^\infty$ and thus $|h'|$ and $|h''|$ are bounded.

Using the triangle inequality, the first term on the RHS of (A.6) satisfies

$$\begin{aligned} &\mathbb{E}_{\mathbf{X}_{1:2}} \left[\left| \ell^2 \mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{1,3:n}} \left[\alpha(\mathbf{X}^{(n)}, \mathbf{Y}_{X_2}^{(n)}) \mathbb{1}_{\mathcal{X}_1}(Y_1) \right] - \upsilon(\ell, X_1) \right| \right] \\ &\leq \ell^2 \mathbb{E}_{\mathbf{X}_{1:n}, \mathbf{Y}_{1,3:n}} \left[\left| \alpha(\mathbf{X}^{(n)}, \mathbf{Y}_{X_2}^{(n)}) - \hat{\alpha}(\mathbf{X}^{(n)}, \mathbf{Y}_{X_2}^{(n)}) \right| \mathbb{1}_{\mathcal{X}_1}(Y_1) \right] \\ &\quad + \mathbb{E}_{\mathbf{X}_{1:2}} \left[\left| \ell^2 \mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{1,3:n}} \left[\hat{\alpha}(\mathbf{X}^{(n)}, \mathbf{Y}_{X_2}^{(n)}) \mathbb{1}_{\mathcal{X}_1}(Y_1) \right] - \upsilon(\ell, X_1) \right| \right] , \end{aligned}$$

where the function $\hat{\alpha}$ is as in Lemma B.4. Using Lemmas B.4 and B.5, the above converges to 0 as $n \to \infty$. It thus only remains to verify that the second term on the right hand side of (A.6) also converges to 0; Lemma B.6 leads us to that conclusion.

B. Appendix : Intermediate results

Lemma B.1. *As $n \to \infty$, we have $\mathbb{E}\left[\left| \alpha(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) - \alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) \right| \right] \to 0$, with α as in Appendix A.1 and $\alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) = 1 \wedge \exp\left\{ \varepsilon_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) \right\}$, with*

$$\begin{aligned} \varepsilon_2(\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) &= \log \frac{f_1(\mu_n + \frac{\tilde{Y}_1}{\sqrt{n}} | \mathbf{x}_{2:n})}{f_1(\mu_n + \frac{\tilde{x}_1}{\sqrt{n}} | \mathbf{x}_{2:n})} + \sum_{i=2}^n \tfrac{\partial}{\partial x} \log f(x | \mu_n + \tfrac{\tilde{Y}_1}{\sqrt{n}}) \Big|_{x = x_i} (Y_i - x_i) \\ &\quad - \frac{\ell^2}{2n} \sum_{i=2}^n \left(\tfrac{\partial}{\partial x} \log f(x | \mu_n + \tfrac{\tilde{Y}_1}{\sqrt{n}}) \Big|_{x = x_i} \right)^2 . \end{aligned} \tag{B.1}$$

Proof. The acceptance function satisfies $\alpha(\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) = 1 \wedge \exp\{\varepsilon(\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\}$, where

$$\varepsilon(\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) = \log \frac{f_1(\mu_n + \frac{\tilde{Y}_1}{\sqrt{n}}) \prod_{i=2}^n f(x_i | \mu_n + \frac{\tilde{Y}_1}{\sqrt{n}})}{f_1(\mu_n + \frac{\tilde{x}_1}{\sqrt{n}}) \prod_{i=2}^n f(x_i | \mu_n + \frac{\tilde{x}_1}{\sqrt{n}})} + \sum_{i=2}^n \left(\log \frac{f(Y_i | \mu_n + \frac{\tilde{Y}_1}{\sqrt{n}})}{f(x_i | \mu_n + \frac{\tilde{Y}_1}{\sqrt{n}})} \right) .$$

Applying obvious changes of variables allows us to express ε in terms of $\mathbf{x}^{(n)}$ and $\mathbf{Y}^{(n)}$:

$$\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) \;\; = \;\; \log \frac{f_1(Y_1|\mathbf{x}_{2:n})}{f_1(x_1|\mathbf{x}_{2:n})} + \sum_{i=2}^{n} \left(\log f(Y_i|Y_1) - \log f(x_i|Y_1) \right).$$

Using a second-order Taylor expansion with respect to Y_i around x_i ($i = 2, \ldots, n$) to reexpress the last term on the right hand side (RHS) leads to

$$\begin{aligned} \varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) \;\; = \;\; & \log \frac{f_1(Y_1|\mathbf{x}_{2:n})}{f_1(x_1|\mathbf{x}_{2:n})} + \sum_{i=2}^{n} \tfrac{\partial}{\partial x_i} \log f(x_i|Y_1)(Y_i - x_i) \\ & + \frac{1}{2} \sum_{i=2}^{n} \tfrac{\partial^2}{\partial U_i^2} \log f(U_i|Y_1)(Y_i - x_i)^2 \end{aligned}$$

for some $U_i \in (x_i, Y_i)$ or $U_i \in (Y_i, x_i)$.

We note that a candidate Y_1 that does not belong to $\mathcal{X}_1$ is automatically rejected by the algorithm, *i.e.* functions α, α_2, α_1, and α^* are identically 0. Applying changes of variables to the function $\varepsilon_2(\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})$ and using the Lispchitz property of $1 \wedge \exp\{\cdot\}$ along with the fact that $Y_i \sim \mathcal{N}(x_i, \ell^2/n)$, $i = 2, \ldots, n$ yield

$$\begin{aligned} \mathbb{E}\left[\left|\alpha(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) - \alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\right|\right] \;\; \leq \;\; & \mathbb{E}\left[\left|\varepsilon(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}) - \varepsilon_2(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})\right| \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\ \leq \;\; & \mathbb{E}\left[\left| \frac{1}{2} \sum_{i=2}^{n} \tfrac{\partial^2}{\partial X_i^2} \log f(X_i|Y_1)(Y_i - X_i)^2 + \frac{\ell^2}{2n} \sum_{i=2}^{n} \left(\tfrac{\partial}{\partial X_i} \log f(X_i|Y_1) \right)^2 \right| \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\ & + \frac{\ell^2}{2} \left(\frac{n-1}{n} \right) \mathbb{E}\left[\left| \tfrac{\partial^2}{\partial U_2^2} \log f(U_2|Y_1) - \tfrac{\partial^2}{\partial X_2^2} \log f(X_2|Y_1) \right| Z_2^2 \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \;, \end{aligned}$$

where $Z_2 = \sqrt{n}(Y_2 - X_2)/\ell \sim \mathcal{N}(0, 1)$, and $\mathbb{1}_{\mathcal{X}_1}(y) = 1$ if $y \in \mathcal{X}_1$ and 0 otherwise. From Proposition C.1 in Appendix C, the first term on the RHS converges to 0 as $n \to \infty$. We now study the second term on the right. Since $Y_2 \to_{a.s.} x_2$, it implies that $U_2 \to_{a.s.} x_2$; from the Continuous Mapping Theorem, we have $\left| \tfrac{\partial^2}{\partial U_2^2} \log f(U_2|Y_1) - \tfrac{\partial^2}{\partial X_2^2} \log f(X_2|Y_1) \right| \to_{a.s.} 0$, for all $Y_1 \in \mathcal{X}_1$. Furthermore,

$$\begin{aligned} \mathbb{E}\left[\left(\tfrac{\partial^2}{\partial U_2^2} \log f(U_2|Y_1) - \tfrac{\partial^2}{\partial X_2^2} \log f(X_2|Y_1) \right)^2 Z_2^4 \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \;\; \leq \;\; & 12\, \mathbb{E}[K^2(Y_1) \mathbb{1}_{\mathcal{X}_1}(Y_1)] \\ \leq \; 24\, \mathbb{E}[(K(Y_1) - K(X_1))^2 \mathbb{1}_{\mathcal{X}_1}(Y_1)] + 24\, \mathbb{E}[K^2(X_1)] \; \leq \; & 24 K^* \frac{\ell^2}{n} + 24\, \mathbb{E}[K^2(X_1)] \; < \; \infty \end{aligned}$$

for some $K^* > 0$ (since $K(x_1)$ satisfies a Lipschitz condition). We conclude, by invoking the Uniform Integrability Theorem, that the second term converges to 0 as $n \to \infty$.

□

Lemma B.2. *As $n \to \infty$, we have $\mathbb{E}\left[\left|\alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) - \alpha_1(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\right|\right] \to 0$, with α_1 as in Appendix A.1 and $\alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})$ as in Lemma B.1.*

Proof. Applying obvious changes of variables to α_1, α_2 and using the Lipschitz property of $1 \wedge \exp\{\cdot\}$ yield

$$\begin{aligned}
\mathbb{E}\left[\left|\alpha_2(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)}) - \alpha_1(\tilde{\mathbf{X}}^{(n)}, \tilde{\mathbf{Y}}^{(n)})\right|\right] \quad &\leq \quad \mathbb{E}\left[\left|\varepsilon_2(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}) - \varepsilon_1(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})\right| \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\
&\leq \mathbb{E}\left[\left|\sum_{i=2}^{n}\left(\tfrac{\partial}{\partial X_i}\log f(X_i|Y_1) - \tfrac{\partial}{\partial X_i}\log f(X_i|X_1)\right)(Y_i - X_i)\right| \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \quad \text{(B.2)} \\
&+ \frac{\ell^2}{2}\left(\frac{n-1}{n}\right)\mathbb{E}\left[\left|\left(\tfrac{\partial}{\partial X_i}\log f(X_i|Y_1)\right)^2 - \left(\tfrac{\partial}{\partial X_i}\log f(X_i|X_1)\right)^2\right| \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] .
\end{aligned}$$

The summation in (B.2) is distributed according to a normal random variable with null mean and variance $\frac{\ell^2}{n}\sum_{i=2}^{n}\left(\frac{\partial}{\partial X_i}\log f(X_i|Y_1) - \frac{\partial}{\partial X_i}\log f(X_i|X_1)\right)^2$. Using Hölder's inequality, the corresponding expectation is bounded by

$$\left\{\ell^2\left(\frac{n-1}{n}\right)\mathbb{E}\left[\left(\tfrac{\partial}{\partial X_i}\log f(X_i|Y_1) - \tfrac{\partial}{\partial X_i}\log f(X_i|X_1)\right)^2 \mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\right\}^{1/2} . \tag{B.3}$$

Since $Y_1 \to_{a.s.} x_1$, we use the Continuous Mapping Theorem to affirm that the integrand converges to 0 almost surely. By assumption, we know that $\mathbb{E}[(\frac{\partial}{\partial X_i}\log f(X_i|X_1))^4] < \infty$. From the proof of Proposition C.1, we also know that $\mathbb{E}[(\frac{\partial}{\partial X_i}\log f(X_i|Y_1))^4 \mathbb{1}_{\mathcal{X}_1}(Y_1)] < \infty$. We can thus use the Uniform Integrability Theorem to deduce that the expectation in (B.3) converges to 0 as $n \to \infty$. The exact same arguments may be used to conclude that the last term in (B.2) converges to 0 as $n \to \infty$. □

Lemma B.3. *As $n \to \infty$ we have $\mathbb{E}_{\mathbf{X}_{1:2}}\left[\left|G_n h(X_2) - G_n^{(1)} h(X_2)\right|\right] \to 0$, where $G_n h(X_2)$ and $G_n^{(1)} h(X_2)$ are in* (A.3) *and* (A.5) *respectively, with $\mathbf{Y}_{x_2}^{(n)} = (Y_1, x_2, Y_3, \ldots, Y_n)$,*

$$g(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = \exp\{\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\}\ \mathbb{1}\left\{\exp\{\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\} < 1\right\} , \tag{B.4}$$

and

$$\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = \log\frac{f_1(Y_1)}{f_1(x_1)} + \log\frac{f(Y_2|Y_1)}{f(x_2|x_1)} + \sum_{i=3}^{n}\left(\log f(Y_i|Y_1) - \log f(x_i|x_1)\right). \tag{B.5}$$

Proof. The acceptance rule in (A.3) may be written $\alpha(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = 1 \wedge \exp\{\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\}$, where the candidates are generated according to $\mathbf{Y}^{(n)} \sim \mathcal{N}(\mathbf{x}^{(n)}, \ell^2 I_n/n)$. We note that a candidate $Y_1 \notin \mathcal{X}_1$ is automatically rejected by the algorithm, and thus corresponds to an acceptance probability that is null. It thus not cause any problem to express the acceptance function as $\alpha(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\mathbb{1}_{\mathcal{X}_1}(Y_1)$ wherever necessary.

We first Taylor expand the acceptance function $\alpha(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = 1 \wedge \exp\{\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\}$ with respect to Y_2 around x_2. As argued in [16], this function is not everywhere differentiable. However, the points $(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})$ at which the derivatives do not exist have a Lebesgue measure that is either null or converging exponentially to 0 as $n \to \infty$; hence this shall not cause any concern when considering expectations of generators. (The latter may happen if f_1 and f are constant over some interval of the state space, for instance, in which case we could have

$\mathbb{P}(\pi(\mathbf{Y}^{(n)}) = \pi(\mathbf{x}^{(n)})) > 0$. The occurence of such values $\mathbf{x}^{(n)}$ however has a probability converging exponentially rapidly to 0 as $n \to \infty$).

The first-order derivative of $\alpha(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})$ with respect to Y_2 is given by

$$\tfrac{\partial}{\partial Y_2}\alpha(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = \tfrac{\partial}{\partial Y_2}\log f(Y_2|Y_1)\ g(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\ ,$$

where the function g is as in (B.4); the second-order derivative is expressed as

$$\tfrac{\partial^2}{\partial Y_2^2}\alpha(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) \quad = \quad \left\{\tfrac{\partial^2}{\partial Y_2^2}\log f(Y_2|Y_1) + \left(\tfrac{\partial}{\partial Y_2}\log f(Y_2|Y_1)\right)^2\right\} g(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}).$$

The generator in (A.3) is thus developed as

$$\begin{aligned}
G_n h(X_2) &= n\mathbb{E}_{Y_2}\left[h(Y_2) - h(X_2)\right]\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[\alpha(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\
&+ n\mathbb{E}_{Y_2}\left[(h(Y_2) - h(X_2))\,(Y_2 - X_2)\right]\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[\tfrac{\partial}{\partial Y_2}\alpha(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})\Big|_{Y_2=X_2}\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\
&+ R_n(\mathbf{X}_{1:2}, U_2),
\end{aligned} \tag{B.6}$$

where

$$R_n(\mathbf{X}_{1:2}, U_2) = \frac{n}{2}\mathbb{E}_{Y_2}\left[(h(Y_2) - h(X_2))\,(Y_2 - X_2)^2\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[\tfrac{\partial^2}{\partial U_2^2}\alpha(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{U_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\right] \tag{B.7}$$

for some $U_2 \in (X_2, Y_2)$ or $U_2 \in (Y_2, X_2)$. This leads to

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}_{1:2}}\left[\left|G_n h(X_2) - G_n^{(1)}h(X_2)\right|\right] \le \mathbb{E}\left[|R_n(\mathbf{X}_{1:2}, U_2)|\right] \\
&+ \quad \mathbb{E}_{\mathbf{X}_{1:2}}\left[\left|n\mathbb{E}_{Y_2}\left[h(Y_2) - h(X_2)\right] - \tfrac{\ell^2}{2}h''(X_2)\right|\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[\alpha(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\right] \qquad \text{(B.8)} \\
&+ \quad \mathbb{E}_{\mathbf{X}_{1:2}}\left[\left|n\mathbb{E}_{Y_2}\left[(h(Y_2) - h(X_2))\,(Y_2 - X_2)\right]\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[\tfrac{\partial}{\partial Y_2}\alpha(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})\Big|_{Y_2=X_2}\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\right.\right. \\
&\qquad\qquad \left.\left. -\ell^2 h'(X_2)\tfrac{\partial}{\partial X_2}\log f(X_2|X_1)\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[g(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\right|\right]\ .
\end{aligned}$$

The remainder term in (B.7) converges to 0 in $\mathcal{L}^1$, as now detailed. By using a first-order Taylor expansion of h with respect to Y_2 around x_2 along with the fact that $h \in C_c^\infty$, it follows that $|h(Y_2) - h(x_2)| \le K_1|Y_2 - x_2|$ for some $K_1 > 0$. Furthermore, since $\frac{\partial}{\partial x_2}\log f(x_2|x_1)$ is Lipschitz continuous on $\mathbb{R}$ for all fixed $x_1 \in \mathcal{X}_1$, then $|\frac{\partial^2}{\partial x_2^2}\log f(x_2|x_1)| \le K(x_1)$. Using the fact that the function g in (B.4) is bounded by 1, we then write

$$\begin{aligned}
\mathbb{E}\left[|R_n(\mathbf{X}_{1:2}, U_2)|\right] \quad &\le \quad \frac{n}{2}K_1\ \frac{2^{3/2}}{\sqrt{\pi}}\frac{\ell^3}{n^{3/2}}\mathbb{E}[K(Y_1)\mathbb{1}_{\mathcal{X}_1}(Y_1)] \\
&\quad + \frac{n}{2}K_1\ \mathbb{E}\left[|Y_2 - X_2|^3\left(\tfrac{\partial}{\partial U_2}\log f(U_2|Y_1)\right)^2\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\ .
\end{aligned}$$

Since $|\frac{\partial}{\partial U_2}\log f(U_2|Y_1)| \le |\frac{\partial}{\partial x_2}\log f(x_2|x_1)| + L(x_2)|Y_1 - x_1| + K(Y_1)|Y_2 - x_2|$ and $(a+b+c)^2 \le 4(a^2 + b^2 + c^2)$ for a, b, and c in $\mathbb{R}$, then

$$\begin{aligned}
\mathbb{E}\left[|R_n(\mathbf{X}_{1:2}, U_2)|\right] \quad &\le \quad \sqrt{\frac{2}{\pi}}K_1\frac{\ell^3}{n^{1/2}}\left\{\mathbb{E}[K(Y_1)\mathbb{1}_{\mathcal{X}_1}(Y_1)] + 4\mathbb{E}\left[\left(\tfrac{\partial}{\partial X_2}\log f(X_2|X_1)\right)^2\right]\right\} \\
&\quad + \sqrt{\frac{2}{\pi}}4K_1\frac{\ell^5}{n^{3/2}}\mathbb{E}[L^2(X_2)] + \sqrt{\frac{32}{\pi}}4K_1\frac{\ell^5}{n^{3/2}}\mathbb{E}[K^2(Y_1)\mathbb{1}_{\mathcal{X}_1}(Y_1)]\ .
\end{aligned}$$

As argued in the proof of Lemma B.1, $\mathbb{E}[K^2(Y_1)\mathbb{1}_{\mathcal{X}_1}(Y_1)] < \infty$; furthermore, the other expectations on the right are finite by assumption. The three terms on the right thus are $\mathcal{O}(n^{-1/2})$, $\mathcal{O}(n^{-3/2})$, and $\mathcal{O}(n^{-3/2})$, which implies that $\mathbb{E}\left[|R_n(\mathbf{X}_{1:2}, U_2)|\right] \to 0$ as $n \to \infty$.

We now turn to the second term on the RHS of (B.8); since the acceptance function takes values in $[0,1]$, this term is bounded by

$$\mathbb{E}_{X_2}\left[\left|n\mathbb{E}_{Y_2}\left[h(Y_2) - h(X_2)\right] - \frac{\ell^2}{2}h''(X_2)\right|\right] \leq \frac{n}{6}\ \mathbb{E}_{X_2}\left[\left|\mathbb{E}_{Y_2}\left[h'''(U_2)(Y_2 - X_2)^3\right]\right|\right]$$

for some $U_2 \in (X_2, Y_2)$ or $U_2 \in (Y_2, X_2)$. The term on the right arises from a third-order Taylor expansion of h with respect to Y_2 around X_2, along with the fact that $Y_2 \sim \mathcal{N}(X_2, \ell^2/n)$. Since $|h'''|$ is bounded by a constant, the previous expression is bounded by $K_2\ell^3/\sqrt{n}$ for some $K_2 > 0$, which converges to 0 as $n \to \infty$.

In a similar fashion, by Taylor expanding h to second order and using the fact that the functions $|h''|$ and g are bounded by $K_3 > 0$ and 1 respectively, the third term on the RHS of (B.8) satisfies

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}_{1:2}}\Bigg[\Bigg|n\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1:n}}&\left[\left(h(Y_2) - h(X_2)\right)(Y_2 - X_2)\tfrac{\partial}{\partial X_2}\log f(X_2|Y_1)g(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\
&\quad -\ell^2 h'(X_2)\tfrac{\partial}{\partial X_2}\log f(X_2|X_1)\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[g(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\Bigg|\Bigg] \\
\leq\ & \ell^2\mathbb{E}\left[|h'(X_2)|\left|\tfrac{\partial}{\partial X_2}\log f(X_2|Y_1) - \tfrac{\partial}{\partial X_2}\log f(X_2|X_1)\right|\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\
&+\frac{1}{\sqrt{2\pi}}K_3\frac{\ell^3}{n^{1/2}}\mathbb{E}\left[\left|\tfrac{\partial}{\partial X_2}\log f(X_2|Y_1)\right|\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\ .
\end{aligned}$$

From the Lipschitz continuity of $\frac{\partial}{\partial x_2}\log f(x_2|x_1)$ and the fact that h' is bounded in absolute value, the first term on the right of the inequality is bounded by $\ell^2 K_4\mathbb{E}[L(X_2)|Y_1 - X_1|] \leq \ell^3\sqrt{2}K_4\mathbb{E}[L(X_2)]/\sqrt{\pi n}$ for some $K_4 > 0$; it is thus $\mathcal{O}(n^{-1/2})$. The second term also is $\mathcal{O}(n^{-1/2})$ since $\mathbb{E}\left[|\frac{\partial}{\partial X_2}\log f(X_2|Y_1)|\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] < \infty$ (from the proof of Proposition C.1).

□

Lemma B.4. *As $n \to \infty$, we have*

$$\mathbb{E}_{\mathbf{X}_{1:n},\mathbf{Y}_{1,3:n}}\left[\left|\alpha(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2}) - \hat{\alpha}(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\right|\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \to 0\ ,$$

where $\alpha(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = 1 \wedge \exp\{\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\}$ with ε as in (B.5) *and $\hat{\alpha}(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = 1 \wedge \exp\{\hat{\varepsilon}(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)})\}$ with*

$$\begin{aligned}
\hat{\varepsilon}(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) = \log\frac{f_1(Y_1)}{f_1(x_1)} + \log\frac{f(Y_2|Y_1)}{f(x_2|x_1)} + \sum_{i=3}^{n}\tfrac{\partial}{\partial x_1}\log f(x_i|x_1)(Y_1 - x_1) \\
+\frac{1}{2}\sum_{i=3}^{n}\tfrac{\partial^2}{\partial x_1^2}\log f(x_i|x_1)(Y_1 - x_1)^2 + \sum_{i=3}^{n}\tfrac{\partial}{\partial x_i}\log f(x_i|x_1)(Y_i - x_i) - \frac{\ell^2}{2n}\sum_{i=3}^{n}\left(\tfrac{\partial}{\partial x_i}\log f(x_i|x_1)\right)^2\ .
\end{aligned} \tag{B.9}$$

Proof. The function ε in (B.5) is reexpressed as

$$\begin{aligned}\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) &= \log \frac{f_1(Y_1)}{f_1(x_1)} + \log \frac{f(Y_2|Y_1)}{f(x_2|x_1)} + \sum_{i=3}^{n} (\log f(Y_i|Y_1) - \log f(Y_i|x_1)) \\ &\quad + \sum_{i=3}^{n} (\log f(Y_i|x_1) - \log f(x_i|x_1)) \,.\end{aligned}$$

Using second-order Taylor expansions with respect to Y_i around x_i ($i = 3, \ldots, n$) to reexpress the last two terms on the right hand side leads to

$$\begin{aligned}\varepsilon(\mathbf{x}^{(n)}, \mathbf{Y}^{(n)}) &= \log \frac{f_1(Y_1)}{f_1(x_1)} + \log \frac{f(Y_2|Y_1)}{f(x_2|x_1)} + \sum_{i=3}^{n} (\log f(x_i|Y_1) - \log f(x_i|x_1)) \\ &\quad + \sum_{i=3}^{n} \left(\tfrac{\partial}{\partial x_i} \log f(x_i|Y_1) - \tfrac{\partial}{\partial x_i} \log f(x_i|x_1) \right) (Y_i - x_i) \\ &\quad + \frac{1}{2} \sum_{i=3}^{n} \left(\tfrac{\partial^2}{\partial U_i^2} \log f(U_i|Y_1) - \tfrac{\partial^2}{\partial U_i^2} \log f(U_i|x_1) \right) (Y_i - x_i)^2 \\ &\quad + \sum_{i=3}^{n} \tfrac{\partial}{\partial x_i} \log f(x_i|x_1)(Y_i - x_i) + \frac{1}{2} \sum_{i=3}^{n} \tfrac{\partial^2}{\partial x_i^2} \log f(x_i|x_1)(Y_i - x_i)^2 \\ &\quad + \frac{1}{2} \sum_{i=3}^{n} \left(\tfrac{\partial^2}{\partial V_i^2} \log f(V_i|x_1) - \tfrac{\partial^2}{\partial x_i^2} \log f(x_i|x_1) \right) (Y_i - x_i)^2\end{aligned}$$

for some $U_i, V_i \in (x_i, Y_i)$ or $U_i, V_i \in (Y_i, x_i)$. Furthermore, by Taylor expanding the third term of the previous expression to second order (with respect to Y_1 around x_1) we obtain

$$\begin{aligned}&\sum_{i=3}^{n} (\log f(x_i|Y_1) - \log f(x_i|x_1)) \\ &\quad = \sum_{i=3}^{n} \tfrac{\partial}{\partial x_1} \log f(x_i|x_1)(Y_1 - x_1) + \frac{1}{2} \sum_{i=3}^{n} \tfrac{\partial^2}{\partial x_1^2} \log f(x_i|x_1)(Y_1 - x_1)^2 \\ &\qquad + \frac{1}{2} \sum_{i=3}^{n} \left(\tfrac{\partial^2}{\partial U_1^2} \log f(x_i|U_1) - \tfrac{\partial^2}{\partial x_1^2} \log f(x_i|x_1) \right) (Y_1 - x_1)^2\end{aligned}$$

for some $U_1 \in (x_1, Y_1)$ or $U_1 \in (Y_1, x_1)$.

Using the Lispchitz property of $1 \wedge \exp\{\cdot\}$ yields

$$\begin{aligned}
&\mathbb{E}\left[\left|1 \wedge \exp\{\varepsilon(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\} - 1 \wedge \exp\{\hat{\varepsilon}(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\}\right| \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\left|\sum_{i=3}^{n} \left(\tfrac{\partial^2}{\partial U_1^2} \log f(X_i|U_1) - \tfrac{\partial^2}{\partial X_1^2} \log f(X_i|X_1)\right)(Y_1 - X_1)^2 \mathbb{1}_{\mathcal{X}_1}(Y_1)\right|\right] \\
&\quad + \mathbb{E}\left[\left|\sum_{i=3}^{n} \left(\tfrac{\partial}{\partial X_i} \log f(X_i|Y_1) - \tfrac{\partial}{\partial X_i} \log f(X_i|X_1)\right)(Y_i - X_i) \mathbb{1}_{\mathcal{X}_1}(Y_1)\right|\right] \\
&\quad + \frac{1}{2}\mathbb{E}\left[\left|\sum_{i=3}^{n} \left(\tfrac{\partial^2}{\partial U_i^2} \log f(U_i|Y_1) - \tfrac{\partial^2}{\partial U_i^2} \log f(U_i|X_1)\right)(Y_i - X_i)^2 \mathbb{1}_{\mathcal{X}_1}(Y_1)\right|\right] \\
&\quad + \mathbb{E}\left[\left|\frac{1}{2}\sum_{i=3}^{n} \tfrac{\partial^2}{\partial X_i^2} \log f(X_i|X_1)(Y_i - X_i)^2 + \frac{\ell^2}{2n}\sum_{i=3}^{n}\left(\tfrac{\partial}{\partial X_i} \log f(X_i|X_1)\right)^2\right|\right] \\
&\quad + \frac{1}{2}\mathbb{E}\left[\left|\sum_{i=3}^{n} \left(\tfrac{\partial^2}{\partial V_i^2} \log f(V_i|X_1) - \tfrac{\partial^2}{\partial X_i^2} \log f(X_i|X_1)\right)(Y_i - X_i)^2\right|\right] . \qquad \text{(B.10)}
\end{aligned}$$

It remains to show that each term on the right hand side converges to 0 as $n \to \infty$. We look at the first term of (B.10). Using the triangle's inequality and the fact that $(Y_1 - X_1) \sim \mathcal{N}(0, \ell^2/n)$, we have

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\left|\sum_{i=3}^{n} \left(\tfrac{\partial^2}{\partial U_1^2} \log f(X_i|U_1) - \tfrac{\partial^2}{\partial X_1^2} \log f(X_i|X_1)\right)(Y_1 - X_1)^2 \mathbb{1}_{\mathcal{X}_1}(Y_1)\right|\right] \\
&\quad \leq \frac{\ell^2}{2}\left(\frac{n-2}{n}\right)\mathbb{E}\left[\left|\tfrac{\partial^2}{\partial U_1^2} \log f(X_3|U_1) - \tfrac{\partial^2}{\partial X_1^2} \log f(X_3|X_1)\right| Z_1^2 \mathbb{1}_{\mathcal{X}_1}(Y_1)\right] ,
\end{aligned}$$

where $Z_1 = \sqrt{n}(Y_1 - x_1)/\ell \sim \mathcal{N}(0, 1)$. Since $|U_1 - X_1| \leq |Y_1 - X_1|$ and $Y_1 \in \mathcal{X}_1$, then $U_1 \in \mathcal{X}_1$; in addition, $Y_1 \to_{a.s.} X_1$ implies that $U_1 \to_{a.s.} X_1$. By the Continuous Mapping Theorem, $|\frac{\partial^2}{\partial U_1^2} \log f(X_3|U_1) - \frac{\partial^2}{\partial X_1^2} \log f(X_3|X_1)|\mathbb{1}_{\mathcal{X}_1}(Y_1) \to_{a.s.} 0$. Since this term is bounded by $2M(X_3) \geq 0$ and that $2\mathbb{E}[M(X_3)Z_1^2] = 2\mathbb{E}[M(X_3)] < \infty$, the Dominated Convergence Theorem can be used to conclude that the first term on the right of (B.10) converges to 0 as $n \to \infty$.

We now consider the second term. Given $x_1, Y_1 \in \mathcal{X}_1$ and $x_i \in \mathbb{R}$ $(i = 3, \ldots, n)$,

$$\begin{aligned}
&\sum_{i=3}^{n} \left(\tfrac{\partial}{\partial x_i} \log f(x_i|Y_1) - \tfrac{\partial}{\partial x_i} \log f(x_i|x_1)\right)(Y_i - x_i) \\
&\quad \sim \ \mathcal{N}\left(0, \frac{\ell^2}{n}\sum_{i=3}^{n}\left(\tfrac{\partial}{\partial x_i} \log f(x_i|Y_1) - \tfrac{\partial}{\partial x_i} \log f(x_i|x_1)\right)^2\right) .
\end{aligned}$$

Computing the expectation of the half-normal distribution, applying Jensen's inequality (for the square root function, which is concave), and then the triangle inequality lead to

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}_{1,3:n},\mathbf{Y}_{1,3:n}}\left[\left|\sum_{i=3}^{n}\left(\tfrac{\partial}{\partial X_i}\log f(X_i|Y_1)-\tfrac{\partial}{\partial X_i}\log f(X_i|X_1)\right)(Y_i-X_i)\mathbb{1}_{\mathcal{X}_1}(Y_1)\right|\right]\\ &\leq \sqrt{\frac{2\ell^2}{\pi}\left(\frac{n-2}{n}\right)}\left(\mathbb{E}_{\mathbf{X}_{1,3},Y_1}\left[\left(\tfrac{\partial}{\partial X_3}\log f(X_3|Y_1)-\tfrac{\partial}{\partial X_3}\log f(X_3|X_1)\right)^2\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\right)^{1/2} .\end{aligned}$$

Since $Y_1 \to_{a.s.} X_1$, then $\left(\frac{\partial}{\partial X}\log f(X|Y_1)-\frac{\partial}{\partial X}\log f(X|X_1)\right)^2 \to_{a.s.} 0$ by the Continuous Mapping Theorem. Furthermore, we know that

$$\begin{aligned}&\mathbb{E}\left[\left(\tfrac{\partial}{\partial X}\log f(X|Y_1)-\tfrac{\partial}{\partial X}\log f(X|X_1)\right)^4\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\\ &\leq \mathbb{E}\left[L^4(X)(Y_1-x_1)^4\right]=3\frac{\ell^4}{n^2}\mathbb{E}\left[L^4(X)\right]<\infty ;\end{aligned}$$

the Uniform Integrability Theorem can then be used to conclude that the second term on the right hand side of (B.10) converges to 0 as $n\to\infty$.

Using the triangle's inequality and the fact that $(Y_i-X_i)\sim\mathcal{N}(0,\ell^2/n)$ $(i=3,\ldots,n)$, the third term on the right hand side of (B.10) is bounded by

$$\frac{\ell^2}{2}\left(\frac{n-2}{n}\right)\mathbb{E}\left[\left|\tfrac{\partial^2}{\partial U_3^2}\log f(U_3|Y_1)-\tfrac{\partial^2}{\partial U_3^2}\log f(U_3|X_1)\right|Z_3^2\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] ,$$

where $Z_3=\sqrt{n}(Y_3-X_3)/\ell\sim\mathcal{N}(0,1)$. Given that $Y_1\to_{a.s.} X_1$, the Continuous Mapping Theorem implies that $\left|\frac{\partial^2}{\partial U_3^2}\log f(U_3|Y_1)-\frac{\partial^2}{\partial U_3^2}\log f(U_3|X_1)\right|\to_{a.s.} 0$ as $n\to\infty$. We again invoke the Uniform Integrability Theorem to conclude that the third term on the right converges to 0 as $n\to\infty$, since

$$\begin{aligned}&\mathbb{E}\left[\left(\tfrac{\partial^2}{\partial U_3^2}\log f(U_3|Y_1)-\tfrac{\partial^2}{\partial U_3^2}\log f(U_3|X_1)\right)^2 Z_3^4\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\\ &\leq 6\mathbb{E}\left[K^2(Y_1)\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]+6\mathbb{E}\left[K^2(X_1)\right]<\infty .\end{aligned}$$

Replacing Y_1 by X_1 in the proof of Proposition C.1, the fourth term on the right of (B.10) is easily seen to converge towards 0 as $n\to\infty$. Finally, the last term is bounded by

$$\frac{\ell^2}{2}\left(\frac{n-2}{n}\right)\mathbb{E}\left[\left|\tfrac{\partial^2}{\partial V_3^2}\log f(V_3|X_1)-\tfrac{\partial^2}{\partial X_3^2}\log f(X_3|X_1)\right|Z_3^2\right] ,$$

with $Z_3=\sqrt{n}(Y_3-X_3)/\ell$. Given that $Y_3\to_{a.s.} X_3$ and $|V_3-X_3|\leq|Y_3-X_3|$, we have $V_3\to_{a.s.} X_3$ and the Continuous Mapping Theorem implies that the integrand converges to 0 almost surely. Furthermore, the integrand is bounded by $2K(X_1)Z_3^2$ and since $2\mathbb{E}[K(X_1)Z_3^2]=2\mathbb{E}[K(X_1)]<\infty$, the Dominated Convergence Theorem is used to conclude the proof.

□

Lemma B.5. *As $n\to\infty$, we have*

$$\mathbb{E}_{\mathbf{X}_{1:2}}\left[\left|\ell^2\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[\hat{\alpha}(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]-\upsilon(\ell,X_1)\right|\right]\to 0 ,$$

where $\hat{\alpha}(\mathbf{x}^{(n)},\mathbf{Y}^{(n)})=1\wedge\exp\{\hat{\varepsilon}(\mathbf{x}^{(n)},\mathbf{Y}^{(n)})\}$ with the function $\hat{\varepsilon}$ as in (B.9) *and the function υ as in* (6).

Proof. We have from the triangle inequality

$$\mathbb{E}_{\mathbf{X}_{1:2}}\left[\left|\ell^2\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{1,3:n}}\left[\hat{\alpha}(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]-\upsilon(\ell,X_1)\right|\right]\leq$$
$$\ell^2\mathbb{E}_{\mathbf{X}_{1:2},Z_1}\left[\left|\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{3:n}}\left[1\wedge\exp\{\hat{\varepsilon}(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\}\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]-2\Phi\left(-\frac{\ell}{2}\gamma^{1/2}(X_1,Z_1)\right)\right|\right]\ ,$$

where $Z_1=\sqrt{n}(Y_1-x_1)/\ell\sim\mathcal{N}(0,1)$. From the boundedness of the absolute value in the above expression, it is sufficient to show that, conditionally on $x_1\in\mathcal{X}_1$, $x_2,Z_1\in\mathbb{R}$,

$$\left|\mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{3:n}}\left[1\wedge\exp\{\hat{\varepsilon}(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\}\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]-2\Phi\left(-\frac{\ell}{2}\gamma^{1/2}(X_1,Z_1)\right)\right|\to 0\quad\text{as}\quad n\to\infty\ .$$

The function $\hat{\varepsilon}$ being evaluated at $Y_2=x_2$, it is reexpressed as

$$\begin{aligned}\hat{\varepsilon}(\mathbf{x}^{(n)},(x_1+\tfrac{\ell}{\sqrt{n}}Z_1,x_2,\mathbf{Y}_{3:n})) &= \log\frac{f_1(x_1+\frac{\ell}{\sqrt{n}}Z_1)}{f_1(x_1)}+\log\frac{f(x_2|x_1+\frac{\ell}{\sqrt{n}}Z_1)}{f(x_2|x_1)} \\ &\quad+\frac{\ell}{\sqrt{n}}\sum_{i=3}^{n}\tfrac{\partial}{\partial x_1}\log f(x_i|x_1)Z_1+\frac{\ell^2}{2n}\sum_{i=3}^{n}\tfrac{\partial^2}{\partial x_1^2}\log f(x_i|x_1)Z_1^2 \\ &\quad+\sum_{i=3}^{n}\tfrac{\partial}{\partial x_i}\log f(x_i|x_1)(Y_i-x_i)-\frac{\ell^2}{2n}\sum_{i=3}^{n}\left(\tfrac{\partial}{\partial x_i}\log f(x_i|x_1)\right)^2\ .\end{aligned}\tag{B.11}$$

In the sequel, we thus condition on $x_1\in\mathcal{X}_1$, $x_2,Z_1\in\mathbb{R}$, and study the convergence of every term in (B.11) as $n\to\infty$. Given any $x_1\in\mathcal{X}_1$ and $Z_1\in\mathbb{R}$, $\exists n^*\geq 1$ such that $x_1+\frac{\ell}{\sqrt{n}}Z_1\in\mathcal{X}_1$ for all $n\geq n^*$; it therefore follows from the continuity of functions that $\log\{f_1(Y_1)/f_1(x_1)\}\to 0$ and $\log\{f(x_2|Y_1)/f(x_2|x_1)\}\to 0$ for any given $x_2\in\mathbb{R}$. We now show that conditionally on x_1,Z_1, the remaining terms are asymptotically distributed according to a normal random variable.

Given any $x_1\in\mathcal{X}_1$, $Z_1\in\mathbb{R}$, applying the Central Limit Theorem to the third term of (B.11) yields

$$\frac{\ell}{\sqrt{n}}Z_1\sum_{i=3}^{n}\tfrac{\partial}{\partial x_1}\log f(X_i|x_1)\to_d\mathcal{N}\left(0,\ell^2Z_1^2\ \mathbb{E}_X\left[\left(\tfrac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right]\right).$$

This follows from the regularity assumptions in Section 2, which imply that $\frac{\partial}{\partial x_1}f(x|x_1)$ is locally integrable and thus that we can differentiate outside of the integral sign to obtain

$$\mathbb{E}_X\left[\tfrac{\partial}{\partial x_1}\log f(X|x_1)\right]=\tfrac{\mathrm{d}}{\mathrm{d}x_1}\int_{\mathbb{R}}f(x|x_1)\ \mathrm{d}x=0.$$

To study the fourth term, we condition on $x_1\in\mathcal{X}_1$, $Z_1\in\mathbb{R}$ and use the SLLN to get

$$\frac{\ell^2}{2n}Z_1^2\sum_{i=3}^{n}\tfrac{\partial^2}{\partial x_1^2}\log f(X_i|x_1)\quad\to_{a.s.}\quad\frac{\ell^2}{2}Z_1^2\ \mathbb{E}_X\left[\tfrac{\partial^2}{\partial x_1^2}\log f(X|x_1)\right]\ .$$

Again from the regularity assumptions, $\frac{\partial^2}{\partial x_1^2}f(x|x_1)$ is locally integrable and thus the following identity holds :

$$\mathbb{E}_X\left[\tfrac{\partial^2}{\partial x_1^2}\log f(X|x_1)\right]+\mathbb{E}_X\left[\left(\tfrac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right]=\tfrac{\mathrm{d}^2}{\mathrm{d}x_1^2}\int_{\mathbb{R}}f(x|x_1)\ \mathrm{d}x=0\ ;$$

therefore, $\mathbb{E}_X\left[\frac{\partial^2}{\partial x_1^2}\log f(X|x_1)\right] = -\mathbb{E}_X\left[\left(\frac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right]$ for all $x_1 \in \mathcal{X}_1$.

Combining the previous developments and making use of Slutsky's Theorem allows us to conclude that given any $x_1 \in \mathcal{X}_1, Z_1 \in \mathbb{R}$,

$$\frac{\ell}{\sqrt{n}} Z_1 \sum_{i=3}^{n} \frac{\partial}{\partial x_1}\log f(X_i|x_1) + \frac{\ell^2}{2n} Z_1^2 \sum_{i=3}^{n} \frac{\partial^2}{\partial x_1^2}\log f(X_i|x_1)$$
$$\to_d \;\; \mathcal{N}\left(-\frac{\ell^2}{2} Z_1^2 \mathbb{E}_X\left[\left(\frac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right],\; \ell^2 Z_1^2 \mathbb{E}_X\left[\left(\frac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right]\right) \;.$$

Now, given any $x_1 \in \mathcal{X}_1$, the last two terms on the right of (B.11) satisfy

$$\frac{\ell}{\sqrt{n}} \sum_{i=3}^{n} \frac{\partial}{\partial X_i}\log f(X_i|x_1) Z_i - \frac{\ell^2}{2n}\sum_{i=3}^{n}\left(\frac{\partial}{\partial X_i}\log f(X_i|x_1)\right)^2,$$
$$\to_p \mathcal{N}\left(-\frac{\ell^2}{2}\mathbb{E}_X\left[\left(\frac{\partial}{\partial X}\log f(X|x_1)\right)^2\right],\; \ell^2 \mathbb{E}_X\left[\left(\frac{\partial}{\partial X}\log f(X|x_1)\right)^2\right]\right) \;;$$

this follows from the WLLN and the fact that $Z_i \sim \mathcal{N}(0,1)$ independently for $i = 3, \ldots, n$.

Given $x_1 \in \mathcal{X}_1$, $Z_1 \in \mathbb{R}$, the two normal random variables just introduced are asymptotically independent (this is easily seen from the fact that $\sqrt{n}(\mathbf{Y}_{3:n} - \mathbf{x}_{3:n})/\ell^2$ is independent of $\mathbf{x}_{3:n}$ $\forall n \geq 3$). We therefore conclude that given any $X_1 \in \mathcal{X}_1$ and $Z_1 \in \mathbb{R}$, $\hat{\varepsilon}(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2}) \to_d \eta(X_1, Z_1)$, where $\eta(x_1, Z_1) \sim \mathcal{N}\left(-\ell^2\gamma(x_1, Z_1)/2, \ell^2\gamma(x_1, Z_1)\right)$, with

$$\gamma(x_1, Z_1) = Z_1^2 \;\mathbb{E}_X\left[\left(\frac{\partial}{\partial x_1}\log f(X|x_1)\right)^2\right] + \mathbb{E}_X\left[\left(\frac{\partial}{\partial X}\log f(X|x_1)\right)^2\right] \;.$$

It easily follows from the fact that $\mathbb{1}_{\mathcal{X}_1}(x_1 + \frac{\ell}{\sqrt{n}}Z_1) \to 1$ given any $x_1 \in \mathcal{X}_1$, $Z_1 \in \mathbb{R}$, Slutsky's Theorem, and the Continuous Mapping Theorem, that $1 \wedge \exp\{\hat{\varepsilon}(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\}\mathbb{1}_{\mathcal{X}_1}(Y_1) \to_d 1 \wedge \exp\{\eta(X_1, Z_1)\}$. From Proposition 2.4 in [17], we know that given x_1, Z_1,

$$\mathbb{E}_\eta[1 \wedge \exp\{\eta(x_1, Z_1)\}] = 2\Phi\left(-\frac{\ell}{2}\gamma^{1/2}(x_1, Z_1)\right) \;.$$

From the convergence in distribution and the boundedness (and thus uniform integrability) of the random variables, the means are known to converge, *i.e.* given any $x_1 \in \mathcal{X}_1$ and $x_2, Z_1 \in \mathbb{R}$

$$\left|\mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{3:n}}\left[1 \wedge \exp\{\hat{\varepsilon}(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\}\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] - 2\Phi\left(-\frac{\ell}{2}\gamma^{1/2}(X_1, Z_1)\right)\right| \to 0 \;,$$

which concludes the proof.

$\square$

Lemma B.6. *As $n \to \infty$, we have*

$$\mathbb{E}_{\mathbf{X}_{1:2}}\left[\left|\frac{\partial}{\partial X_2}\log f(X_2|X_1)\right| \left|\ell^2 \mathbb{E}_{\mathbf{X}_{3:n}, \mathbf{Y}_{1,3:n}}\left[g(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] - \frac{1}{2}\upsilon(\ell, X_1)\right|\right] \to 0 \;,$$

where the function g is as in (B.4).

Proof. Making use of the triangle inequality, we may bound the expectation in the statement of the lemma by

$$\ell^2 \mathbb{E}_{\mathbf{X}_{1:2},Z_1}\left[|\tfrac{\partial}{\partial X_2}\log f(X_2|X_1)| \left| \mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{3:n}}\left[g(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] - \Phi\left(-\frac{\ell}{2}\gamma^{1/2}(X_1,Z_1)\right)\right|\right] \ ,$$

which is itself bounded by $2\mathbb{E}\left[|\frac{\partial}{\partial X_2}\log f(X_2|X_1)|\right] < \infty$ since each term in the difference is bounded by 1 in absolute value. We can thus use the Dominated Convergence Theorem to bring the limit inside the first expectation. To conclude the proof, all is left to do is to verify that given any $X_1 \in \mathcal{X}_1$, $X_2, Z_1 \in \mathbb{R}$,

$$\left| \mathbb{E}_{\mathbf{X}_{3:n},\mathbf{Y}_{3:n}}\left[g(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] - \Phi\left(-\frac{\ell}{2}\gamma^{1/2}(X_1,Z_1)\right)\right| \to 0 \ ,$$

where $Z_1 = \sqrt{n}(Y_1 - X_1)/\ell$.

In the proof of Lemma B.4 we have verified, among other things, that $\mathbb{E}[|\varepsilon(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2}) - \hat{\varepsilon}(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})|\mathbb{1}_{\mathcal{X}_1}(Y_1)] \to 0$ as $n \to \infty$. This $\mathcal{L}^1$-convergence thus entails that $|\varepsilon(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2}) - \hat{\varepsilon}(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})|\mathbb{1}_{\mathcal{X}_1}(Y_1) \to_p 0$. From the proof of Lemma B.5 we know that given any $X_1 \in \mathcal{X}_1$ and $X_2, Z_1 \in \mathbb{R}$, $\hat{\varepsilon}(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1) \to_d \eta(X_1,Z_1)$. Using Slutsky's Theorem, these convergences imply that, conditionally on $X_1 \in \mathcal{X}_1$ and $X_2, Z_1 \in \mathbb{R}$, $\varepsilon(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1) \to_d \eta(X_1,Z_1)$.

From the Continuous Mapping Theorem, we deduce that given any $X_1 \in \mathcal{X}_1$, $Z_1 \in \mathbb{R}$,

$$g(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}_{X_2})\mathbb{1}_{\mathcal{X}_1}(Y_1) \to_d \exp\{\eta(X_1,Z_1)\}\ \mathbb{1}\left\{\exp\{\eta(X_1,Z_1)\} < 1\right\} \ .$$

The function under study is obviously not continuous; however, the discontinuities of the function on the right have null Lebesgue measure and thus the Continuous Mapping Theorem is applicable as stated in [8] (Theorem 5.1 and its corollaries).

By examining the proof of Proposition 2.4 in [17] we obtain, conditionally on $X_1 \in \mathcal{X}_1$, $Z_1 \in \mathbb{R}$,

$$\mathbb{E}_{\eta}\left[\exp\{\eta(X_1,Z_1)\}\ \mathbb{1}\left\{\exp\{\eta(X_1,Z_1)\} < 1\right\}\right] = \Phi\left(-\frac{\ell}{2}\gamma^{1/2}(X_1,Z_1)\right) \ .$$

From the convergence in distribution and the fact that the random variables under consideration are bounded (and thus uniformly integrable), the means are known to converge; this concludes the proof of the lemma. □

C. Appendix

Proposition C.1. *Define*

$$W(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}) \ = \ \frac{1}{2}\sum_{i=2}^{n}\frac{\partial^2}{\partial X_i^2}\log f(X_i|Y_1)(Y_i - X_i)^2 + \frac{\ell^2}{2n}\sum_{i=2}^{n}\left(\frac{\partial}{\partial X_i}\log f(X_i|Y_1)\right)^2 \ ;$$

then, $\mathbb{E}[|W(\mathbf{X}^{(n)},\mathbf{Y}^{(n)})|\mathbb{1}_{\mathcal{X}_1}(Y_1)] \to 0$ *as* $n \to \infty$.

Proof. By Jensen's inequality, $\mathbb{E}[|W|] \leq \sqrt{\mathbb{E}[W^2]}$. Developing the square and taking the expectation with respect to $\mathbf{Y}_{2:n}$ yield

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_{2:n}}\left[W^2(\mathbf{X}^{(n)},\mathbf{Y}^{(n)})\right] &= \frac{\ell^4}{2n^2}\sum_{i=2}^{n}\left(\tfrac{\partial^2}{\partial X_i^2}\log f(X_i|Y_1)\right)^2 \\ &\quad+\frac{\ell^4}{4n^2}\left\{\sum_{i=2}^{n}\left(\tfrac{\partial^2}{\partial X_i^2}\log f(X_i|Y_1)+\left(\tfrac{\partial}{\partial X_i}\log f(X_i|Y_1)\right)^2\right)\right\}^2 ,\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_{2:n}}\left[|W(\mathbf{X}^{(n)},\mathbf{Y}^{(n)})|\right] &\leq \frac{\ell^2}{\sqrt{2n}}\left(\frac{1}{n}\sum_{i=2}^{n}\left(\tfrac{\partial^2}{\partial X_i^2}\log f(X_i|Y_1)\right)^2\right)^{1/2} \\ &\quad+\frac{\ell^2}{2}\left|\frac{1}{n}\sum_{i=2}^{n}\left(\tfrac{\partial^2}{\partial X_i^2}\log f(X_i|Y_1)+\left(\tfrac{\partial}{\partial X_i}\log f(X_i|Y_1)\right)^2\right)\right| .\end{aligned}$$

Reapplying Jensen's inequality on the first term and developing the second term lead to

$$\begin{aligned}\mathbb{E}[|W(\mathbf{X}^{(n)},\mathbf{Y}^{(n)})|\mathbb{1}_{\mathcal{X}_1}(Y_1)] &\leq \frac{\ell^2}{\sqrt{2n}}\left\{\mathbb{E}\left[\left(\tfrac{\partial^2}{\partial X^2}\log f(X|Y_1)\right)^2\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]\right\}^{1/2} \\ &\quad+\frac{\ell^2}{2}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=2}^{n}\left(\tfrac{\partial^2}{\partial X_i^2}\log f(X_i|X_1)+\left(\tfrac{\partial}{\partial X_i}\log f(X_i|X_1)\right)^2\right)\right|\right] \\ &\quad+\frac{\ell^2}{2}\left(\frac{n-1}{n}\right)\mathbb{E}\left[\left|\tfrac{\partial^2}{\partial X_i^2}\log f(X_i|Y_1)-\tfrac{\partial^2}{\partial X_i^2}\log f(X_i|X_1)\right|\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\ &\quad+\frac{\ell^2}{2}\left(\frac{n-1}{n}\right)\mathbb{E}\left[\left|\left(\tfrac{\partial}{\partial X_i}\log f(X_i|Y_1)\right)^2-\left(\tfrac{\partial}{\partial X_i}\log f(X_i|X_1)\right)^2\right|\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] .\end{aligned}$$

The first term on the right is bounded by $\ell^2\left\{\mathbb{E}\left[K^2(Y_1)\mathbb{1}_{\mathcal{X}_1}(Y_1)\right]/(2n)\right\}^{1/2}$, which converges to 0 as $n\to\infty$ from the argument at the end of the proof of Lemma B.1. From Lemma 12 in [2], we know that $\frac{\partial}{\partial x}\log f(x|x_1)\to 0$ as $x\to\pm\infty$, $\forall x_1\in\mathcal{X}_1$; hence, given x_1, we have $\mathbb{E}_X[\frac{\partial^2}{\partial X^2}\log f(X|x_1)+\left(\frac{\partial}{\partial X}\log f(X|x_1)\right)^2]=\int\frac{\partial^2}{\partial x^2}f(x|x_1)\mathrm{d}x=0$ and by the WLLN,

$$\left|\frac{1}{n}\sum_{i=2}^{n}\left(\tfrac{\partial^2}{\partial X_i^2}\log f(X_i|X_1)+\left(\tfrac{\partial}{\partial X_i}\log f(X_i|X_1)\right)^2\right)\right|\to_p 0 .$$

To invoke the Uniform Integrability Theorem for the second term, we use the finiteness of $\mathbb{E}[(\frac{\partial^2}{\partial X^2}\log f(X|X_1))^2]\leq\mathbb{E}[K^2(X_1)]$ and $\mathbb{E}[(\frac{\partial}{\partial X}\log f(X|X_1))^4]$.

From $Y_1\to_{a.s.}x_1$ and the Continuous Mapping Theorem, the integrands of the last two terms are seen to converge to 0 almost surely. Since $\mathbb{E}[K^2(Y_1)\mathbb{1}_{\mathcal{X}_1}(Y_1)]<\infty$ (Section A.2) and $\mathbb{E}[K^2(X_1)]<\infty$, the third term converges to 0 using the Uniform Integrability Theorem. We come to the same conclusion for the fourth term, using $\mathbb{E}[(\frac{\partial}{\partial X}\log f(X|X_1))^4]<\infty$ and

$$\begin{aligned}\mathbb{E}\left[\left(\tfrac{\partial}{\partial X}\log f(X|Y_1)\right)^4\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] &\leq 8\mathbb{E}\left[\left(\tfrac{\partial}{\partial X}\log f(X|Y_1)-\tfrac{\partial}{\partial X}\log f(X|X_1)\right)^4\mathbb{1}_{\mathcal{X}_1}(Y_1)\right] \\ &\quad+8\mathbb{E}\left[\left(\tfrac{\partial}{\partial X}\log f(X|X_1)\right)^4\right] \\ &\leq 24\frac{\ell^4}{n^2}\mathbb{E}\left[L^4(X)\right]+8\mathbb{E}\left[\left(\tfrac{\partial}{\partial X}\log f(X|X_1)\right)^4\right]<\infty .\end{aligned}$$

□