ON THE ROBUSTNESS OF OPTIMAL SCALING FOR RANDOM WALK METROPOLIS ALGORITHMS

by

Mylène Bédard

A thesis submitted in conformity with the requirements

for the degree of Doctor of Philosophy

Graduate Department of Statistics

University of Toronto

On the Robustness of Optimal Scaling for Random Walk Metropolis Algorithms

Mylène Bédard

Department of Statistics, University of Toronto

Ph.D. Thesis, 2006

Abstract

In this thesis, we study the optimal scaling problem for sampling from a target distribution of interest using a random walk Metropolis (RWM) algorithm. In order to implement this method, the selection of a proposal distribution is required, which is assumed to be a multivariate normal distribution with independent components. We investigate how the proposal scaling (i.e. the variance of the normal distribution) should be selected for best performance of the algorithm.

The d-dimensional target distribution we consider is formed of independent components, each of which has its own scaling term $\theta_j^{-2}(d)$ $(j = 1, \ldots, d)$. This constitutes an extension of the d-dimensional *iid* target considered by Roberts, Gelman & Gilks (1997) who showed that for large d, the acceptance rate should be tuned to 0.234 for optimal performance of the algorithm. In a similar fashion, we show that for the aforementioned framework, the relative efficiency of the algorithm can be characterized by its overall acceptance rate.

We first propose a method to determine the optimal form for the proposal

scaling as a function of d. This assists us in producing a necessary and sufficient condition for the algorithm to adopt the same limiting behavior as with *iid* targets, resulting in an asymptotically optimal acceptance rate (AOAR) of 0.234. We show that when this condition is violated the limiting process of the algorithm is altered, yielding AOARs which might drastically differ from the usual 0.234. We also demonstrate that inhomogeneous proposal distributions are sometimes essential to obtain of a nontrivial limit. Finally, we illustrate how our results can be applied to the special case of correlated targets whose distribution is jointly normal.

Specifically, we prove that the sequence of stochastic processes formed by say the i^*-th component of each Markov chain usually converges (as $d \to \infty$) to a Langevin diffusion process whose speed measure varies according to target scaling vector. The demonstrations involve an appropriate rescaling of space and time, as well as $\mathcal{L}^1$ convergence of generators.

We illustrate with a number of examples how practitioners can take advantage of these results.

Acknowledgements

I would like to express my gratitude to all those who gave me the possibility to complete this thesis.

First and foremost, I would like to thank Professor Jeffrey S. Rosenthal, the best supervisor one could wish for. Without his expertise, active involvement, sound advices, and encouragements, this work would not have been possible. Thank you for pushing me, for getting 100% involved and for always having my best interest in mind. I appreciate you both as a supervisor and as a person.

I would like to thank the members of my committee, Professors V. Radu Craiu and X. Sheldon Lin, who somehow managed to make the committee meetings enjoyable. I am also grateful to the external examiner, Professor Wilfrid S. Kendall, for his insightful comments, and to Professor Gareth O. Roberts for useful discussions.

The University of Toronto and the Department of Statistics (both faculty and staff!) have provided great support throughout my graduate studies. Special appreciation goes to Professor Samuel A. Broverman for being so accommodating with TAs (and also just for having such a great sense of humor) and Professor Donald A.S. Fraser for interesting discussions. Thanks also go to Professor Thierry Duchesne for initiating me to research and to Laura Kerr who makes the sixth floor so warm. I am grateful to NSERC (National Sciences and Engineering Research Council of Canada) and Ontario Graduate Scholarship, for providing the necessary financial help.

I would like to thank some friends and colleagues: in particular Hanna Jankowski and Ana-Maria Staicu, but also Sigfrido Iglesias-Gonzalez, Samuel Hikspoors,

Baisong Huang, Xiaoming Liu, Elena Parkhomenko, Mohammed Shakhatreh, John Sheriff, and Zheng Zheng. I am also grateful to Vanessa Dionne, Alexandre Drouin and Amélie Grimard; I benefited from your company. I wish to express many thanks to Daniil Bunimovich, especially for that first ride (wouldn't have been from you, I would have gotten one hell of a ride!)

I am really grateful to Thierry Bédard and Cynthia Martel for their warm welcome, little attentions, and moral support during my frequent (and sometime lengthy) visits to Toronto. I would also like to thank my brother for all the proof-reading. Special appreciation also goes to Nathalie Dumont, who made so many aspects of my life better; her teaching and advices are priceless.

I owe many thanks to my parents, Nicole and Raynald, who have encouraged me all along and who have always been there for me. In particular, I would like to thank my mom for her listening and advices and my dad for his unconditional love. Thank you for believing in me.

Finally, I want to thank Pat, one of the best persons I know. Thank you for always being there, for being so thoughtful, for always making everything okay and most of all, thank you for making me laugh.

Contents

List of Figures

Introduction

In recent years, Markov chain Monte Carlo (MCMC) methods, which constitute powerful tools allowing for data generation from highly complex distributions, have raised a growing enthusiasm. Since their first appearance in the statistical physics literature (see [24]), these techniques have opened new horizons regarding the complexity of models that can be used in practice. As of this day, such methods are extensively used by researchers and practitioners in various fields of application such as biostatistics, computer science, physics, economics, finance, and applied statistics. They have had a deep impact on the progress in statistical genetics for instance, a field often dealing with highly complex hierarchical models. In statistical finance, they have permitted the elaboration of methods for inference and prediction of stochastic volatility models (see [22], for example). In applied statistics, MCMC methods are widely used in the Bayesian environment, where very high dimensional state spaces or extremely complex distributions are challenges that one is likely to encounter.

A number of researchers contributed to the development of these algorithms. Hastings' adaptation (see [21]) of the Metropolis algorithm (see [24]) has resulted in the famous generalization known today as the Metropolis-Hastings algorithm. The particularity of these algorithms resides in the necessity of choosing a pro-

posal distribution for their implementation. The work presented in [17] about the Bayesian restoration of images is famous in the statistical literature, as it constitutes the first instance of the Gibbs sampler algorithm. This algorithm utilizes the full conditional distributions to iteratively sample from a joint distribution. This work has also been one of the first applications of MCMC methods to Bayesian analysis. The data-augmentation algorithm presented in [39] constitutes a popular iterative method to compute Bayesian posterior distributions. A generalization of the Gibbs sampler and data-augmentation algorithms can be found in [18], along with several examples illustrating the use of these techniques. Various MCMC methods and some of their variants are outlined in [40], which also discusses theoretical and practical issues related to these methods.

The main advantage of MCMC methods resides in the simplicity of the underlying principle. In order to randomly sample from a specific probability distribution called the target distribution, it suffices to design a Markov chain whose stationary distribution is the target distribution. The Metropolis-Hastings and the Gibbs sampler algorithms are among the most popular methods of designing the Markov chain. However, there now exist several variations and hybrid methods which have been derived by modifying and/or combining these techniques. Once the Markov chain is created, we simulate it on the computer for a period of time long enough so as to be assured that the chain has reached stationarity. At that point, we record the state of the Markov chain and treat it as a sample point from the target distribution. A data set is thus obtained by either repeating this procedure many times or by recording several states after the burn-in period (and skipping more or less states in between draws, depending if we are looking for an independent

sample or not).

As the complexity of the target distribution intensifies, significant complications occur which have raised a number of questions related to probability theory and Markov chains. Namely, it has become crucial to understand as precisely as possible how fast the chain converges to its stationary distribution ([14], [15]). This has stimulated a great amount of research related to the bounding of convergence rates ([35], [23]), the central limit theory of MCMC algorithms ([34], [32]), as well as the scaling of proposal distributions for Metropolis-Hastings algorithms and the limiting behavior of these algorithms as the dimension of the target distribution increases. Optimal scaling refers to the need of tuning the proposal scaling and other parameters so as to obtain an efficient convergence of the algorithm. Nowadays, the trend has moved towards improving and modifying the existing algorithms using adaptive schemes. This can be seen as a dynamic extension of optimal scaling, where the algorithm attempts to find the optimal parameter values and update them while it runs (see [20]). This movement has spawned much research about whether or not the different adaptations satisfy the required properties for MCMC methods ([1], [2], [33]).

In this dissertation, we shall consider the optimal scaling problem for random walk Metropolis (RWM) algorithms, which constitute the most commonly used class of Metropolis-Hastings algorithms. These algorithms are named after the fact that the kernel driving the chain is a random walk (which is ensured by an appropriate choice of proposal distribution). Their ease of implementation and wide applicability have conferred their popularity to RWM algorithms and they are frequently used nowadays by all levels of practitioners in various fields of ap-

plication. A downside of their versatility is however the potential slowness of their convergence, which calls for an optimization of their performance. Because the efficiency of Metropolis-Hastings algorithms depends crucially on the scaling of the proposal distribution, it is thus fundamental to judiciously choose this parameter.

Informal guidelines for the optimal scaling problem have been proposed among others by [8] and [9], but the first theoretical results have been obtained by [29]. In particular, the authors considered d-dimensional target distributions with *iid* components and studied the asymptotic behavior (as $d \to \infty$) of RWM algorithms with Gaussian proposals. It was proved that under some regularity conditions for the target distribution, the asymptotic acceptance rate should be tuned to be approximately 0.234 for optimal performance of the algorithm. It was also shown that the correct variance for the proposal distribution is of the form ℓ^2/d for some constant ℓ as $d \to \infty$. The simplicity of the obtained asymptotically optimal acceptance rate (AOAR) makes these theoretical results extremely useful in practice. Afterwards, [30] carried out a similar study for Metropolis-adapted Langevin algorithms (MALA), whose proposal distribution uses the gradient of the target density to generate wiser moves. Analogous conclusions were obtained, but with an AOAR of 0.574. In spite of the *iid* assumption for the target components, in both cases the results are believed to be far more robust and to hold under various perturbations of the target distribution. Since their publication these papers have spawned much interest and consequently, other authors have studied diverse extensions of the *iid* model; their conclusions all corroborate the results found in [29] and [30] (see, for instance, [11], [12], [13], [26] and [31]).

The goal of this dissertation is to study the robustness of the optimal scaling

results introduced in [29] when the RWM algorithm is applied to other types of target distributions. In particular, we consider d-dimensional targets with independent but not identically distributed components, where the components possess different scaling terms which are allowed to depend on the dimension of the target distribution. This setting includes some popular statistical models, but it produces distributions having unstable scaling terms (since some of them might converge to 0 or ∞ as the dimension increases). Despite the independence of the various target components, the disparities exhibited by the scaling terms thus constitute a critical distinction with the *iid* case. Furthermore, because Gaussian distributions are invariant under orthogonal transformations, the model described also includes multivariate normal target distributions with correlated components as a special case.

We provide a necessary and sufficient condition under which the algorithm admits the same limiting diffusion process and the same AOAR as those found in [29]. We also prove that when this condition is violated, the acceptance rate 0.234 is not always asymptotically optimal, and we thus present methods for determining the correct AOAR. This addresses the issue raised in Open Problem #3 of [32]. In particular, we show that when there exists a finite number of target scaling terms converging significantly faster than the others, then the asymptotic acceptance rate optimizing the efficiency of the algorithm depends on specific target components only and is smaller than 0.234. Finally, we shall see that when the target distribution possesses too small scaling terms (i.e. scaling terms that are remote from the other ones), then the optimization problem is ill-posed which calls for the use of inhomogeneous proposal scalings.

In order to reach these conclusions, an appropriate rescaling of space and time allows us to obtain a nontrivial limiting process as $d \to \infty$. This is achieved in the first place by determining the appropriate form of the proposal scaling as a function of d, which is now different from the *iid* case. Then, by verifying $\mathcal{L}^1$ convergence of generators, we find the limiting distribution of the sequence of stochastic processes formed by say the i^*-th component of each Markov chain. In the majority of cases, this limiting distribution is a Langevin diffusion process with a certain speed measure. The speed measure then determines if the process behaves as in the *iid* case (same speed measure) or converges to a different limit (different speed measure). Speed measures being the sole measure of efficiency for diffusions, obtaining the AOAR is thus a simple matter of optimizing this quantity. In certain cases however, we might have to deal with certain target components whose limiting distribution remains a RWM algorithm, but with a particular acceptance rule. This arises when dealing with target components whose scaling term is significantly smaller than that of the other components. Since there exist multiple measures of efficiency for discrete-time processes, we choose not to use these components to determine the AOAR.

Throughout this dissertation, we also study how well these asymptotic results serve the practical side; this is achieved by presenting various examples illustrating the application of the theorems proved in this work. This demonstrates that, although of asymptotic nature, the results can be used to facilitate the tuning of algorithms in finite-dimensional problems. We finally apply the results to popular statistical models, and perform some simulation studies aiming to demonstrate their robustness to certain disruptions from the assumed target model.

This thesis is organized as follows. Chapter 1 presents some background information about Metropolis-Hastings algorithms, along with the optimal scaling results for *iid* target distributions. In Chapter 2, we introduce the assumed target model, which consists in an extension of the d-dimensional *iid* target distribution considered in [29]. We also present a method to determine the optimal form for the proposal scaling as a function of the dimension, as well as a discussion about measures of efficiency for Metropolis-Hastings algorithms. The goal of the third chapter is to present the main optimal scaling results, along with some examples. In particular, we first consider the case where the AOAR obtained is the usual 0.234, and we then focus on the case where this rate is not asymptotically optimal anymore. Some extensions and special cases are discussed in Chapter 4; inhomogeneous proposal distributions, among others, shall reveal essential in certain situations. Chapter 5 aims to prove the various theorems presented in the preceding chapters, which is achieved by resorting to lemmas proved in Chapter 6. For the proofs of the theorems presented in this work to be complete, we however require results about weak convergence of stochastic processes which can be found in [16]; Chapter 7 is devoted to the study of these results.

Chapter 1

Metropolis-Hastings Algorithms and Optimal Scaling

The aim of this chapter is to introduce the Metropolis-Hastings algorithm. In particular, we present the idea behind this method, how it is implemented, and the required conditions in order to ensure its convergence to the right distribution. We also familiarize ourselves with various types of Metropolis-Hastings algorithms, among which the random walk Metropolis and Metropolis-adjusted Langevin algorithms. We conclude this chapter by presenting the optimal scaling results obtained by [29] and [30]. This shall provide the tools necessary to the introduction of the main results in the following chapters.

1.1 Construction of the M-H Algorithm

Metropolis-Hastings algorithms are an important class of MCMC algorithms. They provide a way to sample from complex distributions by generating a Markov chain

$X_0, X_1, \ldots$ whose stationary distribution is the distribution we are interested in, referred to as the target distribution π_D. These algorithms allow for a lot of flexibility, as they can be used with basically any probability distribution. Their implementation is characterized by the necessity of choosing a proposal distribution.

Specifically, let π be the density of the target distribution with respect to some reference measure μ. We also denote the state space by $\mathcal{X}$, with a σ-algebra $\mathcal{F}$ of measurable subsets. Note that in what follows, a statement such as $\pi_1(dx) = \pi_2(dx)$ for instance really means $\int_{x\in A} \pi_1(dx) = \int_{x\in A} \pi_2(dx)$ for all $A \in \mathcal{F}$. To avoid trivial cases, we assume that π is not a point-mass function. We want to design a Markov chain having transition probabilities $P(x, dy)$ for $x, y \in \mathcal{X}$ such that

$$\int_{x\in\mathcal{X}} \pi(x) P(x, dy) \mu(dx) = \pi(y) \mu(dy),$$

i.e. where π is stationary for the chain. This can be achieved by making use of the reversibility condition with respect to π. That is, by designing transition probabilities satisfying

$$\pi(x) P(x, dy) \mu(dx) = \pi(y) P(y, dx) \mu(dy) \tag{1.1}$$

for $x, y \in \mathcal{X}$, it is easily seen that π is stationary for the chain since

$$\int_{x\in\mathcal{X}} \pi(x) P(x, dy) \mu(dx) = \pi(y) \mu(dy) \int_{x\in\mathcal{X}} P(y, dx) = \pi(y) \mu(dy).$$

In order to construct the chain, a proposal distribution and an acceptance function are required. We then define a proposal chain on $\mathcal{X}$ with transition

probabilities $Q(x, dy)$ for $x, y \in \mathcal{X}$. We also assume that the transitions are distributed according to a density q with respect to the same measure μ as before, i.e. $Q(x, dy) = q(x, y)\,\mu(dy)$. We are now left to define the acceptance function $\alpha(x, y)$. The idea is to make the chain move through the state space by first proposing a new state using the proposal distribution, and then using the acceptance function to determine if this proposed move is accepted or not. In the case where the proposed move is rejected, the chain just remains at the current state. The transition probabilities $P(x, dy)$ are thus given the form

$$\begin{aligned} P(x, dy) &= (1 - \delta_x(y))\, q(x, y)\, \alpha(x, y)\, \mu(dy) \\ &\quad + \delta_x(y) \left(1 - \int_{y \in \mathcal{X}} q(x, y)\, \alpha(x, y)\, \mu(dy)\right), \end{aligned}$$

where δ_x is the point-mass function at x. The acceptance function $\alpha(x, y)$ being the only undefined object, we can now develop restrictions on it that will guarantee the reversibility of the chain with respect to π. For $x \neq y$ (the case $x = y$ is trivial), we have

$$\begin{aligned} \pi(x)\, q(x, y)\, \alpha(x, y)\, \mu(dx)\, \mu(dy) &= \pi(y)\, q(y, x)\, \alpha(y, x)\, \mu(dy)\, \mu(dx) \\ \Rightarrow \qquad \alpha(y, x) &= \alpha(x, y) \left(\frac{\pi(x)\, q(x, y)}{\pi(y)\, q(y, x)}\right). \end{aligned}$$

Adding $\alpha(x, y)$ on both sides of the equation, we get

$$\alpha(y, x) + \alpha(x, y) = \alpha(x, y) \left(1 + \frac{\pi(x)\, q(x, y)}{\pi(y)\, q(y, x)}\right).$$

Finally, we let $s(x,y) = \alpha(y,x) + \alpha(x,y)$. This yields

$$\alpha(x,y) = \frac{s(x,y)}{\left(1 + \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}\right)} \equiv \frac{s(x,y)}{(1+t(x,y))}, \tag{1.2}$$

where $t(x,y)$ is implicitly defined and $s(x,y)$ is any symmetric function of x and y chosen such that $0 \leq \alpha(x,y) \leq 1$ for all $x, y \in \mathcal{X}$. This thus implies that $0 \leq s(x,y) \leq 1 + t(x,y)$ for all $x, y \in \mathcal{X}$.

Two popular choices for the function $s(x,y)$ are given by

$$s^{(M)}(x,y) = \begin{cases} 1 + t(x,y), & \text{if } t(y,x) \geq 1 \\ 1 + t(y,x), & \text{if } t(y,x) \leq 1 \end{cases}$$

and

$$s^{(B)}(x,y) = 1.$$

If in addition the proposal density is symmetric, i.e. $q(x,y) = q(y,x)$, then using $s^{(M)}(x,y)$ yields the Metropolis algorithm developed in [24], and $s^{(B)}(x,y)$ yields Barker's method (see [3]).

For a given proposal distribution, [28] investigated on the relative merits of different choices for the function $s(x,y)$. It was shown that $s^{(M)}(x,y)$ is the optimal candidate for $s(x,y)$ (in terms of asymptotic variance of estimated quantities), as it results in a function $\alpha(x,y)$ accepting suitable steps more often than other forms, and thus encouraging a better sampling of the states. From now on, we then focus on the Metropolis-Hastings algorithm with $s^{(M)}(x,y)$, and we shall refer to it as the Metropolis algorithm.

1.2 Implementation and Convergence of Metropolis Algorithms

For a given proposal distribution (not necessarily symmetric), the acceptance function of the Metropolis algorithm is given by

$$\alpha(x,y) = \begin{cases} \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right), & \text{if } \pi(x)\,q(x,y) > 0 \\ 1, & \text{if } \pi(x)\,q(x,y) = 0 \end{cases}. \tag{1.3}$$

An advantage of this algorithm is that it depends on the target density only through the ratio $\pi(x)/\pi(y)$. Hence, one only needs to know the target density up to a normalizing constant in order to perform the Metropolis algorithm. Moreover, in the case where the target density at the proposed move is null, i.e. $\pi(y) = 0$, then $\alpha(x,y) = 0$ provided that $\pi(x)\,q(x,y) > 0$. The proposed state is then rejected and this means that the chain almost surely remains in $\mathcal{X}^+ = \{x : \pi(x) > 0\}$ once it is entered.

To perform the Metropolis algorithm, we begin by choosing some starting value $X(0)$. Then given $X(t)$, the state of the chain at time t, a value $Y(t+1)$ is generated from the proposal distribution $Q(X(t), \cdot)$. The probability of accepting the proposed value $Y(t+1)$ as the new value for the chain is $\alpha(X(t), Y(t+1))$, where $\alpha(x,y)$ is defined as in (1.3). If the proposed move is accepted, the chain jumps to $X(t+1) = Y(t+1)$; otherwise, it stays where it is and $X(t+1) = X(t)$. Replacing t by $t+1$, the procedure is repeated until the chain has converged to its stationary distribution. Once the chain has reached stationarity, one way of getting a sample is to record the state of the chain and repeat the whole procedure

say N times in order to get a sample of size N; alternatively, it is possible to run only one big chain and record values after the burn-in period. In the first method, only the last point of each run can be used, since we sample as soon as we are confident enough that the chain has reached stationarity. With long runs, all the data obtained after the burn-in period can be exploited (to possibly obtain more than one sample). The second method thus makes a more efficient usage of the data.

The stationarity of the target distribution for the Markov chain is not sufficient to guarantee the success of the algorithm. We also need the t-step transition probability of the Markov chain,

$$P^t(x, A) = \mathrm{P}(X(t) \in A \,|\, X(0) = x),$$

to converge to $\pi_D(A) = \int_{x \in A} \pi(x)\, \mu(dx)$ for all measurable $A \subseteq \mathcal{X}$ as $t \to \infty$. It is a well-known result that this requirement is satisfied when the Markov chain is ϕ-irreducible and aperiodic. The following theorem and definitions can be found in [36].

Theorem 1.2.1. *If a discrete-time Markov chain on a general state space is ϕ-irreducible and aperiodic, and furthermore has a stationary distribution $\pi_D(\cdot)$, then for π_D-almost every $x \in \mathcal{X}$, we have that*

$$\lim_{t \to \infty} \sup_{A \in \mathcal{F}} \left| P^t(x, A) - \pi_D(A) \right| = 0.$$

For a proof of this result, see [36].

A ϕ-irreducible Markov chain is such that all sets $A \subseteq \mathcal{X}$ with $\phi(A) > 0$ are eventually reachable from any point of the state space.

Definition 1.2.2. *A chain is ϕ-irreducible if there exists a non-zero σ-finite measure ϕ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer $t = t(x, A)$ such that $P^t(x, A) > 0$.*

The aperiodicity condition ensures that the chain does not cycle through some subsets of the state space.

Definition 1.2.3. *A Markov chain is aperiodic if there do not exist $d \geq 2$ disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_d \subseteq \mathcal{X}$ with $\pi_D(\mathcal{X}_i) > 0$, such that $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$, $1 \leq i \leq d-1$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$.*

Therefore, when the Markov chain is both ϕ-irreducible and aperiodic, then $P^t(x, \cdot)$ is close to $\pi_D(\cdot)$ for large t, and it becomes possible to successfully sample from the chain. It is necessary to assess ϕ-irreducibility and aperiodicity on an individual basis for every designed transition law. An essential condition for the Markov chain P to be irreducible is that the proposal chain Q be irreducible as well. Indeed, a Q that is not ϕ-irreducible means that a part of the state space is inaccessible from some starting value and it is therefore impossible to propose moves in this region of the state space. However, since P depends on both Q and π, the ϕ-irreducibility of Q is not sufficient to guarantee the ϕ-irreducibility of P. Nonetheless, the ϕ-irreducibility condition is usually easily verified by choosing ϕ to be the Lebesgue measure on the appropriate region of the state space.

Once we have assessed the ϕ-irreducibility of P, it is generally easy to verify the aperiodicity condition. Let $A = \left\{x : 1 - \int_{y \in \mathcal{X}} q(x, y)\, \alpha(x, y)\, \mu(dy) > 0\right\}$ be

the set of states at which the probability of rejecting the proposed move is positive. If $\phi(A) > 0$ and P is ϕ-irreducible, then the set A will be reachable from any starting value in $\mathcal{X}$. There thus exists a positive probability that the chain stays at the same state two consecutive times, since $P(x, x) > 0$ for all $x \in A$, conferring aperiodicity. As a result, periodic MCMC algorithms are rarely encountered and the aperiodicity condition is verified for virtually any MCMC algorithm.

1.3 Various Types of Metropolis Algorithms

Traditionally, Metropolis algorithms are divided into different classes according to their proposal density. To be easily implemented on a computer, it is crucial to select a proposal distribution from which values can be easily generated. Of course, the closer to the target density is the proposal density, the more efficient is the algorithm. Nonetheless, it is important to keep in mind that MCMC methods are used for sampling from highly complicated distributions, from where the importance of choosing a relatively simple proposal distribution.

Infinitely many distributions can act as proposal distributions; we however restrict our attention to four important classes.

- Symmetric Metropolis algorithms: This category contains the algorithms whose proposal density is symmetric in terms of x and y, i.e. $q(x, y) = q(y, x)$. In this case, the acceptance probability simplifies to

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

The probability of accepting a proposed move thus depends on how small is the

target density at this point compared to the current state of the chain. In the case where the target density at the proposed move is larger, then the chain automatically accepts this new state.

- Random walk Metropolis (RWM) algorithms: The proposal chains included in this class take the form $Y(t+1) = X(t) + Z(t)$, where the random variables $Z(t)$ are independent with common density q. The proposal density thus satisfies $q(x, y) = q(y - x)$. Popular examples of proposal distributions for random walk Metropolis algorithms are uniform, normal or Student distributions centered at x.

- Independence Metropolis algorithms: When a proposal density does not depend on the current state of the chain x, the corresponding chain is described as an independence chain. The proposal density then satisfies $q(x, y) = q(y)$ and the acceptance probability can be expressed as

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)\, q(x)}{\pi(x)\, q(y)}\right) = \min\left(1, \frac{w(y)}{w(x)}\right),$$

where $w(x) = \pi(x)/q(x)$. The function $w(x)$ is identical to the weight function of the importance sampling method (see [37], for example), and the idea behind the algorithm itself turns out to be similar to that method.

Since x is generated from the density q, $\pi(x)$ is generally expected to be small compared to $q(x)$. However, since

$$\int_{x \in \mathcal{X}} \frac{\pi(x)}{q(x)}\, q(x)\, \mu(dx) = 1,$$

then $\pi(x)/q(x)$ equals 1 on average, which is an indication that the ratio some-

times bears large values. The chain thus reproduces the target density by remaining at values with high weights for long periods of time, rising the amount of probability sitting at these points.

- Metropolis-adjusted Langevin algorithms (MALA) : The rationale behind this refined algorithm is to use a discrete approximation to a Langevin diffusion process to create an algorithm proposing wiser moves. The proposal distribution is such that

$$Y(t+1) \sim N\left(X(t) + \frac{\delta^2}{2}\nabla \log \pi(X(t)), \delta^2\right),$$

for some small $\delta^2 > 0$. This algorithm thus attempts to optimize its convergence speed by making use of the gradient $\nabla \log \pi(X(t))$ to push the proposal chain towards higher values of π.

This dissertation focuses on symmetric RWM algorithms, where the proposed moves are normally distributed and centered around the current state of the chain: $Y(t+1) \sim N(X(t), \sigma^2)$. This can easily be extended to multidimensional target distribution settings; in d dimensions for instance, the moves are proposed independently from each other and $\mathbf{Y}^{(d)}(t+1) \sim N\left(\mathbf{X}^{(d)}(t), \sigma^2 I_d\right)$, where I_d is the $d \times d$ identity matrix.

Besides the diversity of problems efficiently solved by the RWM algorithm with Gaussian proposal distribution, the main advantage of this algorithm is the easiness with which values can be generated from its proposal distribution. Another appealing feature is the minimal amount of required information inherent to its implementation, the only essential quantity being the unnormalized target density. In comparison, the Metropolis-adjusted Langevin algorithm (MALA) attempts to

propose smarter moves, but the price to pay is the computation of the gradient $\nabla \log \pi (X(t))$. However, depending on the complexity of the target density, the implementation of this technique can be almost as simple due to the Gaussian form of the proposal distribution.

Among the similarities they share, the RWM algorithm and the MALA both possess only one arbitrary parameter, the scaling of the proposal distribution (σ^2 for the RWM algorithm, δ^2 for MALA). This parameter turns out to play a crucial role in the efficiency of the method, from where the importance of choosing it judiciously. In the case where the scaling parameter is excessively small, the proposed jumps will be too short and therefore simulation will move very slowly to the target distribution in spite of the fact that the proposed moves will be almost all accepted. At the opposite, large scaling values will generate jumps that are far away from the current state of the chain, often in low target density regions. This will result in the rejection of the proposed moves and in a chain that stands still most of the time, compromising the mixing of the states. In order to have some level of performance in the convergence of the algorithm, it then becomes necessary to find the middle ground between these two extremes. This is what we attempt to do in the following chapters. In particular, next section aims to introduce the existing results related to the choice of an optimal scaling value for the proposal distribution.

1.4 Optimal Scaling for *iid* Target Distributions

A number of rules of thumb have been proposed by different authors to handle the proposal scaling issue (among others, see [8], [9]). However, the first theoretical

results have been introduced by [29], who studied the special case of d-dimensional target densities formed of *iid* components, to which they applied the RWM algorithm with Gaussian proposal density. These results shall be introduced in the present section.

First, let the target density

$$\pi\left(\mathbf{x}^{(d)}\right) = \prod_{i=1}^{d} f\left(x_i\right) \tag{1.4}$$

be a d-dimensional product density with respect to Lebesgue measure. The function f is taken to be positive, continuous over $\mathbf{R}$ and is also assumed to belong to C^2, the space of real-valued functions with continuous second derivative. The proposed moves are distributed as $\mathbf{Y}^{(d)}\left(t+1\right) \sim N\left(\mathbf{X}^{(d)}\left(t\right), \sigma^2\left(d\right) I_d\right)$, and then the proposal density can be expressed as

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

The acceptance probability for the proposed moves is

$$\alpha\left(\mathbf{X}^{(d)}\left(t\right), \mathbf{Y}^{(d)}\left(t+1\right)\right) = 1 \wedge \frac{\pi\left(\mathbf{Y}^{(d)}\left(t+1\right)\right)}{\pi\left(\mathbf{X}^{(d)}\left(t\right)\right)},$$

where $\wedge$ denotes the minimum function. Note that in a multidimensional setting, the acceptance probability takes d-dimensional vectors as arguments but returns a scalar. This implies that at a given time, the algorithm either accepts the proposed moves for all d components or rejects all of them. As d increases, the number of proposed moves in a given step obviously increases. Since the moves are proposed

independently for each component, then as d grows and for a given proposal scaling σ^2 the algorithm becomes more likely to propose an unreasonable move for one of the components. Because of the product form of the target density, a single unreasonable move is sufficient to reduce the acceptance probability significantly, causing the rejection of the move and eventually resulting in a chain that remains stuck at some states for long periods of time.

To remedy to this situation, it is sensible to define the proposal variance to be a decreasing function of the dimension. Indeed, if $\sigma^2(d)$ becomes smaller for large values of d, then it reduces the probability of foolish moves in any one dimension and hence improves the convergence speed of the Markov chain. To this end, let $\sigma^2(d) = \ell^2/d$, where ℓ is some positive constant. It turns out that this form is optimal for the proposal variance since it is the only one yielding a nontrivial limit. If the variance is smaller than $O(1/d)$ then the Markov chain explores the state space very slowly while for larger order scalings, the acceptance rate converges to 0 too quickly. The form of the variance being determined, the objective has now evolved in optimizing the choice of the constant ℓ.

Before introducing the optimal scaling results, we have to assess the convergence of the algorithm to the target distribution. It is then necessary to verify that the Markov chain generated is indeed ϕ-irreducible and aperiodic, which is now possible given the information about the target and proposal distributions. In fact, due to the simplicity of the proposal distribution, both conditions are easily verified, implying that the chain will exhibit the desired limiting behavior. We

have, for all $\mathbf{x}^{(d)} \in \mathbf{R}$,

$$P\left(\mathbf{x}^{(d)}, A\right) \geq \int_{\mathbf{y}^{(d)} \in A} q\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) \left(1 \wedge \frac{\pi\left(\mathbf{y}^{(d)}\right)}{\pi\left(\mathbf{x}^{(d)}\right)}\right) d\mathbf{y}^{(d)},$$

where equality holds if $\mathbf{x}^{(d)} \notin A$. Since $q\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) > 0$ for all $\mathbf{x}^{(d)}, \mathbf{y}^{(d)} \in \mathbf{R}^d$, then for all sets A satisfying $\pi_D(A) = \int_{\mathbf{y}^{(d)} \in A} \pi\left(\mathbf{y}^{(d)}\right) d\mathbf{y}^{(d)} > 0$ we have $P\left(\mathbf{x}^{(d)}, A\right) > 0$. The Markov chain is thus π_D-irreducible, since it can reach any set A of positive π_D-probability in a single step.

To prove aperiodicity, it suffices to demonstrate that the Markov chain has a positive probability of remaining at the same state for two consecutive times. Following the discussion of Section 1.2, we want to show that

$$\pi_D\left(\left\{\mathbf{x}^{(d)} : 1 - \int_{\mathbf{y}^{(d)} \in \mathbf{R}^d} q\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) \alpha\left(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) d\mathbf{y}^{(d)} > 0\right\}\right) > 0.$$

Define $M = \left\{\mathbf{x}^{(d)} : \pi\left(\mathbf{x}^{(d)}\right) \leq \pi\left(\mathbf{y}^{(d)}\right), \ \forall \mathbf{y}^{(d)} \in \mathbf{R}^d\right\}$ to be the set containing the values at which the target density reaches its minimum. By the continuity of π on $\mathbf{R}^d$ and since $\pi_D\left(\mathbf{R}^d\right) = 1$, then it must be true that $\pi_D(M) < 1$. For $\mathbf{x}^{(d)} \in M$, the ratio $\pi\left(\mathbf{y}^{(d)}\right)/\pi\left(\mathbf{x}^{(d)}\right)$ is greater or equal to 1 for all $\mathbf{y}^{(d)} \in \mathbf{R}^d$ so $\alpha\left(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) = 1$ and hence $1 - \int_{\mathbf{y}^{(d)} \in \mathbf{R}^d} q\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) \alpha\left(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) d\mathbf{y}^{(d)} = 0$, meaning that when the chain is at a state in M, it will almost surely leave the next period. If $\mathbf{x}^{(d)} \notin M$, then by the continuity of π there is at least a small interval for which the ratio $\pi\left(\mathbf{y}^{(d)}\right)/\pi\left(\mathbf{x}^{(d)}\right)$ is smaller than 1. This implies that $1 - \int_{\mathbf{y}^{(d)} \in \mathbf{R}^d} q\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) \alpha\left(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) d\mathbf{y}^{(d)} > 0$ and since $\pi_D\left(M^c\right) = 1 - \pi_D(M) > 0$, we conclude that the Markov chain is aperiodic.

By its nature, the RWM algorithm is a discrete-time process. Since space

(the proposal scaling) is a function of the dimension of the target distribution, we also have to rescale the time between each step in order to get a nontrivial limiting process as $d \to \infty$. We can make a parallel between our case and Brownian motion expressed as the limit of a simple symmetric random walk. Since we rescaled space through the factor $d^{-1/2}$ (the proposal standard deviation), we have to compensate by speeding up time by a factor of d.

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

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

where $[\cdot]$ denotes the integer part function. That is, instead of proposing only one move, the sped up process has the possibility to move d times during each time interval. In other words, the algorithm proposes a move every $1/d$ time unit and consequently, the process $\left\{\mathbf{Z}^{(d)}(t)\right\}$ is asymptotically continuous as $d \to \infty$. We shall now study the limiting behavior of its first component, i.e. the limit of the sequence of processes $\left\{Z_1^{(d)}(t), t \geq 0\right\}$ as the dimension increases. Note that due to the *iid* assumption on the target components, we would obtain equivalent results should we choose to study a different target component.

In order to state the theorem, we need to introduce further conditions on the density f in (1.4). In addition to be a C^2 density function on $\mathbf{R}^d$, $(\log f(x))'$ must also be Lipschitz continuous (see (A.1)). Moreover, the following two moment

conditions must be verified:

$$\mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^8\right] = \int \left(\frac{f'(x)}{f(x)}\right)^8 f(x)\,dx < \infty$$

and similarly $\mathrm{E}\left[\left(\frac{f''(X)}{f(X)}\right)^4\right] < \infty$.

We denote weak convergence of processes in the Skorokhod topology by $\Rightarrow$, standard Brownian motion at time t by $B(t)$, and the standard normal cumulative distribution function (*cdf*) by $\Phi(\cdot)$.

Theorem 1.4.1. *Consider a random walk Metropolis algorithm with proposal distribution* $\mathbf{Y}^{(d)} \sim N\left(\mathbf{x}^{(d)}, \frac{\ell^2}{d} I_d\right)$ *and applied to a target density as in (1.4). Consider the process* $\left\{Z_1^{(d)}(t), t \geq 0\right\}$ *and let* $\mathbf{X}^{(d)}(0)$ *be distributed according to the target density* π *in (1.4). We have*

$$\left\{Z_1^{(d)}(t), t \geq 0\right\} \Rightarrow \{Z(t), t \geq 0\},$$

where $Z(0)$ *is distributed according to* f *and* $\{Z(t), t \geq 0\}$ *satisfies the Langevin stochastic differential equation (SDE)*

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

Here, $\upsilon(\ell) = 2\ell^2\Phi\left(-\ell\sqrt{I}/2\right)$ *and* $I = \mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]$.

This results says that as $d \to \infty$, the path of $\left\{Z_1^{(d)}(t), t \geq 0\right\}$ behaves according to a Langevin diffusion process. It is interesting to note that the stationary distribution of this diffusion process has density f.

Here, $\upsilon(\ell)$ is sometimes interpreted as the speed measure of the diffusion process. This means that the limiting process can be expressed as a sped up version of $\{U(t), t \geq 0\}$, a Langevin diffusion process with unity speed measure:

$$\{Z(t), t \geq 0\} = \{U(\upsilon(\ell)\, t), t \geq 0\},$$

where $dU(t) = dB(t) + \frac{1}{2}(\log f(U(t)))' dt$.

In fact, letting $s = \upsilon(\ell)\, t$ gives $ds = \upsilon(\ell)\, dt$ and thus

$$\begin{aligned} dU(s) &= (ds)^{1/2} + \frac{1}{2}\frac{d}{dU(s)} \log f(U(s))\, ds \\ &= (\upsilon(\ell)\, dt)^{1/2} + \frac{1}{2}\frac{d}{dU(\upsilon(\ell)\, t)} \log f(U(\upsilon(\ell)\, t))\, \upsilon(\ell)\, dt \\ &= (\upsilon(\ell))^{1/2} dB(t) + \frac{1}{2}\upsilon(\ell)\frac{d}{dZ(t)} \log f(Z(t))\, dt \\ &= dZ(t). \end{aligned}$$

The speed measure of the diffusion being proportional to the mixing rate of the algorithm, it suffices to maximize the function $\upsilon(\ell)$ in order to optimize the efficiency of the algorithm.

We now introduce a quantity which is closely related to the notion of algorithm efficiency. Let the π-average acceptance rate be defined by

$$\begin{aligned} a(d, \ell) &= \mathrm{E}\left[1 \wedge \frac{\pi\left(\mathbf{Y}^{(d)}\right)}{\pi\left(\mathbf{X}^{(d)}\right)}\right] \\ &= \int\int \pi\left(\mathbf{x}^{(d)}\right) \alpha\left(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) q\left(d, \mathbf{x}^{(d)}, \mathbf{y}^{(d)}\right) d\mathbf{x}^{(d)} d\mathbf{y}^{(d)} \end{aligned} \tag{1.5}$$

for the d-dimensional symmetric RWM algorithm. The following corollary intro-

duces the value of ℓ maximizing the speed measure, and thus the efficiency of the RWM algorithm. It also presents the asymptotically optimal acceptance rate (AOAR), which is of great use for applications.

Corollary 1.4.2. *In the setting of Theorem 1.4.1 we have* $\lim_{d\to\infty} a(d,\ell) = a(\ell)$, *where* $a(\ell) = 2\Phi\left(-\ell\sqrt{I}/2\right)$. *Furthermore,* $v(\ell)$ *is maximized at the unique value* $\hat{\ell} = 2.38/\sqrt{I}$ *for which* $a(\hat{\ell}) = 0.234$ *(to three decimal places).*

It is possible to give a simple interpretation to these results. Consider a high-dimensional target distribution as in (1.4) to which is applied the symmetric RWM algorithm defined at the beginning of this section. The value ℓ should then be chosen such that the acceptance rate is close to 0.234 in order to optimize the efficiency of the algorithm. If it is realized that the acceptance rate is substantially larger or smaller than 0.234, the value of $\sigma^2(d)$ should be modified accordingly. This rule is convenient and easy to apply, as the AOAR does not depend on the particular target density considered. It is also interesting to note that although asymptotic, these results work quite well in relatively low dimensions ($d \geq 10$); this is discussed in [31].

The quantity I is a measure of roughness of the density f: the smaller is I, the smoother is thus the density. Consequently, the optimal scaling value $\hat{\ell}$ is inversely proportional to I. In other words, rougher densities require shorter proposed moves.

Similar asymptotic results have been obtained by [30] for sampling from *iid* distributions when using the MALA. They first determined the optimal form of the proposal scaling to be $\sigma^2(d) = \ell^2/d^{1/3}$. Given some target density, the MALA thus tolerates larger proposed moves than the RWM algorithm. They then proved

that if the algorithm is sped up by a factor of $d^{1/3}$, each component in the sequence of processes $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\} = \left\{\mathbf{X}^{(d)}\left(\left[d^{1/3}t\right]\right), t \geq 0\right\}$ converges weakly to a Langevin diffusion process possessing some speed measure $\upsilon(\ell)$ which differs from the RWM algorithm case. Optimizing this new speed measure yields an optimal scaling value $\hat{\ell}$, which corresponds this time to an AOAR of 0.574. This acceptance rate is much higher than that obtained for the RWM algorithm, and this makes sense since the MALA uses the gradient of the log density in order to propose wiser moves. A drawback of this method is however the necessity to compute this quantity for its implementation.

Since the publication of the asymptotic results in [29], several variations of the *iid* model have been considered by different authors. Some of them relaxed the "identically distributed" assumption for the d components forming the target distribution, while others considered specific statistical models. All these researchers agreed on a same conclusion: the results in [29] show robustness to certain perturbations of the *iid* target density. That is, 0.234 seems to hold far more generally than for this type of target only. The goal of this dissertation is to study to what extent these conclusions can be generalized to more general target distributions. We consider a model where the d components, although independent, are not identically distributed. We provide a condition for the RWM algorithm to adopt the same limiting behavior as for *iid* targets. We also show that when this condition is not satisfied, the chain converges to a Langevin diffusion process with different speed measures, yielding AOARs that are smaller than the usual 0.234. These conclusions are the first instance of AOARs differing from 0.234 in the literature.

Chapter 2

Sampling from the Target Distribution

In this chapter, we describe the d-dimensional target distribution setting considered, which consists in an extension of the *iid* model where the different components are independent, but where each one of them has its own scaling term (possibly depending on d). We also introduce a method for determining the optimal form of the proposal variance as a function of the dimension when sampling from this type of target. Measures of efficiency for the algorithm shall finally be discussed but in order to clarify the purpose of the following sections, we begin with an example.

2.1 A Hierarchical Model

We present two examples of distributions satisfying the assumed target model; the first distribution is formed of independent components, while the second one admits a nontrivial correlation structure. Although MCMC methods would not

be necessary for sampling from these particular targets, the aim is to illustrate the importance and the stakes of the chosen target model for statisticians through simple examples.

Consider the case where we wish to sample from the target density

$$\pi\left(d, x_1, \ldots, x_d\right) \propto x_1^4 e^{-x_1}\ x_2^4 e^{-x_2} \prod_{i=3}^{d} x_i^4 e^{-x_i/5\sqrt{d}}. \tag{2.1}$$

The density π is then formed of d independent components, each having a gamma density $f\left(x\right) = \left(24\lambda^5\right)^{-1} x^4 \exp\left(-x/\lambda\right)$, $x > 0$. The parameter λ of the first two components is 1, while that of the other $d-2$ components is $5\sqrt{d}$.

Because the majority of the d scaling parameters get larger as the dimension of the target increases, the impact of selecting an inaccurate value for the proposal variance on the efficiency of the algorithm can be huge. In order to choose a proper value for this parameter, we first have to determine which form should be adopted for the proposal variance as a function of d; then, we can figure out how to optimize this function for greatest efficiency of the method.

For a relatively low-dimensional target density of this sort, say $d = 10$, the density of the last eight components is spread out over $(0, \infty)$ while that of the first two comparatively remains peaked, with their mass concentrated within a much narrower interval of the state space (see Figure 2.1). Choosing a proper variance for the proposal distribution is thus not an easy task: the last 8 components require a large proposal variance for appropriate exploration of their state space, but the selection of too large a value would result in frequent rejection of the proposed moves by the variables X_1 and X_2. A compromise between these requirements then becomes necessary. For this example, the optimal proposal variance turns

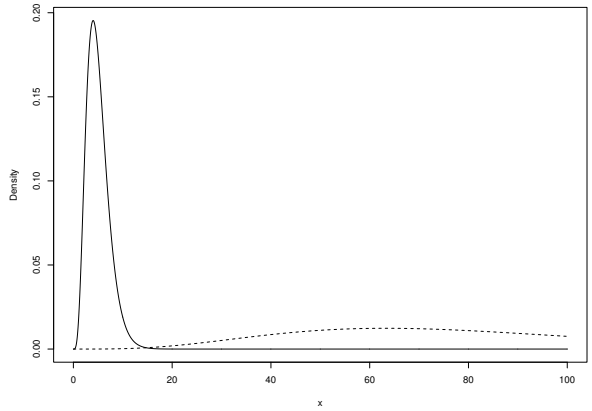

Figure 2.1: Gamma density, $d = 10$. The solid and dashed lines represent the density function for X_i, $i = 1, 2$ ($\lambda = 1$) and X_i, $i = 3, \ldots, 10$ ($\lambda = 5\sqrt{d} = 15.81$) respectively.

out to be close to $\sigma^2 = 61$ for any dimension d, as shall be seen in Example 3.2.3. Furthermore, this results in an AOAR lying around 0.098, and thus substantially smaller than the usual 0.234. In fact, tuning the algorithm to accept a proportion of 23.4% of the proposed moves would reduce the efficiency of the algorithm by about 20% in the present case, from where the importance of determining the right proposal scaling.

The target model we consider also includes, as a special case, any vector of random variables which is jointly distributed according to a multivariate normal density, no matter the correlation structure among the components.

To illustrate this, consider a normal-normal hierarchical model such that $X_1 \sim N(0, 1)$ and $X_i \sim N(X_1, 1)$, $i = 2, \ldots, d$. The joint distribution of such a target

is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix

$$\Sigma_d = \begin{pmatrix} 1 & 1 & \cdots & & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ & & \ddots & & \\ 1 & \cdots & 1 & 2 & 1 \\ 1 & & \cdots & 1 & 2 \end{pmatrix}_{d \times d}.$$

A useful property of multivariate normal distributions is their invariance under orthogonal transformations. It is therefore possible to transform the covariance matrix Σ_d into a diagonal matrix where the diagonal elements consist in the eigenvalues of Σ_d. Applying such a transformation thus results in a multivariate normal distribution with independent components.

Recall that the eigenvalues $\lambda_1, \ldots, \lambda_d$ can be found by solving the equation $|\Sigma_d - \lambda I_d| = 0$, where I_d is the d-dimensional identity matrix. This task is usually simplified by making the matrix $\Sigma_d - \lambda I_d$ triangular and then taking the product of the diagonal elements in order to obtain an expression for the determinant. By using this method, we find that $d-2$ of the eigenvalues are 1, and that the other two are the solutions of the equation

$$\frac{(d+1) \pm \sqrt{(d+1)^2 - 4}}{2} = 0.$$

These two eigenvalues can be approximated by $\frac{1}{d+1}$ and $d+1$. Our problem then reduces to the case of a multivariate normal target distribution with independent components and variances given by $\left(\frac{1}{d+1}, d+1, 1, 1, 1, \ldots,\right)$. This model agrees

with the type of target we consider and in the following chapter, we shall present how the RWM algorithm can be optimized to sample from this particular distribution. In fact, we shall obtain an AOAR of 0.216, which this time is somehow closer to 0.234 (see Example 3.2.4).

This example illustrates an interesting application of our results to multi-dimensional target distributions with correlated components. Of course, in very large dimensions, it becomes difficult to apply an orthogonal transformation and determine the eigenvalues of Σ_d. Nonetheless, our results generally work pretty well in relatively low dimensions. For these very high-dimensional models arising in practice, we are currently investigating the optimal scaling problem for general hierarchical models. Before discussing the optimal scaling results, we however start by introducing the general model for the target density.

2.2 The Target Distribution

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

$$\pi\left(d, \mathbf{x}^{(d)}\right) = \prod_{j=1}^{d} \theta_j\left(d\right) f\left(\theta_j\left(d\right) x_j\right). \tag{2.2}$$

In what follows, we shall refer to $\theta_j^{-2}\left(d\right)$, $j = 1, \ldots, d$ as the scaling terms of the target distribution.

We impose the following regularity conditions on the density f: f is a positive

C^2 function, $(\log f(X))'$ is Lipschitz continuous,

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

and similarly $\mathrm{E}\left[\left(\frac{f''(X)}{f(X)}\right)^2\right] < \infty$. These assumptions are similar, but slightly weaker than those stated in Section 1.4 for the *iid* case. The latter being a special case of the model considered here, we shall then see in Chapter 3 that Theorem 3.1.1 strengthens the results presented in [29] (although other papers already stated that the assumptions in [29] could be weakened, see for instance [31]).

The d target components, although independent, are however not necessarily identically distributed. In particular, we consider the case where the scaling terms $\theta_j^{-2}(d)$ take the following form

$$\boldsymbol{\Theta}^{-2}(d) = \left(\frac{K_1}{d^{\lambda_1}}, \ldots, \frac{K_n}{d^{\lambda_n}}, \underbrace{\frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+1}}{d^{\gamma_1}}}_{c(\mathcal{J}(1,d))}, \ldots, \underbrace{\frac{K_{n+m}}{d^{\gamma_m}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}}_{c(\mathcal{J}(m,d))}\right), \tag{2.3}$$

where $0 \leq n < \infty$, $1 \leq m < \infty$ and $\{K_j, j = 1, \ldots, n+m\}$ are some positive and finite constant terms.

Ultimately, we shall be interested in the limit of the target distribution as $d \to \infty$, and thus in the infinite-dimensional version of $\boldsymbol{\Theta}^{-2}(d)$. The first particularity to notice in (2.3) is that the first n terms appear only once each, while the balance is repeated according to some functions of d. That is, the last $d-n$ terms are separated into m different groups, in each of which the number of terms grows with the dimension. The n scaling terms appearing a finite number of times in (2.3) are denoted K_j/d^{λ_j}, $j = 1, \ldots, n$, while those appearing an infinite number

of times in the limit are denoted K_{n+i}/d^{γ_i} for $i = 1, \ldots, m$. For now, we assume the constants K_{i+n}, $i = 1, \ldots, m$, to be the same for all scaling terms within each of the m groups. This assumption shall be relaxed in Chapter 4.

It shall reveal convenient to rearrange the terms of $\boldsymbol{\Theta}^{-2}(d)$ so that all the different scaling terms appear at one of the first $n + m$ positions:

$$\begin{aligned}\boldsymbol{\Theta}^{-2}(d) \;=\; \Bigg(&\frac{K_1}{d^{\lambda_1}}, \ldots, \frac{K_n}{d^{\lambda_n}}, \frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}, \\ &\frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}, \ldots, \frac{K_{n+1}}{d^{\gamma_1}}, \ldots, \frac{K_{n+m}}{d^{\gamma_m}}\Bigg). \qquad (2.4)\end{aligned}$$

That is, we first enumerate each one of the $n + m$ different scaling terms. Afterwards, we cycle through the remaining ones, i.e. the scaling terms that will appear an infinite number of times in the limit. The m groups to which they belong might however occupy different proportions of the vector $\boldsymbol{\Theta}^{-2}(d)$, and we should make sure to preserve the proportion in which they appear when cycling through them. This method helps to clearly identify each component being studied as $d \to \infty$ without referring to a component that would otherwise be at an infinite position.

To easily refer to the various groups of components whose scaling term appears infinitely often, we define the sets

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

The i-th set thus contains positions of components having a scaling term equal to K_{i+n}/d^{γ_i}. These sets are mutually exclusive and their union satisfies $\dot{\bigcup}_{i=1}^{m} \mathcal{J}(i, d) =$

$\{n+1,\ldots,d\}$. We can then write the d-dimensional product density in (2.2) as

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

Since each of the m groups of scaling terms might occupy different proportions of $\mathbf{\Theta}^{-2}(d)$, we also define the cardinality of the sets $\mathcal{J}(i,d)$: for $i=1,\ldots,m$,

$$\begin{aligned}c\left(\mathcal{J}(i,d)\right) &= d-n-\sum_{j=1,j\neq i}^{m} c\left(\mathcal{J}(j,d)\right) \\ &= \#\left\{j\in\{1,\ldots,d\};\theta_j^{-2}(d)=\frac{K_{i+n}}{d^{\gamma_i}}\right\},\end{aligned} \tag{2.5}$$

where $c\left(\mathcal{J}(i,d)\right)$ is some polynomial function of the dimension which satisfies $\lim_{d\to\infty} c\left(\mathcal{J}(i,d)\right)=\infty$ and subject to the constraint that the total number of components in the target is d.

Without loss of generality, we assume the first n and the next m scaling terms in (2.4) to be respectively arranged according to an asymptotic increasing order. If $\preceq$ means "is asymptotically smaller than or equal to", then we have $\theta_1^{-2}(d)\preceq\ldots\preceq\theta_n^{-2}(d)$ and similarly $\theta_{n+1}^{-2}(d)\preceq\ldots\preceq\theta_{n+m}^{-2}(d)$, which respectively implies $-\infty<\lambda_n\leq\lambda_{n-1}\leq\ldots\leq\lambda_1<\infty$ and $-\infty<\gamma_m\leq\gamma_{m-1}\leq\ldots\leq\gamma_1<\infty$. We finally point out that some of the first n components might have exactly the same scaling term. When this happens, we still refer to them as say K_j/d^{λ_j} and $K_{j+1}/d^{\lambda_{j+1}}$, with $K_j=K_{j+1}$ and $\lambda_j=\lambda_{j+1}$.

According to this ordering, the asymptotically smallest scaling term $\hat{\theta}^{-2}(d)$

obviously has to be either $\theta_1^{-2}(d)$ or $\theta_{n+1}^{-2}(d)$:

$$\hat{\theta}^{-2}(d) = \begin{cases} \frac{K_1}{d^{\lambda_1}}, & \text{if } \lim_{d\to\infty} \frac{K_1/d^{\lambda_1}}{K_{n+1}/d^{\gamma_1}} = 0 \\ \frac{K_{n+1}}{d^{\gamma_1}}, & \text{if } \lim_{d\to\infty} \frac{K_1/d^{\lambda_1}}{K_{n+1}/d^{\gamma_1}} \text{ diverges} \\ \min\left(\frac{K_1}{d^{\lambda_1}}, \frac{K_{n+1}}{d^{\gamma_1}}\right), & \text{if } \lim_{d\to\infty} \frac{K_1/d^{\lambda_1}}{K_{n+1}/d^{\gamma_1}} = \frac{K_1}{K_{n+1}} \end{cases} . \quad (2.6)$$

The simple example that follows should help clarifying the notation just introduced.

Example 2.2.1. Consider a d-dimensional target density as in (2.2) with the following scaling terms: $1/\sqrt{d}$, $4/\sqrt{d}$, 10 and the other ones equally divided among $2\sqrt{d}$ and $d/2$. As the dimension increases, the last two scaling terms are replicated, implying that $n = 3$ and $m = 2$. After respectively ordering the first three and the next two scaling terms according to an asymptotically increasing order, we find

$$\boldsymbol{\Theta}^{-2}(d) = \left(\frac{1}{\sqrt{d}}, \frac{4}{\sqrt{d}}, 10, 2\sqrt{d}, \frac{d}{2}, 2\sqrt{d}, \frac{d}{2}, \ldots\right).$$

All five different scaling terms thus appear at the first five positions.

The cardinality functions for the scaling terms appearing an infinite number of times in the limit are

$$c(1,d) = \#\left\{j \in \{1,\ldots,d\}; \theta_j^{-2}(d) = 2\sqrt{d}\right\} = \left\lceil \frac{d-3}{2} \right\rceil$$

and

$$c(2,d) = \#\left\{j \in \{1,\ldots,d\}; \theta_j^{-2}(d) = \frac{d}{2}\right\} = \left[\frac{d-3}{2}\right],$$

where $\lceil\cdot\rceil$ and $[\cdot]$ denote the ceiling and the integer part functions respectively.

Note however that such rigorousness is superfluous for the application of the results presented next chapter, and it is enough to affirm that both cardinality functions grow according to $d/2$.

It is important to precise that the target model just introduced is not the most general form under which the conclusions of the theorems presented subsequently are satisfied. However, for simplicity's sake, we decided to consider a slightly more restrictive model in the first place. The results for more general cases shall be presented as extensions in Chapter 4.

Our goal is to study the limiting distribution of each component forming the d-dimensional Markov process. To this end, we set the scaling term of the target component of interest equal to 1 ($\theta_{i^*}(d) = 1$). This adjustment, necessary to obtain a nontrivial limiting process, is performed without loss of generality by applying a linear transformation to the target distribution. In particular, when the first component of the chain is studied ($i^* = 1$), we set $\theta_1^{-2}(d) = 1$ and adjust the other scaling terms accordingly. $\mathbf{\Theta}^{-2}(d)$ thus varies according to the component of interest i^* considered.

2.3 The Proposal Distribution and its Scaling

A crucial step in the implementation of RWM algorithms is the determination of the optimal form for the proposal scaling as a function of d. There exist two factors affecting this quantity: the asymptotically smallest scaling term and the fact that some scaling terms appear infinitely often in the limit. If the first factor were ignored, the proposed moves would possibly be too large for the corresponding

component, resulting in high rejection rates and compromising the convergence of the algorithm. The effect of the second factor is that as $d \to \infty$, the algorithm proposes more and more independent moves in a single step, increasing the odds of proposing an improbable move for one of the components. In this case, a drop in the acceptance rate can be overturned by letting $\sigma^2(d)$ be a decreasing function of the dimension.

Combining these two constraints, the optimal form for the variance of our proposal distribution turns out to be $\sigma^2(d) = \ell^2/d^\alpha$, where ℓ^2 is some constant and α is the smallest number satisfying

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{d^\alpha} < \infty \quad \text{and} \quad \lim_{d\to\infty} \frac{d^{\gamma_i} c\left(\mathcal{J}(i,d)\right)}{d^\alpha} < \infty, \quad i = 1, \ldots, m. \tag{2.7}$$

Therefore, at least one of these $m+1$ limits converges to some positive constant, while the other ones converge to 0. Since the scaling term of the component studied is taken to be one (i.e. the scaling term of the component of interest is independent of d), this implies that the largest possible asymptotical form for the proposal variance is $\sigma^2 = \sigma^2(d) = \ell^2$, and hence it never diverges as the dimension grows. In particular, the proposal variance will take its largest form when studying the $O\left(d^{-\lambda_1}\right)$ components, but only if the proposal scaling satisfies $\sigma^2(d) = \ell^2/d^{\lambda_1}$.

Having found the optimal form for the proposal variance, we can thus write $\mathbf{Y}^{(d)} - \mathbf{x}^{(d)} \sim N\left(\mathbf{0}, \frac{\ell^2}{d^\alpha} I_d\right)$. Our goal is now to optimize the choice of the constant ℓ appearing in the proposal variance.

Example 2.3.1. Given a target density as in (2.2) with a vector of scaling terms as in Example 2.2.1, we now determine the optimal form for the proposal variance of the RWM algorithm. Since $n > 0$ and $m = 2$, we have three limits to verify: the

first one involves the first scaling term, which is also the asymptotically smallest one in the present case

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{d^{\alpha}} = \lim_{d\to\infty} \frac{d^{1/2}}{d^{\alpha}} < \infty.$$

The smallest α satisfying the finite property is $1/2$. For the second and third limits, we have

$$\lim_{d\to\infty} \frac{d^{\gamma_1} c\,(1,d)}{d^{\alpha}} = \lim_{d\to\infty} \left(\frac{d-3}{2}\right)\left(\frac{d^{-1/2}}{d^{\alpha}}\right) < \infty$$

and

$$\lim_{d\to\infty} \frac{d^{\gamma_2} c\,(2,d)}{d^{\alpha}} = \lim_{d\to\infty} \left(\frac{d-3}{2}\right)\left(\frac{d^{-1}}{d^{\alpha}}\right) < \infty$$

The smallest α satisfying the finite property is $1/2$ for the second limit and 0 for the third one. Hence, the smallest α satisfying the constraint that all three limits be finite is $1/2$, and thus $\sigma^2\,(d) = \ell^2/\sqrt{d}$.

As mentioned in Section 1.4, RWM algorithms are discrete-time processes and thus on a microscopic level, the chain evolves according to the transition kernel outlined earlier. The proposal scaling (space) being a function of d, an appropriate rescaling of the elapsed time between each step will guarantee that we obtain a nontrivial limiting process as $d \to \infty$. This corresponds to study the model from a macroscopic viewpoint and on this level, we shall see next section that the component of interest most often behaves according to a Langevin diffusion process. The only exception to this happens with the smallest order components, and specifically when $\sigma^2\,(d) = \ell^2$. In that case, the proposal scaling is independent of d and thus a nontrivial limit is obtained without having to apply a time-speeding factor. This means that the process is already moving fast enough, and that we

can expect the limiting process to be of a discrete-time nature.

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

$$\mathbf{Z}^{(d)}(t) = \mathbf{X}^{(d)}([d^{\alpha}t]) = \left(X_1^{(d)}([d^{\alpha}t]), \ldots, X_d^{(d)}([d^{\alpha}t])\right), \tag{2.8}$$

where $[\cdot]$ is the integer part function. In reality, the process $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$ is the continuous-time, constant-interpolation version of the sped up RWM algorithm. The periods of time between each step are thus shorter and instead of proposing only one move, the sped up process proposes on average d^{α} moves during each time interval.

Note that the weak convergence results introduced in this thesis are proved in the Skorokhod topology (see Section 7.1). In this topology, we could equivalently consider a sped up RWM algorithm where the jumps, instead of occurring at regular time intervals, happen according to a Poisson process with rate d^{α}. In fact, we show in Section 5.1 that both this continuous-time version and the discrete-time sped up RWM algorithm possess the same generator. A desirable property of such a process with exponential holding times is that it preserves the time-homogeneous and Markovian attributes of the process. It is even possible to show that this setting is the only one to yield a continuous-time process preserving these properties, which can be justified by the memoryless property of the exponential distribution.

The next chapter shall be devoted to the study of $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$. That is, for each such d-dimensional process we choose a particular component $Z_{i^*}^{(d)}$ and study the limiting behavior of this sequence of processes as the dimension increases.

Even though the process $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$ can equivalently be considered as the constant-interpolation version or the exponential-holding-time version of the sped up RWM algorithm, most often the latter shall reveal more convenient.

2.4 Efficiency of the Algorithm

In order to optimize the mixing of our RWM algorithm, it would be convenient to determine criteria for measuring efficiency. We already mentioned that for diffusion processes, all measures of efficiency are equivalent to optimizing the speed measure of the diffusion. In our case however, diffusions occur as limiting processes only, and we thus still need an efficiency criterion for finite-dimensional RWM algorithms; this shall be useful for studying how well our theoretical results can be applied to finite-dimensional problems.

Recall that the basic idea for calculating the expectation of some function g with respect to the target density $\pi(\cdot)$, i.e.

$$\mu = \mathrm{E}\left[g\left(\mathbf{X}^{(d)}\right)\right] = \int g\left(\mathbf{x}^{(d)}\right)\pi\left(d, \mathbf{x}^{(d)}\right)d\mathbf{x}^{(d)},$$

is to use the generated Markov chain $\mathbf{X}^{(d)}(1), \mathbf{X}^{(d)}(2), \ldots$ to compute the sample average

$$\hat{\mu}_k = \frac{1}{k}\sum_{i=1}^{k} g\left(\mathbf{X}^{(d)}(i)\right)$$

(see for instance [25] and [31]). Just like the Central Limit Theorem for independent variables, the limiting theory for Markov chains then asserts that

$$\sqrt{k}\left(\hat{\mu}_k - \mu\right) \rightarrow_d N\left(0, \sigma^2\right),$$

provided that certain regularity conditions hold. The smaller is the variance σ^2, the more efficient is thus the algorithm for estimating the particular function $g(\cdot)$. Minimizing σ^2 would then be a good way to optimize efficiency, but the important drawback of using such a measure resides in its dependence on the function of interest $g(\cdot)$. Since we do not want to lose generality by specifying such a quantity of interest, we instead choose to base our analysis on the first order efficiency criterion, as used by [30] and [27]. This measures the average squared jumping distance for the algorithm and is defined by

$$\mathrm{E}\left[\left(X_{n+1}^{(d)}(t+1) - X_{n+1}^{(d)}(t)\right)^2\right]. \tag{2.9}$$

Note that we choose to base the first order efficiency criterion on the path of the $(n+1)$-st component of the Markov chain. Since the d components are not all identically distributed, this detail is important (although we could have chosen any of the last $d-n$ components). Indeed, as $d \to \infty$, we shall see in Chapters 3 and 4 (and prove in Chapters 5 and 6) that the path followed by any of the last $d-n$ components of an appropriately rescaled version of the RWM algorithm converges to a diffusion process with some speed measure $v(\ell)$.

For a diffusion process, the only sensible measure of efficiency is its speed measure: optimal efficiency is thus obtained by maximizing this quantity. This means that no matter the efficiency measure selected when working with our finite-dimensional RWM algorithm, it will end up being proportional to the speed measure of the limiting diffusion process as d increases. Any efficiency measure considered in finite dimensions will thus be asymptotically equivalent, including the first order efficiency criterion introduced previously. The fact that we are choosing

first order efficiency here is thus not as important as the fact that we compute it with respect to the path of a component whose limit is a continuous-time process. Indeed, in this case, the effect of choosing a particular efficiency criterion vanishes as d gets larger.

Even though the last $d - n$ terms always converge to some diffusion limit, it might not be the case for the first n components, whose limit could remain discrete as $d \to \infty$. Trying to optimize the proposal scaling by relying on these components would then result in conclusions that are specific to our choice of efficiency measure.

Chapter 3

Optimizing the Sampling Procedure

We now present weak convergence and optimal scaling results for sampling from the target distribution described in Section 2.2, using the RWM algorithm with a proposal distribution as in Section 2.3.

Since we know the results in [29] to be robust to some perturbations of the target density, we might expect these conclusions to be valid when the scaling terms in $\boldsymbol{\Theta}^{-2}(d)$ do not vary too greatly from one another. This first case is considered in Section 3.1, in which we introduce a condition involving $\boldsymbol{\Theta}^{-2}(d)$ and ensuring that the algorithm asymptotically behaves as in [29]. The last two sections focus on the asymptotic behavior of the algorithm when this condition is violated. We obtain a result stating that when there exists at least one scaling term that is significantly smaller than the others, then the limiting process and AOAR are different from those obtained for the *iid* case. Under such circumstances, we can differentiate two particular cases: the first one where the significantly small scaling terms are

also reasonably small versus the other where they are excessively small. In this last case, we shall not only see that it is impossible to optimize the efficiency of the RWM algorithm for high-dimensional distributions, but also that every proposal variance results in an ineffective algorithm. Several examples aiming to illustrate the application of the various theorems are also included.

3.1 The Familiar Asymptotic Behavior

It is now an established fact that 0.234 is the AOAR for target distributions with *iid* components, as demonstrated by [29]. [31] even showed that the *id* assumption could be relaxed to some extent, implying for instance that the same conclusion still applies in the case of a target density as in (2.2), but with scaling vector $\mathbf{\Theta}^{-2}$ independent of d. It is thus natural to wonder how big a discrepancy between the scaling terms is tolerated in order not to violate this established asymptotic behavior.

The following theorem presents explicit asymptotic results allowing us to optimize ℓ^2, the constant term of $\sigma^2(d)$. We first introduce a weak convergence result for the process $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$ in (2.8) and most importantly in practice, we transform the conclusion achieved into a statement about efficiency as a function of acceptance rate, as was done in [29].

As before, we denote weak convergence of processes in the Skorokhod topology by $\Rightarrow$, standard Brownian motion at time t by $B(t)$, and the standard normal *cdf* by $\Phi(\cdot)$. Moreover, recall that the scaling term of the component of interest X_{i^*} is taken to be one ($\theta_{i^*}(d) = 1$) which, as explained in Section 2.2, might require a linear transformation of $\mathbf{\Theta}^{-2}(d)$.

Theorem 3.1.1. *Consider a RWM algorithm with proposal distribution* $\mathbf{Y}^{(d)} \sim N\left(\mathbf{X}^{(d)}, \frac{\ell^2}{d^\alpha} I_d\right)$*, where* α *satisfies (2.7). Suppose that the algorithm is applied to a target density as in (2.2) satisfying the specified conditions on* f*, with* $\theta_j^{-2}(d)$*,* $j = 1, \ldots, d$ *as in (2.4) and* $\theta_{i^*}(d) = 1$*. Consider the* i^**-th component of the process* $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$*, that is* $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\} = \left\{X_{i^*}^{(d)}([d^\alpha t]), t \geq 0\right\}$*, and let* $\mathbf{X}^{(d)}(0)$ *be distributed according to the target density* π *in (2.2).*

We have

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

where $Z(0)$ *is distributed according to the density* f *and* $\{Z(t), t \geq 0\}$ *satisfies the Langevin stochastic differential equation (SDE)*

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

if and only if

$$\lim_{d\to\infty} \frac{\theta_1^2(d)}{\sum_{j=1}^d \theta_j^2(d)} = 0. \tag{3.1}$$

Here,

$$\upsilon(\ell) = 2\ell^2 \Phi\left(-\ell\sqrt{E_R}/2\right) \tag{3.2}$$

and

$$E_R = \lim_{d\to\infty} \sum_{i=1}^m \frac{c(\mathcal{J}(i,d))}{d^\alpha} \frac{d^{\gamma_i}}{K_{n+i}} \mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right], \tag{3.3}$$

with $c(\mathcal{J}(i,d))$ *as in (2.5).*

Intuitively, we might say that when none of the target components possesses a scaling term significantly smaller than those of the other components, the limiting process is the same as that found in [29]. Although the previous statement is one involving the asymptotically smallest scaling term, we notice that the numerator of Condition (3.1) is based on $\theta_1^{-2}(d)$ only, which is not necessarily the asymptotically smallest scaling term. Technically, Condition (3.1) should then really be

$$\lim_{d\to\infty} \frac{\hat{\theta}^2(d)}{\sum_{j=1}^{d} \theta_j^2(d)} = 0,$$

with the reciprocal of $\hat{\theta}^2(d)$ as in (2.6). Instead of explicitly verifying if the previous condition is satisfied, we can equivalently check if Condition (3.1) is still satisfied when $\theta_1^{-2}(d)$ is replaced by $\theta_{n+1}^{-2}(d)$ at the numerator. This is easily assessed given the term $c(\mathcal{J}(1,d))\,\theta_{n+1}^2(d)$ at the denominator and the previous condition is thus simplified as in (3.1).

Recall that the function $a(d,\ell)$ is the π-average acceptance rate defined in (1.5). The following corollary introduces the optimal value $\hat{\ell}$ and AOAR leading to greatest efficiency of the RWM algorithm.

Corollary 3.1.2. *In the setting of Theorem 3.1.1 we have* $\lim_{d\to\infty} a(d,\ell) = a(\ell)$, *where*

$$a(\ell) = 2\Phi\left(-\frac{\ell\sqrt{E_R}}{2}\right).$$

Furthermore, $\upsilon(\ell)$ *is maximized at the unique value* $\hat{\ell} = 2.38/\sqrt{E_R}$ *for which* $a(\hat{\ell}) = 0.234$ *(to three decimal places).*

This result provides valuable guidelines for practitioners. It reveals that when the target distribution has no scaling term that is significantly smaller than the

others (ensured by Condition (3.1)), then the asymptotic acceptance rate optimizing the efficiency of the chain is 0.234. Alternatively, setting the parameter ℓ to the value $2.38/\sqrt{E_R}$ for which $v(\ell)$ is maximized leads to greatest efficiency of the algorithm and the proportion of accepted moves is 0.234. In some situations, finding $\hat{\ell}$ will be easier while in others, tuning the algorithm according to the AOAR will reveal more convenient. In the present case, since the AOAR does not depend on the particular choice of f, it is simpler in practice to monitor the acceptance rate and to tune it to be about 0.234.

The results presented in this section can be applied, for instance, to the case where $f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, which yields a multivariate normal target distribution with independent components. In that case, note that the scaling terms in (2.4) represent the variances of the individual components. The drift and volatility terms of the limiting Langevin diffusion thus become $-Z(t)/2$ and 1 respectively, and the expression for E_R in (3.3) can be simplified since $\mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right] = 1$.

More interestingly however, Theorem 3.1.1 can also be applied to any multivariate normal distribution with covariance matrix Σ, as mentioned in Section 2.1. After having applied the orthogonal transformation to obtain a diagonal covariance matrix formed of the eigenvalues of Σ, these eigenvalues can be used to verify if Condition (3.1) is satisfied, and hence to determine whether or not $2.38/\sqrt{E_R}$ is the optimal scaling value for the proposal distribution. For example, consider a nontrivial covariance matrix Σ where the variance of each component is 2 ($\sigma_i = 2$, $i = 1, \ldots, d$) and where each covariance term is equal to 1 ($\sigma_{ij} = 1$, $j \neq i$). The d eigenvalues of Σ are $(d, 1, \ldots, 1)$ and satisfy Condition (3.1). For a relatively high-dimensional multivariate normal with such a correlation structure,

it is thus optimal to tune the acceptance rate to 0.234. Note however that not all d components mix at the same rate. When studying any of the last $d-1$ components the vector $\boldsymbol{\Theta}^{-2}(d) = (d, 1, \ldots, 1)$ is appropriate, so $\sigma^2(d) = \ell^2/d$ and these components thus mix in $O(d)$ iterations. When studying the first component, we need to linearly transform the scaling vector so that $\theta_1^{-2}(d) = 1$. We then use $\boldsymbol{\Theta}^{-2}(d) = (1, 1/d, \ldots, 1/d)$, so $\sigma^2(d) = \ell^2/d^2$ and this component mixes according to $O(d^2)$.

While the AOAR is independent of the target distribution, $\hat{\ell}$ is not and varies inversely proportionally to E_R. Recall that two different factors influence E_R: the function f itself (through the expectation term in (3.3)) and the scaling terms. The latter can have an effect through their size as a function of d, their constant term, or the proportion of the vector $\boldsymbol{\Theta}^{-2}(d)$ they occupy. Specifically, suppose that $c(i,d)\,\theta_{n+i}^2(d)$ is $O(d^\alpha)$ for some $i \in \{1, \ldots, m\}$, implying that the i-th group has an impact on the value of E_R. Then, the value $\hat{\ell}$ increases with K_{n+i} but is inversely proportional to the proportion of scaling terms included in the group. The following examples shall clarify these concepts.

The next two examples aim to illustrate the impact on $\hat{\ell}$ of choosing different functions f in (2.2) and different settings for the scaling vector $\boldsymbol{\Theta}^{-2}(d)$, the two factors influencing the quantity E_R. The third example presents a situation where the convergence of some components towards the AOAR is extremely slow.

Example 3.1.3. Consider a d-dimensional target distribution as in (2.2) with $f(x) = \exp(-x^2/2)/\sqrt{2\pi}$ and where the scaling terms are equally divided among 1 and $2d$, i.e. $\boldsymbol{\Theta}^{-2}(d) = (1, 2d, \ldots, 1, 2d)$. Referring to the notation introduced in Chapter 2 we find $n = 0$ and $m = 2$, with a proposal scaling of the form

$\sigma^2(d) = \ell^2/d$. Condition (3.1) is verified by computing

$$\lim_{d\to\infty} \frac{1}{1\left(\frac{d}{2}\right) + \frac{1}{2d}\left(\frac{d}{2}\right)} = \lim_{d\to\infty} \frac{4}{2d+1} = 0$$

(in fact the satisfaction of the condition is trivial since $n = 0$), and we can thus optimize the efficiency of the algorithm by setting the acceptance rate to be close to 0.234. Finally, since $\mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right] = 1$ then

$$E_R = \lim_{d\to\infty} \left(\frac{d}{2}(1)\frac{1}{d} + \frac{d}{2}\left(\frac{1}{2d}\right)\frac{1}{d}\right) = \frac{1}{2}$$

and the optimal value for ℓ is $\hat{\ell} = 2.38\sqrt{2} = 3.366$. What is causing an increase of $\hat{\ell}$ with respect to the baseline 2.38 for the case where all components are *iid* standard normal is the fact that only half of the components affect the accept/reject ratio in the limit. Since there are less components ruling the algorithm, a higher value of ℓ is tolerated as optimal.

The first graph in Figure 3.1 presents the relation between first order efficiency in (2.9) and the scaling parameter ℓ^2. The dotted curve has been obtained by performing 100,000 iterations of the RWM algorithm in dimensions 100, and as expected the maximum is located very close to $(3.366)^2 = 11.33$. Furthermore, the data agrees with the theoretical curve (solid line) of $v(\ell)$ in (3.2) versus ℓ^2. For the second graph, we run the RWM algorithm with various values of ℓ and plot first order efficiency as a function of the proportion of accepted moves for the different proposal variances. That is, each point in a given curve is the result of a simulation with a particular value for ℓ. We again performed 100,000 iterations, but this time we repeated the simulations for different dimensions ($d = 10, 20, 50, 100$), outlining

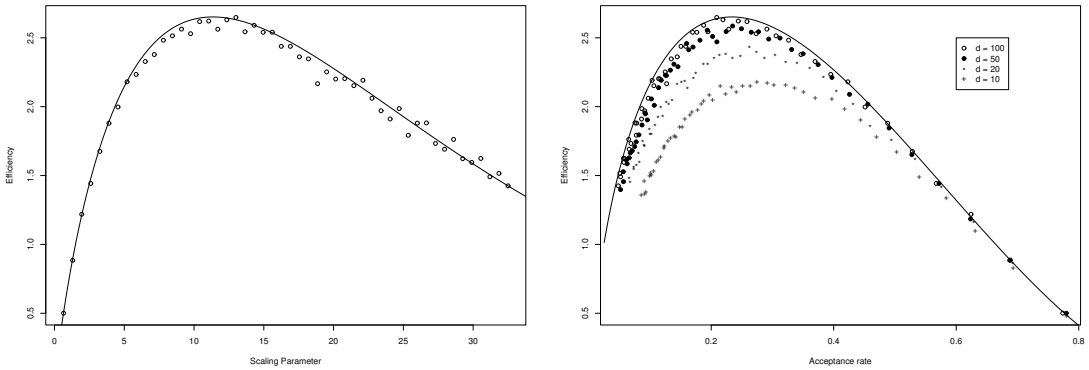

Figure 3.1: Left graph: efficiency of X_1 versus ℓ^2; the dotted line is the result of simulations with $d = 100$. Right graph: efficiency of X_1 versus the acceptance rate; the dotted lines come from simulations in dimensions 10, 20, 50 and 100. In both graphs, the solid line represents the theoretical curve $\upsilon(\ell)$.

the fact that the optimal acceptance rate converges very rapidly to its asymptotic counterpart. The theoretical curve of $\upsilon(\ell)$ versus $a(\ell)$ is represented by the solid line.

We note that efficiency is a relative measure in our case. Consequently, choosing an acceptance rate around 0.05 or 0.5, would necessitate to run the chain 1.5 times as long to obtain the same precision for a particular estimate.

Example 3.1.4. As a second example, we suppose that each of the d components in (2.2) has a gamma density with a shape parameter of 5, that is $f(x) = \frac{1}{24}x^4 \exp(-x)$ for $x > 0$. The scaling vector of the d-dimensional density is taken to be $\boldsymbol{\Theta}^{-2}(d) = \left(\frac{d}{5}, 4, d, 4, 4, d, \ldots\right)$; the first term appears only once, while the second and third ones are repeated infinitely often in the limit and appear in the proportion 2:1.

We thus have $n = 1$, $m = 2$ and $\sigma^2(d) = \ell^2/d$. Condition (3.1) is validated

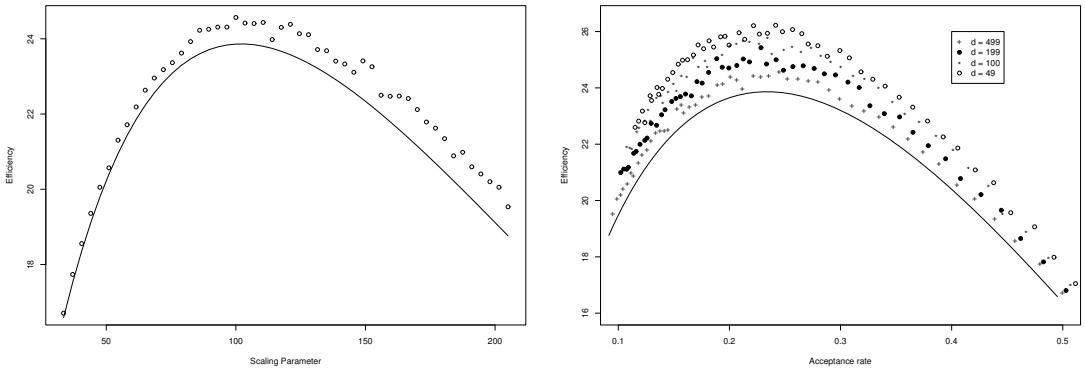

Figure 3.2: Left graph: efficiency of X_2 versus ℓ^2; the dotted curve is the result of a simulation with $d = 500$. Right graph: efficiency of X_2 versus the acceptance rate; the dotted curves come from simulations in various dimensions. In both cases, the solid line represents the theoretical curve $v(\ell)$.

by checking that

$$\lim_{d\to\infty} \frac{\frac{5}{d}}{\frac{5}{d} + \frac{2(d-1)}{3}\left(\frac{1}{4}\right) + \frac{d-1}{3}\left(\frac{1}{d}\right)} = \lim_{d\to\infty} \frac{30}{28 + d + d^2} = 0;$$

furthermore,

$$\mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right] = \mathrm{E}\left[\frac{16}{X^2} - \frac{8}{X} + 1\right] = \frac{1}{3},$$

and then

$$E_R = \lim_{d\to\infty} \frac{1}{3}\left\{\frac{2(d-1)}{3}\left(\frac{1}{4}\right)\frac{1}{d} + \frac{d-1}{3}\left(\frac{1}{d}\right)\frac{1}{d}\right\} = \frac{1}{18}.$$

In Figure 3.2, we find results similar to those of Example 3.1.3. This time we performed 500,000 iterations in dimensions 499 for the graph on the left and in dimensions 49, 100, 199, and 499 for the graph on the right. The optimal value for ℓ^2 is $\hat{\ell}^2 = \left(2.38\sqrt{18}\right)^2 = 101.96$, which the first graph corroborates. The

density f, the constant term 4 and the cardinality function $c(1,d) = 2(d-1)/3$ all contributed to boost the value of $\hat{\ell}$ (compared to a target with *iid* standard normal components). As before, the second graph allows us to verify that the optimal acceptance rate indeed converges to 0.234.

It was shown in the *iid* case that although asymptotic, the results are pretty accurate in small dimensions ($d \geq 10$). In the present case however, this fact is not always verified and care must be exercised in practice. In particular, if there exists a finite number of scaling terms such that λ_j is close to α (but with $\lambda_j < \alpha$, otherwise Condition (3.1) would be violated) then the optimal acceptance rate converges extremely slowly to 0.234 from above. For instance, suppose that $\boldsymbol{\Theta}^{-2}(d) = \left(d^{-\lambda}, 1, \ldots, 1\right)$ with $\lambda < 1$. The proposal variance is then $\sigma^2(d) = \ell^2/d$ and the closer to 1 is λ, the slower is the convergence of the optimal acceptance rate to 0.234. In fact, for a multivariate normal target with $\lambda = 0.75$, the next example shows that d must be as big as $100,000$ for the optimal acceptance rate to be reasonably close to 0.234; simulations also show that for $\alpha - \lambda \geq 0.5$, the asymptotic results are accurate in relatively small dimensions, just as in the *iid* case.

Example 3.1.5. As a last example of the conventional asymptotic behavior, consider the target in (2.2) with f the density of the standard normal distribution and $\boldsymbol{\Theta}^{-2}(d) = (d^{-0.75}, 1, 1, 1, \ldots)$ the vector of scaling terms. Under this setting, we obtain $n = m = 1$ and $\sigma^2(d) = \ell^2/d$; moreover, Condition (3.1) is verified since

$$\lim_{d\to\infty} \frac{d^{0.75}}{d^{0.75} + (d-1)} = 0.$$

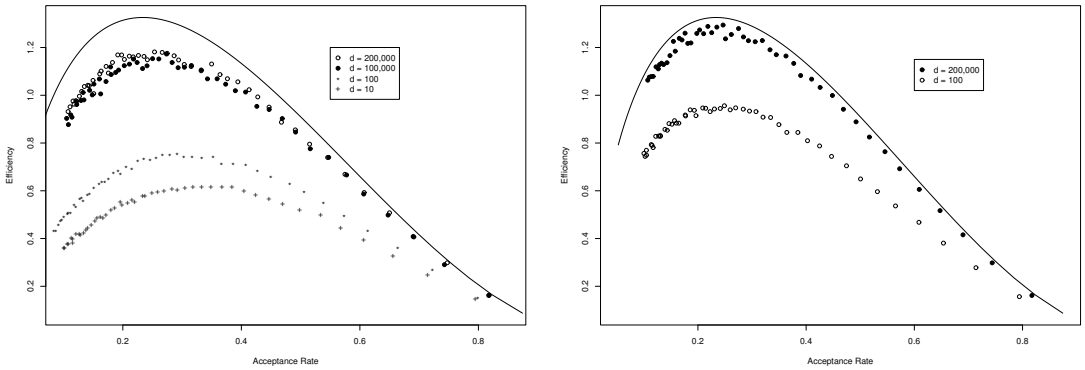

Figure 3.3: Left graph: efficiency of X_1 versus the acceptance rate; the dotted curves are the results of simulations in dimensions 10, 100, 100,000 and 200,000. Right graph: efficiency of X_2 versus the acceptance rate; the dotted curves come from simulations in dimensions 100 and 200,000. In both graphs, the solid line represents the theoretical curve $v(\ell)$.

The quantity E_R being equal to 1, the optimal value for ℓ is then the baseline 2.38.

The particularity of this case resides in the size of $\theta_1^{-2}(d)$, which is somewhat smaller than the other terms but not enough to remain significant as $d \to \infty$. As a consequence, the dimension of the target distribution must be quite large before the asymptotics kick in. In small dimensions, the optimal acceptance rate is thus closer to the case where there exists at least one scaling term significantly smaller than the others, which shall be studied in Section 3.2.

In the two previous examples, similar graphs would have been obtained no matter which component would have been selected to compute first order efficiency. In the present situation, this is still true in the limit. However, as Figure 3.3 demonstrates, the convergence of the component X_1 is much slower than that of the other components.

Even in small dimensions, the optimal acceptance rate of the last $d-1$ com-

ponents is very close to 0.234 so as not to make much difference in practice. For the first component however, setting $d = 100$ yields an optimal acceptance rate around 0.3 and dimensions must be raised as high as 100,000 to get an optimal acceptance rate reasonably close to the asymptotic one. Relying on the first order efficiency of X_1 would then falsely suggest a higher optimal acceptance rate in the present case, from where the importance of basing the efficiency on the $(n+1)$-st component as explained in Section 2.4.

Before moving to the next section, consider the normal-normal hierarchical model presented in Section 2.1. We mentioned that by applying an orthogonal transformation to this model, we obtain a target density of the form (2.2) with scaling vector $\boldsymbol{\Theta}^{-2}(d) = (O(1/d), O(d), 1, 1, \ldots)$. Such a model thus violates Condition (3.1) implying that 0.234 might not be optimal, even though the distribution is normal (see Theorem 4.3.3 of Section 4.3 when dealing with more general $\theta_j(d)$'s). This might seem surprising as multivariate normal distributions have long been believed to behave as *iid* target distributions in the limit. A natural question to ask is then, what happens when Condition (3.1) is not satisfied? Those issues shall be discussed in the next few sections.

3.2 A Reduction of the AOAR

In the presence of a finite number of scaling terms that are significantly smaller than the other ones, choosing a correct proposal variance is a slightly more delicate task. We can think for instance of the densities in Figure 2.1, which seem to promote contradictory characteristics when it comes to the selection of an efficient

proposal variance. In that example, the components X_1 and X_2 are said to rule the algorithm since despite the fact that there is only two of them, they govern the choice for the proposal variance by ensuring that it is not too big as a function of d. When dealing with such target densities, we realize that Condition (3.1) is violated and we then face the complemental case where

$$\lim_{d\to\infty} \frac{\theta_1^2(d)}{\sum_{j=1}^{d} \theta_j^2(d)} > 0. \tag{3.4}$$

According to the form of $\mathbf{\Theta}^{-2}(d)$, the asymptotically smallest scaling term in (2.4) would normally have to be either $\theta_1^{-2}(d)$ or $\theta_{n+1}^{-2}(d)$. However, it is interesting to notice that under the fulfilment of the previous condition this uncertainty is resolved and K_1/d^{λ_1} is smallest for large d. Furthermore, the existence of other target components having a $O\left(d^{\lambda_1}\right)$ scaling term is also possible. In particular, let $b = \max\left(j \in \{1,\ldots,n\}\,;\lambda_j = \lambda_1\right)$; b is then the number of such components, which is finite and at most n.

More can be said about the determination of the proposal variance. Under the fulfilment of Condition (3.4), we show in Section 5.3.1 that λ_1 has to be big enough compared to γ_1 so as to obtain $\sigma^2(d) = \ell^2/d^{\lambda_1}$. In words, this means that the proposal variance is governed by the b asymptotically smallest scaling terms. This then implies that the proposal variance takes its largest form ($\sigma^2(d) = \ell^2$) when studying one of the first b components only. This conclusion is the opposite to that achieved in the previous section, where the form of the proposal variance had to be based on one of the m groups of scaling terms appearing infinitely often in the limit (this is proved in Section 5.2.1).

We now introduce weak convergence results which shall later be used to es-

tablish an equation permitting numerically solving for the optimal ℓ^2 value.

Theorem 3.2.1. *Consider a RWM algorithm with proposal distribution* $\mathbf{Y}^{(d)} \sim N\left(\mathbf{X}^{(d)}, \frac{\ell^2}{d^{\lambda_1}} I_d\right)$. *Suppose that the algorithm is applied to a target density as in (2.2) satisfying the specified conditions on* f, *with* $\theta_j^{-2}(d)$, $j = 1, \ldots, d$ *as in (2.4) and* $\theta_{i^*}(d) = 1$. *Consider the* i^**-th component of the process* $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$, *that is* $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\} = \left\{X_{i^*}^{(d)}\left(\left[d^{\lambda_1} t\right]\right), t \geq 0\right\}$, *and let* $\mathbf{X}^{(d)}(0)$ *be distributed according to the target density* π *in (2.2).*

We have

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

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

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

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

$$\lim_{d \to \infty} \frac{c(\mathcal{J}(i,d)) d^{\gamma_i}}{d^{\lambda_1}} > 0, \tag{3.5}$$

with $c(\mathcal{J}(i,d))$ *as in (2.5).*

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

$$\begin{aligned} \alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right) &= \mathrm{E}_{\mathbf{Y}^{(b)-}, \mathbf{X}^{(b)-}}\left[\Phi\left(\frac{\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right. \\ &\quad \left. + \prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)} \Phi\left(\frac{-\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}}\right)\right]. \end{aligned} \tag{3.6}$$

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

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

where

$$\upsilon(\ell) = 2\ell^2 \mathrm{E}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}} \left[\Phi\left(\frac{\sum_{j=1}^{b} \varepsilon(X_j, Y_j) - \ell^2 E_R/2}{\sqrt{\ell^2 E_R}} \right) \right]. \tag{3.7}$$

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

$$E_R = \lim_{d\to\infty} \sum_{i=1}^{m} \frac{c(\mathcal{J}(i,d))}{d^{\lambda_1}} \frac{d^{\gamma_i}}{K_{n+i}} \mathrm{E}\left[\left(\frac{f'(X)}{f(X)} \right)^2 \right]. \tag{3.8}$$

Interestingly, the b components of smallest order each possess a discrete-time limiting process. In comparison with the other components, they already converge fast enough so a speed-up time factor is superfluous. Furthermore, the acceptance rule of the limiting Metropolis-Hastings algorithm is influenced by the components affecting the form of the proposal variance only. These components are more likely to cause the rejection of the proposed moves, and in that sense they constitute the components having the deepest impact on the rejection rate of the algorithm, ultimately becoming the only ones having an impact as $d \to \infty$. Intuitively, we know that if many components are ruling the algorithm then it will be harder to accept moves. We thus expect the probability of accepting the proposed move y_{i^*} given that we are at state x_{i^*} to get smaller as b and/or E_R get larger.

It is worth noticing the singular form of the acceptance rule, which verifies the

detailed balance condition in (1.1) and can be shown to belong to the Metropolis-Hastings family (i.e. to take the form in (1.2) for some symmetric function $s(x,y)$). In particular when $b=1$ the expectation operator can be dropped and for a general proposal density $q(x,y)$ we obtain

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

The effectiveness (in terms of asymptotic variance) of this new acceptance rule depends on the parameter $\ell^2 E_R$. When $\ell^2 E_R \to \infty$, then we find $\alpha(\ell^2 E_R, x, y) \to 0$, meaning that the chain never moves. At the other extreme, if $\ell^2 E_R = 0$ then this term has no more impact on the acceptance probability and the resulting rule is the usual one, i.e. $1 \wedge \frac{f(y)q(y,x)}{f(x)q(x,y)}$. In [28], the optimal Metropolis-Hastings acceptance rule was shown to be $1 \wedge \frac{f(y)q(y,x)}{f(x)q(x,y)}$ because it favors the mixing of the chain by improving the sampling of all possible states. The efficiency of the modified acceptance rule is thus inversely proportional to its parameter $\ell^2 E_R$.

Figure 3.4 presents the acceptance function $\alpha(\ell^2 E_R, x, y)$ of a symmetric Metropolis-Hastings algorithm (i.e. with $q(x,y) = q(y,x)$) as a function of $\frac{f(y)}{f(x)}$ for various values of the parameter $\ell^2 E_R$. We notice that as $\ell^2 E_R$ increases, it becomes more difficult to accept moves. Furthermore if $\ell^2 E_R > 0$ the fact of drawing a proposed move whose target density is higher than that of the current state does not ensure that the move will be accepted. This thus confirms that our new acceptance rule is not optimal (in terms of asymptotic variance), as proposed moves are more likely to be accepted with the usual acceptance rule. In fact, the acceptance

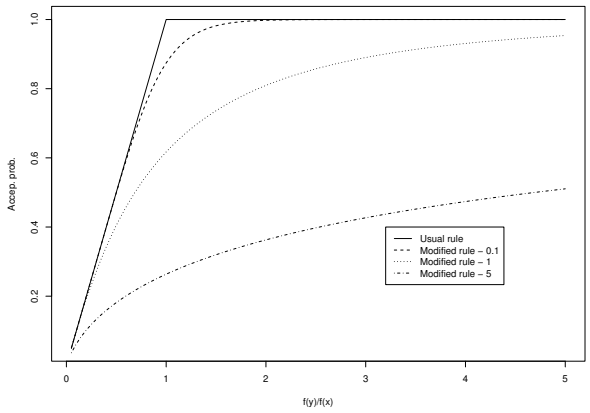

Figure 3.4: Modified acceptance rule $\alpha\left(\ell^2 E_R, x, y\right)$ as a function of the density ratio $f\left(y\right)/f\left(x\right)$ for different values of $\ell^2 E_R$ and when $b = 1$. From the top to the bottom, $\ell^2 E_R$ takes the values 0, 0.1, 1 and 5.

probability of the new rule is seen to be scaled down by the *cdf* function $\Phi\left(\cdot\right)$.

Note that the previous analysis is only valid for a proposal distribution which is independent of both ℓ^2 and E_R. In our particular case where $\mathbf{Y}^{(d)}$ is a function of ℓ, setting $\ell = 0$ is obviously not the optimal choice since this would yield a chain that is static. The optimal value for ℓ thus lies in $(0, \infty)$, but also depends on the particular measure of efficiency selected since we are dealing with a discrete-time process.

For $i^* = b + 1, \ldots, d$, the variance of the proposal distribution is a function of d and a speed-up time factor is then required in order to get a sensible limit. Consequently, we obtain a continuous-time limiting process, and the speed measure of the limiting Langevin diffusion is now different from those found in [29] and Section 3.1. It now depends on exactly the same components as for the discrete-time limit and as we shall see, this alters the value of the AOAR.

Since there are two limiting processes for the same algorithm, we now face

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

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

Corollary 3.2.2. *In the setting of Theorem 3.2.1 we have* $\lim_{d\to\infty} a(d,\ell) = a(\ell)$, *where*

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

An analytical solution for the value $\hat{\ell}$ that maximizes the function $v(\ell)$ cannot be obtained. However, this maximization problem can be easily resolved through the use of numerical methods. For densities f satisfying the regularity conditions mentioned in Section 2.2, $\hat{\ell}$ will be finite and unique. This will thus yield an AOAR $a(\hat{\ell})$ and although an explicit form is not available for this quantity either, we can still draw some conclusions about $\hat{\ell}$ and the AOAR. First, Condition (3.4) ensures the existence of a finite number of components having a scaling term significantly

smaller than the others. Since this constitutes the complement of the case treated in Section 3.1, we know that the variation in the speed measure is directly due to these components. When studying any component X_{i^*} with $i^* \in \{b+1, \ldots, d\}$, we also know that $\theta_j^{-2}(d) \to 0$ as $d \to \infty$ for $j = 1, \ldots, b$ since these scaling terms are of smaller order than $\theta_{i^*} = 1$. Hence, the first b components obviously provoke a reduction of both $\hat{\ell}$ and the AOAR, which is now necessarily smaller than 0.234. In particular, both quantities will get smaller as b increases. The AOAR is unfortunately not independent of the target distribution anymore, and will vary according to the choice of density f in (2.2) and vector $\boldsymbol{\Theta}^{-2}(d)$. It is then easier to optimize the efficiency of the algorithm by determining $\hat{\ell}$ rather than monitoring the acceptance rate, since in any case finding the AOAR implies solving for $\hat{\ell}$.

We now revisit the examples introduced in Section 2.1; this shall illustrate how to solve for the appropriate $\hat{\ell}$ and AOAR using (3.7). In the first example, tuning the acceptance rate to be about 0.234 would result in an algorithm whose performance is substantially less than when using the correct AOAR.

Example 3.2.3. Consider the d-dimensional target density mentioned in (2.1), where each component is distributed according to the gamma density $f(x) = \frac{1}{24}x^4 \exp(-x)$, $x > 0$. Consistent with the notation of Sections 2.2 and 2.3, the scaling vector is taken to be $\boldsymbol{\Theta}^{-2}(d) = (1, 1, 25d, 25d, 25d, \ldots)$, so $n = 2$, $m = 1$ and $\sigma^2(d) = \ell^2$. We remark that the first two scaling terms are significantly smaller than the balance and cause the limit in (3.4) to be positive:

$$\lim_{d\to\infty} \frac{1}{2 + (d-2)\frac{1}{25d}} = \frac{25}{51} > 0.$$

Even though the scaling parameters of X_1 and X_2 are significantly smaller than

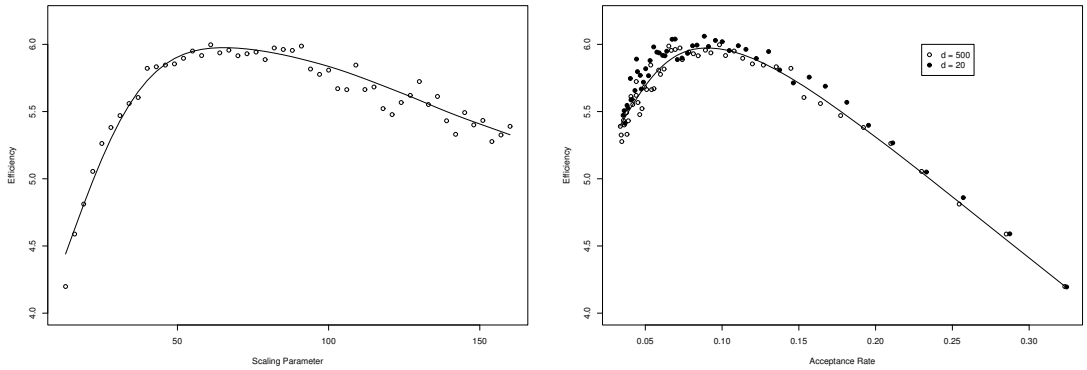

Figure 3.5: Left graph: efficiency of X_3 versus ℓ^2; the dotted curve represents the results of simulations with $d = 500$. Right graph: efficiency of X_3 versus the acceptance rate; the results of simulations with $d = 20$ and 500 are pictured by the dotted curves. In both cases, the theoretical curve $v(\ell)$ is depicted (solid line).

the others, they still share the responsibility of selecting the proposal variance with the other $d-2$ components since

$$\lim_{d\to\infty} c(1,d)\,\theta_3^2(d)\,\sigma_\alpha^2(d) = \lim_{d\to\infty}(d-2)\,\frac{1}{25d} = \frac{1}{25} > 0.$$

Since Conditions (3.4) and (3.5) are satisfied, we thus use (3.7) to optimize the efficiency of the algorithm. After having estimated the expectation term in (3.7) for various values of ℓ, a scan of the vector $v(\ell)$ produces $\hat{\ell}^2 = 61$ and $a(\hat{\ell}) = 0.0980717$. Note that the term $E_R = 1/75$ causes an increase of $\hat{\ell}$, but the components X_1 and X_2 ($b = 2$) force it downwards. This is why $\hat{\ell}^2 < 424.83$, which would be the optimal value for ℓ if X_1 and X_2 were ignored.

Figure 3.5 illustrates the result of 500,000 iterations of a Metropolis algorithm in dimensions 500 for the left graph and in dimensions 20 and 500 for the right one. On both graphs, the maximum occurs close to the theoretical values men-

tioned previously. We note that the AOAR is now quite far from 0.234, and that tuning the proposal scaling so as to produce this acceptance rate would contribute to considerably lessen the performance of the method. In particular, this would generate a drop of at least 20% in the efficiency of the algorithm.

Example 3.2.4. Let the target be the normal-normal hierarchical model considered in Section 2.1. That is, the location parameter satisfies $X_1 \sim N(0,1)$ and the other $d-1$ components are also normally distributed and conditionally independent given their mean X_1, i.e. $X_i \sim N(X_1, 1)$ for $i = 2, \ldots, d$.

Note that in order to deal with a dependent distribution, it is necessary to include the location parameter X_1 in the joint distribution and to update it as a regular component in the algorithm. Otherwise, the target distribution considered would be a $(d-1)$-dimensional *iid* distribution conditional on X_1, which has been studied by [29].

After having applied an orthogonal transformation to this target, we obtain a d-dimensional normal density with independent components and variances given by $\left(\frac{1}{d+1}, d+1, 1, 1, \ldots\right)$.

In this case, $n = 2$, $m = 1$, $\sigma^2(d) = \ell^2/d$ and Condition (3.4) is satisfied:

$$\lim_{d\to\infty} \frac{d+1}{(d+1) + \frac{1}{d+1} + (d-2)} = \frac{1}{2} > 0.$$

In addition, Condition (3.5) is also met since

$$\lim_{d\to\infty} c(1,d)\,\theta_3^2(d)\,\sigma_\alpha^2(d) = \lim_{d\to\infty} \frac{d-2}{d} = 1 > 0.$$

In light of this information, (3.7) shall then be used to optimize the efficiency of

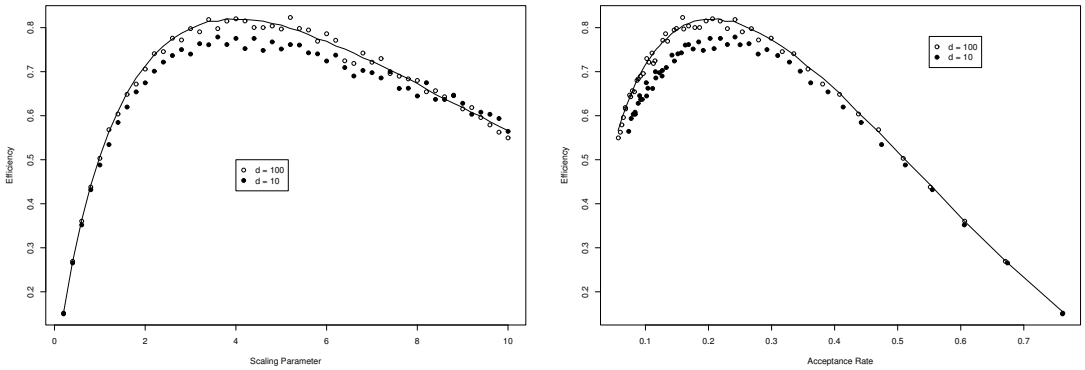

Figure 3.6: Left graph: efficiency of X_3 versus ℓ^2. Right graph: efficiency of X_3 versus the acceptance rate. In both cases, the solid line represents the theoretical curve $v(\ell)$ and the dotted curves portray the results of simulations in dimensions 10 and 100.

the Metropolis algorithm.

Using $E_R = 1$ along with the method described in Example 3.2.3, we estimate the optimal scaling value to be $\hat{\ell}^2 = 3.8$, for which $a(\hat{\ell}) = 0.2158$. The value $\hat{\ell} = 1.95$ thus differs from the baseline 2.38 in Section 3.1, but still yields an AOAR that is somewhat close to 0.234. As before, Figure 3.6 displays graphs of the first order efficiency of X_3 versus ℓ^2 and the acceptance rate respectively; 100,000 iterations were performed in dimensions 10 and 100. The curves obtained emphasize the rapid convergence of finite-dimensional distributions to their asymptotic counterpart, which is represented by the solid line.

In Examples 3.1.3 and 3.1.4, it did not matter which one of the d components was selected to compute first order efficiency, as all of them would have yielded similar efficiency curves. In Example 3.1.5, the choice of the component became important since X_1 had a scaling term much smaller than the others, resulting in a lengthy convergence to the right optimal acceptance rate. In both Examples

3.2.3 and 3.2.4, it is now crucial to choose this component judiciously since X_1 has an asymptotic distribution that remains discrete. The AOAR generated by this sole component is thus specific to the chosen measure of efficiency, which is not representative of the target distribution as a whole. As mentioned in Section 2.4, the $(n+1)$-th component (X_3 in the last two examples) is always a good choice, as is any of the last $d-n$ components.

In theory, it is necessary for the component studied to have a scaling term of order 1 to obtain a nontrivial limiting distribution for this component. In practice, this is not really an issue since we are not dealing with infinite dimensions. Nonetheless, for large dimensions, care must be exercised because extreme scaling terms could result in overflow or underflow when running the algorithm.

3.3 Excessively Small Scaling Terms: An Impasse

We finally consider the remaining situation where there exist b components having scaling terms that are excessively small compared to the others, implying that they are the only ones to have an impact on our choice for $\sigma^2(d)$. This means that if the first n components of the target density were ignored, i.e. by basing our prognostic for α on the last $d-n$ components only, we would opt for a proposal variance which is of larger order. Consequently, there does not exist a group of components with small or numerous enough scaling terms so as to have an equivalent influence on the proposal distribution. The first b components thus become the only ones to have an impact on the accept/reject ratio as the dimension of the target increases.

Theorem 3.3.1. *In the setting of Theorem 3.2.1 but with Condition (3.5) replaced by*

$$\lim_{d\to\infty} \frac{c\left(\mathcal{J}\left(i,d\right)\right) d^{\gamma_i}}{d^{\lambda_1}} = 0 \quad \forall\, i \in \{1,\ldots,m\}\,, \tag{3.9}$$

the conclusions of Theorem 3.2.1 are preserved, but the acceptance rule is now

$$\alpha\left(X_{i^*}, Y_{i^*}\right) \;=\; \mathrm{E}_{\mathbf{Y}^{(b)-},\mathbf{X}^{(b)-}}\left[1 \wedge \prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)}\right] \tag{3.10}$$

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

$$\upsilon\left(\ell\right) = 2\ell^2 \mathrm{P}_{\mathbf{Y}^{(b)},\mathbf{X}^{(b)}}\left(\sum_{j=1}^{b} \varepsilon\left(X_j, Y_j\right) > 0\right) \tag{3.11}$$

for the limiting Langevin diffusion.

As in Theorem 3.2.1 we obtain two different limiting processes, depending on which component we focus. Since the proposal variance is now entirely ruled by the first b components, it means that $E_R = 0$. When $b = 1$, the acceptance rule of the limiting RWM algorithm reduces to the usual rule. In that case, the first component not only becomes independent of the others as $d \to \infty$, but it is completely unaffected by these $d-1$ components in the limit, which move too slowly compared to the pace of X_1. For the last $d-b$ components, the limiting process is continuous and the speed measure of the diffusion is also affected by the first b components only. As mentioned previously, we use the continuous-time limit to attempt optimizing the efficiency of the chain.

Corollary 3.3.2. *In the setting of Theorem 3.3.1, we have* $\lim_{d\to\infty} a\left(d,\ell\right) = a\left(\ell\right)$,

where

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

Attempting to optimize $v(\ell)$ leads to an impasse, since this function is unbounded for basically any smooth density f. That is, $v(\ell)$ increases with ℓ, which implies that $\hat{\ell}$ must be chosen arbitrarily large; examining the function $a(\ell)$ thus leads us to the conclusion that the AOAR is null. This phenomenon can be explained by the fact that the scaling terms of the first b components are much smaller than the others, determining the form of $\sigma^2(d)$ as a function of d. However, the moves generated by a proposal distribution with such a variance are definitely too small for the other components, forcing the parameter ℓ to increase in order to generate reasonable moves for them. In practice, it is thus impossible to find a proposal variance that is small enough for the first b components, but at the same time large enough so as to generate moves that are not compromising the convergence speed of the last $d-b$ components. In Section 3.2, the situation encountered was similar, except that it was possible to achieve an equilibrium between these two constraints. In the current circumstances, the discrepancy between the scaling terms is too large and the disparities are irreconcilable. In theory, we thus obtain a well-defined limiting process, but in practice we reach a useless conclusion as far as the AOAR is concerned. In this case, we can even say that homogeneous proposal scalings will result in an algorithm which is inefficient as $d \to \infty$. We shall see next chapter that for such cases, inhomogeneous proposal scalings constitute a wiser option.

Note that if we were choosing a smaller α for the proposal variance (i.e. a function of larger order for $\sigma^2(d)$), the proposed moves would be too big for the first

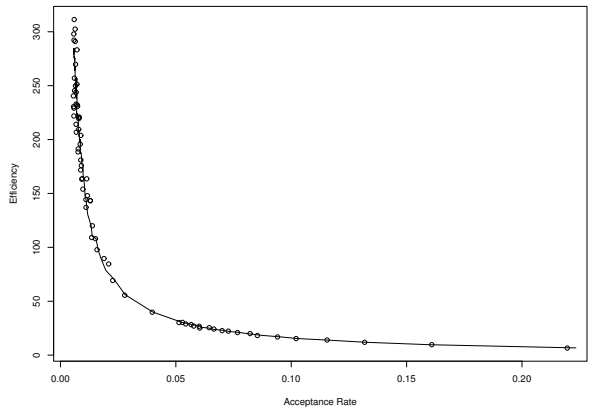

Figure 3.7: Efficiency of X_2 versus the acceptance rate. The solid line represents the theoretical curve $\upsilon(\ell)$ and the dotted line has been obtained by running a Metropolis algorithm in dimensions 101.

b components, resulting in a trivial limiting process (with a generator converging to 0). In fact, by examining the proof of Theorem 3.3.1 (and in particular Proposition A.7 as well as a similar proposition for the case where $\lambda_j > \alpha$), we realize that the proposal variance we consider is the only one to yield a nontrivial limiting process.

Example 3.3.3. Suppose f is the standard normal density and consider a target density as in (2.2) with scaling vector $\mathbf{\Theta}^{-2}(d) = \left(\frac{1}{d^5}, \frac{1}{\sqrt{d}}, 3, \frac{1}{\sqrt{d}}, 3 \ldots\right)$. The particularity of this setting resides in the fact that $\theta_1^{-2}(d)$ is much smaller than the other scaling terms. Note that an immediate consequence of this is the satisfaction of Condition (3.4).

In the present case, $n = 1$, $m = 2$ and the proposal variance is totally governed by the first component as $\sigma^2(d) = \ell^2/d^5$. Since $\theta_1^{-2}(d)$ is the only component to have an impact on the proposal variance then

$$\lim_{d\to\infty} c(1,d)\,\theta_2^2(d)\,\sigma_\alpha^2(d) = \lim_{d\to\infty}\left(\frac{d-1}{2}\right)\sqrt{d}\frac{1}{d^5} = 0$$

and

$$\lim_{d\to\infty} c\left(2,d\right)\theta_3^2\left(d\right)\sigma_\alpha^2\left(d\right) = \lim_{d\to\infty}\left(\frac{d-1}{2}\right)\left(\frac{1}{3}\right)\frac{1}{d^5} = 0,$$

implying that Condition (3.9) is also verified. We must then use (3.11) to determine how to optimize the efficiency of the algorithm.

As explained previously and as illustrated in Figure 3.7, the optimal value for ℓ converges to infinity, resulting in an optimal acceptance rate which converges to 0. Obviously, it is impossible to reach a satisfactory level of efficiency in the limit using the prescribed proposal distribution.

Chapter 4

Inhomogeneous Proposal Scalings and Target Extensions

The optimal scaling results presented in Chapter 3 assumed the homogeneity of the proposal distribution as well as a specific type of target density. The present chapter aims to relax these assumptions to some extent, and to solve the deadlock faced in Section 3.3, where the homogeneity assumption kept the algorithm from converging efficiently.

Before relaxing any assumption, we start by considering the special case where the target distribution is multivariate normal, in which case the theorems of Sections 3.2 and 3.3 can be somehow simplified. Then, in Section 4.2, we study whether or not there is an improvement in the efficiency of the RWM algorithm when applying a certain type of inhomogeneous proposal distributions. Section 4.3 focuses or relaxing the assumed form for $\boldsymbol{\Theta}^{-2}(d)$, while the goal of Section 4 is to present simulation studies to investigate the efficiency of the algorithm applied to more complicated and widely used statistical models.

4.1 Normal Target Density

The results of Sections 3.2 and 3.3 can be simplified when f is taken to be the standard normal density function. Indeed, it then becomes possible to compute the expectations with respect to $\mathbf{X}^{(b)}$ and conditional on $\mathbf{Y}^{(b)}$ in (3.6), (3.7), and (3.11). We obtain the following results.

Theorem 4.1.1. *In the setting of Theorem 3.2.1 but with* $f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, *the conclusions of Theorem 3.2.1 and Corollary 3.2.2 are preserved but with Metropolis-Hastings acceptance rule*

$$\begin{aligned} \alpha\left(\ell^2, x_{i^*}, y_{i^*}\right) &= \mathrm{E}\left[\Phi\left(\frac{\varepsilon\left(x_{i^*}, y_{i^*}\right) - \frac{\ell^2}{2}\left(\sum_{j=1, j\neq i^*}^{b} \frac{\chi_j^2}{K_j} + E_R\right)}{\sqrt{\ell^2\left(\sum_{j=1, j\neq i^*}^{b} \frac{\chi_j^2}{K_j} + E_R\right)}}\right)\right. \\ &\quad \left. + \frac{f\left(y_{i^*}\right)}{f\left(x_{i^*}\right)} \Phi\left(\frac{-\varepsilon\left(x_{i^*}, y_{i^*}\right) - \frac{\ell^2}{2}\left(\sum_{j=1, j\neq i^*}^{b} \frac{\chi_j^2}{K_j} + E_R\right)}{\sqrt{\ell^2\left(\sum_{j=1, j\neq i^*}^{b} \frac{\chi_j^2}{K_j} + E_R\right)}}\right)\right], \end{aligned} \tag{4.1}$$

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

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

In addition, the Langevin speed measure is now given by

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

and the limiting average acceptance rate satisfies

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

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

For the case where some components entirely rule the proposal variance, we find the following result.

Theorem 4.1.2. *In the setting of Theorem 3.3.1 but with $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, the conclusions of Theorem 3.3.1 and Corollary 3.3.2 are preserved but with Langevin speed measure*

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

where χ_j^2, $j = 1, \ldots, b$ are independent chi square random variables with 1 degree of freedom. Furthermore, the limiting average acceptance rate now satisfies

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

When $b = 1$, the limiting process of X_1 is the usual one-dimensional RWM algorithm. As we said before, measures of efficiency are not unique in this case but to understand the situation, suppose we consider first-order efficiency. We then want to maximize the expected square jumping distance, which will result in a better mixing of the chain. The acceptance rate maximizing this quantity is 0.45, as mentioned in [31]. As b increases, more and more components affect the acceptance process and this results in a reduction of the AOAR towards 0. Indeed,

when $b \to \infty$, Condition (3.4) is not satisfied anymore and we find ourselves facing the complemental case introduced in Section 3.1. In such a situation, the proposal scaling $\sigma^2(d) = \ell^2/d^{\lambda_1}$ is then inappropriate (too large) and in order to handle this case, a new rescaling of space and time allows us to find a nontrivial limiting process and an AOAR of 0.234. In the case of Theorem 4.1.1, the acceptance rule is more restrictive and accepting moves is thus harder. First-order efficiency is maximized when the acceptance rate is about 0.35 for $b = 1$, and decreases towards 0 as $b \to \infty$, in which case an appropriate rescaling of space and time is again required to ultimately find an AOAR of 0.234. The difference between both rules thus becomes insignificant for large values of b.

The previous analysis allows us to get some insight about the situation for discrete-time limits. Nonetheless in practice, continuous-time limits must be used to determine the AOAR that should be applied for optimal performance of the algorithm. In Theorem 4.1.1, we notice that the expectation term in the speed measure $v(\ell)$ is decreasing faster than the term $\Phi\left(-\ell\sqrt{E_R}/2\right)$ in (3.2). Consequently, the optimal value $\hat{\ell}$ is bounded above by $2.38/\sqrt{E_R}$ and gets smaller as the number b of components increases. As expected, the diminution of the parameter ℓ is not important enough to outdo the factors χ_i^2/K_i and the AOAR is thus continually smaller than 0.234. This difference is intensified with the growth of b.

The speed measure in Theorem 4.1.2 is particular in the sense that its expectation term does not vanish fast enough to overturn the growth of ℓ^2. The optimal value $\hat{\ell}$ is thus infinite, yielding an AOAR converging to 0. This means that any acceptance rate will result in an algorithm that is inefficient in practice for large d. The best solution is to resort to inhomogeneous proposal distribution, which shall

be discussed next section.

4.2 Inhomogeneous Proposal Scalings: An Alternative

So far, we have assumed the proposal scaling $\sigma^2(d) = \ell^2/d^\alpha$ to be the same for all d components. In such a case, the results obtained in Chapter 3 demonstrate that the components do not all mix at the same rate (unless we are in the *iid* case). Indeed, a particular component X_{i^*} mixes according to $O(d^\alpha)$, where α is determined by applying (2.7) to the scaling vector $\mathbf{\Theta}^{-2}(d)$ with $\theta_{i^*}(d) = 1$. It is natural to wonder if adjusting the proposal variance as a function of d for each component would yield a better performance of the algorithm. An important point to keep in mind is that for $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$ to be a stochastic process, we must speed up time by the same factor for every component. Otherwise, we would face a situation where some components move more frequently than others in the same time interval, and since the acceptance probability of the proposed moves depends on all d components this would violate the definition of a stochastic process. Despite the fact that we have to speed up all components of a given vector by the same factor, we can however use different speeding factors for studying different components (as was done in Chapter 3). That is, we might speed up all components by d^2 (say) when studying X_1, but speed up all components by d (say) when studying X_2.

The inhomogeneous scheme we adopt is the following: we personalize the proposal variances of the last $d - n$ components only, implying that the proposal variances of the first n components are the same as they would have been under

the homogeneity assumption. In order to determine the proposal variances of the last $d-n$ terms, we treat each of the m groups of scaling terms appearing infinitely often as a different portion of the scaling vector and determine the appropriate α for each group.

In particular, consider the $\theta_j(d)$'s appearing in (2.4) and let the proposal variance of X_j be $\sigma_j^2(d) = \ell^2/d^{\alpha_j}$, where $\alpha_j = \alpha$ for $j = 1, \ldots, n$ and α_j is such that $\lim_{d\to\infty} c(\mathcal{J}(i,d)) d^{\gamma_i}/d^{\alpha_j} = 1$ for $j = n+1, \ldots, d$, $j \in \mathcal{J}(i,d)$. In order to study the component X_{i^*}, we still assume that $\theta_{i^*}(d) = 1$, but we now let $\mathbf{Z}^{(d)}(t) = \mathbf{X}^{(d)}([d^{\alpha_{i^*}} t])$. We have the following result.

Theorem 4.2.1. *In the setting of Theorem 3.1.1 but with proposal variances and process* $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$ *as just described, the conclusions of Theorem 3.1.1 and Corollary 3.1.2 are preserved.*

Since the variances are now adjusted, every constant term $K_{n+1}, \ldots, K_{n+m}$ has an impact on the limiting process, yielding a larger value of E_R. Hence, the optimal value $\hat{\ell} = 2.38/\sqrt{E_R}$ is smaller than with homogeneous proposal scalings. When the proposal variances of all components were based on the same value α, the algorithm had to compensate for the fact that α is chosen as small as possible, and thus maybe too small for certain groups of components, with a larger value for $\hat{\ell}^2$. Since the variances are now personalized, a smaller value for $\hat{\ell}$ is more appropriate.

As realized previously, it is possible to face a situation where the efficiency of the algorithm cannot be optimized under homogeneous proposal scalings. This happens when a finite number of scaling terms request a proposal variance of very small order, resulting in an excessively slow convergence of the other components.

To overcome this problem, inhomogeneous proposal scalings will add a touch a personalization and ensure a decent speed of convergence in each direction.

Theorem 4.2.2. *In the settings of Theorems 3.2.1 and 3.3.1 (that is, no matter if Condition (3.5) is satisfied or not) but with proposal variances and process* $\left\{\mathbf{Z}^{(d)}(t), t \geq 0\right\}$ *as just described, the conclusions of Theorem 3.2.1 and Corollary 3.2.2 are preserved.*

In Theorem 4.2.1, it was easily verified that the AOAR is unaffected by the use of inhomogeneous proposals. The same statement does not hold in the present case, although we can still affirm that the AOAR will not be greater than 0.234. Indeed, since ℓ is assumed to be fixed in each direction, the algorithm can hardly do better than for *iid* targets even though the proposal has been personalized. As explained previously, $\hat{\ell}$ is now smaller than with homogeneous proposal scalings since the algorithm does not have to compensate for the fact that $\sigma^2(d) = \ell^2/d^{\lambda_1}$ was maybe too small for certain groups of components. In fact, in the case of Theorem 4.2.2, we expect the AOAR to lie somewhere in between the AOAR obtained under homogeneous proposal scalings and 0.234. The inhomogeneity assumption should then help us solving the problem of Section 3.3, in which case $\hat{\ell}$ was arbitrarily large and the AOAR was null.

Example 4.2.3. We now revisit Example 3.3.3. That is, we let f be the standard normal density and consider a d-dimensional target distribution as in (2.2) with scaling vector $\mathbf{\Theta}^{-2}(d) = \left(\frac{1}{d^5}, \frac{1}{\sqrt{d}}, 3, \frac{1}{\sqrt{d}}, 3 \ldots\right)$.

In the present case, it was shown that the use of homogeneous proposal scalings results in an optimal scaling value converging to infinity, and an AOAR converging towards 0.

To optimize the efficiency of this RWM algorithm, the idea is then to personalize the proposal variances of the last $d-n$ terms, so the last $d-1$ terms in our case. The proposal variance for the first term just stays the same as before, i.e. ℓ^2/d^5. Using the method presented at the beginning of this section, the vector of inhomogeneous proposal scalings is thus $\left(\frac{\ell^2}{d^5}, \frac{\ell^2}{d^{1.5}}, \frac{\ell^2}{d}, \ldots, \frac{\ell^2}{d^{1.5}}, \frac{\ell^2}{d}\right)$. From the results of Section 3.2, we then deduce that

$$E_R = \lim_{d\to\infty} \left(\frac{d-1}{2}\sqrt{d}\frac{1}{d^{1.5}} + \frac{d-1}{2}\left(\frac{1}{3}\right)\frac{1}{d} \right) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3}.$$

Running the Metropolis algorithm for 100,000 iterations in dimensions 101 yields the curves in Figure 4.1, where the solid line again represents the theoretical curve $\upsilon(\ell)$ in (3.7). The theoretical values obtained for $\hat{\ell}^2$ and $a(\hat{\ell})$ are 6 and 0.181 respectively, which agree with the simulations. The inhomogeneous proposal scalings have then contributed to decrease $\hat{\ell}$ while raising the AOAR. Indeed, large values for $\hat{\ell}$ are now inappropriate since components with larger scaling terms now possess proposal variances that are suited to their size, ensuring an reasonable speed of convergence for these components.

As illustrated in the previous example, the adjustment of the proposal variances of the last $d-n$ components also affects the mixing rate of these components. That is, each component X_j with $j \in \mathcal{J}(i,d)$ now mixes according to $O(c(\mathcal{J}(i,d)))$, $i = 1, \ldots, m$, while the first n components still mix according to $O(d^{\alpha})$ (for the particular values of α found when we set $\theta_{i^*}(d) = 1$ for $i^* = 1, \ldots, n$). The inhomogeneous assumption thus results in an improved efficiency for the majority of the last $d-n$ components.

Note that we could also personalize the proposal variances of all d scaling

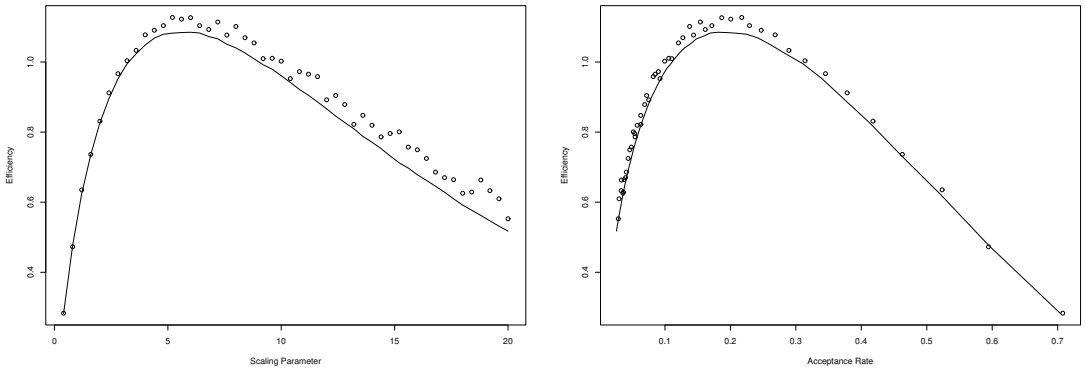

Figure 4.1: Left graph: efficiency of X_2 versus ℓ^2. Right graph: efficiency of X_2 versus the acceptance rate. In both cases, the solid line represents the theoretical curve and the dotted curve is the result of simulations in dimensions 101.

terms, meaning that we could adjust the proposal variances of the first n components as a function of d by setting $\alpha_j = \lambda_j$, $j = 1, \ldots, n$. Such a modification of the proposal scaling vector would result in processes $\left\{ \mathbf{Z}_{i*}^{(d)}(t), t \geq 0 \right\}$ which asymptotically behave as in Theorem 3.2.1 and Corollary 3.2.2. This can be explained by the fact that each of $X_1, \ldots, X_n$ would now mix according to $O(1)$, and thus these components would always affect the limiting distribution of the process of interest. We however feel that the first option presented is a better compromise since in some cases, we might take advantage of the fact that the AOAR is 0.234, which will not happen if the proposal variance of every component is personalized.

4.3 Various Target Extensions

It is important to see how the conclusions of Chapter 3 extend to more general target distribution settings. First, we can relax the assumption of equality among the scaling terms $\theta_j^{-2}(d)$ for $j \in \mathcal{J}(i, d)$. That is, we assume the constant terms within

each of the m groups to be random and come from some distribution satisfying certain moment conditions. In particular, let

$$\begin{aligned}\boldsymbol{\Theta}^{-2}(d) &= \left(\frac{K_1}{d^{\lambda_1}},\ldots,\frac{K_n}{d^{\lambda_n}},\frac{K_{n+1}}{d^{\gamma_1}},\ldots,\frac{K_{n+c(\mathcal{J}(1,d))}}{d^{\gamma_1}},\ldots\right. \\ &\qquad \left.,\frac{K_{n+\sum_{i=1}^{m-1}c(\mathcal{J}(i,d))+1}}{d^{\gamma_m}},\ldots,\frac{K_d}{d^{\gamma_m}}\right). \qquad (4.2)\end{aligned}$$

We assume that $\{K_j, j \in \mathcal{J}(i,d)\}$ are *iid* and chosen randomly from some distribution with $\mathrm{E}\left[K_j^{-2}\right] < \infty$. Without loss of generality, we denote $\mathrm{E}\left[K_j^{-1/2}\right] = a_i$ and $\mathrm{E}\left[K_j^{-1}\right] = b_i$ for $j \in \mathcal{J}(i,d)$. Recall that the scaling term of the component of interest is assumed to be independent of d, and we therefore have $\theta_{i^*}^{-2}(d) = K_{i^*}$.

To support the previous modifications, we now suppose that $-\infty < \gamma_m < \gamma_{m-1} < \ldots < \gamma_1 < \infty$. In addition, we suppose that there does not exist a λ_j, $j = 1,\ldots,n$ equal to one of the γ_i, $i = 1,\ldots,m$. This means that if there is an infinite number of scaling terms with the same power of d, they must necessarily belong to the same of the m groups. We obtain the following results.

Theorem 4.3.1. *Consider the setting of Theorem 3.1.1 with* $\boldsymbol{\Theta}^{-2}(d)$ *as in (4.2) and* $\theta_{i^*} = K_{i^*}^{-1/2}$. *We have*

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

where $Z(0)$ *is distributed according to the density* $\theta_{i^*} f(\theta_{i^*} x)$ *and* $\{Z(t), t \geq 0\}$ *satisfies the Langevin SDE*

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

if and only if

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{\sum_{j=1}^{n} d^{\lambda_j} + \sum_{i=1}^{m} c\left(\mathcal{J}\left(i,d\right)\right) d^{\gamma_i}} = 0. \tag{4.3}$$

Here, $\upsilon\left(\ell\right)$ is as in Theorem 3.1.1 and

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

with

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

Furthermore, the conclusions of Corollary 3.1.2 are preserved.

It is important to notice that Conditions (3.1) and (4.3) are equivalent since the constant terms are assumed to be finite. Condition (4.3) is however easier to verify in the present case due to the randomness of the constant terms.

The fact of admitting a certain level of variability among the scaling terms slightly affects the efficiency of the algorithm. In order to illustrate this, suppose that we transform the scaling vector so as to obtain $\theta_{i^*} = 1$. In that case, we would replace E_R in the previous theorem by

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

Now, suppose that we study a target distribution similar to that described previously but where the K_j's, instead of being random, are equal to $1/a_i^2$ for $j \in \mathcal{J}\left(i,d\right)$. If we suppose, for this particular target, that $\theta_{i^*}\left(d\right) = K_{i^*}^{(1)}$ and that

we transform the scaling vector so that $\theta_{i^*}(d) = 1$, we obtain

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

We now compare the speed measure obtained for the respective Langevin diffusion processes. Specifically, we can reexpress the speed measure in Theorem 4.3.1 as

$$\upsilon\left(\ell\right) = 2\ell^2\Phi\left(-\frac{\ell\sqrt{E_R}}{2}\right) = \frac{E_R^*}{E_R} \times 2\left(\ell^2\frac{E_R}{E_R^*}\right)\Phi\left(-\frac{\sqrt{\ell^2 E_R/E_R^*}\sqrt{E_R^*}}{2}\right),$$

which makes clear that the efficiency of the algorithm as a function of the acceptance rate is identical to (3.2) in Theorem 3.1.1, but now multiplied by the factor E_R^*/E_R. The component X_{i^*} thus mixes according to $O\left(\frac{E_R}{E_R^*}d^{\alpha}\right)$ and since $a_i^2 \leq b_i$, we realize that (at least if $K_{i^*}^{(1)} \leq K_{i^*}$) this component is slowed down by a factor of E_R^*/E_R when compared to the corresponding target where the K_j's are fixed.

In the case where K_j, $j = n+1, \ldots, d$ are random and there exists a finite number of scaling terms remaining significantly small as $d \to \infty$, we have the following result.

Theorem 4.3.2. *Consider the setting of Theorem 3.2.1 (Theorem 3.3.1) with* $\boldsymbol{\Theta}^{-2}(d)$ *as in (4.2),* $\theta_{i^*} = K_{i^*}^{-1/2}$ *and replace Condition (3.4) by*

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{\sum_{j=1}^{n} d^{\lambda_j} + \sum_{i=1}^{m} c\left(\mathcal{J}\left(i,d\right)\right) d^{\gamma_i}} > 0. \tag{4.4}$$

We have

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

where $Z(0)$ *is distributed according to the density* $\theta_{i^*} f(\theta_{i^*} x)$ *and* $\{Z(t), t \geq 0\}$ *is identical to the limit found in Theorem 3.2.1 (Theorem 3.3.1) for the first* b *components, but where it satisfies the Langevin SDE*

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

for the other $d - b$ *components, with* $\upsilon(\ell)$ *as in Theorem 3.2.1 (Theorem 3.3.1).*

For both limiting processes, we now use

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

instead of (3.8) in Theorem 3.2.1, with

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

In addition, the conclusion of Corollary 3.2.2 (Corollary 3.3.2) is preserved.

We note that if the terms K_j's are known, it might be better to scale the proposal distribution proportional to the K_j's. In particular, this would allow us to recover the loss in the efficiency attributed to the introduction of randomness among the scaling terms. In fact, one only needs to know the K_j's of the groups of scaling terms having an impact on $\sigma^2(d)$ (i.e. with $O(c(\mathcal{J}(i,d)) d^{\gamma_i}) = O(d^{\alpha})$), as well as those of the significantly small scaling terms if there is any. This would yield slightly more efficient algorithms, with limiting results similar to those found for each of the three different cases presented in Chapter 3. In particular, this means that this adjustment would not be sufficient to obtain an efficient algorithm

in the presence of extremely small scaling terms (Section 3.3), and inhomogeneous proposal scalings would still be necessary in this case.

The previous results can also be extended to more general functions $c\left(\mathcal{J}\left(i,d\right)\right)$, $i=1,\ldots,m$ and $\theta_j\left(d\right)$, $j=1,\ldots,d$. In order to have sensible limiting theory, we however restrict our attention to functions for which the limit exists as $d\to\infty$. As before, we must also have $c\left(\mathcal{J}\left(i,d\right)\right)\to\infty$ as $d\to\infty$. We can even allow the scaling terms $\left\{\theta_j^{-2}\left(d\right), j\in\mathcal{J}\left(i,d\right)\right\}$ to vary within each of the m groups, provided that they are of the same order. That is, for $j\in\mathcal{J}\left(i,d\right)$ we suppose

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

for some reference function $\theta_i'\left(d\right)$ and some constant K_j coming from the distribution described for Theorems 4.3.1 and 4.3.2. For instance, if $\theta_j\left(d\right)=\sqrt{d^2+d+1}$ for some $j\in\mathcal{J}\left(1,d\right)$ then we obtain $\theta_1'\left(d\right)=d$.

As for the previous two theorems, we assume that if there is infinitely many scaling terms of a certain order they must all belong to one of the m groups. Hence, $\boldsymbol{\Theta}^{-2}\left(d\right)$ contains at least m and at most $n+m$ functions of different order. The positions of the elements belonging to the i-th group are thus given by

$$\mathcal{J}\left(i,d\right)=\left\{j\in\left\{1,\ldots,d\right\};0<\lim_{d\to\infty}\theta_j^{-2}\left(d\right)\theta_i'^{2}\left(d\right)<\infty\right\},\tag{4.5}$$

for $i\in\left\{1,\ldots,m\right\}$. We again suppose that the scaling terms are classified according to an asymptotic increasing order. In particular, the first n terms of $\boldsymbol{\Theta}^{-2}\left(d\right)$ satisfy $\theta_1^{-2}\left(d\right)\prec\ldots\prec\theta_n^{-2}\left(d\right)$ and the order of the following m terms is chosen to satisfy $\theta_1'^{\,-2}\left(d\right)\prec\ldots\prec\theta_m'^{\,-2}\left(d\right)$.

For such target distributions we define the proposal scaling to be $\sigma^2(d) = \ell^2\sigma_\alpha^2(d)$, with $\sigma_\alpha^2(d)$ the function of largest possible order such that

$$\begin{aligned} &\lim_{d\to\infty} \theta_1^2(d)\,\sigma_\alpha^2(d) < \infty, \\ &\lim_{d\to\infty} c(\mathcal{J}(i,d))\,\theta_i'^2(d)\,\sigma_\alpha^2(d) < \infty \quad \text{for } i = 1,\ldots,m. \end{aligned} \tag{4.6}$$

We then have the following results.

Theorem 4.3.3. *Under the setting of Theorem 4.3.1, but with proposal scaling $\sigma^2(d) = \ell^2\sigma_\alpha^2(d)$ where $\sigma_\alpha^2(d)$ satisfies (4.6) and with general functions for $c(\mathcal{J}(i,d))$ and $\theta_j(d)$ as defined previously, the conclusions of Theorem 4.3.1 are preserved, provided that*

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

holds instead of Condition (3.1) and with

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

where $c(\mathcal{J}(i,d))$ is the cardinality function of (4.5).

Interestingly, the asymptotically optimal acceptance rate can be shown to be 0.234 as before.

Theorem 4.3.4. *Under the setting of Theorem 4.3.2, but with proposal scaling $\sigma^2(d) = \ell^2\sigma_\alpha^2(d)$ where $\sigma_\alpha^2(d)$ satisfies (4.6) and with general functions for $c(\mathcal{J}(i,d))$ and $\theta_j(d)$ as defined previously, the conclusions of Theorem 4.3.2 are*

preserved, provided that

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

holds instead of Condition (4.4),

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

holds instead of Condition (3.5), and

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

holds instead of Condition (3.9).

Under this setting, the quantity E_R is now given by

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

where $c(\mathcal{J}(i,d))$ is the cardinality function of (4.5).

Although the AOAR might turn out to be close to the usual 0.234 it is also possible to face a case where this rate is inefficient, from where the importance to determine the correct proposal variance.

These theorems assume quite a general form for the scaling terms of the target distribution and allow for a lot of flexibility. This is important as the results of Chapter 3 cannot always be applied due to the simplistic form of the assumed scaling terms.

4.4 Simulation Studies: Hierarchical Models

This section focuses on applying the results presented to some popular statistical models. The examples illustrate how to deal with more intricate situations, and also demonstrate that the results are robust to certain types of dependence among the components of the target density.

In Section 4.4.1, we show how to optimize the performance of the RWM algorithm for multivariate normal targets with correlated components. In Sections 4.4.2 and 4.4.3 we study the variance components model and the gamma-gamma hierarchical model respectively. Although the results presented in this paper do not directly apply to these last two cases, these examples allow to evaluate their robustness to other types of targets.

4.4.1 Normal Hierarchical Model

The first example we consider is a three-level hierarchical model where all the densities are normal, and whose goal is to illustrate how to deal with more complicated covariance matrices. Indeed, computing the determinant of an intricate covariance matrix is rarely an easy task, and it might reveal challenging to determine how the eigenvalues of high-dimensional target distributions evolve with d. The following example shall hopefully complement Example 3.2.4 of Section 3.2, in which eigenvalues were straight-forwardly determined.

Consider a model with location parameters $\mu_1 \sim N(0,1)$ and $\mu_2 \sim N(\mu_1,1)$. Further suppose that there exist 18 components which are conditionally *iid* given μ_1 and μ_2: half of them (i.e. 9 components) are distributed according to $X_i \sim N(\mu_1,1)$, while the other half satisfies $X_i \sim N(\mu_2,1)$.

Since each of these 20 components is normally distributed, the joint distribution of μ_1, μ_2, X_1, ..., X_{18} will also be a multivariate (20-dimensional) normal distribution, where the unconditional mean turns out to be the null vector. Obtaining the covariance between each pair of components is easily achieved by using conditioning: for the variances, we obtain $\sigma_1^2 = 1$, $\sigma_i^2 = 2$ for $i = 2, \ldots, 11$ and $\sigma_i^2 = 3$ for $i = 12, \ldots, 20$; for the covariances, we get $\sigma_{ij} = 2$ for $i = 2$, $j = 12, \ldots, 20$ (or vice versa) and for $i = 12, \ldots, 20$, $j = 12, \ldots, 20$, $i \neq j$; all the other covariance terms are equal to 1. The covariance matrix then looks like

$$\Sigma_{20} = \begin{pmatrix} 1 & 1 & & \cdots & & 1 & 1 & & \cdots & 1 \\ 1 & 2 & 1 & & \cdots & 1 & 2 & & \cdots & 2 \\ 1 & 1 & 2 & 1 & \cdots & 1 & 1 & \cdots & 1 & 1 \\ \vdots & & 1 & 2 & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & \vdots & \vdots & & \ddots & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 2 & 1 & \cdots & 1 & 1 \\ 1 & 2 & 1 & \cdots & 1 & 1 & 3 & 2 & \cdots & 2 \\ \vdots & & \vdots & \ddots & \vdots & \vdots & 2 & 3 & & \vdots \\ & \vdots & 1 & \cdots & 1 & 1 & \vdots & & \ddots & 2 \\ 1 & 2 & 1 & \cdots & 1 & 1 & 2 & \cdots & 2 & 3 \end{pmatrix}_{20\times 20}.$$

In order to determine which one of the speed measures in (3.2), (3.7) or (3.11) should be used for the optimization problem, we need to know how the eigenvalues of the $d \times d$ matrix Σ_d evolve as a function of d. Finding the eigenvalues of such a covariance matrix can be tedious, especially in large dimensions. A useful method is to transform the matrix into a triangular one, which allows us to determine

Table 4.1: Eigenvalues for Σ_d in various dimensions

d	$\lambda_1(d)$	$\lambda_2(d)$	$\lambda_3(d)$	$\lambda_4(d)$
600	0.003305	0.003329	115.5865	786.4069
800	0.002484	0.002498	153.7839	1048.211
a_i	1.982754	1.997465	0.192644	1.310678
$\frac{a_i}{800}$	0.002478	0.002497	-	-
$800a_i$	-	-	154.1153	1048.542

the determinant by taking the product of the diagonal terms. By applying this method, we find that $d-4$ of the eigenvalues are exactly equal to 1 while the other four are the solutions of the equation

$$\lambda^4 - \left(\frac{3d}{2} - 1\right)\lambda^3 + \left(\frac{d^2}{4} + \frac{d}{2} + 2\right)\lambda^2 - (d+1)\lambda + 1 = 0.$$

Unfortunately, solving for the roots of a polynomial of degree 4 is possible but not straight-forward as there does not exist a nice formula as for polynomials of degree 2.

Determining eigenvalues numerically for any given matrix is easily achieved by using any statistical software (we used R). A way of obtaining the information needed for $\lambda_1(d)$, $\lambda_2(d)$, $\lambda_3(d)$ and $\lambda_4(d)$, the four remaining eigenvalues ordered in ascending order, is thus to examine the numerical eigenvalues of Σ_d in various dimensions. A plot of $\lambda_i(d)$ versus $1/d$ for $i = 1, 2$ clearly shows that the two smallest eigenvalues are linear functions of $1/d$, and satisfy $a_i/d = \lambda_i(d)$ for $i = 1, 2$. Similarly, a plot of $\lambda_i(d)$ versus d for $i = 3, 4$ reveals a relation of the form $a_i d = \lambda_i(d)$ for the two largest eigenvalues.

We can even approximate the constant terms of $\lambda_i(d)$ for $i = 1, \ldots, 4$. Using

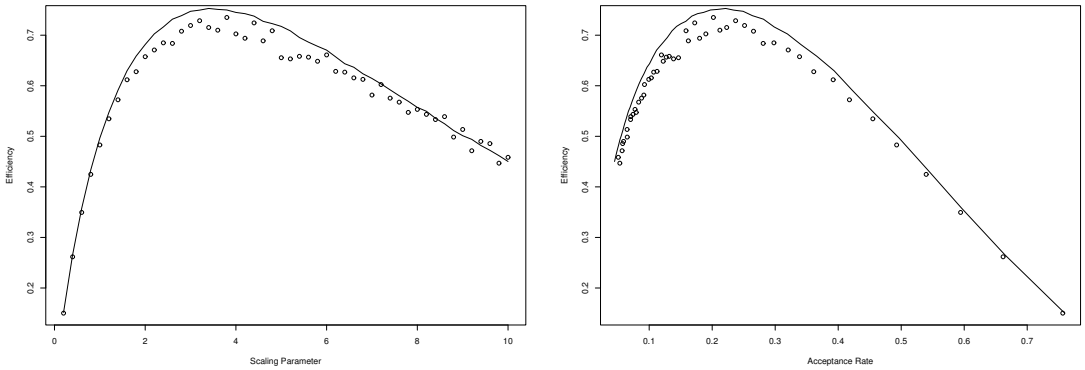

Figure 4.2: Left graph: efficiency of X_3 versus ℓ^2. Right graph: efficiency of X_3 versus the acceptance rate. The solid line represents the theoretical curve, while the dotted curve is the result of the simulation study.

the numbers obtained in dimensions 600 and recorded in Table 4.1, we fit the linear equations stated previously and obtain fitted values for the slopes a_i, $i = 1, \ldots, 4$ (also recorded in the table). The eigenvalues in dimensions 800 are also included along with their fitted counterpart, exhibiting the accuracy of this approach.

Optimizing the efficiency of the algorithm for sampling from the hierarchical model presented previously then reduces to optimizing a 20-dimensional multivariate normal distribution with independent components, null mean and variances equal to $\left(\frac{1.9828}{20}, \frac{1.9975}{20}, 0.1926\,(20), 1.3107\,(20), 1, \ldots, 1\right)$.

It is easily verified that such a vector of scaling terms satisfies Conditions (3.4) and (3.5), and leads to a proposal variance of the form $\sigma^2(\ell) = \ell^2/d$. In light of this information, we should then turn to equation (3.7) to optimize the efficiency of the algorithm. Since $E_R = 1$, we conclude that $\hat{\ell} = 3.4$ and that the AOAR is 0.2214368.

Figure 4.2 presents graphs based on 100,000 iterations of the RWM algorithm, depicting how the first order efficiency of X_5 relates to ℓ^2 and the acceptance

rate respectively. This clearly illustrates that the algorithm behaves similarly to corresponding high-dimensional target distributions (solid curve).

4.4.2 Variance Components Model

It would be interesting to have an idea as to which extend the results presented in this paper are robust to the inclusion of certain relations of dependence between the components of the target density. We are already aware that our results are valid for any multivariate normal target distribution, and thus for any normal hierarchical model where the randomness affects the location parameter (i.e. the mean) only. In this section and the next one, we consider two different hierarchical models engendering distributions that are not jointly normal.

The second simulation study focuses on the variance components model. Let $\mu \sim N(0,1)$, $\sigma_\theta^2 \sim IG(3,1)$ and $\sigma_e^2 \sim IG(2,1)$. The means θ_i are conditionally *iid* given μ, σ_θ^2 and are distributed according to $\theta_i \sim N(\mu, \sigma_\theta^2)$ for $i = 1, \ldots, 30$. The 30 groups of data values are conditionally independent given the mean vector $(\theta_1, \ldots, \theta_{30})$ and the variance σ_e^2, while the values within each group are *iid*. In particular, $Y_{i,j} \sim N(\theta_i, \sigma_e^2)$ for $i = 1, \ldots, 30$ and $j = 1, \ldots, 10$. Graphically, this can be expressed as

$$
\begin{array}{ccccc}
 & & \mu, \sigma_\theta^2 & & \\
 & \swarrow & & \searrow & \\
\theta_1 & & \cdots & & \theta_{30} \\
 & & \sigma_e^2 & & \\
\downarrow & \swarrow & & \searrow & \downarrow \\
Y_{1,1}, \ldots, Y_{1,10} & & \cdots & & Y_{30,1}, \ldots, Y_{30,10}
\end{array}
$$

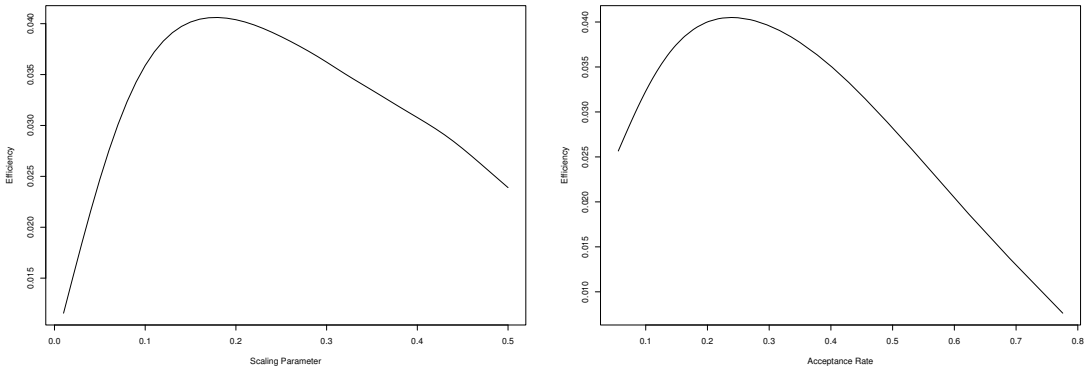

Figure 4.3: Left graph: efficiency of θ_1 versus ℓ^2. Right graph: efficiency of θ_1 versus the acceptance rate.

Under this model, not only the means of the various components are random, but so are the variances. We are interested in the posterior distribution of $\mu, \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_{30}$ given the data $Y_{i,j}$, $i = 1, \ldots, 30$, $j = 1, \ldots, 10$, and thus the target and is such that

$$\pi\left(\mu, \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_{30} \,|\mathbf{Y}\right) \propto \left(\sigma_\theta^2\right)^{-19} \left(\sigma_e^2\right)^{-153} \times \exp\left(-\frac{\mu^2}{2\sigma_0^2} - \frac{1}{\sigma_\theta^2} - \frac{1}{\sigma_e^2} - \sum_{i=1}^{30} \frac{\left(\theta_i - \mu\right)^2}{2\sigma_\theta^2} - \sum_{i=1}^{30}\sum_{j=1}^{10} \frac{\left(\theta_i - Y_{i,j}\right)^2}{2\sigma_e^2}\right). \quad (4.7)$$

The algorithm does not work well when generating moves from a normal proposal to mimic moves from an inverse gamma distribution. In order for the algorithm to behave well, we instead deal with gamma distributions and use the inverse transformation. Hence, it is more convenient to use the target

$$\pi\left(\mu, \varphi_\theta, \varphi_e, \theta_1, \ldots, \theta_{30} \,|\mathbf{Y}\right) \propto \left(\varphi_\theta\right)^{17} \left(\varphi_e\right)^{151} \times \exp\left(-\frac{\mu^2}{2\sigma_0^2} - \varphi_\theta - \varphi_e - \sum_{i=1}^{30} \frac{\varphi_\theta \left(\theta_i - \mu\right)^2}{2} - \sum_{i=1}^{30}\sum_{j=1}^{10} \frac{\varphi_e \left(\theta_i - Y_{i,j}\right)^2}{2}\right) \quad (4.8)$$

To sample from (4.7), it then suffices to sample values from (4.8) using the RWM algorithm, and apply the transformations $\sigma_\theta^2 = 1/\varphi_\theta$ and $\sigma_e^2 = 1/\varphi_e$ in order to retrieve the two variances.

For the sake of the example, the data was simulated from the target distribution and the same sample was used for each simulation. We performed 100,000 iterations of the RWM algorithm and plotted first order efficiency of the fourth component versus ℓ^2 and the acceptance rate. The maximum is located around 0.17 for $\hat{\ell}^2$ and choosing an acceptance rate close to 0.2 then optimizes the efficiency of the algorithm. The AOAR thus seems to lie close to 0.234, but it is hard to tell its exact value from the graph. According to the previous results, we suspect that it might differ from 0.234, which might become clearer when simulating from target distributions possessing a greater number of non-normal components. Since the joint distribution is not normally distributed, we cannot directly use the results introduced in this paper to optimize the performance of the algorithm. Nonetheless, we observe that it seems possible to study the optimal scaling issue not only for hierarchical targets where the mean of normally distributed variables is random, but also for hierarchical targets with more layers and where the variance is random as well.

4.4.3 Gamma-Gamma Hierarchical Model

As a last example, consider a hierarchical target model where the conditionally independent variables are not normally distributed anymore, producing an AOAR substantially smaller than 0.234.

Let $\lambda \sim \Gamma(4, 1)$ and, assuming conditional independence, $X_i \sim \Gamma(4, \lambda)$ for

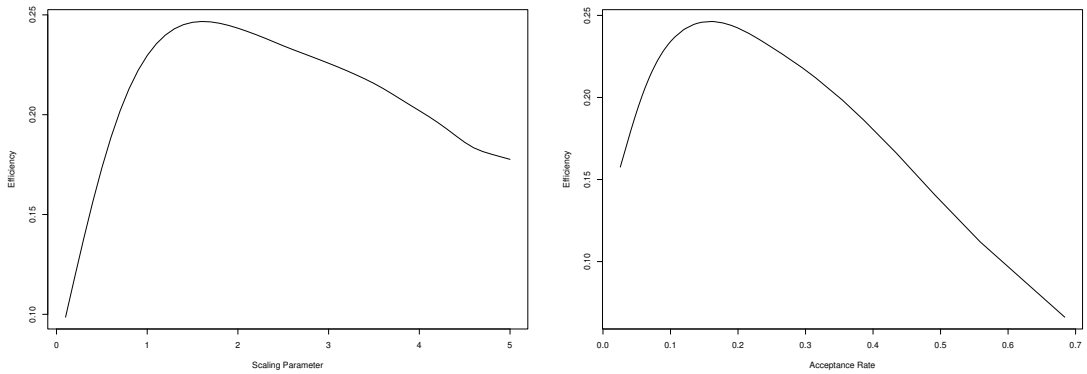

Figure 4.4: Left graph: efficiency of X_1 versus ℓ^2. Right graph: efficiency of X_1 versus the acceptance rate.

$i = 1, \ldots, 20$. The dependency between the 20 variables then appears through the randomness of the scaling parameter λ, and the unconditional 21-dimensional target density satisfies

$$\pi\left(\lambda, x_1, \ldots, x_{20}\right) \propto \lambda^{83} \exp\left(-\lambda\left(1 + \sum_{i=1}^{20} x_i\right)\right) \prod_{i=1}^{20} x_i^3.$$

This time, 10,000,000 iterations of the algorithm were required to reduce Monte Carlo errors and obtain clear curves. Figure 4.4 shows the existence of a finite value of ℓ optimizing the efficiency of the method ($\hat{\ell} = 1.6$), resulting in an optimal acceptance rate lying around 0.16.

This small AOAR appears to corroborate the discussion at the end of last section. That is, it seems feasible to optimize the efficiency of RWM algorithms for general hierarchical target models, but this will yield an AOAR that varies according to the target distribution.

Chapter 5

Weak Convergence of the Rescaled RWM Algorithm: Proofs

In this chapter, we prove the various theorems introduced in Chapters 3 and 4. Detailed proofs are included for Theorems 3.1.1, 3.2.1, and 3.3.1; the demonstrations of the theorems in Chapter 4 being similar, we shall just outline the main differences. In order to be as clear as possible, the proofs have been divided among three chapters. In particular, Chapter 6 contains lemmas building to the proofs that are presented subsequently. That chapter thus complements the present one, and consequently we shall regularly refer to the results it contains.

For the proofs presented subsequently to be complete, it is also essential to refer to various results appearing in [16]; the pillars of the results introduced in this thesis are Theorem 8.2 and Corollary 8.6 of Chapter 4. The first result roughly says that for the finite-dimensional distributions of a sequence of processes to converge

weakly to those of some Markov process, it is enough to verify $\mathcal{L}^1$ convergence of their generators. To reach weak convergence of the stochastic processes themselves, Theorem 7.8 of Chapter 3 in [16] states that it is sufficient to assess relative compactness of the sequence of stochastic processes under study. This is achieved using the second result, Corollary 8.6 of Chapter 4, which introduces conditions for the sequence of stochastic processes considered to be relatively compact. The last chapter is devoted to the study of these results.

For now, our task is then to focus on the proof of the $\mathcal{L}^1$ convergence of the generators. To this end, we base our approach on the proof for the RWM algorithm case in [26]. Note however that the authors instead refer to Corollary 8.7 of Chapter 4 in [16] and prove uniform convergence of generators, which could not be used in the present situation.

The generator is written in term of an arbitrary test function h, which can usually be any smooth function. In our case, we shall most often restrict our attention to functions in C_c^∞, the space of infinitely differentiable functions on compact support. Since the limiting process obtained generally is a diffusion, then C_c^∞ is a core for the generator by Theorem 2.1 of Chapter 8 in [16], meaning that it is representative enough so as to focus on the functions it contains only. This shall also be discussed in Chapter 7.

In order to lighten the formulas, we adopt the following convention for defining vectors: $\mathbf{X}^{(b-a)} = (X_{a+1}, \ldots, X_b)$. The minus sign appearing outside the brackets (e.g. $\mathbf{X}^{(b-a)-}$) means that the component of interest X_{i^*} is excluded from the vector. We also use the following notation for conditional expectations: $\mathrm{E}\left[f\left(X,Y\right)|Y\right] = \mathrm{E}_X\left[f\left(X,Y\right)\right]$. When there is no subscript, the expectation is

taken with respect to all random variables included in the expression.

5.1 Generator of the Rescaled RWM Algorithm

Before proving the different results, we first determine the form of the generator to be used in the demonstrations. Recall that the process we are studying is the i^*-th component of the sped up RWM algorithm. It is interesting to precise that although RWM algorithms are Markovian, the fact that the acceptance probability depends on all d components keeps this one-dimensional process from being Markovian as well. We are aware that the RWM algorithm considered is originally a discrete-time process. We shall see that if we let the time between each proposed move be exponentially distributed with mean $1/d^\alpha$ (i.e. the process jumps according to a Poisson process with rate d^α), then the generator obtained is identical to that of the discrete-time process. In our case, it is thus equivalent to talk about the i^*-th component of the sped up RWM algorithm (and its continuous-time, constant-interpolation version) or its continuous-time version with exponential holding times. Because the results in [16] are expressed in terms of continuous-time processes, it shall then reveal convenient to treat the generator as originating from the latter.

We start by deriving the generator of discrete-time and continuous-time Markov processes. Then, we shall obtain the generator of our (discrete-time) RMW algorithm and make the liaison with its continuous-time version. We should then be in a good position to introduce what we call (maybe abusively since the process in question is not Markovian) the generator of $\{Z_{i^*}(t), t \geq 0\}$, the continuous-time version of the process of interest.

5.1.1 Generator of Markov Processes

We define the generator of a Markov process to be an operator G acting on smooth functions $h : \mathbf{R} \to \mathbf{R}$ such that

$$M(t) = h(X(t)) - \int_0^t Gh(X(s))\, ds, \tag{5.1}$$

is a martingale.

Replacing the integral by a Riemann sum in the previous formula results in an expression for the generator of discrete-time processes. We find that for discrete-time Markov processes, the generator satisfies

$$Gh(X(s)) = \frac{1}{k}\mathrm{E}\left[h(X(s+k)) - h(X(s)) \,|X(s)\right], \tag{5.2}$$

where k is the time interval between successive steps. In order to verify the martingale condition for this generator, it suffices to check that $\mathrm{E}\left[M(t) - M(s) \,|\mathcal{F}(s)\right] = 0$. We have

$$\begin{aligned}
&\mathrm{E}\left[M(t) - M(s) \,|\mathcal{F}(s)\right] \\
&\quad = \mathrm{E}\left[h(X(t)) - h(X(s)) - k\sum_{u=s/k}^{t/k-1} Gh(X(ku)) \,|\mathcal{F}(s)\right] \\
&\quad = \mathrm{E}\left[h(X(t)) - h(X(s)) \,|\mathcal{F}(s)\right] \\
&\qquad - \sum_{u=s/k}^{t/k-1} \mathrm{E}\left[\mathrm{E}\left[h(X(k(u+1))) - h(X(ku)) \,|X(ku)\right]|\,\mathcal{F}(s)\right].
\end{aligned}$$

By the Markov property, we find

$$\begin{aligned}
&\sum_{u=s/k}^{t/k-1} \mathrm{E}\left[\mathrm{E}\left[h\left(X\left(k\left(u+1\right)\right)\right)-h\left(X\left(ku\right)\right)|X\left(ku\right)\right]|\,\mathcal{F}\left(s\right)\right] \\
&= \sum_{u=s/k}^{t/k-1} \mathrm{E}\left[\mathrm{E}\left[h\left(X\left(k\left(u+1\right)\right)\right)-h\left(X\left(ku\right)\right)|\mathcal{F}\left(ku\right)\right]|\,\mathcal{F}\left(s\right)\right] \\
&= \mathrm{E}\left[\sum_{u=s/k}^{t/k-1}\left(h\left(X\left(k\left(u+1\right)\right)\right)-h\left(X\left(ku\right)\right)\right)\middle|\,\mathcal{F}\left(s\right)\right] \\
&= \mathrm{E}\left[h\left(X\left(t\right)\right)-h\left(X\left(s\right)\right)|\mathcal{F}\left(s\right)\right],
\end{aligned}$$

implying that $M\left(t\right)$ is a martingale and that (5.2) is indeed the discrete-time generator of a Markov process.

In a similar fashion, we define the generator of a continuous-time Markov process to be such that

$$\begin{aligned}
Gh\left(X\left(s\right)\right) &= \lim_{k\to 0}\frac{1}{k}\,\mathrm{E}\left[h\left(X\left(s+k\right)\right)-h\left(X\left(s\right)\right)|X\left(s\right)\right] \\
&= \frac{d}{dt}\mathrm{E}\left[h\left(X\left(t\right)\right)|X\left(s\right)\right]\Big|_{t=s}. \qquad (5.3)
\end{aligned}$$

By reproducing the previous development, it is simple to verify that (5.1) is a martingale.

5.1.2 Generator of RWM Algorithms

We now find the generator of our RWM algorithm, which is a d-dimensional discrete-time Markov process. Specifically, we want to find the generator of the sped up version of the algorithm. Since the time interval between each step is

$1/d^{\alpha}$, the generator becomes

$$Gh\left(d,\mathbf{Z}^{(d)}\left(s\right)\right)=d^{\alpha}\mathrm{E}\left[h\left(\mathbf{Z}^{(d)}\left(s+\frac{1}{d^{\alpha}}\right)\right)-h\left(\mathbf{Z}^{(d)}\left(s\right)\right)\left|\mathbf{Z}^{(d)}\left(s\right)\right.\right].$$

Since the process either jumps to the proposed move $\mathbf{Y}^{(d)}\left(d^{\alpha}s+1\right)$ or stays at the current state $\mathbf{Z}^{(d)}\left(s\right)$, we then have

$$\begin{aligned}&\left(h\left(\mathbf{Z}^{(d)}\left(s+\frac{1}{d^{\alpha}}\right)\right)-h\left(\mathbf{Z}^{(d)}\left(s\right)\right)\right)\left|\mathbf{Z}^{(d)}\left(s\right)\right.\\&\quad=\begin{cases}h\left(\mathbf{Y}^{(d)}\left(d^{\alpha}s+1\right)\right)-h\left(\mathbf{Z}^{(d)}\left(s\right)\right), & \text{w.p. }\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\left(d^{\alpha}s+1\right)\right)}{\pi\left(d,\mathbf{Z}^{(d)}\left(s\right)\right)}\right)\\0, & \text{w.p. }1-\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\left(d^{\alpha}s+1\right)\right)}{\pi\left(d,\mathbf{Z}^{(d)}\left(s\right)\right)}\right)\end{cases}.\end{aligned}$$

Using the law of total probabilities, we can express the discrete-time generator of the rescaled RWM algorithm as

$$\begin{aligned}&Gh\left(d,\mathbf{Z}^{(d)}\left(s\right)\right)=\\&\quad d^{\alpha}\ \mathrm{E}_{\mathbf{Y}^{(d)}}\left[\left(h\left(\mathbf{Y}^{(d)}\left(d^{\alpha}s+1\right)\right)-h\left(\mathbf{Z}^{(d)}\left(s\right)\right)\right)\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\left(d^{\alpha}s+1\right)\right)}{\pi\left(d,\mathbf{Z}^{(d)}\left(s\right)\right)}\right)\right].\end{aligned}$$

Note that since $\left\{\mathbf{Z}^{(d)}\left(t\right),t\geq0\right\}$ is just a sped up version of $\left\{\mathbf{X}^{(d)}\left(t\right),t\geq0\right\}$, we can equivalently express the previous expectation in terms on the latter process:

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

Furthermore, since the generator does not depend explicitly on s, but merely on

the last state of the process, we shall subsequently use the simpler expression

$$Gh\left(d,\mathbf{X}^{(d)}\right)=d^{\alpha}\ \mathrm{E}_{\mathbf{Y}^{(d)}}\left[\left(h\left(\mathbf{Y}^{(d)}\right)-h\left(\mathbf{X}^{(d)}\right)\right)\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right)\right]. \tag{5.4}$$

Even though the process we are considering is a discrete-time process, we mentioned previously that the results used to prove weak convergence involve continuous-time Markov processes. The only way to preserve the Markov property of this discrete-time process while making it a continuous-time process is to resort to the memoryless property of the exponential distribution. That is, we let the time between each step be exponentially distributed with mean $1/d^{\alpha}$, so the process will jump at times of a Poisson process with rate d^{α}.

For a d-dimensional continuous-time (sped up) Markov process, we have

$$Gh\left(\mathbf{x}^{(d)}\right)=\lim_{k\to 0}\frac{1}{k}\mathrm{E}\left[h\left(\mathbf{Z}^{(d)}\left(s+k\right)\right)-h\left(\mathbf{Z}^{(d)}\left(s\right)\right)\left|\mathbf{Z}^{(d)}\left(s\right)=\mathbf{x}^{(d)}\right.\right].$$

In order to have the process moving to $\mathbf{Y}^{(d)}$ at time $s+k$ given that the process is at $\mathbf{x}^{(d)}$ at time s, the Poisson process must first jump, which happens with probability $d^{\alpha}k+o\left(k\right)$. Then, given that the process jumps, we have a probability $\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{x}^{(d)}\right)}\right)$ that the process actually goes to $\mathbf{Y}^{(d)}$, otherwise it stays where it is. The generator thus becomes

$$\begin{aligned}Gh\left(d,\mathbf{x}^{(d)}\right) &= \lim_{k\to 0}\frac{1}{k}\ \mathrm{E}\left[\left(h\left(\mathbf{Y}^{(d)}\right)-h\left(\mathbf{x}^{(d)}\right)\right)\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{x}^{(d)}\right)}\right)\left(d^{\alpha}k+o\left(k\right)\right)\right]\\ &= d^{\alpha}\mathrm{E}\left[\left(h\left(\mathbf{Y}^{(d)}\right)-h\left(\mathbf{x}^{(d)}\right)\right)\left(1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{x}^{(d)}\right)}\right)\right]. \end{aligned}\tag{5.5}$$

As we can see, the generator of the continuous-time RWM algorithm in (5.5)

is identical to the (rescaled) discrete-time version in (5.4). Because the results of this thesis are proven in the Skorokhod topology, we are allowed to transform our discrete-time algorithm into an equivalent continuous-time process, and to use the latter to derive weak convergence results. The particularity of this topology is that it allows a small deformation of the time scale, justifying why we can use these two processes interchangeably. The Skorokhod topology shall be briefly described in Section 7.1.

5.1.3 Generator of the i^*-th Component

Although we obtained the generator of RWM algorithms, the process under study is formed of the i^*-th component of this d-dimensional algorithm only. As mentioned previously, this process is not Markovian by itself, but is part of a d-dimensional Markov process. Keeping this in mind, we now make an abuse of notation and define what we call the generator of the i^*-th component of a multidimensional Markov process as

$$Gh\left(X_{i^*}^{(d)}(s)\right) = \lim_{k\to 0}\frac{1}{k}\mathrm{E}\left[h\left(X_{i^*}^{(d)}(s+k)\right) - h\left(X_{i^*}^{(d)}(s)\right)\middle|\mathcal{F}^{X_{i^*}^{(d)}}(s)\right].$$

In spite of the fact that $\left\{X_{i^*}^{(d)}(t), t \geq 0\right\}$ is not Markovian, the previous expression does satisfy the martingale condition in (5.1). In particular, we can show that $M(t)$ is an $\mathcal{F}^{X_{i^*}^{(d)}}(t)$-martingale when G is defined as above; this can be verified by using a method similar to that presented in Section 5.1.1.

Starting from the definition of our sped up RWM algorithm, we then find that $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\}$, the i^*-th component of the d-dimensional sped up process, has

generator

$$Gh\left(d, X_{i^*}\right) = d^{\alpha}\mathrm{E}_{\mathbf{Y}^{(d)},\mathbf{X}^{(d)-}}\left[\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right)\left(1 \wedge \frac{\pi\left(d, \mathbf{Y}^{(d)}\right)}{\pi\left(d, \mathbf{X}^{(d)}\right)}\right)\right], \tag{5.6}$$

where the expectation is taken with respect to all random variables included in the expression except X_{i^*}.

Since the generator of our RWM algorithm is identical to that of its continuous-time version, this shall be the same for the generator of this one component. As mentioned previously, we can then associate the generator in (5.6) to the continuous-time version of the sped up RWM algorithm with exponential holding times. This detail shall allow us to use the results in [16].

5.1.4 Generators' Dilemma

Traditionally, optimal scaling results have been derived by studying the limit of $Gh\left(d, \mathbf{X}^{(d)}\right)$, the generator of the (sped up) RWM algorithm. In particular, to study only one component of this algorithm, researchers use the following expression for the generator:

$$\begin{aligned} Gh\left(d, \mathbf{X}^{(d)}\right) &= \lim_{k \to 0} \frac{1}{k}\mathrm{E}\left[h\left(Z_{i^*}^{(d)}\left(s+k\right)\right) - h\left(Z_{i^*}^{(d)}\left(s\right)\right) \middle| \mathcal{F}^{\mathbf{Z}^{(d)}}\left(s\right)\right] \\ &= d^{\alpha}\mathrm{E}_{\mathbf{Y}^{(d)}}\left[\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right)\left(1 \wedge \frac{\pi\left(d, \mathbf{Y}^{(d)}\right)}{\pi\left(d, \mathbf{X}^{(d)}\right)}\right)\right]. \end{aligned} \tag{5.7}$$

That is, they study the limiting distribution of one particular component of the algorithm and to do so, they have access to all the past information of the d-dimensional process, $\mathcal{F}^{\mathbf{Z}^{(d)}}\left(s\right)$. This makes much sense, as the ultimate goal in

deriving these weak convergence results is to find an optimal scaling value which is, by assumption, the same for every component; this optimal value should then take into account the moves of the whole process and not just one component.

So far, this method has served the case pretty well. Indeed, the limiting distributions obtained have been Markovian in the limit, and thus independent of the other components of the algorithm. In fact, the limiting distributions found have yielded a simple optimization problem: optimize the speed measure of the limiting Langevin diffusion, and find an AOAR of 0.234.

In this work, $Gh\left(d, X_{i^*}\right)$ in (5.6) is used to prove $\mathcal{L}^1$ convergence of the generators and study the limiting behavior of the RWM algorithm. Using such a generator is equivalent to studying the "marginal" distribution of the i^*-th component, i.e. studying the limiting behavior of this component without knowing what happens to the other $d-1$ components of the process. We thus possess information on the path of the i^*-th component only, $\mathcal{F}^{Z_{i^*}^{(d)}}(s)$. From a mathematical viewpoint, it does not cause any problem to study this type of "marginal" process. However, since the proposal scaling value is assumed to be the same for every component of the algorithm, then using this generator to solve the optimal scaling problem might not seem as intuitive as the usual method. For this reason, we now justify this choice of generator.

In the cases considered in the literature, where the the limiting processes are Markovian given $\mathcal{F}^{\mathbf{Z}^{(d)}}(t)$, note that using the generator in (5.6) to study the marginal distribution of every component of the algorithm would yield the same conclusions. Indeed, since the limiting processes obtained are Markovian, they do not depend on the variables X_i, $i \neq i^*$; taking an extra expectation of the generator

in (5.7) with respect to those variables thus leaves the processes intact, implying that both methods are equivalent. This is exactly what happened in Section 3.1, where there did not exist a finite number of components with significantly small scaling terms.

The particularity of our problem is that in certain cases (see Sections 3.2 and 3.3), there exists a finite number of components that are affecting the accept/reject ratio of the algorithm in the limit; consequently, the limiting distribution obtained by using a generator as in (5.7) depends specifically on these b components having significantly small scaling terms. As far as the first b components are concerned, they weakly converge to a b-dimensional discrete-time RWM algorithm. Hence, it would not be a wise move to use the limiting distribution of those components to determine the optimal scaling value $\hat{\ell}$, because as mentioned in Section 2.4, efficiency measures are not unique for discrete-time processes; we thus have to base our analysis on the last $d-b$ components.

Given $\mathcal{F}^{\mathbf{Z}^{(d)}}(t)$, the limiting distribution of the i^*-th component ($i^* = b+1, \ldots, d$) is thus not Markovian in the strict sense, but is part of a multidimensional Markov process. Since the limiting distribution of the last $d-b$ components each depends on the first b components, we then face a problem when it comes to determining the optimal scaling value. First, because we obtain a limiting diffusion process whose drift and volatility terms depend on $\ell^2, X_1, \ldots, X_b$ (that is, we cannot factorize a speed measure for the diffusion), it becomes difficult to determine how the equation can be optimized. Secondly, even it we had a nice optimization problem (i.e. a speed measure for the diffusion), then the optimal scaling value $\hat{\ell}$ would be a function of $X_1, \ldots, X_b$, and would thus depend on the current state

of the algorithm. Since we want to determine a global value $\hat{\ell}$ for the algorithm, the natural way to reach this goal would then be to take the expectation of the various $\hat{\ell}\,(X_1, \ldots, X_b)$. In the present case, we unfortunately cannot obtain those quantities; however, we can overturn this problem by considering the probability to obtain a particular diffusion, and thus take the expectation of the limiting process with respect to the variables $X_1, \ldots, X_b$. Looking for a global value $\hat{\ell}$ is then equivalent to studying some component given its own past only. In other words, we do not want to possess any information about the other components, in order to avoid obtaining a limiting process that depends on them.

By applying this method, we see in Sections 3.2 and 3.3 that we obtain a limiting Langevin diffusion with some speed measure which is a function of ℓ only, and which can be used to find an optimal value $\hat{\ell}$ and an AOAR. Of course, $\hat{\ell}$ might not be optimal at a given state (i.e. for specific values of $X_1, \ldots, X_b$); however, $\hat{\ell}$ will yield better results than any other constant ℓ. It is also important to precise that the reason why such a method makes sense is due, among other things, to the fact that each of the last $d-b$ components possesses the same limiting distribution; if it were not the case, then it would be impossible to obtain a sensible value $\hat{\ell}$ by studying the marginal limit of one component only.

From the previous discussion, we realize that the limiting process obtained by considering the generator in (5.6) might be different from that obtained when working with (5.7). In general, we can say that if (5.7) weakly converges to a Markovian process, then the limit will be identical when working with (5.6); however, the converse is not necessarily true. From a mathematical viewpoint, it does not cause any problem to study one limit or the other. When it comes to the

optimal scaling problem however, it might become necessary to use (5.6) in order to obtain a global value $\hat{\ell}$, as was the case in Sections 3.2 and 3.3. For a matter of consistency, the weak convergence results in this thesis are all derived by using (5.6), even though (5.7) would have worked equally well in Section 3.1.

5.2 The Familiar Asymptotic Behavior

We now present the proofs of the results presented in Section 3.1, where the conclusions are identical to those established for the *iid* case.

5.2.1 Restrictions on the Proposal Scaling

The first step of the proof is to transform Condition (3.1) into a statement about the proposal variance and its parameter α. For this condition to be satisfied, we must equivalently have

$$\begin{aligned}\lim_{d\to\infty}\frac{K_1}{d^{\lambda_1}}\left(\frac{d^{\lambda_1}}{K_1}+\ldots+\frac{d^{\lambda_n}}{K_n}\right)\\ +\lim_{d\to\infty}\frac{K_1}{d^{\lambda_1}}\left(c\left(\mathcal{J}\left(1,d\right)\right)\frac{d^{\gamma_1}}{K_{n+1}}+\ldots+c\left(\mathcal{J}\left(m,d\right)\right)\frac{d^{\gamma_m}}{K_{n+m}}\right)=\infty.\end{aligned}$$

Letting $b = \max\left(j\in\{1,\ldots,n\}\,;\lambda_j=\lambda_1\right)$, i.e. the number of components with a scaling term of the same order as that of X_1, we obtain

$$\lim_{d\to\infty}\frac{K_1}{d^{\lambda_1}}\left(\frac{d^{\lambda_1}}{K_1}+\ldots+\frac{d^{\lambda_n}}{K_n}\right) \;=\; 1+\sum_{j=2}^{b}\frac{K_1}{K_j}<\infty.$$

To have an overall limit that is infinite, there must then exist at least one $i\in$

$\{1, \cdots, m\}$ such that

$$\lim_{d\to\infty} \frac{c\left(\mathcal{J}\left(i,d\right)\right) d^{\gamma_i}}{d^{\lambda_1}} = \infty. \tag{5.8}$$

This implies that the form of the proposal variance, i.e. the choice of the parameter α, must be based on one of the groups of scaling terms appearing infinitely often. In other words, it cannot possibly be based on K_1/d^{λ_1}, the smallest scaling term appearing a fixed number of times. If it was, this would mean that

$$\lim_{d\to\infty} \frac{c\left(\mathcal{J}\left(i,d\right)\right) d^{\gamma_i}}{d^{\alpha}} = \lim_{d\to\infty} \frac{c\left(\mathcal{J}\left(i,d\right)\right) d^{\gamma_i}}{d^{\lambda_1}} = \infty,$$

for all i for which (5.8) was diverging, which would contradict the definition of α. Therefore when Condition (3.1) is satisfied, it follows that $\lim_{d\to\infty} d^{\lambda_1}/d^{\alpha} = 0$ and $\theta_1^{-2}\left(d\right)$ does not have any impact on the determination of the parameter α. This thus implies that α is always strictly greater than 0, no matter which component is under study.

5.2.2 Proof of Theorem 3.1.1

We now demonstrate that the generator in (5.6) converges in $\mathcal{L}^1$ to that of the Langevin diffusion. To this end, we shall use results appearing in Sections 6.1, 6.2, and 6.3.

We need to show that for an arbitrary test function $h \in C_c^{\infty}$,

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

where $G_L(X_{i^*}) = \upsilon(\ell)\left[\frac{1}{2}h''(X_{i^*}) + \frac{1}{2}h'(X_{i^*})(\log f(X_{i^*}))'\right]$ is the generator of a Langevin diffusion process with speed measure $\upsilon(\ell)$ as in Theorem 3.1.1.

We begin by introducing a third generator $\widetilde{G}h(d, X_{i^*})$ (as in (6.1) of Lemma 6.1.2) that is asymptotically equivalent to the original generator $Gh(d, X_{i^*})$. By the triangle's inequality,

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

From Lemma 6.1.2, the first expectation on the RHS converges to 0 as $d \to \infty$. To prove the theorem, we are thus left to show $\mathcal{L}^1$ convergence of the generator $\widetilde{G}h(d, X_{i^*})$ to that of the Langevin diffusion.

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

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

Since the function h has compact support, it implies that h itself and its derivatives are bounded in absolute value by some constant. Therefore, $\mathrm{E}\left[|h''(X_{i^*})|\right]$ and $\mathrm{E}\left[\left|h'(X_{i^*})(\log f(X_{i^*}))'\right|\right]$ are both bounded by K, say. Using Lemmas 6.2.1 and 6.3.1, we then conclude that the first absolute difference on the RHS goes to

0 as $d \to \infty$; we reach the same conclusion for the second absolute difference by applying Lemmas 6.2.2 and 6.3.2.

5.3 The New Limiting Behaviors

In this section, we prove results admitting conclusions which differ from the *iid* case; these results were introduced in Sections 3.2 and 3.3.

5.3.1 Restrictions on the Proposal Scaling

Condition (3.4) ensures that there exists a finite number of scaling terms significantly smaller than the others. An impact of this condition is that it also determines the variance of the proposal distribution as a function of d, which we now demonstrate. For this condition to be verified, we must equivalently have

$$\lim_{d\to\infty} \frac{K_1}{d^{\lambda_1}} \left(\frac{d^{\lambda_1}}{K_1} + \ldots + \frac{d^{\lambda_n}}{K_n} \right) + \lim_{d\to\infty} \frac{K_1}{d^{\lambda_1}} \left(c\left(\mathcal{J}\left(1,d\right)\right) \frac{d^{\gamma_1}}{K_{n+1}} + \ldots + c\left(\mathcal{J}\left(m,d\right)\right) \frac{d^{\gamma_m}}{K_{n+m}} \right) < \infty.$$

It must then be true that

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

in which case $\lim_{d\to\infty} \sum_{j=1}^{d} \hat{\theta}^{-2}\left(d\right) \theta_j^2\left(d\right) = 1 + \sum_{j=2}^{b} K_1/K_j < \infty$, where $b = \max\left(j \in \left\{1,\ldots,n\right\}; \lambda_j = \lambda_1\right)$. In other words, $\theta_1^{-2}\left(d\right)$ is not only the asymptotically smallest scaling term, but it is also small enough so as to act of proposal variance for the algorithm, implying that $\sigma^2\left(d\right) = \ell^2/d^{\lambda_1}$.

5.3.2 Proof of Theorem 3.2.1

We now show that the generator

$$Gh\left(d, X_{i^*}\right) = d^{\lambda_1} \mathrm{E}_{\mathbf{Y}^{(d)}, \mathbf{X}^{(d)-}} \left[\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right) \right) \left(1 \wedge \frac{\pi\left(d, \mathbf{Y}^{(d)}\right)}{\pi\left(d, \mathbf{X}^{(d)}\right)} \right) \right] \quad (5.9)$$

converges in $\mathcal{L}^1$ to the generator of a one-dimensional Metropolis-Hastings algorithm with a particular acceptance rule in some cases, and to that of a Langevin diffusion in other cases. To this end, we shall use the results appearing in Sections 6.1, 6.2 and 6.4.

We first need to show that for $i^* \in \{1, \ldots, b\}$ and an arbitrary test function $h \in \overline{C}$, the space of bounded and continuous functions on $\mathbf{R}$,

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

where

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

with acceptance rule

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

Since the component of interest is assumed to have a scaling term equal to 1, we have $\lambda_j = \lambda_1 = 0$ for $j = 1, \ldots, b$ and the proposal scaling is $\sigma^2(d) = \ell^2$. Since $\sigma^2(d)$ does not depend on the dimension of the target, there is thus no need to

speed up the RWM algorithm in order to obtain a nontrivial limit, justifying the discrete-time nature of the limiting process.

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

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

with $z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ as in (6.6). By the triangle's inequality,

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

and the first expectation on the RHS converges to 0 as $d \to \infty$ by Lemma 6.1.3. To complete the proof, we are then left to show $\mathcal{L}^1$ convergence of the third generator $\widehat{G}h\left(d, X_{i^*}\right)$ to that of the modified Metropolis-Hastings algorithm.

Substituting explicit expressions for the generators and using the triangle's inequality along with the fact that the function h is bounded give

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

where K is chosen such that $\left|h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right| < K$. Since the expectation on the RHS converges to 0 as ascertained by Lemma 6.4.1, this proves the first part of Theorem 3.2.1.

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

arbitrary test function $h \in C_c^\infty$,

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

where $G_L\left(X_{i^*}\right) = \upsilon\left(\ell\right)\left[\frac{1}{2}h''\left(X_{i^*}\right) + \frac{1}{2}h'\left(X_{i^*}\right)\left(\log f\left(X_{i^*}\right)\right)'\right]$ and $\upsilon\left(\ell\right)$ is as in (3.7).

Similar to the first part of the proof, we introduce a third generator $\widetilde{G}h\left(d, X_{i^*}\right)$ as in Lemma 6.1.2. From this same lemma, we conclude that this new generator is asymptotically equivalent to the original generator and hence we have $\mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - \widetilde{G}h\left(d, X_{i^*}\right)\right|\right] \to 0$ as $d \to \infty$. We then complete the proof by showing that this third generator also converges in mean to the generator of the Langevin diffusion.

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

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

Since the function h has compact support, this implies that both $\mathrm{E}\left[\left|h''\left(X_{i^*}\right)\right|\right]$ and $\mathrm{E}\left[\left|h'\left(X_{i^*}\right)\left(\log f\left(X_{i^*}\right)\right)'\right|\right]$ are bounded by K, say. Using Lemmas 6.2.1 and 6.4.2, we then conclude that the first term on the RHS goes to 0 as $d \to \infty$. We reach the

same conclusion for the second term by applying Lemma 6.2.2 along with Lemma 6.4.3.

5.3.3 Proof of Theorem 3.3.1

The proof of Theorem 3.3.1 is essentially the same as that of Theorem 3.2.1. In fact, all we need to do is to use Lemma 6.5.1 in place of Lemma 6.4.1, and to replace Lemmas 6.4.2 and 6.4.3 by Lemma 6.5.2.

5.4 Inhomogeneity and Extensions

In Chapter 4, we presented various extensions of the results introduced in Chapter 3. Since the proofs of these extensions can be carried in a similar fashion to those presented previously, we shall only outline the main differences.

5.4.1 Proofs of Theorems 4.1.1 and 4.1.2

The proof of Theorem 4.1.1 is almost identical to the proof of Theorem 3.2.1. For the discrete-time limit, it suffices to use Lemma 6.6.1 instead of Lemma 6.4.1 to achieve the desired conclusion. For the continuous-time limit, the proof also stays the same but a modification is needed in Lemmas 6.4.2 and 6.4.3. That is, the body of these proofs remains unchanged but to find the appropriate speed measure, we use the conditional distribution for $z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ developed in (6.20) rather than that developed in (6.15). Theorem 4.1.2 then follows by letting $E_R = 0$ in Theorem 4.1.1.

5.4.2 Proofs of Theorems 4.3.1 and 4.3.2

Most of the proofs is very similar to those of Theorems 3.1.1, 3.2.1, and 3.3.1. The main difference happens when working with any one of the m groups formed of infinitely many components. Since the constant terms are now random, we cannot factorize the scaling terms of components belonging to a same group. This difficulty is however easily overcome by changes of variables and the use of conditional expectations; for instance, a typical situation we face is

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

where $\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}$ is the vector containing the random variables $\{X_j, j\in\mathcal{J}\left(i,d\right)\}$ and similarly for $\boldsymbol{\Theta}^{(d)}_{\mathcal{J}(i,d)}$. Instead of carrying the term $\theta^2_{n+i}\left(d\right)=d^{\gamma_i}/K_{n+i}$, we thus carry $b_i d^{\gamma_i}$.

5.4.3 Proofs of Theorems 4.3.3 and 4.3.4

The general forms of the functions $c\left(\mathcal{J}\left(i,d\right)\right)$, $i=1,\ldots,m$ and $\theta_j\left(d\right)$, $j=1,\ldots,d$ necessitate a fancier notation, but do not affect the body of the proofs. What alters the demonstrations is rather the fact that the scaling terms $\theta_j\left(d\right)$ for $j\in\mathcal{J}\left(i,d\right)$ are allowed to be different functions of the dimension as long as they are of the

same order. Because of this particularity of the model, we have to write $\theta_j(d) = K_j^{-1/2}\theta_i'(d)\frac{\theta_j^*(d)}{\theta_i'(d)}$, where $\theta_j^*(d)$ is implicitly defined. We can then carry with the proofs as usual, factoring the term $b_i(\theta_i'(d))^2$ instead of $\theta_{n+i}^2(d)$ in Theorems 3.1.1 and 3.2.1 (or $b_i d^{\gamma_i}$ in Theorems 4.3.1 and 4.3.2). Since $\lim_{d\to\infty}\theta_j^*(d)/\theta_i'(d) = 1$ for $j \in \mathcal{J}(i,d)$, the rest of the proofs can be repeated with minor modifications.

Chapter 6

Weak Convergence of the Rescaled RWM Algorithm: Lemmas

This chapter presents and demonstrates the results used in the proofs of Chapter 5. In particular, Section 6.1 focuses on the asymptotically equivalent generators $\widetilde{G}h\left(d, X_{i^*}\right)$ and $\widehat{G}h\left(d, X_{i^*}\right)$; similarly, Section 6.2 introduces drift and volatility terms which are asymptotically equivalent to those of $\widetilde{G}h\left(d, X_{i^*}\right)$. The remaining sections aim to simplify the expressions for the drift and volatility terms found in Section 6.2, as well as for the acceptance rule of $\widehat{G}h\left(d, X_{i^*}\right)$. Sections 6.3, 6.4, and 6.5 are respectively associated with Theorems 3.1.1, 3.2.1, and 3.3.1, while Section 6.6 is related to Theorems 4.1.1 and 4.1.2 (where the target is assumed to be normally distributed).

6.1 Asymptotically Equivalent Generators

6.1.1 Approximation Term

The following lemma shall be of great use to prove the results presented in Section 6.1.3, as well as for the demonstrations of several of the subsequent lemmas.

Lemma 6.1.1. *For $i = 1, \ldots, m$, let*

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

where $Y_j\,|X_j \sim N\left(X_j, \ell^2/d^\alpha\right)$ and X_j, $j = 1, \ldots, d$ are independently distributed according to the density $\theta_j\left(d\right) f\left(\theta_j\left(d\right) x_j\right)$. Then for $i = 1, \ldots, m$

$$\mathrm{E}\left[\left|W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right] \to 0 \text{ as } d \to \infty.$$

Proof. By Jensen's inequality

$$\mathrm{E}_{\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}}\left[\left|W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right] \leq \sqrt{\mathrm{E}_{\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}}\left[W_i^2\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right]}.$$

Developing the square and computing the expectation conditional on $\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}$ yield

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}}\left[W_i^2\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)},\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right]=\frac{\ell^4}{4d^{2\alpha}}\times\\
&\quad\left\{3\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\right)^2+\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\right)^4\right.\\
&\quad+2\sum_{k\in\mathcal{J}(i,d)}\sum_{j=\mathcal{J}_{k+1}(i,d)}^{\mathcal{J}_{c(\mathcal{J}(i,d))}(i,d)}\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\frac{d^2}{dX_k^2}\log f\left(\theta_k\left(d\right)X_k\right)\\
&\quad+2\sum_{k\in\mathcal{J}(i,d)}\sum_{j=\mathcal{J}_{k+1}(i,d)}^{\mathcal{J}_{c(\mathcal{J}(i,d))}(i,d)}\left(\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\right)^2\left(\frac{d}{dX_k}\log f\left(\theta_k\left(d\right)X_k\right)\right)^2\\
&\quad\left.+2\sum_{k\in\mathcal{J}(i,d)}\sum_{j\in\mathcal{J}(i,d)}\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\left(\frac{d}{dX_k}\log f\left(\theta_k\left(d\right)X_k\right)\right)^2\right\}.
\end{aligned}$$

The previous expression can be reexpressed as

$$\begin{aligned}
\mathrm{E}_{\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}}\left[W_i^2\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)},\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right]&=\frac{\ell^4}{2d^{2\alpha}}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\right)^2\\
&\quad+\frac{\ell^4}{4d^{2\alpha}}\left\{\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)+\left(\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\right)^2\right)\right\}^2,
\end{aligned}$$

and hence

$$\begin{aligned}
&\sqrt{\mathrm{E}_{\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}}\left[W_i^2\left(d,\mathbf{X}^{(d)}_{\mathcal{J}(i,d)},\mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right]}\\
&\quad\leq\quad\frac{\ell^2}{\sqrt{2}d^{\alpha}}\left(\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\right)^2\right)^{1/2}\\
&\qquad+\frac{\ell^2}{2d^{\alpha}}\left|\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)+\left(\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\right)^2\right)\right|.
\end{aligned}$$

Using changes of variables, the unconditional expectation then satisfies

$$\begin{aligned}
&\mathrm{E}\left[\left|W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right] \\
&\quad\leq \frac{\ell^2}{\sqrt{2}d^{\alpha}}\theta^2_{n+i}(d)\sqrt{c(\mathcal{J}(i,d))}\mathrm{E}\left[\left(\frac{d^2}{dX^2}\log f(X)\right)^2\right]^{1/2} \\
&\qquad+\frac{\ell^2}{2d^{\alpha}}\theta^2_{n+i}(d)\, c(\mathcal{J}(i,d)) \\
&\qquad\times \mathrm{E}\left[\left|\frac{1}{c(\mathcal{J}(i,d))}\sum_{j\in\mathcal{J}(i,d)}\left(\frac{d^2}{dX_j^2}\log f(X_j)+\left(\frac{d}{dX_j}\log f(X_j)\right)^2\right)\right|\right].
\end{aligned}$$

Since $d^{\alpha} > d^{\gamma_i}\sqrt{c(\mathcal{J}(i,d))}$ and the first expectation on the RHS is bounded by some constant, the first term converges to 0 as $d\to\infty$. Given that the expression $\theta^2_{n+i}(d)\, c(\mathcal{J}(i,d))/d^{\alpha}$ is $O(1)$ for at least one $i\in\{1,\ldots,m\}$, we must also show that the second expectation on the RHS converges to 0. Because $(\log f(x))'$ is Lipschitz continuous, then Lemma A.1 affirms that $f'(x)\to 0$ as $x\to\pm\infty$. Consequently,

$$\mathrm{E}\left[\frac{d^2}{dX^2}\log f(X)+\left(\frac{d}{dX}\log f(X)\right)^2\right]=\mathrm{E}\left[\frac{f''(X)}{f(X)}\right]=\int f''(x)\,dx=0.$$

Since $\mathrm{Var}\left(\frac{f''(X)}{f(X)}\right)=\mathrm{E}\left[\left(\frac{f''(X)}{f(X)}\right)^2\right]<\infty$, then by the Weak Law of Large Numbers (WLLN) we find that

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

as $d\to\infty$.

We now want to verify if we could bring the limit inside the expectation. By

independence between the X_j's, we find

$$\begin{aligned} \mathrm{E}\left[\left(S_i\left(d\right)\right)^2\right] &= \mathrm{E}\left[\left(\frac{1}{c\left(\mathcal{J}\left(i,d\right)\right)}\sum_{j\in\mathcal{J}(i,d)}\frac{f''\left(X_j\right)}{f\left(X_j\right)}\right)^2\right] \\ &= \frac{1}{c\left(\mathcal{J}\left(i,d\right)\right)}\mathrm{E}\left[\left(\frac{f''\left(X\right)}{f\left(X\right)}\right)^2\right], \end{aligned}$$

which is finite for all d. Then, as $a \to \infty$,

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

Since the uniform integrability condition is satisfied (see, for instance, [10], [19] or [36]), we find $\lim_{d\to\infty}\mathrm{E}\left[\left|S_i\left(d\right)\right|\right] = \mathrm{E}\left[\lim_{d\to\infty}\left|S_i\left(d\right)\right|\right] = 0$, which completes the proof of the lemma. □

6.1.2 Continuous-Time Generator

We now introduce the generator $\widetilde{G}h\left(d, X_{i^*}\right)$, which is asymptotically equivalent to the generator in (5.6) for the cases where the RWM algorithm is sped up by a factor $d^\alpha > 1$. It is interesting to note that $\widetilde{G}h\left(d, X_{i^*}\right)$ is the generator of a Langevin diffusion process with a drift and volatility depending on d.

Lemma 6.1.2. *For any function $h \in C_c^\infty$, let*

$$\begin{aligned} \widetilde{G}h\left(d, X_{i^*}\right) = \frac{1}{2}\ell^2 h''\left(X_{i^*}\right)\mathrm{E}\left[1 \wedge e^{\sum_{j=1,j\neq i^*}^d \varepsilon(d,X_j,Y_j)}\right] \\ + \ell^2 h'\left(X_{i^*}\right)\left(\log f\left(X_{i^*}\right)\right)'\mathrm{E}\left[e^{\sum_{j=1,j\neq i^*}^d \varepsilon(d,X_j,Y_j)}; \sum_{j=1,j\neq i^*}^d \varepsilon\left(d, X_j, Y_j\right) < 0\right], \end{aligned} \tag{6.1}$$

where

$$\varepsilon\left(d, X_j, Y_j\right) = \log \frac{f\left(\theta_j\left(d\right) Y_j\right)}{f\left(\theta_j\left(d\right) X_j\right)}. \tag{6.2}$$

If $\alpha > 0$ *as defined in (2.7), then* $\lim_{d\to\infty} \mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - \widetilde{G}h\left(d, X_{i^*}\right)\right|\right] = 0.$

Proof. As mentioned in Section 5.1.3, the generator of the i^*-th component of the sped up RWM algorithm is given by

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

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

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

where $\varepsilon\left(X_{i^*}, Y_{i^*}\right) = \log \frac{f(Y_{i^*})}{f(X_{i^*})}$ and $\varepsilon\left(d, X_j, Y_j\right) = \log \frac{f(\theta_j(d)Y_j)}{f(\theta_j(d)X_j)}$. We can thus express the generator as

$$Gh\left(d, X_{i^*}\right) = d^{\alpha}\mathrm{E}_{Y_{i^*}}\left[\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right)\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{\sum_{j=1}^{d} \varepsilon(d, X_j, Y_j)}\right]\right]. \tag{6.3}$$

We shall compute the outside expectation. To this effect, a Taylor expansion of the minimum function with respect to Y_{i^*} and around X_{i^*} is used. Since f is a C^2 density function, the minimum function in (6.3) is twice differentiable as well except at the countable number of points where $\sum_{j=1}^{d} \varepsilon\left(d, X_j, Y_j\right) = 0$. However, this does not affect the value of the expectation since the set of values at which the derivatives do not exist has Lebesgue probability 0. The first and second derivatives of the minimum function are

$$\frac{\partial}{\partial Y_{i^*}} 1 \wedge e^{\sum_{j=1}^{d} \varepsilon(d, X_j, Y_j)} = \begin{cases} \frac{\partial}{\partial Y_{i^*}} \varepsilon\left(X_{i^*}, Y_{i^*}\right) e^{\sum_{j=1}^{d} \varepsilon(d, X_j, Y_j)} & \text{if } \sum_{j=1}^{d} \varepsilon\left(d, X_j, Y_j\right) < 0 \\ 0 & \text{if } \sum_{j=1}^{d} \varepsilon\left(d, X_j, Y_j\right) > 0 \end{cases},$$

and

$$\frac{\partial^2}{\partial Y_{i^*}^2} 1 \wedge e^{\sum_{j=1}^{d} \varepsilon(d, X_j, Y_j)} = \begin{cases} \left(\frac{\partial^2}{\partial Y_{i^*}^2} \varepsilon\left(X_{i^*}, Y_{i^*}\right) + \left(\frac{\partial}{\partial Y_{i^*}} \varepsilon\left(X_{i^*}, Y_{i^*}\right) \right)^2 \right) e^{\sum_{j=1}^{d} \varepsilon(d, X_j, Y_j)} & \text{if } \sum_{j=1}^{d} \varepsilon\left(d, X_j, Y_j\right) < 0 \\ 0 & \text{if } \sum_{j=1}^{d} \varepsilon\left(d, X_j, Y_j\right) > 0 \end{cases}.$$

Expressing the inner expectation in (6.3) as a function of these derivatives, we find

$$\begin{aligned} &\mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[1 \wedge e^{\sum_{j=1}^{d} \varepsilon(d, X_j, Y_j)} \right] = \mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[1 \wedge e^{\sum_{j=1, j \neq i^*}^{d} \varepsilon(d, X_j, Y_j)} \right] \\ &+ \left(Y_{i^*} - X_{i^*}\right) \left(\log f\left(X_{i^*}\right)\right)' \mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[e^{\sum_{j=1, j \neq i^*}^{d} \varepsilon(d, X_j, Y_j)}; \sum_{j=1, j \neq i^*}^{d} \varepsilon\left(d, X_j, Y_j\right) < 0 \right] \\ &+ \frac{1}{2} \left(Y_{i^*} - X_{i^*}\right)^2 \left(\left(\left(\log f\left(U_{i^*}\right)\right)' \right)^2 + \left(\log f\left(U_{i^*}\right)\right)'' \right) \mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[e^{g(U_{i^*})}; g\left(U_{i^*}\right) < 0 \right], \end{aligned}$$

where $g(U_{i^*}) = \varepsilon(X_{i^*}, U_{i^*}) + \sum_{j=1, j\neq i^*}^{d} \varepsilon(d, X_j, Y_j)$ for some $U_{i^*} \in (X_{i^*}, Y_{i^*})$ or (Y_{i^*}, X_{i^*}). Using this expansion, the generator becomes

$$
\begin{aligned}
& Gh(d, X_{i^*}) \\
= \; & d^{\alpha} \mathrm{E}_{Y_{i^*}} \left[(h(Y_{i^*}) - h(X_{i^*})) \right] \mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[1 \wedge e^{\sum_{j=1, j\neq i^*}^{d} \varepsilon(d, X_j, Y_j)} \right] \\
& + d^{\alpha} (\log f(X_{i^*}))' \mathrm{E}_{Y_{i^*}} \left[(h(Y_{i^*}) - h(X_{i^*})) (Y_{i^*} - X_{i^*}) \right] \\
& \times \mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[e^{\sum_{j=1, j\neq i^*}^{d} \varepsilon(d, X_j, Y_j)}; \sum_{j=1, j\neq i^*}^{d} \varepsilon(d, X_j, Y_j) < 0 \right] \\
& + \frac{d^{\alpha}}{2} \mathrm{E}_{Y_{i^*}} \Big[(h(Y_{i^*}) - h(X_{i^*})) (Y_{i^*} - X_{i^*})^2 \left((\log f(U_{i^*}))' \right)^2 \\
& \qquad \times \mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[e^{g(U_{i^*})}; g(U_{i^*}) < 0 \right] \Big] \\
& + \frac{d^{\alpha}}{2} \mathrm{E}_{Y_{i^*}} \Big[(h(Y_{i^*}) - h(X_{i^*})) (Y_{i^*} - X_{i^*})^2 (\log f(U_{i^*}))'' \\
& \qquad \times \mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[e^{g(U_{i^*})}; g(U_{i^*}) < 0 \right] \Big] .
\end{aligned}
$$

Expressing $h(Y_{i^*}) - h(X_{i^*})$ as a three-term Taylor's expansion, we obtain

$$
\begin{aligned}
h(Y_{i^*}) - h(X_{i^*}) &= h'(X_{i^*}) (Y_{i^*} - X_{i^*}) \\
&\quad + \frac{1}{2} h''(X_{i^*}) (Y_{i^*} - X_{i^*})^2 + \frac{1}{6} h'''(V_{i^*}) (Y_{i^*} - X_{i^*})^3 ,
\end{aligned} \tag{6.4}
$$

for some V_{i^*} lying between X_{i^*} and Y_{i^*}. Since the function h has compact support, then h itself and its derivatives are bounded in absolute value by some positive constant (say K), which gives

$$
d^{\alpha} \mathrm{E}_{Y_{i^*}} \left[h(Y_{i^*}) - h(X_{i^*}) \right] \; \leq \; \frac{\ell^2}{2} h''(X_{i^*}) + \frac{\ell^3}{6} \sqrt{\frac{8}{\pi}} \frac{K}{d^{\alpha/2}},
$$

along with

$$d^{\alpha}\mathrm{E}_{Y_{i^*}}\left[h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\left(Y_{i^*}-X_{i^*}\right)\right] \;\leq\; \ell^2 h'\left(X_{i^*}\right)+\frac{\ell^4}{2d^{\alpha}}K.$$

Substituting these expressions in the equation for $Gh\left(d,X_{i^*}\right)$, noticing that all expectations computed with respect to $\mathbf{Y}^{(d)-}$ and $\mathbf{X}^{(d)-}$ are bounded by one and that $\left|\left(\log f\left(U_{i^*}\right)\right)''\right|$ is bounded by some positive constant K, we obtain

$$\begin{aligned}\left|Gh\left(d,X_{i^*}\right)-\widetilde{G}h\left(d,X_{i^*}\right)\right| \leq{}& \frac{\ell^3}{6}\sqrt{\frac{8}{\pi}}\frac{K}{d^{\alpha/2}}+\frac{\ell^4}{2d^{\alpha}}K\left|\left(\log f\left(X_{i^*}\right)\right)'\right| \\ &+\frac{d^{\alpha}}{2}\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left(Y_{i^*}-X_{i^*}\right)^2\left(\left(\log f\left(U_{i^*}\right)\right)'\right)^2\right] \\ &+\frac{d^{\alpha}}{2}K\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left(Y_{i^*}-X_{i^*}\right)^2\right]. \qquad (6.5)\end{aligned}$$

Using a two-term Taylor expansion around X_{i^*}, we find

$$\begin{aligned}\left(\log f\left(U_{i^*}\right)\right)' &= \left(\log f\left(X_{i^*}\right)\right)'+\left(\log f\left(V_{i^*}\right)\right)''\left(U_{i^*}-X_{i^*}\right) \\ &\leq \left|\left(\log f\left(X_{i^*}\right)\right)'\right|+K\left|Y_{i^*}-X_{i^*}\right|,\end{aligned}$$

where $V_{i^*}\in\left(X_{i^*},U_{i^*}\right)$ or $\left(U_{i^*},X_{i^*}\right)$. Using (6.4), the expectation on the third line of (6.5) becomes

$$\begin{aligned}&d^{\alpha}\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left(Y_{i^*}-X_{i^*}\right)^2\left(\left(\log f\left(U_{i^*}\right)\right)'\right)^2\right] \\ &\leq\; d^{\alpha}\left(\left(\log f\left(X_{i^*}\right)\right)'\right)^2\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left(Y_{i^*}-X_{i^*}\right)^2\right] \\ &\quad+d^{\alpha}K\left|\left(\log f\left(X_{i^*}\right)\right)'\right|\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left|Y_{i^*}-X_{i^*}\right|^3\right] \\ &\quad+d^{\alpha}K\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left(Y_{i^*}-X_{i^*}\right)^4\right]\end{aligned}$$

for some $K > 0$. Again using (6.4), we first find

$$\begin{aligned} & d^{\alpha}\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left(Y_{i^*}-X_{i^*}\right)^2\right] \\ \leq\ & \frac{\ell^3}{d^{\alpha/2}}\sqrt{\frac{8}{\pi}}K+\frac{3\ell^4}{2d^{\alpha}}K+\frac{\ell^5}{d^{3\alpha/2}}\sqrt{\frac{32}{\pi}}\frac{K}{3}, \end{aligned}$$

then

$$\begin{aligned} & d^{\alpha}\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left|Y_{i^*}-X_{i^*}\right|^3\right] \\ \leq\ & 3\frac{\ell^4}{d^{\alpha}}K+\frac{\ell^5}{d^{3\alpha/2}}\sqrt{\frac{32}{\pi}}K+\frac{\ell^6}{d^{2\alpha}}\frac{5}{2}K, \end{aligned}$$

and finally

$$\begin{aligned} & d^{\alpha}\mathrm{E}_{Y_{i^*}}\left[\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\left(Y_{i^*}-X_{i^*}\right)^4\right] \\ \leq\ & 2\sqrt{\frac{32}{\pi}}\frac{\ell^5}{d^{3\alpha/2}}K+\frac{15}{2}\frac{\ell^6}{d^{2\alpha}}K+\sqrt{\frac{128}{\pi}}\frac{\ell^7}{d^{5\alpha/2}}K. \end{aligned}$$

We can then simplify (6.5) further and write

$$\begin{aligned} \left|Gh\left(d,X_{i^*}\right)-\widetilde{G}h\left(d,X_{i^*}\right)\right| \leq\ & K\left(\frac{\ell^3}{d^{\alpha/2}}+\frac{\ell^4}{d^{\alpha}}+\frac{\ell^5}{d^{3\alpha/2}}\right)\left(\left(\log f\left(X_{i^*}\right)\right)'\right)^2 \\ +\ & K\left(\frac{\ell^4}{d^{\alpha}}+\frac{\ell^5}{d^{3\alpha/2}}+\frac{\ell^6}{d^{2\alpha}}\right)\left(1+\left|\left(\log f\left(X_{i^*}\right)\right)'\right|\right)+K\frac{\ell^3}{d^{\alpha/2}}+K\frac{\ell^7}{d^{5\alpha/2}}, \end{aligned}$$

for some constant $K > 0$. By assumption $\mathrm{E}\left[\left(\left(\log f\left(X_{i^*}\right)\right)'\right)^2\right] < \infty$, so it follows that $\mathrm{E}\left[\left|Gh\left(d,X_{i^*}\right)-\widetilde{G}h\left(d,X_{i^*}\right)\right|\right]$ converges to 0 as $d \to \infty$. □

6.1.3 Discrete-Time Generator

The discrete-time generator $\widehat{G}h\left(d, X_{i^*}\right)$ introduced in this section shall reveal a good asymptotic approximation of the generator in (5.6), but only when the RWM algorithm does not require a speed-up time factor.

Lemma 6.1.3. *For any function $h \in \overline{C}$, the space of bounded and continuous functions on $\mathbf{R}$, let*

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

with

$$z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right) = \sum_{j=1}^{n} \varepsilon\left(d, X_j, Y_j\right) + \sum_{i=1}^{m} \sum_{j \in \mathcal{J}(i,d)} \frac{d}{dX_j} \log f\left(\theta_j\left(d\right) X_j\right)\left(Y_j - X_j\right) - \frac{\ell^2}{2} \sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right). \tag{6.6}$$

Here, $\varepsilon\left(d, X_j, Y_j\right)$ is as in (6.2) and for $i = 1, \ldots, m$, we let

$$R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right) = \frac{1}{d^{\alpha}} \sum_{j \in \mathcal{J}(i,d)} \left(\frac{d}{dX_j} \log f\left(\theta_j\left(d\right) X_j\right)\right)^2, \tag{6.7}$$

where $\mathbf{X}^{(d)}_{\mathcal{J}(i,d)}$ is the vector containing the random variables $\{X_j, j \in \mathcal{J}\left(i, d\right)\}$. If $\alpha = 0$, then we have $\lim_{d \to \infty} \mathrm{E}\left[\left|Gh\left(d, X_{i^*}\right) - \widehat{G}h\left(d, X_{i^*}\right)\right|\right] = 0$.

Proof. As mentioned in the previous lemma, the generator of the process considered satisfies

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

and we can express the inner expectation as

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

Using a three-term Taylor expansion to express the difference of the log functions, we obtain

$$\begin{aligned}\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\frac{\pi\left(d,\mathbf{Y}^{(d)}\right)}{\pi\left(d,\mathbf{X}^{(d)}\right)}\right]&=\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1\wedge\exp\left\{\sum_{j=1}^{n}\varepsilon\left(d,X_j,Y_j\right)\right.\right.\\&+\sum_{i=1}^{m}\sum_{j\in\mathcal{J}(i,d)}^{d}\left[\frac{d}{dX_j}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)\right.\\&\left.\left.\left.+\frac{1}{2}\frac{d^2}{dX_j^2}\log f\left(\theta_j\left(d\right)X_j\right)\left(Y_j-X_j\right)^2+\frac{1}{6}\frac{d^3}{dU_j^3}\log f\left(\theta_j\left(d\right)U_j\right)\left(Y_j-X_j\right)^3\right]\right\}\right].\end{aligned}\tag{6.8}$$

for some $U_j\in(X_j,Y_j)$ or (Y_j,X_j). Note that to compare the terms in the exponential function with $z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$, we have also grouped the last $d-n$ components according to their scaling term.

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

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

By the Lipschitz property of the function $1 \wedge e^x$ (see Proposition A.4 in the Appendix) and noticing that the first two terms of the function $z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ cancel out with the first two terms of the exponential term in (6.8), we obtain

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

Noticing that the first double summation consists in the random variables $W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)$'s of Lemma 6.1.1, we find

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\left\{1 \wedge \exp\left(\log \frac{\pi\left(d, \mathbf{Y}^{(d)}\right)}{\pi\left(d, \mathbf{X}^{(d)}\right)}\right)\right\} - \left\{1 \wedge e^{z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)}\right\}\right|\right] \\
&\leq \mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\sum_{i=1}^{m} W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right] \\
&\quad + \frac{1}{6}\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[\left|\sum_{i=1}^{m} \sum_{j \in \mathcal{J}(i,d)} \frac{d^3}{dU_j^3} \log f\left(\theta_j\left(d\right) U_j\right)\left(Y_j - X_j\right)^3\right|\right] \\
&\leq \sum_{i=1}^{m} \mathrm{E}\left[\left|W_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}, \mathbf{Y}^{(d)}_{\mathcal{J}(i,d)}\right)\right|\right] + \sum_{i=1}^{m} c\left(\mathcal{J}\left(i,d\right)\right) \ell^3 K \frac{d^{3\gamma_i/2}}{d^{3\alpha/2}} \qquad (6.9)
\end{aligned}$$

for some constant $K > 0$.

We are now ready to verify $\mathcal{L}^1$ convergence of the generators. By the triangle's

inequality and (6.9), we observe

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

Because $h\in\overline{C}$, there exists a constant such that $\left|h\left(Y_{i^*}\right)-h\left(X_{i^*}\right)\right|\leq K$ and thus Lemma 6.1.1 implies that the previous expression converges to 0 as $d\to\infty$. □

6.2 The Asymptotically Equivalent Continuous-Time Generator

6.2.1 Asymptotically Equivalent Volatility

The goal of the following result is to replace the volatility term of the generator $\widetilde{G}h\left(d,X_{i^*}\right)$ by an asymptotically equivalent, but more convenient expression.

Lemma 6.2.1. *If $\alpha>0$, we have*

$$\lim_{d\to\infty}\left|\mathrm{E}\left[1\wedge e^{\sum_{j=1,j\neq i^*}^{d}\varepsilon(d,X_j,Y_j)}\right]-\mathrm{E}\left[1\wedge e^{z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)}\right]\right|=0,$$

where $\varepsilon(d, X_j, Y_j)$ *is as in (6.2) and*

$$z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right) = \sum_{j=1, j\neq i^*}^{n} \varepsilon(d, X_j, Y_j) \tag{6.10}$$
$$+ \sum_{i=1}^{m} \sum_{j\in\mathcal{J}(i,d), j\neq i^*} \frac{d}{dX_j} \log f\left(\theta_j(d) X_j\right)(Y_j - X_j) - \frac{\ell^2}{2} \sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right),$$

where $R_i\left(d, \mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)$ *is as in (6.7), except that the* i^**-th component is explicitly excluded from the sum. That is,* $z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)$ *is the version of (6.6) where* X_{i^*} *is omitted.*

Proof. The proof is basically the same as the proof of the previous lemma with $h = 1$, the only difference lying in the fact that the i^*-th component is now omitted. □

6.2.2 Asymptotically Equivalent Drift

The following result aims to replace the drift term of $\widetilde{G}h(d, X_{i^*})$ in Lemma 6.1.2 by an asymptotically equivalent, but more convenient expression.

Lemma 6.2.2. *We have*

$$\lim_{d\to\infty} \left| \mathrm{E}\left[e^{\sum_{j=1, j\neq i^*}^{d} \varepsilon(d, X_j, Y_j)}; \sum_{j=1, j\neq i^*}^{d} \varepsilon(d, X_j, Y_j) < 0 \right] \right.$$
$$\left. - \mathrm{E}\left[e^{z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)}; z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right) < 0 \right] \right| = 0, \tag{6.11}$$

where the functions $\varepsilon(d, X_j, Y_j)$ *and* $z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)$ *are as in (6.2) and (6.10) respectively.*

Proof. First, let

$$T(x) = e^x \mathbf{1}_{(x<0)} = \begin{cases} e^x, & x < 0 \\ 0, & x \geq 0 \end{cases}.$$

It is important to realize that the function $T(x)$ is not Lipschitz, which keeps us from reproducing the proof of Lemma 6.2.1. Now, let

$$A\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right) = T\left(\sum_{j=1, j\neq i^*}^{d} \varepsilon\left(d, X_j, Y_j\right)\right) - T\left(z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)\right)$$

and

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

We shall show that $A\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right) \rightarrow_p 0$, and then use this result to prove convergence of expectations.

In order to simplify the expressions involved in the following development, we shall omit the arguments $\mathbf{Y}^{(d)-}$ and $\mathbf{X}^{(d)-}$ in the functions $\varepsilon(\cdot)$, $z(\cdot)$ and $A(\cdot)$. We have

$$\begin{aligned} \mathrm{P}\left(|A(d)| \geq \delta(d)\right) &= \mathrm{P}\left(|A(d)| \geq \delta(d); \sum \varepsilon(d) \geq 0; z(d) \geq 0\right) \\ &\quad + \mathrm{P}\left(|A(d)| \geq \delta(d); \sum \varepsilon(d) < 0; z(d) < 0\right) \\ &\quad + \mathrm{P}\left(|A(d)| \geq \delta(d); \sum \varepsilon(d) \geq 0; z(d) < 0\right) \\ &\quad + \mathrm{P}\left(|A(d)| \geq \delta(d); \sum \varepsilon(d) < 0; z(d) \geq 0\right). \end{aligned}$$

We can bound the third term on the RHS by

$$\begin{aligned}
&\mathrm{P}\left(|A(d)| \geq \delta(d); \sum \varepsilon(d) \geq 0; z(d) < 0\right) \\
&\leq \mathrm{P}\left(\sum \varepsilon(d) \geq 0; z(d) < 0; \left|\sum \varepsilon(d) - z(d)\right| < \delta(d)\right) \\
&\quad + \mathrm{P}\left(\sum \varepsilon(d) \geq 0; z(d) < 0; \left|\sum \varepsilon(d) - z(d)\right| \geq \delta(d)\right),
\end{aligned}$$

and similarly for the fourth term. Also note that if $\sum \varepsilon(d) \geq 0$ and $z(d) \geq 0$, or $\sum \varepsilon(d) < 0$ and $z(d) < 0$, then $|A(d)| \leq |\sum \varepsilon(d) - z(d)|$. Therefore,

$$\begin{aligned}
\mathrm{P}(|A(d)| \geq \delta(d)) \leq\ & \mathrm{P}\left(\left|\sum \varepsilon(d) - z(d)\right| \geq \delta(d)\right) \\
& + \mathrm{P}\left(\sum \varepsilon(d) \geq 0; z(d) < 0; \left|\sum \varepsilon(d) - z(d)\right| < \delta(d)\right) \\
& + \mathrm{P}\left(\sum \varepsilon(d) < 0; z(d) \geq 0; \left|\sum \varepsilon(d) - z(d)\right| < \delta(d)\right).
\end{aligned}$$

Since $\sum \varepsilon(d)$ and $z(d)$ are of different sign but the difference between them must be less than $\delta(d)$, we can bound the last two terms and obtain

$$\begin{aligned}
&\mathrm{P}(|A(d)| \geq \delta(d)) \\
&\leq \mathrm{P}\left(\left|\sum \varepsilon(d) - z(d)\right| \geq \delta(d)\right) \\
&\quad + \mathrm{P}(-\delta(d) < z(d) < 0) + \mathrm{P}(0 \leq z(d) < \delta(d)) \\
&= \mathrm{P}\left(\left|\sum \varepsilon(d) - z(d)\right| \geq \delta(d)\right) + \mathrm{P}(-\delta(d) < z(d) < \delta(d)). \qquad (6.12)
\end{aligned}$$

By Markov's inequality and the proof of Lemma 6.1.2, then as $d \to \infty$ the first

term on the RHS of (6.12) satisfies

$$\begin{aligned}
&\mathrm{P}\left(\left|\sum_{j=1,j\neq i^*}^{d}\varepsilon\left(d,X_j,Y_j\right)-z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\right|\geq\delta\left(d\right)\right)\\
&\quad\leq\frac{1}{\delta\left(d\right)}\mathrm{E}\left[\left|\sum_{j=1,j\neq i^*}^{d}\varepsilon\left(d,X_j,Y_j\right)-z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\right|\right]\leq\sqrt{\delta\left(d\right)}\rightarrow 0.
\end{aligned}$$

Now consider the second term on the RHS of (6.12). From the proof of Lemma 6.3.1, we know the distribution of $z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\left|\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}\right.$. Using conditional theory, we have

$$\begin{aligned}
&\mathrm{P}\left(\left|z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\right|<\delta\left(d\right)\right)\\
&\quad=\ \mathrm{E}_{\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}}\left[\mathrm{P}_{\mathbf{Y}^{(d-n)-}}\left(\left|z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\right|<\delta\left(d\right)\right)\right].
\end{aligned}$$

Focusing on the conditional probability, we write

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

Using the convergence results developed in the proof of Lemma 6.3.1 along with the fact that $\delta\left(d\right)\rightarrow 0$ as $d\rightarrow\infty$, we deduce that

$$\mathrm{P}_{\mathbf{Y}^{(d-n)-}}\left(\left|z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\right|<\delta\left(d\right)\right)\rightarrow 0.$$

Using the Bounded Convergence Theorem, we then find that the unconditional probability $\mathrm{P}\left(\left|z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)\right| < \delta\left(d\right)\right)$ converges to 0 as well. Since we showed that $\mathrm{P}\left(\left|A\left(d\right)\right| \geq \delta\left(d\right)\right) \to 0$ as $d \to \infty$, then we can use the Bounded Convergence Theorem to verify (6.11). □

6.3 Volatility and Drift for the Familiar Limit

We now simplify the expressions for the drift and volatility obtained in the previous section. This shall, by the same fact, allow us to determine the speed measure of the limiting Langevin diffusion process.

6.3.1 Simplified Volatility

Lemma 6.3.1. *If Condition (3.1) is satisfied, then*

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

where $z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)$ *and* E_R *are as in (6.10) and (3.3) respectively.*

Proof. Making use of conditioning allows us to write

$$\begin{aligned}&\mathrm{E}\left[1 \wedge \exp\left(z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)\right)\right]\\ &= \mathrm{E}_{\mathbf{Y}^{(n)-}, \mathbf{X}^{(d)-}}\left[\mathrm{E}_{\mathbf{Y}^{(d-n)-}}\left[1 \wedge \exp\left(z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)\right)\right]\right]. \qquad (6.13)\end{aligned}$$

To solve the inner expectation, we need to find the distribution of the random variable $z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)\left|\mathbf{Y}^{(n)-}, \mathbf{X}^{(d)-}\right.$. Since $\left(Y_j - X_j\right)|X_j$, $j = 1, \cdots, d$, are

iid and normally distributed with mean 0 and variance ℓ^2/d^α, then

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

Applying Proposition A.5 in the Appendix allows us to express the inner expectation in (6.13) in terms of $\Phi(\cdot)$, the *cdf* of a standard normal random variable

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

We are then left to evaluate the expectation of $M(\cdot)$. Again using conditional expectations, we find

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

From Proposition A.6, we find that both terms included in the function $M\left(d, \mathbf{Y}^{(n)-}, \mathbf{X}^{(d)-}\right)$ have the same inner expectation. The unconditional expec-

tation thus simplifies to

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

We now find the limit of the term inside the function $\Phi(\cdot)$ as $d \to \infty$. From Proposition A.7, we have that $\varepsilon(d, X_j, Y_j)$ converges in probability to 0 for all $j \in \{1, \ldots, n\}$ but excluding $j = i^*$. Similarly, we use Proposition A.8 to conclude that $\sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right) \to_p E_R$. Furthermore, $E_R > 0$ since there exists at least one $i \in \{1, \ldots, m\}$ such that $\lim_{d\to\infty} c\left(\mathcal{J}(i,d)\right) d^{\gamma_i}/d^{\alpha} > 0$.

By applying Slutsky's Theorem, the Continuous Mapping Theorem and by recalling that convergence in probability and convergence in distribution are equivalent when the limit is a constant, we conclude that

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

Since $\Phi(\cdot)$ is positive and bounded by 1, we finally use the Bounded Conver-

gence Theorem to find

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

which completes the proof of the lemma. □

6.3.2 Simplified Drift

Lemma 6.3.2. *If Condition (3.1) is satisfied, then*

$$\lim_{d\to\infty}\left|\mathrm{E}\left[e^{z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)};z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)<0\right]-\Phi\left(-\frac{\ell\sqrt{E_R}}{2}\right)\right|=0,$$

where the functions $\varepsilon\left(d,X_j,Y_j\right)$ and $z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)$ are as in (6.2) and (6.10) respectively.

Proof. The proof is similar to that of Lemma 6.3.1 and for this reason, we just outline the differences. We know the distribution of $z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\left|\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}\right.$

from the proof of Lemma 6.3.1, so we can use Proposition A.5 to obtain

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

Applying Proposition A.6, we find

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

Repeating the last part of the proof of Lemma 6.3.1 completes the demonstration of the present lemma. □

6.4 Acceptance Rule, Volatility and Drift for the New Limit

6.4.1 Modified Acceptance Rule

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

Lemma 6.4.1. *If $\alpha = 0$ and Conditions (3.4) and (3.5) are satisfied, then*

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

as $d \to \infty$, with $z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ as in (6.6) and $\alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right)$ as in (3.6).

Proof. We first use conditional expectations to obtain

$$\begin{aligned}
&\mathrm{E}_{\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}}\left[1 \wedge e^{z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right] \\
&\quad = \mathrm{E}_{\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}}\left[\mathrm{E}_{\mathbf{Y}^{(d-n)}}\left[1 \wedge e^{z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)}\right]\right]. \qquad (6.14)
\end{aligned}$$

In order to evaluate the inner expectation, we need to find the distribution of the function $z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ conditional on $\mathbf{Y}^{(n)}$ and $\mathbf{X}^{(d)}$. From the proof of Lemma 6.3.1 and from the fact that $\alpha = 0$, we have

$$\begin{aligned}
&z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right) \Big| \mathbf{Y}^{(n)}, \mathbf{X}^{(d)} \sim \\
&N\left(\sum_{j=1}^{n} \varepsilon\left(d, X_j, Y_j\right) - \frac{\ell^2}{2}\sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right),\ \ell^2 \sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right)\right). \qquad (6.15)
\end{aligned}$$

Applying Proposition A.5, we can express the inner expectation in (6.14) in

terms of $\Phi(\cdot)$, the *cdf* of a standard normal random variable

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

We need to study the convergence of every term included in previous function. Condition (3.4) implies that $\theta_1^{-2}(d)$ is the asymptotically smallest scaling term and along with $\alpha = 0$, this means that the fastest converging component has an $O(1)$ scaling term. However there might be a finite number of other components also having an $O(1)$ scaling term. Recall that b is the number of such components and in the present case is defined as $b = \max\left(j \in \{1,\ldots,n\}; \lambda_j = \lambda_1 = 0\right)$. It is thus pointless to study the convergence of these b variables since they are independent of d. However, we can study the convergence of the other $n-b$ components and from Proposition A.7 we know that $\varepsilon\left(d,X_j,Y_j\right) \to_p 0$ for $j = b+1,\ldots,n$ since $\lambda_j < 0$. Similarly, we can use Proposition A.8 and Condition (3.5) to conclude that $\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right) \to_p E_R > 0$, with E_R as in (3.8).

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

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

Using the triangle's inequality, we obtain

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

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

6.4.2 Simplified Volatility

Lemma 6.2.1 established that $\ell^2\mathrm{E}\left[1 \wedge e^{z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)}\right]$ is an asymptotically valid volatility term for the continuous-time generator $\widetilde{G}h\left(d,X_{i^*}\right)$. We now wish to find a convenient expression for this volatility term as $d \rightarrow \infty$. This is achieved in the following lemma.

Lemma 6.4.2. *If Conditions (3.4) and (3.5) are satisfied, then*

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

where $z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)$ *is as in (6.10) and* E_R *is as in (3.8).*

Proof. From the proof of Lemma 6.3.1, we have

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

By Proposition A.7, we have $\varepsilon\left(d,X_j,Y_j\right) \rightarrow_p 0$ since $\lambda_j < \lambda_1$ for $j = b+1,\ldots,n$. From Proposition A.8, we also know that $\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right) \rightarrow_p E_R$, where E_R is as in (3.8) and is strictly positive by Condition (3.5). Applying Slutsky's and the Continuous Mapping Theorems thus yields

$$\begin{aligned}
\Phi\left(\frac{\sum_{j=1,j\neq i^*}^{n}\varepsilon\left(d,X_j,Y_j\right)-\frac{\ell^2}{2}\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)}{\sqrt{\ell^2\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)}}\right) &\rightarrow_p \\
\Phi\left(\frac{\sum_{j=1,j\neq i^*}^{b}\varepsilon\left(X_j,Y_j\right)-\frac{\ell^2}{2}E_R}{\sqrt{\ell^2 E_R}}\right). &\qquad (6.16)
\end{aligned}$$

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

6.4.3 Simplified Drift

Lemma 6.2.2 introduced a drift term that is asymptotically equivalent to the drift term of the continuous-time generator $\widetilde{G}h\left(d,X_{i^*}\right)$ in (6.1). The goal of the following lemma is to determine a simple expression for this new drift term as $d \rightarrow \infty$.

Lemma 6.4.3. *If Conditions (3.4) and (3.5) are satisfied, then*

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

where $z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)$ *and* E_R *are as in (6.10) and (3.8) respectively.*

Proof. The proof of this result is similar to that of Lemma 6.4.2 and for this reason, we shall not repeat every detail. Since we know the conditional distribution of $z\left(d,\mathbf{Y}^{(d)-},\mathbf{X}^{(d)-}\right)\left|\mathbf{Y}^{(n)-},\mathbf{X}^{(d)-}\right.$, we can use Proposition A.5 to obtain

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

From Proposition A.9, the unconditional expectation simplifies to

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

Using (6.16) along with the Bounded Convergence Theorem completes the proof of the lemma. □

6.5 Acceptance Rule, Volatility and Drift for the Limit with Unbounded Speed

This section determines the acceptance rule of the limiting Metropolis-Hastings algorithm, as well as the drift and volatility terms of the Langevin diffusion process in the case where the first b components are the only ones to govern the proposal variance.

6.5.1 Modified Acceptance Rule

Lemma 6.5.1. *If $\alpha = 0$ and Conditions (3.4) and (3.9) are satisfied, then*

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

as $d \to \infty$, with $z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right)$ as in (6.6) and $\alpha\left(X_{i^*},Y_{i^*}\right)$ as in (3.10).

Proof. We first note that $z\left(d,\mathbf{Y}^{(d)},\mathbf{X}^{(d)}\right) \to_p \sum_{j=1}^{b} \varepsilon\left(X_j,Y_j\right)$, since

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

the inequality has been obtained by applying Chebychev's inequality and the last equality has been obtained from the conditional distribution in (6.15).

Since $1 \wedge e^{z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)}$ is a continuous and bounded function, we then conclude that $\mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}}\left[1 \wedge e^{z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)}\right] \rightarrow_p \mathrm{E}_{\mathbf{Y}^{(b)-}, \mathbf{X}^{(b)-}}\left[1 \wedge \prod_{j=1}^{b} \frac{f(\theta_j Y_j)}{f(\theta_j X_j)}\right]$. □

6.5.2 Simplified Volatility and Drift

Lemma 6.5.2. *If Conditions (3.4) and (3.9) are satisfied, then*

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

where $z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)$ *is as in (6.10) and* E_R *is as in (3.8). Furthermore,*

$$\lim_{d \rightarrow \infty} \left| \mathrm{E}\left[e^{z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)}; z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right) < 0\right] - \mathrm{P}\left(\prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)} > 1\right) \right| = 0.$$

Proof. It suffices to use the fact that

$$\mathrm{E}\left[1 \wedge e^{z\left(d, \mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}\right)}\right] \rightarrow_p \mathrm{E}_{\mathbf{Y}^{(b)}, \mathbf{X}^{(b)}}\left[1 \wedge \prod_{j=1}^{b} \frac{f\left(\theta_j Y_j\right)}{f\left(\theta_j X_j\right)}\right]$$

along with the following decomposition

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

and Proposition A.9 to conclude that

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

The approach is the same for the second limit. □

6.6 Acceptance Rule for Normal Targets

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

Lemma 6.6.1. *If* $f(x) = (2\pi)^{-1/2} \exp(x^2/2)$, $\alpha = 0$ *and if Conditions (3.4) and (3.5) are satisfied, then*

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

with $z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)$ *as in (6.6) and* $\alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right)$ *as in (4.1).*

Proof. We first use conditional expectations to obtain

$$\begin{aligned} & \mathrm{E}_{\mathbf{Y}^{(d)-}, \mathbf{X}^{(d)-}} \left[1 \wedge e^{z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)} \right] \\ = \; & \mathrm{E}_{\mathbf{Y}^{(n)-}, \mathbf{X}^{(d-n)}} \left[\mathrm{E}_{\mathbf{Y}^{(d-n)}, \mathbf{X}^{(n)-}} \left[1 \wedge e^{z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right)} \right] \right]. \end{aligned} \tag{6.17}$$

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

$$\begin{aligned} & \left(\sum_{i=1}^{m} \sum_{j \in \mathcal{J}(i,d)} \frac{d}{dX_j} \log f\left(\theta_j(d) X_j\right)\left(Y_j - X_j\right) - \frac{\ell^2}{2} \sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right) \right) \Bigg| \mathbf{X}^{(d-n)} \\ \sim \; & N\left(-\frac{\ell^2}{2} \sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right), \; \ell^2 \sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)}_{\mathcal{J}(i,d)}\right) \right). \end{aligned} \tag{6.18}$$

Due to the normal form of the target components, it is possible to find the distribu-

tion of $\sum_{j=1,j\neq i^*}^{n} \varepsilon\left(d, X_j, Y_j\right)$. Since $Y_j \left| X_j \right. \sim N\left(X_j, \ell^2\right)$ and $X_j \sim N\left(0, K_j/d^{\lambda_j}\right)$, we obtain $Y_j = X_j + U_j$, where $U_j \sim N\left(0, \ell^2\right)$ is independent of X_j for $j = 1, \ldots, n$. We then have

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

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

$$\left(\frac{\ell^2}{2}\widetilde{U}_j^2 + \ell \widetilde{U}_j \widetilde{X}_j\right) \left| \widetilde{U}_j \right. \sim N\left(\frac{\ell^2}{2}\widetilde{U}_j^2, \ell^2 \widetilde{U}_j^2\right). \tag{6.19}$$

Combining (6.18) and (6.19), we obtain the conditional distribution

$$\begin{aligned} & z\left(d, \mathbf{Y}^{(d)}, \mathbf{X}^{(d)}\right) \left| \mathbf{Y}^{(n)}, X_{i^*}, \mathbf{X}^{(d-n)} \right. \\ & \sim N\left(\varepsilon\left(X_{i^*}, Y_{i^*}\right) - \frac{\ell^2}{2}\left(\sum_{j=1,j\neq i^*}^{n} \widetilde{U}_j^2 + \sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right)\right), \right. \\ & \qquad \left. \ell^2\left(\sum_{j=1,j\neq i^*}^{n} \widetilde{U}_j^2 + \sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right)\right)\right). \end{aligned} \tag{6.20}$$

Applying Proposition A.5, we can express the inner expectation in (6.17) in

terms of $\Phi(\cdot)$

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

We need to study the convergence of every term appearing in the preceding equation. From Proposition A.8, we know that $\sum_{i=1}^{m}R_i\left(d,\mathbf{X}_{\mathcal{J}(i,d)}^{(d)}\right) \to_p E_R$. Because the variance of the components $\widetilde{U}_1,\ldots,\widetilde{U}_b$ does not vary with the dimension of the target, it is not relevant to talk about convergence for these variables. We can however study the convergence of the components $\widetilde{U}_{b+1},\ldots,\widetilde{U}_n$. Using Chebychev's inequality, we have for all $\epsilon > 0$

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

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

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

variables with 1 degree of freedom, we conclude that

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

Using the triangle's inequality, we therefore obtain

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

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

Chapter 7

Weak Convergence of Stochastic Processes

The theorems and proofs presented in this thesis are based on a specific result providing general conditions for a sequence of processes to converge weakly (in the Skorokhod topology) to some Markov process. This result (Theorem 8.2 of Chapter 4 in [16]) exploits the fact that Markov processes can be expressed according to a corresponding operator semigroup, and can then be characterized by the generator of this semigroup. The present chapter thus aims to clarify these concepts and to justify why this particular result is applicable in our situation.

We first introduce the basics about operator semigroups, and relate them to Markov processes. We then discuss some notions about weak convergence of probability measures, before studying the details of the key convergence result. We conclude this chapter with the concept of core, which greatly simplifies the proofs presented in Chapters 5 and 6. Most of the results presented in this chapter can be found in [16]. Since there is quite a lot of notation involved, we found it simpler

to use the same notation as that of the book for introducing the following results. We shall however relate the theory to our specific situation when required.

7.1 Skorokhod Topology

The space $D_E[0,\infty)$ is defined to be the space of functions $x : [0,\infty) \to E$ admitting discontinuities of the first kind. A function x is said to have a discontinuity of the first kind at t if $x(t-)$ and $x(t+)$ exist but differ and $x(t)$ lies between them. It has become conventional to assume, when this can be done without altering the finite-dimensional distributions, that sample paths of this type of stochastic processes are right continuous with left limits (or càdlàg). Therefore, for each $t \geq 0$ and each function x in $D_E[0,\infty)$, $\lim_{x\to t+} x(s) = x(t)$ and $\lim_{x\to t-} x(s) \equiv x(t-)$ exists. Note that the process $\left\{ Z_{i^*}^{(d)}(t), t \geq 0 \right\}$, which admits a countable number of discontinuities and is right continuous, perfectly fits into this setting.

The availability of complete and separable metric spaces is an asset when dealing with results about convergence of probability measures. Therefore, it would be appropriate to define a metric on $D_E[0,\infty)$ that satisfies these two characteristics. Let (E,r) denote a metric space; it is possible to define a metric on $D_E[0,\infty)$ (say ρ) under which this metric space is separable if E is separable, and complete if (E,r) is complete (for more details, see [16] on pages 116-118).

The topology induced on $D_E[0,\infty)$ by the metric ρ is called the Skorokhod topology, in which the weak convergence results of this thesis have been proven. The particularity of this topology is that besides admitting a small perturbation of the ordinates (as for the uniform topology), it also allows a small deformation of the time scale. This turns out to be quite useful, as this allows us to transform our

discrete-time RWM algorithm into a continuous-time one, and to use the latter to derive weak convergence results.

7.2 Operator Semigroups

In this section, definitions and basic properties related to operator semigroups, which provide a primary tool in the study of Markov processes, are introduced. Bear in mind that L denotes a real Banach space with norm $\|\cdot\|$.

Definition 7.2.1. *A one-parameter family* $\{T(t); t \geq 0\}$ *of bounded linear operators on a Banach space* L *is called a semigroup if*

(a) $T(0) = I$*, the identity matrix;*

(b) $T(s+t) = T(t)\,T(s)$ *for all* $s, t \geq 0$.

A semigroup is thus an operator varying with t. Two desirable properties of semigroups are to be a strongly continuous and a contraction semigroup.

Definition 7.2.2. *A semigroup* $\{T(t)\}$ *on* L *is said to be strongly continuous if*

$$\lim_{t \to 0} T(t)\,h = h, \quad \forall h \in L.$$

Definition 7.2.3. *A contraction semigroup is such that* $\|T(t)\| \leq 1$ *for all* $t \geq 0$*, where*

$$\|T(t)\| = \sup_{h \in L,\ h \neq 0} \frac{\|T(t)\,h\|}{\|h\|}$$

is the operator norm.

The generator of the semigroup can sometimes be used to determine a semigroup on L; it is defined as follows.

Definition 7.2.4. *The (infinitesimal) generator of a semigroup $\{T(t)\}$ on L is the linear operator A defined by*

$$Ah = \lim_{t \to 0} \left(\frac{T(t)h - h}{t} \right).$$

The domain $\mathcal{D}(A)$ of A is the subspace of all $h \in L$ for which this limit exists.

The next proposition says that strongly continuous contraction semigroups are uniquely determined by their corresponding generator.

Proposition 7.2.5. *Let $\{T(t)\}$ and $\{S(t)\}$ be strongly continuous contraction semigroups on L with generators A and B, respectively. If $A = B$, then $T(t) = S(t)$ for all $t \geq 0$.*

Proof. The proof of this proposition can be found in [16], on page 15. □

The graph of a linear operator A is given by $\mathcal{G}(A) = \{(h, Ah) : h \in \mathcal{D}(A)\} \subset L \times L$. We say that A is single-valued since $(0, g) \in \mathcal{G}(A)$ implies $g = 0$. More generally, we can define a multivalued operator on L as an arbitrary subset A of $L \times L$ with domain $\mathcal{D}(A) = \{h : (h, g) \in A \text{ for some } g\}$ and range $\mathcal{R}(A) = \{g : (h, g) \in A \text{ for some } h\}$. This concept allows us to extend the notion of generator of a semigroup, and to define the full generator $\hat{A}$ of a measurable contraction semigroup $\{T(t)\}$ on L by

$$\hat{A} = \left\{ (h, g) \in L \times L : T(t)h - h = \int_0^t T(s)g ds, \quad t \geq 0 \right\}. \tag{7.1}$$

We shall see next section that the full generator $\hat{A}$ is a useful way to characterize Markov processes.

7.3 Markov Processes and Semigroups

Basic theory about time-homogeneous continuous-time Markov processes on general state spaces is now introduced, and related to operator semigroups.

Consider the metric space (E, r). The collection of all real-valued, Borel measurable functions on E is denoted $M(E)$. The Banach space of bounded functions with norm $\|h\| = \sup_{x \in E} |h(x)|$ is $B(E) \subset M(E)$. Finally, $\overline{C}(E) \subset B(E)$ is the subspace of bounded continuous functions.

Let $\{X(t), t \geq 0\}$ be a stochastic process defined on a probability space $(\Omega, \mathcal{F}, P)$ with values in E, and let $\mathcal{F}^X(t) = \sigma(X(s); s \leq t)$. Then, X is a Markov process if

$$\mathrm{P}\left(X(t+s) \in \Gamma \left| \mathcal{F}^X(t) \right.\right) = \mathrm{P}\left(X(t+s) \in \Gamma \left| X(t) \right.\right),$$

for all $s, t \geq 0$ and $\Gamma \in \mathcal{B}(E)$, the Borel σ-algebra corresponding to the metric space E. Basically, this means that even if all the movements of the process up to time t are known, the probability of the process being in some set Γ at time $t+s$ only depends on the position of the process at time t (i.e. on the latest information available). Since this equality is true for all Borel sets in $\mathcal{B}(E)$, this implies that

$$\mathrm{E}\left[h(X(t+s)) \left| \mathcal{F}^X(t) \right.\right] = \mathrm{E}\left[h(X(t+s)) \left| X(t) \right.\right],$$

for all $s, t \geq 0$ and $h \in B(E)$.

A function $P(t, x, \Gamma)$ defined on $[0, \infty) \times E \times \mathcal{B}(E)$ is said to be a time-

homogeneous transition function if

(1) $P(t,x,\cdot) \in \mathcal{P}(E)$, the set of Borel probability measures on E, for $(t,x) \in [0,\infty) \times E$;

(2) $P(0,x,\cdot) = \delta_x$, the point-mass at x, for $x \in E$;

(3) $P(\cdot,\cdot,\Gamma) \in B([0,\infty) \times E)$, for $\Gamma \in \mathcal{B}(E)$;

(4) $P(t+s,x,\Gamma) = \int P(t,x,dy)\, P(s,y,\Gamma)$, for $s,t \geq 0$, $x \in E$, and $\Gamma \in \mathcal{B}(E)$.

The last property is referred to as the Chapman-Kolmogorov property.

$P(t,x,\Gamma)$ is a transition function for a time-homogeneous Markov process X if

$$\mathrm{P}\left(X(t+s) \in \Gamma \,\middle|\, \mathcal{F}^X(t)\right) = P(s, X(t), \Gamma), \tag{7.2}$$

for all $s,t \geq 0$ and $\Gamma \in \mathcal{B}(E)$. When a Markov process is not time-homogeneous, such a relation is not true since the term on the RHS varies with t. That is, the fact that the latest information is at time t_1 or t_2 yields different probabilities if $t_1 \neq t_2$ (i.e. $P(s, X(t_1), \Gamma) \neq P(s, X(t_2), \Gamma)$). It then becomes necessary to define a new function $P(t, s, X(t), \Gamma)$ that also depends on t.

The relation in (7.2) can equivalently be expressed in terms of expectations. For a function h belonging to $B(E)$,

$$\mathrm{E}\left[h(X(t+s)) \,\middle|\, \mathcal{F}^X(t)\right] = \int h(y)\, P(s, X(t), dy),$$

for all $s,t \geq 0$.

The initial distribution of X is defined to be a probability measure $\nu \in \mathcal{P}(E)$, given by $\nu(\Gamma) = \mathrm{P}(X(0) \in \Gamma)$. A transition function and an initial distribution

for X are sufficient to determine the finite-dimensional distributions of X. These are defined as

$$\begin{aligned} &\mathrm{P}\left(X\left(0\right)\in\Gamma_0, X\left(t_1\right)\in\Gamma_1, ..., X\left(t_n\right)\in\Gamma_n,\right) = \\ &\int_{\Gamma_0} ... \int_{\Gamma_{n-1}} P\left(t_n - t_{n-1}, y_{n-1}, \Gamma_n\right) P\left(t_{n-1} - t_{n-2}, y_{n-2}, dy_{n-1}\right) \\ &\qquad \cdots P\left(t_1, y_0, dy_1\right) \nu\left(dy_0\right). \end{aligned} \tag{7.3}$$

Moreover, if the metric space (E, r) is complete and separable, then we can affirm that for a specific transition function $P\left(t, x, \Gamma\right)$ there exists a Markov process X on E whose finite-dimensional distributions are uniquely determined by (7.3) (see Theorem 1.1 of Chapter 4 in [16]). Since it is a well-known fact that the metric space $\left(\mathbf{R}^d, \|\cdot\|\right)$ satisfies these two conditions for all $d \geq 1$ (here, $\|\cdot\|$ denotes the Euclidean metric), the previous statement is applicable to the Markov processes we consider.

Unfortunately, formulas for transition functions can only rarely be obtained. Consequently, it becomes necessary to find alternate ways of specifying Markov processes. One method is to use the concept of operator semigroups. It is possible to define an operator semigroup as follows

$$T\left(t\right)h\left(x\right) = \int h\left(y\right) P\left(t, x, dy\right). \tag{7.4}$$

It can be verified that $\{T\left(t\right)\}$ defines a measurable contraction semigroup on $B\left(E\right)$. The first condition of Definition 7.2.1 is indeed satisfied, since

$$T\left(0\right)h\left(x\right) = \int h\left(y\right) P\left(0, x, dy\right) = \int h\left(y\right) \delta_x\left(dy\right) = h\left(x\right).$$

For the second condition, we find

$$\begin{aligned} T(t)\,T(s)\,h(x) &= T(t)\int h(y)\,P(s,x,dy) \\ &= \int\int h(y)\,P(s,z,dy)\,P(t,x,dz)\,. \end{aligned}$$

Applying Fubini's Theorem (this is possible since both h and the transition function are bounded),

$$T(t)\,T(s)\,h(x) = \int h(y)\int P(s,z,dy)\,P(t,x,dz)\,,$$

and by the Chapman-Kolmogorov property, we have for all $s,t \geq 0$

$$T(t)\,T(s)\,h(x) = \int h(y)\,P(s+t,x,dy) = T(s+t)\,h(x)\,.$$

Furthermore, since

$$\begin{aligned} \|T(t)\| &= \sup_{h\in B(E),\ h\neq 0} \frac{\|T(t)\,h\|}{\|h\|} \\ &= \sup_{h\in B(E),\ h\neq 0} \left(\frac{\sup_{x\in\mathbf{R}} |T(t)\,h(x)|}{\sup_{x\in\mathbf{R}} |h(x)|} \right) \\ &= \sup_{h\in B(E),\ h\neq 0} \left(\frac{\sup_{x\in\mathbf{R}} \left|\int h(y)\,P(t,x,dy)\right|}{\sup_{x\in\mathbf{R}} \left|\int h(y)\,\delta_x(dy)\right|} \right) \\ &\leq 1, \end{aligned}$$

then $\{T(t)\}$ is a contraction semigroup. Note that the inequality follows from the fact that the numerator is necessarily smaller than the denominator, since the term at the denominator puts all its mass on one value, chosen to be the supremum.

For the particular types of Markov processes we consider (i.e. continuous-time version of the d-dimensional RWM algorithm and Langevin diffusion), we can even show that the corresponding semigroups $\{T(t)\}$ are strongly continuous. With reference to (7.4), we say that an E-valued Markov process X corresponds to $\{T(t)\}$ if

$$\mathrm{E}\left[h\left(X\left(t+s\right)\right)\left|\mathcal{F}^{X}\left(t\right)\right.\right]=T\left(s\right)h\left(X\left(t\right)\right),$$

for all $s,t\geq 0$ and $h\in L\subset B\left(E\right)$.

For the RWM algorithm, we find by using the law of total probabilities that

$$\begin{aligned}\lim_{t\to 0}\mathrm{E}\left[h\left(\mathbf{Z}^{(d)}\left(t\right)\right)\left|\mathbf{Z}^{(d)}\left(0\right)\right.\right]&=\\ \lim_{t\to 0}\mathrm{E}\left[h\left(\mathbf{Z}^{(d)}\left(t\right)\right)\left|\mathbf{Z}^{(d)}\left(0\right),N\left(t\right)\geq 1\right.\right]\mathrm{P}\left(N\left(t\right)\geq 1\right)&\\ +\lim_{t\to 0}\mathrm{E}\left[h\left(\mathbf{Z}^{(d)}\left(t\right)\right)\left|\mathbf{Z}^{(d)}\left(0\right),N\left(t\right)=0\right.\right]\mathrm{P}\left(N\left(t\right)=0\right),&\end{aligned}$$

where $N(t)$ is the jump (Poisson) process. Given that the process jumps, the expectation in the first term on the RHS does not depend on t. Since $\mathrm{P}\left(N\left(t\right)\geq 1\right)=d^{\alpha}t+o\left(t\right)$, then the first term of the RHS converges to 0 as $t\to 0$. Given that the process does not jump, then the process remains at the same state and we have $\mathrm{E}\left[h\left(\mathbf{Z}^{(d)}\left(t\right)\right)\left|\mathbf{Z}^{(d)}\left(0\right),N\left(t\right)=0\right.\right]=h\left(\mathbf{Z}^{(d)}\left(0\right)\right)$. Since $\mathrm{P}\left(N\left(t\right)=0\right)=1-d^{\alpha}t-o\left(t\right)$, this implies that

$$\lim_{t\to 0}\mathrm{E}\left[h\left(\mathbf{Z}^{(d)}\left(t\right)\right)\left|\mathbf{Z}^{(d)}\left(0\right)\right.\right]=h\left(\mathbf{Z}^{(d)}\left(0\right)\right)$$

and the semigroup $\{T(t)\}$ corresponding to the Markov process $\left\{\mathbf{Z}^{(d)}\left(t\right),t\geq 0\right\}$ is strongly continuous.

For the Langevin diffusion process $\{Z(t), t \geq 0\}$ (in fact for any diffusion process), we have that $\forall \epsilon > 0$,

$$\mathrm{P}\left(|Z(t) - Z(0)| > \epsilon \,|Z(0) = x\right) = o(t)$$

(see [19] on page 493). Since convergence in probability implies convergence in distribution, we then obtain $\lim_{t \to 0} \mathrm{E}\left[h(Z(t)) \,|Z(0)\right] = h(Z(0))$ for $h \in B(E)$. Therefore, the semigroup $\{T(t)\}$ on $B(E)$ corresponding to the Langevin diffusion process is strongly continuous.

In our specific case, we know that the finite-dimensional distributions of each of our Markov processes are determined by a corresponding semigroup $\{T(t)\}$ (see Theorem 1.1 and Proposition 1.6 of Chapter 4 in [16]); this thus implies that they are in turn determined by a corresponding full generator $\hat{A}$. Combining (7.1) and (7.4), we can then conclude that for some Markov process, the pairs of functions (h, g) belonging to $\hat{A}$ are such that $M(t) = h(X(t)) - \int_0^t g(X(s))\, ds$ is an $\mathcal{F}_t^X$-martingale.

Since both Markov processes we consider correspond to strongly continuous contraction semigroups, we even know from Proposition 7.2.5 that they are uniquely determined by their corresponding single-valued generator A (and also their initial distribution ν). Therefore, when studying the (continuous-time) RWM algorithm and the Langevin diffusion, we can focus on functions $h \in L$ such that $M(t) = h(X(t)) - \int_0^t Ah(X(s))\, ds$ is an $\mathcal{F}_t^X$-martingale, which corresponds to the argument used in Section 5.1. In that section we verified, using the martingale property, that the generator of a continuous-time Markov process is an operator

G acting on smooth functions $h : \mathbf{R} \to \mathbf{R}$ such that

$$Gh(x) = \lim_{k \to 0} \frac{1}{k} \mathrm{E}\left[h\left(X\left(t+k\right)\right) - h\left(X\left(t\right)\right) | X\left(t\right) = x\right].$$

By expressing the Markov process in terms of its semigroup, we find

$$\begin{aligned} Gh(x) &= \lim_{k \to 0} \frac{1}{k} \left\{T\left(k\right) h\left(x\right) - T\left(0\right) h\left(x\right)\right\} \\ &= \lim_{k \to 0} \frac{1}{k} \left\{T\left(k\right) h\left(x\right) - h\left(x\right)\right\}, \end{aligned}$$

which implies that $Gh = \lim_{k \to 0} \left(\frac{T(k)h - h}{k}\right)$ and hence both definitions are equivalent.

7.4 Convergence of Probability Measures

We now present some concepts related to the convergence of probability measures. We first introduce the definition of weak convergence, which is of major importance to us. Recall that $\overline{C}\left(E\right)$ is the subspace of real-valued, Borel measurable bounded continuous functions on E. Moreover, $\mathcal{P}\left(E\right)$ is the set of Borel probability measures on E.

Definition 7.4.1. *A sequence of probability measures $\{P_n\} \subset \mathcal{P}\left(E\right)$ is said to converge weakly to $P \in \mathcal{P}\left(E\right)$ if*

$$\lim_{n \to \infty} \int h \, dP_n = \int h \, dP, \qquad \forall h \in \overline{C}\left(E\right). \tag{7.5}$$

Equivalently, a sequence $\{X_n\}$ of E-valued random variables is said to converge in

distribution to the E-valued random variable X if

$$\lim_{n\to\infty} \mathrm{E}\left[h\left(X_n\right)\right] = \mathrm{E}\left[h\left(X\right)\right], \qquad \forall h \in \overline{C}\left(E\right).$$

Weak convergence is denoted by $P_n \Rightarrow P$ and convergence in distribution by $X_n \Rightarrow X$.

We now define two useful properties about sets of functions.

Definition 7.4.2. *A set $M \subset \overline{C}\left(E\right)$ is called separating if whenever $P, Q \in \mathcal{P}\left(E\right)$ and*

$$\int h \; dP = \int h \; dQ, \tag{7.6}$$

for $h \in M$, we have $P = Q$.

In words, it suffices to verify (7.6) for the functions belonging to the separating set M in order to conclude that P and Q really are equal.

Definition 7.4.3. *A set $M \subset \overline{C}\left(E\right)$ is called convergence determining if whenever $\{P_n\} \subset \mathcal{P}\left(E\right)$, $P \in \mathcal{P}\left(E\right)$, and*

$$\lim_{n\to\infty} \int h \; dP_n = \int h \; dP, \tag{7.7}$$

for $h \in M$, the sequence of probability measures $\{P_n\}$ converges weakly to P as n goes to ∞ (i.e. $P_n \Rightarrow P$).

Therefore, if the set M is convergence determining, it is enough to verify (7.7) for the functions belonging to M in order to assess weak convergence.

We should note that if $M \subset \overline{C}\left(E\right)$ is convergence determining, then M is separating. This statement is easily verified. If M is convergence determining,

then we only need to check (7.7) for the functions in M to conclude that $\{P_n\}$ converges weakly to P. However, suppose there exists another probability measure Q such that $P \neq Q$ and such that for all $h \in M$ we have $\int h \, dP = \int h \, dQ$. In that case, since we verified (7.7) for the functions in M only, then the weak limit could also be Q. But this contradicts the fact that M is convergence determining, and so M must be separating.

As noticed in the previous definitions, sets of functions play an important role when assessing convergence of probability measures. In order to study the weak convergence results that shall be presented next section, we need to introduce further definitions related to sets of functions and their properties.

Definition 7.4.4. *An algebra of functions on E is a subset C_a of the space of all continuous functions on E such that for all functions $h, g \in C_a$, we have that $hg \in C_a$, $h + g \in C_a$, and for all constants c, $ch \in C_a$.*

The following example presents an example of an algebra of functions on $\mathbf{R}$.

Example 7.4.5. Consider $\overline{C}^{\infty}(\mathbf{R})$, the subspace of infinitely differentiable bounded functions on $\mathbf{R}$. We show that $\overline{C}^{\infty}(\mathbf{R})$ is an algebra of functions on $\mathbf{R}$.

We start by verifying the last condition. If h is bounded, then ch also is. Furthermore, if h is infinitely differentiable, then $(ch)^{(n)} = ch^{(n)}$ for all $n \in \mathbf{N}$, so ch is infinitely differentiable.

For the second condition, we have that if h and g are in $\overline{C}^{\infty}(\mathbf{R})$, then $h + g$ is bounded as well. Also, the n-th derivative of $h + g$ is $(h + g)^{(n)} = h^{(n)} + g^{(n)}$ for all $n \in \mathbf{N}$, so $h + g$ is infinitely differentiable.

Finally, if both h and g are bounded, then hg clearly bears the same property.

Moreover, $(hg)' = h'g + hg'$ and more generally, the n-th derivative of hg is

$$(hg)^{(n)} = \sum_{i=0}^{n} \binom{n}{i} h^{(i)} g^{(n-i)}$$

for all $n \in \mathbf{N}$. Hence, hg is infinitely differentiable and since all three conditions are verified, then $\overline{C}^{\infty}(\mathbf{R})$ is an algebra of functions on $\mathbf{R}$.

Before moving to the notion of relative compactness of a family of stochastic processes, we present two additional properties of collections of functions.

Definition 7.4.6. *A collection of functions $M \subset \overline{C}(E)$ is said to separate points if for every $x, y \in E$ with $x \neq y$ there exists $f \in M$ such that $f(x) \neq f(y)$.*

Definition 7.4.7. *A collection of functions $M \subset \overline{C}(E)$ is said to strongly separate points if for every $x \in E$ and $\delta > 0$ there exists a finite set $\{f_1, \ldots, f_k\} \subset M$ such that*

$$\inf_{y: r(x,y) \geq \delta} \max_{1 \leq i \leq k} |f_i(y) - f_i(x)| > 0. \tag{7.8}$$

Roughly, the second definition says that if we take some $x \in E$ and $\delta > 0$, there exists a finite set of functions in M such that at least one of the functions f_i belonging to the finite set satisfies $f_i(x) \neq f_i(y)$, for some y that yields the "smallest" difference, but such that $r(x, y) \geq \delta$. It is thus pretty clear from the definitions that if M strongly separates points, then M separates points. Also, in the case where M is a finite set that separates points, then obviously M strongly separates points.

It is easily verified that the set of functions $\overline{C}^{\infty}(\mathbf{R})$ considered in the previous example separates points. To realize this, consider the function $f(x) =$

$\exp(-1/x^2)\,\mathbf{1}_{(x>0)} - \exp(-1/x^2)\,\mathbf{1}_{(x<0)}$; f is thus strictly increasing, bounded below by -1, and above by 1. Moreover, since $f(0) = 0$, then this function is infinitely differentiable. Since f clearly separates points by itself, then obviously $\overline{C}^{\infty}(\mathbf{R})$ separates points as well. Because $\overline{C}^{\infty}(\mathbf{R})$ contains a finite subset of functions which separates points (formed by f only), then it also strongly separates points.

By using a method similar to that of Example 7.4.5, it is easy to show that $C_c^{\infty}(\mathbf{R})$, the subspace of infinitely differentiable functions on $\mathbf{R}$ with compact support, is an algebra of functions. Furthermore, for every $x \in \mathbf{R}$, we can find $f \in C_c^{\infty}(\mathbf{R})$ such that $f(y) < f(x)$ (or equivalently $f(x) < f(y)$) for all y with $y \neq x$; that is, x is the only point reaching the absolute maximum (or minimum) of the function f. It then follows that $C_c^{\infty}(\mathbf{R})$ is an algebra that strongly separates points.

An important concept when talking of weak convergence is that of relative compactness. Indeed, for a sequence of probability measures to converge weakly, it is necessary that they be relatively compact. We say that a family of probability distributions $\{P_n\} \subset \mathcal{P}(E)$ is relatively compact if the closure of $\{P_n\}$ in $\mathcal{P}(E)$ is compact. We have the following result.

Lemma 7.4.8. *Let $\{P_n\} \subset \mathcal{P}(E)$ be relatively compact, let $P \in \mathcal{P}(E)$, and let $M \subset \overline{C}(E)$ be separating. If*

$$\lim_{n\to\infty} \int h\, dP_n = \int h\, dP,$$

holds for $h \in M$, then $P_n \Rightarrow P$.

Proof. The proof of this lemma can be found in [16], p.112. □

This basically says that if a sequence of probability measures is relatively compact and there exists a separating set of functions in $\overline{C}(E)$, then this set of functions is also convergence determining. Hence, if a sequence of probability measures is relatively compact, then we only need to consider the functions belonging to a separating set to assess weak convergence of probability measures.

In this work, we are interested in stochastic processes with sample paths in $D_E[0,\infty)$. It this case, the notion of relative compactness of a sequence of probability measures can equivalently be expressed in terms of a sequence of stochastic processes.

Definition 7.4.9. *Let $\{X_n\}$ be a family of stochastic processes with sample paths in $D_E[0,\infty)$, and let $\{P_n\} \subset \mathcal{P}(D_E[0,\infty))$ be the family of associated probability distributions. We say that $\{X_n\}$ is relatively compact if $\{P_n\}$ is relatively compact.*

In order to define criteria for assessing relative compactness of processes in $D_E[0,\infty)$, it becomes necessary to define compact subsets for collections of step functions. The following corollary presents necessary and sufficient conditions for a sequence of stochastic processes with sample paths in $D_E[0,\infty)$ to be relatively compact.

Corollary 7.4.10. *Let (E,r) be complete and separable, and let $\{X_n\}$ be a sequence of processes with sample paths in $D_E[0,\infty)$. Then $\{X_n\}$ is relatively compact if and only if the following two conditions hold:*

(a) For every $\eta > 0$ and rational $t \geq 0$, there exists a compact set $\Gamma_{\eta,t} \subset E$ such that

$$\liminf_{n\to\infty} \mathrm{P}\left(X_n(t) \in \Gamma^{\eta}_{\eta,t}\right) \geq 1-\eta.$$

(b) For every $\eta > 0$ and $T > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathrm{P}\left(w\left(X_n, \delta, T\right) \geq \eta\right) \leq \eta,$$

with

$$w\left(x, \delta, T\right) = \inf_{t_i} \max_{i} \sup_{s,t\in[t_{i-1},t_i)} r\left(x\left(s\right), x\left(t\right)\right),$$

where $\{t_i\}$ ranges over all partitions of the form $0 = t_0 < t_1 < \cdots < t_{k-1} < T \leq t_k$ with $\min_{1\leq i\leq k}\left(t_i - t_{i-1}\right) > \delta$ and $k \geq 1$.

Proof. The proof of this corollary can be found in [16], on page 130. □

If the sequence $\{X_n\}$ is relatively compact, then the weaker compact containment condition holds. That is, for every $\eta > 0$ and $T > 0$ there is a compact set $\Gamma_{\eta,T} \subset E$ such that

$$\liminf_{n\to\infty} \mathrm{P}\left(X_n\left(t\right) \in \Gamma_{\eta,T} \quad \text{for} \quad 0 \leq t \leq T\right) \geq 1 - \eta.$$

We conclude this section with the result stating that for a sequence of relatively compact processes, weak convergence of the stochastic processes follows from the weak convergence of their finite-dimensional distributions.

Theorem 7.4.11. *Let E be separable, and let X_n, $n = 1, 2, \ldots$, and X be processes with sample paths in $D_E\left[0, \infty\right)$. If $\{X_n\}$ is relatively compact and there exists a dense set $D \subset [0, \infty)$ such that $\left(X_n\left(t_1\right), \ldots, X_n\left(t_k\right)\right) \Rightarrow \left(X\left(t_1\right), \ldots, X\left(t_k\right)\right)$ for every finite set $\{t_1, \ldots, t_k\} \subset D$, then $X_n \Rightarrow X$.*

Proof. The proof of this theorem can be found in [16], on page 132. □

As mentioned at the beginning of Chapter 5 and as we shall see next section, verifying $\mathcal{L}^1$ convergence of the generators leads us to the weak convergence of the finite-dimensional distributions only. To reach weak convergence of the stochastic processes, the previous result affirms that we should verify a priori the relative compactness of $\{X_n\}$. Unfortunately, it is often difficult to verify the necessary and sufficient conditions for relative compactness presented in Corollary 7.4.10. In the case of our RWM algorithm (or, more specifically, its component of interest X_{i^*}), the first condition is easily assessed through a continuity of probabilities argument, but the second is far more difficult to verify. Consequently, it becomes necessary to refer to alternate results. This is where the compact containment condition, as well as the different characteristics of collections of functions defined previously in this section, come into action. Indeed, there exist various results introducing sufficient conditions for assessing relative compactness, and which involve the tools introduced previously. The results we use for the RWM shall be presented next section, and applied to our particular case.

7.5 Weak Convergence to a Markov Process

The following results about weak convergence of stochastic processes are the foundations of the work presented in this thesis. They allow us to deduce that proving $\mathcal{L}^1$ convergence of the generators is sufficient to assess weak convergence of the specific stochastic processes we consider in this thesis. The first result leads us to weak convergence of the finite-dimensional distributions, and is complemented by the second one which verifies relative compactness of the sequence of stochastic processes considered.

For $n = 1, 2, \ldots$ let $\mathcal{G}^n(t)$ be a complete filtration, and let $\mathcal{L}_n$ be the space of real-valued $\mathcal{G}^n(t)$-progressive processes ξ satisfying

$$\sup_{t \leq T} \mathrm{E}\left[|\xi(t)|\right] < \infty$$

for each $T > 0$. Let $\hat{\mathcal{A}}_n$ be the collection of pairs $(\xi, \varphi) \in \mathcal{L}_n \times \mathcal{L}_n$ such that

$$\xi(t) - \int_0^t \varphi(s)\, ds \tag{7.9}$$

is a $\mathcal{G}^n(t)$-martingale.

Note that a $\mathcal{G}^n(t)$-progressive process is such that for each $t \geq 0$, the restriction of X to $[0,t] \times \Omega$ is $\mathcal{B}[0,t] \times \mathcal{G}^n(t)$-measurable. From [16], we know that every right-continuous $\mathcal{G}^n(t)$-adapted process (that is, $X(t)$ is $\mathcal{G}^n(t)$-adapted if it is $\mathcal{G}^n(t)$-measurable for each $t \geq 0$) is $\mathcal{G}^n(t)$-progressive. Therefore, this implies that our sequence of processes $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\}$, $d = 1, 2, \ldots$ is $\mathcal{F}^{Z_{i^*}^{(d)}}(t)$-progressive.

Theorem 7.5.1. *Let (E, r) be complete and separable. Let $A \subset \overline{C}(E) \times \overline{C}(E)$ be linear, and suppose the closure of A generates a strongly continuous contraction semigroup $\{T(t)\}$ on $L \equiv \mathcal{D}(A)$. Suppose X_n, $n = 1, 2, \ldots$, is a $\mathcal{G}^n(t)$-progressive E-valued process, X is a Markov process corresponding to $\{T(t)\}$, and $X_n(0) \Rightarrow X(0)$. Let $M \subset \overline{C}(E)$ be separating and suppose either L is separating and $\{X_n(t)\}$ is relatively compact for each $t \geq 0$, or L is convergence determining. Then the following are equivalent:*

(a) The finite-dimensional distributions of X_n converge weakly to those of X.

(b) For each $(h, g) \in A$,

$$\lim_{n\to\infty} \mathrm{E}\left[\left(h\left(X_n\left(t+s\right)\right) - h\left(X_n\left(t\right)\right) - \int_t^{t+s} g\left(X_n\left(u\right)\right) du\right) \prod_{i=1}^{k} f_i\left(X_n\left(t_i\right)\right)\right] = 0$$

for all $k \geq 0$, $0 \leq t_1 < t_2 < \ldots < t_k \leq t < t+s$, *and* $f_1, \ldots, f_k \in M$.

(c) For each $(h, g) \in A$ *and* $T > 0$, *there exist* $(\xi_n, \varphi_n) \in \hat{\mathcal{A}}_n$ *such that*

$$\sup_n \sup_{s\leq T} \mathrm{E}\left[\left|\xi_n\left(s\right)\right|\right] < \infty, \tag{7.10}$$

$$\sup_n \sup_{s\leq T} \mathrm{E}\left[\left|\varphi_n\left(s\right)\right|\right] < \infty, \tag{7.11}$$

$$\lim_{n\to\infty} \mathrm{E}\left[\left(\xi_n\left(t\right) - h\left(X_n\left(t\right)\right)\right) \prod_{i=1}^{k} f_i\left(X_n\left(t_i\right)\right)\right] = 0, \tag{7.12}$$

$$\lim_{n\to\infty} \mathrm{E}\left[\left(\varphi_n\left(t\right) - g\left(X_n\left(t\right)\right)\right) \prod_{i=1}^{k} f_i\left(X_n\left(t_i\right)\right)\right] = 0, \tag{7.13}$$

for all $k \geq 0$, $0 \leq t_1 < t_2 < \ldots < t_k \leq t \leq T$, *and* $f_1, \ldots, f_k \in M$.

Proof. For a proof of this theorem, see [16] on page 227. □

The previous theorem establishes weak convergence of the finite-dimensional distributions of some sequence of processes to those of a certain Markov process. In our case, we choose to verify condition (c) in order to prove the weak convergence result. First, since the limiting Markov process corresponds to a strongly continuous contraction semigroup, then by Proposition 7.2.5 it is uniquely determined by its single-valued generator A; it thus suffices to verify the conditions in (c) for each pair of functions (h, Ah), where $h \in \mathcal{D}\left(A\right) \subset \overline{C}\left(E\right)$.

Furthermore, note that (7.12) and (7.13) are implied by

$$\lim_{n\to\infty} \mathrm{E}\left[\left|\xi_n\left(t\right) - h\left(X_n\left(t\right)\right)\right|\right] = \lim_{n\to\infty} \mathrm{E}\left[\left|\varphi_n\left(t\right) - g\left(X_n\left(t\right)\right)\right|\right] = 0. \tag{7.14}$$

Indeed, since $f_1, \cdots, f_k \in M \subset \overline{C}\left(E\right)$, where $\overline{C}\left(E\right)$ is the set of bounded continuous functions on E, then $f_i \leq B$ (say). Using the triangle's inequality then yields

$$\begin{aligned}
&\mathrm{E}\left[\left(\xi_n\left(t\right) - h\left(X_n\left(t\right)\right)\right)\prod_{i=1}^{k} f_i\left(X_n\left(t_i\right)\right)\right] \\
&\leq \mathrm{E}\left[\left|\left(\xi_n\left(t\right) - h\left(X_n\left(t\right)\right)\right)\prod_{i=1}^{k} f_i\left(X_n\left(t_i\right)\right)\right|\right] \\
&\leq B^k \mathrm{E}\left[\left|\xi_n\left(t\right) - h\left(X_n\left(t\right)\right)\right|\right],
\end{aligned}$$

and similarly for (7.13). We thus use (7.14) instead of (7.12) and (7.13).

Now, let's apply this theorem to the particular case we are studying in this thesis. First, suppose that the process $\left\{X_d\left(t\right), t \geq 0\right\}$ is in fact the process $\left\{Z_{i^*}^{(d)}\left(t\right), t \geq 0\right\}$ formed by the i^*-th component of a d-dimensional Markov process. If we let $\mathcal{G}^d\left(t\right) = \mathcal{F}^{Z_{i^*}^{(d)}}\left(t\right)$, then we have access to the past movements of the component of interest X_{i^*} only (see the discussion in Section 5.1.4). Setting $\xi_d\left(t\right) = h\left(Z_{i^*}^{(d)}\left(t\right)\right)$ and $\varphi_d\left(t\right) = \lim_{k\to 0} \frac{1}{k}\mathrm{E}\left[h\left(Z_{i^*}^{(d)}\left(t+k\right)\right) - h\left(Z_{i^*}^{(d)}\left(t\right)\right) \middle| \mathcal{F}^{Z_{i^*}^{(d)}}\left(t\right)\right]$, it is easily verified using the developments in Section 5.1.3 that the martingale condition in (7.9) is satisfied (with respect to $\mathcal{F}^{Z_{i^*}^{(d)}}\left(t\right)$). Because $h \in \overline{C}$, then (7.10) is trivial; given that (7.12) is exactly equal to 0, then the only equations left to verify are (7.11) and (7.13) (or the second part of (7.14)), which in our case are

respectively given by $\sup_d \mathrm{E}\left[|Gh\left(d, X_{i^*}\right)|\right] < \infty$ and

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

where $Gh\left(d, X_{i^*}\right)$ is as in (5.6). Furthermore, since we know that the limiting process is either a Langevin diffusion or a Metropolis-Hastings algorithm, then $Ah\left(X_{i^*}\left(t\right)\right)$ is either given by $G_L h\left(X_{i^*}\right)$ in Section 5.2.2 with appropriate speed measure $\upsilon\left(\ell\right)$ or by $G_{MH} h\left(X_{i^*}\right)$ in Section 5.3.2 with appropriate acceptance rule $\alpha\left(\ell^2, X_{i^*}, Y_{i^*}\right)$.

All is left to do to assess weak convergence of the finite-dimensional distributions is thus to determine for which set of functions L the previous two conditions should be verified. From Theorem 7.5.1, we know that $L \subset \overline{C}$; for the case where the limiting process is a Metropolis-Hastings algorithm, we can prove the relations for all $h \in \overline{C}$. Since $\overline{C}$ is clearly separating (see the discussion in Section 7.4) then Theorem 7.5.1 can be applied without any concern. For the case where the limit is a Langevin diffusion process however, it is not as simple to verify $\mathcal{L}^1$ convergence of the generators for all continuous and bounded functions h. To overcome this problem, we resort to the convenient concept of core (see Section 7.6) and conclude that it is sufficient to verify both conditions for all functions $h \in C_c^\infty$.

In Chapters 5 and 6, we concentrated on verifying $\mathcal{L}^1$ convergence of the generators. However, to apply Theorem 7.5.1, we see that there is a second condition that still needs to be checked. We shall see shortly that this condition is satisfied, as a stronger statement is verified at the end of this section when talking about relative compactness.

In order to conclude that the sequence of processes converges weakly to the

Markov process in question, Theorem 7.4.11 states that we should assess relative compactness of this same sequence of processes. To do so, we use the following result.

Corollary 7.5.2. *Suppose in Theorem 7.5.1 that the X_n and X have sample path in $D_E[0,\infty)$, and there is an algebra $C_a \subset L$ that separates points. Suppose either that the compact containment condition holds for $\{X_n\}$ or that C_a strongly separates points. If $\{(\xi_n, \varphi_n)\}$ in condition (c) can be selected so that*

$$\lim_{n\to\infty} \mathrm{E}\left[\sup_{t\in\mathbf{Q}\,\cap\,[0,T]} |\xi_n(t) - h(X_n(t))|\right] = 0 \tag{7.15}$$

and

$$\limsup_{n\to\infty} \mathrm{E}\left[\|\varphi_n\|_{p,T}\right] < \infty \quad \text{for some} \quad p \in (1,\infty], \tag{7.16}$$

where $\|\varphi_n\|_{p,T} = \left[\int_0^T |\varphi_n(t)|^p\, dt\right]^{1/p}$ if $p < \infty$ and $\|\varphi_n\|_{\infty,T} = \text{ess sup}_{0\le t\le T} |\varphi_n(t)|$, then $X_n \Rightarrow X$.

Proof. For a proof of this theorem, see [16] on page 232. □

To verify the conditions of this result, it will be necessary to focus on our specific problem. From Section 7.4, we know that $C_c^\infty \subset L$ is an algebra that strongly separates points. From our previous choice of (ξ_d, φ_d), we know that $\left|\xi_d(t) - h\left(Z_{i^*}^{(d)}(t)\right)\right|$ is exactly 0, so (7.15) is trivial. To verify the second condition, we set $p = 2$ and realize that by Jensen's inequality,

$$\mathrm{E}\left[\left(\int_0^T \varphi_d^2(t)\, dt\right)^{1/2}\right] \le \left(\mathrm{E}\left[\int_0^T \varphi_d^2(t)\, dt\right]\right)^{1/2}.$$

We first concentrate on the integral appearing on the RHS, keeping in mind the

fact that $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\}$ is a step function and proposes moves according to a Poisson process with rate d^α; furthermore, $\varphi_d(t) = Gh(d, X_{i^*})$.

Using Theorem 5.2 in [38], it is easy to show that

$$\mathrm{E}\left[\sum_{i=1}^{N(T)} \varphi_d^2(T_i)\right] = \mathrm{E}\left[\sum_{i=1}^{N} \varphi_d^2(U_i)\right] = d^\alpha \int_0^T \varphi_d^2(t)\, dt,$$

where T_i is the time of the n-th event of the Poisson process $\{N(t), t \geq 0\}$ having rate d^α, and where $U_1, U_2, \ldots$ is a sequence of *iid* uniform $(0, T)$ random variables that is independent of N, a Poisson random variable with mean λT. Therefore, we find

$$\begin{aligned}\mathrm{E}\left[\int_0^T \varphi_d^2(t)\, dt\right] &= \frac{1}{d^\alpha}\mathrm{E}\left[\mathrm{E}_N\left[\mathrm{E}_{U_1,U_2,\ldots}\left[\sum_{i=1}^{N} \varphi_d^2(U_i)\right]\right]\right] \\ &= \frac{1}{d^\alpha}\mathrm{E}\left[\mathrm{E}_N\left[N\mathrm{E}_{U_1}\left[\varphi_d^2(U_1)\right]\right]\right].\end{aligned}$$

Recall that we chose $\varphi_d(t) = Gh(d, X_{i^*}(t))$; since by assumption $\mathbf{X}^{(d)}(0)$ is distributed according to the target density π which is stationary for the Markov chain, then the unconditional distribution of $X_{i^*}(t)$ has density f for all $t \geq 0$. We then obtain

$$\mathrm{E}\left[\int_0^T \varphi_d^2(t)\, dt\right] = \frac{1}{d^\alpha}\mathrm{E}[N]\,\mathrm{E}\left[\varphi_d^2(0)\right] = T\mathrm{E}\left[\varphi_d^2(0)\right].$$

If we can show that $\mathrm{E}[\varphi_d^2(0)]$ is bounded by some constant for all $d \geq 1$, then (7.16) will be verified (and this will imply by the same fact that (7.11) is satisfied as well). For the case where the limiting process is a discrete-time RWM algorithm, this follows directly from the definition of $Gh(d, X_{i^*})$ in (5.9). Since $\alpha = \lambda_1 = 0$ in

this case and $h \in \overline{C}$, then the expectation is bounded by some constant. For the case where the limiting process is a Langevin diffusion process, $\alpha > 0$ and thus we have to develop (5.6) further if we want to reach the same conclusion. From the proof of Lemma 6.1.2, we know that

$$\begin{aligned} |Gh(d, X_{i^*})| &\leq \left|\widetilde{G}h(d, X_{i^*})\right| + K\left(\frac{\ell^3}{d^{\alpha/2}} + \frac{\ell^4}{d^{\alpha}} + \frac{\ell^5}{d^{3\alpha/2}}\right)\left((\log f(X_{i^*}))'\right)^2 \\ &\quad + K\left(\frac{\ell^4}{d^{\alpha}} + \frac{\ell^5}{d^{3\alpha/2}} + \frac{\ell^6}{d^{2\alpha}}\right)\left(1 + \left|(\log f(X_{i^*}))'\right|\right) \\ &\quad + K\frac{\ell^3}{d^{\alpha/2}} + K\frac{\ell^7}{d^{5\alpha/2}} \end{aligned}$$

for some positive constant K. From the definition of $\widetilde{G}h(d, X_{i^*})$ in (6.1) and using the fact that both its expectation terms are bounded, along with the fact that h and its derivatives are bounded in absolute value, we obtain

$$\begin{aligned} |Gh(d, X_{i^*})| &\leq K + K\left|(\log f(X_{i^*}))'\right| + K\frac{1}{d^{\alpha/2}} \\ &\quad + K\frac{1}{d^{\alpha/2}}\left((\log f(X_{i^*}))'\right)^2 + K\frac{1}{d^{\alpha}}\left|(\log f(X_{i^*}))'\right| \\ &\leq K + K\left|(\log f(X_{i^*}))'\right| + K\frac{1}{d^{\alpha/2}}\left((\log f(X_{i^*}))'\right)^2, \end{aligned}$$

for some $K > 0$. By assumption (see Section 2.2), we know that $\mathrm{E}\left[\left((\log f(X))'\right)^4\right] < \infty$, which implies that $\mathrm{E}\left[\varphi_d^2(0)\right] = \mathrm{E}\left[(Gh(d, X_{i^*}))^2\right] < \infty$ for all d; therefore, (7.16) follows.

Since the conditions of Corollary 7.5.2 are satisfied, we then proved that $\left\{Z_{i^*}^{(d)}(t), t \geq 0\right\}$ is relatively compact and thus the weak convergence of the finite-dimensional distributions in Theorem 7.5.1 implies the weak convergence of the stochastic processes.

7.6 Cores

We finally briefly introduce the notion of cores, which was mentioned in the previous section and sometimes constitutes a useful alternative to the domain of an operator.

Definition 7.6.1. *A subspace D of $\mathcal{D}(A)$ is said to be a core for A if the closure of the restriction of A to D is equal to A, that is $\overline{A|_D} = A$.*

More intuitively, a core is a representative subspace of the domain of an operator which can be used in its place to simplify the problem.

It turns out that a core can be found for the generator of the Langevin diffusion process considered in this thesis. Consequently, instead of verifying $\mathcal{L}^1$ convergence of the generators for all functions h in the domain of G_L, we shall be allowed to work with functions belonging to this core only.

Specifically, we are given that the generator of a diffusion process satisfies:

$$A = \frac{1}{2} a(x) \frac{d^2}{dx^2} + b(x) \frac{d}{dx},$$

where $a(x)$ and $b(x)$ are the volatility and drift terms respectively (see [36], for instance). The domain of A is C^2, the subspace of twice differentiable functions on $\mathbf{R}$.

Theorem 7.6.2. *Suppose $a \in C^2$, $a \geq 0$, a'' is bounded, and $b : \mathbf{R} \to \mathbf{R}$ is Lipschitz continuous. Then, C_c^∞ is a core for A.*

Proof. For the proof of this theorem, refer to [16] on pages 371-372. □

The generator of a Langevin diffusion is given by

$$G_L h(x) = \frac{1}{2} h''(x) + \frac{1}{2} (\log f(x))' h'(x).$$

Since $a(x) = 1/2$ is a constant, it is easily verified that $a(x) \in C^2$, is strictly positive and also has a bounded second derivative. Moreover, since $(\log f)'$ is Lipschitz continuous by assumption, then C_c^∞ is a core for G_L.

Since the Langevin diffusion is uniquely determined by its generator G_L and from the definition of a core, we know that C_c^∞ is convergence determining (this also follows from Theorem 4.5 of Chapter 3 in [16], along with the fact that C_c^∞ is an algebra that strongly separates points); from the discussion in Section 7.4, C_c^∞ is thus also separating. Using Lemma 7.5.1, it is thus sufficient to prove $\mathcal{L}^1$ convergence of generators for functions $h \in C_c^\infty$ in order to achieve weak convergence of the processes $\left\{ Z_{i^*}^{(d)}(t), t \geq 0 \right\}$ to the Langevin diffusion as $d \to \infty$.

Conclusion

The well-known acceptance rate 0.234 has long been believed to hold under certain perturbations of the target density. In this thesis, we extended the *iid* work of [29] to a more general setting where the scaling term of each target component is allowed to depend on the dimension of the target distribution. The particularity of our results is that they provide, for the specified target setting, a necessary and sufficient condition under which the (rescaled) RWM algorithm has an asymptotic behavior that is identical to the *iid* case, yielding an AOAR of 0.234. This condition ensures that the target distribution does not admit components having scaling terms that are significantly smaller than the others. We also proved that when this condition is not verified, the process of interest has a different limiting behavior, yielding AOARs that are smaller than 0.234. In particular, we introduced an equation from which the appropriate AOAR can be numerically solved for optimal performance of the algorithm. These results are the first to admit limiting processes and AOARs that are different from those found by [29] for RWM algorithms. This work should then act as a warning for practitioners, who should be aware that the usual 0.234 might be inefficient even with seemingly regular targets.

It is worth mentioning that although asymptotic, the results presented in this paper work well in relatively small dimensions. In addition, the results provided

about the optimal form for the proposal variance as a function of d constitute useful guidelines in practice. The results of this paper are then relatively easy to apply for practitioners, as it suffices to verify which conditions are satisfied, and then numerically optimize the appropriate equation to find the optimal scaling value.

As a special case, our results can be used to determine the AOAR for virtually any correlated multivariate normal target distribution. Contrarily to what seemed to be a common belief, multivariate normal distributions do not always adopt a conventional limiting behavior. However, a drastic variation in the AOAR seems to be more common as target distributions get further from normality. In general, AOARs for multivariate normal targets appear to lie relatively close to 0.234, regardless of the correlation structure existing among the components. As discussed in Section 3.3 however, even for the most regular target distributions, an extremely small scaling term causes the algorithm to be inefficient and forces us to resort to inhomogeneous proposal distributions. The AOAR obtained under this method is then not necessarily close to 0.234.

Since our results can be used to optimize any multivariate normal target distribution, this also includes cases where the target is a hierarchical model possessing a distribution which is jointly normal. This raises the question as to whether RWM algorithms can be optimized and similar results derived for broader hierarchical target models. The examples presented in Section 4.4 seem to answer this question positively, but AOARs appear to differ from 0.234 with increasing significance as the distribution gets further from normality. The optimization problem for general hierarchical models is presently under investigation (see [7]).

Appendix A. Miscellaneous Results for the Lemmas Proofs

This appendix contains various results used in the proofs of the lemmas in Chapter 6. The results are presented in the same order as they appear in the text. Note that Propositions A.4 and A.5 can be found in [29].

Lemma A.1. *Let f be a C^2 probability density function (pdf). If $(\log f(x))'$ is Lipschitz continuous, i.e.*

$$\sup_{x,y\in\mathbf{R},x\neq y} \frac{\left|\frac{f'(x)}{f(x)} - \frac{f'(y)}{f(y)}\right|}{|x-y|} < \infty, \tag{A.1}$$

then $f'(x) \to 0$ as $x \to \pm\infty$.

Proof. The asymptotic behavior of a C^2 *pdf* can be one of three things:

1) $f(x) \to 0$ as $x \to \pm\infty$ and $f'(x) \to 0$ as $x \to \pm\infty$;

2) $f(x) \to 0$ as $x \to \pm\infty$ and $f'(x) \not\to 0$ as $x \to \pm\infty$;

3) $f(x) \not\to 0$ as $x \to \pm\infty$ and $f'(x) \not\to 0$ as $x \to \pm\infty$.

We prove that if we are in case (2) or (3), then $(\log f(x))'$ is not Lipschitz continuous, thus implying that (1) is the only possible option.

Case (2): We might face the case where the density f converges to 0, but where its first derivative f' does not. Since $f \to 0$, then $\forall \epsilon > 0$ there exists $x_0(\epsilon) \in \mathbf{R}$ such that $\forall x \geq x_0(\epsilon)$, we have $f(x) < \epsilon$.

Also, since $f' \nrightarrow 0$, then for all $\epsilon > 0$ we can find $x^* \geq x_0(\epsilon) + 1$ such that (a) $f'(x^*) < -\limsup |f'|/2$ or (b) $f'(x^*) > \limsup |f'|/2$. We now demonstrate that in both cases, the Lipschitz continuity assumption is violated, ruling out the option where $f \to 0$ but $f' \nrightarrow 0$.

Case (a): $f'(x^*) < -\limsup |f'|/2$. Note that a function taking the value ϵ at time 0 and with a slope of $-\epsilon$ will reach 0 at time 1. Since f is C^2, then $\forall 0 < \epsilon < \limsup |f'|/2$ there exists at least one value $y < x^*$ with $x^* - y \leq 1$ such that $f'(y) = -\epsilon$. If this was not true, this would mean that $f'(x) < -\epsilon$ for $x \in (x^* - 1, x^*)$ and then f would cross 0, violating the fact that $f \geq 0$. Now, let y^* be the largest of those y's, which implies that $f(y^*) > f(x^*)$. Given $0 < \epsilon < \limsup |f'|/2$, we then have

$$\begin{aligned}\sup_{x,y \in \mathbf{R}, x \neq y} \frac{\left|\frac{f'(x)}{f(x)} - \frac{f'(y)}{f(y)}\right|}{|x-y|} &\geq \frac{\left|\frac{f'(x^*)}{f(x^*)} - \frac{f'(y^*)}{f(y^*)}\right|}{1} \geq \left|\frac{f'(x^*)}{f(x^*)} - \frac{-\epsilon}{f(x^*)}\right| \\ &\geq \left|\frac{-\limsup |f'|/2 + \epsilon}{f(x^*)}\right| \geq \left|\frac{\limsup |f'|/2 - \epsilon}{\epsilon}\right|.\end{aligned}$$

Since this is true for all $0 < \epsilon < \limsup |f'|/2$, then

$$\sup_{x,y \in \mathbf{R}, x \neq y} \frac{\left|\frac{f'(x)}{f(x)} - \frac{f'(y)}{f(y)}\right|}{|x-y|} = \infty$$

and the Lipschitz continuity assumption is violated.

Case (b): $f'(x^*) > \limsup |f'|/2$. In a similar fashion, we note that a function

starting at 0 and with slope equal to ϵ will reach ϵ after one unit of time. Since f is C^2, then $\forall 0 < \epsilon < \limsup |f'|/2$ there exists at least one value $y > x^*$ with $y - x^* \leq 1$ such that $f'(y) = \epsilon$; if this was not true, then f would cross ϵ, which would contradict the fact that $f(x) < \epsilon$ for all $x \geq x_0(\epsilon)$. Now, let y^* be the smallest such value, which implies that $f(y^*) > f(x^*)$. Given $0 < \epsilon < \limsup |f'|/2$ and repeating what was done in (a), we obtain

$$\sup_{x,y \in \mathbf{R}, x \neq y} \frac{\left| \frac{f'(x)}{f(x)} - \frac{f'(y)}{f(y)} \right|}{|x - y|} \geq \left| \frac{\limsup |f'|/2 - \epsilon}{\epsilon} \right|,$$

and therefore the Lipschitz continuity assumption is again violated.

Note that although we focused on the behavior of f as $x \to \infty$ in both (a) and (b), we can repeat the same reasoning for $x \to -\infty$.

Case (3): We might also face the case where f does not converge to 0. Since f is continuous, positive, and $\int f = 1$, we then have that $\forall \epsilon > 0$, there exists $x_0(\epsilon) \in \mathbf{R}$ such that $f(x) < \epsilon$ for $x \geq x_0(\epsilon)$, except on a set A_ϵ of Lebesgue measure $\lambda(A_\epsilon) < \epsilon$. In other words, we can find $x_0(\epsilon) \in \mathbf{R}$ such that $f(x) < \epsilon$ for the majority of $x \geq x_0(\epsilon)$, and such that the other $x \geq x_0(\epsilon)$ with $f(x) \geq \epsilon$ have a probability smaller than ϵ of occurring, i.e. $\mathrm{P}(\{x : x \geq x_0(\epsilon), f(x) \geq \epsilon\}) < \epsilon$.

Since $(-\infty, \epsilon)$ is an open set and f is continuous, it follows that the set $B = \{x \in \mathbf{R} : f(x) < \epsilon\}$ must be an open set as well (in our case, this set is a countable union of open intervals). We then conclude that the complement of this set ($B^c = \{x \in \mathbf{R} : f(x) \geq \epsilon\}$) must be formed of closed intervals (which might include singletons). Since $A_\epsilon = B^c \cap [x_0, \infty)$, then it is also formed of closed intervals over which $f(x) \geq \epsilon$.

Since f is a C^2 density, then $f(x) < \infty$ for all $x \in \mathbf{R}$. Also, since $f \nrightarrow 0$, then for all $\epsilon > 0$ we can find an interval $[x(\epsilon), y(\epsilon)]$ in A_ϵ where the maximum value reached by the function f over this interval ($h(\epsilon)$ say) is such that $h(\epsilon) > \limsup |f|/2$. There might be more than just one value in this interval for which f attains its maximum, but for all of those values we will have $f'(x) = 0$. Since the maximum and minimum values taken by f over the interval $[x(\epsilon), y(\epsilon)]$ are $h(\epsilon)$ and ϵ respectively (since $f(x(\epsilon)) = f(y(\epsilon)) = \epsilon$), then $\sup_{x\in\mathbf{R}} f'(x) \geq \frac{h(\epsilon)-\epsilon}{y(\epsilon)-x(\epsilon)} > \frac{h(\epsilon)-\epsilon}{\epsilon}$, where the last inequality comes from the fact that $|y(\epsilon) - x(\epsilon)| = \lambda([x(\epsilon), y(\epsilon)]) \leq \lambda(A_\epsilon) < \epsilon$; hence, $\sup_{x\in\mathbf{R}} \frac{f'(x)}{f(x)} > \frac{h(\epsilon)-\epsilon}{\epsilon h(\epsilon)}$. Since this is true for all $\epsilon > 0$, then $\sup_{x\in\mathbf{R}} \frac{f'(x)}{f(x)} = \infty$.

With this information in hand, we now verify if the Lipschitz continuity assumption is satisfied. Given $\epsilon > 0$, we take y to be one of the points in $[x(\epsilon), y(\epsilon)]$ such that $f(y) = h(\epsilon)$ and $f'(y) = 0$. We then have

$$\sup_{x,y\in\mathbf{R},x\neq y} \frac{\left|\frac{f'(x)}{f(x)} - \frac{f'(y)}{f(y)}\right|}{|x-y|} \geq \sup_{x\in\mathbf{R}} \frac{\left|\frac{f'(x)}{f(x)} - 0\right|}{|x(\epsilon) - y(\epsilon)|} > \sup_{x\in\mathbf{R}} \frac{\left|\frac{f'(x)}{f(x)} - 0\right|}{\epsilon} = \infty,$$

and since this inequality must be verified $\forall \epsilon > 0$, then we conclude that the Lipschitz continuity assumption is violated. Note that we have considered the case where $x \to \infty$, but we can repeat a similar reasoning for the case where $x \to -\infty$. □

Proposition A.2. *If $Z \sim N(0, \sigma^2)$, then*

$$\mathrm{E}\left[Z^k\right] = 0, \qquad k = 1, 3, 5, ... \tag{A.2}$$

$$\mathrm{E}\left[Z^k\right] = \sigma^k \prod_{i=0}^{\frac{k}{2}-1} (k - (2i+1)), \qquad k = 2, 4, 6, ... \tag{A.3}$$

Proof. Since this distribution is symmetric and centered at 0, then for $k = 1, 3, 5, ...$

$$\mathrm{E}\left[Z^k\right] = \int_{\mathbf{R}} \left(2\pi\sigma^2\right)^{-1/2} z^k \exp\left(-\frac{z^2}{2\sigma^2}\right) dz = 0.$$

For the case where k is even,

$$\begin{aligned} \mathrm{E}\left[Z^k\right] &= \int_{\mathbf{R}} \left(2\pi\sigma^2\right)^{-1/2} z^k \exp\left(-\frac{z^2}{2\sigma^2}\right) dz \\ &= 2\int_0^{\infty} \left(2\pi\sigma^2\right)^{-1/2} z^k \exp\left(-\frac{z^2}{2\sigma^2}\right) dz. \end{aligned}$$

Letting $u = z^2$, we get $du = 2z\ dz$ and

$$\begin{aligned} \mathrm{E}\left[Z^k\right] &= \int_0^{\infty} \left(2\pi\sigma^2\right)^{-1/2} u^{\frac{k-1}{2}} \exp\left(-\frac{u}{2\sigma^2}\right) du \\ &= \left(2\pi\sigma^2\right)^{-1/2} \int_0^{\infty} u^{\frac{k-1}{2}} \exp\left(-\frac{u}{2\sigma^2}\right) du. \end{aligned}$$

The integrand is the unormalized density of a $\Gamma\left(\frac{k+1}{2}, \frac{1}{2\sigma^2}\right)$, and then

$$\int_0^{\infty} u^{\frac{k-1}{2}} \exp\left(-\frac{u}{2\sigma^2}\right) du = \Gamma\left(\frac{k+1}{2}\right) \left(2\sigma^2\right)^{\frac{k+1}{2}}.$$

Since $\Gamma(x+1) = x\Gamma(x)$ and since $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, we have that

$$\Gamma\left(\frac{k+1}{2}\right) = \sqrt{\pi}\ 2^{-\frac{k}{2}} \prod_{i=0}^{\frac{k}{2}-1} \left(k - (2i+1)\right).$$

Therefore, for $k = 2, 4, 6, ...,$

$$\begin{aligned} \mathrm{E}\left[Z^k\right] &= \left(2\pi\sigma^2\right)^{-1/2} \sqrt{\pi}\ 2^{-\frac{k}{2}} \prod_{i=0}^{\frac{k}{2}-1} \left(k - (2i+1)\right) \left(2\sigma^2\right)^{\frac{k+1}{2}} \\ &= \sigma^k \prod_{i=0}^{\frac{k}{2}-1} \left(k - (2i+1)\right). \end{aligned}$$

□

Proposition A.3. *If* $Z \sim N\left(0, \sigma^2\right)$, *then*

$$\begin{aligned} \mathrm{E}\left[|Z|^k\right] &= \frac{2^{\frac{k}{2}}}{\sqrt{\pi}}\ \Gamma\left(\frac{k+1}{2}\right) \sigma^k, && k = 1, 3, 5, ... \\ \mathrm{E}\left[|Z|^k\right] &= \mathrm{E}\left[Z^k\right], && k = 2, 4, 6, ... \end{aligned} \tag{A.4}$$

Proof. For $k = 1, 3, 5, ...,$

$$\mathrm{E}\left[|Z|^k\right] = 2 \int_0^\infty \left(2\pi\sigma^2\right)^{-1/2} z^k \exp\left(-\frac{z^2}{2\sigma^2}\right) dz.$$

Letting $u = z^2$, then $du = 2z\ dz$ and

$$\begin{aligned} \mathrm{E}\left[|Z|^k\right] &= \left(2\pi\sigma^2\right)^{-1/2} \int_0^\infty u^{\frac{k-1}{2}} \exp\left(-\frac{u}{2\sigma^2}\right) du \\ &= \left(2\pi\sigma^2\right)^{-1/2} \Gamma\left(\frac{k+1}{2}\right) \left(2\sigma^2\right)^{\frac{k+1}{2}} \\ &= \frac{2^{\frac{k}{2}}}{\sqrt{\pi}}\ \Gamma\left(\frac{k+1}{2}\right) \sigma^k. \end{aligned}$$

□

Proposition A.4. *The function $g(x) = 1 \wedge e^x$ is Lipschitz with coefficient 1. That is, for all $x, y \in \mathbf{R}$, $|g(x) - g(y)| \leq |x - y|$.*

Proof. The graph of $g(x)$ versus x is convex, increases from 0 to 1 over $(-\infty, 0)$, and then remains constant at 1 over the positive axis.

We then have three possible cases:

1. If $x, y > 0$, we have $|g(x) - g(y)| = |1 - 1| = 0 \leq |x - y|$.

2. If $x, y \leq 0$, we have

$$\frac{|g(x) - g(y)|}{|x - y|} = \frac{|e^x - e^y|}{|x - y|} \leq 1 \quad \Rightarrow \quad |g(x) - g(y)| \leq |x - y|,$$

since $\frac{d}{dx} e^x = e^x \leq 1$ for all $x \leq 0$ (and then the slope is always ≤ 1).

Note that in the special case where $y = 0$ and $x < 0$ (or equivalently $x = 0$ and $y < 0$), we have

$$1 - e^x \leq |x|. \tag{A.5}$$

3. Finally, if $x \leq 0$, $y > 0$ (or $x > 0$, $y \leq 0$) then the LHS is the same as in (A.5) but the RHS is bigger, i.e. $|g(x) - g(y)| = 1 - e^x \leq |x| \leq |x - y|$.

□

Proposition A.5. *If $A \sim N(\mu, \sigma^2)$, then*

$$\mathrm{E}\left[1 \wedge e^A\right] = \Phi\left(\frac{\mu}{\sigma}\right) + \exp\left(\mu + \frac{\sigma^2}{2}\right) \Phi\left(-\sigma - \frac{\mu}{\sigma}\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Proof. By the law of total probabilities, we get

$$\begin{aligned} \mathrm{E}\left[1 \wedge e^{A}\right] &= \mathrm{E}\left[1 \wedge e^{A}; A \leq 0\right] + \mathrm{E}\left[1 \wedge e^{A}; A > 0\right] \\ &= \mathrm{E}\left[e^{A}; A \leq 0\right] + \Phi\left(\frac{\mu}{\sigma}\right). \end{aligned}$$

The first term satisfies

$$\begin{aligned} &\mathrm{E}\left[e^{A}; A \leq 0\right] \\ &= \int_{-\infty}^{0} \exp\left(a\right) \left(2\pi\sigma^{2}\right)^{-1/2} \exp\left(\frac{-1}{2\sigma^{2}}\left(a-\mu\right)^{2}\right) da \\ &= \exp\left(\frac{-\mu^{2}}{2\sigma^{2}}\right) \int_{-\infty}^{0} \left(2\pi\sigma^{2}\right)^{-1/2} \exp\left(\frac{-1}{2\sigma^{2}}\left(a^{2} - 2\left(\mu+\sigma^{2}\right)a\right)\right) da. \end{aligned}$$

Completing the square in the exponential term, we get

$$\begin{aligned} \mathrm{E}\left[e^{A}; A \leq 0\right] &= \exp\left(\frac{\left(\mu+\sigma^{2}\right)^{2}}{2\sigma^{2}} - \frac{\mu^{2}}{2\sigma^{2}}\right) \\ &\quad \times \int_{-\infty}^{0} \left(2\pi\sigma^{2}\right)^{-1/2} \exp\left(\frac{-1}{2\sigma^{2}}\left(a - \left(\mu+\sigma^{2}\right)\right)^{2}\right) da \\ &= \exp\left(\mu + \frac{\sigma^{2}}{2}\right) \Phi\left(-\frac{\mu}{\sigma} - \sigma\right). \end{aligned}$$

□

Proposition A.6. *Let* $\mathbf{Y}^{(d)} \left| \mathbf{X}^{(d)} \right. \sim N\left(\mathbf{X}^{(d)}, \sigma^{2}\left(d\right) I_{d}\right)$*, where* $\mathbf{X}_{j}$ *is distributed according to the density* $\theta_{j}\left(d\right) f\left(\theta_{j}\left(d\right) x_{j}\right)$ *for* $j = 1, \ldots, d$*. If* $\varepsilon\left(d, X_{j}, Y_{j}\right)$ *is as in*

(6.2), then we have

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

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

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

Using Fubini's Theorem and swapping $\mathbf{y}^{(n-)}$ and $\mathbf{x}^{(n-)}$ then yield the desired result.

□

Proposition A.7. *Let $\varepsilon\left(d,X_j,Y_j\right)$, $j=1,\ldots,n$ be as in (6.2). If $\lambda_j<\alpha$, then $\varepsilon\left(d,X_j,Y_j\right)\rightarrow_p 0$.*

Proof. By Taylor's Theorem, we have for some $U_j\in(X_j,Y_j)$ or (Y_j,X_j)

$$\begin{aligned}
&\mathrm{E}\left[\left|\varepsilon\left(d,X_j,Y_j\right)\right|\right]=\mathrm{E}\left[\left|\left(\log f\left(\theta_j\left(d\right)X_j\right)\right)'\left(Y_j-X_j\right)+\right.\right.\\
&\quad\left.\left.\frac{1}{2}\left(\log f\left(\theta_j\left(d\right)X_j\right)\right)''\left(Y_j-X_j\right)^2+\frac{1}{6}\left(\log f\left(\theta_j\left(d\right)U_j\right)\right)'''\left(Y_j-X_j\right)^3\right|\right].
\end{aligned}$$

Applying changes of variables and using the fact that $\left|(\log f(X))''\right|$ and $\left|(\log f(U))'''\right|$ are bounded by a constant, we get for some $K > 0$

$$\mathrm{E}\left[\left|\varepsilon\left(d, X_j, Y_j\right)\right|\right] \leq \ell \frac{d^{\lambda_j/2}}{d^{\alpha/2}} K \mathrm{E}\left[\left|(\log f(X))'\right|\right] + \left(\ell^2 \frac{d^{\lambda_j}}{d^{\alpha}} + \ell^3 \frac{d^{3\lambda_j/2}}{d^{3\alpha/2}}\right) K.$$

By assumption, $\mathrm{E}\left[\left|(\log f(X))'\right|\right]$ is bounded by some finite constant. Since $\lambda_j < \alpha$, the previous expression converges to 0 as $d \to \infty$. To complete the proof of the proposition we use Markov's inequality and find that $\forall \epsilon > 0$, $\mathrm{P}\left(\left|\varepsilon\left(d, X_j, Y_j\right)\right| \geq \epsilon\right) \leq \mathrm{E}\left[\left|\varepsilon\left(d, X_j, Y_j\right)\right|\right] / \epsilon \to 0$ as $d \to \infty$. □

Proposition A.8. *Let $R_i\left(d, \mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)$, $i = 1, \ldots, m$ be as in (6.7). We have $\sum_{i=1}^{m} R_i\left(d, \mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right) \to_p E_R$, where E_R is as in (3.3).*

Proof. For $i = 1, \ldots, m$, we have

$$\begin{aligned}
\mathrm{E}\left[R_i\left(d, \mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)\right] &= \frac{1}{d^{\alpha}} \int_{\mathbf{R}} \ldots \int_{\mathbf{R}} \sum_{j \in \mathcal{J}(i,d), j \neq i^*} \left(\frac{d}{dx_j} \log \theta_j(d)\, f\left(\theta_j(d)\, x_j\right)\right)^2 \\
&\qquad \times \prod_{k \in \mathcal{J}(i,d), k \neq i^*} \theta_k(d)\, f\left(\theta_k(d)\, x_k\right) dx_k \\
&= \frac{\theta^2_{n+i}(d)}{d^{\alpha}} \sum_{j \in \mathcal{J}(i,d), j \neq i^*} \int_{\mathbf{R}} \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\, dx,
\end{aligned}$$

and writing the integral as an expectation yields

$$\mathrm{E}\left[R_i\left(d, \mathbf{X}^{(d)-}_{\mathcal{J}(i,d)}\right)\right] = \frac{c\left(\mathcal{J}(i,d)\right)}{d^{\alpha}} \frac{d^{\gamma_i}}{K_{n+i}} \mathrm{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right].$$

In the limit, we obtain

$$\begin{aligned} E_R &\equiv \lim_{d\to\infty} \sum_{i=1}^{m} \mathrm{E}\left[R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)\right] \\ &= \lim_{d\to\infty} \sum_{i=1}^{m} \frac{c\left(\mathcal{J}\left(i,d\right)\right)}{d^{\alpha}} \frac{d^{\gamma_i}}{K_{n+i}} \mathrm{E}\left[\left(\frac{f'\left(X\right)}{f\left(X\right)}\right)^2\right] < \infty. \end{aligned}$$

Since all X_j's are independent, the variance satisfies

$$\begin{aligned} &\mathrm{Var}\left(\sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)\right) = \\ &\sum_{i=1}^{m} \frac{1}{d^{2\alpha}} \sum_{j\in\mathcal{J}(i,d), j\neq i^*} \mathrm{Var}\left(\left[\left(\log\theta_j\left(d\right) f\left(\theta_j\left(d\right) X_j\right)\right)'\right]^2\right), \end{aligned}$$

and using the fact that $\mathrm{Var}\left(X\right) \leq \mathrm{E}\left[X^2\right]$ along with a change of variable yield

$$\begin{aligned} \mathrm{Var}\left(\sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)\right) &\leq \sum_{i=1}^{m} \frac{1}{d^{2\alpha}} \sum_{j\in\mathcal{J}(i,d), j\neq i^*} \theta_j^4\left(d\right) \mathrm{E}\left[\left[\left(\log f\left(X\right)\right)'\right]^4\right] \\ &= \sum_{i=1}^{m} \frac{1}{d^{2\alpha}} \frac{d^{2\gamma_i}}{K_{n+i}^2} c\left(\mathcal{J}\left(i,d\right)\right) \mathrm{E}\left[\left(\frac{f'\left(X\right)}{f\left(X\right)}\right)^4\right]. \end{aligned}$$

By assumption, we know that the expectation involved in the previous expression is finite. Since $c\left(\mathcal{J}\left(i,d\right)\right) d^{\gamma_i} \leq d^{\alpha}$ and $c\left(\mathcal{J}\left(i,d\right)\right) \to \infty$ as $d \to \infty$ for $i = 1, \ldots, m$, the variance thus converges to 0 as $d \to \infty$.

To conclude the proof of the lemma, we use Chebychev's inequality and find that for all $\epsilon > 0$

$$\mathrm{P}\left(\left|\sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right) - E_R\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \mathrm{Var}\left(\sum_{i=1}^{m} R_i\left(d, \mathbf{X}_{\mathcal{J}(i,d)}^{(d)-}\right)\right) \to 0$$

as $d \to \infty$. □

Proposition A.9. *Let* $\mathbf{X}_j$ *be distributed according to the density* $\theta_j f(\theta_j x_j)$ *for* $j = 1, \ldots, d$ *and also let* $\mathbf{Y}^{(d)} \left| \mathbf{X}^{(d)} \right. \sim N\left(\mathbf{X}^{(d)}, \sigma^2(d) I_d\right)$. *We have*

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

Proof. Developing the LHS leads to

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

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

Appendix B: R Functions

In this section, we present examples of the programming involved in the implementation of the RWM algorithm.

The function **gam** that follows has been used to obtain the curves from the finite-dimensional RWM algorithm in Figure 3.5, while the function **gamtheo** has produced the theoretical curves of the same example. The last function presented, **norm**, has been used to sample from the (finite-dimensional) normal-normal hierarchical model of Example 3.2.4 (see Figure 3.6).

```
gam <- function(it = 500000, dim = 100, ep = 0, n = 50) {
     ell <- 10 + (1:n) * 3
     vsd <- ell/dim^ ep
     a <- 5
     lamb <- c(rep(1/sqrt(dim)/5, dim - 2), 1, 1)
     ll <- n * dim
     x <- rgamma(ll, a, rep(lamb, n))
     xp1 <- x[((0:(n-1)) * dim) + 1]
     sx1 <- 0
     accr <- rep(0, n)
     for(i in 1:(it - 1)) {
          y <- rnorm(ll, mean = x, sd = rep(vsd^ 0.5, each = dim))
          y[y < 0] <- rep(0, ll)[y < 0]
          vv <- cumsum((y == 0) * 1)[(1:n) * dim]
          vv <- diff(c(0, vv))
          y <- (rep(vv, each = dim) == 0) * y
          yl <- y
```

```
        yl[yl == 0] <- x[yl == 0]
        suml <- cumsum(log(yl) - log(x))[(1:n) * dim]
        suml <- diff(c(0, suml))
        sumxy <- cumsum(rep(lamb, n) * (yl - x))[(1:n) * dim]
        sumxy <- diff(c(0, sumxy))
        alphal <- pmin((a - 1) * suml - sumxy, 0) + (vv != 0)
        * (-10000000)
        rrl <- log(runif(n))
        x[rep(rrl, each = dim) < rep(alphal, each = dim)] <-
        y[rep(rrl, each = dim) < rep(alphal, each = dim)]
        sx1 <- sx1 + (x[((0:(n-1)) * dim) + 1] - xp1)^ 2
        xp1 <- x[((0:(n-1)) * dim) + 1]
        accr <- accr + (rrl < alphal)
    }
    accr <- accr/it
    par(mfrow = c(2, 1))
    plot(ell, dim^ ep * sx1/it)
    plot(accr, dim^ ep * sx1/it)
    return(ell, accr, dim^ ep * sx1/it)
}


gamtheo <- function(it = 500000, dim = 5000, ep = 0, n = 50) {
    ell <- 10 + (1:n) * 3
    ER <- 1/75
    vsd <- ell/dim^ ep
    a <- 5
    esp <- 0
    for(i in (1:it)){
        x <- rgamma(2 * n, a, 1)
        y <- rnorm(2 * n, x, rep(vsd^ 0.5, each = 2))
        eps <- matrix(log(dgamma(y, a, 1)/dgamma(x, a, 1)),
        2, n)
        eps <- (eps[1, ] + eps[2, ] - ell * ER/2)/sqrt(ell
        * ER)
        esp <- esp + pnorm(eps)
    }
    ellhat <- ell[which.max(2 * ell * esp/it)]
    OAR <- 2 * esp[which.max(2 * ell * esp/it)]/it
    plot(ell, 2 * ell *esp/it, pch = 16)
    return(ell, 2 * esp/it, 2 * ell * esp/it, which.max(2
```

```
     * ell * esp/it), ellhat, OAR)
}


hiernorm <- function(it = 100000, dim = 10, n = 50, ell
= (1:50) * 0.2) {
     vsd <- ell/dim
     mu1 <- 0
     s1 <- 1
     s2 <- 1
     ll <- n * dim
     mp <- rnorm(n, mu1, s1)
     x <- rnorm(ll, rep(mp, each = dim), s2)
     xp1 <- x[((0:(n-1)) * dim) + 1]
     sx <- 0
     accr <- rep(0, n)
     for(i in 1:(it - 1)) {
          y1 <- rnorm(ll, mean = x, sd = rep(vsd^ 0.5,
          each = dim))
          y2 <- rnorm(n, mean = mp, sd = vsd^ 0.5)
          sumxy <- cumsum((x - rep(mp, each = dim))^ 2 - (y1 -
          rep(y2, each = dim))^ 2)[(1:n) * dim]
          sumxy <- diff(c(0, sumxy))
          alpha1 <- pmin(exp((sumxy/s2 + ((mp - mu1)^ 2 - (y2 -
          mu1)^ 2)/s1)/2), 1)
          rr1 <- runif(n)
          mp[rr1 < alpha1] <- y2[rr1 < alpha1]
          x[rep(rr1, each = dim) < rep(alpha1, each = dim)] <-
          y1[rep(rr1, each = dim) < rep(alpha1, each = dim)]
          sx <- sx + (x[((0:(n-1)) * dim) + 1] - xp1)^ 2
          xp1 <- x[((0:(n-1)) * dim) + 1]
          accr <- accr + (rr1 < alpha1)
     }
     accr <- accr/it
     par(mfrow = c(1, 1))
     plot(ell, dim * sx/it)
     plot(accr, dim * sx/it)
     return(ell, accr, dim * sx/it)
}
```

Bibliography

[1] Andrieu, C., Moulines, E. (2003). On the ergodicity properties of some adaptive Markov chain Monte Carlo algorithms. *To appear in Ann. Appl. Probab.*

[2] Atchadé, Y.F., Rosenthal, J.S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli.* **11**, 815-28.

[3] Barker, A.A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Aust. J. Phys.* **18**, 119-33.

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

[5] Bédard, M. (2006). Optimal Acceptance Rates for Metropolis Algorithms: Moving Beyond 0.234. Submitted for publication.

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

[7] Bédard, M. (2006). On the Optimization of Metropolis Algorithms for Hierarchical Target Distributions. In preparation.

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

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

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

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

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

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

[14] Cowles, M.K., Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.* **91**, 883-904.

[15] Cowles, M.K., Roberts, G.O., Rosenthal, J.R. (1999). Possible biases induced by MCMC convergence diagnostics. *J. Stat. Comput. Simul.* **64**, 87-104.

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

[17] Geman, S., Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Trans. Pattern Analysis and Machine Intelligence.* **6**, 721-41.

[18] Gelfand, A.E., Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

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

[20] Haario, H., Saksman, E., Tamminene, J. (2001). An adaptive Metropolis algorithm. *Bernoulli.* **7**, 223-42.

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

[22] Jacquier, E., Polson, N.G., Rossi, P.E. (2002). Bayesian analysis of stochastic volatility models. *J. Bus. Econom. Statist.* **20**, 69-87.

[23] Mengersen, K.L., Tweedie, R.L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101-21.

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

[25] Mira, A., Geyer, C.J. (1999). Ordering Monte Carlo Markov Chains. Technical Report 632, School of Statistics, Univ. Minnesota

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

[27] Pasarica, C., Gelman A. (2003). Adaptively scaling the Metropolis algorithm using expected squared jumped distance. Technical Report, Department of Statistics, Columbia University.

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

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

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

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

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

[33] Roberts, G.O., Rosenthal, J.S. (2005). Coupling and ergodicity of adaptive MCMC. Preprint.

[34] Roberts, G.O., Tweedie, R.L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika.* **83**, 95-110.

[35] Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **90**, 558-66.

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

[37] Ross, S.M. (1997). *Simulation.* Academic Press, California.

[38] Ross, S.M. (2003). *Introduction to Probability Models, 8th ed.* Academic Press, California.

[39] Tanner, M., Wong, W. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**, 528-50.

[40] Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701-62.