THE UNIVERSITY OF
WARWICK

AUTHOR: **Alexandre H. Thiéry** DEGREE: **Ph.D.**

TITLE: **Scaling Analysis of MCMC algorithms**

DATE OF DEPOSIT: .....

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: .....

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

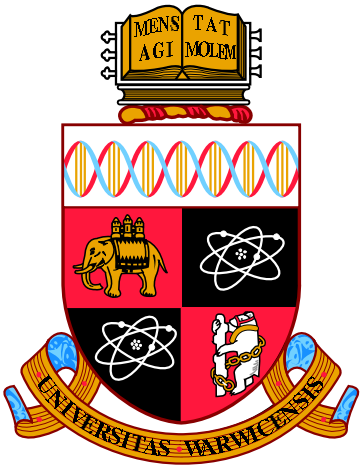
.....

.....

.....

.....

.....



Scaling Analysis of MCMC algorithms

by

Alexandre H. Thiéry

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

February 2013

THE UNIVERSITY OF WARWICK

Contents

List of Figures

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisors Gareth Roberts and Andrew Stuart for their enthusiasm, encouragement and for introducing me to such exciting topics. Their support and guidance have been invaluable and I have very much enjoyed our regular meetings. The discussions and feedback during the course of this thesis were immensely appreciated. I am also grateful to them both for their trust in allowing me the freedom to follow my ideas.

I am indebted to all of my collaborators and colleagues over these three years at Warwick. Special thanks go to Alex Beskos, Natesh Pillai, Krys Latuszynski and Sebastian Vollmer.

I will remember all the friends met during these last few years. Thank you Ashish for all these coffees and this unforgettable trip, Emmanuel for being such a good friend, Miro and Holly for the BBQs and these wonderful summer days, the salsa team Kris, Anton and Tom for making Leamington what it is, Martin for all these squash breaks, Mano for the tennis-ski-pool sessions, Maks for all those years in Paris.

Words fail me to express my gratitude to my parents for their constant support and for their teachings. My parents in laws are simply the best. Indeed, this thesis would have never been possible without the constant encouragements and love that my wife brings me everyday.

To Thanh and Sophia.

Declarations

This declaration confirms that this thesis is original and sole work of the author alone. The thesis does not include any previous material submitted by any other researcher in any form not acknowledged as required by existing regulations. No material contained in this thesis has been used elsewhere for publication prior the production of this work. This declaration also officially affirms that this thesis is being submitted for the degree of Doctor of Philosophy of the University of Warwick only and not to any other similar institution of higher learning for the same purposes.

Alexandre Thiéry

Abstract

Markov Chain Monte Carlo (MCMC) methods have become a workhorse for modern scientific computations. Practitioners utilize MCMC in many different areas of applied science yet very few rigorous results are available for justifying the use of these methods. The purpose of this dissertation is to analyse random walk type MCMC algorithms in several limiting regimes that frequently occur in applications. Scaling limits arguments are used as a unifying method for studying the asymptotic complexity of these MCMC algorithms. Two distinct strands of research are developed: (a) We analyse and prove diffusion limit results for MCMC algorithms in high or infinite dimensional state spaces. Contrarily to previous results in the literature, the target distributions that we consider do not have a product structure; this leads to Stochastic Partial Differential Equation (SPDE) limits. This proves among other things that optimal proposals results already known for product form target distributions extend to much more general settings. We then show how to use these MCMC algorithms in an infinite dimensional Hilbert space in order to imitate a gradient descent without computing any derivative. (b) We analyse the behaviour of the Random Walk Metropolis (RWM) algorithm when used to explore target distributions concentrating on the neighbourhood of a low dimensional manifold of $\mathbb{R}^n$. We prove that the algorithm behaves, after being suitably rescaled, as a diffusion process evolving on a manifold.

Chapter 1

Introduction

1.1 Motivation

The use of Markov Chain Monte Carlo (MCMC) methods for high dimensional and intractable computations has revolutionised applied mathematics in general and Bayesian statistics in particular. MCMC has been called one of the ten most important algorithms of the twentieth century [Cip00]. Since its first appearance in the statistical physics literature [MRTT53], MCMC techniques have opened new horizons in various fields of application such as biostatistics, computer science, physics, economics, finance, and applied statistics.

The power of MCMC methods reside in the simplicity of the underlying principles and the wide range of applications: in order to (approximately) compute expectations with respect to a given probability distribution called the *target distribution*, it suffices to build a Markov chain that is ergodic with respect to this target distribution and let the Markov chain run long enough. Moreover, the Metropolis-Hastings algorithm shows that it is straightforward to construct Markov chains that are ergodic with respect to a given target distribution: the problem is choice! Indeed, the choice of the Markov kernel can drastically influence the performance of the algorithm. For complex target distributions, it has become crucial to understand as precisely as possible how fast the Markov chain converges to equilibrium.

The design, tuning and analysis of efficient Markov chains lead to fascinating mathematics and rest upon a surprisingly wide range of ideas including representation theory [DS81; DH92], Fourier analysis [Dia88], micro-local analysis [DL09], functional analysis [SC97], partial differential equations [BCG08], optimal transport [EMM12], stochastic partial differential equations [HSVW05; HSV07], Riemannian geometry [GC11].

There exists a large literature on MCMC methods and practitioners now have many different Markov kernels to choose from. In practice, the ease of implementation and wide applicability have conferred their popularity to random walk type proposals. A downside of their versatility is however the potential slowness of their convergence, which calls for an analysis of their performances. In this dissertation, we shall study the behaviour of various random walk type algorithms in several limiting regime.

1.2 Contributions of the thesis

Some of the research contributions included in this dissertation can be summarised as follows.

1. We build upon ideas of [MPS11] and develop a framework for proving diffusion limits for MCMC algorithms that is general enough to tackle infinite dimensional examples. The main result is Proposition 3.3.1. Contrarily to other approaches, Markov chains that do not evolve at stationarity can be analysed without difficulties. This framework is subsequently used in chapters 4 and 5 for obtaining infinite dimensional diffusion limits.

2. We significantly extend the analysis of the Metropolis Adjusted Langevin Algorithm (MALA). Our result (theorem 4.2.4) can tackle non product form target densities and describes infinite dimensional scaling limits.

3. In chapter 5 we design an algorithm that imitates a gradient flow in an infinite dimensional Hilbert space. The algorithm does not need to compute any gradient. A rigorous analysis (theorem 5.2.2) of the algorithm through a scaling arguments is obtained.

4. We analyse MCMC methods that are designed to evolve on infinite dimensional state space. We adopt the 'optimize then discretize' viewpoint and produce an MCMC algorithm whose performances do not degenerate as the dimension of the discretisation increases.

5. In chapter 6 we analyse MCMC algorithm on target distributions that are concentrated on the neighbourhood of a low dimensional manifold. To the best of our knowledge, this is the first time that this setting that often appears in practice is analysed.

6. We develop a method of proof based on a separation of time scales for analysing MCMC algorithms that can lead to non constant volatility diffusion limits (theorem 6.2.3).

1.3 Organisation of the thesis

The rest of the dissertation is organised as follows.

- **Chapter 2**
  We give a brief reminder on the MCMC method with special emphasis on defining the Metropolis-Hastings algorithm on a general space. We then describe classical results on convergence of Markov chains (different notions of convergence, Markov CLT, spectral methods, conductance bounds). The concept of Expected Squared Jumping Distance (ESJD), of special importance in this dissertation, is then introduced. The chapter is concluded by a general presentation of the *scaling limit method* for analysing MCMC algorithms.

- **Chapter 3**
  Gaussian measures on infinite dimensional Hilbert spaces are introduced. We then described how one can define a distribution as a change of probability with respect to such a Gaussian measure and then find finite dimensional approximations of it. We then give a very brief introduction to stochastic differential equations (SDEs) evolving in a Hilbert space. With the applications that we have in mind, only a very restricted class of SDEs is described. We then conclude the chapter by proving a diffusion-approximation result (proposition 3.3.1) that is used at several places in the dissertation for proving infinite dimensional scaling limit results.

- **Chapter 4**
  This chapter is joint work with Andrew Stuart and Natesh Pillai and is based on the article [PST12]. We prove a diffusion limit for the output of the Metropolis Adjusted Langevin Algorithm (MALA) towards a Hilbert space valued SDE, when applied to $N-$dimensional approximations of an infinite dimensional target distribution. This implies, among other things, that the complexity of the MALA algorithm scales as $\mathcal{O}(N^{1/3})$ with the dimension N of the approximation. Moreover we show that the speed of the limiting diffusion is maximized for an average acceptance probability of 0.574, just as in the i.i.d product scenario [RR98]. Thus in this regard, our work is the first extension of the remarkable results in [RR98] for the Langevin algorithm to target measures

which are not of product form. This adds theoretical weight to the results observed in computational experiments [BR07; RR01; Béd08] which demonstrate the robustness of the optimality criteria developed in [RGG97; RR98].

- **Chapter 5**
  This chapter is joint work with Andrew Stuart and Natesh Pillai and is based on the article [PST]. There are many applications where it is of interest to find global or local minima of a functional

$$J(x) = \frac{1}{2}\|C^{-1/2}x\|^2 + \Psi(x) \tag{1.3.1}$$

  where C is a self-adjoint, positive and trace-class linear operator on a Hilbert space $\mathcal{H}$ and $\Psi : \mathcal{H} \to \mathbb{R}$ is a functional of interest. Gradient flow or steepest descent is a natural approach to this problem, but in its basic form requires computation of the gradient of Ψ which, in some applications, may be an expensive or complex task. In addition, when multiple minima are present, it may be important to include noise within the algorithm in order to allow escape from local minima. We show in this chapter how a noisy gradient descent can emerge from certain carefully specified random walks, when combined with a Metropolis-Hastings accept-reject mechanism, with tunable noise level τ. We analyse this algorithm through a scaling limit argument.

- **Chapter 6**
  This chapter is joint work with Alex Beskos and Gareth Roberts and is based on the article [BRT13]. It often happens in applied probability that one needs to explore a target distribution π that is concentrated on a very narrow subset of a state space. In this chapter, we consider the continuous setting where the target distribution π lives in the n-dimensional euclidean space $\mathbb{R}^n$ and concentrates on the neighbourhood of a low dimensional manifold $\mathcal{M}$; this means that there exists $\varepsilon \ll 1$ such that the ε-neighbourhood $A_\varepsilon := \{x \in \mathbb{R}^n : d(x, \mathcal{M}) < \varepsilon\}$ of the manifold $\mathcal{M}$ verifies $\pi(A_\varepsilon) \approx 1$. A Random Walk Metropolis (RWM) Markov chain will tend to walk along the manifold. The purpose of this chapter is to quantify this behaviour. In this chapter we focus on the limiting regime when the thickness ε of the neighbourhood A_ε of the limiting manifold $\mathcal{M}$ converges to zero. The influence of the size of the jumps is analysed by adopting the Expected Squared Jumping Distance ESJD as measure of efficiency and by proving diffusion limits. The main finding is that in the majority of the cases, in order to explore the target distribution, it is optimal to choose the size of the jumps of the same order of magnitude as the

thickness ε. For this choice, we prove a diffusion limit result (Theorem 6.2.3). This gives quantitative estimates on the complexity of the RWM algorithm when applied to target concentrating near a manifold. For simplicity, all the rigorous results are proved for the case where the manifold $\mathcal{M}$ is flat. We present conjectures and numerical illustrations for the general case. To the best of our knowledge, this is the first time that a diffusion approximation for MCMC algorithm leads to a diffusion limit with non-constant volatility. The proof is based on a time-scale separation argument.

1.4 Notation

We use the standard convention whereby capital letters denote random variables, whereas lower case letters are used for their values. We adopt a slight abuse of notation by referring to densities as distributions, and where convenient, employ the measure-theoretic notations $\mu(A) = \int_A \mu(dx)$ and $\mu(f) = \int f(x)\,\mu(dx)$.

Throughout the paper we use the following notation in order to compare sequences and to denote conditional expectations.

- Two positive sequences $\{\alpha_n\}_{n\geq 0}$ and $\{\beta_n\}_{n\geq 0}$ are equivalent, $\alpha_n \sim \beta_n$, if the following limit holds $\lim_{n\to\infty} \alpha_n/\beta_n = 1$.
- Two sequences $\{\alpha_n\}_{n\geq 0}$ and $\{\beta_n\}_{n\geq 0}$ satisfy $\alpha_n \lesssim \beta_n$ if there exists a constant $K > 0$ satisfying $\alpha_n \leq K\beta_n$ for all $n \geq 0$. The notations $\alpha_n \asymp \beta_n$ means that $\alpha_n \lesssim \beta_n$ and $\beta_n \lesssim \alpha_n$.
- Two sequences of real functions $\{f_n\}_{n\geq 0}$ and $\{g_n\}_{n\geq 0}$ defined on the same set D satisfy $f_n \lesssim g_n$ if there exists a constant $K > 0$ satisfying $f_n(x) \leq K g_n(x)$ for all $n \geq 0$ and all $x \in D$. The notation $f_n \asymp g_n$ means that $f_n \lesssim g_n$ and $g_n \lesssim f_n$.
- The notation $\mathbb{E}_x\big[f(X,\xi)\big]$ denotes expectation with respect to ξ conditionally upon the event $X = x$.

Chapter 2

Probabilistic toolbox

2.1 Markov chain Monte Carlo methods

We assume the reader familiar with the basic Markov Chain Monte Carlo (MCMC) method [MRTT53; Has70]. See [Tie94; SR93; Dia09] for an introduction, [GRS96; RC04; Liu08] for book-length treatments and the reference [MT93] for technical developments. In this section we give a quick reminder of MCMC methods on general state spaces and introduce the main notations. In this thesis, we will be dealing with MCMC on infinite dimensional Hilbert spaces and special care is necessary to properly define the Metropolis-Hastings in such situations.

Consider a measured space $(\mathcal{X}, \mathcal{B})$, a σ-finite probability distribution π on $\mathcal{X}$ and a proposal kernel $q(x, dy)$. For each $x \in \mathcal{X}$ the quantity $q(x, \cdot)$ defines a probability distribution on $\mathcal{X}$. The MCMC algorithm requires an accept-reject function $\alpha : \mathcal{X} \times \mathcal{X} \to [0, 1]$. If the current state is x, then a candidate y for the next state is generated from $q(x, dy)$ and accepted with probability $\alpha(x, y)$. The resulting transition kernel

$$P(x, dy) = \alpha(x, y)\, q(x, dy) + \big(1 - \alpha(x)\big)\, \delta_x(dy) \tag{2.1.1}$$

where δ_x is the Dirac mass at x and $\alpha(x) := \int_y \alpha(x, y)\, q(x, dy)$ is the mean acceptance probability at x. The accept-reject function $\alpha(\cdot, \cdot)$ is chosen so that the transition kernel $P(x, dy)$ is reversible with respect to the target distribution π,

$$\int_{A \times B} \pi(dx)\, P(x, dy) = \int_{A \times B} \pi(dy)\, P(y, dx) \tag{2.1.2}$$

for all measurable subsets $A, B \subset \mathcal{X}$. Suppose that there exists a symmetric domi-

nating measure $\nu(dx, dy)$ on $\mathcal{X} \times \mathcal{X}$ and write $f(x, y)$ for the Radon-Nikodym derivative of the measure $\pi(dx)\, q(x, dy)$ with respect to $\nu(dx, dy)$. It is proved in [Tie94] that under the assumption that the accept-reject function α satisfies

$$\alpha(x, y) f(x, y) = \alpha(y, x)\, f(y, x) \tag{2.1.3}$$

the Markov kernel $P(x, dy)$ is reversible with respect to π. The Metropolis-Hastings algorithms corresponds to the choice

$$\alpha(x, y) \;=\; \min\Big\{1, \frac{f(y, x)}{f(x, y)}\Big\}.$$

The possibility that the denominator of the above ratio is zero is not a concern since for such pair (x, y) there is zero probability to propose such a move. In such a situation, any value for the quotient $\frac{f(y,x)}{f(x,y)}$ can be chosen without affecting the reversibility condition (2.1.3). It should be noted that reversibility with respect to π does not imply that the Markov chain converges (see next section for the different notions of convergence) to π. The most usual situation is where there is a common dominating measure μ with $\pi(dx) = \pi(x)\, \mu(dx)$ and $q(x, dy) = q(x, y)\, \mu(dy)$. The choice $\nu = \mu \otimes \mu$ shows that in this case one can choose

$$\alpha(x, y) \;=\; \min\Big\{1, \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)}\Big\}.$$

This remark is implicitly used at several places in this thesis where μ is a Gaussian measure on an infinite dimensional Hilbert space.

2.2 Convergence of Markov chains

This section is a quick reminder on the different ways of measuring the speed at which a Markov chain convergences to equilibrium. Since we are mainly interested in Metropolis-Hastings Markov chains, the focus is on reversible Markov chains. The main purpose of this section is to show that it is in general extremely difficult to obtain accurate rates of convergence. In this thesis, we investigate situations where this analysis becomes possible through diffusion limits arguments.

2.2.1 Convergence theorem

Consider a Markov chain $X = \{X_k\}_{k \geq 0}$ on the state space $\mathcal{X}$ with transition probabilities $P(x, dy)$. The Markov chain is assumed to be reversible with respect to the

probability π,

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

Since X is reversible with respect to π, the distribution π is also a stationary distribution in the sense that if $X_0 \overset{\mathcal{D}}{\sim} \pi$ then $X_k \overset{\mathcal{D}}{\sim} \pi$ for $k \geq 1$. To measure how quickly the Markov chain converges to equilibrium, one needs to introduce a metric on the space of probability distributions on $\mathcal{X}$. A popular choice is the *total variation distance* defined by $d_{TV}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$ where the supremum runs over all the measurable subsets $A \subset \mathcal{X}$. This choice of distance is especially useful for studying Markov chains since it fits very well with the concept of coupling that is an important ingredient of many convergence results. It is interesting to notice (Proposition 3 of [RR04]) that a Markov operator P is always a contraction in the space of probability measures in the sense that $d_{TV}\big(\mu P, \nu P\big) \leq d_{TV}(\mu,\nu)$ for any probability measure μ and ν. Under *irreducibility* and *aperiodicity* conditions, the Markov chain X converges to equilibrium in a sense made precise below.

Definition 2.2.1. (Irreducibility and periodicity)

- *A Markov chain is φ-irreducible if there exists a non-zero σ-finite measure φ on $\mathcal{X}$ such that for any subset $A \subset \mathcal{X}$ with positive measure $\varphi(A) > 0$, and for all $x \in \mathcal{X}$ there exists a positive integer $n = n(x,A)$ such that $P^n(x,A) > 0$.*
- *A Markov chain with stationary distribution π is aperiodic if there does not exist a period $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_d$ of positive π-measure satisfying $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ (indices modulo d).*

The irreducibility condition informally means that any state can be reached if one waits long enough. The aperiodicity condition means that the Markov chain does not move following cycles. As in the more classical finite state setting, the irreducibility and aperiodicity conditions ensure that the Markov chain converges to equilibrium.

Theorem 2.2.2. (General convergence theorem) *Let X be a Markov chain on a state space $\mathcal{X}$ with countably generated σ-algebra. Suppose that the Markov chain has a stationary distribution π and is φ-irreducible and aperiodic.*

- *For π-a.e. $x \in \mathcal{X}$ the sequence of probability measures $P^n(x,dy)$ converges towards π in the total variation distance, $\lim_{n\to\infty} d_{TV}\big(P^n(x,\cdot),\pi\big) = 0$.*

- *Let $f : \mathcal{X} \to \mathbb{R}$ a π-integrable function $\pi(|f|) < \infty$. For π-a.e. initial state $X_0 = x \in \mathcal{X}$, the law of large numbers holds, $\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} f(X_j) = \pi(f)$ almost surely. The aperiodicity condition is not needed for this result.*

A possible proof (see [MT93]) consists in showing that the irreducibility condition implies the existence of a small set. One can then use a coupling argument to finish the proof. The notion of mixing time $\tau := \inf_{t\geq 0} \sup_\mu d_{TV}(\mu P^t, \pi) \leq \frac{1}{4}$ is often introduced in the literature. Indeed, since one can prove that the function $d(t) := 2\times\sup_\mu d_{TV}(\mu P^t, \pi)$ is sub-multiplicative $d(s+t) \leq d(s)\, d(t)$, once the mixing time has been reached the Markov chain then converges to equilibrium exponentially quickly. The conclusion of Theorem 2.2.2 is only qualitative and thus typically of no great value since it does not describe the speed at which the convergence takes place. For a more precise description of convergence to equilibrium, we introduce the notion of *uniform* and *geometric* ergodicity.

Definition 2.2.3. (Uniform and geometric ergodicity) *Let X be Markov chain on $\mathcal{X}$ with stationary distribution π.*

- *The Markov chain is uniformly ergodic if there exists a constant $M < \infty$ such that $d_{TV}\big(P^n(x,\cdot), \pi\big) \leq M\, \rho^n$ for some $\rho < 1$.*
- *The Markov chain is geometrically ergodic if $d_{TV}\big(P^n(x,\cdot), \pi\big) \leq M(x)\, \rho^n$ for some constant $\rho < 1$ and function M satisfying $M(x) < \infty$ for π-a.e. $x \in \mathcal{X}$.*

Markov chains on continuous state spaces are very rarely uniformly ergodic. In practice, to prove that a Markov chain is geometrically ergodic one can sometimes use the notion of *drift condition*, which is a variant of the geometric *Foster-Lyapunov* condition [Fos53].

Definition 2.2.4. (Drift Condition) *The Markov kernel P satisfies a drift condition if there are constants $0 < \lambda < 1$ and $b < \infty$, and a Lyapunov function $V : \mathcal{X} \to [1, +\infty]$ satisfying*

$$PV \leq \lambda V + b\, 1_C$$

where C is a small set for the Markov kernel P.

We remind the reader that a small set $C \subset \mathcal{X}$ for the Markov kernel P is a set C such that there exist a constant $\varepsilon > 0$, an integer $n_0 \geq 1$ and a probability measure ν such that $P^{n_0}(x, A) \geq \varepsilon\nu(A)$ for every measurable subset $A \subset \mathcal{X}$ and $x \in C$. The drift condition 2.2.4 quantifies the way in which the process $\{V(X_k)\}_{k\geq 0}$

behaves as a supermartingale before that the Markov chain X enters the small set C. On average, the quantity $V(X_k)$ decreases at rate λ when outside of C, implying (since $V(x) \geq 1$) that the Markov chain X returns geometrically quickly to the small set C. When inside the small set C, a regeneration happens with probability at least ε. The drift condition 2.2.4 allows it to be shown by a coupling argument that the chain is geometrically ergodic.

Theorem 2.2.5. *Consider a φ-irreducible, aperiodic Markov chain X with stationary distribution π. Suppose that the drift condition 2.2.4 is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$ and a π-a.e. finite Lyapunov function V. Then the Markov chain X is geometrically ergodic.*

Establishing a drift condition is more an art than a general method, though. The interested reader is referred to [MT93] for a thorough description of this approach.

To estimate the statistical fluctuation of the estimator $S_N(f) = N^{-1} \sum_{k=0}^{N-1} f(X_j)$, it is useful to establish conditions that ensure that a central limit theorem holds. For reversible Markov chains, a central limit holds as soon as the asymptotic variance σ_f^2 is finite.

Theorem 2.2.6. *[KV86]* **(Kipnis-Varadhan)** *Let X be an ergodic Markov chain reversible with respect to the probability distribution π and $f \in L^2(\pi)$. Suppose that the chain is started at stationarity $X_0 \overset{\mathcal{D}}{\sim} \pi$ and that the asymptotic variance exists and is finite,*

$$\lim_{N\to\infty} \frac{1}{N} \, Var\Big(f(X_0) + \ldots + f(X_{N-1})\Big) = \sigma_f^2 < \infty.$$

The sequence $\sqrt{N}\,\big(S_N(f) - \pi(f)\big)$ converges weakly to a centred Gaussian distribution with variance σ_f^2. On every finite time interval, the sequence of rescaled processes

$$B^N(t) := \frac{1}{\sqrt{N}} \sum_{k<tN} \Big(f(X_k) - \pi(f)\Big)$$

converges in the Skorohod topology to a Brownian motion with variance σ_f^2.

Since the quantity $\frac{1}{N} \operatorname{Var}\Big(f(X_0) + \ldots + f(X_{N-1})\Big)$ also equals $\operatorname{Var}_\pi(f)\,\big(1 + 2\sum_{k=1}^{N-1}(1-k/N)\rho_f(k)\big)$ where $\rho_f(k)$ is the correlation at stationarity between $f(X_j)$ and $f(X_{j+k})$, it follows that σ_f^2 is finite if $\sum_{k=1}^{\infty}(1 - k/N)\big|\rho_f(k)\big| < \infty$. In other

words, if the autocorrelation sequence $\{\rho_f(k)\}_{k\geq 1}$ converges quickly enough to zero, a Markov central limit theorem holds. Corollary 2.1 of [RR97] shows that geometric ergodicity 2.2.3 of the reversible chain X is enough to guaranty that σ_f^2 exists and is finite for any functional $f \in L^2(\pi)$. For non reversible chains, one needs the stronger assumption that $f \in L^{2+\varepsilon}(\pi) < 0$ for some $\varepsilon > 0$ to ensure that a Markov central limit theorem holds. In the next section we give an expression for σ_f^2 in terms of the spectral decomposition of the Markov operator P.

2.2.2 Spectral analysis and consequences

This section describes the spectral approach to the study of the convergence of Markov chains. To keep the exposition simple and avoid the subtleties inherent to the spectral theory of linear operators on general Hilbert spaces, we only consider the case where the state space is discrete and finite $\mathcal{X} = \{1, 2, \ldots, n\}$. This limited setting is general enough to illustrate the main ideas of the general theory. Reversibility $\pi(x)\,P(x,y) = \pi(y)\,P(y,x)$ of the Markov chain X with respect to the probability distribution π implies that the Markov kernel P can be regarded as a self-adjoint linear operator, or matrix, on $L^2(\mathcal{X}, \pi)$. The operator P acts on functions $f : \mathcal{X} \to \mathbb{R}$ and measures μ on $\mathcal{X}$ as $\mu Pf = \sum_{x,y} \mu(x)P(x,y)f(y)$. For any functions $f, g : \mathcal{X} \to \mathbb{R}$ we have $\langle Pf, g\rangle_\pi = \langle f, Pg\rangle_\pi$ where $\langle\cdot,\cdot\rangle_\pi$ is the usual inner product in $L^2(\pi)$.

Since P is a self-adjoint operator in $L^2(\pi)$, there exists an orthonormal eigenbasis $(\varphi_1, \ldots, \varphi_n)$ of $L^2(\pi)$ with $P\varphi_j = \lambda_j\varphi_j$ for the real eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Since P is a Markov operator, the eigenvalues are less than one in absolute value. We assume that P is irreducible so that π is the unique invariant distribution; this ensures that $\lambda_1 = 1$ with $\varphi_1 = (1, 1, \ldots, 1)/\|(1, 1, \ldots, 1)\|_{L^2(\pi)} = (1, 1, \ldots, 1)$ and $\lambda_2 = 1 - \lambda_{\text{gap}} < 1$. The difference $\lambda_{\text{gap}} > 0$ between the first two eigenvalues is called the *spectral gap* of the Markov transition operator P. We also assume that P is aperiodic, which ensures than $\lambda_n > -1$. In other words, all the eigenvalues λ_j for $2 \leq j \leq n$ are strictly less than one in absolute value. With these notations we have $\pi(f) := \mathbb{E}_\pi(f) = \langle f, \varphi_1\rangle_\pi$ for any function $f : \mathcal{X} \to \mathbb{R}$.

Since the Markov chain is assumed to be reversible with respect to π and irreducible and aperiodic, the law of large numbers holds. For any function $f : \mathcal{X} \to \mathbb{R}$, the sequence $S_N(f) := (f(X_0) + \ldots + f(x_{N-1}))/N$ converges almost surely to $\pi(f)$ as $n \to \infty$. The knowledge of the spectrum of the operator P provides a quantitative estimate of the convergence of the empirical means to the expected

value $\pi(f)$. The function f can be decomposed on the orthonormal eigenbasis of P in the sense that $f = \pi(f)\,\varphi_1 + \sum_{j=2}^n \alpha_j \varphi_j$ (with $\alpha_j = \langle f, \varphi_j \rangle_\pi$). This implies that for any time $t \geq 0$ we have the decomposition $P^t f = \pi(f)\,\varphi_1 + \sum_{j=2}^n \lambda_j^t \alpha_j \varphi_j$. Notice that the term $\sum_{j=2}^n \lambda_j^t \alpha_j \varphi_j$ decreases geometrically quickly to zero at rate $\rho = \max_{j=2}^n |\lambda_j| < 1$. A similar computation would show that for any probability measure μ the sequence μP^t converges at rate ρ to the invariant distribution π. The quantity ρ is the spectral radius of the restriction P_0 of P to the subspace $L_0^2(\pi) = \big\{ f \in L^2(\pi) \,:\, \pi(f) = 0 \big\}$. The definition of uniform ergodicity 2.2.3 shows that on a finite state space an irreducible ergodic Markov chain is uniformly ergodic with rate ρ given by the spectral radius of P_0. It is common practice in the literature to study a lazy version $P_{\text{lazy}} = \frac{1}{2}\big(I + P\big)$ of the Markov operator P to ensure that the eigenvalues of P_{lazy} are non-negative and in this case we have $\rho = 1 - \lambda_{\text{gap}}$. Other conditions (e.g. Lemma 3.1 of [Bax05]) can sometimes help ensure that eigenvalues of P are non-negative. The advantage of having non-negative eigenvalues is that there exists many different techniques for estimating the spectral gap λ_{gap} while it is often more difficult to study the smallest eigenvalue λ_n. According to the central limit Theorem 2.2.6, the sequence $\sqrt{N}\,\big(S_N(f) - \pi(f)\big)$ converges weakly to a centred Gaussian distribution with variance σ_f^2. If the Markov chain is started at stationarity, the variance of $\sqrt{N}\,\big(S_N(f) - \pi(f)\big)$ can also be expressed as

$$\sum_{j=2}^{n} \big(1 + 2 \sum_{t=1}^{N-1} (1 - t/N) \lambda_j^t\big)\, \alpha_j^2 \to \sum_{j=2}^{n} \frac{1 + \lambda_j}{1 - \lambda_j}\, \alpha_j^2 =: \sigma_f^2.$$

This follows from the observation that the variance of $\sqrt{N}\,\big(S_N(f) - \pi(f)\big)$ equals $\frac{1}{N}\,\mathbb{E}\big(\sum_{t=0}^{N-1} f_0(X_t)\big)^2$, with $f_0 := f - \pi(f)\varphi_1 = \sum_{j=2}^n \alpha_j \varphi_j$, and at stationarity we have $\mathbb{E}\big[f_0(X_i) f_0(X_j)\big] = \langle f_0, P^{|j-i|} f_0 \rangle_{L^2(\pi)} = \sum_{j=2}^n \lambda_j^{|j-i|} \alpha_j^2$. Notice that it might happen that $\sigma_f^2 < \text{Var}_\pi(f) = \sum_{j=2}^n \alpha_j^2$. Indeed, if all the eigenvalues are positive we have $\sigma_f^2 > \text{Var}_\pi(f)$ for any function $f : \mathcal{X} \to \mathbb{R}$. Since $0 \leq \frac{1+\lambda}{1-\lambda} < (2 - \lambda_{\text{gap}})\lambda_{\text{gap}}^{-1}$ for $-1 \leq \lambda \leq 1 - \lambda_{\text{gap}}$, it follows that the asymptotic variance satisfies

$$\sigma_f^2 \leq (2 - \lambda_{\text{gap}}) \text{Var}_\pi(f) / \lambda_{\text{gap}}$$

for any function f. The bound is sharp since it is achieved if f is the multiple of the second eigenfunction φ_2. The spectral gap λ_{gap} is thus of great importance to the statistical analysis of MCMC algorithms. The variational characterisation of

the second eigenvalue

$$\lambda_2 = \sup_{L_0^2(\pi)} \langle f, Pf\rangle_\pi / \mathrm{Var}(f)$$

where $L_0^2(\pi)$ is the subspace of functions satisfying $\pi(f) = 0$ shows that the spectral gap $\lambda_{\mathrm{gap}} = 1 - \lambda_2$ can also be expressed as $\lambda_{\mathrm{gap}} = \inf_{L_0^2(\pi)} \langle f, (I-P)f\rangle_\pi / \mathrm{Var}(f)$. Using the reversibility of the operator P one can check that this also reads

$$\lambda_{\mathrm{gap}} = \inf \ \frac{1}{2}\,\mathbb{E}_\pi\big(f(X_1) - f(X_0)\big)^2 / \mathrm{Var} f =: \inf \ \mathcal{D}(f)/\mathrm{Var} f \tag{2.2.1}$$

with $X_0 \overset{\mathcal{D}}{\sim} \pi$. Notice that in Equation 2.2.1 one does not need to restrict the set of test functions to $L_0^2(\pi)$ since the term $f(X_1) - f(X_0)$ is not affected by the operation $f \mapsto f - \pi(f)$. The quantity $\mathcal{D}(f) := \frac{1}{2}\,\mathbb{E}_\pi\big(f(X_1) - f(X_0)\big)^2$ is called the *Dirichlet form* associated to the π-reversible transition operator P. The variational characterisation of the spectral gap 2.2.1 is useful since it immediately gives an upper bound for the spectral gap; indeed, the bound $\lambda_{\mathrm{gap}} \leq \mathcal{D}(f)/\mathrm{Var} f$ holds for any non trivial test function f. If one considers the set of test functions of the form $f = 1_A$ where $A \subset \mathcal{X}$, we quickly arrive to the notion of *conductance*. Indeed, one can check that for $f = 1_A$ we have $\mathcal{D}(f) = Q(A, A^c)$ where

$$Q(A, A^c) = \sum_{x,y} \pi(x) P(x,y) 1_A(x) 1_{A^c}(y)$$

is the probability, at stationarity, that the Markov chain jumps from the set A to its complement A^c. Indeed, this also reads $Q(A, A^c) = \mathbb{E}\big[1_A(X_k) 1_{A^c}(X_{k+1})\big]$ with $X_k \overset{\mathcal{D}}{\sim} \pi$. Also, since $\mathrm{Var}(f) = \pi(A)\,\pi(A^c)$ is bigger than $\min\big(\pi(A), \pi(A^c)\big)/2$, one can upper bound the spectral gap by $\lambda_{\mathrm{gap}} \leq 2\Phi(A)$ where we have defined $\Phi(A) := Q(A, A^c)/\min\big(\pi(A), \pi(A^c)\big)$. This leads to the upper bound $\lambda_{\mathrm{gap}} \leq 2\Phi$ with $\Phi := \inf_{A \subset \mathcal{X}} \Phi(A)$. Nevertheless, it is often of much greater interest to lower bound the spectral gap (e.g. to prove that a Markov chain mixes quickly). It was independently proved in [SJ89] and [LS88] that the quantity Φ also provides a lower bound for the spectral gap,

$$\frac{\Phi^2}{2} \leq \lambda_{\mathrm{gap}} \leq 2\Phi. \tag{2.2.2}$$

Since the quantity Φ is defined through an infimum, one can generally only find upper bounds for Φ, which makes the lower bound in the Cheeger's inequality (2.2.2) not as useful as one might think at first sight. Indeed, it is nevertheless a great way

of establishing negative results and prove that the spectral gap is small. In the same spirit, Proposition 2.16 of [HSV11] shows that the *acceptance probability* for Metropolis-Hastings algorithms is related to to spectral gaps through the bound

$$\lambda_{\text{gap}} \leq 2 \inf_{x \in \mathcal{X}} \alpha(x). \tag{2.2.3}$$

In equation (2.2.3), the quantity $\alpha(x)$ denotes the acceptance probability of a Metropolis-Hastings Markov chain as defined in equation (2.1.1). Equation (2.2.3) already shows that the tuning of the mean acceptance probability of MCMC algorithms is of fundamental importance to the analysis of MCMC algorithms. For example, if the mean acceptance probability is exponentially small then the spectral gap is exponentially small. This idea is a motivation for several results of this thesis.

2.3 Optimal proposals

To compare different MCMC algorithms, we need to discuss how to measure the efficiency of a particular MCMC transition kernel. Consider a target distribution π on the state space $\mathcal{X}$ and two reversible Markov chains X and Y with respective Markov transition kernel P_X and P_Y. Suppose further that these two Markov chains are geometrically ergodic so that a central limit holds for any function $f \in L^2(\pi)$,

$$N^{-\frac{1}{2}} \Big(\sum_0^{N-1} f(X_i) - \pi(f) \Big) \Rightarrow \mathrm{N}(0, \sigma^2_{X,f}) \quad \text{and} \quad N^{-\frac{1}{2}} \Big(\sum_0^{N-1} f(Y_i) - \pi(f) \Big) \Rightarrow \mathrm{N}(0, \sigma^2_{Y,f}).$$

Naturally, it would be natural to say that the Markov chain X is more efficient than the Markov chain Y if for any function $f \in L^2(\pi)$ the asymptotic variances $\sigma^2_{X,f}$ is less than $\sigma^2_{Y,f}$. In this case, we say that the kernel P_X dominates the kernel P_Y in the efficiency ordering and write $P_X \succcurlyeq P_Y$. This is indeed a very strong condition; in a finite state space setting, two Markov kernels P and Q that are reversible with respect to a probability distribution π can be ordered as $P_X \succcurlyeq P_Y$ if, and only if, their eigenvalues can be ordered as $\lambda^P_k \leq \lambda^Q_k$ for every $1 \leq k \leq n$ [Mir01; MG99]. It was proved by Peskun [Pes73] for finite state spaces, and by Tierney [Tie98] for general state spaces, that a sufficient condition for $P_X \succcurlyeq P_Y$ is that $P_X(z, A) \geq P_Y(z, A)$ for all $z \in \mathcal{X}$ and subset $A \subset \mathcal{X}$ with $z \notin A$; in other words, a sufficient condition for $P_X \succcurlyeq P_Y$ is that P_X dominates P_Y off the diagonal. Indeed, this condition is very strong and not often useful for comparing two Markov kernels P_X and P_Y. It does happen more often than not that two different Markov kernels P_X and P_Y reversible with respect to the same target probability measure π

cannot be compared through this criterion in the sense that one find two functions $f, g \in L^2(\pi)$ such that $\sigma^2_{X,f} > \sigma^2_{Y,f}$ and $\sigma^2_{X,g} < \sigma^2_{Y,g}$.

Another solution for comparing two reversible Markov transition P_X and P_Y reversible with respect to the same target probability π would be to compare their spectral gaps $\lambda_{\text{gap}}(X)$ and $\lambda_{\text{gap}}(Y)$. In view of our discussion of spectral gaps 2.2.2, we could say that P_X is more efficient than P_Y if the spectral gap $\lambda_{\text{gap}}(X)$ of P_X is larger than the spectral gap $\lambda_{\text{gap}}(Y)$ of P_Y. This is indeed a valid approach but spectral gaps are notoriously hard quantities to estimate.

In this thesis, we are mainly interested in MCMC methods which proceed via local moves. In other words, the proposals are small perturbations of the current state of the Metropolis-Hastings Markov chain. For complex target distributions, this is often the only type of proposals that can be efficiently implemented and the scale of the increments often has a dramatic influence on the complexity of the resulting MCMC algorithm. A simple heuristic suggests the existence of an "optimal scale": smaller values of the proposal variance lead to high acceptance rates but the chain does not move much even when accepted, and therefore may not be efficient. Larger values of the proposal variance lead to larger moves, but then the acceptance probability is tiny. The optimal scale for the proposal variance strikes a balance between making large moves and still having a reasonable acceptance probability. The next two sections introduces two related approaches to investigating the "optimal scale" for MCMC algorithms that evolves through local moves.

2.3.1 Expected squared jumping distance

Consider a Markov chain $X = \{X_k\}_{k\geq 0}$ evolving on the Hilbert space $\mathcal{X}$. The Markov chain X is assumed to be ergodic with invariant probability π. The Expected Squared Jumping Distance (ESJD) is the expected size of the squared jump size $\|X_{k+1} - X_k\|^2$ between two consecutive steps of the Markov chain X,

$$\text{ESJD} := \mathbb{E}\big\|X_{k+1} - X_k\big\|^2, \tag{2.3.1}$$

when the Markov chain is assumed to evolve at stationarity $X_k \overset{\mathcal{D}}{\sim} \pi$. Algebra reveals that the ESJD can also be expressed as $\text{ESJD} := 2\big(\rho(0) - \rho(1)\big)$ where $\rho(r)$ is the covariance $\mathbb{E}\langle X_k, X_{k+r}\rangle$. Since at stationarity the quantity $\rho(0) = \mathbb{E}\|X\|^2$ depends on the target probability π only (and not on the Markov kernel), maximising the ESJD is equivalent to minimising the first covariance coefficient $\rho(1) = \mathbb{E}\langle X_k, X_{k+1}\rangle$.

The ESJD has the advantage of being relatively straightforward to study in situations where it would typically be impossible to obtain meaningful information on the spectral gap of the Markov kernel or on the Monte Carlo asymptotic variance. In several cases, the ESJD can be used for analysing situations where the scaling approach (see section 2.3.2) fails. Indeed, one major disadvantage of the scaling limit approach is its reliance on asymptotics in the dimensionality of the problem; the majority of the results obtained through the scaling approach considers high dimensional limits where each coordinate evolves asymptotically independently from the others. On the contrary, the ESJD can tackle situations where coordinates are highly correlated. For example, the article [SR09] gives a non-asymptotic formula for the ESJD of the Random Walk Metropolis (RWM) algorithm on spherically symmetric unimodal distributions; as a corollary, it is proved in this case that there exists a unique scale that maximises the ESJD. In the same line, [She13] gives conditions under which the 0.234 rule of [RGG97] holds for much more general target distributions than the one that can be analysed through diffusion limits. The article [NR11] gives non-asymptotic results for several random walk type MCMC algorithms with non-Gaussian proposals. In [BRS09], the ESJD is used to analyse non-product form target distributions that are discretisation of infinite dimensional probability distributiond; chapter 4 of this thesis gives diffusion limit justifications of some results of [BRS09].

One should nevertheless keep in mind that the ESJD analysis gives in general much less intuition on the behaviour of a MCMC algorithm than a scaling limit result. Indeed, the understanding of the subtle path properties of a MCMC algorithm that can be gained through a scaling limit result are typically completely unavailable through an ESJD analysis. As described in [RR01], in many situations a diffusion limit result can be seen as a rigorous justification for using the ESJD approach.

2.3.2 Scaling analysis of a sequence of Markov processes

The majority of the results presented in this thesis are of the following type. We consider a sequence of Markov chains $x^N = \{x^{k,N}\}_{k\geq 0}$. The Markov chain x^N evolves on a state space $\mathcal{X}_N$ that might depend on the index $N \geq 1$. We are interested in the asymptotic behaviour of these Markov chains as the index N goes to infinity, say. To this end, we introduce a sequence of transformations $\nu_N : \mathcal{X}_N \to \mathcal{X}$ that map the different state spaces $\mathcal{X}_N$ to a fixed state space $\mathcal{X}$. In general, notice that the transformed process $\nu_N(x^N)$ does not enjoy the Markov property. The transformation ν_N is generally chosen so that one can find a diffusive time scale

$\Delta t = \Delta t(N)$ such that the sequence of time rescaled processes

$$z^N(t) := \nu_N(x^{\lfloor t/\Delta t \rfloor, N})$$

converges weakly to a limiting $\mathcal{X}$-valued non trivial process z[1]. If z^N converges weakly to a suitable stationary process then it is natural to deduce that the complexity of the MCMC algorithm based on the Markov chain x^N is inversely proportional to the diffusive time scale $\Delta t(N)$. This weak convergence is denoted as $z^N \Rightarrow z$ in the rest of this thesis. The limiting process z is often described in this thesis as the solution of a stochastic differential equation. Indeed, more general limiting processes are possible. We sometimes choose to index the different state spaces and Markov chain by a parameter $\varepsilon \to 0$ instead of $N \to \infty$.

1. The idea of finding diffusion limits for MCMC methods was pioneered by Roberts and co-workers in [RGG97]; see [RR01] for an overview. In this article, a target distribution π^N on $\mathcal{X}_N := \mathbb{R}^N$ with product form

$$\pi^N(x_1, \ldots, x_N) = \prod_{i=1}^{N} e^{A(x_i)} \tag{2.3.2}$$

is explored through MCMC simulations. If the current position of the Metropolis-Hastings Markov chain is $x^{k,N} = (x_1^{k,N}, \ldots, x_N^{k,N}) \in \mathbb{R}^N$, the Random Walk Metropolis (RWM) algorithm proposes a move distributed as $x^* = x^{k,N} + \frac{\ell}{\sqrt{N}}\,\xi$ where $\xi \overset{\mathcal{D}}{\sim} \mathrm{N}(0, I_N)$ is a standard Gaussian random variable in $\mathbb{R}^N$. The constant $\ell > 0$ is a tuning parameter and the scale $1/\sqrt{N}$ ensures that as the dimension grows to infinity $N \to \infty$ the mean acceptance probability of the algorithm is bounded away from zero and from one. The seminal paper [RGG97] considers the so called first-coordinate transformation $\nu_N : \mathbb{R}^N \to \mathbb{R}$ that maps a vector $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$ to its first coordinate $\nu_N(x_1, \ldots, x_N) := x_1$. It is proved in this article that under mild assumptions the choice of diffusive time scale $\Delta t := 1/N$ leads to a diffusion limit. In other words, the sequence of $\mathbb{R}$-valued continuous time processes z^N defined as

$$z^N(t) = x_1^{\lfloor Nt \rfloor, N}$$

converges in a suitable sense to the solution z of a stochastic differential equation that is ergodic with respect to the probability distribution $e^{A(x)}\,dx$ on the

[1] the notation $\lfloor x \rfloor$ stands for the largest integer less or equal to $x \in \mathbb{R}$, also known as the *floor function*

real line. The stochastic differential equation that described z is of the form

$$dz_t = h(\ell)\,\mu(z_t)\,dt + \sqrt{h(\ell)}\,\sigma(z_t)\,dW_t$$

where $\mu : \mathbb{R} \to \mathbb{R}$ is a drift function and $\sigma : \mathbb{R} \to \mathbb{R}$ is a volatility function that both do not depend on the tuning parameter ℓ. The function μ and σ only depend on the potential A. The speed function $h(\ell)$ is strictly positive and converges to zero as $\ell \to 0$ and $\ell \to \infty$. It has a unique maximum ℓ^*. This reveals that in order to maximise the efficiency of the algorithm as the dimension N grows to infinity, one should choose the tuning parameter ℓ close to the optimal value ℓ^*. Maybe surprisingly, the optimal value ℓ^* is the only value that leads to an asymptotic acceptance probability of 0.234 (to three decimal places); this gives an easy way for practitioners to tune their RWM algorithms: choose the size of the jumps so that the mean acceptance probability is close to 23%. Indeed, this optimal scaling result has only been proven for very restricted and simple class of target probability distributions and is not expected to hold for more complicated distributions. MCMC algorithm targeting probability distributions with multiple modes or exhibiting different scales or intricate local structures are in general very difficult to tune. The seminal result of [RGG97] has initiated a large literature on scaling limits for MCMC algorithms. The articles [RR98; RR01; BDM10; BDM12; NR06] consider more complex proposals, [Béd07; BR07] study the robustness of the 0.234 rule, [BRS09; BRSV08; BRS09; HSVW05; HSV07; MPS11; SVW04] uses scaling arguments in infinite dimensional settings, [CRR05; JLM12] study the initial transient phase, [NR08; NRY12] examine the case of discontinuous target distributions.

2. In chapter 4 and 5 we consider the following setup. An infinite dimensional target distribution π on a separable Hilbert space $\mathcal{H}$ is discretised. The target distribution π on $\mathcal{H}$ is defined through its Radon-Nikodym derivative $\frac{d\pi}{d\pi_0}(x) \propto \exp\big(-\Psi(x)\big)$ with respect to a Gaussian measure π_0. A discretised version π^N of π is introduced in order to approximate π on the finite memory of a computer. The discretised version π^N of π lives on a finite dimensional linear subspace of $\mathcal{H}$. A Metropolis Markov chain x^N that evolves on $\mathcal{X}_N := \mathcal{H}$ using local moves (RWM or MALA – defined in the sequel) is used to explore the target distribution π^N. We prove in this thesis that the identity mapping $\nu_N : \mathcal{H} \to \mathcal{H} =: \mathcal{X}$ and a diffusive time scale of the form $\Delta t := N^{-\gamma}$ leads to

a diffusion limit. More specifically, the sequence of processes z^N defined as

$$z^N(t) = x^{\lfloor N^\gamma t \rfloor, N},$$

where $\gamma > 0$ is an exponent whose value needs to be discussed, converges in a suitable sense to the solution of a $\mathcal{H}$-valued stochastic differential equation that is ergodic with respect to the probability distribution π.

3. In chapter 6 we consider a sequence of target distributions π^ε on $\mathbb{R}^n$ that concentrate, as ε goes to zero, on an neighbourhood of a (fixed) manifold $\mathcal{M}$ of dimension strictly inferior to n (the dimension n is fixed while $\varepsilon \to 0$). In this setting the state space is fixed, $\mathcal{X} = \mathcal{X}^\varepsilon := \mathbb{R}^n$. The distribution π^ε is explored through a Random Walk Metropolis (RWM) Markov chain $x^\varepsilon = \{x^{k,\varepsilon}\}_{k\geq 0}$. For simplicity, we limit our analysis to the case where the manifold $\mathcal{M}$ is flat *i.e* is an affine subspace of $\mathbb{R}^n$. We prove that if π^ε concentrates on a neighbourhood of thickness ε (to be defined rigorously in the sequel) and the standard deviation of the jumps of x^ε is of order ε, the choice of diffusive time scale $\Delta t = \varepsilon^2$ leads to a diffusion limit. In other words, we prove that the sequence of processes z^ε defined as

$$z^\varepsilon(t) = x^{\lfloor t/\varepsilon^2 \rfloor, \varepsilon},$$

converges in a suitable sense to the solution z of a $\mathbb{R}^n$-valued stochastic differential equation. The limiting diffusion evolves on the limiting manifold $\mathcal{M}$. We characterise its invariant distribution.

Chapter 3

Infinite dimensional methods

3.1 Gaussian measures

3.1.1 Gaussian measures on Hilbert spaces

We give in this section a brief introduction to Gaussian measures on Hilbert space. See [DPZ92] for a more developed account of the general theory. Let $\mathcal{H}$ be a separable Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and associated norm $\|x\|^2 = \langle x, x \rangle$. A $\mathcal{H}$-valued random variable X is said to be Gaussian if for any vector $v \in \mathcal{H}$ the scalar random variable $\langle X, v \rangle$ is a real Gaussian random variable. The mean $m \in \mathcal{H}$ is the unique vector satisfying $\mathbb{E}\langle X, v \rangle = \langle m, v \rangle$ for any vector $v \in \mathcal{H}$. The Gaussian random variable X is centred if $\langle X, v \rangle$ is centred for any $v \in \mathcal{H}$. The covariance operator C is the nonnegative symmetric bilinear map $C : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ defined by

$$C(u, v) = \mathrm{Cov}\Big(\langle X, u \rangle, \langle X, v \rangle\Big).$$

This implicitly defines by duality (Riesz representation) a linear mapping[1] $C : \mathcal{H} \to \mathcal{H}$ defined by $C(u, v) = \langle u, Cv \rangle = \langle Cu, v \rangle$. The Gaussian distribution on $\mathcal{H}$ with mean μ and covariance C is denoted by $\mathrm{N}(\mu, C)$. Fernique's theorem [Fer75] states that any Gaussian measure enjoys nice integrability properties; there exists an exponent $\alpha > 0$ such that $\mathbb{E}\big[\exp\big(\alpha \|X\|^2\big)\big] < \infty$. It follows that $\mathbb{E}[\|X\|^2] < \infty$ from which it follows that $C : \mathcal{H} \to \mathcal{H}$ is a trace class operator in the sense that for any orthonormal basis $\{e_j\}_{j \geq 1}$ of the Hilbert space $\mathcal{H}$ we have $\mathrm{Tr}(C) := \sum_j \langle e_j, Ce_j \rangle < \infty$. This is because $\sum_j \langle e_j, Ce_j \rangle = \sum_j \mathbb{E}\langle X, e_j \rangle^2 = \mathbb{E}\|X\|^2$. Since a trace class operator is compact [DS63], the spectral analysis of compact symmetric operators on

[1] by abuse of notation we use the same symbol to denote the bilinear operator $C(u, v) = \mathrm{Cov}\Big(\langle X, u \rangle, \langle X, v \rangle\Big)$ and the associated linear operator $C : \mathcal{H} \to \mathcal{H}$

Hilbert spaces reveals that there exists an orthonormal basis $\{\varphi_j\}_{j\geq 1}$ and eigenvalues[2] $\{\lambda_j^2\}_{j\geq 1}$ such that

$$C\varphi_j = \lambda_j^2\,\varphi_j \qquad \text{and} \qquad \mathrm{Tr}(C) = \sum_{j\geq 1}\lambda_j^2 < \infty.$$

We refer to this orthonormal eigenbasis as the Karhunen-Loève basis. Any vector $x \in \mathcal{H}$ can be decomposed on the Karhunen-Loève basis as

$$x = \sum_{j\geq 1} x_j\,\varphi_j \tag{3.1.1}$$

where $x_j := \langle x, \varphi_j\rangle$. This decomposition shows that the centred Gaussian random variable X with covariance operator C has the same law as the infinite sum

$$X = \sum_{j\geq 1}\langle X, \varphi_j\rangle\,\varphi_j \quad \overset{\mathcal{D}}{\sim} \quad \sum_{j\geq 1}\lambda_j\,\xi_j\,\varphi_j \tag{3.1.2}$$

where $\{\xi_j\}_{j\geq 1}$ is an i.i.d sequence of standard $\mathrm{N}(0,1)$ Gaussian random variables. This expansion of the Gaussian random variable X as an infinite sum is usually called the *Karhunen-Loève* expansion (see [DPZ92], section *White noise expansions*). We now give two examples of particular interest for our purposes.

- **Brownian motion**
  Consider a finite horizon $T < \infty$ and Brownian paths $\{W_t\}_{t\in[0,T]}$ on the interval $[0,T]$. This defines a centred Gaussian measure on $\mathcal{H} = L^2([0,T])$ since for any function $f \in L^2([0,T])$ the random variable $\langle f, W\rangle = \int_0^T f(t)\,W(t)\,dt$ is a centred Gaussian random variable. Since $\mathbb{E}[W(s)W(t)] = \min(s,t)$ it follows that the covariance operator is given by

  $$C(f,g) = \iint_{[0,T]^2} f(s)g(t)\,\min(s,t)\,ds\,dt.$$

  The associated linear operator C is the integral operator that maps a function $f \in L^2([0,T])$ to the function $C(f) \in L^2([0,T])$ given by $C(f)(t) = \int_{s=0}^T f(s)\,\min(s,t)\,ds$. One can then find the eigen-decomposition of this integral operator. The normalised eigenfunctions are $\varphi_k(t) = \sqrt{2/T}\,\sin(t/\lambda_k)$ with eigenvalue $\lambda_k^2 = \left(\frac{T}{(k-\frac{1}{2})\pi}\right)^2$ for $k \geq 1$ (see [DP05] for details). The

[2] We choose the eigenvalues to be $\{\lambda_j^2\}_{j\geq 1}$ and not $\{\lambda_j\}_{j\geq 1}$ in order to simplify the writing of the expansion (3.1.2) and emphasise that the eigenvalues are nonnegative

Karhunen-Loève expansion on $L^2([0,T])$ of a Brownian motion thus reads

$$t \mapsto \frac{\sqrt{2\,T}}{\pi} \sum_{k\geq 1} \frac{\xi_k}{k-\frac{1}{2}} \sin\Big((k-\frac{1}{2})\pi\, t/T\Big)$$

where $\{\xi_k\}_{k\geq 1}$ are i.i.d standard Gaussian random variables. In other words, a Brownian trajectory can be seen as a random superposition of sinusoidal functions with increasing frequencies.

- **Brownian bridge**
  Similarly, one can consider the Karhunen-Loève expansion of a Brownian bridge on $[0,T]$. Indeed, a Brownian bridge $\{B_t\}_{t\in[0,T]}$ on $[0,T]$ defines a Gaussian measure on $L^2([0,T])$ with covariance operator

$$C(f,g) = \iint_{[0,T]^2} f(s)g(t)\Big(\min(s,t) - \frac{st}{T}\Big)\, ds\, dt$$

  and associated linear operator $C(f)(t) = \int_{s=0}^{T} f(s)\big(\min(s,t) - \frac{st}{T}\big)\, ds$. One can diagonalise this operator and satisfy that the normalised eigenfunctions are $\varphi_k(t) = \sqrt{2/T}\,\sin(t/\lambda_k)$ with eigenvalue $\lambda_k^2 = \Big(\frac{T}{k\pi}\Big)^2$ for $k \geq 1$. The Karhunen-Loève expansion on $L^2([0,T])$ of a Brownian bridge thus reads

$$t \mapsto \frac{\sqrt{2\,T}}{\pi} \sum_{k\geq 1} \frac{\xi_k}{k} \sin\Big(k\pi\, t/T\Big)$$

  where $\{\xi_k\}_{k\geq 1}$ are i.i.d standard Gaussian random variables (see [DP05] for details)

3.1.2 Regularity subspaces: the spaces $\mathcal{H}^r$

In the sequel (Chapter 4 and 5), we will be interested in studying a target probability measure π defined through its Radon-Nikodym derivative

$$\frac{d\pi}{d\pi_0}(x) \propto \exp\big\{-\Psi(x)\big\} \tag{3.1.3}$$

with respect to a Gaussian measure π_0 on a Hilbert space $\mathcal{H}$. Nevertheless, in the applications that we have in mind, it does happen more often than not that the function Ψ is not defined on the whole Hilbert space $\mathcal{H}$ but only on a smaller subspace $\mathcal{H}^r \subset \mathcal{H}$ that enjoys better regularity properties. Indeed, the function Ψ only needs to be defined on the support of π_0 for the change of measure that

defines π to make sense. In this section we describe how to properly define a family $\mathcal{H}^r$ of such linear subspaces of $\mathcal{H}$ and give ways to ensure that the support of the Gaussian measure π_0 is a subset $\mathcal{H}^r \subset \mathcal{H}$. For $r > 0$ the space $\mathcal{H}^r$ is a strict linear subspace of $\mathcal{H}$. For $r > 0$, the space $\mathcal{H}^{-r}$ can be interpreted as the dual of $\mathcal{H}^r$.

For every $x \in \mathcal{H}$ we have expansion (3.1.1) of x on the Hilbert basis $\{\varphi_j\}_{j\geq 1}$. Using this expansion, we define Sobolev-like spaces $\mathcal{H}^r, r \in \mathbb{R}$, with the inner-products and norms defined by

$$\langle x, y\rangle_r \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} j^{2r} x_j y_j, \qquad \|x\|_r^2 \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} j^{2r}\, x_j^2. \tag{3.1.4}$$

For $r \geq 0$, the space $\mathcal{H}^r$ is defined as the subset of vectors $x \in \mathcal{H}$ that have finite $\|\cdot\|_r$ norm. Notice that $\mathcal{H}^0 = \mathcal{H}$ and $\mathcal{H}^r \subset \mathcal{H} \subset \mathcal{H}^{-r}$ for any $r > 0$. The Hilbert-Schmidt norm $\|\cdot\|_C$ associated to the covariance operator C with eigen-decomposition $C\varphi_j = \lambda_j^2\, \varphi_j$ is defined as

$$\|x\|_C^2 = \sum_j \lambda_j^{-2} x_j^2.$$

For $x, y \in \mathcal{H}^r$, the outer product operator in $\mathcal{H}^r$ is the operator $x \otimes_{\mathcal{H}^r} y : \mathcal{H}^r \to \mathcal{H}^r$ defined by $(x \otimes_{\mathcal{H}^r} y)z \stackrel{\text{def}}{=} \langle y, z\rangle_r\, x$ for every $z \in \mathcal{H}^r$. For $r \in \mathbb{R}$, let B_r denote the operator which is diagonal in the basis $\{\varphi_j\}_{j\geq 1}$ with diagonal entries j^{2r}. The operator B_r satisfies $B_r\, \varphi_j = j^{2r}\varphi_j$ so that $B_r^{\frac{1}{2}}\, \varphi_j = j^r \varphi_j$. The operator B_r lets us alternate between the Hilbert space $\mathcal{H}$ and the Sobolev spaces $\mathcal{H}^r$ via the identities $\langle x, y\rangle_r = \langle B_r^{\frac{1}{2}} x, B_r^{\frac{1}{2}} y\rangle$. Since $\|B_r^{-1/2}\varphi_k\|_r = \|\varphi_k\| = 1$, we deduce that $\{B_r^{-1/2}\varphi_k\}_{k\geq 0}$ forms an orthonormal basis for $\mathcal{H}^r$.

We now describe a sufficient condition that ensures that a Gaussian measure π_0 with covariance C is supported in $\mathcal{H}^r$. For a positive, self-adjoint operator $D : \mathcal{H} \mapsto \mathcal{H}$, we define its trace in $\mathcal{H}^r$ by

$$\operatorname*{Tr}_{\mathcal{H}^r}(D) \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \langle (B_r^{-\frac{1}{2}}\varphi_j), D(B_r^{-\frac{1}{2}}\varphi_j)\rangle_r. \tag{3.1.5}$$

Since $\operatorname{Tr}_{\mathcal{H}^r}(D)$ does not depend on the choice of the $\mathcal{H}^r$-orthonormal basis [DS63], the operator D is said to be trace class in $\mathcal{H}^r$ if $\operatorname{Tr}_{\mathcal{H}^r}(D) < \infty$ for some, and hence any, orthonormal basis of $\mathcal{H}^r$. Let us define the operator $C_r \stackrel{\text{def}}{=} B_r^{1/2}\, C\, B_r^{1/2}$. Notice that $\operatorname{Tr}_{\mathcal{H}^r}(C_r) = \sum_{j=1}^{\infty} \lambda_j^2\, j^{2r}$. In section 2.1 of [MPS11] it is shown that under the

condition

$$\operatorname*{Tr}_{\mathcal{H}^r}(C_r) < \infty, \tag{3.1.6}$$

the support of $\pi_0 \overset{\mathcal{D}}{\sim} \mathrm{N}(0, C)$ is included in $\mathcal{H}^r$ in the sense that π_0-almost every $x \in \mathcal{H}$ belongs to $\mathcal{H}^r$. Furthermore, the induced distribution of π_0 on $\mathcal{H}^r$ is identical to that of a centered Gaussian measure on $\mathcal{H}^r$ with covariance operator C_r. This means that for $\xi \overset{\mathcal{D}}{\sim} \pi_0$, the following identity $\mathbb{E}\big[\langle \xi, u\rangle_r \langle \xi, v\rangle_r\big] = \langle u, C_r v\rangle_r$ holds for any two functions $u, v \in \mathcal{H}^r$. Thus in what follows, we alternate between the Gaussian measures $\mathrm{N}(0, C)$ on $\mathcal{H}$ and $\mathrm{N}(0, C_r)$ on $\mathcal{H}^r$, for those r for which (3.1.6) holds.

3.1.3 Change of measure

Our goal is to sample from a measure π defined through the change of probability formula (3.1.3). As described in section 3.1.2, the condition $\mathrm{Tr}_{\mathcal{H}^r}(C_r) < \infty$ implies that the measure π_0 has full support on $\mathcal{H}^r$, that is, $\pi_0(\mathcal{H}^r) = 1$. Consequently, if $\mathrm{Tr}_{\mathcal{H}^r}(C_r) < \infty$, the function Ψ needs only to be defined on $\mathcal{H}^r$ in order for the change of probability formula (3.1.3) to be valid. In this section we give assumptions on the decay of the eigenvalues of the covariance operator C of π_0 that ensure the existence of a real number $s > 0$ such that π_0 has full support on $\mathcal{H}^s$. The function Ψ is assumed to be defined on $\mathcal{H}^s$ for some exponent $s > 0$ and we impose regularity assumptions on Ψ that ensure that the probability distribution π is not too different from π_0, when projected into directions associated with φ_j for j large. For each $x \in \mathcal{H}^s$ the derivative $\nabla\Psi(x)$ is an element of the dual $(\mathcal{H}^s)^* \cong \mathcal{H}^{-s}$ of $\mathcal{H}^s$ comprising linear functions on $\mathcal{H}^s$. However, we may identify $(\mathcal{H}^s)^*$ with $\mathcal{H}^{-s}$ and view $\nabla\Psi(x)$ as an element of $\mathcal{H}^{-s}$ for each $x \in \mathcal{H}^s$. With this identification, the following identity holds,

$$\|\nabla\Psi(x)\|_{\mathcal{L}(\mathcal{H}^s,\mathbb{R})} = \|\nabla\Psi(x)\|_{-s}.$$

This is because $\|\nabla\Psi(x)\|_{\mathcal{L}(\mathcal{H}^s,\mathbb{R})} = \sup_{\|y\|_s \le 1} \sum_j \lambda_j^{2s} \langle \nabla\Psi(x), \varphi_j\rangle y_j$ and the last expression can be re-arranged as $\sum_j \lambda_j^{2s} \langle \nabla\Psi(x), \varphi_j\rangle y_j = \sum_j \lambda_j^{-2s} \langle \nabla\Psi(x), \varphi_j\rangle (\lambda_j^{4s} y_j)$. Similarly, the second derivative $\partial^2\Psi(x)$ can be identified as an element of $\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})$. To avoid technicalities we assume that Ψ is quadratically bounded, with first derivative linearly bounded and second derivative globally bounded.

Assumptions 3.1.1. *The covariance operator C and function Ψ satisfy the following:*

1. ***Decay of Eigenvalues*** λ_j^2 ***of*** C***:*** *there is an exponent* $\kappa > \frac{1}{2}$ *such that*

$$\lambda_j \asymp j^{-\kappa}. \tag{3.1.7}$$

2. ***Assumptions on*** Ψ*: the function* Ψ *is defined on* $\mathcal{H}^s$ *for some exponent* $s \in [0, \kappa - 1/2)$*. There exist constants* $M_i \in \mathbb{R}, i \leq 4$ *such that for all* $x \in \mathcal{H}^s$ *we have*

$$M_1 \leq \Psi(x) \leq M_2 \left(1 + \|x\|_s^2\right) \tag{3.1.8}$$

$$\|\nabla\Psi(x)\|_{-s} \leq M_3 \left(1 + \|x\|_s\right) \tag{3.1.9}$$

$$\|\partial^2\Psi(x)\|_{\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})} \leq M_4. \tag{3.1.10}$$

Remark 3.1.2. *The condition* $\kappa > \frac{1}{2}$ *ensures that the covariance operator* C *is trace class in* $\mathcal{H}$ *since in this case* $\mathrm{Tr}(C) \lesssim \sum_j j^{-2\kappa} < \infty$*. The same reasoning gives that the operator* C_r *is trace-class in* $\mathcal{H}^r$ *for any* $r < \kappa - \frac{1}{2}$*. It follows that* π_0 *has full measure in* $\mathcal{H}^r$ *for any* $r \in [0, \kappa - 1/2)$*. In particular* π_0 *has full support on* $\mathcal{H}^s$*.*

Remark 3.1.3. *The function* $\Psi(x) = \frac{1}{2}\|x\|_s^2$ *satisfies assumptions 3.1.1. Indeed, it is defined on* $\mathcal{H}^s$ *and its derivative at* $x \in \mathcal{H}^s$ *is given by* $\nabla\Psi(x) = \sum_{j\geq 0} j^{2s} x_j \varphi_j \in \mathcal{H}^{-s}$ *with* $\|\nabla\Psi(x)\|_{-s} = \|x\|_s$*. The second derivative* $\partial^2\Psi(x) \in \mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})$ *is the linear operator that maps* $u \in \mathcal{H}^s$ *to* $\sum_{j\geq 0} j^{2s}\langle u, \varphi_j\rangle\varphi_j \in \mathcal{H}^s$ *and its norm satisfies* $\|\partial^2\Psi(x)\|_{\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})} = 1$ *for any* $x \in \mathcal{H}^s$*.*

Since the eigenvalues $\{\lambda_j^2\}_{j\geq 1}$ of C decrease as $\lambda_j \asymp j^{-\kappa}$, the operator C has a smoothing; $C^\alpha h$ gains $2\alpha\kappa$ orders of regularity in the sense that the $\mathcal{H}^\beta$-norm of $C^\alpha h$ is controlled by the $\mathcal{H}^{\beta-2\alpha\kappa}$-norm of $h \in \mathcal{H}$. Indeed, under Assumption 3.1.1, the following estimates holds

$$\|h\|_C \asymp \|h\|_\kappa \qquad \text{and} \qquad \|C^\alpha h\|_\beta \asymp \|h\|_{\beta-2\alpha\kappa}. \tag{3.1.11}$$

The proof follows the methodology used to prove Lemma 3.3 of [MPS11]. The reader is referred to this text for more details. This estimate is used at several places in the sequel. In chapters 4 and 5 we will consider stochastic differential equations evolving in $\mathcal{H}^s$ with a drift of the form $d(x) \stackrel{\text{def}}{=} -\left(x + C\nabla\Psi(x)\right)$. The methods of proof that we will be using exploit the fact that the drift function $d : \mathcal{H}^s \to \mathcal{H}^s$ is a Lipschitz function under mild assumptions on the function Ψ. The next lemma gives such sufficient conditions.

Lemma 3.1.4. *Let assumptions 3.1.1 hold.*

1. *The function* $d(x) \stackrel{def}{=} -\Big(x + C\nabla\Psi(x)\Big)$ *is well defined and globally Lipschitz on* $\mathcal{H}^s$,

$$\|d(x) - d(y)\|_s \lesssim \|x - y\|_s \qquad \forall x, y \in \mathcal{H}^s. \tag{3.1.12}$$

2. *The second order remainder term in the Taylor expansion of* Ψ *satisfies*

$$\big|\Psi(y) - \Psi(x) - \langle \nabla\Psi(x), y - x\rangle\big| \lesssim \|y - x\|_s^2 \qquad \forall x, y \in \mathcal{H}^s \tag{3.1.13}$$

Proof. See Equation 3.5 of [MPS11]. □

3.2 Stochastic differential equations in Hilbert spaces

Consider a Gaussian measure on the separable Hilbert space $\mathcal{H}$ with covariance operator C and Karhunen-Loève eigen-basis $C\varphi_j = \lambda_j^2 \varphi_j$. The Brownian motion on $\mathcal{H}$ with covariance C is the continuous time stochastic process $W : [0; +\infty) \to \mathcal{H}$ defined by

$$W_t = \sum_{j \geq 1} \lambda_j \beta_j(t)\, \varphi_j$$

where $\{\beta_j\}_{j\geq 1}$ is a family of independent real standard Brownian motions. In other words, each coordinate in the Karhunen-Loève eigen-basis evolves as an independent brownian motion. The Brownian motion W is a $\mathcal{H}$-valued centred Gaussian process with almost-sure continuous paths. It is characterised by its autocorrelation structure; one can verify that for any two vectors $u, v \in \mathcal{H}$ the following formula holds,

$$\mathrm{Cov}\big(\langle W_s, u\rangle, \langle W_t, v\rangle\big) = \min(s, t)\, \langle u, Cv\rangle.$$

This directly follows from the usual Brownian autocorrelation structure $\mathrm{Cov}(\beta_s, \beta_t) = \min(s, t)$. At time $t > 0$ the Brownian motion W_t has a Gaussian distribution with covariance operator tC. Indeed, if the Gaussian measure $\mathrm{N}(0, C)$ has full measure in $\mathcal{H}^s$, the Brownian motion with covariance C can be seen as a Brownian motion in $\mathcal{H}^s$. The solution of the $\mathcal{H}$-valued stochastic differential equation

$$dX_t = \mu(X_t)\, dt + \sigma\, dW_t \tag{3.2.1}$$

where $\mu : \mathcal{H} \to \mathcal{H}$ is a drift function, $\sigma \in \mathbb{R}$ is a fixed constant and W is a brownian motion with covariance C is defined as the solution of the integral equation

$$X_t = X_0 + \int_0^t \mu(X_s)\, ds + \sigma\, W_t \qquad \forall t > 0. \tag{3.2.2}$$

The advantage of dealing with a volatility function that is a fixed constant $\sigma \in \mathbb{R}$ is that in order to define the solution to the stochastic differential equation (3.2.1), one does not need to use the theory of stochastic calculus with respect to Hilbert space valued continuous martingales; a simple integral equation of the form (3.2.2) suffices. Since the noise enters (3.2.1) additively, the induced Itô map Θ which takes Brownian trajectories and initial conditions into solutions is continuous in the supremum-in-time topology (Lemma 3.2.1). This fact, which would not be true if (3.2.1) had multiplicative noise, allows us to employ an argument simpler than the more general techniques often used; for a Lipschitz drift function μ, the usual Picard iteration approach gives existence and uniqueness of solutions to the stochastic differential equation (3.2.1). For a fixed time horizon $T > 0$, the Itô map $\Theta : \mathcal{H} \times C([0,T],\mathcal{H}) \to C([0,T],\mathcal{H})$, where $C([0,T],\mathcal{H})$ denotes the linear space of continuous function from the interval $[0,T]$ to $\mathcal{H}$, is the function that maps the pair $(x_0, w) \in \mathcal{H} \times C([0,T],\mathcal{H})$ to the solution $x = \Theta(x_0, w)$ of the integral equation $x_t = x_0 + \int_0^t \mu(x_s)\, ds + \sigma\, w_t$ for all $t \in [0,T]$. We now prove that the Itô map Θ is continuous if $C([0,T],\mathcal{H})$ is endowed with the supremum norm $\|x\|_\infty = \max_{t\in[0,T]} \|x_t\|$.

Lemma 3.2.1. (Continuity of the Itô map)
Let $\mathcal{H}$ be a Hilbert space and suppose that the drift function $\mu : \mathcal{H} \to \mathcal{H}$ is Lipschitz. The Itô map $\Theta : \mathcal{H} \times C([0,T],\mathcal{H}) \to C([0,T],\mathcal{H})$ associated to the integral equation (3.2.2) is continuous if $C([0,T],\mathcal{H})$ is endowed with the supremum topology.

Proof. The proof follows the Picard iteration approach for proving the Cauchy-Lipschitz existence and uniqueness theorem of ODE theory. See the proof of Lemma 3.7 of [MPS11] for details. □

Lemma 3.2.1 shows that the solutions of the stochastic differential equation (3.2.1) can be constructed as image under the Itô map Θ of a Brownian motion in $\mathcal{H}$ with covariance C. This explicit construction is at the basis of several weak convergence results described in chapters 4 and 5.

3.3 Diffusion-approximation

The paper [MPS11] developed an approach for deriving diffusion limits for MCMC methods, using ideas from numerical analysis. In this section we build upon these techniques to derive a general framework for proving diffusion limits in very general settings. We prove in particular a general diffusion-approximation result that will be used at several places in the sequel. We consider a sequence of $\mathcal{H}$-valued Markov chains $x^N = \{x^{k,N}\}_{k\geq 0}$ and a sequence of time steps $\Delta t = \Delta t(N)$ that converges to zero. For time $t \geq 0$ satisfying $k\Delta t \leq t < (k+1)\Delta t$, we define the accelerated version $\bar{z}^N$ of x^N and its continuous interpolant z^N by

$$\begin{cases} \bar{z}^N(t) &= x^{k,N} \\ z^N(t) &= \frac{(k+1)\Delta t - t}{\Delta t} x^{k,N} + \frac{t - k\Delta t}{\Delta t} x^{k+1,N}. \end{cases} \tag{3.3.1}$$

Notice that the process z^N has continuous sample paths and $z^N(k\Delta t) = \bar{z}^N(k\Delta t)$ for any indices $k, N \geq 0$. In words, the process $\bar{z}^N$ is a continuous time and piecewise constant accelerated version (by a factor $1/\Delta t$) of the process x^N. The process z^N is the continuous (piecewise affine) version of the process $\bar{z}^N$. We introduce the following martingale-drift decomposition of the Markov chain x^N,

$$x^{k+1,N} - x^{k,N} = d^N(x^{k,N})\,\Delta t + \sqrt{\Delta t}\,\Gamma^{k,N} \tag{3.3.2}$$

where $d^N : \mathcal{H} \to \mathcal{H}$ is a deterministic function and $\Gamma^N = \{\Gamma^{k,N}\}_{k\geq 0}$ is a $\mathcal{H}$-valued martingale difference (i.e. $M_n := \sum_{k\leq n} \Gamma^{k,N}$ is a martingale). Equation (3.3.2) is another way of writting the identity $x^{k+1,N} - x^{k,N} = \mathbb{E}[x^{k+1,N} - x^{k,N}\,|x^{k,N}] + \big(x^{k+1,N} - x^{k,N} - \mathbb{E}[x^{k+1,N} - x^{k,N}\,|x^{k,N}]\big)$ with $d^N(x^{k,N}) = \mathbb{E}[x^{k+1,N} - x^{k,N}\,|x^{k,N}]/\Delta t$ and $\Gamma^{k,N} = \big(x^{k+1,N} - x^{k,N} - \mathbb{E}[x^{k+1,N} - x^{k,N}\,|x^{k,N}]\big)/\sqrt{\Delta t}$. In applications, the scaling factor Δt is chosen such that the drift function d^N and the martingale difference term $\Gamma^{k,N}$ behave well as $N \to \infty$. The rescaled martingale W^N is defined as

$$W^N(t) = \sqrt{\Delta t} \sum_{j=0}^{k} \Gamma^{j,N} \;+\; \frac{t - k\Delta t}{\sqrt{\Delta t}}\, \Gamma^{k+1,N} \tag{3.3.3}$$

for $k\Delta t \leq t < (k+1)\Delta t$. Notice that the process W^N has continuous (piecewise affine) sample paths. The next proposition is the main result of this section states that if the sequence W^N converges to a Brownian motion and the sequence of deterministic functions d^N converges to a limiting Lipschitz function $\mu : \mathcal{H} \to \mathcal{H}$ then the accelerated process z^N converges to the solution of a $\mathcal{H}$-valued stochastic differential equation. The proof is inspired by the machinery developed to prove the

main theorem of [MPS11].

Proposition 3.3.1. (General diffusion approximation for Markov chains) *Consider a separable Hilbert space $\big(\mathcal{H}, \langle\cdot,\cdot\rangle\big)$, a finite time horizon $T > 0$ and a sequence of $\mathcal{H}$-valued Markov chains $x^N = \{x^{k,N}\}_{k\geq 0}$. Suppose that the drift-martingale decomposition (3.3.2) satisfies the following conditions.*

1. **Convergence of initial conditions:** *the sequence of initial distributions converges in distribution to a probability measure π that has a finite first moment, $\mathbb{E}_\pi \|X\| < \infty$.*

2. **Invariance principle:** *the sequence $(x^{0,N}, W^N)$ defined by equation (3.3.3) converges weakly in $\mathcal{H}\times C([0,T]$ to (z^0, W) where $z^0 \overset{\mathcal{D}}{\sim} \pi$ and W is a Brownian motion in $\mathcal{H}$, independent from z^0, with covariance operator C.*

3. **Convergence of the drift:** *there exists a globally Lipschitz function $\mu : \mathcal{H} \to \mathcal{H}$ such that the following limit holds in probability,*

$$\lim_{N\to\infty} \int_0^T \Big\| d^N(\bar{z}^N(u)) - \mu(z^N(u)) \Big\| \, du = 0,$$

with processes $\bar{z}^N$ and z^N defined by Equation (3.3.1).

Under these three conditions the sequence of rescaled interpolants $z^N \in C([0,T],\mathcal{H})$ defined by equation (3.3.1) converges weakly in $C([0,T],\mathcal{H})$ to the solution of the $\mathcal{H}$-valued stochastic differential equation

$$dz = \mu(z)\, dt + dW$$

with initial condition $z(0) \overset{\mathcal{D}}{\sim} \pi$. Here W is a Brownian motion in $\mathcal{H}$ with covariance C.

Remark 3.3.2. *Indeed, the conclusion remains valid if the martingale-drift decomposition reads $x^{k+1,N} - x^{k,N} = C_1\, d^N(x^{k,N})\, \Delta t + C_2\sqrt{\Delta t}\, \Gamma^{k,N}$ for two constants $C_1, C_2 \in \mathbb{R}$. In this case and under the same assumptions the sequence z^N of rescaled Markov chains converges weakly to the solution of the stochastic differential equation $dz = C_1\, \mu\big(z(t)\big)\, dt + C_2\, dW$.*

Remark 3.3.3. *There are many scaling limit results for MCMC algorithm available in the literature. Except notable exceptions [CRR05; JLM12], virtually all these results [RGG97; RR98; RR01; Béd07; NR11; NRY12; BDM12; BDM10; BRS09; BPR+13; BPS04] assume that the algorithm is started at stationarity. Proposition*

3.3.1 does not rely on such an assumption. We prove a scaling limit result without stationarity assumptions in chapter 5.

Proof. The process $\bar{z}^N(t)$ verifies

$$\begin{aligned} z^N(t) &= x^{0,N} + \int_0^t d^N(\bar{z}^N(u))\,du + W^N(t) \\ &= z^{0,N} + \int_0^t \mu(z^N(u))\,du + \widehat{W}^N(t) \end{aligned} \tag{3.3.4}$$

where the process $W^N \in C([0,T],\mathcal{H})$ is defined by equation (3.3.3) and

$$\widehat{W}^N(t) = W^N(t) + \int_0^t \Big(d^N(\bar{z}^N(u)) - \mu(z^N(u))\Big)\,du.$$

Define the Itô map $\Theta\colon \mathcal{H} \times C([0,T];\mathcal{H}) \to C([0,T];\mathcal{H})$ that maps (z_0, W) to the unique solution $z \in C([0,T],\mathcal{H})$ of the integral equation

$$z(t) = z_0 + \int_0^t \mu(z(u))\,du + W(t), \qquad \forall t \in [0,T].$$

Equation (3.3.4) thus also reads $z^N = \Theta(x^{0,N}, \widehat{W}^N)$. The proof of the diffusion approximation is accomplished through the following steps.

- **The Itô map $\Theta : \mathcal{H} \times C([0,T],\mathcal{H}) \to C([0,T],\mathcal{H})$ is continuous.**
  Since $\mu : \mathcal{H} \to \mathcal{H}$ is globally Lipschitz, Lemma 3.2.1 applies.

- **The pair $(x^{0,N}, \widehat{W}^N)$ converges weakly to (z^0, W).**
  In a Hilbert space, Slutsky's theorem [GS01] states that if the sequence of random variables $\{A_n\}_{n\in\mathbb{N}}$ converges weakly to the random variable A and the sequence $\{B_n\}_{n\in\mathbb{N}}$ converges in probability to zero then the sequence $\{A_n + B_n\}_{n\in\mathbb{N}}$ converges weakly to A. It is assumed that $(x^{0,N}, W^N)$ converges weakly to (z^0, W) in $\mathcal{H} \times C([0,T],\mathcal{H})$. Since the quantity $\int_0^T \|d^N(\bar{z}^N(u)) - \mu(z^N(u))\|_s\,du$ is assumed to converge in probability to 0 as $N \to \infty$ and $\widehat{W}^N(t) = W^N(t) + \int_0^t \Big(d^N(\bar{z}^N(u)) - \mu(z^N(u))\Big)\,du$, it thus follows that the sequence $(x^{0,N}, \widehat{W}^N)$ converges weakly to (z^0, W) in $\mathcal{H} \times C([0,T],\mathcal{H})$ as $N \to \infty$.

- **Continuous mapping argument.**
  The sequence $(x^{0,N}, \widehat{W}^N)$ converges weakly in $\mathcal{H} \times C([0,T],\mathcal{H})$ to (z^0, W) and the Itô map $\Theta\colon \mathcal{H} \times C([0,T],\mathcal{H}) \to C([0,T],\mathcal{H})$ is a continuous function. The

continuous mapping theorem thus shows that $z^N = \Theta(x^{0,N}, \widehat{W}^N)$ converges weakly to $z = \Theta(z^0, W)$, finishing the proof of Proposition 3.3.1.

□

3.4 MCMC on Hilbert spaces

The Bayesian approach to inverse problems is a natural framework for analysing frequently occurring situations [Fit91; BS09; HSV10; Stu10; CRSW12]. When the object of interest is a function, the posterior distribution is a measure on a space of functions. In the examples that we have in mind, the function space of interest can be endowed with the structure of a Hilbert space. In this section we give a brief description of several applied problems where this viewpoint is natural. In these examples the posterior measure π has a density with respect to a Gaussian reference measure π_0 on a Hilbert space $\mathcal{H}$. In other words, the posterior distribution π can be described as a change of measure of the form (3.1.3). For the change of probability $\frac{d\pi}{d\pi_0}(x) \propto \exp\big\{-\Psi(x)\big\}$ to make sense we require that the potential $\Psi : \mathcal{H} \to \mathbb{R}$ is defined π_0-almost surely. The covariance operator of the Gaussian measure π_0 is the linear operator $C : \mathcal{H} \to \mathcal{H}$. The mean of π_0 is denoted by $m \in \mathcal{H}$. The success of using Gaussian priors to model an unknown function stems largely from the model flexibility they afford. With analogy to the finite dimensional setting, it is instructive (though not formally correct) to write the prior Gaussian density as $\pi_0(x) \propto \exp\{-\frac{1}{2}\langle x-m, C^{-1}x-m\rangle\}$, which can also be written $\pi_0(x) \propto \exp\{\frac{1}{2}\langle x-m, \mathcal{L}x-m\rangle\}$ where the inverse $\mathcal{L}$ of $-C$ is known as the precision operator. Using this notation, the informal expression for the density of posterior distribution π is

$$\pi(x) \propto \exp\{-\Psi(x) + \frac{1}{2}\langle x-m, \mathcal{L}x-m\rangle\}.$$

In many of our applications $\mathcal{L}$ will be a differential operator. We now give several examples leading to posterior distributions that can be seen as a change of measure with respect to Gaussian measure living on an infinite dimensional Hilbert space. More details can be found in [BS09; CRSW12].

- **Bayesian Inverse problems**
  Suppose that one tries to reconstruct an unknown function $x \in \mathcal{H}$ from observed data y. We assume that the data $y \in \mathbb{R}^d$ is obtained by applying an (possibly non-linear) operator $\mathcal{G} : \mathcal{H} \to \mathbb{R}^d$ to the function x and adding the

realisation of mean zero Gaussian random variable with covariance Σ,

$$y = \mathcal{G}(x) + \xi, \qquad \xi \overset{\mathcal{D}}{\sim} \mathrm{N}(0, \Sigma).$$

The operator $\mathcal{G}$ is sometimes called the observation operator in the applied literature. Adopting a Gaussian random field priors $\mathrm{N}(0, C)$ on the unknown function $x \in \mathcal{H}$, Bayes' theorem shows that the posterior distribution is a Gaussian change of measure of the form $\frac{d\pi}{d\pi_0}(x) \propto \exp\{-\Psi(x)\}$ with

$$\Psi(x) = \frac{1}{2}\big\|\mathcal{G}(x) - y\big\|_{\Sigma}^2.$$

We have used the standard notation $\|z\|_{\Sigma}^2 := \langle z, \Sigma^{-1} z\rangle^2$. The article [BS09] describes examples including Lagrangian data assimilation and geophysical modelling where the operator $\mathcal{G}$ involves solving a partial differential equation. In practice, the computation of the quantity $\mathcal{G}(x)$ might be very expensive (e.g. involves solving a PDE) and it is important to design efficient MCMC algorithms that enjoy high mean acceptance probability. Indeed, it is computationally very inefficient to consider a proposal $x \mapsto x'$, compute the quantity $\mathcal{G}(x')$ and then reject the proposal x'.

- **Molecular dynamics**
  A common approach for describing the movement of a molecule is that of a Brownian dynamics. The atomic position x of the molecule is a vector in $\mathbb{R}^{Nd}$ where N is the number of atoms in the molecule and d the spatial dimension. It is modelled by the Langevin diffusion

$$dx_t = -\nabla U(x_t)\, dt + \sqrt{2\tau}\, dW \tag{3.4.1}$$

  where $U : \mathbb{R}^{Nd} \to \mathbb{R}$ is a potential describing the physical system. The process W is a standard Brownian motion in $\mathbb{R}^{Nd}$ and $\tau > 0$ the temperature. The invariant distribution of this dynamics has a density with respect to the Lebesgue measure proportional to $\exp\{-U(x)/\tau\}$. For small temperature $\tau \ll 1$, the solution of the Langevin diffusion spends most of its time near the minima of the potential U and transitions between these minima are rare events. The time between two transitions is exponentially long in the inverse temperature $1/\tau$ [FW12] so that it is computationally infeasible to simply solve the SDE forward and hope to observe a transition. Instead we may condition on this rare event occurring. To this end, let T be a finite time horizon and $x_{\pm}$ de-

note two minima of the potential U. We consider the Langevin dynamics (3.4.1) conditioned on the event $x(0) = x_-$ and $x(T) = x_+$. The probability measure π governing the conditioned Langevin diffusion (3.4.1) has density in $\mathcal{H} = L^2([0;T], \mathbb{R}^{Nd})$ with respect to the Brownian bridge measure π_0 arising in the case of vanishing potential $U = 0$. Girsanov's theorem gives that the measure π can be described by a Gaussian change of probability [HSV07] of the form $\frac{d\pi}{d\pi_0}(x) \propto \exp\{-\Psi_\tau(x)\}$ with potential

$$\Psi_\tau(x) := \frac{1}{2\tau} \int_0^T \left(\frac{1}{2} \left| \nabla U(x_t) \right|^2 - \tau \Delta U(x_t) \right) dt$$

The Brownian bridge measure π_0 is the law of a Brownian bridge with volatility $\sqrt{2\tau}$ starting at x_- at time $t = 0$ and ending at x_+ at time T.

- **Signal processing**
  It is often of interest to identify a hidden signal $\{x_t\}_{t\in[0,T]}$ given some observation y. In applications of interest, the hidden process x can be modelled by a Markov process. In the continuous time and continuous state space setting where the hidden process x evolves in $\mathbb{R}^n$, it is convenient to describe its dynamics by a diffusion of the form

$$dx_t = f(x_t)\, dt + dW_t$$

  and initial condition $x_0 \overset{\mathcal{D}}{\sim} \zeta$. The smoothing problem consists in finding the distribution of x_t given all the observations y available on $[0, T]$. In the case where y consists in discrete observations, this gives rise to conditioned diffusions very similar to the previous example. Another frequently occurring setting consists in modelling the observation process y as a solution of a stochastic differential equation of the form

$$dy_t = g(x_t)\, dt + dB_t.$$

  where B is a Brownian motion possibly correlated to W. The smoothing problem can be formulated as determining a probability measure π on $\mathcal{H} = L^2([0, T], \mathbb{R}^m)$ describing the conditional distribution of the hidden process $\{x_t\}_{t\in[0,T]}$ conditionally upon the observation process $\{y_t\}_{t\in[0,T]}$. Girsanov's formula shows that under mild assumption and if the initial distribution ζ is Gaussian, the distribution π can be described as a Gaussian change of measure of the form $\frac{d\pi}{d\pi_0}(x) \propto \exp\{-\Psi(x)\}$. Here π_0 is a Gaussian measure derived from

the original problem in the case where the functions f and g are set to zero. Details can be in [HSVW05; HSV07].

A key idea [SVW04; HSVW05; HSV07; BRSV08; BRS09; CRSW12] for constructing MCMC algorithms targeting infinite dimensional distribution that are Gaussian change of measures of the form $\frac{d\pi}{d\pi_0}(x) \propto \exp\{-\Psi(x)\}$ is to design proposals based on discretizations of $\mathcal{H}$-valued stochastic differential equations which are reversible with respect to either the reference measure π_0 or to the full target measure π. The simplest $\mathcal{H}$-valued diffusion that is reversible with respect to the Gaussian measure $\pi_0 = \mathrm{N}(0, C)$ might be the Ornstein-Uhlenbeck diffusion

$$dX_t = -X_t\,dt + \sqrt{2}\,dW_t$$

where W is a Brownian motion in $\mathcal{H}$ with covariance C. One of the advantages of designing proposals based on an Ornstein-Uhlenbeck diffusion is that exact discretizations are available. In other words, there is no need to resort to Euler-Maryama approximations (or higher order schemes). Consequently, one can design proposals that are exactly reversible with respect to π_0. An accept-reject mechanism is then necessary to transform these proposals into an algorithm that is reversible with respect to π. This line of work is explored in chapter 5. Nonetheless, one of the drawbacks of proposals based on the Ornstein-Uhlenbeck diffusion is that no information contained in the potential Ψ is taken into account. Instead, one can design proposals that are based on discretisations of the Langevin diffusion

$$dX_t = -\big(X_t + C\nabla\Psi(X_t)\big)\,dt + \sqrt{2}\,dW_t.$$

As proved in [DPZ92; HSVW05; HSV07], this Langevin diffusion is reversible with respect to the distribution π. Contrary to proposals based on an Ornstein-Uhlenbeck, the Langevin proposals take into account information contained in the potential Ψ. If one could construct exact discretisations of the Langevin diffusion, one could in theory simulate a Markov chain that is exactly reversible with respect to π. Nevertheless, it is in general not possible to construct exact discretizations of a Langevin diffusion and one thus have to resort to approximations. This leads to algorithms which do not scale well with the dimensionality of the problem. Questions related to this phenomenon are investigated in chapter 4.

Chapter 4

Scaling Analysis of Metropolis Adjusted Langevin Algorithm

This chapter is joint work with Andrew Stuart and Natesh Pillai and is based on the paper [PST12].

4.1 Introduction

Sampling probability distributions π^N in $\mathbb{R}^N$ for N large is of interest in numerous applications arising in applied probability and statistics. The Markov Chain Monte Carlo (MCMC) methodology [RC04] provides a framework for many algorithms which affect this sampling. It is hence of interest to quantify the computational cost of MCMC methods as a function of dimension N. The simplest class of target measures for which analysis can be carried out are perhaps product-form target distributions π^N with density of the type

$$\frac{d\pi^N}{d\lambda^N}(x) = \Pi_{i=1}^N f(x_i). \tag{4.1.1}$$

Here $\lambda^N(dx)$ is the N-dimensional Lebesgue measure and $f(x)$ is a one-dimensional probability density function. Thus π^N has the form of an i.i.d. product. The scaling analysis of local move MCMC algorithms evolving on product form densities (4.1.1) is described in the seminal papers [RGG97; RR98]. Two widely used proposals are the random walk proposal (obtained from the discrete approximation of Brownian motion)

$$y = x + \sqrt{2\delta}Z^N, \quad Z^N \sim \mathrm{N}(0, \mathrm{I}_N), \tag{4.1.2}$$

and the Langevin proposal (obtained from the time discretization of the Langevin diffusion)

$$y = x + \delta \nabla \log \pi^N(x) + \sqrt{2\delta}\, Z^N, \qquad Z^N \sim \mathrm{N}(0, \mathrm{I}_N)\ . \tag{4.1.3}$$

Here 2δ is the proposal variance, a parameter quantifying the size of the discrete time increment; we will consider "local proposals" for which δ is small. The Markov chain corresponding to proposal (4.1.2) is the Random Walk Metropolis (RWM) algorithm [MRTT53], and the Markov transition rule constructed from the proposal (4.1.3) is known as the Metropolis Adjusted Langevin Algorithm (MALA) [RC04]. This chapter is aimed at analyzing the computational complexity of the MALA algorithm in high dimensions.

A fruitful way to quantify the computational cost of these Markov chains which proceed via local proposals is to determine the "optimal" size of increment δ as a function of dimension N (the precise notion of optimality is discussed below). The optimal scale for the proposal variance strikes a balance between making large moves and still having a reasonable acceptance probability. In order to quantify this idea, we will carry out a scaling analysis of the MALA algorithm in high dimensions. The reader is referred to section 2.3.2 for more details on this method. We define a continuous interpolant z^N of the Markov chain x^N by

$$z^N(t) = \left(\frac{t}{\Delta t} - k\right) x^{k+1,N} + \left(k + 1 - \frac{t}{\Delta t}\right) x^{k,N} \tag{4.1.4}$$

for $k\Delta t \le t < (k+1)\Delta t$. Notice that z^N is an accelerated version of x^N. In order to prove a diffusion limit, we choose the proposal variance to satisfy $\delta = \ell \Delta t$, with $\Delta t = N^{-\gamma}$ setting the diffusive scale in terms of dimension and the parameter ℓ a "tuning" parameter which is independent of the dimension N. Key questions, then, concern the choice of γ and ℓ. If z^N converges weakly to a suitable stationary diffusion process then it is natural to deduce that the number of Markov chain steps required in stationarity is inversely proportional to the proposal variance, and hence grows like N^γ. The parametric dependence of the limiting diffusion process then provides a selection mechanism for ℓ. A research program along these lines was initiated by Roberts and co-workers in the pair of papers [RGG97; RR98]. These papers concerned the RWM and MALA algorithms respectively when applied to the target (4.1.1). In both cases it was shown that the projection of z^N into any single fixed coordinate direction x_i converges weakly in $C([0,T];\mathbb{R})$ to z, the scalar

diffusion process

$$dz_t = h(\ell)[\log f(z_t)]' \, dt + \sqrt{2h(\ell)} \, dW_t \tag{4.1.5}$$

for $h(\ell) > 0$ a constant determined by the parameter ℓ from the proposal variance. For RWM the scaling of the proposal variance to achieve this limit is determined by the choice $\gamma = 1$ ([RGG97]) whilst for MALA $\gamma = \frac{1}{3}$ ([RR98]). The analysis shows that the number of steps required to sample the target measure grows as $\mathcal{O}(N)$ for RWM, but only as $\mathcal{O}(N^{\frac{1}{3}})$ for MALA. This quantifies the efficiency gained by use of MALA over RWM, and in particular from employing local moves informed by the gradient of the logarithm of the target density. A second important feature of the analysis is that it suggests that the optimal choice of ℓ is that which maximizes $h(\ell)$. This value of ℓ leads in both cases to a universal (independent of $f(\cdot)$) optimal average acceptance probability (to three significant figures) of 0.234 for RWM and 0.574 for MALA.

These theoretical analysis have had a huge practical impact as the optimal acceptance probabilities send a concrete message to practitioners: one should "tune" the proposal variance of the RWM and MALA algorithms so as to have acceptance probabilities of 0.234 and 0.574 respectively. However, practitioners use these tuning criteria far outside the class of target distributions given by (4.1.1). It is natural to ask whether they are wise to do so. Extensive simulations (see [RR01; SFR10]) show that these optimality results also hold for more complex target distributions. Furthermore, a range of subsequent theoretical analyses confirmed that the optimal scaling ideas do indeed extend beyond (4.1.1); these papers studied slightly more complicated models such as products of one dimensional distributions with different variances and elliptically symmetric distributions ([Béd07; Béd09; BPS04; CRR05]). However the diffusion limits obtained remain essentially one-dimensional in all of these extensions.[1] In this section we study considerably more complex target distributions which are not of the product form and the limiting diffusion takes values in an infinite dimensional space.

Our perspective on these problems is motivated by applications such as Bayesian nonparametric statistics, for example in application to inverse problems [Stu10], and the theory of conditioned diffusions [HSV10]. In both these areas the target measure of interest, π, is on an infinite dimensional real separable Hilbert space $\mathcal{H}$ and, for Gaussian priors (inverse problems) or additive noise (diffusions) is absolutely continuous with respect to a Gaussian measure π_0 on $\mathcal{H}$ with mean

[1] The paper [BR00] contains an infinite dimensional diffusion limit, but we have been unable to employ the techniques of that paper.

zero and covariance operator C. This framework for the analysis of MCMC in high dimensions was first studied in the papers [BRSV08; BRS09; BS09] and is described in more depth in section 3.1. The Radon-Nikodym derivative defining the target measure is assumed to have the form

$$\frac{d\pi}{d\pi_0}(x) = M_\Psi \exp(-\Psi(x)) \tag{4.1.6}$$

for a real-valued function $\Psi : \mathcal{H}^s \mapsto \mathbb{R}$ defined on a subspace $\mathcal{H}^s \subset \mathcal{H}$ that contains the support of the reference measure π_0; here M_Ψ is a normalizing constant. We are interested in studying MCMC methods applied to finite dimensional approximations of this measure found by projecting onto the first N eigenfunctions of the covariance operator C of the Gaussian reference measure π_0.

It is proved in [DPZ92; HSVW05; HSV07] that the measure π is invariant for $\mathcal{H}-$valued SDEs (or stochastic PDEs – SPDEs) with the form

$$dz_t = -h(\ell)\left(z_t + C\nabla\Psi(z_t)\right)dt + \sqrt{2h(\ell)}\,dW_t \tag{4.1.7}$$

where W is a Brownian motion (see [DPZ92]) in $\mathcal{H}$ with covariance operator C and $h(\ell) > 0$ is a positive constant. In [MPS11] the RWM algorithm is studied when applied to a sequence of finite dimensional approximations of π as in (4.1.6). The continuous time interpolant of the Markov chain z^N given by (4.1.4) is shown to converge weakly to z solving (4.1.7) in $C([0,T];\mathcal{H}^s)$. Furthermore, as for the i.i.d target measure the scaling of the proposal variance which achieves this scaling limit is inversely proportional to N (*i.e.* corresponds to the exponent $\gamma = 1$) and the speed of the limiting diffusion process is maximized at the same universal acceptance probability of 0.234 that was found in the i.i.d case. Thus, remarkably, the i.i.d. case has been of fundamental importance in understanding MCMC methods applied to complex infinite dimensional probability measures arising in practice. We use the framework developed in section 3.3 to prove a scaling limit result.

To the best of our knowledge, the only paper to consider the optimal scaling for the MALA algorithm for non-product targets is [BPS04], in the context of non-linear regression. In [BPS04] the target measure has a structure similar to that of the mean field models studied in statistical mechanics and hence behaves asymptotically like a product measure when the dimension goes to infinity. Thus the diffusion limit obtained in [BPS04] is finite dimensional.

The main contribution of this chapter is the proof of a diffusion limit for the output of the MALA algorithm, suitably interpolated, to the SPDE (4.1.7), when applied to $N-$dimensional approximations of the target measures (4.1.6) with

proposal variance inversely proportional to $N^{\frac{1}{3}}$. Moreover we show that the speed $h(\ell)$ of the limiting diffusion is maximized for an average acceptance probability of 0.574, just as in the i.i.d product scenario [RR98]. Thus in this regard, our work is the first extension of the remarkable results in [RR98] for the Langevin algorithm to target measures which are not of product form. This adds theoretical weight to the results observed in computational experiments which demonstrate the robustness of the optimality criteria developed in [RGG97; RR98]. In particular the paper [BRSV08] shows numerical results indicating the need to scale time-step as a function of dimension to obtain $\mathcal{O}(1)$ acceptance probabilities.

In section 4.2 we state the main theorem of this section, having defined precisely the setting in which it holds. Section 4.3 contains the proof of the main theorem, postponing the proof of a number of key technical estimates to section 4.4. In section 4.5 we conclude by summarising and providing the outlook for further research in this area.

4.2 Main theorem

This section is devoted to stating the main theorem of the chapter. We are interested in infinite dimensional target distributions that can be defined as a Gaussian change of measure on a separable Hilbert space. Section 4.2.1 describes this family of target distribution. The reader is referred several times to section 3.1 where the theory of Gaussian measures on Hilbert spaces is introduced. We then define in subsection 4.2.2 the MCMC algorithm that will be analysed. We then discuss in subsection 4.2.3 how the choice of scaling used in the theorem emerges from study of the acceptance probabilities. Finally, in subsection 4.2.4, we state the main theorem.

4.2.1 Target distribution

The reader is referred to section 3.1 for background on Gaussian measures. Let $\mathcal{H}$ be a separable Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and associated norm $\|x\|^2 = \langle x, x \rangle$. We consider a centered Gaussian measure π_0 with covariance operator $C : \mathcal{H} \to \mathcal{H}$. The operator C is trace class and is diagonalisable in an orthonormal Hilbert basis $\{\varphi_j\}_{j \geq 1}$ that will be referred to as *Karhunen-Loève eigen-basis*,

$$C\varphi_j = \lambda_j^2 \, \varphi_j \qquad \text{and} \qquad \mathrm{Tr}(C) = \sum_{j \geq 1} \lambda_j^2 < \infty.$$

In other words, the eigenvalues of the covariance operator C are $\{\lambda_j^2\}_{j \geq 1}$. Any vector $x \in \mathcal{H}$ can be decomposed on the Karhunen-Loève basis as $x = \sum_{j \geq 1} x_j \, \varphi_j$ where

$x_j = \langle x, \varphi_j \rangle$. This decomposition shows the infinite sum $\sum_{j\geq 1} \lambda_j \, \xi_j \, \varphi_j$ where $\{\xi_j\}_{j\geq 1}$ is an i.i.d sequence of standard $\mathrm{N}(0,1)$ Gaussian random variables is distributed as π_0. This expansion of the Gaussian measure π_0 is usually called the *Karhunen-Loève* expansion. Our goal is to sample from a measure π defined through the change of probability formula

$$\frac{d\pi}{d\pi_0}(x) \propto \exp\{-\Psi(x)\} \tag{4.2.1}$$

with respect to the Gaussian measure π_0. The change of probability is assumed to satisfy assumption 3.1.1. This means that there exists an exponent $s \geq 0$ such that the support of π_0 is included in $\mathcal{H}^s$ and that the function Ψ is well defined on $\mathcal{H}^s$ and satisfies various regularity estimates. The Sobolev-like subspace $\mathcal{H}^s$ is rigorously defined in section 3.1.2.

We are interested in finite dimensional approximations of the probability distribution π. To this end, we introduce the vector space spanned by the first N eigenfunctions of the covariance operator,

$$X^N \stackrel{\text{def}}{=} \mathrm{span}\big\{\varphi_1, \varphi_2, \ldots, \varphi_N\big\}.$$

Notice that $X^N \subset \mathcal{H}^r$ for any $r \in [0; +\infty)$. In particular, X^N is a subspace of $\mathcal{H}^s$. Next, we define N-dimensional approximations of the function $\Psi(\cdot)$ and of the reference measure π_0. To this end, we introduce the orthogonal projection on X^N denoted by $P^N : \mathcal{H}^s \mapsto X^N \subset \mathcal{H}^s$. The function $\Psi(\cdot)$ is approximated by the function $\Psi^N : X^N \mapsto \mathbb{R}$ defined by

$$\Psi^N \stackrel{\text{def}}{=} \Psi \circ P^N. \tag{4.2.2}$$

The approximation π_0^N of the reference measure π_0 is the Gaussian measure on X^N given by the law of the random variable

$$\pi_0^N \stackrel{\mathcal{D}}{\sim} \sum_{j=1}^{N} \lambda_j \xi_j \varphi_j \;=\; (C^N)^{\frac{1}{2}} \, \xi^N$$

where ξ_j are i.i.d standard Gaussian random variables, $\xi^N = \sum_{j=1}^N \xi_j \varphi_j$ and $C^N = P^N \circ C \circ P^N$. Consequently we have $\pi_0^N = \mathrm{N}(0, C^N)$. Finally, one can define the approximation π^N of π by the change of probability formula

$$\frac{d\pi^N}{d\pi_0^N}(x) = M_{\Psi^N} \exp\big\{ - \Psi^N(x)\big\} \tag{4.2.3}$$

where M_{Ψ^N} is a normalization constant. Under the assumptions 3.1.1, the normalizing constants M_{Ψ^N} are uniformly bounded and we use this fact to obtain uniform bounds on moments of functions in $\mathcal{H}$ under π^N. Moreover, as N goes to infinity, the sequence of probability distributions π^N converges weakly to the distribution π. This claims are made rigorous in the following lemma.

Lemma 4.2.1. (Finite dimensional approximation π^N of π) *Under the assumptions 3.1.1 the normalization constants M_{Ψ^N} are uniformly bounded. For any measurable function $f : \mathcal{H} \mapsto \mathbb{R}$, we have $\mathbb{E}^{\pi^N}\big[|f(x)|\big] \lesssim \mathbb{E}^{\pi_0}\big[|f(x)|\big]$. The sequence of probability measures π^N converges weakly in $\mathcal{H}^s$ towards the probability distribution π.*

Proof. The first part is contained in Lemma 3.5 of [MPS11]. Let us prove that $\pi^N \Longrightarrow \pi$. We need to show that for any bounded continuous function $g : \mathcal{H}^s \to \mathbb{R}$ we have $\lim_{N\to\infty} \mathbb{E}^{\pi^N}[g(x)] = \mathbb{E}^{\pi}[g(x)]$ where

$$\mathbb{E}^{\pi^N}[g(x)] = \mathbb{E}^{\pi_0^N}[g(x)M_{\Psi^N}e^{-\Psi^N(x)}] = \mathbb{E}^{\pi_0}[g(P^N x)M_{\Psi^N}e^{-\Psi(P^N x)}].$$

Since g is bounded, Ψ is lower bounded and since the normalization constants are uniformly bounded, the dominated convergence theorem shows that it suffices to show that $g(P^N x)M_{\Psi^N}e^{-\Psi(P^N x)}$ converges π_0-almost surely to $g(x)M_{\Psi}e^{-\Psi(x)}$. For this in turn it suffices to show that $\Psi(P^N\, x)$ converges π_0-almost surely to $\Psi(x)$ as this also proves almost sure convergence of the normalization constants. By (3.1.9) we have

$$|\Psi(P^N\, x) - \Psi(x)| \lesssim (1 + \|x\|_s + \|P^N x\|_s)\|P^N\, x - x\|_s.$$

But $\lim_{N\to\infty} \|P^N\, x - x\|_s \to 0$ for any $x \in \mathcal{H}^s$, by dominated convergence, and the result follows. □

Fernique's theorem [DPZ92] implies that for any exponent $p \geq 0$ we have $\mathbb{E}^{\pi^0}\big[\|x\|_s^p\big] < \infty$. It thus follows from lemma 4.2.1 that for any $p \geq 0$

$$\sup_N \Big\{\mathbb{E}^{\pi^N}\big[\|x\|_s^p\big] : N \in \mathbb{N}\Big\} \;<\; \infty.$$

This estimate is repeatedly used in the sequel. Notice that the probability distribution π^N is supported on X^N and has Lebesgue density[2] on X^N equal to

$$\pi^N(x) \;\propto\; \exp\Big(-\frac{1}{2}\|x\|_{C^N}^2 - \Psi^N(x)\Big). \tag{4.2.4}$$

[2]For ease of notation we do not distinguish between a measure and its density, nor do we distinguish between the representation of the measure in X^N or in coordinates in $\mathbb{R}^N$

In formula (4.2.4), the Hilbert-Schmidt norm $\|\cdot\|_{C^N}$ on X^N is given by the scalar product $\langle u, v\rangle_{C^N} = \langle u, (C^N)^{-1}v\rangle$ for all $u, v \in X^N$. The operator $\mathcal{C}^N$ is invertible on X^N because the eigenvalues of C are assumed to be strictly positive. The quantity $C^N\nabla \log \pi^N(x)$ is repeatedly used in the text and in particular appears in the function $\mu^N(x)$ given by

$$\mu^N(x) = -\Big(P^N x + C^N \nabla \Psi^N(x)\Big) \tag{4.2.5}$$

which, up to an additive constant, is $C^N\nabla \log \pi^N(x)$. This function is the drift of an ergodic Langevin diffusion that leaves π^N invariant. Similarly, one defines the function $\mu : \mathcal{H}^s \to \mathcal{H}^s$ given by

$$\mu(x) = -\Big(x + C\nabla \Psi(x)\Big) \tag{4.2.6}$$

which can informally be seen as $C\nabla \log \pi(x)$, up to an additive constant. In the sequel, Lemma 4.4.1 shows that, for π_0-almost every $x \in \mathcal{H}$, we have $\lim_{N\to\infty} \mu^N(x) = \mu(x)$. This quantifies the manner in which the function μ^N is an approximation of the function μ.

The next lemma gathers various regularity estimates on the function $\Psi(\cdot)$ and $\Psi^N(\cdot)$ that are repeatedly used in the sequel. These are simple consequences of assumptions 3.1.1 and proofs can be found in section 3.1 and 3.2 of [MPS11].

Lemma 4.2.2. (Properties of Ψ) *Let the function $\Psi(\cdot)$ satisfy assumptions 3.1.1 and consider the function $\Psi^N(\cdot)$ defined by equation (4.2.2). The following estimates hold.*

1. *The functions $\Psi^N : \mathcal{H}^s \to \mathbb{R}$ satisfy the same conditions imposed on Ψ given by equations (3.1.8), (3.1.9) and (3.1.10) with constants that can be chosen independent of N.*

2. *The function $C\nabla\Psi : \mathcal{H}^s \to \mathcal{H}^s$ is globally Lipschitz on $\mathcal{H}^s$: there exists a constant $M_5 > 0$ such that*

$$\|\mathcal{C}\nabla\Psi(x) - \mathcal{C}\nabla\Psi(y)\|_s \leq M_5 \|x - y\|_s \qquad \forall x, y \in \mathcal{H}^s.$$

   Moreover, the functions $C^N\nabla\Psi^N : \mathcal{H}^s \to \mathcal{H}^s$ also satisfy this estimate with a constant that can be chosen independent of N.

3. *The function $\Psi : \mathcal{H}^s \to \mathbb{R}$ satisfies a second order Taylor formula[3]. There*

[3]We extend $\langle\cdot,\cdot\rangle$ from an inner-product on $\mathcal{H}$ to the dual pairing between $\mathcal{H}^{-s}$ and $\mathcal{H}^s$.

exists a constant $M_6 > 0$ *such that*

$$\Psi(y) - \Big(\Psi(x) + \langle \nabla\Psi(x), y - x\rangle\Big) \leq M_6 \|x - y\|_s^2 \qquad \forall x, y \in \mathcal{H}^s. \tag{4.2.7}$$

Moreover, the functions $\Psi^N(\cdot)$ *also satisfy this estimates with a constant that can be chosen independent of* N*.*

Remark 4.2.3. *Regularity Lemma 4.2.2 shows in particular that the function* $\mu : \mathcal{H}^s \to \mathcal{H}^s$ *defined by (4.2.6) is globally Lipschitz on* $\mathcal{H}^s$*. Similarly, it follows that* $C^N\nabla\Psi^N : \mathcal{H}^s \to \mathcal{H}^s$ *and* $\mu^N : \mathcal{H}^s \to \mathcal{H}^s$ *given by (4.2.5) are globally Lipschitz with Lipschitz constants that can be chosen uniformly in* N*.*

4.2.2 The algorithm

The MALA algorithm is defined in this section. This method is motivated by the fact that the probability measure π^N defined by equation (4.2.3) is invariant with respect to the Langevin diffusion process

$$dz_t = \mu^N(z_t)\, dt + \sqrt{2}\, dW_t^N, \tag{4.2.8}$$

where W^N is a Brownian motion in $\mathcal{H}$ with covariance operator C^N. The drift function $\mu^N : \mathcal{H}^s \to \mathcal{H}^s$ is the gradient of the log-density of π^N, as described by equation (4.2.5). The idea of the MALA algorithm is to make a proposal based on Euler-Maruyama discretization of the diffusion (4.2.8). To this end we consider, from state $x \in X^N$, proposals $y \in X^N$ given by

$$y - x = \delta\, \mu^N(x) + \sqrt{2\delta}\, (C^N)^{\frac{1}{2}} \xi^N \qquad \text{where} \qquad \delta = \ell N^{-\frac{1}{3}} \tag{4.2.9}$$

with $\xi^N = \sum_{i=1}^N \xi_i\, \varphi_i$ and $\xi_i \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1)$. Notice that $(C^N)^{\frac{1}{2}}\xi^N \overset{\mathcal{D}}{\sim} \mathrm{N}(0, C^N)$. The quantity δ is the time-step in an Euler-Maruyama discretization of (4.2.8). We introduce a related parameter

$$\Delta t := \ell^{-1}\delta = N^{-\frac{1}{3}}$$

which will be the natural time-step for the limiting diffusion process derived from the proposal above, after inclusion of an accept-reject mechanism. The scaling of Δt, and hence δ, with N will ensure that the average acceptance probability is bounded away from 0 and 1 as N grows. This is discussed in more detail in section 4.2.3. The quantity $\ell > 0$ is a fixed parameter which can be chosen to maximize the speed of the limiting diffusion process: see the discussion in the introduction and after the

main theorem below.

We will study the Markov chain $x^N = \{x^{k,N}\}_{k\geq 0}$ resulting from Metropolizing this proposal when it is started at stationarity: the initial position $x^{0,N}$ is distributed as π^N and thus lies in X^N. Therefore, the Markov chain evolves in X^N; as a consequence, only the first N components of an expansion in the eigenbasis of C are non-zero and the algorithm can be implemented in $\mathbb{R}^N$. However, the analysis is cleaner when written in $\mathcal{H}^s$. The acceptance probability only depends on the first N coordinates of x and y and has the form

$$\alpha^N(x,\xi^N) = 1 \wedge \frac{\pi^N(y)T^N(y,x)}{\pi^N(x)T^N(x,y)} = 1 \wedge \exp\left(Q^N(x,\xi^N)\right) \tag{4.2.10}$$

where the proposal y is given by equation (4.2.9). The function $T^N(\cdot,\cdot)$ is the density of the Langevin proposals (4.2.9) and is given by

$$T^N(x,y) \propto \exp\Big\{ -\frac{1}{4\delta}\|y - x - \delta\mu^N(x)\|^2_{C^N}\Big\}$$

The local mean acceptance probability $\alpha^N(x)$ is defined by

$$\alpha^N(x) = \mathbb{E}_x\big[\alpha^N(x,\xi^N)\big]. \tag{4.2.11}$$

It is the expected acceptance probability when the algorithm stands at $x \in \mathcal{H}$. The Markov chain $x^N = \{x^{k,N}\}_{k\geq 0}$ can also be expressed as

$$\begin{cases} y^{k,N} & = x^{k,N} + \delta\mu^N(x^{k,N}) + \sqrt{2\delta}\,(C^N)^{\frac{1}{2}}\,\xi^{k,N} \\ x^{k+1,N} & = \gamma^{k,N}y^{k,N} + (1-\gamma^{k,N})\,x^{k,N} \end{cases} \tag{4.2.12}$$

where $\xi^{k,N}$ are i.i.d samples distributed as ξ^N and $\gamma^{k,N} = \gamma^N(x^{k,N},\xi^{k,N})$ creates a Bernoulli random sequence with k^{th} success probability $\alpha^N(x^{k,N},\xi^{k,N})$. We may view the Bernoulli random variable as $\gamma^{k,N} = 1_{\{U^k<\alpha^N(x^{k,N},\xi^{k,N})\}}$ where $U^k \overset{\mathcal{D}}{\sim} \text{Uniform}(0,1)$ is independent from $x^{k,N}$ and $\xi^{k,N}$. The quantity Q^N defined in equation (4.2.10) may be expressed as

$$\begin{aligned} Q^N(x,\xi^N) = &-\frac{1}{2}\Big(\|y\|^2_{C^N} - \|x\|^2_{C^N}\Big) - \Big(\Psi^N(y) - \Psi^N(x)\Big) \\ &- \frac{1}{4\delta}\Big\{\|x - y - \delta\mu^N(y)\|^2_{C^N} - \|y - x - \delta\mu^N(x)\|^2_{C^N}\Big\}. \end{aligned} \tag{4.2.13}$$

As will be seen in the next section, a key idea behind our diffusion limit is that, for large N, the quantity $Q^N(x,\xi^N)$ behaves like a Gaussian random variable independent of the current position x.

In summary, the Markov chain that we have described in $\mathcal{H}^s$ is, when projected onto X^N, equivalent to a standard MALA algorithm on $\mathbb{R}^N$ for the Lebesgue density (4.2.4). Recall that the target measure π in (4.1.6) is the invariant measure of the SPDE (4.1.7). Our goal is to obtain an invariance principle for the continuous interpolant (4.1.4) of the Markov chain $x^N = \{x^{k,N}\}_{k\geq 0}$ started in stationarity, *i.e*, to show weak convergence in $C([0,T];\mathcal{H}^s)$ of $z^N(t)$ to the solution $z(t)$ of the SPDE (4.1.7), as the dimension $N \to \infty$.

4.2.3 Optimal scale $\gamma = \frac{1}{3}$

In this section, we informally describe why the optimal scale for the MALA proposals (4.2.9) is given by the exponent $\gamma = \frac{1}{3}$. For product-form target probability described by equation (4.1.1), the optimality of the exponent $\gamma = \frac{1}{3}$ was first obtained in [RR98]. For further discussion, see also [BRS09]. To keep the exposition simple in this explanatory subsection we focus on the case $\Psi(\cdot) = 0$. The analysis is similar with a non-vanishing function $\Psi(\cdot)$, because absolute continuity ensures that the effect of $\Psi(\cdot)$ is small compared to the dominant Gaussian effects described here. Inclusion of non-vanishing $\Psi(\cdot)$ is carried out in Lemma 4.4.3.

In the case $\Psi(\cdot) = 0$, straightforward algebra shows that the acceptance probability $\alpha^N(x,\xi^N) = 1 \wedge e^{Q^N(x,\xi^N)}$ satisfies

$$Q^N(x,\xi^N) = -\frac{\ell\Delta t}{4}\Big(\|y\|^2_{C^N} - \|x\|^2_{C^N}\Big).$$

For $\Psi(\cdot) = 0$ and $x \in X^N$, the proposal y is distributed as $y = (1-\ell\Delta t)x + \sqrt{2\ell\Delta t}(C^N)^{\frac{1}{2}}\xi^N$. It follows that

$$\begin{aligned}\|y\|^2_{C^N} - \|x\|^2_{C^N} = -2\ell\Delta t\left(\|x\|^2_{C^N} - \|(C^N)^{\frac{1}{2}}\xi^N\|^2_{C^N}\right) + (\ell\Delta t)^2\|x\|^2_{C^N} \\ + 2\sqrt{2\ell\Delta t}(1-\Delta t)\,\langle x,(C^N)^{\frac{1}{2}}\xi^N\rangle_{C^N}.\end{aligned}$$

The details can be found in the proof of lemma 4.4.3. Since the Markov chain $x^N = \{x^{k,N}\}_{k\geq 0}$ evolves in stationarity, for all $k \geq 0$ we have $x^{k,N} \overset{\mathcal{D}}{\sim} \pi^N = \mathrm{N}(0,C^N)$. Therefore, with $x \overset{\mathcal{D}}{\sim} \mathrm{N}(0,C^N)$ and $\xi^N \overset{\mathcal{D}}{\sim} \mathrm{N}(0,C^N)$, the law of large numbers shows that both $\|x\|^2_{C^N}$ and $\|(C^N)^{\frac{1}{2}}\xi^N\|^2_{C^N}$ are of order $\mathcal{O}(N)$, whilst the central limit theorem shows that $\langle x,(C^N)^{\frac{1}{2}}\xi^N\rangle_{C^N} = \mathcal{O}(N^{\frac{1}{2}})$ and $\|x\|^2_{C^N} - \|(C^N)^{\frac{1}{2}}\xi^N\|^2_{C^N} = \mathcal{O}(N^{\frac{1}{2}})$. For $\Delta t = \ell N^{-\gamma}$ and $\gamma < \frac{1}{3}$, it follows

$$Q^N(x,\xi^N) = -\frac{(\ell\Delta t)^3}{4}\|x\|^2_{C^N} + \mathcal{O}(N^{\frac{1}{2}-\frac{3\gamma}{2}}) \;\approx\; -\frac{\ell^3}{4}N^{1-3\gamma},$$

which shows that the acceptance probability is exponentially small of order $\exp\big(-\frac{\ell^3}{4}N^{1-3\gamma}\big)$. The same argument shows that for $\gamma > \frac{1}{3}$ we have $Q^N(x,\xi^N) \to 0$, which shows that the average acceptance probability converges to 1. For the critical exponent $\gamma = \frac{1}{3}$ the acceptance probability is of order $\mathcal{O}(1)$. In fact Lemma 4.4.3 shows that for $\gamma = \frac{1}{3}$, even when $\Psi(\cdot)$ is non-zero, the quantity $Q^N(x,\xi^N)$ can be approximated by a Gaussian random variable $\mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2})$. This approximation is key to derivation of the diffusion limit. In summary, choosing $\gamma > \frac{1}{3}$ leads to exponentially small acceptance probabilities: almost all the proposals are rejected so that the expected squared jumping distance $\mathbb{E}_{\pi^N}[\|x^{k+1,N} - x^{k,N}\|^2]$ converges exponentially quickly to 0 as the dimension N goes to infinity. On the other hand, for any exponent $\gamma \geq \frac{1}{3}$, the acceptance probabilities are bounded away from zero: the Markov chain moves with jumps of size $\mathcal{O}(N^{-\frac{\gamma}{2}})$ and the expected squared jumping distance is of order $\mathcal{O}(N^{-\gamma})$. If we adopt the expected squared jumping distance as measure of efficiency, the optimal exponent is thus given by $\gamma = \frac{1}{3}$. This viewpoint is analyzed further in [BRS09].

4.2.4 Statement of main theorem

The main result of this chapter describes the behavior of the MALA algorithm for the optimal scale $\gamma = \frac{1}{3}$; the proposal variance is given by $2\delta = 2\ell N^{-\frac{1}{3}}$. In this case, Lemma 4.4.3 and 4.4.4 show that the local mean acceptance probability $\alpha^N(x,\xi^N) = 1 \wedge e^{Q^N(x,\xi^N)}$ is such that $Q^N(x,\xi^N)$ converges to $Z_\ell \overset{\mathcal{D}}{\sim} \mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2})$ in the Wasserstein metric. As a consequence, the asymptotic mean acceptance probability of the MALA algorithm can be explicitly computed as a function of the parameter $\ell > 0$,

$$\alpha(\ell) \overset{\text{def}}{=} \lim_{N\to\infty} \mathbb{E}^{\pi^N}\big[\alpha^N(x,\xi^N)\big] = \mathbb{E}\big[1 \wedge e^{Z_\ell}\big]. \tag{4.2.14}$$

This result is rigorously proved as Corollary 4.4.5. We then define the "speed function"

$$h(\ell) = \ell \times \alpha(\ell). \tag{4.2.15}$$

Note that the time step made in the proposal is $\delta = \ell\Delta t$ and that, if this is accepted a fraction $\alpha(\ell)$ of the time, then a naive argument invoking independence shows that the effective time-step is reduced to $h(\ell)\Delta t$. This is made rigorous in theorem 4.2.4 which shows that the quantity $h(\ell)$ is the asymptotic speed function of the limiting diffusion obtained by rescaling the sequence of Metropolis-Hastings Markov chains $\{x^N\}_{N\geq 1}$.

Theorem 4.2.4. (Main theorem) *Let the reference measure π_0 and the function $\Psi(\cdot)$ satisfy assumptions 3.1.1. Consider the MALA algorithm (4.2.12) with initial condition $x^{0,N} \overset{\mathcal{D}}{\sim} \pi^N$. Let $z^N(t)$ be the piecewise linear, continuous interpolant of the MALA algorithm as defined in (4.1.4), with $\Delta t = N^{-\frac{1}{3}}$. Then $z^N(t)$ converges weakly in $C([0,T],\mathcal{H}^s)$ to the diffusion process $z(t)$ given by*

$$dz_t = -h(\ell)\big(z_t + C\nabla\Psi(z_t)\big)\,dt + \sqrt{2\,h(\ell)}\,dW_t \tag{4.2.16}$$

with initial distribution $z(0) \overset{\mathcal{D}}{\sim} \pi$.

We now explain the following two important implications of this result.

- Since time has to be accelerated by a factor $(\Delta t)^{-1} = N^{\frac{1}{3}}$ in order to observe a diffusion limit, it follows that in stationarity the work required to explore the invariant measure scales as $\mathcal{O}(N^{\frac{1}{3}})$.
- The speed at which the invariant measure is explored, again in stationarity, is maximized by choosing ℓ so as to maximize $h(\ell)$; this is achieved at an average acceptance probability 0.574. From a practical point of view, this shows that when dealing with target distributions that are discretisations of infinite dimensional Gaussian change of measure, one should "tune" the proposal variance of the MALA algorithm so as to have a mean acceptance probability of 0.574. This result holds mainly because we are considering target distributions that are dominated by their Gaussian components i.e. the functional Ψ appearing in the change of probability formula (4.2.1) has nice growth and regularity properties. Indeed, for more complex target distributions, there is no reason why such an optimality result should hold.

The first implication follows from (4.1.4) since this shows that $\mathcal{O}(N^{\frac{1}{3}})$ steps of the MALA Markov chain (4.2.12) are required for $z^N(t)$ to approximate $z(t)$ on a time interval $[0,T]$ long enough for $z(t)$ to have explored its invariant measure. To understand the second implication, note that if $Z(t)$ solves (4.2.16) with $h(\ell) \equiv 1$ then, in law, $z(t) = Z\big(h(\ell)t\big)$. This result suggests choosing the value of ℓ that maximizes the speed function $h(\cdot)$ since $z(t)$ will then explore the invariant measure as fast as possible. For practitioners, who often tune algorithms according to the acceptance probability, it is relevant to express the maximization principle in terms of the asymptotic mean acceptance probability $\alpha(\ell)$. Figure 4.2.4 shows that the speed function $h(\cdot)$ is maximized for an optimal acceptance probability of $\alpha^\star = 0.574$, to three decimal places. This is precisely the argument used in [RR98] for the case of product target measures and it is remarkable that the optimal acceptance

probability identified in that context is also optimal for the non-product measures studied in this section.

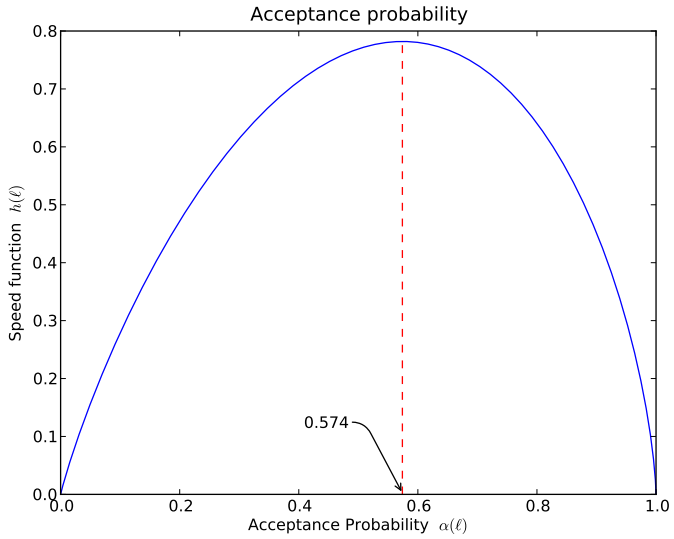


Figure 4.1: Optimal acceptance probability = 0.574

4.3 Proof of main theorem

In subsection 4.3.1 we outline the proof strategy and introduce the drift-martingale decomposition of our discrete-time Markov chain which underlies it. In subsection 4.3.2 we use the general diffusion approximation result stated in Proposition 3.3.1 of section 3.3 to prove the main theorem of this paper, pointing to section 4.4 for the key estimates required.

4.3.1 Proof strategy

To communicate the main ideas, we give a heuristic of the proof before proceeding to give full details in subsequent sections. Let us first examine a simpler situation: consider a scalar Lipschitz function $\mu : \mathbb{R} \to \mathbb{R}$ and two scalar constants $\ell, c > 0$. The usual theory of diffusion approximation for Markov processes [EK86] shows that the sequence $x^N = \{x^{k,N}\}$ of Markov chains

$$x^{k+1,N} - x^{k,N} = \mu(x^{k,N})\,\ell N^{-\frac{1}{3}} + \sqrt{2\ell N^{-\frac{1}{3}}}\; c^{\frac{1}{2}}\xi^k,$$

with i.i.d. $\xi^k \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1)$ converges weakly, when interpolated using a time-acceleration factor of $N^{\frac{1}{3}}$, to the scalar diffusion $dz_t = \ell\mu\big(z_t\big)\,dt + \sqrt{2\ell}\,dW_t$ where W is a Brownian motion with variance $\mathrm{Var}\big(W(t)\big) = ct$. Also, if γ^k is an i.i.d. sequence of Bernoulli random variables with success rate $\alpha(\ell)$, independent from the Markov chain x^N, one can prove that the sequence $x^N = \big\{x^{k,N}\big\}$ of Markov chains given by

$$x^{k+1,N} - x^{k,N} = \gamma^k \left\{\mu(x^{k,N})\ell N^{-\frac{1}{3}} + \sqrt{2\ell N^{-\frac{1}{3}}}\; c^{\frac{1}{2}}\xi^k\right\}$$

converges weakly, when interpolated using a time-acceleration factor $N^{\frac{1}{3}}$, to the diffusion

$$dz_t = h(\ell)\mu\big(z_t\big)\,dt + \sqrt{2h(\ell)}\,dW_t$$

where the speed function is given by $h(\ell) = \ell\alpha(\ell)$. This shows that the Bernoulli random variables $\big\{\gamma^k\big\}_{k\geq 0}$ have slowed down the original Markov chain by a factor $\alpha(\ell)$. The proof of theorem 4.2.4 is an application of this idea in a slightly more general setting. The following complications arise.

- Instead of working with scalar diffusions, the result holds for a Hilbert space-valued diffusion. The correlation structure between the different coordinates is not present in the preceding simple example and has to be taken into account.
- Instead of working with a single drift function μ, a sequence of approximations μ^N converging to μ has to be taken into account.
- The Bernoulli random variables $\gamma^{k,N}$ are not i.i.d. and have an autocorrelation structure. On top of that, the Bernoulli random variables $\gamma^{k,N}$ are not independent from the Markov chain $x^{k,N}$. This is the main difficulty in the proof.
- It should be emphasized that the main theorem uses the fact that the MALA Markov chain is started at stationarity. This in particular implies that $x^{k,N} \overset{\mathcal{D}}{\sim} \pi^N$ for any $k \geq 0$, which is crucial to the proof of the invariance principle as it allows us to control the correlation between $\gamma^{k,N}$ and $x^{k,N}$.

The rigorous proof of the main result 4.2.4 is based on Proposition 3.3.1. To this end, we need to introduce a martingale-drift decomposition of the Markov chain $x^N = \{x^{k,N}\}_{k\geq 0}$ and obtain a good understanding of the accept-reject mechanism of the MALA algorithm. The acceptance probability of proposal (4.2.9) is equal to $\alpha^N(x,\xi^N) = 1 \wedge e^{Q^N(x,\xi^N)}$ and the quantity $\alpha^N(x) = \mathbb{E}_x[\alpha^N(x,\xi^N)]$ given by (4.2.11)

represents the mean acceptance probability when the Markov chain x^N stands at x. For our proof it is important to understand how the acceptance probability $\alpha^N(x, \xi^N)$ depends on the current position x and on the source of randomness ξ^N. Recall quantity Q^N defined in equation (4.2.13). The main observation underlying the proof of our main result is that $Q^N(x, \xi^N)$ can be approximated by a Gaussian random variable

$$Q^N(x, \xi^N) \;\approx\; Z_\ell \tag{4.3.1}$$

where $Z_\ell \overset{\mathcal{D}}{\sim} \mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2})$. These approximations are made rigorous in Lemma 4.4.3 and Lemma 4.4.4. Therefore, the Bernoulli random variable $\gamma^N(x, \xi^N)$ with success probability $1 \wedge e^{Q^N(x,\xi^N)}$ can be approximated by a Bernoulli random variable, independent of x, with success probability equal to

$$\alpha(\ell) = \mathbb{E}\big[1 \wedge e^{Z_\ell}\big]. \tag{4.3.2}$$

Thus, the limiting acceptance probability of the MALA algorithm is as given in equation (4.3.2). We now introduce the drift-martingale decomposition of the Markov chain x^N. Recall that $\Delta t = N^{-\frac{1}{3}}$. With this notation we introduce the drift function $d^N : \mathcal{H}^s \to \mathcal{H}^s$ given by

$$d^N(x) = \big(h(\ell)\Delta t\big)^{-1}\mathbb{E}\big[x^{1,N} - x^{0,N} \,|x^{0,N} = x\big] \tag{4.3.3}$$

and the martingale difference array $\{\Gamma^{k,N} : k \geq 0\}$ defined by $\Gamma^{k,N} = \Gamma^N(x^{k,N}, \xi^{k,N})$ with

$$\Gamma^{k,N} = \big(2h(\ell)\Delta t\big)^{-\frac{1}{2}}\Big(x^{k+1,N} - x^{k,N} - h(\ell)\Delta t\; d^N(x^{k,N})\Big). \tag{4.3.4}$$

The normalization constant $h(\ell)$ defined in equation (4.2.15) ensures that the drift function μ^N and the martingale difference array $\{\Gamma^{k,N}\}$ are asymptotically independent from the parameter ℓ. The drift-martingale decomposition of the Markov chain $\{x^{k,N}\}_k$ then reads

$$x^{k+1,N} - x^{k,N} = h(\ell)\, d^N(x^{k,N})\, \Delta t + \sqrt{2h(\ell)\Delta t}\; \Gamma^{k,N}. \tag{4.3.5}$$

In order to use the general diffusion-approximation result given by Proposition 3.3.1, one needs to quantify how close the approximate drift function $\mu^N(\cdot)$ is from the limiting drift function $\mu(\cdot)$ defined by equation (4.2.6) and prove a Brownian scaling

limit for the sequence of processes W^N,

$$W^N(t) = \sqrt{\Delta t} \sum_{j=0}^{k} \Gamma^{j,N} + \frac{t - k\Delta t}{\sqrt{\Delta t}} \Gamma^{k+1,N} \tag{4.3.6}$$

for $k\Delta t \leq t < (k+1)\Delta t$. This is done in Lemma 4.4.6 and Proposition 4.4.9.

4.3.2 Proof of main theorem

The proof of Theorem 4.2.4 consists in checking that the three conditions needed for Proposition 3.3.1 to apply are satisfied by the sequence of drift-martingale decompositions 4.3.5 of the MALA Markov chains (4.2.12) evolving in the separable Hilbert space $\mathcal{H}^s$. The key estimates are proved in section 4.4.

1. By Lemma 4.2.1 the sequence of probability measures π^N converges weakly in $\mathcal{H}^s$ to the probability measure π that verifies $\mathbb{E}^\pi \|X\|_s < \infty$.

2. Proposition 4.4.9 below proves that $(x^{0,N}, W^N)$ converges weakly in $\mathcal{H}^s \times C([0,T], \mathcal{H}^s)$ to (z^0, W), where W is a Brownian motion in $\mathcal{H}^s$ with covariance C_s independent from $z^0 \overset{\mathcal{D}}{\sim} \pi$.

3. Proposition 4.2.2 below shows that $\mu : \mathcal{H}^s \to \mathcal{H}^s$ is a Lipschitz function. Since the Markov chain $x^N = \{x^{k,N}\}_{k\geq 0}$ evolves at stationarity (and thus $x^{k,N} \overset{\mathcal{D}}{\sim} \pi^N$ for all $k \geq 0$), we have

$$\begin{aligned}
\mathbb{E} \int_0^T \left\| d^N(\bar{z}^N(u)) - \mu(z^N(u)) \right\|_s du &\lesssim \Delta t \sum_{k\Delta t \leq T} \mathbb{E} \left\| d^N(x^{k,N}) - \mu(x^{k+1,N}) \right\|_s \\
&= \Delta t \sum_{k\Delta t \leq T} \mathbb{E} \left\| d^N(x^{0,N}) - \mu(x^{1,N}) \right\|_s \lesssim T \times \mathbb{E} \left\| d^N(x^{0,N}) - \mu(x^{1,N}) \right\|_s \\
&\lesssim \mathbb{E} \left\| d^N(x^{0,N}) - \mu(x^{0,N}) \right\|_s + \left\| \mu(x^{0,N}) - \mu(x^{1,N}) \right\|_s .
\end{aligned}$$

   Lemma 4.4.6 implies that the drift function $d^N(x)$ verifies $\lim_N \mathbb{E}^{\pi^N} \left\| d^N(x) - \mu(x) \right\|_s = 0$. Also, since the function $\mu(\cdot)$ is globally Lipschitz in $\mathcal{H}^s$, we have that $\mathbb{E} \left\| \mu(x^{0,N} - \mu(x^{1,N}) \right\|_s \lesssim \mathbb{E} \left\| x^{0,N} - x^{1,N} \right\|_s \to 0$. This implies that, as $N \to \infty$, the quantity $\int_0^T \left\| d^N(\bar{z}^N(u)) - \mu(z^N(u)) \right\|_s du$ converges in expectation and thus in probability to zero.

The three assumptions needed for Proposition 3.3.1 to apply are satisfied, which concludes the proof of Theorem 4.2.4.

4.4 Key Estimates

Subsection 4.4.1 contains some technical lemmas of use throughout. In section 4.4.2 we study the large N Gaussian approximation of the acceptance probability and establish that this acceptance probability is asymptotic independent of the current state of the Markov chain. This approximation is then used in subsections 4.4.3 and 4.4.4 to give quantitative versions of the heuristics $d^N(\cdot) \approx \mu(\cdot)$ The section concludes with subsection 4.4.5 in which we prove an invariance principle for W^N given by (4.3.6).

4.4.1 Technical lemmas

The first lemma shows that, for π_0-almost every function $x \in \mathcal{H}^s$, the approximation $\mu^N(x) \approx \mu(x)$ holds as N goes to infinity.

Lemma 4.4.1. (μ^N converges π_0-almost surely to μ) *Let assumptions 3.1.1 hold. The sequences of functions $\mu^N : \mathcal{H}^s \to \mathcal{H}^s$ satisfies*

$$\pi_0\Big(\Big\{x \in \mathcal{H}^s : \lim_{N\to\infty} \|\mu^N(x) - \mu(x)\|_s = 0 \Big\}\Big) = 1.$$

Proof. It is enough to verify that for any $x \in \mathcal{H}^s$ the quantity $\|P^N x - x\|_s = 0$ and the quantity $\|CP^N\nabla\Psi(P^N x) - C\nabla\Psi(x)\|_s = 0$ converge to zero as N goes to infinity.

- Let us prove the first equation. For $x \in \mathcal{H}^s$ we have $\sum_{j\geq 1} j^{2s}x_j^2 < \infty$ so that

$$\lim_{N\to\infty} \|P^N x - x\|_s^2 = \lim_{N\to\infty} \sum_{j=N+1}^{\infty} j^{2s}x_j^2 = 0. \tag{4.4.1}$$

- To prove the second equation one can start by using the triangle inequality,

$$\begin{aligned}\|CP^N\nabla\Psi(P^N x) - C\nabla\Psi(x)\|_s \leq \|CP^N\nabla\Psi(P^N x) - CP^N\nabla\Psi(x)\|_s \\ + \|CP^N\nabla\Psi(x) - C\nabla\Psi(x)\|_s.\end{aligned}$$

  The same proof as Lemma 4.2.2 reveals that $CP^N\nabla\Psi : \mathcal{H}^s \to \mathcal{H}^s$ is globally Lipschitz, with a Lipschitz constant that can be chosen independent from N. Consequenly, Equation (4.4.1) shows that $\|CP^N\nabla\Psi(P^N x) - CP^N\nabla\Psi(x)\|_s \lesssim \|P^N x - x\|_s \to 0$. Also, since $z = \nabla\Psi(x) \in \mathcal{H}^{-s}$ we have $\|\nabla\Psi(x)\|_{-s}^2 = \sum_{j\geq 1} j^{-2s}z_j^2 < \infty$. The eigenvalues of C satisfy $\lambda_j^2 \asymp j^{-2\kappa}$ with $s < \kappa - \frac{1}{2}$.

Consequently,

$$\|CP^N\nabla\Psi(x) - C\nabla\Psi(x)\|_s^2 = \sum_{j=N+1}^{\infty} j^{2s}(\lambda_j^2 z_j)^2 \lesssim \sum_{j=N+1}^{\infty} j^{2s-4\kappa} z_j^2$$
$$= \sum_{j=N+1}^{\infty} j^{4(s-\kappa)} j^{-2s} z_j^2 \leq \frac{1}{(N+1)^{4(\kappa-s)}} \|\nabla\Psi(x)\|_{-s}^2 \ \to \ 0.$$

□

Next lemma shows that the size of the jump $y - x$ is of order $\sqrt{\Delta t}$.

Lemma 4.4.2. *Consider y given by (4.2.9). Under assumptions 3.1.1, for any $p \geq 1$ we have*

$$\mathbb{E}_x^{\pi^N}\left[\|y - x\|_s^p\right] \lesssim (\Delta t)^{\frac{p}{2}} \cdot (1 + \|x\|_s^p).$$

Proof. Under assumption 3.1.1 the function μ^N is globally Lipschitz on $\mathcal{H}^s$, with Lipschitz constant that can be chosen independent from N. Thus

$$\|y - x\|_s \lesssim \Delta t(1 + \|x\|_s) + \sqrt{\Delta t}\, \|\mathcal{C}^{\frac{1}{2}}\xi^N\|_s.$$

We have $\mathbb{E}^{\pi^0}\left[\|\mathcal{C}^{\frac{1}{2}}\xi^N\|_s^p\right] \leq \mathbb{E}^{\pi^0}\left[\|\zeta\|_s^p\right]$, where $\zeta \overset{\mathcal{D}}{\sim} \mathrm{N}(0, \mathcal{C})$. From Fernique's theorem [DPZ92] it follows that $\mathbb{E}^{\pi^0}\left[\|\zeta\|_s^p\right] < \infty$ so that the expectation $\mathbb{E}^{\pi^0}\left[\|\mathcal{C}^{\frac{1}{2}}\xi^N\|_s^p\right]$ is uniformly bounded as a function of N, proving the lemma. □

4.4.2 Gaussian approximation of Q^N

In this section we prove several preliminary results that shed some lights on the Gaussian behaviour of the quantity Q^N. These estimates are at the heart of the proof of Theorem 4.2.4. Recall the quantity Q^N defined in Equation (4.2.13). This section proves that Q^N has a Gaussian behavior in the sense that

$$Q^N(x, \xi^N) = Z^N(x, \xi^N) \ + \ i^N(x, \xi^N) \ + \ \mathbf{e}^N(x, \xi^N) \tag{4.4.2}$$

where the quantities Z^N and i^N are equal to

$$Z^N(x, \xi^N) = -\frac{\ell^3}{4} - \frac{\ell^{\frac{3}{2}}}{\sqrt{2}} N^{-\frac{1}{2}} \sum_{j=1}^{N} \lambda_j^{-1} \xi_j x_j \tag{4.4.3}$$

$$i^N(x, \xi^N) = \frac{1}{2}(\ell\Delta t)^2 \Big(\|x\|_{C^N}^2 - \|(C^N)^{\frac{1}{2}}\xi^N\|_{C^N}^2 \Big) \tag{4.4.4}$$

with i^N and e^N small. Thus the principal contributions to Q^N comes from the random variable $Z^N(x,\xi^N)$. Notice that, for each fixed $x \in \mathcal{H}^s$, the random variable $Z^N(x,\xi^N)$ is Gaussian. Furthermore, the Karhunen-Loève expansion of π_0 shows that for π_0-almost every choice of function $x \in \mathcal{H}$ the sequence $\left\{Z^N(x,\xi^N)\right\}_{N\geq 1}$ converges in law to the distribution of $Z_\ell \overset{\mathcal{D}}{\sim} \mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2})$. The next lemma rigorously bounds the error terms $\mathbf{e}^N(x,\xi^N)$ and $i^N(x,\xi^N)$: we show that i^N is an error term of order $\mathcal{O}(N^{-\frac{1}{6}})$ and $\mathbf{e}^N(x,\xi)$ is an error term of order $\mathcal{O}(N^{-\frac{1}{3}})$. In Lemma 4.4.4 we then quantify the convergence of $Z^N(x,\xi^N)$ to Z_ℓ.

Lemma 4.4.3. (Gaussian Approximation) *Let $p \geq 1$ be an integer. Under assumptions 3.1.1 the error terms i^N and $\boldsymbol{e}^N$ in the Gaussian approximation (4.4.2) satisfy*

$$\left(\mathbb{E}^{\pi^N}\left[|i^N(x,\xi^N)|^p\right]\right)^{\frac{1}{p}} = \mathcal{O}(N^{-\frac{1}{6}}) \qquad \textit{and} \qquad \left(\mathbb{E}^{\pi^N}\left[|\boldsymbol{e}^N(x,\xi^N)|^p\right]\right)^{\frac{1}{p}} = \mathcal{O}(N^{-\frac{1}{3}}). \tag{4.4.5}$$

Proof. For notational clarity, without loss of generality, we suppose $p = 2q$. The quantity Q^N is defined in Equation (4.2.13) and expanding terms leads to

$$Q^N(x,\xi^N) \;=\; I_1 \;+\; I_2 \;+\; I_3$$

where the quantities I_1, I_2 and I_3 are given by

$$\begin{aligned}
I_1 &= -\frac{1}{2}\left(\|y\|^2_{C^N} - \|x\|^2_{C^N}\right) - \frac{1}{4\ell\Delta t}\left(\|x - y(1-\ell\Delta t)\|^2_{C^N} - \|y - x(1-\ell\Delta t)\|^2_{C^N}\right)\\
I_2 &= -\Big(\Psi^N(y) - \Psi^N(x)\Big) - \frac{1}{2}\Big(\langle x - y(1-\ell\Delta t), C^N\nabla\Psi^N(y)\rangle_{C^N}\\
&\qquad\qquad - \langle y - x(1-\ell\Delta t), C^N\nabla\Psi^N(x)\rangle_{C^N}\Big)\\
I_3 &= -\frac{\ell\Delta t}{4}\Big\{\|C^N\nabla\Psi^N(y)\|^2_{C^N} - \|C^N\nabla\Psi^N(x)\|^2_{C^N}\Big\}.
\end{aligned}$$

The term I_1 arises purely from the Gaussian part of the target measure π^N and from the Gaussian part of the proposal. The two other terms I_2 and I_3 come from the change of probability involving the function Ψ^N. We start by simplifyng the expression for I_1, and then return to estimate the terms I_2 and I_3.

$$\begin{aligned}
I_1 &= -\frac{1}{2}\left(\|y\|^2_{C^N} - \|x\|^2_{C^N}\right) - \frac{1}{4\ell\Delta t}\left(\|(x-y) + \ell\Delta t\; y)\|^2_{C^N} - \|(y-x) + \ell\Delta t\; x)\|^2_{C^N}\right)\\
&= -\frac{1}{2}\left(\|y\|^2_{C^N} - \|x\|^2_{C^N}\right) - \frac{1}{4\ell\Delta t}\left(2\ell\Delta t[\|x\|^2_{C^N} - \|y\|^2_{C^N}] + (\ell\Delta t)^2[\|y\|^2_{C^N} - \|x\|^2_{C^N}]\right)\\
&= -\frac{\ell\Delta t}{4}\Big(\|y\|^2_{C^N} - \|x\|^2_{C^N}\Big).
\end{aligned}$$

The term I_1 is $\mathcal{O}(1)$ and constitutes the main contribution to Q^N. Before analyzing I_1 in more detail, we show that I_2 and I_3 are of order $N^{-\frac{1}{3}}$,

$$\Big(\mathbb{E}^{\pi^N}[I_2^{2q}]\Big)^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{1}{3}}) \qquad \text{and} \qquad \Big(\mathbb{E}^{\pi^N}[I_3^{2q}]\Big)^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{1}{3}}). \tag{4.4.6}$$

- We expand I_2 and use the bound on the remainder of the Taylor expansion of Ψ described in Equation (4.2.7),

$$\begin{aligned} I_2 = -\Big\{\Psi^N(y) - [\Psi^N(x) + \langle \nabla\Psi^N(x), y-x\rangle]\Big\} + \frac{1}{2}\langle y-x, \nabla\Psi^N(y) - \nabla\Psi^N(x)\rangle \\ + \frac{\ell\Delta t}{2}\Big\{\langle x, \nabla\Psi^N(x)\rangle - \langle y, \nabla\Psi^N(y)\rangle\Big\} = A_1 + A_2 + A_3. \end{aligned}$$

  Equation (4.2.7) and Lemma 4.4.2 show that

$$\mathbb{E}^{\pi^N}[A_1^{2q}] \lesssim \mathbb{E}^{\pi^N}[\|y-x\|_s^{4q}] \lesssim (\Delta t)^{2q}\,\mathbb{E}^{\pi^N}[1+\|x\|_s^{4q}] \lesssim (\Delta t)^{2q} = \Big(N^{-\frac{1}{3}}\Big)^{2q},$$

  where we have used the fact that $\mathbb{E}^{\pi^N}[\|x\|_s^{4q}] \lesssim \mathbb{E}^{\pi_0}[\|x\|_s^{4q}] < \infty$. Assumption 3.1.1 states that $\partial^2\Psi$ is uniformly bounded in $\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})$ so that

$$\begin{aligned} \|\nabla\Psi(y) - \nabla\Psi(y)\|_{-s} &= \Big\| \int_0^1 \partial^2\Psi\big(x + t(y-x)\big)\cdot(y-x)\,dt\Big\|_{-s} \\ &\leq \int_0^1 \|\partial^2\Psi\big(x+t(y-x)\big)\cdot(y-x)\|_{-s}\,dt \leq M_4\int_0^1 \|y-x\|_s\,dt. \end{aligned} \tag{4.4.7}$$

  This proves that $\|\nabla\Psi^N(y) - \nabla\Psi^N(x)\|_{-s} \lesssim \|y-x\|_s$. Consequently, Lemma 4.4.2 shows that

$$\begin{aligned} \mathbb{E}^{\pi^N}[A_2^{2q}] &\lesssim \mathbb{E}^{\pi^N}\Big[\|y-x\|_s^{2q}\cdot\|\nabla\Psi^N(y) - \nabla\Psi^N(x)\|_{-s}^{2q}\Big] \\ &\lesssim \mathbb{E}^{\pi^N}\Big[\|y-x\|_s^{4q}\Big] \lesssim (\Delta t)^{2q}\,\mathbb{E}^{\pi^N}\Big[1+\|x\|_s^{4q}\Big] \lesssim (\Delta t)^2 \;=\; N^{-\frac{2q}{3}}. \end{aligned}$$

  Under assumptions 3.1.1, for any $z \in \mathcal{H}^s$ we have $\|\nabla\Psi^N(z)\|_{-s} \lesssim 1 + \|z\|_s$. Therefore we have $\mathbb{E}^{\pi^N}[A_3^{2q}] \lesssim (\Delta t)^{2q}$ and

$$\Big(\mathbb{E}^{\pi^N}[I_2^{2q}]\Big)^{\frac{1}{2q}} \;\lesssim\; \Big(\mathbb{E}^{\pi^N}[A_1^{2q} + A_2^{2q} + A_3^{2q}]\Big)^{\frac{1}{2q}} \;=\; \mathcal{O}(N^{-\frac{1}{3}}).$$

- Lemma 4.2.2 states $C^N\nabla\Psi^N : \mathcal{H}^s \to \mathcal{H}^s$ is globally Lipschitz, with a Lipschitz constant that can be chosen uniformly in N. Therefore,

$$\|C^N\nabla\Psi^N(z)\|_s \lesssim 1 + \|z\|_s. \tag{4.4.8}$$

Since $\|C^N \nabla \Psi^N(z)\|^2_{C^N} = \langle \nabla \Psi^N(z), C^N \nabla \Psi^N(z) \rangle$, the bound (3.1.9) gives

$$\begin{aligned}
\mathbb{E}^{\pi^N}\left[I_3^{2q}\right] &\lesssim \Delta t^{2q}\, \mathbb{E}\Big[\langle \nabla \Psi^N(x), C^N \nabla \Psi^N(x)\rangle^q + \langle \nabla \Psi^N(y), C^N \nabla \Psi^N(y)\rangle^q\Big] \\
&\lesssim \Delta t^{2q}\, \mathbb{E}^{\pi^N}\Big[(1+\|x\|_s)^{2q} + (1+\|y\|_s)^{2q}\Big] \\
&\lesssim \Delta t^{2q}\, \mathbb{E}^{\pi^N}\Big[1 + \|x\|_s^{2q} + \|y\|_s^{2q}\Big] \;\lesssim\; \Delta t^{2q} \;=\; N^{-\frac{2q}{3}},
\end{aligned}$$

which concludes the proof of Equation (4.4.6).

We now simplify further the expression for I_1 and demonstrate that it has a Gaussian behaviour. We use the definition of the proposal y given in Equation (4.2.9) to expand I_1. For $x \in X^N$ we have $P^N x = x$. Therefore, for $x \in X^N$,

$$\begin{aligned}
I_1 &= -\frac{\ell \Delta t}{4}\Big(\|(1-\ell\Delta t)x - \ell \Delta t\, C^N \nabla \Psi^N(x) + \sqrt{2\ell\Delta t}\,(C^N)^{\frac{1}{2}}\xi^N\|^2_{C^N} - \|x\|^2_{C^N}\Big) \\
&= Z^N(x,\xi^N) \;+\; i^N(x,\xi^N) \;+\; B_1 \;+\; B_2 \;+\; B_3 \;+\; B_4.
\end{aligned}$$

with $Z^N(x,\xi^N)$ and $i^N(x,\xi^N)$ given by Equation (4.4.3) and (4.4.4) and

$$\begin{aligned}
B_1 &= \frac{\ell^3}{4}\big(1 - \frac{\|x\|^2_{C^N}}{N}\big) \\
B_2 &= -\frac{\ell^3}{4} N^{-1}\Big\{\|C^N \nabla \Psi^N(x)\|^2_{C^N} + 2\langle x, \nabla \Psi^N(x)\rangle\Big\} \\
B_3 &= \frac{\ell^{\frac{5}{2}}}{\sqrt{2}} N^{-\frac{5}{6}} \langle x + C^N \nabla \Psi^N(x), (C^N)^{\frac{1}{2}}\xi^N\rangle_{C^N} \\
B_4 &= \frac{\ell^2}{2} N^{-\frac{2}{3}} \langle x, \nabla \Psi^N(x)\rangle.
\end{aligned}$$

The quantity Z^N is the leading term. For each fixed value of $x \in \mathcal{H}^s$ the term $Z^N(x,\xi^N)$ is Gaussian. Below, we prove that quantity i^N is $\mathcal{O}(N^{-\frac{1}{6}})$. We now establish that each B_j is $\mathcal{O}(N^{-\frac{1}{3}})$,

$$\mathbb{E}^{\pi^N}\left[B_j^{2q}\right]^{\frac{1}{2q}} \;=\; \mathcal{O}(N^{-\frac{1}{3}}) \qquad\qquad j = 1, \ldots, 4. \tag{4.4.9}$$

- Lemma 4.2.1 shows that $\mathbb{E}^{\pi^N}[\big(1 - \frac{\|x\|^2_{C^N}}{N}\big)^{2q}] \lesssim \mathbb{E}^{\pi_0}[\big(1 - \frac{\|x\|^2_{C^N}}{N}\big)^{2q}]$. Under π_0, the random variable $\frac{\|x\|^2_{C^N}}{N}$ is distributed as $\frac{\rho_1^2 + \ldots + \rho_N^2}{N}$ where $\rho_1, \ldots, \rho_N$ are independent and identcally distributed $\mathrm{N}(0,1)$ Gaussian random variables. Consequently, $\mathbb{E}^{\pi^N}[B_1^{2q}]^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{1}{2}})$.

- The term $\|C^N \nabla \Psi^N(x)\|^{2q}_{C^N}$ has already been bounded while proving $\mathbb{E}^{\pi^N}[I_3^{2q}] \lesssim \left(N^{-\frac{1}{3}}\right)^{2q}$. Equation (3.1.9) gives the bound $\|\nabla \Psi^N(x)\|_{-s} \lesssim 1 + \|x\|_s$ and

shows that $\mathbb{E}^{\pi^N}\big[\langle x, \nabla\Psi^N(x)\rangle^{2q}\big]$ is uniformly bounded as a function of N. Therefore $\mathbb{E}^{\pi^N}\big[B_2^{2q}\big]^{\frac{1}{2q}} = \mathcal{O}(N^{-1})$.

- We have $\langle C^N\nabla\Psi^N(x), (C^N)^{\frac{1}{2}}\xi^N\rangle_{C^N} = \langle \nabla\Psi^N(x), (C^N)^{\frac{1}{2}}\xi^N\rangle$ so that

$$\mathbb{E}^{\pi^N}[\langle C^N\nabla\Psi^N(x), (C^N)^{\frac{1}{2}}\xi^N\rangle_{C^N}^{2q}] \lesssim \mathbb{E}^{\pi^N}[\|\nabla\Psi^N(x)\|_{-s}^{2q}\cdot\|(C^N)^{\frac{1}{2}}\xi^N\|_s^{2q}] \lesssim 1.$$

  By Lemma 4.2.1, one can suppose $x \overset{\mathcal{D}}{\sim} \pi_0$ and $\langle x, (C^N)^{\frac{1}{2}}\xi^N\rangle_{C^N} \overset{\mathcal{D}}{\sim} \sum_{j=1}^N \rho_j\xi_j$ where $\rho_1,\dots,\rho_N$ are independent and identically distributed $\mathrm{N}(0,1)$ Gaussian random variables. Consequently $\Big(\mathbb{E}^{\pi^N}\big[\langle x, (C^N)^{\frac{1}{2}}\xi^N\rangle_{C^N}^{2q}\big]\Big)^{\frac{1}{2q}} = \mathcal{O}(N^{\frac{1}{2}})$, which proves that $\mathbb{E}^{\pi^N}\big[B_3^{2q}\big]^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{5}{6}+\frac{1}{2}}) = \mathcal{O}(N^{-\frac{1}{3}})$.

- The bound $\|\nabla\Psi^N(x)\|_{-s} \lesssim 1 + \|x\|_s$ ensures that $\Big(\mathbb{E}^{\pi^N}\big[B_4^{2q}\big]\Big)^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{2}{3}})$.

Define the quantity $\mathbf{e}^N(x,\xi^N) = I_2 + I_3 + B_1 + B_2 + B_3 + B_4$ so that Q^N can also be expressed as

$$Q^N(x,\xi^N) = Z^N(x,\xi^N) + i^N(x,\xi^N) + \mathbf{e}^N(x,\xi^N).$$

Equations (4.4.6) and (4.4.9) show that $\mathbf{e}^N$ satisfies $\Big(\mathbb{E}^{\pi^N}\big[\mathbf{e}^N(x,\xi^N)^{2q}\big]\Big)^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{1}{3}})$. We now prove that i^N is $\mathcal{O}(N^{-\frac{1}{6}})$. By Lemma 4.2.1, $\mathbb{E}^{\pi^N}[i^N(x,\xi^N)^{2q}] \lesssim \mathbb{E}^{\pi_0}[i^N(x,\xi^N)^{2q}]$. If $x \overset{\mathcal{D}}{\sim} \pi_0$ we have

$$i^N(x,\xi^N) = \frac{\ell^2}{2}N^{-\frac{2}{3}}\Big\{\|x\|_{C^N}^2 - \|(C^N)^{\frac{1}{2}}\xi^N\|_{C^N}^2\Big\} = \frac{\ell^2}{2}N^{-\frac{2}{3}}\sum_{j=1}^N(\rho_j^2 - \xi_j^2).$$

where $\rho_1,\dots,\rho_N$ are i.i.d $\mathrm{N}(0,1)$ Gaussian random variables. Since $\mathbb{E}\big[\big\{\sum_{j=1}^N(\rho_j^2 - \xi_j^2)\big\}^{2q}\big] \lesssim N^q$ it follows that $\mathbb{E}^{\pi^N}\big[i^N(x,\xi^N)^{2q}\big]^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{2}{3}+\frac{1}{2}}) = \mathcal{O}(N^{-\frac{1}{6}})$, which ends the proof of Lemma 4.4.3 □

The next lemma quantifies the fact that $Z^N(x,\xi^N)$ is asymptotically independent from the current position x.

Lemma 4.4.4. (Asymptotic independence) *Let $p \geq 1$ be a positive integer and $f : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function. Consider any error terms $\boldsymbol{e}_\star^N(x,\xi)$ satisfying $\lim_{N\to\infty}\mathbb{E}^{\pi^N}\big|\boldsymbol{e}_\star^N(x,\xi^N\big|^p = 0$. Define the functions $\bar{f}^N : \mathbb{R}\to\mathbb{R}$ and the constant $\bar{f}\in\mathbb{R}$ by*

$$\bar{f}^N(x) = \mathbb{E}_x\Big[f\big(Z^N(x,\xi^N) + \boldsymbol{e}_\star^N(x,\xi^N)\big)\Big]$$

and $\bar{f} = \mathbb{E}[f(Z_\ell)]$. *The function* f^N *is highly concentrated around its mean in the sense that*

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\Big[|\bar{f}^N(x) - \bar{f}|^p\Big] = 0.$$

Proof. Define the function $F : \mathbb{R} \times [0;\infty) \to \mathbb{R}$ by $F(\mu,\sigma) = \mathbb{E}\big[f(\rho_{\mu,\sigma})\big]$ with $\rho_{\mu,\sigma} \overset{\mathcal{D}}{\sim} \mathrm{N}(\mu,\sigma^2)$. The function F satisfies $\big|F(\mu_1,\sigma_1) - F(\mu_2,\sigma_2)\big| \lesssim |\mu_2-\mu_1| + |\sigma_2-\sigma_1|$ for any choice $\mu_1,\mu_2 \in \mathbb{R}$ and $\sigma_1,\sigma_2 \geq 0$. Indeed,

$$\begin{aligned}\big|F(\mu_1,\sigma_1) - F(\mu_2,\sigma_2)\big| &= \big|\mathbb{E}\big[f(\mu_1+\sigma_1\,\rho_{0,1}) - f(\mu_2+\sigma_2\,\rho_{0,1})\big]\big| \\ &\leq \mathbb{E}\Big[|\mu_2-\mu_1| + |\sigma_2-\sigma_1|\cdot|\rho_{0,1}|\Big] \lesssim |\mu_2-\mu_1| + |\sigma_2-\sigma_1|.\end{aligned}$$

We have $\mathbb{E}_x[Z^N(x,\xi^N)] = \mathbb{E}[Z_\ell] = -\frac{\ell^3}{4}$ while the variances are given by $\mathrm{Var}\big[Z^N(x,\xi^N)\big] = \frac{\ell^3}{2}\frac{\|x\|^2_{C^N}}{N}\,\mathrm{Var}\big[Z_\ell\big] = \frac{\ell^3}{2}$. Therefore, using Lemma 4.2.1,

$$\begin{aligned}\mathbb{E}^{\pi^N}\Big[\big|\bar{f}^N(x) - \bar{f}\big|^p\Big] &= \mathbb{E}^{\pi^N}\Big[\big|\mathbb{E}_x\big[f\big(Z^N(x,\xi^N) + \mathbf{e}^N_\star(x,\xi^N)\big) - f(Z_\ell)\big]\big|^p\Big] \\ &\lesssim \mathbb{E}^{\pi^N}\Big[\big|\mathbb{E}_x\big[f\big(Z^N(x,\xi^N)\big) - f(Z_\ell)\big]\big|^p\Big] + \mathbb{E}^{\pi^N}\big[|\mathbf{e}^N_\star(x,\xi^N)|^p\big] \\ &= \mathbb{E}^{\pi^N}\Big[\big|F(-\frac{\ell^3}{4},\mathrm{Var}\big[Z^N(x,\xi^N)\big]^{\frac12}) - F(-\frac{\ell^3}{4},\mathrm{Var}\big[Z_\ell\big]^{\frac12})\big|^p\Big] \\ &\quad + \mathbb{E}^{\pi^N}\big[|\mathbf{e}^N_\star(x,\xi^N)|^p\big] \\ &\lesssim \mathbb{E}^{\pi^N}\Big[\big|\mathrm{Var}\big[Z^N(x,\xi^N)\big]^{\frac12} - \mathrm{Var}\big[Z_\ell\big]^{\frac12}\big|^p\Big] + \mathbb{E}^{\pi^N}\big[|\mathbf{e}^N_\star(x,\xi^N)|^p\big] \\ &\lesssim \mathbb{E}^{\pi_0}\Big|\Big\{\frac{\|x\|^2_{C^N}}{N}\Big\}^{\frac12} - 1\Big|^p + \mathbb{E}^{\pi^N}\big[|\mathbf{e}^N_\star(x,\xi^N)|^p\big] \to 0.\end{aligned}$$

In the last step we have used the fact that if $x \overset{\mathcal{D}}{\sim} \pi_0$ then $\frac{\|x\|^2_{C^N}}{N} \overset{\mathcal{D}}{\sim} \frac{\rho_1^2+\ldots+\rho_N^2}{N}$ where $\rho_1,\ldots,\rho_N$ are i.i.d Gaussian random variables $\mathrm{N}(0,1)$ so that $\mathbb{E}^{\pi_0}\Big|\Big\{\frac{\|x\|^2_{C^N}}{N}\Big\}^{\frac12} - 1\Big|^p \to 0$. □

Corollary 4.4.5. *Let* $p \geq 1$ *be a positive. The local mean acceptance probability* $\alpha^N(x)$ *defined in Equation* (4.2.11) *satisfies*

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\big[|\alpha^N(x) - \alpha(\ell)|^p\big] = 0.$$

Proof. The function $f(z) = 1 \wedge e^z$ is 1-Lipschitz and $\alpha(\ell) = \mathbb{E}[f(Z_\ell)]$. Also,

$$\alpha^N(x) = \mathbb{E}_x\Big[f(Q^N(x,\xi^N))\Big] = \mathbb{E}_x\big[f(Z^N(x,\xi^N) + \mathbf{e}^N_\star(x,\xi^N)\Big]$$

with $\mathbf{e}_\star^N(x,\xi^N) = i^N(x,\xi^N)+\mathbf{e}^N(x,\xi^N)$. Lemma 4.4.3 shows that $\lim_{N\to\infty} \mathbb{E}^{\pi^N}[\mathbf{e}_\star^N(x,\xi)^p] = 0$ and therefore lemma 4.4.4 gives the conclusion. □

4.4.3 Drift approximation

This section proves that the approximate drift function $d^N : \mathcal{H}^s \to \mathcal{H}^s$ defined in Equation (4.3.3) converges to the drift function $\mu : \mathcal{H}^s \to \mathcal{H}^s$ of the limiting diffusion (4.2.16).

Lemma 4.4.6. (Drift Approximation) *Let assumptions 3.1.1 hold. The drift function $d^N : \mathcal{H}^s \to \mathcal{H}^s$ converges to μ in the sense that*

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\left[\left\|d^N(x) - \mu(x)\right\|_s^2\right] = 0.$$

Proof. Recall that $\{\varphi_j\}_{j\geq 1}$ is an orthonormal basis of the Hilbert space $\mathcal{H}$. For notational convenience we introduce the quantity $\hat{\varphi}_j = j^{-s}\varphi_j$ so that $\{\hat{\varphi}_j\}_{j\geq 1}$ is an orthonormal basis of $\mathcal{H}^s$. The approximate drift d^N is given by Equation (4.3.3). The definition of the local mean acceptance probability $\alpha^N(x)$ given by Equation (4.2.11) show that d^N can also be expressed as

$$d^N(x) = \left(\alpha^N(x)\alpha(\ell)^{-1}\right)\mu^N(x) + \sqrt{2\ell}h(\ell)^{-1}(\Delta t)^{-\frac{1}{2}}\varepsilon^N(x)$$

where $\mu^N(x) = -\left(P^N x + C^N\nabla\Psi^N(x)\right)$ and the term $\varepsilon^N(x)$ is defined by

$$\varepsilon^N(x) \;=\; \mathbb{E}_x\left[\gamma^N(x,\xi^N)\,\mathcal{C}^{\frac{1}{2}}\xi^N\right] \;=\; \mathbb{E}_x\left[\left(1\wedge e^{Q^N(x,\xi^N)}\right)\mathcal{C}^{\frac{1}{2}}\xi^N\right].$$

To prove Lemma 4.4.6 it suffices to verify that

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\left[\left\|\left(\alpha^N(x)\alpha(\ell)^{-1}\right)\mu^N(x) - \mu(x)\right\|_s^2\right] = 0 \tag{4.4.10}$$

$$\lim_{N\to\infty} (\Delta t)^{-1}\,\mathbb{E}^{\pi^N}\left[\left\|\varepsilon^N(x)\right\|_s^2\right] = 0. \tag{4.4.11}$$

- Let us first prove Equation (4.4.10). The triangle inequality and Cauchy-Schwarz inequality show that

$$\left(\mathbb{E}^{\pi^N}\left[\left\|\left(\alpha^N(x)\alpha(\ell)^{-1}\right)\mu^N(x) - \mu(x)\right\|_s^2\right]\right)^2 \lesssim \mathbb{E}[|\alpha^N(x)-\alpha(\ell)|^4]\cdot\mathbb{E}^{\pi^N}[\|\mu^N(x)\|_s^4] + \mathbb{E}^{\pi^N}[\|\mu^N(x)-\mu(x)\|_s^4].$$

  By Remark 4.2.3 $\mu^N : \mathcal{H}^s \to \mathcal{H}^s$ is Lipschitz, with a Lipschitz constant that can be chosen independent of N. It follows that $\sup_N \mathbb{E}^{\pi^N}[\|\mu^N(x)\|_s^4] < \infty$.

Lemma 4.4.4 and Corollary 4.4.5 show that $\mathbb{E}[|\alpha^N(x)-\alpha(\ell)|^4] \to 0$. Therefore,

$$\lim_{N\to\infty} \mathbb{E}[|\alpha^N(x) - \alpha(\ell)|^4] \cdot \mathbb{E}^{\pi^N}[\|\mu^N(x)\|_s^4] = 0.$$

The functions μ^N and μ are globally Lipschitz on $\mathcal{H}^s$, with a Lipschitz constant that can be chosen independent from N, so that $\|\mu^N(x)-\mu(x)\|_s^4 \lesssim (1+\|x\|_s^4)$. Lemma 4.4.1 proves that the sequence of functions $\{\mu^N\}$ converges π_0-almost surely to $\mu(x)$ in $\mathcal{H}^s$ and lemma 4.2.1 show that $\mathbb{E}^{\pi^N}[\|\mu^N(x) - \mu(x)\|_s^4] \lesssim \mathbb{E}^{\pi_0}[\|\mu^N(x)-\mu(x)\|_s^4]$. It thus follows from the dominated convergence theorem that

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}[\|\mu^N(x) - \mu(x)\|_s^4] = 0.$$

This concludes the proof of the Equation (4.4.10).

- Let us prove Equation (4.4.11). If the Bernoulli random variable $\gamma^N(x,\xi^N)$ were independent from the noise term $(C^N)^{\frac{1}{2}}\xi^N$, it would follow that $\varepsilon^N(x) = 0$. In general $\gamma^N(x,\xi^N)$ is not independent from $(C^N)^{\frac{1}{2}}\xi^N$ so that $\varepsilon^N(x)$ is not equal to zero. Nevertheless, as quantified by Lemma 4.4.4, the Bernoulli random variable $\gamma^N(x,\xi^N)$ is asymptotically independent from the current position x and from the noise term $(C^N)^{\frac{1}{2}}\xi^N$. Consequently, we can prove in Equation (4.4.13) that the quantity $\varepsilon^N(x)$ is small. To this end, we establish that each component $\langle\varepsilon(x),\hat{\varphi}_j\rangle_s^2$ satisfies

$$\mathbb{E}^{\pi^N}\left[\langle\varepsilon^N(x),\hat{\varphi}_j\rangle_s^2\right] \quad\lesssim\quad N^{-1}\mathbb{E}^{\pi^N}[\langle x,\hat{\varphi}_j\rangle_s^2] + N^{-\frac{2}{3}}(j^s\lambda_j)^2. \quad (4.4.12)$$

Summation of Equation (4.4.12) over $j = 1,\dots,N$ leads to

$$\begin{aligned}\mathbb{E}^{\pi^N}\left[\|\varepsilon^N(x)\|_s^2\right] &\lesssim N^{-1}\mathbb{E}^{\pi^N}\left[\|x\|_s^2\right] + N^{-\frac{2}{3}}\operatorname*{Tr}_{\mathcal{H}^s}(C_s) \qquad (4.4.13)\\ &\lesssim N^{-\frac{2}{3}},\end{aligned}$$

which gives the proof of Equation (4.4.11). To prove Equation (4.4.12) for a fixed index $j \in \mathbb{N}$, the quantity $Q^N(x,\xi)$ is decomposed as a sum of a term independent from ξ_j and another remaining term of small magnitude. To this end we introduce

$$\begin{cases} Q^N(x,\xi^N) &= Q_j^N(x,\xi^N) + Q_{j,\perp}^N(x,\xi^N)\\ Q_j^N(x,\xi^N) &= -\frac{1}{\sqrt{2}}\ell^{\frac{3}{2}}N^{-\frac{1}{2}}\lambda_j^{-1}x_j\xi_j - \frac{1}{2}\ell^2N^{-\frac{2}{3}}\lambda_j^2\xi_j^2 + \mathbf{e}^N(x,\xi^N).\end{cases} \quad (4.4.14)$$

The definitions of $Z^N(x,\xi^N)$ and $i^N(x,\xi^N)$ in Equation (4.4.3) and (4.4.4) readily show that $Q^N_{j,\perp}(x,\xi^N)$ is independent from ξ_j. The noise term satisfies $\mathcal{C}^{\frac{1}{2}}\xi^N = \sum_{j=1}^N (j^s\lambda_j)\xi_j\hat{\varphi}_j$. Since $Q^N_{j,\perp}(x,\xi^N)$ and ξ_j are independent and $z \mapsto 1 \wedge e^z$ is 1-Lipschitz, it follows that

$$\begin{aligned}
\langle \varepsilon^N(x), \hat{\varphi}_j\rangle_s^2 &= (j^s\lambda_j)^2 \left(\mathbb{E}_x\big[\big(1 \wedge e^{Q^N(x,\xi^N)}\big)\,\xi_j\big]\right)^2 \\
&= (j^s\lambda_j)^2 \left(\mathbb{E}_x\big[\big[\big(1 \wedge e^{Q^N(x,\xi^N)}\big) - \big(1 \wedge e^{Q^N_{j,\perp}(x,\xi^N)}\big)\big]\,\xi_j\big]\right)^2 \\
&\lesssim (j^s\lambda_j)^2 \mathbb{E}_x\big[|Q^N(x,\xi^N)) - Q^N_{j,\perp}(x,\xi^N)|^2\big] = (j^s\lambda_j)^2\mathbb{E}_x\big[Q^N_j(x,\xi^N)^2\big].
\end{aligned}$$

By Lemma 4.4.3 $\mathbb{E}^{\pi^N}\big[\mathbf{e}^N(x,\xi^N)^2\big] \lesssim N^{-\frac{2}{3}}$. Therefore,

$$\begin{aligned}
(j^s\lambda_j)^2\mathbb{E}^{\pi^N}\big[Q^N_j(x,\xi^N)^2\big] &\lesssim (j^s\lambda_j)^2\Big\{N^{-1}\lambda_j^{-2}\mathbb{E}^{\pi^N}\big[x_j^2\xi_j^2\big] \\
&\qquad + N^{-\frac{4}{3}}\mathbb{E}^{\pi^N}\big[\lambda_j^4\xi_j^4\big] + \mathbb{E}^{\pi^N}\big[\mathbf{e}^N(x,\xi)^2\big]\Big\} \\
&\lesssim N^{-1}\,\mathbb{E}^{\pi^N}\big[(j^s x_j)^2\xi_j^2\big] + (j^s\lambda_j)^2(N^{-\frac{4}{3}} + N^{-\frac{2}{3}}) \\
&\lesssim N^{-1}\,\mathbb{E}^{\pi^N}\big[\langle x,\hat{\varphi}_j\rangle_s^2\big] + (j^s\lambda_j)^2 N^{-\frac{2}{3}},
\end{aligned}$$

which finishes the proof of Equation (4.4.12).

□

4.4.4 Noise approximation

We remind the reader that the family $\{\hat{\varphi}_j := j^{-s}\varphi_j\}_{j\geq 1}$ is an orthonormal basis of $\mathcal{H}^s$ while $\{\varphi_j\}_{j\geq 1}$ is an orthonormal basis of $\mathcal{H}$. Recall the definition (4.3.4) of the martingale difference $\Gamma^{k,N}$. In this section we estimate the error in the approximation $\Gamma^{k,N} \approx \mathrm{N}(0, C_s)$ where C_s has been defined in section 3.1.2 as the covariance of π_0 when seen as a Gaussian measure on $\mathcal{H}^s$. To this end we introduce the covariance operator

$$D^N(x) = \mathbb{E}_x\Big[\Gamma^{k,N} \otimes_{\mathcal{H}^s} \Gamma^{k,N}\ |x^{k,N} = x\Big].$$

For any $x,u,v \in \mathcal{H}^s$ the operator $D^N(x)$ satisfies $\mathbb{E}\Big[\langle\Gamma^{k,N},u\rangle_s\langle\Gamma^{k,N},v\rangle_s\ |x^{k,N} = x\Big] = \langle u, D^N(x)v\rangle_s$. The next lemma gives a quantitative version of the approximation $D^N(x) \approx C_s$.

Lemma 4.4.7. *Let assumptions 3.1.1 hold. For any pair of indices $i,j \geq 0$ the*

operator $D^N(x) : \mathcal{H}^s \to \mathcal{H}^s$ *satisfies*

$$\lim_{N\to\infty} \quad \mathbb{E}^{\pi^N}\big|\langle\hat{\varphi}_i, D^N(x)\hat{\varphi}_j\rangle_s - \langle\hat{\varphi}_i, C_s\hat{\varphi}_j\rangle_s\big| \;=\; 0 \tag{4.4.15}$$

and, furthermore,

$$\lim_{N\to\infty} \quad \mathbb{E}^{\pi^N}\big|\operatorname*{Tr}_{\mathcal{H}^s}(D^N(x)) - \operatorname*{Tr}_{\mathcal{H}^s}(C_s)\big| \;=\; 0. \tag{4.4.16}$$

Proof. The martingale difference $\Gamma^N(x,\xi)$ is given by

$$\begin{aligned}\Gamma^N(x,\xi) =& \alpha(\ell)^{-\frac{1}{2}}\gamma^N(x,\xi)\,\mathcal{C}^{\frac{1}{2}}\xi \\ &+ \frac{1}{\sqrt{2}}\alpha(\ell)^{-\frac{1}{2}}(\ell\Delta t)^{\frac{1}{2}}\Big\{\gamma^N(x,\xi)\mu^N(x) - \alpha(\ell)d^N(x)\Big\}.\end{aligned} \tag{4.4.17}$$

We only prove Equation (4.4.16); the proof of Equation (4.4.15) is essentially identical but easier. Remark 4.2.3 shows that the functions $\mu, \mu^N : \mathcal{H}^s \to \mathcal{H}^s$ are globally Lipschitz and Lemma 4.4.6 shows that $\mathbb{E}^{\pi^N}\big[\|d^N(x) - \mu(x)\|_s^2\big] \to 0$. Therefore

$$\mathbb{E}^{\pi^N}\Big[\|\gamma^N(x,\xi)\mu^N(x) - \alpha(\ell)d^N(x)\|_s^2\Big] \;\lesssim\; 1, \tag{4.4.18}$$

which implies that the second term on the right-hand-side of Equation (4.4.17) is $\mathcal{O}\big(\sqrt{\Delta t}\big)$. Since $\operatorname{Tr}_{\mathcal{H}^s}(D^N(x)) = \mathbb{E}_x\big[\|\Gamma^N(x,\xi)\|_s^2\big]$, Equation (4.4.18) implies that

$$\mathbb{E}^{\pi^N}\Big[\big|\alpha(\ell)\operatorname*{Tr}_{\mathcal{H}^s}(D^N(x)) - \mathbb{E}_x\big[\|\gamma^N(x,\xi)\,\mathcal{C}^{\frac{1}{2}}\xi\|_s^2\big]\big|\Big] \lesssim (\Delta t)^{\frac{1}{2}}.$$

Consequently, to prove Equation (4.4.16) it suffices to verify that

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\Big[\big|\mathbb{E}_x\big[\|\gamma^N(x,\xi)\,\mathcal{C}^{\frac{1}{2}}\xi\|_s^2\big] - \alpha(\ell)\operatorname*{Tr}_{\mathcal{H}^s}(C_s)\big|\Big] \;=\; 0. \tag{4.4.19}$$

We have $\mathbb{E}_x\big[\|\gamma^N(x,\xi)\,\mathcal{C}^{\frac{1}{2}}\xi\|_s^2\big] \;=\; \sum_{j=1}^N (j^s\lambda_j)^2\mathbb{E}_x\big[\big(1\wedge e^{Q^N(x,\xi)}\big)\xi_j^2\big]$. Therefore, to prove Equation (4.4.19) it suffices to establish

$$\lim_{N\to\infty} \sum_{j=1}^N (j^s\lambda_j)^2\mathbb{E}^{\pi^N}\Big[\big|\mathbb{E}_x\big[\big(1\wedge e^{Q^N(x,\xi)}\big)\xi_j^2\big] - \alpha(\ell)\big|\Big] = 0. \tag{4.4.20}$$

Since $\sum_{j=1}^\infty (j^s\lambda_j)^2 < \infty$ and $\big|1\wedge e^{Q^N(x,\xi)}\big| \le 1$, the dominated convergence theorem shows that (4.4.20) follows from

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\Big[\big|\mathbb{E}_x\big[\big(1\wedge e^{Q^N(x,\xi)}\big)\xi_j^2\big] - \alpha(\ell)\big|\Big] = 0 \qquad \forall j \ge 0. \tag{4.4.21}$$

We now prove Equation (4.4.21). As in the proof of Lemma 4.4.6, we use the decomposition $Q^N(x,\xi) = Q^N_j(x,\xi) + Q^N_{j,\perp}(x,\xi)$ where $Q^N_{j,\perp}(x,\xi)$ is independent from ξ_j. Therefore, since $\mathrm{Lip}(f) = 1$,

$$\begin{aligned}\mathbb{E}_x\big[\big(1\wedge e^{Q^N(x,\xi)}\big)\xi_j^2\big] &= \mathbb{E}_x\big[\big(1\wedge e^{Q^N_{j,\perp}(x,\xi)}\big)\xi_j^2\big] + \mathbb{E}_x\big[\big[\big(1\wedge e^{Q^N(x,\xi)}\big) - \big(1\wedge e^{Q^N_{j,\perp}(x,\xi)}\big)\big]\xi_j^2\big]\\ &= \mathbb{E}_x\big[1\wedge e^{Q^N_{j,\perp}(x,\xi)}\big] + \mathcal{O}\Big(\big\{\mathbb{E}_x\big[|Q^N(x,\xi)) - Q^N_{j,\perp}(x,\xi)|^2\big]\big\}^{\frac{1}{2}}\Big)\\ &= \mathbb{E}_x\big[1\wedge e^{Q^N_{j,\perp}(x,\xi)}\big] + \mathcal{O}\Big(\big\{\mathbb{E}_x\big[Q^N_j(x,\xi)^2\big]\big\}^{\frac{1}{2}}\Big).\end{aligned}$$

Lemma 4.4.4 ensures that, for $f(\cdot) = 1\wedge\exp(\cdot)$,

$$\lim_{N\to\infty}\ \mathbb{E}^{\pi^N}\Big[\big|\mathbb{E}_x\big[f(Q^N_{j,\perp}(x,\xi))\big] - \alpha(\ell)\big|\Big] = 0$$

and the definition of $Q^N_i(x,\xi)$ readily shows that $\lim_{N\to\infty}\ \mathbb{E}^{\pi^N}\big[Q^N_j(x,\xi)^2\big] = 0$. This concludes the proof of Equation (4.4.21) and thus ends the proof of Lemma 4.4.7. □

Corollary 4.4.8. *More generally, for any fixed vector $h\in\mathcal{H}^s$, the following limit holds,*

$$\lim_{N\to\infty}\quad \mathbb{E}^{\pi^N}\big|\langle h, D^N(x)h\rangle_s - \langle h, C_s h\rangle_s\big| \;=\; 0. \tag{4.4.22}$$

Proof. If $h = \hat\varphi_i$, this is precisely the content of Lemma 4.4.7. More generally, by linearity, Lemma 4.4.7 shows that this is true for $h = \sum_{i\le N}\alpha_i\hat\varphi_i$, where $N\in\mathbb{N}$ is a fixed integer. For a general vector $h\in\mathcal{H}^s$, we can use the decomposition $h = h^* + e^*$ where $h^* = \sum_{j\le N}\langle h,\hat\varphi_j\rangle_s\,\hat\varphi_j$ and $e^* = h - h^*$. It follows that

$$\begin{aligned}&\Big|\Big(\langle h, D^N(x)\ h\rangle_s - \langle h, C_s h\rangle_s\Big) - \Big(\langle h^*, D^N(x)\ h^*\rangle_s - \langle h^*, C_s h^*\rangle_s\Big)\Big|\\ &\quad\le \Big|\langle h+h^*, D^N(x)\ (h-h^*)\rangle_s - \langle h+h^*, C_s\ (h-h^*)\rangle_s\Big|\\ &\quad\le 2\|h\|_s\cdot\|h-h^*\|_s\cdot\Big(\operatorname*{Tr}_{\mathcal{H}^s}(D^N(x)) + \operatorname*{Tr}_{\mathcal{H}^s}(C_s)\Big),\end{aligned}$$

where we have used the fact that for an non-negative self-adjoint operator $D:\mathcal{H}^s\to\mathcal{H}^s$ we have $\langle u, Dv\rangle_s \le \|u\|_s\cdot\|v\|_s\cdot\mathrm{Tr}_{\mathcal{H}^s}(D)$. lemma 4.4.7 implies that $\mathbb{E}^{\pi^N}[\mathrm{Tr}_{\mathcal{H}^s}(D^N(x))] < \infty$ and assumption 3.1.1 ensures that $\mathrm{Tr}_{\mathcal{H}^s}(C_s) < \infty$. Consequently, $\lim_{N\to\infty}\mathbb{E}^{\pi^N}\big|\langle h, D^N(x)\ h\rangle - \langle h, C_s h\rangle\big|$ is less than a constant multiple of $\lim_{N\to\infty}\mathbb{E}^{\pi^N}\big|\langle h^*, D^N(x)\ h^*\rangle - \langle h^*, C_s\ h^*\rangle\big| \;+\; \|h-h^*\|_s$, which equals $\|h-h^*\|_s$.

Therefore, we have

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N} \left| \langle h, D^N(x)\, h\rangle - \langle h, C_s h\rangle \right| \lesssim \|h - h^*\|_s.$$

Because $\|h - h^*\|_s$ can be chosen arbitrarily small, the conclusion follows. □

4.4.5 Martingale invariance principle

This section proves that the process W^N defined in Equation (4.3.6) converges to a Brownian motion.

Proposition 4.4.9. *Let assumptions 3.1.1 hold. Let $z^0 \sim \pi$ and $W^N(t)$ the process defined in equation (4.3.6) and $x^{0,N} \overset{\mathcal{D}}{\sim} \pi^N$ the starting position of the Markov chain x^N. Then*

$$(x^{0,N}, W^N) \Longrightarrow (z^0, W), \tag{4.4.23}$$

where $\Longrightarrow$ denotes weak convergence in $\mathcal{H}^s \times C([0,T];\mathcal{H}^s)$, and W is a $\mathcal{H}^s$-valued Brownian motion with covariance operator C_s. Furthermore the limiting Brownian motion W is independent of the initial condition z^0.

Proof. Remember that we have defined the quantity $\hat{\varphi}_j = j^{-s}\varphi_j$ so that $\{\hat{\varphi}_j\}_{j\geq 1}$ is an orthonormal basis of $\mathcal{H}^s$. As a first step, we show that W^N converges weakly to W. As described in [MPS11], a consequence of proposition 5.1 of [Ber86] shows that in order to prove that W^N converges weakly to W in $C([0,T];\mathcal{H}^s)$ it suffices to prove that for any $t \in [0,T]$ and any pair of indices $i, j \geq 0$ the following three limits hold in probability, the third for any $\varepsilon > 0$,

$$\lim_{N\to\infty} \Delta t \sum_{k=1}^{k_N(T)} \mathbb{E}\Big[\|\Gamma^{k,N}\|_s^2 \,|\mathcal{F}^{k,N}\Big] = T \operatorname*{Tr}_{\mathcal{H}^s}(C_s) \tag{4.4.24}$$

$$\lim_{N\to\infty} \Delta t \sum_{k=1}^{k_N(t)} \mathbb{E}\Big[\langle \Gamma^{k,N}, \hat{\varphi}_i\rangle_s \langle \Gamma^{k,N}, \hat{\varphi}_j\rangle_s \,|\mathcal{F}^{k,N}\Big] = t\, \langle \hat{\varphi}_i, C_s \hat{\varphi}_j\rangle_s \tag{4.4.25}$$

$$\lim_{N\to\infty} \Delta t \sum_{k=1}^{k_N(T)} \mathbb{E}\Big[\|\Gamma^{k,N}\|_s^2 1_{\{\|\Gamma^{k,N}\|_s^2 \geq \Delta t\, \varepsilon\}} \,|\mathcal{F}^{k,N}\Big] = 0 \tag{4.4.26}$$

where $k_N(t) = \lfloor \frac{t}{\Delta t} \rfloor$, $\{\hat{\varphi}_j\}$ is an orthonormal basis of $\mathcal{H}^s$ and $\mathcal{F}^{k,N}$ is the natural filtration of the Markov chain $\{x^{k,N}\}$. The proof follows from the estimate on $D^N(x) = \mathbb{E}\Big[\Gamma^{0,N} \otimes \Gamma^{0,N} \,|x^{0,N} = x\Big]$ presented in Lemma 4.4.7 For the sake of simplicity, we will write $\mathbb{E}_k[\,\cdot\,]$ instead of $\mathbb{E}[\,\cdot\,|\mathcal{F}^{k,N}]$. We now prove that the three conditions are satisfied.

- **Condition** (4.4.24)
  It is enough to prove that $\lim \mathbb{E}\Big|\Big\{\frac{1}{\lfloor N^{\frac{1}{3}}\rfloor}\sum_{k=1}^{\lfloor N^{\frac{1}{3}}\rfloor}\mathbb{E}_k\big[\|\Gamma^{k,N}\|_s^2\big]\Big\} - \mathrm{Tr}_{\mathcal{H}^s}(C_s)\Big| = 0$ where

$$\mathbb{E}_k\Big[\|\Gamma^{k,N}\|_s^2\Big] \;=\; \mathbb{E}_k\sum_{j=1}^{N}\Big[\langle\hat{\varphi}_j, D^N(x^{k,N})\hat{\varphi}_j\rangle_s\Big] \;=\; \mathbb{E}_k \operatorname*{Tr}_{\mathcal{H}^s}(D^N(x^{k,N})).$$

  Because the Metropolis-Hastings algorithm preserves stationarity and $x^{0,N} \overset{\mathcal{D}}{\sim} \pi^N$ it follows that $x^{k,N} \overset{\mathcal{D}}{\sim} \pi^N$ for any $k \geq 0$. Therefore, for all $k \geq 0$ we have $\mathrm{Tr}_{\mathcal{H}^s}(D^N(x^{k,N})) \overset{\mathcal{D}}{\sim} \mathrm{Tr}_{\mathcal{H}^s}(D^N(x))$ where $x \overset{\mathcal{D}}{\sim} \pi^N$. Consequently, the triangle inequality shows that

$$\mathbb{E}\Big|\Big\{\frac{1}{\lfloor N^{\frac{1}{3}}\rfloor}\sum_{k=1}^{\lfloor N^{\frac{1}{3}}\rfloor}\mathbb{E}_k\|\Gamma^{k,N}\|^2\Big\} - \operatorname*{Tr}_{\mathcal{H}^s}(C_s)\Big| \quad\leq\quad \mathbb{E}^{\pi^N}\Big|\operatorname*{Tr}_{\mathcal{H}^s}\big(D^N(x)\big) - \operatorname*{Tr}_{\mathcal{H}^s}(C_s)\Big| \to 0$$

  where the last limit follows from Lemma 4.4.7.

- **Condition** (4.4.25)
  It is enough to prove that

$$\lim\; \mathbb{E}^{\pi^N}\Big|\Big\{\frac{1}{\lfloor N^{\frac{1}{3}}\rfloor}\sum_{k=1}^{\lfloor N^{\frac{1}{3}}\rfloor}\mathbb{E}_k\Big[\langle\Gamma^{k,N},\hat{\varphi}_i\rangle_s\langle\Gamma^{k,N},\hat{\varphi}_j\rangle_s\Big]\Big\} - \langle\hat{\varphi}_i, C_s\hat{\varphi}_j\rangle_s\Big| = 0$$

  where $\mathbb{E}_k\Big[\langle\Gamma^{k,N},\hat{\varphi}_i\rangle_s\langle\Gamma^{k,N},\hat{\varphi}_j\rangle_s\Big] = \langle\hat{\varphi}_i, D^N(x^{k,N})\hat{\varphi}_j\rangle_s$. Because $x^{k,N} \overset{\mathcal{D}}{\sim} \pi^N$ the conclusion again follows from Lemma 4.4.7.

- **Condition** (4.4.26)
  For all $k \geq 1$ we have $x^{k,N} \overset{\mathcal{D}}{\sim} \pi^N$ so that

$$\mathbb{E}^{\pi^N}\Big|\frac{1}{\lfloor N^{\frac{1}{3}}\rfloor}\sum_{k=1}^{\lfloor N^{\frac{1}{3}}\rfloor}\mathbb{E}_k[\|\Gamma^{k,N}\|_s^2 1_{\|\Gamma^{k,N}\|_s^2 \geq N^{\frac{1}{3}}\varepsilon}]\Big| \;\leq\; \mathbb{E}^{\pi^N}\|\Gamma^{0,N}\|_s^2 1_{\{\|\Gamma^{0,N}\|_s^2 \geq N^{\frac{1}{3}}\varepsilon\}}.$$

  Equation (4.4.17) shows that for any power $p \geq 0$ we have $\sup_N \mathbb{E}^{\pi^N}\big[\|\Gamma^{0,N}\|_s^p\big] < \infty$. Therefore the sequence $\{\|\Gamma^{0,N}\|_s^2\}$ is uniformly integrable, which shows that

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\|\Gamma^{0,N}\|_s^2 1_{\{\|\Gamma^{0,N}\|_s^2 \geq N^{\frac{1}{3}}\varepsilon\}} = 0.$$

The three hypothesis are satisfied, proving that W^N converges weakly in $C([0,T];\mathcal{H}^s)$ to a Brownian motion W in $\mathcal{H}^s$ with covariance C_s. Therefore, Corollary 4.4 of [MPS11] shows that the sequence $\left\{(x^{0,N}, W^N)\right\}_{N\geq 1}$ converges weakly to (z^0, W) in $\mathcal{H}\times C([0,T],\mathcal{H}^s)$. This finishes the proof of Proposition 4.4.9. □

4.5 Conclusion

We have studied the application of the MALA algorithm to sample from measures defined via density with respect to a Gaussian measure on Hilbert space. We prove that a suitably interpolated and scaled version of the Markov chain has a diffusion limit in infinite dimensions. There are two main conclusions which follow from this theory: firstly this work shows that, in stationarity, the MALA algorithm applied to an $N-$dimensional approximation of the target will take $\mathcal{O}(N^{\frac{1}{3}})$ steps to explore the invariant measure; secondly the MALA algorithm will be optimized at an average acceptance probability of 0.574. We have thus significantly extended the work [RR98] which reaches similar conclusions in the case of i.i.d. product targets. In contrast we have considered target measures with significant correlation, with structure motivated by a range of applications. As a consequence our limit theorems are in an infinite dimensional Hilbert space and we have employed an approach to the derivation of the diffusion limit which differs significantly from that used in [RR98]. This approach was developed in [MPS11] to study diffusion limits for the RWM algorithm.

There are many possible developments of this work. We list several of these.

- In [BPR$^+$13] it is shown that the Hybrid Monte Carlo algorithm (HMC) requires, for target measures of the form (4.1.1), $\mathcal{O}(N^{\frac{1}{4}})$ steps to explore the invariant measure. However there is no diffusion limit in this case. Identifying an appropriate limit, and extending analysis to the case of target measures (4.2.3) provides a challenging avenue for exploration.

- In the i.i.d product case it is known that, if the Markov chain is started "far" from stationarity, a fluid limit (ODE) is observed [CRR05]. It would be interesting to study such limits in the present context.

- Combining the analysis of MCMC methods for hierarchical target measures [Béd09] with the analysis herein provides a challenging set of theoretical questions, as well as having direct applicability.

- It should also be noted that, for measures absolutely continuous with respect to a Gaussian, there exist new non-standard versions of RWM [BS09], MALA

[BRSV08] and HMC [BPSSS11] for which the acceptance probability does not degenerate to zero as dimension N increases. These methods may be expensive to implement when the Karhunen-Loève basis is not known explicitly, and comparing their overall efficiency with that of standard RWM, MALA and HMC is an interesting area for further study.

- It is natural to ask whether analysis similar to that undertaken here could be developed for Metropolis-Hastings methods applied to other reference measures with a non-Gaussian product structure. In particular the Besov priors of [LSS09] provide an interesting class of such reference measures, and the paper [DHS12] provides a machinery for analyzing change of measure from the Besov prior, analogous to that used here in the Gaussian case. Another intertesting class of reference measures are those used in the study of uncertainty quantification for elliptic PDEs: these have the form of an infinite product of compactly supported uniform distributions; see [SS12].

Chapter 5

Gradient flow without gradient

This chapter is joint work with Andrew Stuart and Natesh Pillai and is based on the paper [PST].

5.1 Introduction

There are many applications where it is of interest to find global or local minima of a functional

$$J(x) = \frac{1}{2}\|C^{-1/2}x\|^2 + \Psi(x) \tag{5.1.1}$$

where C is a self-adjoint, positive and trace-class linear operator on a Hilbert space $\mathcal{H}$. The functional $J : \mathcal{H} \to \mathbb{R}$ has been written under the form (5.1.1) in order to emphasise that it can be seen as a perturbation of the quadratic potential $x \mapsto \frac{1}{2}\|C^{-1/2}x\|^2$. This remark is especially important in infinite dimensional settings where the operator C will play the role of the covariance operator of a Gaussian measure. Gaussian measures in infinite dimensional spaces are the natural analogue of the Lebesgue measures in finite dimensional settings. Gradient flow or steepest descent is a natural approach to this problem, but in its basic form requires computation of the gradient of Ψ which, in some applications, may be an expensive or a complex task. In addition, when multiple minima are present, it may be important to include noise within the algorithm in order to allow escape from local minima. The purpose of this chapter is to show how a noisy gradient descent can emerge from certain carefully specified random walks, when combined with a Metropolis-Hastings accept-reject mechanism, with tunable noise level τ. Furthermore, the algorithms that we study are Markov chain-Monte Carlo methods which are reversible and invariant with respect to a probability measure π^τ (for which probability maximizers occur where J is minimized) and are hence of interest in their own right; the noisy

gradient descent provides a way to analyze the efficiency of the resulting algorithms.

In the finite state [KGV83; Čer85] or finite dimensional context [Gem85; GH86; HKS89] the idea of using random walks, with accept-reject, to perform global optimization is a well-known idea which goes by the name of simulated-annealing; see the review [BT93] for further references. The novelty of our work is that the theory is developed on an infinite dimensional Hilbert space, leading to an algorithm which is robust to finite dimensional approximation: we adopt the "optimize then discretize" viewpoint (see [HPUU08], Chapter 3). We emphasize that discretizing, and then applying standard simulated annealing techniques in $\mathbb{R}^N$ to optimize, can lead to algorithms which degenerate as N increases. The diffusion limit proved in [MPS11] provides a concrete example of this phenomenon for the standard random walk approach to sampling the measure π^τ. The work in this chapter shows that small changes in the standard random walk algorithm can result in large efficiency gains when sampling the measure π^τ, and relatedly when minimizing J via simulated annealing.

The algorithms we construct have two basic building blocks: drawing samples from the centred Gaussian measure $N(0, C)$ and evaluating Ψ. By judiciously combining these ingredients we generate (approximately) a noisy gradient flow for J with tunable temperature parameter controlling the size of the noise. In finite dimensions the basic idea behind simulated annealing is built from Metropolis-Hastings methods which have an invariant measure with Lebesgue density proportional to $\exp\bigl(-J(x)/\tau\bigr)$. By adapting the temperature $\tau \in (0, \infty)$ according to an appropriate cooling schedule it is possible to locate global minima of J. The essential challenge in transfering this idea to infinite dimensions is that there is no Lebesgue measure. This issue can be circumvented by working with measures defined via their density with respect to a Gaussian measure, and for us the natural Gaussian measure on $\mathcal{H}$ is

$$\pi_0^\tau = \mathrm{N}(0, \tau\, C). \tag{5.1.2}$$

The quadratic form $\|x\|_C^2 := \|C^{-\frac{1}{2}}x\|^2$ is the Cameron-Martin norm corresponding to the Gaussian measure π_0^τ. Given π_0^τ we may then define the (in general non-Gaussian) measure π^τ via its Radon-Nikodym derivative with respect to π^τ:

$$\frac{d\pi^\tau}{d\pi_0^\tau}(x) \propto \exp\Bigl(-\frac{\Psi(x)}{\tau}\Bigr).$$

Note that if $\mathcal{H}$ is finite dimensional then π^τ has Lebesgue density proportional to

$\exp\big(-J(x)/\tau\big)$.

Our basic strategy will be to construct a Markov chain which is π^τ invariant and to show that a piecewise linear interpolant of the Markov chain converges weakly (in the sense of probability measures) to the desired noisy gradient flow in an appropriate parameter limit. To motivate the Markov chain we first observe that the Ornstein-Uhlenbeck diffusion in $\mathcal{H}$ given by

$$\begin{cases} dz & = -z\,dt + \sqrt{2\tau}dW \\ z(0) & = x \end{cases} \tag{5.1.3}$$

where W is a Brownian motion in $\mathcal{H}$ with covariance operator equal to C, is reversible and ergodic with respect to π_0^τ given by (5.1.2) [DPZ96]. If $t > 0$ then the exact solution of this equation is given by

$$\begin{aligned} z(t) &= e^{-t}x + \sqrt{\Big(\tau(1-e^{-2t})\Big)}\xi \\ &= \big(1-2\delta\big)^{\frac{1}{2}}x + \sqrt{2\delta\tau}\xi. \end{aligned} \tag{5.1.4}$$

where ξ is a Gaussian random variable drawn from $N(0,C)$ and $\delta = \frac{1}{2}(1-e^{-2t})$. Given a current state x of our Markov chain we will propose to move to $z(t)$ given by this formula, for some choice of $t > 0$, or equivalently $\delta \in (0, \frac{1}{2})$. Notice that if $\Psi = 0$, $\pi^\tau = \pi_0^\tau$, and therefore the auto-regressive (AR(1)) process (5.1.4) converges to the Gaussian invariant measure π_0^τ. For a nontrivial functional Ψ such that π^τ is absolutely continuous with respect to π_0^τ, one needs an "accept-reject" mechanism to adjust for the change of measure and converge to the invariant measure π^τ. The "proposed move" $x \mapsto y := \big(1-2\delta\big)^{\frac{1}{2}}x + \sqrt{2\delta\tau}\xi$ given by equation (5.1.4) will be accepted or rejected with probability found from pointwise evaluation of Ψ given by,

$$\alpha^\delta(x,\xi) = 1 \wedge \exp\Big(-\frac{1}{\tau}\big(\Psi(y) - \Psi(x)\big)\Big). \tag{5.1.5}$$

(see Section 5.2 for more details) resulting in a Markov chain $\{x^{k,\delta}\}_{k\in\mathbb{Z}^+}$. This Markov chain corresponds to the preconditioned random walk method P-RWM introduced in [BRSV08], one of a family of Metropolis-Hastings methods defined on the Hilbert space $\mathcal{H}$ and reversible and invariant with respect to π^τ. See also section 3.4 for motivations behind this choice of proposals.

From the output of the P-RWM Metropolis-Hastings method we construct a

continuous interpolant of the Markov chain defined by

$$z^\delta(t) = \frac{1}{\delta}(t - t_k)\, x^{k+1,\delta} + \frac{1}{\delta}(t_{k+1} - t)\, x^{k,\delta} \tag{5.1.6}$$

for $t_k \le t < t_{k+1}$ with $t_k \stackrel{\text{def}}{=} k\delta$. In other words, the process z^δ is a continuous and accelerated (by a factor $1/\delta$) version of the Markov chain x^δ. The main result of the chapter is that as $\delta \to 0$ the Hilbert-space valued process z^δ converges weakly to z solving the Hilbert space valued SDE

$$dz = -\Big(z + C\nabla\Psi(z)\Big)\, dt + \sqrt{2\tau} dW. \tag{5.1.7}$$

This diffusion is reversible, ergodic and satisfies a law of large numbers with respect to the measure π^τ [DPZ92; HSVW05; HSV07]. Since small ball probabilities under π^τ are maximized when centred at minimizers of J, the result thus shows that the algorithm will generate sequences which concentrate near minimizers of J. Varying τ according to a cooling schedule then results in a simulated annealing method on Hilbert space. Weak convergence results for the approximation of stochastic equations in infinite dimensions may be found in the numerical analysis literature. For the heat equation and variants see [Sha03; DP09; GKL09; KLL12], for dispersive and nondispersive wave problems see [Hau10; dBD06] and for delay equations see [BS05; BKMS08]. These papers rely on use of the Kolmogorov equation to establish weak convergence and do not typically deliver convergence on pathspace, but rather convergence of functionals at a given fixed time. In contrast our approach proves weak convergence on pathspace, and does not use the Kolmogorov equation; rather we use the machinery developed in section 3.3 that is based on an invariance principle for Brownian motion in Hilbert space [Ber86], coupled with the preservation of weak convergence under continuous mappings. However, our approach does not deliver rates of weak convergence.

Let us give a heuristic to see why the gradient flow emerges through the pointwise computation of Ψ and the accept-reject mechanism. Note that for $\delta \ll 1$ we have $-\frac{1}{\tau}\big(\Psi(y) - \Psi(x)\big) \approx -\sqrt{\frac{2\delta}{\tau}}\langle\Psi(x), \xi\rangle$ so that we see from (5.1.5) that the acceptance probability can be approximated by

$$\alpha^\delta(x, \xi) \approx 1 \wedge \exp\Big(-\sqrt{\frac{2\delta}{\tau}}\langle\Psi(x), \xi\rangle\Big). \tag{5.1.8}$$

This induces a bias towards accepting moves for which the the Gaussian random variable ξ, which is independent of x, aligns with the negative gradient of Ψ. Formalizing this heuristic is the content of Section 5.3.

Because the SDE (5.1.7) does not possess the smoothing property, almost sure fine scale properties under its invariant measure π^τ are not necessarily reflected at any finite time. For example if C is the covariance operator of Brownian motion or Brownian bridge then the quadratic variation of draws from the invariant measure, an almost sure quantity, is not reproduced at any finite time in (5.1.7) unless $z(0)$ has this quadratic variation; the almost sure property is approached asymptotically as $t \to \infty$. This behaviour is reflected in the underlying Metropolis-Hastings Markov chain P-RWM which approximates (5.1.7), where the almost sure property is only reached asymptotically as $k \to \infty$. In section 5.4 of this chapter we will show that almost sure quantities such as the quadratic variation under P-RWM satisfy a limiting linear ODE with globally attractive steady state given by the value of the quantity under π^τ. This gives quantitative information about the rate at which the P-RWM algorithm approaches statistical equilibrium.

One might wonder why we constructed the proposals based on the discretizations of an Ornstein-Uhlenbeck diffusion while we could just have considered discretizations from a Brownian motion. A standard random walk method S-RWM would use the proposal

$$x + \sqrt{\delta\tau}\,\xi\,, \tag{5.1.9}$$

in place of (5.1.4), which is the discretization of a Brownian motion in $\mathcal{H}$ with covariance C. This, however, leads to the accept-reject formula

$$\alpha^\delta(x,\xi) = 1 \wedge \exp\Big(-\frac{1}{\tau}\big(J\big(x+\sqrt{\delta\tau}\xi\big) - J(x)\big)\Big)$$

in place of (5.1.5). Unfortunately $J(x)$ is almost surely infinite with respect to x drawn from π^τ if $\mathcal{H}$ is infinite dimensional; consequently the S-RWM algorithm is only defined after finite dimensional approximation of the space but not well defined as a infinite dimensional Hilbert space valued MCMC algorithm. From this point of view, P-RWM is the right generalization of the random walk proposal from finite dimensions since it does not suffer from any such restriction. We return to this point in Section 5.6.

Section 5.2 contains a precise definition of the Markov chain $\{x^{k,\delta}\}_{k\in\mathbb{Z}^+}$, together with statement and proof of the weak convergence theorem that is the main result of the chapter. Section 5.3 contains proof of the lemmas which underly the weak convergence theorem. In section 5.4 we state and prove the limit theorem for almost sure quantities such as quadratic variation; such results are often termed "fluid limits" in the applied probability literature. An example is presented in

section 5.5. We conclude in section 5.6.

5.2 Main theorem

This section contains a precise statement of the algorithm, statement of the main theorem showing that piecewise linear interpolant of the output of the algorithm converges weakly to a noisy gradient flow, and proof of the main theorem. The proof of various technical lemmas is defered to section 5.3.

5.2.1 P-RWM algorithm

The reader is referred to section 3.1 for background on Gaussian measures. Let $\mathcal{H}$ be a separable Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and associated norm $\|x\|^2 = \langle x, x \rangle$. We consider a covariance operator $C : \mathcal{H} \to \mathcal{H}$ that is trace class and diagonalisable in an orthonormal Hilbert basis $\{\varphi_j\}_{j \geq 1}$ that will be referred to as *Karhunen-Loève eigen-basis*,

$$C\varphi_j = \lambda_j^2 \, \varphi_j \qquad \text{and} \qquad \mathrm{Tr}(C) = \sum_{j \geq 1} \lambda_j^2 < \infty.$$

In other words, the eigenvalues of the covariance operator C are $\{\lambda_j^2\}_{j \geq 1}$. Any vector $x \in \mathcal{H}$ can be decomposed on the Karhunen-Loève basis as $x = \sum_{j \geq 1} x_j \, \varphi_j$ where $x_j = \langle x, \varphi_j \rangle$. Consider a potential function $\Psi : \mathcal{H} \to \mathbb{R}$. We assume that the pair (C, Ψ) satisfies assumption 3.1.1. This means that there exists an exponent $s \geq 0$ such that for every $\tau > 0$ the support of the Gaussian measure $\pi_0^\tau := \mathrm{N}(0, \tau C)$ is included in $\mathcal{H}^s$ and that the function Ψ is well defined on $\mathcal{H}^s$ and satisfies various regularity estimates. The Sobolev-like subspace $\mathcal{H}^s$ is rigorously defined in section 3.1.2. As described in section 3.1, one can define the operator $C_s : \mathcal{H}^s \to \mathcal{H}^s$ such that the Gaussian measure $\mathrm{N}(0, C)$ in $\mathcal{H}$ can also be described as the Gaussian measure $\mathrm{N}(0, C_s)$ in $\mathcal{H}_s$. One can define a probability distribution π^τ on $\mathcal{H}$ through the formula

$$\frac{\mathrm{p}i^\tau}{d\pi_0^\tau}(x) \propto \exp\{-\Psi(x)/\tau\}. \tag{5.2.1}$$

Notice that the support of π^τ is included in $\mathcal{H}^s$ for every temperature $\tau > 0$.

We now define the Markov chain in $\mathcal{H}^s$ which is reversible with respect to the measure π^τ given by equation (5.2.1). This is the Metropolis-Hastings method introduced in [BRSV08] and refered to there as the P-RWM algorithm. Let $x \in \mathcal{H}^s$ be the current position of the Markov chain. The proposal candidate y is given by

(5.1.4), so that

$$y = \big(1-2\delta\big)^{\frac{1}{2}} x + \sqrt{2\delta\tau}\,\xi \qquad \text{where} \qquad \xi = \mathrm{N}(0, C) \tag{5.2.2}$$

and $\delta \in (0, \frac{1}{2})$ is a small parameter which we will send to zero in order to obtain the noisy gradient flow. In equation (5.2.2), the random variable ξ is chosen to be *independent* of x. As described in [BRSV08] (see also [CDS12; Stu10]), at temperature $\tau \in (0, \infty)$ the Metropolis-Hastings acceptance probability for the proposal y is given by

$$\alpha^{\delta}(x, \xi) = 1 \wedge \exp\Big(-\frac{1}{\tau}\big(\Psi(y) - \Psi(x)\big)\Big). \tag{5.2.3}$$

For future use, we define the local mean acceptance probability at the current position x via the formula

$$\alpha^{\delta}(x) \;=\; \mathbb{E}_x\big[\alpha^{\delta}(x, \xi)\big]. \tag{5.2.4}$$

The chain is then reversible with respect to π^{τ}. The Markov chain $x^{\delta} = \{x^{k,\delta}\}_{k\geq 0}$ can be written as

$$x^{k+1,\delta} = \gamma^{k,\delta} y^{k,\delta} + (1 - \gamma^{k,\delta})\, x^{k,\delta} \tag{5.2.5}$$

with $y^{k,\delta} = \big(1-2\delta\big)^{\frac{1}{2}}\, x^{k,\delta} + \sqrt{2\delta\tau}\xi^k$. Here the ξ^k are iid Gaussian random variables $\mathrm{N}(0, C)$ and the $\gamma^{k,\delta}$ are Bernoulli random variables which account for the accept-reject mechanism of the Metropolis-Hastings algorithm,

$$\gamma^{k,\delta} \overset{\text{def}}{=} \gamma^{\delta}(x^{k,\delta}, \xi^k) \;\overset{\mathcal{D}}{\sim}\; \text{Bernoulli}\Big(\alpha^{\delta}(x^{k,\delta}, \xi^k)\Big). \tag{5.2.6}$$

The function $\gamma^{\delta}(x, \xi)$ can be expressed as $\gamma^{\delta}(x, \xi) = \mathbb{I}_{\{U < \alpha^{\delta}(x,\xi)\}}$ where $U \overset{\mathcal{D}}{\sim} \text{Uniform}(0, 1)$ is independent from any other source of randomness. The next lemma will be repeatedly used in the sequel. It states that the size of the jump $y - x$ is of order $\sqrt{\delta}$.

Lemma 5.2.1. *Under assumptions 3.1.1 and for any integer $p \geq 1$ the following inequality*

$$\mathbb{E}_x\big[\|y - x\|_s^p\big]^{\frac{1}{p}} \quad \lesssim \quad \delta\,\|x\|_s + \sqrt{\delta} \quad \lesssim \quad \sqrt{\delta}\,\big(1 + \|x\|_s\big)$$

holds for any $\delta \in (0, \frac{1}{2})$.

Proof. The definition of the proposal (5.2.2) shows that $\|y - x\|_s^p \lesssim \delta^p\ \|x\|_s^p +$

$\delta^{\frac{p}{2}}\,\mathbb{E}\big[\|\xi\|_s^p\big]$. Fernique's theorem [DPZ92] shows that $\mathbb{E}\big[\|\xi\|_s^p\big] < \infty$. This gives the conclusion. □

5.2.2 Main theorem

Fix a time horizon $T > 0$ and a temperature $\tau \in (0, \infty)$. The piecewise linear interpolant z^δ of the Markov chain (5.2.5) is defined by equation (5.1.6). The following is the main result of this section. Note that "weakly" refers to weak convergence of probability measures.

Theorem 5.2.2. *Let assumptions 3.1.1 hold. Let the Markov chain x^δ start at fixed position $x_* \in \mathcal{H}^s$. Then the sequence of processes z^δ converges weakly to z in $C([0,T];\mathcal{H}^s)$, as $\delta \to 0$, where z solves the $\mathcal{H}^s$-valued stochastic differential equation*

$$\begin{cases} dz &= -\Big(z + C\nabla\Psi(z)\Big)\,dt + \sqrt{2\tau}dW \\ z_0 &= x_* \end{cases} \tag{5.2.7}$$

and W is a Brownian motion in $\mathcal{H}^s$ with covariance operator equal to C_s.

For conceptual clarity, we derive Theorem 5.2.2 as a consequence of the general diffusion-approximation result that is the content of Proposition 3.3.1. To this end, one needs to establish a martingale-drift decomposition of the Markov chain x^δ,

$$x^{k+1,\delta} = x^{k,\delta} + d^\delta(x^{k,\delta})\,\delta + \sqrt{2\tau\delta}\,\Gamma^\delta(x^{k,\delta},\xi^k) \tag{5.2.8}$$

where the approximate drift d^δ and volatility term $\Gamma^\delta(x,\xi^k)$ are given by

$$\begin{aligned} d^\delta(x) &= \delta^{-1}\,\mathbb{E}\big[x^{k+1,\delta} - x^{k,\delta}\;|x^{k,\delta} = x\big] \\ \sqrt{2\tau\delta}\,\Gamma^\delta(x,\xi^k) &= \gamma^\delta(x,\xi^k)\,\big\{(\sqrt{1-2\delta}-1)\,x + \sqrt{2\tau\delta}\,\xi^k\big\} \;-\; d^\delta(x)\,\delta. \end{aligned} \tag{5.2.9}$$

Notice that $\big\{\Gamma^{k,\delta}\big\}_{k\geq 0}$, with $\Gamma^{k,\delta} \stackrel{\text{def}}{=} \Gamma^\delta(x^{k,\delta},\xi^k) = (2\tau\delta)^{-1/2}\,\big(x^{k+1,\delta} - x^{k,\delta} - \mathbb{E}\big[x^{k+1,\delta} - x^{k,\delta}\;|x^{k,\delta}\big]\big)$, is a martingale difference array in the sense that $M^{k,\delta} = \sum_{j=0}^k \Gamma^{j,\delta}$ is a martingale adapted to the natural filtration $\mathcal{F}^\delta = \{\mathcal{F}^{k,\delta}\}_{k\geq 0}$ of the Markov chain x^δ. The parameter δ represents a time increment. We define the piecewise linear rescaled noise process by

$$W^\delta(t) = \sqrt{\delta}\,\sum_{j=0}^{k}\Gamma^{j,\delta} \;+\; \frac{t-t_k}{\sqrt{\delta}}\,\Gamma^{k+1,\delta} \qquad \text{for} \qquad t_k \leq t < t_{k+1}. \tag{5.2.10}$$

In order to apply Proposition 3.3.1, one needs to show that as δ goes to zero the sequence of drift functions $d^\delta(\cdot)$ converges to a limiting drift function $\mu(x) := -\big(x + C\nabla\Psi(x)\big)$ and that the rescaled noise process W^δ converges weakly to a Brownian motion as δ goes to zero. We now rigorously prove that the three conditions necessary for Proposition 3.3.1 to hold are satisfied.

1. **Convergence of initial conditions:** this is obvious since for any value of δ the Markov chain x^δ starts at the fixed position $x^{0,\delta} = x_*$.

2. **Invariance principle:** lemma 5.3.7, proved in section 5.3.5, shows that under assumptions 3.1.1 the sequence of processes W^δ converges weakly in $C([0,T], \mathcal{H}^s)$ to a Brownian motion W in $\mathcal{H}^s$ with covariance C_s. Since the starting position $x^{0,\delta} = x_*$ is deterministic, this indeed implies that the sequence of pairs $(x^{0,\delta}, W^\delta)$ converges weakly in $\mathcal{H}^s \times C([0,T], \mathcal{H}^s)$ to the random variable (x_*, W).

3. **Convergence of the drift:** we prove that the quantity $\int_0^T \big\|d^\delta(\bar{z}^\delta(u)) - \mu(z^\delta(u))\big\|\, du$ converges to zero in expectation as $\delta \to 0$ where $\bar{z}^\delta$ is the piecewise constant interpolant of x^δ accelerated by a factor $1/\delta$ and $\mu(x) = -\big(x + C\nabla\Psi(x)\big)$. To this end, we bound the quantity $\big\|d^\delta(\bar{z}^\delta(u)) - \mu(z^\delta(u))\big\|$ by the sum of $\big\|d^\delta(\bar{z}^\delta(u)) - \mu(\bar{z}^\delta(u))\big\|$ and $\big\|\mu(\bar{z}^\delta(u)) - \mu(z^\delta(u))\big\|$. Lemma 5.3.3, proved in section 5.3.2, shows that under assumptions 3.1.1 the sequence of approximate drift function $d^\delta(\cdot)$ satisfies the bound $\|d^\delta(x) - \mu(x)\|_s^p \lesssim \delta^{\frac{p}{2}}(1 + \|x\|_s^{2p})$ for any integer $p \geq 1$. This shows that $\int_0^T \big\|d^\delta(\bar{z}^\delta(u)) - \mu(z^\delta(u))\big\|\, du$ is less than a constant multiple of $\delta^{3/2} \sum_{k\delta<T}(1 + \|x^{k,\delta}\|_s^2)$ and the a-priori estimate of Lemma 5.3.6 shows that this quantity converges to zero in expectation as $\delta \to 0$. To finish, one needs to show that $\int_0^T \big\|\mu(\bar{z}^\delta(u)) - \mu(z^\delta(u))\big\|\, du$ converges to zero in expectation. Since the drift function $\mu(\cdot)$ is globally Lipschitz on $\mathcal{H}^s$ and Lemma 5.2.1 states that $\mathbb{E}\|x^{k+1,\delta} - x^{k,\delta}\| \lesssim \delta^{\frac{1}{2}}\big(1 + \|x^{k,\delta}\|\big)$ it follows that $\mathbb{E}\big[\int_0^T \big\|\mu(\bar{z}^\delta(u)) - \mu(z^\delta(u))\big\|\, du\big]$ is less than a constant multiple of $\delta^{3/2}\,\mathbb{E}\big[\sum_{k\delta<T}(1 + \|x^{k,\delta}\|_s)\big]$. The a-priori estimate of Lemma 5.3.6 again shows that this quantity goes to zero in expectation.

In conclusion, we have proved that under assumptions 3.1.1 the diffusion approximation result presented in Proposition 3.3.1 can be applied. This finishes the proof of Theorem 5.2.2.

5.3 Key estimates

This section contains the proof of various technical lemmas which are used in the previous section.

5.3.1 Acceptance probability asymptotics

This section describes a first order expansion of the acceptance probability. The approximation $\alpha^\delta(x,\xi) \approx \bar{\alpha}^\delta(x,\xi)$ where

$$\bar{\alpha}^\delta(x,\xi) = 1 - \sqrt{\frac{2\delta}{\tau}}\langle \nabla\Psi(x), \xi\rangle \mathbb{I}_{\{\langle \nabla\Psi(x),\xi\rangle > 0\}} \tag{5.3.1}$$

is valid for $\delta \ll 1$. The quantity $\bar{\alpha}^\delta$ has the advantage over α^δ of being very simple to analyse: explicit computations are available. This will be exploited in section 5.3.2. The quality of the approximation (5.3.1) is rigorously quantified in the next lemma.

Lemma 5.3.1. (Acceptance probability estimate)
Let assumptions 3.1.1 hold. For any integer $p \geq 1$ the quantity $\bar{\alpha}^\delta(x,\xi)$ satisfies

$$\mathbb{E}_x\big[|\alpha^\delta(x,\xi) - \bar{\alpha}^\delta(x,\xi)|^p\big] \lesssim \delta^p\,(1 + \|x\|_s^{2p}). \tag{5.3.2}$$

Proof. Let us introduce the two 1-Lipschitz functions $h, h_* : \mathbb{R} \to \mathbb{R}$ defined by

$$h(x) = 1 \wedge e^x \qquad \text{and} \qquad h_*(x) = 1 + x\,1_{\{x<0\}}. \tag{5.3.3}$$

The function h_* is a first order approximation of h in a neighbourhood of zero and we have

$$\alpha^\delta(x,\xi) \;=\; h\Big(-\frac{1}{\tau}\{\Psi(y) - \Psi(x)\}\Big) \qquad \text{and} \qquad \bar{\alpha}^\delta(x,\xi) \;=\; h_*\Big(-\sqrt{\frac{2\delta}{\tau}}\,\langle \nabla\Psi(x), \xi\rangle\Big)$$

where the proposal y is a function of x and ξ, as described in equation (5.2.2). Since $h_*(\cdot)$ is close to $h(\cdot)$ in a neighbourhood of zero, the proof is finished once it is proved that $-\frac{1}{\tau}\{\Psi(y) - \Psi(x)\}$ is close to $-\sqrt{\frac{2\delta}{\tau}}\,\langle \nabla\Psi(x), \xi\rangle$. We have $\mathbb{E}_x\Big[|\alpha^\delta(x,\xi) - \bar{\alpha}^\delta(x,\xi)|^p\Big] \lesssim A_1 + A_2$ where the quantities A_1 and A_2 are given by

$$A_1 = \mathbb{E}_x\Big[\Big|h\Big(-\frac{1}{\tau}\{\Psi(y) - \Psi(x)\}\Big) \;-\; h\Big(-\sqrt{\frac{2\delta}{\tau}}\,\langle \nabla\Psi(x), \xi\rangle\Big)\Big|^p\Big]$$
$$A_2 = \mathbb{E}_x\Big[\Big|h\Big(-\sqrt{\frac{2\delta}{\tau}}\,\langle \nabla\Psi(x), \xi\rangle\Big) \;-\; h_*\Big(-\sqrt{\frac{2\delta}{\tau}}\,\langle \nabla\Psi(x), \xi\rangle\Big)\Big|^p\Big].$$

By Lemma 3.1.4, the first order Taylor approximation of Ψ is controlled, $\big|\Psi(y) - \Psi(x) - \langle \nabla\Psi(x), y-x\rangle\big| \lesssim \|y-x\|_s^2$. The definition of the proposal y given in equation (5.2.2) shows that $\|(y-x) - \sqrt{2\delta\tau}\xi\|_s \lesssim \delta\|x\|_s$. Assumptions 3.1.1 state that for $z \in \mathcal{H}^s$ we have $\langle \nabla\Psi(x), z\rangle \lesssim \big(1+\|x\|_s\big)\cdot\|z\|_s$. Since the function $h(\cdot)$ is 1-Lipschitz it follows that

$$\begin{aligned} A_1 &= \mathbb{E}_x\Big[\big|h\big(-\frac{1}{\tau}\{\Psi(y)-\Psi(x)\}\big) - h\big(-\sqrt{\frac{2\delta}{\tau}}\,\langle\nabla\Psi(x),\xi\rangle\big)\big|^p\Big] \\ &\lesssim \mathbb{E}_x\Big[\big|\Psi(y)-\Psi(x)-\langle\nabla\Psi(x),y-x\rangle\big|^p \;+\; \big|\langle\nabla\Psi(x),y-x-\sqrt{2\delta\tau}\xi\rangle\big|^p\Big] \\ &\lesssim \mathbb{E}_x\Big[\|y-x\|_s^{2p} \;+\; (1+\|x\|_s^p)\cdot(\delta\,\|x\|_s)^p\Big] \;\lesssim\; \delta^p\,(1+\|x\|_s^{2p}). \end{aligned} \tag{5.3.4}$$

Lemma 5.2.1 has been used to control the size of $\mathbb{E}_x\big[\|y-x\|^p\big]$. To bound A_2, notice that for $z \in \mathbb{R}$ we have $|h(z) - h_*(z)| \le \frac{1}{2}\,z^2$. Therefore the quantity A_2 can be bounded by

$$\begin{aligned} A_2 &\lesssim \mathbb{E}_x\Big[|\sqrt{\delta}\,\langle\nabla\Psi(x),\xi\rangle|^{2p}\Big] \;\lesssim\; \delta^p\,\mathbb{E}_x\Big[(1+\|x\|_s^{2p})\,\|\xi\|_s^{2p}\Big] \\ &\lesssim\; \delta^p\,(1+\|x\|_s^{2p}). \end{aligned} \tag{5.3.5}$$

Estimates (5.3.4) and (5.3.5) together give equation (5.3.2). □

Recall the local mean acceptance probability defined by $\alpha^\delta(x) = \mathbb{E}_x[\alpha^\delta(x,\xi)]$ in equation (5.2.4). Define the approximate local mean acceptance probability by $\bar\alpha^\delta(x) \stackrel{\text{def}}{=} \mathbb{E}_x[\bar\alpha^\delta(x,\xi)]$. We now use Lemma 5.3.1 to approximate the local mean acceptance probability $\alpha^\delta(x)$.

Corollary 5.3.2. *Let assumptions 3.1.1 hold. For any integer $p \ge 1$ the following estimates hold,*

$$\big|\alpha^\delta(x) - \bar\alpha^\delta(x)\big| \;\lesssim\; \delta\,(1+\|x\|_s^2) \tag{5.3.6}$$

$$\mathbb{E}_x\Big[\;\big|\alpha^\delta(x,\xi) \;-\; 1\big|^p\Big] \;\lesssim\; \delta^{\frac{p}{2}}\,(1+\|x\|_s^p) \tag{5.3.7}$$

Proof. Let us prove equations (5.3.6) and (5.3.7).

- Lemma 5.3.1 and Jensen's inequality give equation (5.3.6).
- To prove (5.3.7), one can suppose $\delta^{\frac{p}{2}}\|x\|_s^p \le 1$. Indeed, if $\delta^{\frac{p}{2}}\|x\|_s^p \ge 1$, we have

$$\mathbb{E}_x\Big[\big|\alpha^\delta(x,\xi) \;-\; 1\big|^p\Big] \;\lesssim\; 1 \;\le\; \delta^{\frac{p}{2}}\|x\|_s^p \;\le\; \delta^{\frac{p}{2}}\,(1+\|x\|_s^p),$$

which gives the result. We thus suppose from now on that $\delta^{\frac{p}{2}}\|x\|_s \leq 1$. Under assumptions 3.1.1 we have $\|\nabla\Psi(x)\|_{-s} \lesssim 1 + \|x\|_s$. Lemma 3.1.4 shows that for all $x, y \in \mathcal{H}^s$ we have $\big|\Psi(y) - \Psi(x) - \langle\nabla\Psi(x), y - x\rangle\big| \lesssim \|y - x\|_s^2$. The function $h(x) = 1 \wedge e^x$ is 1-Lipschitz, $\alpha^\delta(x,\xi) = h\big(-\frac{1}{\tau}[\Psi(y) - \Psi(x)]\big)$ and $h(0) = 1$. Consequently,

$$\begin{aligned}\mathbb{E}_x\Big[\big|\alpha^\delta(x,\xi) \;-\; 1\big|^p\Big] &= \mathbb{E}_x\Big[\big|h\big(-\frac{1}{\tau}[\Psi(y) - \Psi(x)]\big) \;-\; h(0)\big|^p\Big] \\ &\lesssim \mathbb{E}_x\big[|\Psi(y) - \Psi(x)|^p\big] \;\lesssim\; \mathbb{E}_x\big[|\langle\nabla\Psi(x), y - x\rangle|^p + \|y - x\|_s^{2p}\big] \\ &\lesssim (1 + \|x\|_s^p)\cdot\mathbb{E}_x\big[\|y - x\|_s^p\big] + \mathbb{E}_x\big[\|y - x\|_s^{2p}\big].\end{aligned}$$

By Lemma 5.2.1, for any integer $\beta \geq 1$ we have $\mathbb{E}_x\big[\|y - x\|_s^\beta\big] \lesssim \delta^\beta\|x\|_s^\beta + \delta^{\frac{\beta}{2}}$ so that the assumption $\delta^{\frac{p}{2}}\|x\|_s^p \leq 1$ leads to

$$\begin{aligned}\mathbb{E}_x\Big[\big|\alpha^\delta(x) \;-\; 1\big|^p\Big] &\lesssim (1 + \|x\|_s^p)\cdot(\delta^p\|x\|_s^p + \delta^{\frac{p}{2}}) + (\delta^{2p}\|x\|_s^{2p} + \delta^p) \\ &\lesssim (1 + \|x\|_s^p)\cdot(\delta^{\frac{p}{2}} + \delta^{\frac{p}{2}}) + (\delta^p + \delta^p) \lesssim \delta^{\frac{p}{2}}\;(1 + \|x\|_s^p).\end{aligned}$$

This finishes the proof of Corollary 5.3.2.

□

5.3.2 Drift estimates

The main result of this section is a quantitative bound on the difference between the approximate drift function $d^\delta(\cdot)$ and the limiting drift function $\mu(x) = -\big(x + C\nabla\Psi(x)\big)$.

Lemma 5.3.3. (Drift estimate)
Let assumptions 3.1.1 hold and let $p \geq 1$ be an integer. Then the following estimate is satisfied,

$$\|d^\delta(x) - \mu(x)\|_s^p \;\lesssim\; \delta^{\frac{p}{2}}(1 + \|x\|_s^{2p}). \tag{5.3.8}$$

Moreover, the approximate drift d^δ is linearly bounded in the sense that

$$\|d^\delta(x)\|_s \;\lesssim\; 1 + \|x\|_s. \tag{5.3.9}$$

Before giving a proof of Lemma 5.3.3, we establish a preliminary result on the approximate acceptance probability $\bar{\alpha}^\delta(x,\xi)$. We will use these explicit computations, together with quantification of the error committed in replacing α^δ by

$\bar{\alpha}^\delta$, to estimate the mean drift (in this section) and the diffusion term (in the next section).

Lemma 5.3.4. *The approximate acceptance probability $\bar{\alpha}^\delta(x,\xi)$ satisfies*

$$\sqrt{\frac{2\tau}{\delta}}\,\mathbb{E}_x\Big[\bar{\alpha}^\delta(x,\xi)\cdot\xi\Big] \quad = \quad -C\nabla\Psi(x) \qquad\qquad \forall x\in\mathcal{H}^s.$$

Proof. Let $u = \sqrt{\frac{2\tau}{\delta}}\,\mathbb{E}_x\Big[\bar{\alpha}^\delta(x,\xi)\cdot\xi\Big] \in \mathcal{H}^s$. To prove $\langle u,v\rangle = -\langle C\nabla\Psi(x),v\rangle$. To this end, use the decomposition $v=\alpha\nabla\Psi(x)+w$ where $\alpha\in\mathbb{R}$ and $w\in\mathcal{H}^{-s}$ satisfies $\langle C\nabla\Psi(x),w\rangle=0$. Since $\xi\overset{\mathcal{D}}{\sim}\mathrm{N}(0,C)$ the two Gaussian random variables

$$Z_\Psi \overset{\text{def}}{=} \langle\nabla\Psi(x),\xi\rangle \qquad\text{and}\qquad Z_w \overset{\text{def}}{=} \langle w,\xi\rangle$$

are independent. Indeed, (Z_Ψ,Z_w) is a Gaussian vector in $\mathbb{R}^2$ with $\mathrm{Cov}(Z_\Psi,Z_w)=0$. It thus follows that

$$\begin{aligned}
\langle u,v\rangle &= -2\,\langle\mathbb{E}_x\big[\langle\nabla\Psi(x),\xi\rangle 1_{\{\langle\nabla\Psi(x),\xi\rangle>0\}}\cdot\xi\ ,\ \alpha\nabla\Psi(x)+w\rangle\big]\\
&= -2\,\mathbb{E}_x\Big[\alpha Z_\Psi^2 1_{\{Z_\Psi>0\}} \ +\ Z_w\,Z_\Psi 1_{\{Z_\Psi>0\}}\Big] = -2\alpha\,\mathbb{E}_x\Big[Z_\Psi^2 1_{\{Z_\Psi>0\}}\Big]\\
&= \ -\alpha\mathbb{E}_x\Big[Z_\Psi^2\Big] = -\alpha\langle C\nabla\Psi(x),\nabla\Psi(x)\rangle \ = \ \langle -C\nabla\Psi(x),\alpha\nabla\Psi(x)+w\rangle\\
&= -\langle C\nabla\Psi(x),v\rangle,
\end{aligned}$$

which concludes the proof of Lemma 5.3.4. □

We now use this explicit computation to give a proof of the drift estimate Lemma 5.3.3.

Proof of Lemma 5.3.3. The function d^δ defined by equation (5.2.9) can also be expressed as

$$\begin{aligned}
d^\delta(x) &= \ \Big\{\frac{(1-2\delta)^{\frac12}-1}{\delta}\alpha^\delta(x)\,x\Big\} + \Big\{\sqrt{\frac{2\tau}{\delta}}\,\mathbb{E}_x[\alpha^\delta(x,\xi)\,\xi]\Big\} \qquad (5.3.10)\\
&= \ B_1+B_2,
\end{aligned}$$

where the mean local acceptance probability $\alpha^\delta(x)$ has been defined in equation (5.2.4) and the two terms B_1 and B_2 are studied below. To prove equation (5.3.8), it suffices to establish that

$$\begin{cases}
\|B_1+x\|_s^p \ \lesssim\ \delta^{\frac p2}(1+\|x\|_s^{2p})\\
\|B_2+C\nabla\Psi(x)\|_s^p \ \lesssim\ \delta^{\frac p2}(1+\|x\|_s^{2p}).
\end{cases} \qquad (5.3.11)$$

We now establish these two bounds.

- Lemma 5.3.1 and Corollary 5.3.2 show that

$$\begin{aligned}\|B_1 + x\|_s^p &= \Big\{\frac{(1-2\delta)^{\frac{1}{2}}-1}{\delta}\alpha^\delta(x)+1\Big\}^p \|x\|_s^p \\ &\lesssim \Big\{\Big|\frac{(1-2\delta)^{\frac{1}{2}}-1}{\delta}-1\Big|^p + \big|\alpha^\delta(x)-1\big|^p\Big\}\|x\|_s^p \\ &\lesssim \Big\{\delta^p + \delta^{\frac{p}{2}}(1+\|x\|_s^p)\Big\}\|x\|_s^p \;\lesssim\; \delta^{\frac{p}{2}}(1+\|x\|_s^{2p}).\end{aligned} \tag{5.3.12}$$

- Lemma 5.3.1 shows that

$$\begin{aligned}\|B_2 + C\nabla\Psi(x)\|_s^p &= \Big\|\sqrt{\frac{2\tau}{\delta}}\,\mathbb{E}_x[\alpha^\delta(x,\xi)\,\xi] + C\nabla\Psi(x)\Big\|_s^p \\ &\lesssim \delta^{-\frac{p}{2}}\big\|\mathbb{E}_x[\{\alpha^\delta(x,\xi)-\bar\alpha^\delta(x,\xi)\}\,\xi]\big\|_s^p + \Big\|\underbrace{\sqrt{\frac{2\tau}{\delta}}\,\mathbb{E}_x[\bar\alpha^\delta(x,\xi)\,\xi] + C\nabla\Psi(x)}_{=0}\Big\|_s^p.\end{aligned} \tag{5.3.13}$$

  By Lemma 5.3.4, the second term on the right hand is equal to zero. Consequently, Cauchy Schwarz' inequality implies that

$$\begin{aligned}\|B_2 + C\nabla\Psi(x)\|_s^p &\lesssim \delta^{-\frac{p}{2}}\mathbb{E}_x[\big|\alpha^\delta(x,\xi)-\bar\alpha^\delta(x,\xi)\big|^2]^{\frac{p}{2}} \\ &\lesssim \delta^{-\frac{p}{2}}\Big(\delta^2(1+\|x\|_s^4)\Big)^{\frac{p}{2}} \;\lesssim\; \delta^{\frac{p}{2}}(1+\|x\|_s^{2p}).\end{aligned}$$

Estimates (5.3.12) and (5.3.13) give equation (5.3.11). To complete the proof we establish the bound (5.3.9). The expression (5.3.10) shows that it suffices to verify

$$\sqrt{\frac{2\tau}{\delta}}\,\mathbb{E}_x[\alpha^\delta(x,\xi)\,\xi] \quad\lesssim\quad 1+\|x\|_s.$$

To this end, we use Lemma 5.3.4 and Corollary 5.3.2. By Cauchy-Schwarz,

$$\begin{aligned}\Big\|\sqrt{\frac{2\tau}{\delta}}\,\mathbb{E}_x\Big[\alpha^\delta(x,\xi)\cdot\xi\Big]\Big\|_s &= \Big\|\sqrt{\frac{2\tau}{\delta}}\,\mathbb{E}_x\Big[[\alpha^\delta(x,\xi)-1]\cdot\xi\Big]\Big\|_s \\ &\lesssim \delta^{-\frac{1}{2}}\,\mathbb{E}_x\Big[(\alpha^\delta(x,\xi)-1)^2\Big]^{\frac{1}{2}} \;\lesssim\; 1+\|x\|_s,\end{aligned}$$

which concludes the proof of Lemma 5.3.3. □

5.3.3 Noise estimates

In this section we estimate the error in the approximation $\Gamma^{k,\delta} \approx \mathrm{N}(0, C_s)$ in $\mathcal{H}^s$ where C_s has been defined in section 3.1.2 as the covariance of $\mathrm{N}(0, C)$ when seen

as a Gaussian measure on $\mathcal{H}^s$. To this end, let us introduce the covariance operator $D^\delta(x)$ of the martingale difference Γ^δ,

$$D^\delta(x) = \mathbb{E}\Big[\Gamma^{k,\delta} \otimes_{\mathcal{H}^s} \Gamma^{k,\delta} \,|x^{k,\delta} = x\Big].$$

For any $x, u, v \in \mathcal{H}^s$ the operator $D^\delta(x)$ satisfies

$$\mathbb{E}\Big[\langle \Gamma^{k,\delta}, u\rangle_s \langle \Gamma^{k,\delta}, v\rangle_s \,|x^{k,\delta} = x\Big] \;=\; \langle u, D^\delta(x) v\rangle_s.$$

The next lemma gives a quantitative version of the approximation $D^\delta(x) \approx C_s$.

Lemma 5.3.5. (Noise estimates)
Let assumptions 3.1.1 hold. For any pair of indices $i, j \geq 1$, the martingale difference term $\Gamma^\delta(x,\xi)$ satisfies

$$|\langle \hat{\varphi}_i, D^\delta(x)\, \hat{\varphi}_j\rangle_s \;-\; \langle \hat{\varphi}_i, C_s\, \hat{\varphi}_j\rangle_s| \;\lesssim\; \delta^{\frac{1}{8}} \cdot \big(1 + \|x\|_s\big) \tag{5.3.14}$$

$$|\operatorname*{Tr}_{\mathcal{H}^s}\big(D^\delta(x)\big) \;-\; \operatorname*{Tr}_{\mathcal{H}^s}\big(C_s\big)| \;\lesssim\; \delta^{\frac{1}{8}} \cdot \big(1 + \|x\|_s^2\big). \tag{5.3.15}$$

Proof. The martingale difference $\Gamma^\delta(x,\xi)$ defined in equation (5.2.9) can also be expressed as

$$\Gamma^\delta(x,\xi) = \xi + F(x,\xi)$$

where the error term $F(x,\xi) = F_1(x,\xi) + F_2(x,\xi)$ is given by

$$F_1(x,\xi) = (2\tau\delta)^{-\frac{1}{2}}\big((1-2\delta)^{\frac{1}{2}} - 1\big)\,\big(\gamma^\delta(x,\xi) \;-\; \mathbb{E}_x[\gamma^\delta(x,\xi)]\big)x$$
$$F_2(x,\xi) = \big(\gamma^\delta(x,\xi) - 1\big)\cdot\xi \;-\; \mathbb{E}_x\big[\gamma^\delta(x,\xi)\cdot\xi\big].$$

We now prove that the quanity $F(x,\xi)$ satisfies

$$\mathbb{E}_x\Big[\|F(x,\xi)\|_s^2\Big] \lesssim \delta^{\frac{1}{4}}\;(1 + \|x\|_s^2) \tag{5.3.16}$$

- We have $\delta^{-\frac{1}{2}}\big((1-2\delta)^{\frac{1}{2}} - 1\big) \lesssim \delta^{\frac{1}{2}}$ and $|\gamma^\delta(x,\xi)| \leq 1$. Consequently,

$$\mathbb{E}_x\Big[\|F_1(x,\xi)\|_s^2\Big] \lesssim \delta\;\|x\|_s^2 \tag{5.3.17}$$

- Let us now prove that F_2 satisfies

$$\mathbb{E}_x\Big[\|F_2(x,\xi)\|_s^2\Big] \lesssim \delta^{\frac{1}{4}}\;(1 + \|x\|^{\frac{1}{2}}). \tag{5.3.18}$$

To this end, use the decomposition

$$\begin{aligned}\mathbb{E}_x\Big[\|F_2(x,\xi)\|_s^2\Big] &\lesssim \mathbb{E}_x\Big[|\gamma^\delta(x,\xi)-1|^2\cdot\|\xi\|_s^2\Big] \;+\; \|\mathbb{E}_x\big[\gamma^\delta(x,\xi)\cdot\xi\big]\|_s^2\\ &= I_1 \;+\; I_2.\end{aligned}$$

The Cauchy-Schwarz inequality shows that $I_1 \lesssim \mathbb{E}_x\Big[|\gamma^\delta(x,\xi)-1|^4\Big]^{\frac{1}{2}}$ where the Bernoulli random variable $\gamma^\delta(x,\xi)$ can be expressed as $\gamma^\delta(x,\xi) = \mathbb{1}_{\{U<\alpha^\delta(x,\xi)\}}$ where $U \overset{\mathcal{D}}{\sim} \text{Uniform}(0,1)$ is independent from any other source of randomness. Consequently

$$\mathbb{E}_x\Big[|\gamma^\delta(x,\xi)-1|^4\Big] \;=\; \mathbb{E}_x\big[\mathbb{1}_{\{\gamma^\delta(x,\xi)=0\}}\big] \;=\; 1-\alpha^\delta(x)$$

where the mean local acceptance probability $\alpha^\delta(x)$ is defined by $\alpha^\delta(x) = \mathbb{E}_x[\alpha^\delta(x,\xi)] \in [0,1]$. The convexity of the function $x \to |1-x|$ ensures that

$$\big|1-\alpha^\delta(x)\big| \;=\; \big|1-\mathbb{E}_x\big[\alpha^\delta(x,\xi)\big]\big| \;\le\; \mathbb{E}_x\big[\big|1-\alpha^\delta(x,\xi)\big|\big] \;\lesssim\; \delta^{\frac{1}{2}}\,(1+\|x\|)$$

where the last inequality follows from Corollary 5.3.2. This proves that $I_1 \lesssim \delta^{\frac{1}{4}}\,(1+\|x\|^{\frac{1}{2}})$. To bound I_2, it suffices to notice

$$\begin{aligned}I_2 = \;\|\mathbb{E}_x\big[\gamma^\delta(x,\xi)\cdot\xi\big]\|_s^2 \;&=\; \|\mathbb{E}_x\big[\big(\gamma^\delta(x,\xi)-1\big)\cdot\xi\big]\|_s^2\\ \lesssim \mathbb{E}_x\Big[|\gamma^\delta(x,\xi)-1|^2\cdot\|\xi\|_s^2\Big] \;&=\; I_1\end{aligned}$$

so that $I_2 \lesssim I_1 \lesssim \delta^{\frac{1}{4}}\,(1+\|x\|^{\frac{1}{2}})$ and $\mathbb{E}_x\Big[\|F_2(x,\xi)\|_s^2\Big] \lesssim \delta^{\frac{1}{4}}\,(1+\|x\|^{\frac{1}{2}})$.

Combining equation (5.3.17) and (5.3.18) gives equation (5.3.16).

Let us now describe how equations (5.3.12) and (5.3.13) follow from the estimate (5.3.16).

- We have $\mathbb{E}[\langle\hat\varphi_i,\xi\rangle_s\langle\hat\varphi_j,\xi\rangle_s] = \langle\hat\varphi_i, C_s\;\hat\varphi_j\rangle_s$ and $\mathbb{E}_x[\langle\hat\varphi_i,\Gamma^\delta(x,\xi)\rangle_s\langle\hat\varphi_j,\Gamma^\delta(x,\xi)\rangle_s] = \langle\hat\varphi_i, D^\delta(x)\;\hat\varphi_j\rangle_s$ with $\Gamma^\delta(x,\xi) = \xi + F(x,\xi)$. Consequently,

$$\begin{aligned}\langle\hat\varphi_i, D^\delta(x)\;\hat\varphi_j\rangle_s \;-\; \langle\hat\varphi_i, C_s\;\hat\varphi_j\rangle_s &= \mathbb{E}_x[\langle\hat\varphi_i,F(x,\xi)\rangle_s\langle\hat\varphi_j,F(x,\xi)\rangle_s]\\ &\quad+\mathbb{E}_x[\langle\hat\varphi_i,\xi\rangle_s\langle\hat\varphi_j,F(x,\xi)\rangle_s]\\ &\quad+\mathbb{E}_x[\langle\hat\varphi_i,F(x,\xi)\rangle_s\langle\hat\varphi_j,\xi\rangle_s].\end{aligned}$$

We have $|\langle\hat\varphi_i,F(x,\xi)\rangle_s| \le \|F(x,\xi)\|_s$ and Cauchy Schwarz's inequality proves

that

$$\mathbb{E}_x[\langle\hat{\varphi}_i, F(x,\xi)\rangle_s\langle\hat{\varphi}_j,\xi\rangle_s]^2 \leq \mathbb{E}_x[\|F(x,\xi)\|_s\,\|\xi\|_s]^2 \lesssim \mathbb{E}_x[\|F(x,\xi)\|_s^2].$$

It thus follows from equation (5.3.16) that

$$\begin{aligned}|\langle\hat{\varphi}_i, D^\delta(x)\;\hat{\varphi}_j\rangle_s \;-\; \langle\hat{\varphi}_i, C_s\;\hat{\varphi}_j\rangle_s| &\lesssim \mathbb{E}_x\big[\|F(x,\xi)\|_s^2\big] + \mathbb{E}_x\big[\|F(x,\xi)\|_s^2\big]^{\frac{1}{2}} \\ &\lesssim \delta^{\frac{1}{8}}\,(1+\|x\|_s),\end{aligned}$$

finishing the proof of (5.3.12).

- We have $\mathrm{Tr}_{\mathcal{H}^s}(C_s) = \mathbb{E}[\|\xi\|_s^2]$ and $\mathrm{Tr}_{\mathcal{H}^s}(D^\delta(x)) = \mathbb{E}[\|\Gamma^\delta(x,\xi)\|_s^2]$. Estimate (5.3.16) thus shows that

$$\begin{aligned}|\operatorname*{Tr}_{\mathcal{H}^s}\big(D^\delta(x)\big) \;-\; \operatorname*{Tr}_{\mathcal{H}^s}\big(C_s\big)| &= \big|\mathbb{E}[\|\Gamma^\delta(x,\xi)\|_s^2 - \|\xi\|_s^2]\big| \\ &= \big|\mathbb{E}[\|\xi \;+\; F(x,\xi)\|_s^2 - \|\xi\|_s^2]\big| \\ &\lesssim \big|\mathbb{E}[\langle 2\xi \;+\; F(x,\xi), F(x,\xi)\rangle_s\big| \;\lesssim\; \mathbb{E}[\|2\xi \;+\; F(x,\xi)\|_s\,\|F(x,\xi)\|_s] \\ &\lesssim \mathbb{E}[4\|\xi\|_s^2 + \|F(x,\xi)\|_s^2]^{\frac{1}{2}} \cdot \mathbb{E}[\|F(x,\xi)\|_s^2]^{\frac{1}{2}} \\ &\lesssim \Big(1+\delta^{\frac{1}{4}}\,(1+\|x\|_s^2)\Big)^{\frac{1}{2}} \cdot \Big(\delta^{\frac{1}{8}}\,(1+\|x\|_s)\Big) \;\lesssim\; \delta^{\frac{1}{8}}(1+\|x\|_s^2),\end{aligned}$$

finishing the proof of (5.3.13).

□

5.3.4 A-priori bound

We prove in this section an a-priori bound of the L^p norm of the Markov chain x^δ. This shows among other things that the rescaled process z^δ given by equation (5.1.6) does not blow up in finite time.

Lemma 5.3.6. (A priori bound)
Consider a fixed time horizon $T > 0$ and an integer $p \geq 1$. Under assumptions 3.1.1 the following bound holds,

$$\sup\quad \Big\{\delta\cdot\mathbb{E}\Big[\sum_{k\delta\leq T}\|x^{k,\delta}\|_s^p\Big]\;:\delta\in(0,\frac{1}{2})\Big\} \;<\; \infty. \tag{5.3.19}$$

Proof. Without loss of generality, assume that $p = 2n$ for some potitive integer $n \geq 1$. We now prove that there exist constants $\alpha_1, \alpha_2, \alpha_3 > 0$ satisfying

$$\mathbb{E}[\|x^{k,\delta}\|_s^{2n}] \leq (\alpha_1 + \alpha_2 k\,\delta)e^{\alpha_3 k\,\delta}. \tag{5.3.20}$$

Lemma 5.3.6 is a straightforward consequence of equation (5.3.20) since this implies that

$$\delta \sum_{k\delta<T} \mathbb{E}[\|x^{k,\delta}\|_s^{2n}] \;\leq\; \delta \sum_{k\delta<T} (\alpha_1 + \alpha_2 k\,\delta) e^{\alpha_3 k\,\delta} \;\asymp\; \int_0^T (\alpha_1 + \alpha_2\, t)\; e^{\alpha_3\, t} < \infty.$$

For notational convenience, let us define $V^{k,\delta} = \mathbb{E}\big[\|x^{k,\delta}\|_s^{2n}\big]$. To prove equation (5.3.20), it suffices to establish that

$$V^{k+1,\delta} - V^{k,\delta} \leq K\,\delta \cdot \big(1 + V^{k,\delta}\big), \tag{5.3.21}$$

where $K > 0$ is a constant independent from $\delta \in (0, \frac{1}{2})$. Indeed, iterating inequality (5.3.21) leads to the bound (5.3.20), for some computable constants $\alpha_1, \alpha_2, \alpha_3 > 0$. The definition of V^k shows that

$$\begin{aligned} V^{k+1,\delta} - V^{k,\delta} &= \mathbb{E}\big[\|x^{k,\delta} + (x^{k+1,\delta} - x^{k,\delta})\|_s^{2n} - \|x^{k,\delta}\|_s^{2n}\big] \\ &= \mathbb{E}\Big[\Big\{\|x^{k,\delta}\|_s^2 + \|x^{k+1,\delta} - x^{k,\delta}\|_s^2 + 2\langle x^{k,\delta}, x^{k+1,\delta} - x^{k,\delta}\rangle_s\Big\}^n - \|x^{k,\delta}\|_s^{2n}\Big] \end{aligned} \tag{5.3.22}$$

where the increment $x^{k+1,\delta} - x^{k,\delta}$ is given by

$$x^{k+1,\delta} - x^{k,\delta} \;=\; \gamma^{k,\delta}\,\Big((1-2\delta)^{\frac{1}{2}} - 1\Big)x^{k,\delta} + \sqrt{2\delta}\,\gamma^{k,\delta}\xi^k. \tag{5.3.23}$$

To bound the right-hand-side of equation (5.3.22), we use a binomial expansion and control each term. To this end, we establish the following estimate: for all integers $i, j, k \geq 0$ satisfying $i + j + k = n$ and $(i, j, k) \;\neq\; (n, 0, 0)$ the following inequality holds,

$$\begin{aligned} \mathbb{E}\big[\big(\|x^{k,\delta}\|_s^2\big)^i \big(\|x^{k+1,\delta} - x^{k,\delta}\|_s^2\big)^j \big(\langle x^{k,\delta}, x^{k+1,\delta} - x^{k,\delta}\rangle_s\big)^k\big] \\ \lesssim \delta(1 + V^{k,\delta}). \end{aligned} \tag{5.3.24}$$

To prove equation (5.3.24), we separate two different cases.

- Let us suppose $(i, j, k) = (n-1, 0, 1)$. Lemma 5.3.3 states that the approximate drift has a linearly bounded growth so that $\big\|\mathbb{E}\big[x^{k+1,\delta} - x^{k,\delta} \;|x^{k,\delta}\big]\big\|_s = \delta \times \|d^\delta(x^{k,\delta})\|_s \lesssim \delta\,(1 + \|x^{k,\delta}\|_s)$. Consequently, we have

$$\begin{aligned} \mathbb{E}\Big[\Big(\|x^{k,\delta}\|_s^2\Big)^{n-1} \langle x^{k,\delta}, x^{k+1,\delta} - x^{k,\delta}\rangle_s\Big] &\lesssim \mathbb{E}\Big[\|x^{k,\delta}\|_s^{2(n-1)}\; \|x^{k,\delta}\|_s \Big(\delta\,(1 + \|x^{k,\delta}\|_s\Big)\Big] \\ &\lesssim \delta(1 + V^{k,\delta}). \end{aligned}$$

  This proves equation (5.3.24) in the case $(i, j, k) = (n-1, 0, 1)$.

- Let us suppose $(i,j,k) \notin \Big\{(n,0,0),(n-1,0,1)\Big\}$. Because for any integer $p \geq 1$,

$$\mathbb{E}_x\Big[\|x^{k+1,\delta} - x^{k,\delta}\|_s^p\Big]^{\frac{1}{p}} \lesssim \delta^{\frac{1}{2}}\ (1 + \|x\|_s)$$

it follows from Cauchy-Schwarz inequality that

$$\mathbb{E}\Big[\Big(\|x^{k,\delta}\|_s^2\Big)^i\ \Big(\|x^{k+1,\delta} - x^{k,\delta}\|_s^2\Big)^j\ \Big(\langle x^{k,\delta}, x^{k+1,\delta} - x^{k,\delta}\rangle_s\Big)^k\Big] \ \lesssim\ \delta^{j+\frac{k}{2}}\ (1 + V^{k,\delta}).$$

Since we have supposed that $(i,j,k) \notin \Big\{(n,0,0),(n-1,0,1)\Big\}$ and $i+j+k = n$, it follows that $j + \frac{k}{2} \geq 1$. This concludes the proof of equation (5.3.24),

The binomial expansion of equation (5.3.22) and the bound (5.3.24) show that

$$V^{k+1,\delta} - V^{k,\delta} \ \lesssim\ \delta\,(1 + V^{k,\delta}).$$

This proves equation (5.3.21), which concludes the proof of Lemma 5.3.6. □

5.3.5 Invariance principle

Combining the noise estimates of Lemma 5.3.5 and the a priori bound of Lemma 5.3.6, we show that under assumptions 3.1.1 the sequence of rescaled noise processes defined in equation 5.2.10 converges weakly to a Brownian motion.

Lemma 5.3.7. (Invariance Principle)
Let assumptions 3.1.1 hold. Then the rescaled noise process $W^\delta(t)$ defined in equation (5.2.10) converges weakly in $C([0,T];\mathcal{H}^s)$ to a $\mathcal{H}^s$-valued Brownian motion W covariance operator C_s.

Proof. As described in [Ber86] [Proposition 5.1], in order to prove that W^δ converges weakly to W in $C([0,T];\mathcal{H}^s)$ it suffices to prove that for any $t \in [0,T]$ and any pair of indices $i,j \geq 0$ the following three limits hold in probability,

$$\lim_{\delta\to 0}\ \delta \sum_{k\delta<t} \mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^2\ |x^{k,\delta}\Big] \ =\ t \cdot \operatorname*{Tr}_{\mathcal{H}^s}(C_s) \tag{5.3.25}$$

$$\lim_{\delta\to 0}\ \delta \sum_{k\delta<t} \mathbb{E}\Big[\langle\Gamma^{k,\delta},\hat{\varphi}_i\rangle_s\langle\Gamma^{k,\delta},\hat{\varphi}_j\rangle_s\ |x^{k,\delta}\Big] \ =\ t\,\langle\hat{\varphi}_i, C_s\hat{\varphi}_j\rangle_s \tag{5.3.26}$$

$$\lim_{\delta\to 0}\ \delta \sum_{k\delta<T} \mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^2\ \mathbb{1}_{\{\|\Gamma^{k,\delta}\|_s^2 \geq \delta^{-1}\,\varepsilon\}}\ |x^{k,\delta}\Big] \ =\ 0 \qquad \forall \varepsilon > 0. \tag{5.3.27}$$

We now check that these three conditions are indeed satisfied.

- Condition (5.3.25): since $\mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^2 \;|x^{k,\delta}\Big] = \mathrm{Tr}_{\mathcal{H}^s}(D^\delta(x^{k,\delta}))$, lemma 5.3.5 shows that

$$\mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^2 \;|x^{k,\delta}\Big] = \underset{\mathcal{H}^s}{\mathrm{Tr}}(C_s) + \mathbf{e}_1^\delta(x^{k,\delta})$$

  where the error term $\mathbf{e}_1^\delta$ satisfies $|\mathbf{e}_1^\delta(x)| \lesssim \delta^{\frac{1}{8}}\ (1 + \|x\|_s^2)$. Consequently, to prove condition (5.3.25) it suffices to establish that

$$\lim_{\delta\to 0} \mathbb{E}\Big[\big|\delta \sum_{k\delta<T} \mathbf{e}_1^\delta(x^{k,\delta})\big|\Big] = 0.$$

  We have $\mathbb{E}\big[\big|\delta \sum_{k\delta<T} \mathbf{e}_1^\delta(x^{k,\delta})\big|\big] \lesssim \delta^{\frac{1}{8}}\Big\{\delta \cdot \mathbb{E}\Big[\sum_{k\delta<T}\ (1 + \|x^{k,\delta}\|_s^2)\Big]\Big\}$ and the apriori bound presented in Lemma 5.3.6 shows that

$$\sup_{\delta\in(0,\frac{1}{2})} \quad \Big\{\delta \cdot \mathbb{E}\Big[\sum_{k\delta<T} (1 + \|x^{k,\delta}\|_s^2)\Big]\Big\} < \infty.$$

  Consequently $\lim_{\delta\to 0}\ \mathbb{E}\big[\big|\delta \sum_{k\delta<T} \mathbf{e}_1^\delta(x^{k,\delta})\big|\big] = 0$, and the conclusion follows.

- Condition (5.3.26): lemma 5.3.5 states that

$$\mathbb{E}_k\Big[\langle\Gamma^{k,\delta}, \hat{\varphi}_i\rangle_s\langle\Gamma^{k,\delta}, \hat{\varphi}_j\rangle_s\Big] = \langle\hat{\varphi}_i, C_s\hat{\varphi}_j\rangle_s + \mathbf{e}_2^\delta(x^{k,\delta})$$

  where the error term $\mathbf{e}_2^\delta$ satisfies $|\mathbf{e}_2^\delta(x)| \lesssim \delta^{\frac{1}{8}}\ (1 + \|x\|_s)$. The exact same approach as the proof of Condition (5.3.25) gives the conclusion.

- Condition (5.3.27): from Cauchy-Schwarz and Markov's inequalities it follows that

$$\begin{aligned}
\mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^2\ \mathbb{1}_{\{\|\Gamma^{k,\delta}\|_s^2 \geq \delta^{-1}\,\varepsilon\}}\Big] &\leq \mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^4\Big]^{\frac{1}{2}} \cdot \mathbb{P}\Big[\|\Gamma^{k,\delta}\|_s^2 \geq \delta^{-1}\,\varepsilon\Big]^{\frac{1}{2}} \\
&\leq \mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^4\Big]^{\frac{1}{2}} \cdot \Big\{\frac{\mathbb{E}\big[\|\Gamma^{k,\delta}\|_s^4\big]}{(\delta^{-1}\,\varepsilon)^2}\Big\}^{\frac{1}{2}} \\
&\leq \frac{1}{\varepsilon^2}\ \delta^2 \cdot \mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^4\Big].
\end{aligned}$$

  Consequently we have

$$\mathbb{E}\Big[\Big|\delta \sum_{k\delta<T} \mathbb{E}\Big[\|\Gamma^{k,\delta}\|_s^2\ \mathbb{1}_{\{\|\Gamma^{k,\delta}\|_s^2 \geq \delta^{-1}\,\varepsilon\}}\ |x^{k,\delta}\Big]\ \Big|\Big] \leq \frac{1}{\varepsilon^2}\ \delta^2\ \Big\{\delta \cdot \mathbb{E}\Big[\sum_{k\delta<T}\ \|x^{k,\delta}\|_s^4\Big]\Big\}$$

and the conclusion again follows from the a priori bound Lemma 5.3.6.

□

5.4 Quadratic variation

As discussed in the introduction, the SPDE (5.1.7), and the Metropolis-Hastings algorithm P-RWM which approximates it for small δ, do not satisfy the smoothing property and so almost sure properties of the limit measure π^τ are not necessarily seen at finite time. To illustrate this point, we introduce in this section a functional $V : \mathcal{H} \to \mathbb{R}$ that is well defined on a dense subset of $\mathcal{H}$ and such that $V(X)$ is π^τ-almost surely well defined and such that $\mathbb{P}\big(V(X) = 1\big) = \tau$ for $X \overset{\mathcal{D}}{\sim} \pi^\tau$. The quantity V corresponds to the usual quadratic variation if π_0 is the Wiener measure. We show that the quadratic variation like quantity $V(x^{k,\tau})$ of a P-RWM Markov chain converges as $k \to \infty$ to the almost sure quantity τ. We then prove that piecewise linear interpolation of this quantity solves, in the small δ limit, a linear ODE (the "fluid limit") whose globally attractive stable state is the almost sure quantity τ. This quantifies the manner in which the P-RWM method approaches statistical equilibrium.

5.4.1 Definition and properties

Under assumptions 3.1.1, the Karhunen-Loéve expansion and the strong Law of Large Numbers show that π_0-almost every $x \in \mathcal{H}$ satisfies

$$\lim_{N\to\infty} N^{-1} \sum_{j=1}^{N} \frac{\langle x, \varphi_j\rangle^2}{\lambda_j^2} = 1.$$

This motivates the definition of the quadratic variation like quantities

$$V_-(x) \overset{\text{def}}{=} \liminf_{N\to\infty} N^{-1} \sum_{j=1}^{n} \frac{\langle x, \varphi_j\rangle^2}{\lambda_j^2} \qquad \text{and} \qquad V_+(x) \overset{\text{def}}{=} \limsup_{N\to\infty} N^{-1} \sum_{j=1}^{n} \frac{\langle x, \varphi_j\rangle^2}{\lambda_j^2}.$$

When these two quantities are equal the vector $x \in \mathcal{H}$ is said to possess a *quadratic variation* $V(x)$ defined as $V(x) = V_-(x) = V_+(x)$. Consequently, π_0-almost every $x \in \mathcal{H}$ possesses a quadratic variation $V(x) = 1$. It is a straightforward consequence that π_0^τ-almost every and π^τ-almost every $x \in \mathcal{H}$ possesses a quadratic variation $V(x) = \tau$. Strictly speaking this only coincides with quadratic variation when C is the covariance of a (possibly conditioned) Brownian motion; however we use

the terminology more generally in this section. The next lemma proves that the quadratic variation $V(\cdot)$ behaves as it should do with respect to additivity.

Lemma 5.4.1. (Quadratic Variation Additivity)
Consider a vector $x \in \mathcal{H}$ and a Gaussian random variable $\xi \overset{\mathcal{D}}{\sim} \pi_0$ and a real number $\alpha \in \mathbb{R}$. Suppose that the vector $x \in \mathcal{H}$ possesses a finite quadratic variation $V(x) < +\infty$. Then almost surely the vector $x + \alpha\xi \in \mathcal{H}$ possesses a quadratic variation that is equal to

$$V(x + \alpha\xi) \;=\; V(x) + \alpha^2.$$

Proof. Let us define $V_N \overset{\text{def}}{=} N^{-1} \sum_1^N \frac{\langle x,\varphi_j\rangle \cdot \langle \xi,\varphi_j\rangle}{\lambda_j^2}$. To prove Lemma 5.4.1 it suffices to prove that almost surely the following limit holds

$$\lim_{N\to\infty} V_N \;=\; 0.$$

The Borel-Cantelli lemma shows that it suffices to prove that for every fixed $\varepsilon > 0$ we have $\sum_{N\geq 1} \mathbb{P}\big[\big|V_N\big| > \varepsilon\big] < \infty$. Notice then that V_N is a centred Gaussian random variables with variance

$$\mathrm{Var}(V_N) = \frac{1}{N}\left(N^{-1}\sum_1^N \frac{\langle x,\varphi_j\rangle^2}{\lambda_j^2}\right) \asymp \frac{V(x)}{N}.$$

The Markov's inequality yields that $\mathbb{P}\big[\big|V_N\big| > \varepsilon\big] \lesssim \varepsilon^{-4}\,\mathbb{E}\big[V_N^4\big] \lesssim \frac{1}{N^2}$ from which it follows that $\sum_{N\geq 1} \mathbb{P}\big[\big|V_N\big| > \varepsilon\big] < \infty$, finishing the proof of the lemma. □

5.4.2 Large k behaviour of quadratic variation for P-RWM

The P-RWM algorithm at temperature $\tau > 0$ and discretization parameter $\delta > 0$ proposes a move from x to y according to the dynamics

$$y = (1-2\delta)^{\frac{1}{2}} x + (2\delta\tau)^{\frac{1}{2}}\,\xi \qquad \text{with} \qquad \xi \overset{\mathcal{D}}{\sim} \pi_0.$$

This move is accepted with probability $\alpha^\delta(x,y)$. In this case, Lemma 5.4.1 shows that if the quadratic variation $V(x)$ exists then the quadratic variation of the proposed move $y \in \mathcal{H}$ exists and satisfies

$$\frac{V(y) - V(x)}{\delta} \;=\; -2(V(x) - \tau). \tag{5.4.1}$$

Consequently, one can prove that for any finite time step $\delta > 0$ and temperature $\tau > 0$ the quadratic variation of the MCMC algorithm converges to τ.

Proposition 5.4.2. (Limiting Quadratic Variation) *Let assumptions 3.1.1 hold and $\{x^{k,\delta}\}_{k\geq 0}$ be the Markov chain of section 5.2.1. Then almost surely the quadratic variation of the Markov chain converges to τ,*

$$\lim_{k\to\infty} V(x^{k,\delta}) \;=\; \tau.$$

Proof. Let us first show that the number of accepted moves is infinite. If this were not the case, the Markov chain would eventually reach a position $x^{k,\delta} = x \in \mathcal{H}$ such that all subsequent proposals $y^{k+l} = (1-2\delta)^{\frac{1}{2}}\, x^k + (2\tau\delta)^{\frac{1}{2}}\, \xi^{k+l}$ would be refused. This means that the i.i.d. Bernoulli random variables $\gamma^{k+l} = \text{Bernoulli}\big(\alpha^\delta(x^k, y^{k+l})\big)$ satisfy $\gamma^{k+l} = 0$ for all $l \geq 0$. This can only happen with probability 0. Indeed, since $\mathbb{P}[\gamma^{k+l} = 1] > 0$, one can use Borel-Cantelli lemma to show that almost surely there exists $l \geq 0$ such that $\gamma^{k+l} = 1$. To conclude the proof of the proposition, notice then that the sequence $\{u_k\}_{k\geq 0}$ defined by $u_{k+1} - u_k \;=\; -2\delta(u_k - \tau)$ converges to τ. □

5.4.3 Fluid limit for quadratic variation of P-RWM

To gain further insight into the rate at which the limiting behaviour of the quadratic variation is observed for P-RWM we derive an ODE "fluid limit" for the Metropolis-Hastings algorithm. We introduce the continuous time process $t \mapsto v^\delta(t)$ defined as continuous piecewise linear interpolation of the the process $k \mapsto V(x^{k,\delta})$; for $t_k \leq t < t_{k+1}$ we define

$$v^\delta(t) = \frac{1}{\delta}\,(t - t_k)\, V(\, x^{k+1,\delta}\,) + \frac{1}{\delta}\,(t_{k+1} - t)\, V(\, x^{k,\delta}\,). \tag{5.4.2}$$

Since the acceptance probability of P-RWM approaches 1 as $\delta \to 0$ (see Corollary 5.3.2) equation (5.4.1) shows heuristically that the trajectories of of the process $t \mapsto v^\delta(t)$ should be well approximated by the solution of the (non stochastic) differential equation

$$\dot{v} = -2\,(v - \tau) \tag{5.4.3}$$

We prove such a result, in the sense of convergence in probability in $C([0,T];\mathbb{R})$:

Theorem 5.4.3. (Fluid Limit For Quadratic Variation) *Let assumptions 3.1.1 hold. Let the Markov chain x^δ start at fixed position $x_* \in \mathcal{H}^s$. Assume that $x_* \in \mathcal{H}$*

possesses a finite quadratic variation, $V(x_*) < \infty$. Then the function $v^\delta(t)$ converges in probability in $C([0,T],\mathbb{R})$, as δ goes to 0, to the solution of the differential equation (5.4.3) with initial condition $v_0 = V(x_*)$.

As already indicated, the heart of the proof of the result consists in showing that the acceptance probability of the algorithm converges to 1 as δ goes to 0. We prove such a result as Lemma 5.4.4 below, and then proceed to prove Theorem 5.4.3. To this end we introduce $t^\delta(k)$, the number of accepted moves:

$$t^\delta(k) \stackrel{\text{def}}{=} \sum_{l \le k} \gamma^{l,\delta},$$

where $\gamma^{l,\delta} = \text{Bernoulli}(\alpha^\delta(x,y))$ is the Bernoulli random variable defined in equation (5.2.6). Since the acceptance probability of the algorithm converges to 1 as $\delta \to 0$, the approximation $t^\delta(k) \approx k$ holds. In order to prove a fluid limit result on the interval $[0,T]$ one needs to prove that the quantity $\left|t^\delta(k) - k\right|$ is small when compared to δ^{-1}. The next lemma shows that such bounds hold uniformly on the interval $[0,T]$.

Lemma 5.4.4. (Number of Accepted Moves) *Let assumptions 3.1.1 hold. The number of accepted moves $t^\delta(\cdot)$ verifies*

$$\lim_{\delta \to 0} \quad \sup\left\{\delta \cdot \left|t^\delta(k) - k\right| \ : 0 \le k \le T\delta^{-1}\right\} = 0$$

where the convergence holds in probability.

The proof of Lemma 5.4.4 consists in showing first that for any $\varepsilon > 0$ one can find a ball of radius $R(\varepsilon)$ around 0 in $\mathcal{H}^s$,

$$B_0(R(\varepsilon)) = \left\{x \in \mathcal{H}_s : \|x\|_s \le R(\varepsilon)\right\},$$

such that with probability $1 - 2\varepsilon$ we have $x^{k,\delta} \in B_0(R(\varepsilon))$ and $y^{k,\delta} \in B_0(R(\varepsilon))$ for all $0 \le k \le T\delta^{-1}$. As is described below, the existence of such a ball follows from the bound

$$\mathbb{E}\left[\sup_{t \in [0,T]} \|x(t)\|_s\right] \ < \ +\infty \tag{5.4.4}$$

where $t \mapsto x(t)$ is the solution of the stochastic differential equation (5.2.7). For the sake of completeness, we include a proof of equation (5.4.4). The solution $t \mapsto x(t)$ of the stochastic differential equation (5.2.7) satisfies $x(t) = \int_0^t d\big(x(u)\big)\, du + \sqrt{2\tau}\, W(t)$ for all $t \in [0,T]$ where the drift function $\mu(x) = -\big(x + C\nabla\Psi(x)\big)$ is globally Lipschitz

on $\mathcal{H}^s$, as described in Lemma 3.1.4. Consequently $\|\mu(x)\|_s \leq A(1+\|x\|_s)$ for some positive constant $A > 0$. The triangle inequality then shows that

$$\|x(t)\|_s \leq A \int_0^t \big(1 + \|x(u)\|_s\big)\, du + \sqrt{2\tau}\|W(t)\|_s. \tag{5.4.5}$$

By Gronwall's inequality we obtain

$$\sup_{[0,T]} \|x(t)\|_s \leq (AT + \sup_{[0,T]} \|W(t)\|_s)\left[1 + ATe^{AT}\right]. \tag{5.4.6}$$

Since $\mathbb{E}[\sup_{[0,T]} \|W(t)\|_s] < \infty$, the bound (5.4.4) is proved.

Proof of Lemma 5.4.4. The proof consists in showing that the the acceptance probability of the algorithm is sufficiently close to 1 so that approximation $t^\delta(k) \approx k$ holds. The argument can be divided into 3 main steps. In the first part, we show that we can find a finite ball $B(0, R(\varepsilon))$ such that the trajectory of the Markov chain $\{x^{k,\delta}\}_{k \leq T\delta^{-1}}$ remains in this ball with probability at least $1 - 2\varepsilon$. This observation is useful since the function Ψ is Lipschitz on any ball of finite radius in $\mathcal{H}^s$. In the second part, using the fact that Ψ is Lipschitz on $B(0, R(\varepsilon))$, we find a lower bound for the acceptance probability α^δ. Then, in the last step, we use a moment estimate to prove that one can make the lower bound uniform on the interval $0 \leq k \leq T\delta^{-1}$.

- **Restriction to a Ball of Finite Radius**
  First, we show that with high probability the trajectory of the MCMC algorithm stays in a ball of finite radius. The functional $x \mapsto \sup_{t\in[0,T]} \|x(t)\|_s$ is continuous on $C([0,T], \mathcal{H}_s)$ and $\mathbb{E}\big[\sup_{t\in[0,T]} \|x(t)\|_s\big] < \infty$ for $t \mapsto x(t)$ following the stochastic differential equation (5.2.7), as proved in equation (5.4.4). Consequently, the weak convergence of z^δ to the solution of (5.2.7) encapsulated in Theorem 5.2.2 shows that $\mathbb{E}\big[\sup_{k<T\delta^{-1}} \|x^{k,\delta}\|_s\big]$ can be bounded by a finite universal constant independent from δ. Given $\varepsilon > 0$, Markov inequality thus shows that one can find a radius $R_1 = R_1(\varepsilon)$ large enough so that the inequality

  $$\mathbb{P}\big[\|x^{k,\delta}\|_s < R_1 \quad \text{for all} \quad 0 \leq k \leq T\delta^{-1}\big] > 1 - \varepsilon \tag{5.4.7}$$

  for any $\delta \in (0, \frac{1}{2})$. By Fernique's Theorem there exists $\alpha > 0$ such that $\mathbb{E}[e^{\alpha\|\xi\|_s^2}] < \infty$. This implies that $\mathbb{P}[\|\xi\|_s > r] \lesssim e^{-\alpha r^2}$. Therefore, if $\{\xi_k\}_{k\geq 0}$ are i.i.d. Gaussian random variables distributed as $\xi \overset{\mathcal{D}}{\sim} \pi_0$, the union bound

shows that

$$\mathbb{P}\big[\|\sqrt{\delta}\xi_k\|_s \le r \quad \text{for all} \quad 0 \le k \le T\delta^{-1}\big] \gtrsim 1 - T\delta^{-1}\exp(-\alpha\delta^{-1}r^2).$$

This proves that one can choose $R_2 = R_2(\varepsilon)$ large enough in such a manner that

$$\mathbb{P}\big[\|\sqrt{\delta}\xi_k\|_s < R_2 \quad \text{for all} \quad 0 \le k \le T\delta^{-1}\big] > 1 - \varepsilon \tag{5.4.8}$$

for any $\delta \in (0, \frac{1}{2})$. At temperature $\tau > 0$ the MCMC proposals are given by $y^{k,\delta} = (1-2\delta)^{\frac{1}{2}}x^{k,\delta} + (2\delta\tau)^{\frac{1}{2}}\,\xi_k$. It thus follows from the bounds (5.4.7) and (5.4.8) that with probability at least $(1-2\varepsilon)$ the vectors $x^{k,\delta}$ and $y^{k,\delta}$ belong to the ball $B_0(R(\varepsilon)) = \{x \in \mathcal{H}_s : \|x\|_s < R(\varepsilon)\}$ for $0 \le k \le T\delta^{-1}$ where radius $R(\varepsilon)$ is given by $R(\varepsilon) = R_1(\varepsilon) + R_2(\varepsilon)$.

- **Lower Bound for Acceptance Probability**
  We now give a lower bound for the acceptance probability $\alpha^\delta(x^{k,\delta}, \xi^k)$ that the move $x^{k,\delta} \to y^{k,\delta}$ is accepted. Assumptions 3.1.1 state that $\|\nabla\Psi(x)\|_{-s} \lesssim 1 + \|x\|_s$. Therefore, the function $\Psi : \mathcal{H}^s \to \mathbb{R}$ is Lipschitz on $B_0(R(\varepsilon))$,

$$\|\Psi\|_{\mathrm{lip},\varepsilon} \stackrel{\text{def}}{=} \sup\left\{\frac{|\Psi(y) - \Psi(x)|}{\|y - x\|_s} \,:\, x, y \in B_0(R(\varepsilon))\right\} < \infty.$$

  One can thus bound the acceptance probability $\alpha^\delta(x^{k,\delta}, \xi^k) = 1 \wedge \exp\big(-\tau^{-1}[\Psi(y^{k,\delta}) - \Psi(y^{k,\delta})]\big)$ for $x^{k,\delta}, y^{k,\delta} \in B_0(R(\varepsilon))$. Since the function $z \mapsto 1 \wedge e^{-\tau^{-1}z}$ is Lipschitz with constant τ^{-1}, the definition of $\|\Psi\|_{\mathrm{lip},\varepsilon}$ shows that the bound

$$\begin{aligned} 1 - \alpha^\delta(x^{k,\delta}, \xi^k) &\le \tau^{-1}\,\|\Psi\|_{\mathrm{lip},\varepsilon}\,\|y^{k,\delta} - x^{k,\delta}\|_s \\ &\le \tau^{-1}\,\|\Psi\|_{\mathrm{lip},\varepsilon}\left\{[(1-2\delta)^{\frac{1}{2}} - 1]\,\|x^{k,\delta}\|_s + (2\delta\tau)^{\frac{1}{2}}\,\|\xi^k\|\right\} \\ &\lesssim \sqrt{\delta}\,(1 + \|\xi^k\|_s) \end{aligned}$$

  holds for every $x^{k,\delta}, y^{k,\delta} \in B_0(R(\varepsilon))$. Hence, there exists a constant $K = K(\varepsilon)$ such that $\widehat{\alpha}^\delta(\xi^k) = 1 - K\sqrt{\delta}\,(1 + \|\xi^k\|_s)$ satisfies $\alpha^\delta(x^{k,\delta}, \xi^k) > \widehat{\alpha}^\delta(\xi^k)$ for every $x^{k,\delta}, y^{k,\delta} \in B_0(R(\varepsilon))$. Since the trajectory of the MCMC algorithm stays in the ball $B_0(R(\varepsilon))$ with probability at least $1 - 2\varepsilon$ the inequality

$$\mathbb{P}[\alpha^\delta(x^{k,\delta}, \xi^k) > \widehat{\alpha}^\delta(\xi^k) \quad \text{for all} \quad 0 \le k \le T\delta^{-1}] \;>\; 1 - 2\varepsilon.$$

  holds for every $\delta \in (0, \frac{1}{2})$.

- **Second Moment Method**
  To prove that $t^\delta(k)$ does not deviate too much from k, we show that its expectation satisfies $\mathbb{E}[t^\delta(k)] \approx k$ and we then control the error by bounding the variance. Since the Bernoulli random variable $\gamma^{k,\delta} = \text{Bernoulli}(\alpha^\delta(x^{k,\delta}\xi^k))$ are not independent, the variance of $t^\delta(k) = \sum_{l\le k}\gamma^{l,\delta}$ is not easily computable. We thus introduce i.i.d. auxiliary random variables $\widehat{\gamma}^{k,\delta}$ such that

$$\sum_{l\le k}\widehat{\gamma}^{l,\delta} = \widehat{t}^\delta(k) \quad \approx \quad t^\delta(k) = \sum_{l\le k}\gamma^{l,\delta}.$$

  As described below, the behaviour of $\widehat{t}^\delta(k)$ is readily controlled since it is a sum of i.i.d. random variables. The proof then exploits the fact that $\widehat{t}^\delta(k)$ is a good approximation of $t^\delta(k)$.

  The Bernoulli random variables $\gamma^{k,\delta}$ can be described as $\gamma^{k,\delta} = \mathbb{1}\big(U_k < \alpha^\delta(x^{k,\delta}\xi^k)\big)$ where $\{U_k\}_{k\ge0}$ are i.i.d. random variables uniformly distributed on $(0,1)$. As a consequence, with probability at least $1-2\varepsilon$, the random variables $\widehat{\gamma}^{k,\delta} = \mathbb{1}\big(U_k < \widehat{\alpha}^\delta\big)$ satisfy $\gamma^{k,\delta} \ge \widehat{\gamma}^{k,\delta}$ for all $0 \le k \le T\delta^{-1}$. Therefore, with probability at least $1-2\varepsilon$, we have $t^{\delta(k)} \ge \widehat{t}^{\delta(k)}$ for all $0 \le k \le T\delta^{-1}$ where $\widehat{t}^{\delta(k)} = \sum_{l\le k}\widehat{\gamma}^{l,\delta}$. Consequently, since $t^{\delta(k)} \le k$, to prove Lemma 5.4.4 it suffices to show instead that the following limit in probability holds,

$$\lim_{\delta\to0} \quad \sup\big\{\delta\cdot\big|\widehat{t}^\delta(k) - k\big| \ : 0 \le k \le T\delta^{-1}\big\} = 0. \tag{5.4.9}$$

  Contrary to the random variables $\{\gamma^{k,\delta}\}_{k\ge0}$, the random variables $\{\widehat{\gamma}^{k,\delta}\}_{k\ge0}$ are i.i.d. and are thus easily controlled. By Doob's inequality we have

$$\mathbb{P}\Big[\sup\big\{\delta\cdot\big|\widehat{t}^\delta(k) - \mathbb{E}[\widehat{t}^\delta(k)]\big| \ : \ 0\le k\le T\delta^{-1}\big\} \ > \ \eta\Big] \ \le \ 2\,\frac{\mathrm{Var}\big(\widehat{t}^\delta(T\delta^{-1})\big)}{(\delta^{-1}\eta)^2} \ \le \ 2\,\frac{\delta T}{\eta^2}.$$

  Since $\mathbb{E}[\widehat{t}^\delta(k)] = k\cdot\big\{1 - K\sqrt{\delta}\,(1+\mathbb{E}[\|\xi^k\|_s])\big\}$, equation (5.4.9) follows. This finishes the proof of Lemma 5.4.4.

□

We now complete the proof of Theorem 5.4.3 using the key Lemma 5.4.4.

Proof of Theorem 5.4.3. The proof consists in proving that the trajectory of the quadratic variation process behaves as if all the move were accepted. The main

ingredient is the uniform lower bound on the acceptance probability given by Lemma 5.4.4.

Recall that $v^\delta(k\delta) = V(x^{k,\delta})$. Consider the piecewise linear function $\widehat{v}^\delta(\cdot) \in C([0,T],\mathbb{R})$ defined by linear interpolation of the values $\hat{v}^\delta(k\delta) = u^\delta(k)$ and where the sequence $\{u^\delta(k)\}_{k\geq 0}$ satisfies $u^\delta(0) = V(x_*)$ and

$$u^\delta(k+1) - u^\delta(k) \;=\; -2\delta\,(u^\delta(k) - \tau).$$

The value $u^\delta(k) \in \mathbb{R}$ represents the quadratic variation of $x^{k,\delta}$ if the k first moves of the MCMC algorithm had been accepted. One can readily check that as δ goes to zero the sequence of continuous functions $\hat{v}^\delta(\cdot)$ converges in $C([0,T],\mathbb{R})$ to the solution $v(\cdot)$ of the differential equation (5.4.3). Consequently, to prove Theorem 5.4.3 it suffices to show that for any $\varepsilon > 0$ we have

$$\lim_{\delta\to 0}\; \mathbb{P}\Big[\sup\Big\{\big|V(x^{k,\delta}) - u^\delta(k)\big| \;:\; k \leq \delta^{-1}T\Big\} \;>\; \varepsilon\Big] \;=\; 0. \tag{5.4.10}$$

The definition of the number of accepted moves $t^\delta(k)$ is such that $V(x^{k,\delta}) = u^\delta(t^\delta(k))$. Note that

$$u^\delta(k) = (1-2\delta)^k u_0 + \big(1 - (1-2\delta)^k\big)\tau. \tag{5.4.11}$$

Hence, for any integers $t_1, t_2 \geq 0$, we have $\big|u^\delta(t_2) - u^\delta(t_1)\big| \leq \big|u^\delta(|t_2 - t_1|) - u^\delta(0)\big|$ so that

$$\big|V(x^{k,\delta}) - u^\delta(k)\big| \;=\; \big|u^\delta(t^\delta(k)) - u^\delta(k)\big| \;\leq\; \big|u^\delta(k - t^\delta(k)) - u^\delta(0)\big|.$$

Equation (5.4.11) shows that $|u^\delta(k) - u^\delta(0)| \lesssim \big(1 - (1-2\delta)^k\big)$. This implies that

$$\big|V(x^{k,\delta}) - u^\delta(k)\big| \;\lesssim\; 1 - (1-2\delta)^{k - t^\delta(k)} \;\lesssim\; 1 - (1-2\delta)^{\delta^{-1}S}$$

where $S = \sup\big\{\delta\cdot\big|t^\delta(k) - k\big| \;:\; 0 \leq k \leq T\delta^{-1}\big\}$. Since for any $a > 0$ we have $1 - (1-2\delta)^{a\delta^{-1}} \to 1 - e^{-2a}$, equation (5.4.10) follows if one can prove that as δ goes to 0 the supremum S converges to 0 in probability: this is precisely the content of Lemma 5.4.4. This concludes the proof of Theorem 5.4.3. □

5.5 Numerical results

This section presents numerical simulations for the minimisation of a functional $J(\cdot)$ defined on the Sobolev space $H_0^1(\mathbb{R}) \subset C^0([0,1]) \subset L^2(0,1)$. Functions $x \in H_0^1([0,1])$ are continuous and satisfy $x(0) = x(1) = 0$. For a given real parameter

$\lambda > 0$, the functional $J : H_0^1([0,1]) \to \mathbb{R}$ is composed of two competitive terms, as follows:

$$J(x) \;=\; \frac{1}{2}\int_0^1 \left|\dot{x}(s)\right|^2 ds \;+\; \frac{\lambda}{4}\int_0^1 \left(x(s)^2 - 1\right)^2 ds. \tag{5.5.1}$$

The first term penalises functions that deviate from being flat, whilst the second term penalises functions that deviate from one in absolute value. Critical points of the functional $J(\cdot)$ solve the following Euler-Lagrange equation:

$$\ddot{x} + \lambda\, x(1 - x^2) = 0 \qquad \text{with} \qquad x(0) = x(1) = 0. \tag{5.5.2}$$

Clearly $x \equiv 0$ is a solution for all $\lambda \in \mathbb{R}^+$. If $\lambda \in (0, \pi^2)$ then this is the unique solution of the Euler-Lagrange equation and is the global minimizer of J. For each integer k there is a supercritical bifurcation at parameter value $\lambda = k^2\pi^2$. For $\lambda > \pi^2$ there are two minimizers, both of one sign and one being minus the other. The three different solutions of (5.5.2) which exist for $\lambda = 2\pi^2$ are displayed in Figure 5.5, at which value the zero (blue dotted) solution is a saddle point, and the two green solutions are the global minimizers of J. These properties of J are overviewed in, for example, [Hen81]. We will show how these global minimizers can emerge from an algorithm whose only ingredients are an ability to evaluate Ψ and to sample from the Gaussian measure with Cameron-Martin norm $\int_0^1 |\dot{x}(s)|^2 ds$. We emphasize that we are not advocating this as the optimal method for solving the Euler-Lagrange equations (5.5.2). We have chosen this example for its simplicity, in order to illustrate the key ingredients of the theory developed in this chapter.

The P-RWM algorithm to minimize J given by (5.5.1) is implemented on $L^2([0,1])$. Recall from section 5.1 that the Gaussian measure $\mathrm{N}(0, C)$ may be identified by finding the covariance operator for which the $H_0^1([0,1])$ norm $\|x\|_C^2 \stackrel{\text{def}}{=} \int_0^1 \left|\dot{x}(s)\right|^2 ds$ is the Cameron-Martin norm. In [HSVW05] it is shown that the Wiener bridge measure $\mathbb{W}_{0\to 0}$ on $L^2([0,1])$ has precisely this Cameron-Martin norm; indeed it is demonstrated that C^{-1} is the densely defined operator $-\frac{d^2}{ds^2}$ with domain $D(C^{-1}) = H^2([0,1]) \cap H_0^1([0,1])$. In this regard it is also instructive to adopt the physicists viewpoint that

$$\mathbb{W}_{0\to 0}(dx) \propto \exp\Big(-\frac{1}{2}\int_0^1 \left|\dot{x}(s)\right|^2 ds\Big)\, dx$$

although, of course, there is no Lebesgue measure in infinite dimensions. Using an

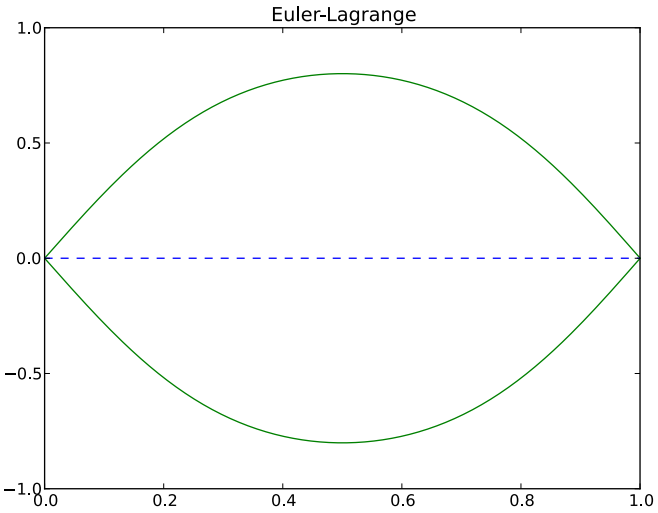


Figure 5.1: The three solutions of the Euler-Lagrange equation (5.5.2) for $\lambda = 2\pi^2$. Only the two non-zero solutions are global minima of the functional $J(\cdot)$. The dotted solution is a local maximum of $J(\cdot)$.

integration by parts, together with the boundary conditions on $H_0^1([0,1])$, then gives

$$\mathbb{W}_{0\to 0}(dx) \propto \exp\left(\frac{1}{2}\int_0^1 x(s)\frac{d^2x}{ds^2}(s)\,ds\right)dx$$

and the inverse of C is clearly identified as the differential operator above. See [CH06] for a basic discussion of the physicists viewpoint on Wiener measure. For a given temperature parameter τ the Wiener bridge measure $\mathbb{W}_{0\to 0}^{\tau}$ on $L^2([0,1])$ is defined as the law of $\{\sqrt{\tau}\,W(t)\}_{t\in[0,1]}$ where $\{W(t)\}_{t\in[0,1]}$ is a standard Brownian bridge on $[0,1]$ drawn from $\mathbb{W}_{0\to 0}$.

The posterior distribution $\pi^{\tau}(dx)$ is defined by the change of probability formula

$$\frac{d\pi^{\tau}}{d\mathbb{W}_{0\to 0}^{\tau}}(x) \;\propto\; e^{-\Psi(x)} \qquad \text{with} \qquad \Psi(x) = \frac{\lambda}{4}\int_0^1 \left(x(s)^2 - 1\right)^2 ds.$$

Notice that $\pi_0^{\tau}\big(H_0^1([0,1)\big) = \pi^{\tau}\big(H_0^1([0,1)\big) = 0$ since a Brownian bridge is almost surely not differentiable anywhere on $[0,1]$. It is for this reason that the algorithm is implemented on $L^2([0,1])$ even though the functional $J(\cdot)$ is defined on the Sobolev space $H_0^1([0,1])$. In terms of assumptions 3.1.1(1) we have $\kappa = 1$ and the measure π_0^{τ} is supported on $\mathcal{H}^r$ if and only if $r < \frac{1}{2}$. note also that $H_0^1([0,1]) = \mathcal{H}^1$. Assumption

3.1.1(2) is satisfied for any choice $s \in [\frac{1}{4}, \frac{1}{2})$ because $\mathcal{H}^s$ is embedded into $L^4([0,1])$ for $s \geq \frac{1}{4}$. We add here that assumptions 3.1.1(3-4) do not hold globally, but only locally on bounded sets, but the numerical results below will indicate that the theory developed in this chapter is still relevant and could be extended to nonlocal versions of assumptions 3.1.1(3-4), with considerable further work.

Following section 5.2.1, the P-RWM Markov chain at temperature $\tau > 0$ and time discretization $\delta > 0$ proposes moves from x to y according to

$$y = (1 - 2\delta)^{\frac{1}{2}} x \; + \; (2\delta\tau)^{\frac{1}{2}} \xi$$

where $\xi \in C([0,1], \mathbb{R})$ is a standard Brownian bridge on $[0,1]$. The move $x \to y$ is accepted with probability $\alpha^\delta(x,\xi) = 1 \wedge \exp\left(-\tau^{-1}[\Psi(y) - \Psi(x)]\right)$. Figure 5.5 displays the convergence of the Markov chain $\{x^{k,\delta}\}_{k\geq 0}$ to a minimiser of the functional $J(\cdot)$. Note that this convergence is not shown with respect to the space $H_0^1([0,1])$ on which J is defined, but rather in $L^2([0,1])$; indeed $J(\cdot)$ is almost surely infinite when evaluated at samples of the P-RWM algorithm, precisely because $\pi_0^\tau\left(H_0^1([0,1))\right) = 0$, as discussed above.

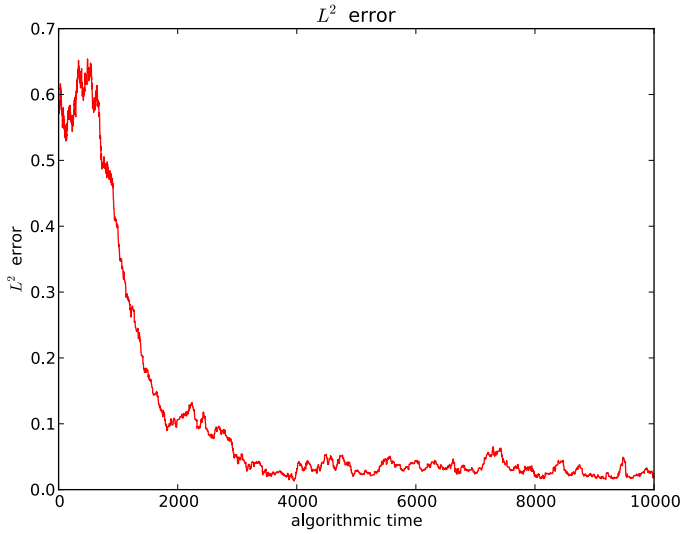


Figure 5.2: P-RWM parameters: $\lambda = 2\pi^2$, $\delta = 1.10^{-2}$, $\tau = 1.10^{-2}$. The algorithm is started at the zero function, $x^{0,\delta}(t) = 0$ for $t \in [0,1]$. After a transient phase, the algorithm fluctuates around a global minimiser of functional $J(\cdot)$. The L^2 error $\|x^{k,\delta} - (\text{minimiser})\|_{L^2}$ is plotted as a function of the algorithmic time k.

Of course the algorithm does not converge *exactly* to a minimiser of $J(\cdot)$, but fluctuates in a neighbourhood of it. As described in the introduction of this section,

in a finite dimensional setting the target probability distribution π^τ has Lebesgue density proportional to $\exp\big(-\tau^{-1}J(x)\big)$. This intuitively shows that the size of the fluctuations around the minimum of the functional $J(\cdot)$ are of size proportional to $\sqrt{\tau}$. Figure 5.5 shows this phenomenon on log-log scales: the asymptotic mean error $\mathbb{E}\big[\|x-(\text{minimiser})\|_2\big]$ is displayed as a function of the temperature τ. Figure 5.6 illustrates Theorem 5.4.3. One can observe the path $\{v^\delta(t)\}_{t\in[0,T]}$ for a finite time step discretization parameter δ as well as the limiting path $\{v(t)\}_{t\in[0,T]}$ that is solution of the differential equation (5.4.3).

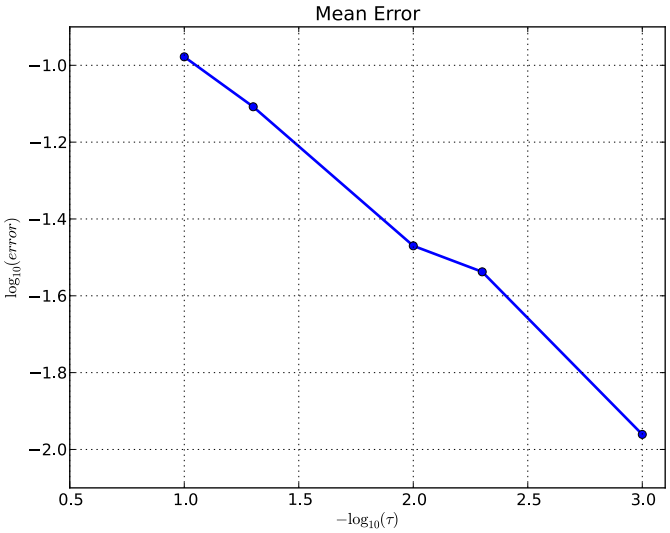


Figure 5.3: Mean error $\mathbb{E}\big[\|x-(\text{minimiser})\|_2\big]$ as a function of the temperature τ.

5.6 Conclusion

There are different perspectives on the material contained in this chapter, including optimization, numerical analysis and statistics. We now detail these perspectives.

- **Optimization:** We have demonstrated a class of algorithms to minimize the functional J given by (5.1.1). The assumptions 3.1.1 encode the intuition that the quadratic part of J dominates. Under these assumptions we study the properties of an algorithm which requires only the evaluation of Ψ and the ability to draw samples from Gaussian measures with Cameron-Martin norm given by the quadratic part of J. We demonstrate that, in a certain parameter limit, the algorithm behaves like a noisy gradient flow for the functional J and that, furthermore, the size of the noise can be controlled systematically.

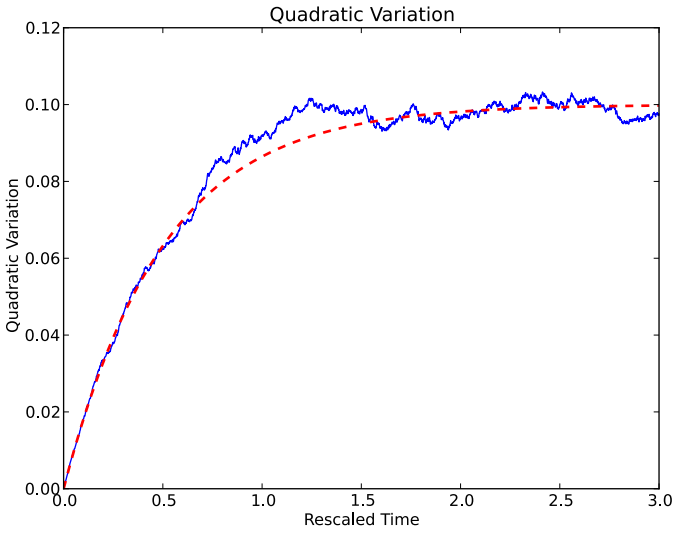


Figure 5.4: P-RWM parameters: $\lambda = 2\,\pi^2$, $\tau = 1.10^{-1}$, $\delta = 1.10^{-3}$ and the algorithm starts at $x^{k,\delta} = 0$. The rescaled quadratic variation process (full line) behaves as the solution of the differential equation (dotted line), as predicted by Theorem 5.4.3. The quadratic variation converges to τ, as described by Proposition 5.4.2.

Thus we have constructed a simulated annealing algorithm on Hilbert space, and connected this to a diffusion process (SDE), a connection made in finite dimensions in [GH86]. The advantage of constructing algorithms on Hilbert space is that they are robust to finite dimensional approximation. We turn to this point in the next bullet.

- **Numerical Analysis:** The algorithm that we use is a Metropolis-Hastings method with an Ornstein-Uhlenbeck proposal which we refer to here as P-RWM. The proposal takes the form for $\xi \sim \mathrm{N}(0, C)$,

$$y = \big(1 - 2\delta\big)^{\frac{1}{2}} x + \sqrt{2\delta\tau}\xi$$

given in (5.1.4). As described in the introduction, the proposal is constructed in such a way that the algorithm is defined on infinite dimensional Hilbert space and may be viewed as a natural analogue of a random walk Metropolis-Hastings method for measures defined via density with respect to a Gaussian. Let us contrast this with the standard random walk method S-RWM with proposal

$$y = x + \sqrt{2\delta\tau}\xi.$$

Although the proposal for S-RWM differs only through a multiplicative factor in the systematic component, and thus implementation of either is practically identical, the S-RWM method is not defined on infinite dimensional Hilbert space as mentioned in the Introduction. This turns out to matter if we compare both methods when applied in $\mathbb{R}^N$ for $N \gg 1$, as would occur if approximating a problem in infinite dimensional Hilbert space: in this setting the S-RWM method requires the choice $\delta = \mathcal{O}(N^{-1})$ to see the diffusion (SDE) limit [MPS11] and so requires $\mathcal{O}(N)$ steps to see $\mathcal{O}(1)$ decrease in the objective function, or to draw independent samples from the target measure; in contrast the P-RWM produces a diffusion limit for $\delta \to 0$ independently of N and so requires $\mathcal{O}(1)$ steps to see $\mathcal{O}(1)$ decrease in the objective function, or to draw independent samples from the target measure. Mathematically this last point is manifest in the fact that we may take the limit $N \to \infty$ (and work on the infinite dimensional Hilbert space) followed by the limit $\delta \to 0$.

- **Statistics:** The target distribution π^τ can be viewed as a posterior distribution in the context of Bayesian inverse problems, and nonparametric regression in particular. Here the goal is to perform statistical estimation of an unknown function from observations subject to noise. The measure π_0^τ is the prior distribution which quantifies the information the statistician has (from experts, knowledge about the regularity of the function, etc) before observing the data. The functional Ψ denotes the log-likelihood. S-RWM algorithms of the kind discussed in this chapter are routinely implemented in applied statistics for drawing samples from the measure π^τ. Our results demonstrate that it is immensely beneficial to modify these algorithms to the P-RWM algorithm, which we have derived using the "optimize then discretize" point of view; this will result in an $\mathcal{O}(N)$ increase in the efficiency of the algorithm when implemented on finite dimensional approximation spaces of dimension N.

Chapter 6

Random walk on ridge densities

This chapter is joint work with Alex Beskos and Gareth Roberts and is based on the paper [BRT13].

6.1 Introduction

In often happens in applied probability that one needs to explore a target distribution π that is concentrated on a very narrow subset of a state space $\mathcal{X}$. This informally means that there exists a small subset $A \subset \mathcal{X}$ where the mass concentrates in the sense that $\pi(A) \approx 1$; smallness can be given several interpretations. In a discrete setting, this is very common. Examples include the sampling of contingency tables, the sampling of a q-colouring of a graph, the Ising-Potts model on a lattice at low temperature.

In this chapter, we consider the continuous setting where the target distribution π lives in the n-dimensional euclidean space $\mathcal{X} = \mathbb{R}^n$ and concentrates on the neighbourhood of a low dimensional manifold $\mathcal{M}$; this means that there exists $\varepsilon \ll 1$ such that the ε-neighbourhood $A_\varepsilon := \{x \in \mathbb{R}^n : d(x, \mathcal{M}) < \varepsilon\}$ of the manifold $\mathcal{M}$ verifies $\pi(A_\varepsilon) \approx 1$. A Random Walk Metropolis (RWM) Markov chain will tend to walk along the manifold. The purpose of this chapter is to quantify this behaviour. While it has often been suggested in the literature to use tempering or adaptive methods to handle these ridges [Nea01; HST01], they remain a celebrated challenge for new Monte Carlo methods [CMMR12]. These strong geometric features commonly occur in non-identifiable models. A frequently occurring situation is the following. An unknown vector $x \in \mathbb{R}^n$ is measured through a possibly non-linear function $f : \mathbb{R}^n \to \mathbb{R}^d$ with $d < n$. Since the dimension d of the observation is strictly less than the dimension n of the unknown data $x \in \mathbb{R}^n$, there is generally no

hope to perfectly reconstruct the data: information is lost through the low dimensional measurement. This leads to non-identifiability issues. An additive Gaussian noise of intensity ε might then corrupt the measurement $f(x)$. The noisy and low dimensional observation $y \in \mathbb{R}^d$ of the unknown data $x \in \mathbb{R}^n$ can thus be modelled by the equation $y = f(x) + \xi$ with $\xi \overset{\mathcal{D}}{\sim} \mathrm{N}(0, \varepsilon^2 I_d)$. In the absence of noise (*i.e* the case $\varepsilon = 0$) and without prior knowledge on the data, any vector belonging to the set $\mathcal{M} := \{z \in \mathbb{R}^n : f(z) = y\}$ is equally likely to have given rise to the observation y. Under mild assumptions [1] on the measurement function $f : \mathbb{R}^n \to \mathbb{R}^d$, the subset $\mathcal{M}$ is a manifold of dimension $n - d$. The dimension of the manifold $\mathcal{M}$ can informally be thought of as the dimension of the non-identifiability. By imposing a prior distribution π_0 on the unknown data, we create weak identifiability. Compressed sensing [Don06; Can08], 'large p small n' problems [FHT09; Tib96] and Bayesian approach to inverse problems [Stu10; Fit91] can be seen as variations on the same theme.

In this chapter we focus on the limiting regime when the thickness ε of the neighbourhood A_ε of the limiting manifold $\mathcal{M}$ converges to zero. To this end, we introduce a family π_ε of distributions on $\mathbb{R}^n$ and a limiting manifold $\mathcal{M}$. The distribution π_ε concentrates on neighbourhood of thickness ε of $\mathcal{M}$. A rigorous definition is given below. We use the Random Walk Metropolis (RWM) algorithm to explore π_ε. The influence of the size of the jumps is analysed by adopting the Expected Squared Jumping Distance ESJD as measure of efficiency. The main finding (see Theorem 6.2.1 and discussion that follows) is that in the majority of the cases, in order to explore π_ε it is optimal to choose the size of the jumps of the same order of magnitude as the thickness ε. For this choice, we prove a diffusion limit result (Theorem 6.2.3). This gives quantitative estimates on the complexity of the RWM algorithm when applied to target concentrating near a manifold. For simplicity, all the rigorous results are proved for the case where the manifold $\mathcal{M}$ is flat. The diffusion limits that are proven in this section are local in nature; we thus believe that analogous results hold for general manifold since any smooth manifold can be regarded as flat (first order approximation) in the neighbourhood of any point. These conjectures for the general case are discussed with numerical illustrations in section 6.6. To the best of our knowledge, this is the first time that a diffusion approximation for MCMC *trajectories* leads to a diffusion limit with non-constant volatility. The related article [JLM12] investigates the diffusion limit of an empirical system of particles where each coordinate can be seen as a one dimensional MCMC algorithm. The interaction is done through the global

[1] e.g. assumptions that ensure that the implicit function theorem holds.

accept-reject mechanism common to the whole system. The *measure valued* limiting diffusion of [JLM12] also has a non constant diffusion coefficient; while related, their result describing a limiting flow of measures is very different from our diffusion limit that described the diffusive behaviour of a single MCMC trajectory. Our proof is based on a time-scale separation argument that borrows ideas from [NRY12].

6.2 Main Results

6.2.1 Distributions concentrating near a manifold

As explained above, the rigorous results are proved in the case when the manifold is flat *i.e.* an affine subspace of dimension n_x of $\mathbb{R}^n = \mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$. We will model this scenario as follows. For each $\varepsilon > 0$ we consider the target distribution $\pi_\varepsilon : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \mapsto \mathbb{R}$ with density with respect to the n-dimensional Lebesgue measure

$$\pi_\varepsilon(x, y) = \pi(x)\, \pi_\varepsilon(y|x) = e^{A(x)}\, e^{B(x,y/\varepsilon)}/\varepsilon^{n_y} \ , \tag{6.2.1}$$

with $\varepsilon > 0$ being 'small' and $n_x + n_y = n$. The x-marginal has density $e^{A(x)}$ independently of the parameter $\varepsilon > 0$. This is a scaled version of the probability distribution π_1 with density $e^{A(x)}e^{B(x,y)}$. Notice that as $\varepsilon \to 0$, the sequence of distributions π_ε concentrates on the linear subspace $\mathcal{M} := \{(x, y) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} : y = 0\}$. The linear subspace $\mathcal{M}$ has dimension n_x. As described in the introduction, the dimension n_x can be thought of as the dimension of the non-identifiability; the dimension n_y can be thought of as the effective dimension of the (possibly noisy) observation. The term $B(x, y/\varepsilon)$ shows that the parameter ε can be thought of as the thickness of the neighbourhood of $\mathcal{M}$ where the distribution π_ε concentrates. In the situation mentioned in the introduction, the parameter ε also describes the intensity of the noise. To obtain samples from π_ε we consider the Random-Walk Metropolis (RWM) algorithm proposing moves

$$\begin{pmatrix} X'_\varepsilon \\ Y'_\varepsilon \end{pmatrix} = \begin{pmatrix} X_\varepsilon \\ Y_\varepsilon \end{pmatrix} + h(\varepsilon) \begin{pmatrix} Z_x \\ Z_y \end{pmatrix} \ , \tag{6.2.2}$$

for scaling factor $h(\varepsilon)$ and Gaussian noise $(Z_x, Z_y) \overset{\mathcal{D}}{\sim} \mathrm{N}(0, I_{n_x+n_y})$. The scaling factor $h(\varepsilon)$ describes the size of the jumps of the RWM algorithm. The influence of $h(\varepsilon)$ on the efficiency of the algorithm is analysed in section 6.2.2. When the context is clear, we write (X, Y) instead of $(X_\varepsilon, Y_\varepsilon)$. We introduce the rescaled coordinate

U_ε and the associate rescaled proposal U'_ε,

$$U_\varepsilon = Y_\varepsilon/\varepsilon \quad \text{and} \quad U'_\varepsilon = U_\varepsilon + \tfrac{h(\varepsilon)}{\varepsilon} Z_y \ .$$

Notice that if $(X_\varepsilon, Y_\varepsilon) \overset{\mathcal{D}}{\sim} \pi_\varepsilon$ then $(X_\varepsilon, U_\varepsilon) := (X_\varepsilon, Y_\varepsilon/\varepsilon)$ is distributed as π_1 with density $e^{A(x)}e^{B(x,u)}$. In the sequel, we interchangeably use the random variables $(X_\varepsilon, Y_\varepsilon)$ and $(X_\varepsilon, U_\varepsilon)$, keeping in mind that $U_\varepsilon := Y_\varepsilon/\varepsilon$. To finish the description of the MCMC algorithm, we need to choose an accept-reject function F. One can choose any $(0,1]$-valued function F satisfying the reversibility condition

$$e^r F(-r) = F(r) \tag{6.2.3}$$

for all $r \in \mathbb{R}$. We choose to work with a general accept-reject function F for conceptual clarity and to emphasise that our results do not depend on the usual Metropolis-Hastings function $F_{\mathrm{MH}}(r) = \min(1, e^r)$. The move $(X, Y) \mapsto (X', Y')$, or equivalently $(X, U) \mapsto (X', U')$, is then accepted with probability $F\Big(\log \frac{\pi_\varepsilon(X',Y')}{\pi_\varepsilon(X,Y)}\Big)$. For the usual Metropolis-Hastings accept-reject mechanism, the acceptance probability indeed also reads $a(X, U, Z_x, Z_y) = \min\big(1, \frac{\pi_\varepsilon(X',U')}{\pi_\varepsilon(X,U)}\big)$. For target density (6.2.1) the acceptance probability reads

$$a(X, U, Z_x, Z_y) = F\Big(A(X') - A(X) + B(X', U') + B(X, U)\Big). \tag{6.2.4}$$

Notice that any function F satisfying the reversibility condition (6.2.3) is dominated by the Metropolis-Hasting function F_{MH} in the sense that the inequality $F(r) \leq F_{\mathrm{MH}}(r)$ holds for any $r \in \mathbb{R}$.

6.2.2 Expected Squared Jumping Distance

In this section we choose to work with the Expected Squared Jumping Distance (ESJD) as an index of the efficiency of MCMC algorithms, as it allows for transparent, explicit calculations. See section 2.3.1 and [RR01; BRS09; PG10] and references therein for motivations behind the ESJD. Since only the x-coordinate matters in the limiting regime $\varepsilon \to 0$ (because the y-coordinates is of order $\mathcal{O}(\varepsilon)$ under π_ε), only the x-coordinate is taken into account in the definition of the ESJD. Consequently, we consider instead the modified ESJD instead,

$$\mathrm{ESJD}(\varepsilon) = \mathbb{E}\Big[\big\|X_{k+1} - X_k\big\|^2\Big].$$

The expectation is taken at stationarity $(X_k, Y_k) \overset{\mathcal{D}}{\sim} \pi_\varepsilon$. We analyse the asymptotic behaviour of the ESJD for different choices of scaling factor $h(\varepsilon) = \varepsilon^\gamma$.

Theorem 6.2.1. (Asymptotic analysis of the ESJD) *Let $\gamma \geq 0$ be a nonnegative exponent. Assume that the scaling factor is of the form $h(\varepsilon) = \varepsilon^\gamma$. In the limiting regime $\varepsilon \to 0$ we have*

$$\mathit{ESJD}(\varepsilon) \asymp \varepsilon^{2\gamma + n_y \max(0, 1-\gamma)}.$$

Proof. The proof is routine and thus only sketched. If $\gamma > 1$ the mean acceptance probability converges to 1 (since the proposed jumps are of size $\mathcal{O}(\varepsilon^\gamma)$ and π_ε concentrates on a neighbourhood of thickness ε around $\mathcal{M}$) and consequently $\mathrm{ESJD}(\varepsilon) \sim h(\varepsilon)^2$. In other words, for $\gamma > 1$ we have $\mathrm{ESJD}(\varepsilon) \sim \varepsilon^{2\gamma}$. In general, the $\mathrm{ESJD}(\varepsilon)$ is equivalent to $h(\varepsilon)^2 \, \mathbb{E}_{\pi_\varepsilon}\Big[F\Big(B(X, \varepsilon^{-1}Y + \varepsilon^{-1}h(\varepsilon)Z_y) - B(X, \varepsilon^{-1}Y)\Big)\Big]$, which is proportional to the integral

$$h(\varepsilon)^{2-d_y}\varepsilon^{d_y} \iiint_{x,y,z} F\Big(B(x, y+z) - B(x, y)\Big) e^{-\frac{\|z\|^2}{2(h(\varepsilon)/\varepsilon)^2}} e^{A(x)+B(x,y)} \, dx \, dy \, dz.$$

For $\gamma < 1$ we have $h(\varepsilon)/\varepsilon \to \infty$ and the triple integral converges to a constant that does not depend on ε. This shows that for $\gamma < 1$ we have $\mathrm{ESJD}(\varepsilon) \asymp h(\varepsilon)^{2-d_y}\varepsilon^{d_y} = \varepsilon^{d_y + \gamma(2-d_y)}$. For $\gamma = 1$, the situation is even simpler since the triple integral does not depend on ε. □

The behaviour of $\varepsilon^{2\gamma + n_y \max(0,1-\gamma)}$ depends on the dimension n_y of the y-coordinate. Maximisation of the ESJD is equivalent to minimisation of the quantity $2\gamma + n_y \max(0, 1-\gamma)$.

- For $n_y = 1$, the optimal exponent is $\gamma_* = 0$ and in this case we have

$$\mathrm{ESJD}(\varepsilon) \asymp \varepsilon.$$

- For $n_y = 2$, any exponent $0 \leq \gamma_* \leq 1$ leads to the asymptotics

$$\mathrm{ESJD}(\varepsilon) \asymp \varepsilon^2.$$

- For $n_y \geq 3$, the optimal exponent is $\gamma_* = 1$ and in this case we have

$$\mathrm{ESJD}(\varepsilon) \asymp \varepsilon^2.$$

An important corollary is that for $n_y \geq 3$, it is optimal (if the ESJD is adopted as a measure of effiviency) to choose a scaling factor $h(\varepsilon)$ of order $\mathcal{O}(\varepsilon)$. In other words, for $n_y \geq 3$ (*i.e.* when the effective dimension of the observation is at least equal to three) it is optimal to choose the jump size $h(\varepsilon)$ of the same order of magnitude as the thickness ε. To go further in this direction, we analyse the behaviour of the algorithm when the scaling factor is of the form $h(\varepsilon) = \ell\,\varepsilon$ for some tuning parameter $\ell > 0$. One can verify that the asymptotic behaviour of the ESJD as a function of the parameter $\ell > 0$ is given by

$$\lim_{\varepsilon \to 0} \frac{\mathrm{ESJD}(\varepsilon, \ell)}{\varepsilon^2} \;=\; \ell^2\, \mathbb{E}\Big[F\big(B(X, U + \ell\, Z_Y) - B(X, U)\big)\Big]$$

where $(X, U) \overset{\mathcal{D}}{\sim} \pi_1$ is the probability distribution with density $e^{A(x)} e^{B(x,u)}$ and $Z_Y \overset{\mathcal{D}}{\sim} \mathrm{N}(0, I_{n_y})$. There is typically no closed form available for the optimal tuning parameter $\ell_* = \mathrm{argmax}_\ell \big\{\ell \mapsto \ell^2\, \mathbb{E}\big[F\big(B(X, U + \ell\, Z_Y) - B(X, U)\big)\big]\big\}$. Contrary to previous optimality results [RGG97; RR98; RR01; Béd07], the value of ℓ_* depends on the form of the target distribution π_ε.

6.2.3 Diffusion limit

As described in section 6.2.2, for dimension $n_y \geq 3$ it is optimal to choose a scaling factor $h(\varepsilon)$ of order $\mathcal{O}(\varepsilon)$. To go further in this direction, we study in this section the behaviour of the RWM algorithm, as $\varepsilon \to 0$, for a scaling factor of the form $h(\varepsilon) = \ell\,\varepsilon$ where $\ell > 0$ is a tuning parameter. In order to state our main result, it is useful to introduce the quantity

$$a_0(x, \ell) = \int_{\mathbb{R}^{n_y}} \mathbb{E}\Big[F\Big(B(x, u + \ell\; Z_y) - B(x, u)\Big)\Big]\, e^{B(x,u)}\, du \tag{6.2.5}$$

as well as the time scale $T(\varepsilon) = \varepsilon^{-2}$. The quantity $a_0(x, \ell)$ is the limiting acceptance probability, as $\varepsilon \to 0$, of the RWM algorithm when conditioned on the x-coordinate. Indeed, one can verify that $a_0(x, \ell)$ can also be expressed as

$$a_0(x, \ell) = \lim_{\varepsilon \to 0} \mathbb{E}_{\pi_\varepsilon}\big[a(X_\varepsilon, U_\varepsilon, Z_x, Z_y)\,|X_\varepsilon = x\big].$$

As it will become clear from our diffusion limit analysis, the time scale $T(\varepsilon)$ is the natural time scale on which the x-coordinate process $\{X_{\varepsilon,k}\}_{k \geq 0}$ evolves. Our main result states that the accelerated processes

$$\widetilde{X}_{\varepsilon,t} := X_{\varepsilon, \lfloor t \cdot T(\varepsilon) \rfloor} \tag{6.2.6}$$

converges weakly, as $\varepsilon \to 0$, to a non-trivial diffusion process. For this reason, $T(\varepsilon)$ is called 'diffusive time scale' in the sequel. For a density π on $\mathbb{R}^n$ and volatility function $\sigma : \mathbb{R}^n \to (0; \infty)$ we introduce the function $\text{drift}(\pi, \sigma^2) : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\text{drift}(\pi, \sigma^2) : \quad x \mapsto \frac{1}{2}\Big(\sigma^2 \nabla \log \pi(x) + \nabla \sigma^2(x)\Big).$$

Under mild assumptions [2] on the density π and the volatility function σ, the function $\text{drift}(\pi, \sigma^2)$ is such that the diffusion process

$$dD_t = \text{drift}(\pi, \sigma^2)(D_t)\, dt + \sigma(D_t)\, dW_t$$

is reversible with respect to the probability distribution π. The case $\sigma \equiv \text{Cst}$ corresponds to the Langevin diffusion $dD = \frac{\sigma^2}{2} \nabla \log \pi \, dt + \sigma \, dW$. For our main scaling limit to hold, we assume regularity and growth assumptions on the functions $A : \mathbb{R}^{n_x} \to \mathbb{R}$ and $B : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \to \mathbb{R}$. These conditions are mainly technical.

Assumptions 6.2.2. (Growth and Regularity Assumptions on π)
The first two derivatives of the functions $A : \mathbb{R}^{n_x} \to \mathbb{R}$ and $B : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \to \mathbb{R}$ are bounded by a polynomial of degree $p \geq 1$. Moreover, there exists an exponent $\eta > 0$ such that the following moment condition holds,

$$\mathbb{E}_{\pi_1}\big[\, (1 + \|X\| + \|U\|)^{2p+\eta} \,\big] \;<\; \infty, \tag{6.2.7}$$

where $(X, U) \overset{\mathcal{D}}{\sim} \pi_1$ has density $e^{A(x)} e^{B(x,u)}$.

Assumptions 6.2.2 implies the existence of an integer $p \geq 1$ such that the norm of the quantities $A(x)$ and $B(x, u)$ and their first two derivatives are less than a constant multiple of $\big(1 + \|x\| + \|u\|\big)^p$. This estimate is used at several places in the proof of our main result. The main theorem of this section is the following. The proof is described in section 6.3.

Theorem 6.2.3. *Let $T > 0$ be a fixed finite time horizon. Assume that assumptions 6.2.2 hold and that the RWM algorithm is started in stationarity, $(X_{\varepsilon,0}, Y_{\varepsilon,0}) \overset{\mathcal{D}}{\sim} \pi_\varepsilon$. As $\varepsilon \to 0$, the sequence of accelerated processes $\{\widetilde{X}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly in the Skorohod space $D([0, T], \mathbb{R}^{n_x})$ to the diffusion process $\{D_t\}_{t\in[0,T]}$ specified as the solution of the stochastic differential equation*

$$dD_t = \mathit{drift}\big(\pi, \sigma^2\big)(D_t)\, dt + \sigma(D_t)\, dW_t \tag{6.2.8}$$

[2] e.g. the functions $\text{drift}(\pi, \sigma^2) : \mathbb{R}^n \to \mathbb{R}^n$ and σ are Lipschitz and linearly bounded

where W is a standard Brownian motion in $\mathbb{R}^{n_x}$. The local volatility function is given by $\sigma^2(x,\ell) = \ell^2\, a_0(x,\ell)$. The initial distribution is $D_0 \overset{\mathcal{D}}{\sim} \pi$.

The diffusion (6.2.8) is ergodic and reversible with respect to π. The diffusive time scale $T(\varepsilon) = \varepsilon^{-2}$ shows that the algorithmic complexity of the RWM grows as $\mathcal{O}(\varepsilon^{-2})$ as the thickness ε goes to zero. The limiting rescaled ESJD can directly be read from the volatility coefficient of the limiting diffusion (6.2.8),

$$\lim_{\varepsilon \to 0} \frac{\mathrm{ESJD}(\varepsilon, \ell)}{\varepsilon^2} = \mathbb{E}\big[\sigma^2(X, \ell)\big]$$

where $X \overset{\mathcal{D}}{\sim} \pi$ and $\sigma^2(x,\ell) = \ell^2\, a_0(X,\ell)$. In the case where the function $(x,y) \mapsto B(x,y)$ does not depend on the x-coordinate, the limiting acceptance probability a_0 does not depend on the local position x anymore, $a_0(x,\ell) = a_0(\ell)$. In this case the optimal value for the parameter ℓ is given by $\ell_* = \mathrm{argmax}\ \ell^2\, a_0(\ell)$, which leads to a 0.234-type optimality result as described in [RGG97]. In general, the optimisation of the limiting ESJD is difficult. The Dirichlet form [Fuk80] associated to the diffusion (6.2.8) reads

$$\mathcal{D}(\varphi) := \frac{1}{2} \int_{\mathbb{R}^{n_x}} \big\|\nabla\varphi(x)\big\|^2\, \sigma^2(x,\ell)\ \pi(dx).$$

The spectral gap of the diffusion (6.2.8) equals $\lambda = \sup_\varphi\ \mathcal{D}(\varphi)$ where the supremum runs over the class of smooth test functions satisfying $\pi(\varphi) = 0$ and $\pi(\varphi^2) = 1$. The maximisation of the ESJD is equivalent to maximising the Dirichlet form over the class of affine functions. In general, the maximisation of the spectral gap and the maximisation of the ESJD thus lead to different answers.

In an attempt to reconcile the different notions of optimality, we adopt slightly more general proposals. The variance of the proposals is allowed to depend on the current position; the tuning parameter $\ell = \ell(x) > 0$ is now allowed to depend on the x-coordinate,

$$\begin{pmatrix} X'_\varepsilon \\ Y'_\varepsilon \end{pmatrix} = \begin{pmatrix} X_\varepsilon \\ Y_\varepsilon \end{pmatrix} + \ell(x)\, \varepsilon \begin{pmatrix} Z_x \\ Z_y \end{pmatrix}. \tag{6.2.9}$$

In other words, when the RWM Markov chain stands at $(x,y) \in \mathbb{R}^n$, a Gaussian jump of size $\ell(x)\,\varepsilon$ is proposed. We now state assumptions on the function $x \mapsto \ell(x)$ that allows diffusion limit results to holds.

Assumptions 6.2.4. (Regularity Assumptions on $x \mapsto \ell(x)$)

The function ℓ is positive, bounded away from zero and infinity. The first two derivatives of ℓ are also bounded.

Under the regularity assumption 6.2.4 on the function $x \mapsto \ell(x)$ the analogue of Theorem 6.2.3 holds. We choose to work in this limited setup so that the proof of the next theorem is a straightforward adaption of Theorem 6.2.3. The accelerated version (6.2.6) of the x-coordinate process converges to a diffusion process.

Theorem 6.2.5. *Let $T > 0$ be a fixed finite time horizon. Assume that assumptions 6.2.2 and 6.2.4 hold and that the RWM algorithm is started in stationarity. As $\varepsilon \to 0$, the sequence of processes $\{\widetilde{X}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly in the Skorohod space $D([0,T],\mathbb{R}^{n_x})$ to the diffusion process $\{D_t\}_{t\in[0,T]}$ specified as the solution of the stochastic differential equation*

$$dD_t = \mathit{drift}\big(\pi,\sigma^2\big)(D_t)\,dt + \sigma(D_t)\,dW_t \tag{6.2.10}$$

where W is a standard Brownian motion in $\mathbb{R}^{n_x}$. The local volatility function is given by $\sigma^2(x) = \ell^2(x)\,a_0\big(x,\ell(x)\big)$. The initial distribution is $D_0 \overset{\mathcal{D}}{\sim} \pi$.

The only difference with Theorem 6.2.3 is the form of the volatility function σ. As before, the limiting distribution (6.2.10) is reversible with respect to π and the Dirichlet form reads

$$\mathcal{D}(\varphi) := \frac{1}{2}\int_{\mathbb{R}^{n_x}} \big\|\nabla\varphi(x)\big\|^2\,\ell^2(x) a_0\big(x,\ell(x)\big)\,\pi(dx).$$

Since the parameter $\ell = \ell(x)$ is a function of the x-coordinate, the optimal choice $\ell_*(x)$ for the tuning parameter ℓ is

$$\ell_*(x) := \mathrm{argmax}_{\ell>0}\,\ell^2\,a_0(x,\ell). \tag{6.2.11}$$

As described in [RR12], the choice (6.2.11) maximises the ESJD, the spectral gap of the limiting diffusion (6.2.10) and the asymptotic variance of MCMC estimators.

6.3 Proof of Theorem 6.2.3

Let us first give a high-level description of the proof. We introduce an intermediate time scale $\widetilde{T}(\varepsilon) = \varepsilon^{-\gamma}$ and the sub-sampled process $\{(s_{\varepsilon,k}, v_{\varepsilon,k})\}_{k\geq 0}$ defined as

$$\begin{aligned}(S_{\varepsilon,0},\, S_{\varepsilon,1},\, S_{\varepsilon,2},\, \ldots) &= (X_{\varepsilon,0}, X_{\varepsilon,\lfloor\widetilde{T}(\varepsilon)\rfloor},\, X_{\lfloor\varepsilon,2\,\widetilde{T}(\varepsilon)\rfloor},\, \ldots)\\ (V_{\varepsilon,0},\, V_{\varepsilon,1},\, V_{\varepsilon,2},\, \ldots) &= (U_{\varepsilon,0}, U_{\varepsilon,\lfloor\widetilde{T}(\varepsilon)\rfloor},\, U_{\varepsilon,\lfloor 2\,\widetilde{T}(\varepsilon)\rfloor},\, \ldots).\end{aligned}$$

The value of the exponent $0 < \gamma < 2$ is discussed later. The intuition behind the time scale $\widetilde{T}(\varepsilon)$ is that on this scale the x-process evolves slowly (i.e. $S_{\varepsilon,k} \approx S_{\varepsilon,k+1}$) while the y-process has the time to mix (i.e. $V_{\varepsilon,k+1}$ is approximately independent from $V_{\varepsilon,k}$). The time scale $\widetilde{T}(\varepsilon)$ is intermediate between the original time scale (i.e. the non accelerated process) and the diffusive time scale $T(\varepsilon)$. We then prove that the sequence of accelerated processes

$$\widetilde{S}_{\varepsilon,t} = S_{\varepsilon,\lfloor t \cdot T(\varepsilon)/\widetilde{T}(\varepsilon) \rfloor} \tag{6.3.1}$$

converges weakly in $D([0,T],\mathbb{R}^{n_x})$ to the limiting diffusion (6.2.8). Since the value of $\widetilde{X}_{\varepsilon,t}$ is close to the value of $\widetilde{S}_{\varepsilon,t}$, the diffusion limit for the process $\{\widetilde{X}_{\varepsilon,t}\}_{t\in[0,T]}$ easily follows.

We now proceed to the rigorous proof of Theorem 6.2.3. To simplify the presentation, we assume that the accept-reject function F satisfying equation (6.2.3) is continuously differentiable with bounded first derivative and second derivative. In particular, F is a Lipschitz function. We denote by $\mathcal{L}$ the generator of the diffusion process in Theorem 6.2.3. The generator $\widetilde{\mathcal{L}}_\varepsilon$ of the process $\widetilde{S}_{\varepsilon,t}$ is

$$\begin{aligned} \widetilde{\mathcal{L}}_\varepsilon \varphi(x,u) &= \mathbb{E}\Big[\frac{\varphi(S_{\varepsilon,1}) - \varphi(S_{\varepsilon,0})}{\widetilde{T}(\varepsilon)/T(\varepsilon)} \,\Big|\, S_{\varepsilon,0} = x, V_{\varepsilon,0} = u \Big] \\ &\equiv \frac{1}{\widetilde{T}(\varepsilon)} \mathbb{E}\Big[\sum_{j=0}^{\lfloor \widetilde{T}(\varepsilon) \rfloor - 1} \mathcal{L}_\varepsilon \varphi(X_{\varepsilon,j}, U_{\varepsilon,j}) \,\Big|\, X_{\varepsilon,0} = x, U_{\varepsilon,0} = u \Big] \end{aligned} \tag{6.3.2}$$

where $\mathcal{L}_\varepsilon$ is the one-step generator of the process $\widetilde{X}_{\varepsilon,t}$,

$$\mathcal{L}_\varepsilon \varphi(x,u) = \mathbb{E}\Big[\frac{\varphi(X_{\varepsilon,1}) - \varphi(X_{\varepsilon,0})}{1/T(\varepsilon)} \,\Big|\, X_{\varepsilon,0} = x, U_{\varepsilon,0} = u \Big] \,. \tag{6.3.3}$$

The second equality of equation (6.3.2) follows from the telescoping expansion $S_{\varepsilon,1} - S_{\varepsilon,0} = \sum_{j=0}^{\lfloor \widetilde{T}(\varepsilon) \rfloor - 1} (X_{\varepsilon,j+1} - X_{\varepsilon,j})$. In other words, the generator $\widetilde{\mathcal{L}}_\varepsilon$ is the average of the one-step generator $\mathcal{L}_\varepsilon$ over $\widetilde{T}(\varepsilon)$ steps. The process X_ε needs to be accelerated by a factor $T(\varepsilon)$ in order to observe a non trivial diffusion limit. Similarly, since the process S_ε is an accelerated (by a factor $\widetilde{T}(\varepsilon)$) version of X_ε,the process S_ε needs to be accelerated by a factor $T(\varepsilon)/\widetilde{T}(\varepsilon)$ in order to observe the same non trivial diffusion limit. For clarity, we now proceed in two steps. First, in section 6.3.1 we prove that the sequence $\{\widetilde{S}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly, as $\varepsilon \to 0$, to the limiting diffusion (6.2.8). We then explain in section 6.3.2 how this implies the convergence

of the sequence $\{\widetilde{X}_{\varepsilon,t}\}_{t\in[0,T]}$ to the same limiting diffusion, finishing the proof of Theorem 6.2.3.

6.3.1 The sequence $\widetilde{S}_{\varepsilon,t}$ converges weakly to the limiting diffusion (6.2.8)

To prove that the sequence of processes $\{\widetilde{S}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly in the Skorohod space $D([0,T],\mathbb{R}^{n_x})$ to the diffusion process (6.2.8), we use an approach based on generators. We prove that the sequence of generators $\widetilde{\mathcal{L}}_\varepsilon$ of the processes $\widetilde{S}_\varepsilon$ converges (in a sense made precise below) to the generator $\mathcal{L}$ of the limiting diffusion (6.2.8). Since the class $\mathcal{C}$ of smooth and compactly supported functions form a core for the generator $\mathcal{L}$ of the diffusion (6.2.8), to prove that the sequence of processes $\{\widetilde{S}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly to the diffusion process (6.2.8) it suffices to check that the following two conditions are satisfied.

1. The sequence of $\{\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t}\}_{t\in[0,T]}$ is relatively weakly compact under the appropriate topology. To this end, it suffices to prove that for any smooth and compactly supported test function $\varphi\in\mathcal{C}$ we have

$$\sup\Big\{\mathbb{E}_{\pi_1}\Big[\Big|\widetilde{\mathcal{L}}_\varepsilon\varphi(X,U)\Big|\Big] \;:\varepsilon\in(0,1)\Big\} \;<\; \infty. \tag{6.3.4}$$

   with $(X,U)\overset{\mathcal{D}}{\sim}\pi_1$. This is Equation 8.9 of Theorem 8.2 of [EK86]. This implicitly uses the fact that if the Markov chain is started at stationarity, $(\widetilde{S}_{\varepsilon,0},\widetilde{V}_{\varepsilon,0})\overset{\mathcal{D}}{\sim}\pi_1$, then for any fixed time $t>0$ we also have $(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t})\overset{\mathcal{D}}{\sim}(X,U)\overset{\mathcal{D}}{\sim}\pi_1$.

2. The sequence of generators $\widetilde{\mathcal{L}}_\varepsilon$ converges to the generator $\mathcal{L}$ of the limiting diffusion (6.2.8) in the sense that for any smooth and compactly supported test function $\varphi\in\mathcal{C}$ the following holds,

$$\lim_{\varepsilon\to0}\;\mathbb{E}_{\pi_1}\Big[\Big|\widetilde{\mathcal{L}}_\varepsilon\varphi(X,U)-\mathcal{L}\varphi(X)\Big|\Big] \;=\; 0 \tag{6.3.5}$$

   with $(X,U)\overset{\mathcal{D}}{\sim}\pi_1$. This is Equation 8.13 of Remark 8.3 of [EK86], again using the fact that if the Markov chain is started at stationarity then for any fixed time $t>0$ we also have $(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t})\overset{\mathcal{D}}{\sim}(X,U)\overset{\mathcal{D}}{\sim}\pi_1$

The conditions (6.3.4) and (6.3.5) are enough to guaranty the convergence of the sequence of processes $\{\widetilde{S}_{\varepsilon_t}\}_{t\in[0,T]}$ to the limiting diffusion (6.2.8) mainly because the class of test functions $\mathcal{C}$ is representative enough (i.e. it is a core of the generator $\mathcal{L}$ of the diffusion (6.2.8)) in order to characterise weak convergence. For more details

on this method of proof based on generators, see Chapter 4 of [EK86] and articles [Béd07; RGG97; RR98].

As a first step toward equations (6.3.4) and (6.3.5), we prove that for test function $\varphi \in \mathcal{C}$ the following limit exists, $\lim_{\varepsilon\to 0} \mathcal{L}_\varepsilon\varphi(x,u) = \mathcal{A}(\varphi,x,u)$. The limiting quantity $\mathcal{A}(\varphi,x,u)$ is not in general the generator of a Markov process. It is given by

$$\mathcal{A}(\varphi,x,u) = \ell^2\,\mathbb{E}\Big[F'(\delta B)\nabla_x\big\{A(x)+B(x,u+\ell Z_y)\big\}\Big]\,\nabla\varphi(x) + \frac{1}{2}\ell^2\,\mathbb{E}\Big[F(\delta B)\Big]\,\Delta\varphi(x) \tag{6.3.6}$$

where for notational convenience we have defined $\delta B = B(x,u+\ell Z_y) - B(x,u)$. The next proposition, whose proof is postponed to section 6.5.1, gives a quantitative rate for the convergence of $\mathcal{L}_\varepsilon\varphi(x,u)$ towards $\mathcal{A}(\varphi,x,u)$.

Proposition 6.3.1. *Let assumptions 6.2.2 be satisfied and $\varphi \in \mathcal{C}$ be a test function. The following identity holds*

$$\mathcal{L}_\varepsilon\varphi(x,u) = \mathcal{A}(\varphi,x,u) + \boldsymbol{e}_1(x,u,\varepsilon) \tag{6.3.7}$$

with the error term satisfying $\lim_{\varepsilon\to 0}\mathbb{E}\big|\boldsymbol{e}_1(X,U,\varepsilon)\big| = 0$ where $(X,U) \overset{\mathcal{D}}{\sim} \pi_1$.

Plugging estimate (6.3.7) into the telescoping equation (6.3.2), it follows that the generator $\widetilde{\mathcal{L}}_\varepsilon$ verifies

$$\widetilde{\mathcal{L}}_\varepsilon\varphi(x,u) = \frac{1}{\widetilde{T}(\varepsilon)}\,\mathbb{E}_{x,u}\Big[\sum_{j=0}^{\lfloor\widetilde{T}(\varepsilon)\rfloor-1}\mathcal{A}(\varphi,X_{\varepsilon,j},U_{\varepsilon,j})\Big] + \mathbf{e}_2(x,u,\varepsilon).$$

Again, the error term satisfies $\lim_{\varepsilon\to 0}\mathbb{E}\big|\mathbf{e}_2(X,U,\varepsilon)\big| = 0$ with $(X,U) \overset{\mathcal{D}}{\sim} \pi_1$. We now define a coupling between the original Markov chain $(X_{\varepsilon,k},Y_{\varepsilon,k})_{\geq 0}$ starting from $X_{\varepsilon,0} = x_0$ and a new Markov chain $\{x_0, Y_{x,\varepsilon,k}\}_{k\geq 0}$ starting from the same position i.e. satisfying $\big(X_{\varepsilon,0},Y_{\varepsilon,0}\big) = \big(x_0,Y_{x,\varepsilon,0}\big)$. Contrarily to the original Markov chain, the x-coordinate of the new Markov chain remains still. The y-coordinate of the new Markov chain is a RWM Markov chain targeting the probability distribution $\pi_\varepsilon(Y\in dy\,|X=x_0)$ on $\mathbb{R}^{n_y}$ with density proportional to $e^{B(x_0,y/\varepsilon)}$. The proposals of the new Markov chain $\{Y_{x_0,\varepsilon,k}\}_{k\geq 0}$ are

$$Y'_{x_0,\varepsilon,k} \;=\; Y_{x_0,\varepsilon,k} + h(\varepsilon)\,Z_{y,k}\;. \tag{6.3.8}$$

Since $h(\varepsilon) = \ell\varepsilon$, equation (6.3.8) also reads $U'_{x_0,\varepsilon,k} = U_{x_0,\varepsilon,k} + \ell\, Z_{y,k}$ where we have defined the rescaled y-coordinate process $U_{x_0,\varepsilon,k} := Y_{x_0,\varepsilon,k}/\varepsilon$. The rescaled Markov chain $\{U_{x_0,\varepsilon,k}\}_{k\geq 0}$ is a RWM Markov chain with target distribution $\pi_1(U \in du\,|X = x_0)$ that has $e^{B(x_0,u)}$ as density. Notice that the same source of randomness $Z_{y,k}$ is used to construct the two processes $\{Y_{\varepsilon,k}\}_{k\geq 0}$ and $\{Y_{x_0,\varepsilon,k}\}_{k\geq 0}$. The accept-reject mechanism can be described by

$$\left\{\begin{array}{ll} U_{\varepsilon,k+1} - U_{\varepsilon,k} & = \ell Z_{y,k} \cdot \mathbb{1}\left[\,\xi_k \leq a(X_{\varepsilon,k}, U_{\varepsilon,k}, Z_{x,k}, Z_{y,k})\,\right] \\ U_{x_0,\varepsilon,k+1} - U_{x_0,\varepsilon,k} & = \ell Z_{y,k} \cdot \mathbb{1}\left[\,\xi_k \leq a(x_0, U_{x_0,\varepsilon,k}, 0, Z_{y,k})\,\right], \end{array}\right. \tag{6.3.9}$$

where $\{\xi_k\}_{k\geq 0}$ are i.i.d. random variables uniformly distributed on $(0,1)$ and independent from all other sources of randomness. In other words, the same source of randomness $\{\xi_k\}_{k\geq 0}$ is used for the accept-reject mechanisms of the two Markov chains $(X_{\varepsilon,k}, U_{\varepsilon,k})$ and $(X_{\varepsilon,x_0,k}, U_{x_0,\varepsilon,k})$. The next proposition, whose proof is postponed to section 6.5.2, shows that the error committed by replacing $U_{\varepsilon,k}$ by $U_{x_0,\varepsilon,k}$ and by fixing the x-coordinate (i.e. by replacing $X_{\varepsilon,k}$ by $X_{\varepsilon,0} = x$) in equation (6.3.7) in negligible. This is mainly because for $k \leq \widetilde{T}(\varepsilon)$ iterations, the heuristic $X_{\varepsilon,k} \approx X_{\varepsilon,0}$ holds.

Proposition 6.3.2. *Let assumptions 6.2.2 be satisfied and $\varphi \in \mathcal{C}$ be a test function. Suppose further that the exponent γ has been chosen so that $0 < \gamma < \frac{1}{2p+1}$ where $p \geq 1$ is given by assumptions 6.2.2. Then the following identity holds*

$$\widetilde{\mathcal{L}}_\varepsilon\varphi(x,u) = \frac{1}{\widetilde{T}(\varepsilon)}\,\mathbb{E}_{x,u}\Big[\sum_{j=0}^{\lfloor\widetilde{T}(\varepsilon)\rfloor-1} \mathcal{A}(\varphi, x, U_{x,\varepsilon,j})\Big] + \boldsymbol{e}_3(x,u,\varepsilon) \tag{6.3.10}$$

with the error term satisfying $\lim_{\varepsilon\to 0}\mathbb{E}\big|\boldsymbol{e}_3(X,U,\varepsilon)\big| = 0$ where $(X,U) \overset{\mathcal{D}}{\sim} \pi_1$.

The advantage of representation (6.3.10) over equation (6.3.2) is that the right-hand side only involves the process $\{U_{x,\varepsilon,k}\}_{k\geq 0}$. In other, we have an expression of the type $\widetilde{\mathcal{L}}_\varepsilon\varphi(x,u) = \mathbb{E}\Big[\frac{1}{N}\sum_{k=0}^{N-1}\Phi(U_{x,\varepsilon,j})\Big]$ + (negligible error) with $N = \widetilde{T}(\varepsilon)$ and $\Phi(\cdot) = \mathcal{A}(\varphi, x, \cdot)$. Consequently, since $\{U_{x,\varepsilon,k}\}_{k\geq 0}$ is an ergodic Markov chain with invariant distribution $\pi_1(U \in du\,|X = x)$, the ergodic theorem for Markov chains shows that $\widetilde{\mathcal{L}}_\varepsilon\varphi(x,u)$ converges to $\mathbb{E}\big[\mathcal{A}(\varphi,X,U)\,\big|X = x\big]$ as $\varepsilon \to 0$. The next lemma show that $\mathbb{E}\big[\mathcal{A}(\varphi,X,U)\,\big|X = x\big] = \mathcal{L}\varphi(x)$, where $\mathcal{L}$ is the generator of the limiting diffusion (6.2.8).

Lemma 6.3.3. *Let $\mathcal{L}$ be the generator of the limiting diffusion (6.2.8) and $\mathcal{A}(\varphi,x,u)$ be the quantity defined in equation 6.3.6. For any test function $\varphi \in \mathcal{C}$ we have*

ideneity holds,

$$\mathbb{E}\big[\mathcal{A}(\varphi, X, U)\,\big|X = x\big] = \mathcal{L}\varphi(x)$$

where $(X, U) \overset{\mathcal{D}}{\sim} \pi_1$.

The proof of Lemma 6.3.3 is a routine calculation whose details can be found in section 6.5.3. It follows that as $\varepsilon \to 0$ the quantity $\widetilde{\mathcal{L}}_\varepsilon\varphi(x,u)$ converges to $\mathcal{L}\varphi(x)$. This result is the content of the next proposition whose detailed proof can be found in section 6.5.4.

Proposition 6.3.4. *Let assumptions 6.2.2 be satisfied and* $\varphi \in \mathcal{C}$ *be a test function. The following limit holds,*

$$\widetilde{\mathcal{L}}_\varepsilon\varphi(x,u) = \mathcal{L}\varphi(X) + e_4(x,u,\varepsilon),$$

with the error term satisfying $\lim_{\varepsilon\to 0}\mathbb{E}\big|\boldsymbol{e}_4(X,U,\varepsilon)\big| = 0$ *where* $(X,U) \overset{\mathcal{D}}{\sim} \pi_1$.

Notice that equation (6.3.5) is a rewording of proposition 6.3.4. To finish the proof of Theorem 6.2.3 it thus remains to verify if assumptions 6.2.2 are satisfied then equation (6.3.4) holds. Thanks to proposition 6.3.4, it suffices to check that $\mathbb{E}\big|\mathcal{L}\varphi(X)\big| < \infty$ with $X \overset{\mathcal{D}}{\sim} \pi$, which easily follows from assumptions 6.2.2.

6.3.2 The sequence $\widetilde{X}_{\varepsilon,t}$ converges weakly to the limiting diffusion (6.2.8)

In section 6.3.1 we have proved that the sequence of processes $\{\widetilde{S}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly in $D([0,T],\mathbb{R}^{n_x})$ to the limiting diffusion (6.2.8). This also implies that the sequence of processes $\{\widetilde{S}_{\varepsilon,t}\}_{t\in[0,T]}$ converges towards the same diffusion if one can establish that the process $\widetilde{X}_\varepsilon$ is close to the process $\widetilde{S}_\varepsilon$ in the sense that

$$\lim_{\varepsilon\to 0}\mathbb{E}\Big[\sup_{t\in[0,T]}\big\|\widetilde{X}_{\varepsilon,t} - \widetilde{S}_{\varepsilon,t}\big\|\Big] = 0. \tag{6.3.11}$$

By definition of the processes $\widetilde{X}_\varepsilon$ and $\widetilde{S}_\varepsilon$ we have that $\widetilde{X}_{\varepsilon,t} = X_{\varepsilon,\lfloor t\cdot T(\varepsilon)\rfloor}$ and $\widetilde{S}_{\varepsilon,t} = X_{\varepsilon,\lfloor\lfloor t\cdot T(\varepsilon/\widetilde{T}(\varepsilon)\rfloor\cdot\widetilde{T}(\varepsilon)\rfloor}$. For any $\alpha,\beta>0$ we have $\alpha-\beta-1 \le \lfloor\lfloor\alpha/\beta\rfloor\cdot\beta\rfloor \le \alpha+\beta+1$ and $\alpha-1\le\lfloor\alpha\rfloor\le\alpha+1$. Choosing $\alpha = t\cdot T(\varepsilon)$ and $\beta = \widetilde{T}(\varepsilon)$ leads to the inequality $\Big|\lfloor t\cdot T(\varepsilon)\rfloor - \lfloor\lfloor t\cdot T(\varepsilon)/\widetilde{T}(\varepsilon)\rfloor\cdot\widetilde{T}(\varepsilon)\rfloor\Big| \le \widetilde{T}(\varepsilon)+2$. The triangular inequality and the bound $\big\|X_{\varepsilon,j+1} - X_{\varepsilon,j}\big\| \le h(\varepsilon)\,\|Z_{x,j}\|$ imply that

$$\sup_{t\in[0,T]}\big\|\widetilde{X}_{\varepsilon,t} - \widetilde{S}_{\varepsilon,t}\big\| \le \sup_{0\le i,j\le T\cdot T(\varepsilon)}\Big\{\big\|X_{\varepsilon,i} - X_{\varepsilon,j}\big\| \;:\; |i-j|\le\widetilde{T}(\varepsilon)+2\Big\}$$

$$\leq \big(\widetilde{T}(\varepsilon)+2\big) \times \sup\Big\{ \big\| X_{\varepsilon,j+1} - X_{\varepsilon,j} \big\| \ : \ j \leq T\cdot T(\varepsilon) \Big\}$$
$$\leq h(\varepsilon)\,\big(\widetilde{T}(\varepsilon)+2\big) \times \sup\Big\{ \big\| Z_{x,j} \big\| \ : \ j \leq T\cdot T(\varepsilon) \Big\}.$$

One can then use the following lemma to control the expectation of the supremum of independent Gaussian random variables.

Lemma 6.3.5. *Let $\{Z_j\}_{j\geq 0}$ be independent $\mathbb{R}^d$-valued random variables distributed as $Z \overset{\mathcal{D}}{\sim} N(0, I_d)$. There exists a constant C_d depending on the dimension d only such that for any $n \geq 0$ the following inequality holds,*

$$\mathbb{E}\Big[\sup\Big\{ \big\| Z_{x,j} \big\| \ : \ j \leq n \Big\} \Big] \leq C_d\,\big(1+\sqrt{\log n}\,\big)$$

Proof. For convenience, we introduce the notation $M_n = \max\{\|Z_j\| \ : j \leq n\}$. Jensen's convexity inequality shows that for any $\lambda > 0$,

$$e^{\lambda\,\mathbb{E}[M_n]} \leq \mathbb{E}[e^{\lambda M_n}] = \mathbb{E}[\sup_{j\leq n} e^{\lambda\,\|Z_j\|}] \leq \sum_{j\leq n} \mathbb{E}[e^{\lambda\,\|Z_j\|}] \tag{6.3.12}$$

Since $\|Z_j\| = \big(\sum_{i=1}^d |Z_j^i|^2\big)^{1/2}$ is less $\sum_{i=1}^d |Z_j^i|$ we can use the bound $\mathbb{E}[e^{\lambda\,\|Z_j\|}] \leq \big(\mathbb{E}[e^{\lambda\,|\xi|}]\big)^d$ where $\xi \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1)$ is a standard scalar Gaussian random variables. Plugging the bound $\mathbb{E}[e^{\lambda\,|\xi|}] \leq 2\,\mathbb{E}[e^{\lambda\,\xi}] = 2e^{\lambda^2/2}$ into (6.3.12) leads to the inequality $e^{\lambda\,\mathbb{E}[M_n]} \leq 2^d n e^{d\,\lambda^2/2}$. Consequently, for any $\lambda > 0$ the inequality $\mathbb{E}[M_n] \leq \frac{d\,\log 2}{\lambda} + \frac{\log n}{\lambda} + \frac{d}{2}\lambda$ holds. The choice $\lambda = \sqrt{\log n}$ directly leads to the conclusion. □

Since $h(\varepsilon) = \ell\,\varepsilon$, $T(\varepsilon) = \varepsilon^{-2}$ and $\widetilde{T}(\varepsilon) = \varepsilon^{-\gamma}$, Lemma 6.3.5 shows that

$$h(\varepsilon)\,\big(\widetilde{T}(\varepsilon)+2\big) \times \mathbb{E}\Big[\sup\Big\{ \big\| Z_{x,j} \big\| \ : \ j \leq T\cdot T(\varepsilon) \Big\} \Big] \to 0$$

for any choice of exponent $0 < \gamma < 1$. This finishes the proof of equation (6.3.11) and concludes the proof of Theorem 6.2.3.

6.4 Proof of Theorem 6.2.5

The proof is entirely similar to the proof of Theorem 6.2.3. We only describe the modifications necessary to deal with this more general setting. We define the quantites $S_\varepsilon, \widetilde{S}_\varepsilon, \mathcal{L}_\varepsilon\varphi, \widetilde{\mathcal{L}}_\varepsilon\varphi$ the same way by equations (6.3.1),(6.3.2),(6.3.3). The acceptance probability of the move $(X,U) \to (X',U')$ reads

$$F \circ \log\Big(\frac{\pi_\varepsilon(X',U')\,p_\varepsilon\big((X',U') \to (X,U)\big)}{\pi_\varepsilon(X,U)\,p_\varepsilon\big((X,U) \to (X',U')\big)} \Big)$$

where $p_\varepsilon\big((X,U)\to(X',U')$ is the likelihood of the move $(X,U)\to(X'U')$. Proposition 6.3.1 still holds but the limiting quantity $\mathcal{A}(\varphi,x,u)=\lim_{\varepsilon\to 0}\mathcal{L}_\varepsilon\varphi(x,u)$ now reads

$$\mathcal{A}(\varphi,x,u)=\mathbb{E}\Big[F'(\delta B)\times\Big(\ell^2(x)\nabla_x\big\{A(x)+B(x,u+\ell Z_y)\big\}+\nabla_x\ell^2(x)\Big)\Big]\,\nabla\varphi(x)\\ +\frac{1}{2}\ell(x)^2\,\mathbb{E}\Big[F(\delta B)\Big]\,\Delta\varphi(x).$$

The proof uses a Taylor expansion of $\mathcal{L}_\varepsilon\varphi(x,u)$ and assumption 6.2.2,6.2.4 is exploited to give a control on the error terms. Under boundedness assumptions on the function $x\mapsto\ell(x)$ the coupling Proposition 6.3.2 is still valid and the rest of the proof is identical to the proof of Theorem 6.2.3.

6.5 Technical lemmas

Under assumptions 6.2.2 the first two derivatives of the functions $x\mapsto A(x)$ and $(x,u)\mapsto B(x,u)$ are bounded by a polynomial of degree p. The mean value theorem shows that

$$\begin{aligned}\big|A(x+\delta)-A(x)-\nabla_xA(x)\cdot\delta\big|&\lesssim(1+\|x\|^p+\|x+\delta\|^p)\,\|\delta\|^2\\&\lesssim(1+\|x\|^p+\|\delta\|^p)\,\|\delta\|^2.\end{aligned}\tag{6.5.1}$$

The second inequality is because $(\alpha+\beta)^p\lesssim\alpha^p+\beta^p$ for any scalars $\alpha,\beta\geq 0$. One can write a similar bound for the function B,

$$\big|B(x+\delta,u)-B(x,u)-\nabla_xB(x,u)\cdot\delta\big|\lesssim(1+\|x\|^p+\|u\|^p+\|\delta\|^p)\,\|\delta\|^2\tag{6.5.2}$$

6.5.1 Proof of Proposition 6.3.1

With the choice of scaling function $h(\varepsilon)=\ell\,\varepsilon$, the proposal $(X',U')=(X+\ell\varepsilon\,Z_x,U+\ell\,Z_y)$ is accepted with probability $a(X,U,Z_x,Z_y)=F\big(A(X')-A(X)+B(X',U')-B(X,U)\big)$. The one-step generator $\mathcal{L}_\varepsilon$ defined in equation (6.3.3) thus also reads

$$\mathcal{L}_\varepsilon\varphi(x,u)=\mathbb{E}\Big[\frac{\varphi(X')-\varphi(X)}{\varepsilon^2}\,a(X,U,Z_x,Z_y)\,\big|(X,U)=(x,u)\Big].\tag{6.5.3}$$

We then expand the two terms $\varphi(X')-\varphi(X)$ and $a(X,U,Z_x,Z_y)$ in powers of ε and control the error terms thanks to equations (6.5.7) and (6.5.2). For a smooth and

compactly supported test function $\varphi \in \mathcal{C}$ we have

$$\varphi(x') - \varphi(x) = \ell\varepsilon\langle\nabla\varphi(x), Z_x\rangle + \frac{\ell^2\varepsilon^2}{2}\langle Z_x, \nabla^2\varphi(x)Z_x\rangle + \mathcal{O}(\varepsilon^2\,\|Z_x\|^3). \quad (6.5.4)$$

where $(x', u') = (x + \ell\varepsilon Z_x, u + \ell Z_y)$. Equations (6.5.7) and (6.5.2) shows that

$$\begin{aligned} A(x') - A(x) + B(x', u') - B(x, u) = \delta B + \ell\varepsilon\langle\nabla A(x) + \nabla_x B(x, u'), Z_x\rangle \\ + \mathcal{O}\Big(\varepsilon^2(1 + \|x\|^p + \|u\|^p + \|Z_x\|^p + \|Z_y\|^p) \times \|Z_x\|^2\Big) \end{aligned}$$

where the leading term has been defined as $\delta B = B(x, u') - B(x, u)$. Plugging this back into the definition of the acceptance probability $a(X, U, Z_x, Z_y)$ and exploiting the fact that by assumptions the function F has its first two derivatives bounded one obtains

$$\begin{aligned} a(x, u, Z_x, Z_y) = F(\delta B) + \ell\varepsilon\, F'(\delta B)\,\langle\nabla_x\big\{A(x) + \nabla_x B(x, u')\big\}, Z_x\rangle \quad (6.5.5) \\ + \mathcal{O}\Big(\varepsilon^2(1 + \|x\|^{2p} + \|u\|^{2p} + \|Z_x\|^{2p} + \|Z_y\|^{2p}) \times \|Z_x\|^2\Big) \end{aligned}$$

Bringing together the Taylor expansions (6.5.4) and (6.5.5) into the definition of the one step generator (6.5.3) shows that $\mathcal{L}_\varepsilon\varphi(x, u) = \mathcal{A}(\varphi, x, u) + e_1(x, u, \varepsilon)$ where the errors terms verifies $\mathbb{E}[e_1(X, U, \varepsilon)] \to 0$ as $\varepsilon \to 0$ and

$$\begin{aligned} \mathcal{A}(\varphi, x, u) = \ell^2\,\mathbb{E}\Big[F'(\delta B)\langle\nabla\varphi(x), Z_x\rangle\,\langle\nabla_x\big\{A(x) + \nabla_x B(x, u')\big\}, Z_x\rangle\Big] \\ + \frac{1}{2}\ell^2\,\mathbb{E}\Big[F(\Delta B)\langle Z_x, \nabla^2\varphi(x)Z_x\rangle\Big]. \end{aligned}$$

Since $\delta B = B(x, u + \ell Z_y) - B(x, u)$ is independent from the random variable Z_x, the quantity $\mathcal{A}(\varphi, x, u)$ can also be written as

$$\begin{aligned} \mathcal{A}(\varphi, x, u) = \ell^2\,\mathbb{E}\Big[F'(\delta B)\nabla_x\big\{A(x) + \nabla_x B(x, u + \ell Z_y)\big\}\Big]\,\nabla\varphi(x) \quad (6.5.6) \\ + \frac{1}{2}\ell^2\,\mathbb{E}\Big[F(\delta B)\Big]\,\Delta\varphi(x). \end{aligned}$$

This finishes the proof of Proposition 6.3.1.

6.5.2 Proof of Proposition 6.3.2

For simplicity, in this section we write X_0, Y_0, U_0 instead of $X_{\varepsilon,0}, Y_{\varepsilon,0}, U_{\varepsilon,0}$. Under the regularity ssumptions 6.2.2 on the function $x \mapsto A(x)$ and $(x, u) \mapsto B(x, u)$ and their derivatives one can check that for any smooth and compactly test function

$\varphi \in \mathcal{C}$ the following bounds hold,

$$\begin{aligned} \mathcal{A}(\varphi, x, u) &\lesssim 1 + \|x\|^p + \|u\|^p \\ \big\| \nabla_x \mathcal{A}(\varphi, x, u) \big\| &\lesssim 1 + \|x\|^{2p} + \|u\|^{2p}. \end{aligned} \tag{6.5.7}$$

Therefore, we have $\mathbb{E}\big[\big|\mathcal{A}(\varphi, X, U)\big|\big] < \infty$ and $\mathbb{E}\big[\big|\partial_x \mathcal{A}(\varphi, X, U)\big|\big] < \infty$ for $(X, U) \overset{\mathcal{D}}{\sim} \pi_1$. In view of Proposition 6.3.1, in order to prove Proposition 6.3.2 it suffices to establish that if the RWM chain is started at stationarity, i.e. $(X_0, U_0) \overset{\mathcal{D}}{\sim} \pi_1$, then for any test function $\varphi \in \mathcal{C}$ we have

$$\lim_{\varepsilon \to 0} \quad \frac{1}{\widetilde{T}(\varepsilon)} \sum_{j=0}^{\lfloor \widetilde{T}(\varepsilon) \rfloor - 1} \mathbb{E}\Big[\Big| \mathcal{A}(\varphi, X_0, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{\varepsilon, j}) \Big| \Big] = 0. \tag{6.5.8}$$

The expectation is less than the sum of $\mathbb{E}\big|\mathcal{A}(\varphi, X_0, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{X_0, \varepsilon, j})\big|$ and $\mathbb{E}\big|\mathcal{A}(\varphi, X_j, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{\varepsilon, j})\big|$. We now control each one of these terms separately.

From the regularity estimate for the function $\mathcal{A}(\varphi, x, u)$ presented in equation (6.5.6) we can deduce that $\big|\mathcal{A}(\varphi, X_0, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{X_0, \varepsilon, j})\big|$ is less than a constant multiple of

$$\big(1 + \|X_0\|^{2p} + \|X_{\varepsilon, j}\|^{2p} + \|U_{X_0, \varepsilon, j}\|^{2p}\big) \, \|X_0 - X_{\varepsilon, j}\|. \tag{6.5.9}$$

One can bound the quantity $\|X_0 - X_j\|$ using the fact that $\|X_{k+1} - X_k\| \leq \ell\varepsilon \|Z_k\|$ for any $k \geq 0$, which leads to the inequality $\mathbb{E}\|X_0 - X_j\| \lesssim j\varepsilon$ and similarly $\mathbb{E}\|X_j\|^{2p} \lesssim \mathbb{E}\|X_0\|^{2p} + j\varepsilon^{2p}$. Plugging this back into (6.5.9) shows that $\mathbb{E}\big|\mathcal{A}(\varphi, X_0, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{X_0, \varepsilon, j})\big| \lesssim j^{2p+1}\, \varepsilon$ so that

$$\frac{1}{\widetilde{T}(\varepsilon)} \sum_{j=0}^{\lfloor \widetilde{T}(\varepsilon) \rfloor - 1} \mathbb{E}\big|\mathcal{A}(\varphi, X_0, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{X_0, \varepsilon, j})\big| \lesssim \widetilde{T}(\varepsilon)^{2p+1}.$$

Since $\widetilde{T}(\varepsilon) = \varepsilon^{-\gamma}$, the upper bound also reads $\varepsilon^{1-\gamma(2p+1)}$, which goes to zero for any exponent γ satisfying $0 < \gamma < \frac{1}{2p+1}$.

To finish the proof, we now verify that $\frac{1}{\widetilde{T}(\varepsilon)} \sum_{j=0}^{\lfloor \widetilde{T}(\varepsilon) \rfloor - 1} \mathbb{E}\big|\mathcal{A}(\varphi, X_{\varepsilon, j}, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{\varepsilon, j})\big|$ also converges to zero as ε goes to zero. Since the difference $\big|\mathcal{A}(\varphi, X_{\varepsilon, j}, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{\varepsilon, j})\big|$ also equals $\mathbb{1}(U_{X_0, \varepsilon, j} \neq U_{\varepsilon, j}) \times \big|\mathcal{A}(\varphi, X_{\varepsilon, j}, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{\varepsilon, j})\big|$, equation (6.5.7) shows that $\big|\mathcal{A}(\varphi, X_{\varepsilon, j}, U_{X_0, \varepsilon, j}) - \mathcal{A}(\varphi, X_{\varepsilon, j}, U_{\varepsilon, j})\big|$ is less than a con-

stant multiple of

$$\mathbb{1}(U_{X_0,\varepsilon,j} \neq U_{\varepsilon,j}) \Big(1 + \|X_{\varepsilon,j}\|^p + \|U_{\varepsilon,j}\|^p + \|U_{X_0,\varepsilon,j}\|^p\Big).$$

Since $(X_0, U_0) \overset{\mathcal{D}}{\sim} \pi_1$, the stationarity of the RWM algorithm and assumption 6.2.2 show that the expectation $\mathbb{E}[1 + \|X_{\varepsilon,j}\|^p + \|U_{\varepsilon,j}\|^p + \|U_{X_0,\varepsilon,j}\|^p]$ is finite and does not depend on ε. Consequently, the Cauchy-Schwarz inequality implies that $\mathbb{E}\big|\mathcal{A}(\varphi, X_{\varepsilon,j}, U_{X_0,\varepsilon,j}) - \mathcal{A}(\varphi, X_{\varepsilon,j}, U_{\varepsilon,j})\big| \lesssim \mathbb{P}\big(U_{X_0,\varepsilon,j} \neq U_{\varepsilon,j}\big)^{1/2}$. To finish the proof, it thus remains to verify that

$$\lim_{\varepsilon \to 0} \frac{1}{\widetilde{T}(\varepsilon)} \sum_{j=0}^{\lfloor \widetilde{T}(\varepsilon) \rfloor - 1} \mathbb{P}\big(U_{X_0,\varepsilon,j} \neq U_{\varepsilon,j}\big)^{1/2} \;=\; 0.$$

To this end, we will bound the quantity $\mathbb{P}\big(U_{X_0,\varepsilon,j} \neq U_{\varepsilon,j}\big)$. The definition (6.3.9) of the coupling between the Markov chains $\{(X_{\varepsilon,k}, U_{\varepsilon,k}\}_{k\geq 0}$ and $\{(X_0, U_{X_0,\varepsilon,k}\}_{k\geq 0}$ shows that $U_{\varepsilon,k} = U_{X_0,\varepsilon,k}$ if, and only if, the proposals $U'_{\varepsilon,j}$ and $U'_{X_0,\varepsilon,j}$ for $j \leq k-1$ have all been accepted or rejected at the same time. This happens if $\mathbb{1}\,[\,\xi_k \leq a(X_{\varepsilon,j}, U_{\varepsilon,j}, Z_{x,j}, Z_{y,j})\,] = \mathbb{1}\,[\,\xi_k \leq a(X_0, U_{X_0,\varepsilon,j}, 0, Z_{y,j})\,]$ for all $j \leq k-1$. The probability $\mathbb{P}\big(U_{X_0,\varepsilon,j} \neq U_{\varepsilon,j}\big)$ thus equals

$$\mathbb{E}\Big[1 - \prod_{j=0}^{k-1} \big(1 - \big|a(X_{\varepsilon,j}, U_{\varepsilon,j}, Z_{x,j}, Z_{y,j}) - a(X_0, U_{\varepsilon,j}, 0, Z_{y,j})\big|\big)\Big],$$

which is inferior to $\sum_{j=0}^{k-1} \mathbb{E}\big|a(X_{\varepsilon,j}, U_{\varepsilon,j}, Z_{x,j}, Z_{y,j}) - a(X_0, U_{\varepsilon,j}, 0, Z_{y,j})\big|$. Since $a(x, u, Z_x, Z_y) = F\big(A(x + Z_x) - A(x) + B(x + Z_x, u + Z_y) - b(x, u)\big)$ and F is a Lipschitz function and one can control the quantity $A(x + Z_x) - A(x)$ and $B(x + Z_x, u + Z_y) - b(x, u)$ thanks to assumptions 6.2.2, routine algebra shows that $\mathbb{E}\big|a(X_{\varepsilon,j}, U_{\varepsilon,j}, Z_{x,j}, Z_{y,j}) - a(X_0, U_{\varepsilon,j}, 0, Z_{y,j})\big|$ is less than a constant multiple of

$$(1 + \|X_0\|^p + \|X_{\varepsilon,j}\|^p + \|U_{\varepsilon,j}\|^p + \|Z_{x,j}\|^p + \|Z_{y,j}\|^p) \cdot (\varepsilon\|Z_{x,j}\| + \|X_{\varepsilon,j} - X_0\|).$$

By the triangle inequality we have $\|X_{\varepsilon,j} - X_0\| \lesssim \varepsilon\big(\|Z_{x,0}\| + \ldots + \|Z_{x,j-1}\|\big)$. Therefore, using Cauchy-Schwarz inequality and Holder's inequality, it follows that $\mathbb{E}\big|a(X_{\varepsilon,j}, U_{\varepsilon,j}, Z_{x,j}, Z_{y,j}) - a(X_0, U_{\varepsilon,j}, 0, Z_{y,j})\big| \lesssim j\varepsilon$ and therefore $\mathbb{P}\big(U_{X_0,\varepsilon,j} \neq U_{\varepsilon,j}\big) \lesssim \sum_{j=0}^{k-1} k\varepsilon \lesssim k^2\,\varepsilon$. Because $\widetilde{T}(\varepsilon) = \varepsilon^{-\gamma}$, putting all these inequalities together leads

to the bound

$$\frac{1}{\widetilde{T}(\varepsilon)} \sum_{j=0}^{\lfloor \widetilde{T}(\varepsilon) \rfloor - 1} \mathbb{E}\big|\mathcal{A}(\varphi, X_{\varepsilon,j}, U_{X_0,\varepsilon,j}) - \mathcal{A}(\varphi, X_{\varepsilon,j}, U_{\varepsilon,j})\big| \lesssim \varepsilon^{\frac{1}{2}-\gamma}. \qquad (6.5.10)$$

Equations (6.5.9) and (6.5.10) together imply that equation (6.5.8) holds. This finishes the proof of Proposition 6.3.2.

6.5.3 Proof of Lemma 6.3.3

To keep this exposition as simple as possible, we suppose that $\ell = 1$ and $n_x = n_y = 1$. The multi-dimensional case is entirely similar. The proof of Lemma 6.3.3 consists in verifying that for all $x \in \mathbb{R}$ the following identity holds,

$$\int_{u\in\mathbb{R}} \mathcal{A}(\varphi, x, u)\, e^{B(x,u)}\, du \;=\; \mathcal{L}\varphi(x) \qquad (6.5.11)$$

where $\mathcal{L}$ is the generator of the limiting diffusion (6.2.8), $\varphi \in \mathcal{C}$ is a test function in the core of $\mathcal{L}$ and $\mathcal{A}(\varphi, x, u)$ reads

$$\mathcal{A}(\varphi, x, u) = \mathbb{E}\Big[F'(\delta B)\big(A'(x) + \partial_x B(x, u+Z)\big)\Big]\, \varphi'(x) + \frac{1}{2}\,\mathbb{E}\Big[F(\delta B)\Big]\, \varphi^{''}(x)$$

where $\delta B = B(x, u+Z) - B(x,u)$ and $Z \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1)$. The proof is a routine calculation that is based on the symmetry of the Gaussian distribution ,i.e. $Z \overset{\mathcal{D}}{\sim} -Z$ for $Z \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1)$, and exploits the fact that the accept-reject function F verifies the reversibility condition (6.2.3). More specifically, the derivative of equation (6.2.3) also reads

$$F(r) = F'(r) + e^r F'(-r) \qquad \text{for all} \qquad r \in \mathbb{R}. \qquad (6.5.12)$$

This identity also hods for the Metropolis-Hastings function $F_{\mathrm{MH}}(r) = \min(1, e^r)$ but has to be interpreted in the sense of distributions. In the scalar case $n_x = 1$ with $\ell = 1$, the generator of (6.2.8) reads $\mathcal{L}\varphi(x) = \frac{1}{2}\big(a_0(x)\, A'(x) + a_0'(x)\big)\, \varphi'(x) + \frac{1}{2} a_0(x)\, \varphi^{''}(x)$ where $a_0(x) := a_0(x, 1)$ is the mean acceptance probability $a_0(x) = \int_{u\in\mathbb{R}} \mathbb{E}[F(\delta B)]\, e^{B(x,u)}\, du$. To prove Equation (6.5.11) it suffices to verify that

$$\mathbb{E}\big[F'(\delta B)\partial_x B(x, u+Z)\big] = \frac{1}{2}\, a_0'(x) \quad \text{and} \quad \mathbb{E}\big[F'(\delta B)\big] = \frac{1}{2} a_0(x). \quad (6.5.13)$$

Let us prove that the first identity holds. Assumptions 6.2.2 justify the following derivation under the integral sign,

$$\partial_x a_0(x) = \int \mathbb{E}[F'(\delta B)\left(\partial_x B(x, u+Z) - \partial_x B(x,u)\right) + F(\delta B)\partial_x B(x,u)]\, e^{B(x,u)}\, du$$

Equation (6.5.12) shows that one can $F(\delta B)$ also equals $F'(\delta B)+F(-\delta B)e^{\delta B}$. Since $e^{\delta B}e^{B(x,u)} = e^{B(x,u+Z)}$, we have

$$\partial_x a_0(x) = \int \mathbb{E}[F'(\delta B)\partial_x B(x,u+Z)e^{B(x,u)}]\, du + \int \mathbb{E}[F'(-\delta B)\partial_x B(x,u)e^{B(x,u+Z)}]\, du.$$

The symmetry of the Gaussian distribution $Z \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1)$ then shows that

$$\int \mathbb{E}[F'(\delta B)\partial_x B(x,u+Z)e^{B(x,u)}]\, du = \int \mathbb{E}[F'(-\delta B)\partial_x B(x,u)e^{B(x,u+Z)}].$$

This concludes the proof of the first identity of (6.5.13). The proof of the second identity is similar and more simple, and thus omitted.

6.5.4 Proof of Proposition 6.3.4

To prove Proposition 6.3.4 one needs to establish that

$$\lim_{N\to\infty} \mathbb{E}\Big|\frac{1}{N}\sum_{k=1}^{N} \mathcal{A}(\varphi, X, U_{X,k}) - \mathcal{L}\varphi(X)\Big| = 0 \tag{6.5.14}$$

where $(X, U_{X,0}) \overset{\mathcal{D}}{\sim} \pi_1$ and for every $x_0 \in \mathbb{R}^{n_x}$ the process $\{U_{x_0,k}\}_{k\geq 0}$ is a RWM Markov chain with target distribution $\pi_1(U \in du \,|X = x_0) = e^{B(x_0,u)}\, du$ as described by equation (6.3.9). The bound (6.5.7) implies that $\mathbb{E}\big|\mathcal{A}(\varphi, X, U)\big| < \infty$ so that for π-almost every $x_0 \in \mathbb{R}^{n_x}$ we have

$$\int_{u\in\mathbb{R}^{n_y}} \big|\mathcal{A}(\varphi, x_0, u)\big|\, e^{B(x_0,u)}\, du < \infty.$$

For such a $x_0 \in \mathbb{R}^{n_x}$, the ergodic theorem for Markov chain shows that the set $\mathcal{S}_{x_0}$ of starting position $U_{x_0,0} = u_0 \in \mathbb{R}^{n_y}$ such that $\frac{1}{N}\sum_{k=1}^{N} \mathcal{A}(\varphi, x_0, U_{x_0,k})$ converges almost surely to $\mathcal{L}\varphi(x_0)$ has full measure $\int_{\mathcal{S}_{x_0}} e^{B(x_0,u)}\, du = 1$. Consequently, the

convergence

$$\Big|\frac{1}{N}\sum_{k=1}^{N}\mathcal{A}(\varphi,X,U_{X,k})-\mathcal{L}\varphi(X)\Big|\to 0 \tag{6.5.15}$$

holds almost surely with $(X,U_{X,0})\overset{\mathcal{D}}{\sim}\pi_1$. To prove that the convergence actually holds in L^1 as described by equation (6.5.14) it suffices to prove that the sequence $\Big\{\frac{1}{N}\sum_{k=1}^{N}\mathcal{A}(\varphi,X,U_{X,k})-\mathcal{L}\varphi(X)\Big\}_{N\geq 1}$ is uniformly integrable. To this end, one can prove that the family is bounded in L^2,

$$\sup_{N\geq 1}\mathbb{E}\Big|\frac{1}{N}\sum_{k=1}^{N}\mathcal{A}(\varphi,X,U_{X,k})-\mathcal{L}\varphi(X)\Big|^2 \;<\; \infty. \tag{6.5.16}$$

By stationarity we have $(X,U_{X,k})\overset{\mathcal{D}}{\sim}\pi_1$ for any $k\geq 0$. The generalised mean inequality thus implies that

$$\begin{aligned}\mathbb{E}\Big|\frac{1}{N}\sum_{k=1}^{N}\mathcal{A}(\varphi,X,U_{X,k})-\mathcal{L}\varphi(X)\Big|^2 &\leq \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\Big|\mathcal{A}(\varphi,X,U_{X,k})-\mathcal{L}\varphi(X)\Big|^2\\ &\lesssim \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\big|\mathcal{A}(\varphi,X,U_{X,k})\big|^2+\mathbb{E}\big|\mathcal{L}\varphi(X)\big|^2\\ &\lesssim \mathbb{E}\big|\mathcal{A}(\varphi,X,U_{X,0})\big|^2+\mathbb{E}\big|\mathcal{L}\varphi(X)\big|^2<\infty\end{aligned}$$

where the last inequality follows from the growth assumptions on A and B. This finishes the proof of the L^2-boundedness (6.5.16) which in turn finishes the proof of Proposition 6.3.4.

6.6 Conjectures

In this chapter, all the proofs of diffusion limit results have been described under the assumption that the limiting manifold $\mathcal{M}$ is flat. The advantage of this assumption is that it is relatively straightforward to described the limiting diffusion process since it is simply a diffusion process evolving on a linear subspace of $\mathbb{R}^n$ *i.e.* there is no need of stochastic geometry to define the limiting process. In the more general setting where the limiting manifold $\mathcal{M}$ is possibly non-linear, a diffusion limit of the same type is conjectured to hold. Indeed, since a general manifold is locally flat and our diffusion limits results are local in nature, all the results proved in this chapter are expected to hold for non-flat limiting manifolds. In particular, to

explore a target probability distribution π_ε concentrated on an ε-neighbourhood of a limiting manifold $\mathcal{M} \subset \mathbb{R}^n$ with dimension $\dim(\mathcal{M}) \leq n-3$, a RWM algorithm should use jumps of size with magnitude $\mathcal{O}(\varepsilon)$. The resulting Markov chain behaves like a diffusion walking on the manifold $\mathcal{M}$ if time is accelerated by a factor ε^{-2}.

To illustrate this point, we carry numerical simulations on a non-linear toy problem. We consider the sequence of distributions π_ε on $\mathbb{R}^2$ with density proportional to

$$\pi_\varepsilon(x,y) \propto \exp\Big\{ - \frac{(x^2 + y^2/4 - 1)^2}{2\,\varepsilon^2} \Big\}. \tag{6.6.1}$$

The distribution π_ε is concentrated on an ε-neighbourhood of the ellipse $\mathcal{M} := \{(x,y) \in \mathbb{R}^2 \,:\, x^2 + \frac{y^2}{4} = 1\}$. The RWM algorithm with jump of size ε is used to explore π_ε. Proposals are given by $(x', y') = (x + \varepsilon\, Z_x, y + \varepsilon\, Z_y)$ with $(Z_x, Z_y) \overset{\mathcal{D}}{\sim} \mathrm{N}(0, I_2)$ and the accept-reject mechanism is given by the usual Metropolis-Hastings function $F_{\mathrm{MH}}(r) = \min(1, e^r)$. A generalisation of Theorem 6.2.3 would imply that as $\varepsilon \to 0$ the algorithmic complexity of this RWM algorithm scales as $\mathcal{O}(\varepsilon^{-2})$. To illustrate this conjecture, the algorithm is started at $(X_{\varepsilon,0}, Y_{\varepsilon,0}) = (0,2) \in \mathcal{M}$ and the hitting time $\tau_\varepsilon := \inf\{k \geq 0 : Y_{\varepsilon,k} \leq 1\}$ is recorded. It is expected that τ_ε is of order $\mathcal{O}(\varepsilon^{-2})$. Figure 6.6 is Monte-Carlo estimate of $\mathbb{E}[\tau_\varepsilon]$ as ε converges to zero.

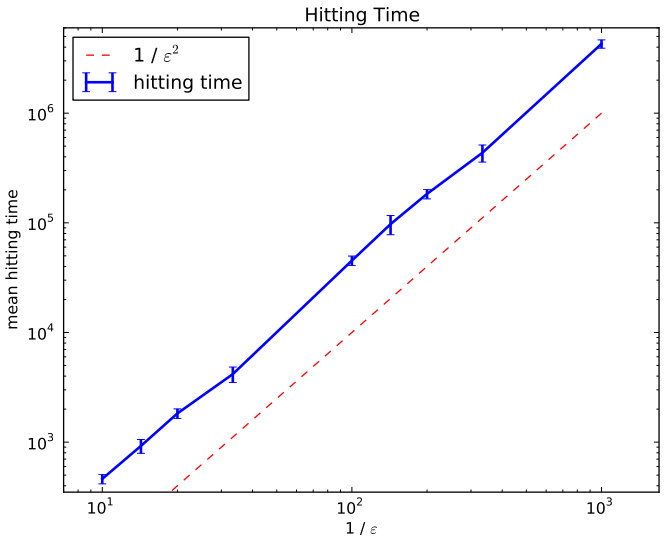


Figure 6.1: RWM algorithm with proposals $(x', y') = (x + \varepsilon Z_x, y + \varepsilon Z_y)$ is used to explore the target distribution 6.6.1. The algorithm is started at $(X_{\varepsilon,0}, Y_{\varepsilon,0}) = (0, 2) \in \mathcal{M}$ and the mean hitting time $\tau_\varepsilon := \inf\{k \geq 0 : Y_{\varepsilon,k} \leq 1\}$ is analysed. The blue curve represents the mean hitting time $\pm$ two standard deviations. The dotted red line represents ε^{-2}.

Bibliography

[Bax05] P. Baxendale. Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Annals of Applied Probability*, 15(1B):700–738, 2005.

[BCG08] D. Bakry, P. Cattiaux, and A. Guillin. Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *Journal of Functional Analysis*, 254(3):727–759, 2008.

[BDM10] M. Bédard, R. Douc, and E. Moulines. Scaling analysis of delayed rejection MCMC methods. *Methodology and Computing in Applied Probability*, pages 1–28, 2010.

[BDM12] M. Bédard, R. Douc, and E. Moulines. Scaling analysis of multiple-try MCMC methods. *Stochastic Processes and their Applications*, 122(3):758–786, 2012.

[Béd07] M. Bédard. Weak convergence of Metropolis algorithms for non-iid target distributions. *Annals of Applied Probability*, 17(4):1222–1244, 2007.

[Béd08] M. Bédard. Efficient sampling using Metropolis algorithms: Applications of optimal scaling results. *Journal of Computational and Graphical Statistics*, 17(2):312–332, 2008.

[Béd09] M. Bédard. On the optimal scaling problem of Metropolis algorithms for hierarchical target distributions. *Preprint*, 2009.

[Ber86] E. Berger. Asymptotic behaviour of a class of stochastic approximation procedures. *Probability Theory and Related Fields*, 71(4):517–552, 1986.

[BKMS08] E. Buckwar, R. Kuske, S. Mohammed, and T. Shardlow. Weak convergence of the Euler scheme for stochastic differential delay equations. *LMS Journal of Computation and Mathematics*, 11(-1):60–99, 2008.

[BPR+13] A. Beskos, N. Pillai, G. Roberts, J. Sanz-Serna, and A. Stuart. Optimal tuning of hybrid monte carlo. to appear, 2013.

[BPS04] L. Breyer, M. Piccioni, and S. Scarlatti. Optimal scaling of MALA for nonlinear regression. *Annals of Applied Probability*, 14(3):1479–1505, 2004.

[BPSSS11] A. Beskos, F. Pinski, J. Sanz-Serna, and A. Stuart. Hybrid monte carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121:2201–2230, 2011.

[BR00] L. Breyer and G. Roberts. From Metropolis to diffusions: Gibbs states and optimal scaling. *Stochastic Process and their Applications*, 90(2):181–206, 2000.

[BR07] M. Bédard and J. Rosenthal. Optimal scaling of Metropolis algorithms: is 0.234 as robust as is believed? *preprint, July*, 2007.

[BRS09] A. Beskos, G. Roberts, and A. Stuart. Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions. *Annals of Applied Probability*, 19(3):863–898, 2009.

[BRSV08] A. Beskos, G. Roberts, A. Stuart, and J. Voss. An MCMC method for diffusion bridges. *Stochastics and Dynamics*, 8(3):319–350, 2008.

[BRT13] A. Beskos, G. Roberts, and A. Thiéry. MCMC on ridge densities. Working version, 2013.

[BS05] E. Buckwar and T. Shardlow. Weak approximation of stochastic differential delay equations. *IMA Journal of Numerical Analysis*, 25(1):57, 2005.

[BS09] A. Beskos and A. Stuart. MCMC methods for sampling function space. In *Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, ICIAM07, Editors Rolf Jeltsch and Gerhard Wanner*, pages 337–364. European Mathematical Society, 2009.

[BT93] D. Bertsimas and J. Tsitsiklis. Simulated annealing. *Statistical Science*, 8(1):10–15, 1993.

[Can08] E. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathématiques*, 346(9):589–592, 2008.

[CDS12] S. Cotter, M. Dashti, and A. Stuart. Variational data assimilation using targeted random walks. *International Journal for Numerical Methods in Fluids*, 68:403–421, 2012.

[Čer85] V. Černỳ. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985.

[CH06] A. Chorin and O. Hald. *Stochastic Tools in Mathematics and Science*, volume 1 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer, New York, 2006.

[Cip00] B. Cipra. The best of the 20th century: Editors name top 10 algorithms. *SIAM news*, 33(4):1–2, 2000.

[CMMR12] J. Cornuet, J. Marin, A. Mira, and C. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

[CRR05] O. Christensen, G. Roberts, and J. Rosenthal. Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268, 2005.

[CRSW12] S. Cotter, G. Roberts, A. Stuart, and D. White. Mcmc methods for functions: modifying old algorithms to make them run faster. *arXiv preprint arXiv:1202.0709*, 2012.

[dBD06] A. de Bouard and A. Debussche. Weak and strong order of convergence of a semi-discrete scheme for the stochastic nonlinear schrödinger equation. *Applied Mathematics and Optimization*, 54(3):369–399, 2006.

[DH92] P. Diaconis and P. Hanlon. Eigen-analysis for some examples of the Metropolis algorithm. *Contemporary Mathematics*, 138:99–117, 1992.

[DHS12] M. Dashti, S. Harris, and A. Stuart. Besov priors for Bayesian inverse problems. *Inverse Problems and Imaging*, 6:183–200, 2012.

[Dia88] P. Diaconis. Applications of non-commutative Fourier analysis to probability problems. *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, pages 51–100, 1988.

[Dia09] P. Diaconis. The Markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

[DL09] P. Diaconis and G. Lebeau. Micro-local analysis for the Metropolis algorithm. *Mathematische Zeitschrift*, 262(2):411–447, 2009.

[Don06] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[DP05] G. Da Prato. *An introduction to infinite-dimensional analysis*. Springer, 2005.

[DP09] A. Debussche and J. Printems. Weak order for the discretization of the stochastic heat equation. *Mathematics of Computation*, 78(266):845–863, 2009.

[DPZ92] G. Da Prato and J. Zabczyk. *Stochastic equations in infinite dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.

[DPZ96] G. Da Prato and J. Zabczyk. *Ergodicity for Infinite Dimensional Systems*. Cambridge University Press, 1996.

[DS63] N. Dunford and J. Schwartz. *Linear Operators Part II: Spectral Theory*. Interscience Publishers, New York, London, 1963.

[DS81] P. Diaconis and M. Shahshahani. Generating a random permutation with random transpositions. *Probability Theory and Related Fields*, 57(2):159–179, 1981.

[EK86] S.N. Ethier and T.G. Kurtz. *Markov processes: Characterization and convergence*, volume 6. Wiley New York, 1986.

[EMM12] T. El Moselhy and Y. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231:7815–7850, 2012.

[Fer75] X. Fernique. Regularité des trajectoires des fonctions aléatoires Gaussiennes. In *Ecole d'Eté de Probabilités de Saint-Flour 1974*, pages 1–96. Springer, 1975.

[FHT09] J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics*, 2009.

[Fit91] B.G. Fitzpatrick. Bayesian analysis in inverse problems. *Inverse problems*, 7:675, 1991.

[Fos53] F. Foster. On the stochastic matrices associated with certain queuing processes. *Annals of Mathematical Statistics*, 24(3):355–360, 1953.

[Fuk80] M. Fukushima. *Dirichlet forms and Markov processes*. North-Holland Amsterdam, 1980.

[FW12] M. Freidlin and A. Wentzell. *Random perturbations of dynamical systems*, volume 260. Springer, 2012.

[GC11] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[Gem85] D. Geman. Bayesian image analysis by adaptive annealing. *IEEE Transactions on Geoscience and Remote Sensing*, 1:269–276, 1985.

[GH86] S. Geman and C. Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24:1031, 1986.

[GKL09] M. Geissert, M. Kovács, and S. Larsson. Rate of weak convergence of the finite element method for the stochastic heat equation with additive noise. *BIT Numerical Mathematics*, 49(2):343–356, 2009.

[GRS96] W. Gilks, S. Richardson, and D. Spiegelhalter. Introducing Markov chain Mnote Carlo. *Markov chain Monte Carlo in practice*, pages 1–19, 1996.

[GS01] G.R. Grimmett and D.R. Stirzaker. *Probability and random processes*. Oxford University Press, USA, 2001.

[Has70] W. Hastings. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[Hau10] E. Hausenblas. Weak approximation of the stochastic wave equation. *Journal of computational and applied mathematics*, 235(1):33–58, 2010.

[Hen81] D. Henry. *Geometric Theory of Semi-linear Parabolic Equations*, volume 61. Springer-Verlag, 1981.

[HKS89] R. Holley, S. Kusuoka, and D. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *Journal of Functional Analysis*, 83(2):333–347, 1989.

[HPUU08] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*. Springer Verlag, 2008.

[HST01] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[HSV07] M. Hairer, A. Stuart, and J. Voss. Analysis of SPDEs arising in path sampling. Part II: the nonlinear case. *Annals of Applied Probability*, 17(5-6):1657–1706, 2007.

[HSV10] M. Hairer, A. Stuart, and J. Voss. Signal processing problems on function space: Bayesian formulation, stochastic pdes and effective MCMC methods. *The Oxford Handbook of Nonlinear Filtering, Editors D. Crisan and B. Rozovsky*, 2010.

[HSV11] M. Hairer, A. Stuart, and S. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *arXiv preprint arXiv:1112.1392*, 2011.

[HSVW05] M. Hairer, A. Stuart, J. Voss, and P. Wiberg. Analysis of SPDEs arising in path sampling. Part i: the Gaussian case. *Communications in Mathematical Sciences*, 3:587–603, 2005.

[JLM12] B. Jourdain, T. Lelièvre, and B. Miasojedow. Optimal scaling for the transient phase of the random walk Metropolis algorithm: the mean-field limit. *arXiv preprint arXiv:1210.7639*, 2012.

[KGV83] S. Kirkpatrick, D. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[KLL12] Mi. Kovács, S. Larsson, and F. Lindgren. Weak convergence of finite element approximations of linear stochastic evolution equations with additive noise. *BIT Numerical Mathematics*, 52(1):85–108, 2012.

[KV86] C. Kipnis and S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986.

[Liu08] J. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer, New York, 2008.

[LS88] G. Lawler and A. Sokal. Bounds on the l2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Transactions of the American Mathematical Society*, 309(2):557–580, 1988.

[LSS09] M. Lassas, E. Saksman, and S. Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3:87–122, 2009.

[MG99] A. Mira and C. Geyer. Ordering monte carlo Markov chains. *School of Statistics, University of Minnesota. technical report*, 1999.

[Mir01] A. Mira. Ordering and improving the performance of monte carlo Markov chains. *Statistical Science*, 16(4):340–350, 2001.

[MPS11] J. Mattingly, N. Pillai, and A. Stuart. SPDE Limits of the Random Walk Metropolis Algorithm in High Dimensions. *Annals of Applied Probability*, 2011.

[MRTT53] N. Metropolis, A.W. Rosenbluth, M.N. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[MT93] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag, London, 1993.

[Nea01] R. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[NR06] P. Neal and G. Roberts. Optimal scaling for partially updating MCMC algorithms. *Annals of Applied Probability*, 16(2):475–515, 2006.

[NR08] P. Neal and G. Roberts. Optimal scaling for random walk Metropolis on spherically constrained target densities. *Methodology and Computing in Applied Probability*, 10(2):277–297, 2008.

[NR11] P. Neal and G. Roberts. Optimal scaling of random walk Metropolis algorithms with non-Gaussian proposals. *Methodology and Computing in Applied Probability*, 13(3):583–601, 2011.

[NRY12] P. Neal, G. Roberts, and W.K. Yuen. Optimal scaling of random walk Metropolis algorithms with discontinuous target densities. *Annals of Applied Probability*, 22(5):1880–1927, 2012.

[Pes73] PH Peskun. Optimum monte carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

[PG10] C. Pasarica and A. Gelman. Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, 20(1):343, 2010.

[PST] N. Pillai, A. Stuart, and A. Thiéry. On the random walk Metropolis algorithm for Gaussian random field priors and gradient flow. Submitted.

[PST12] N. Pillai, A. Stuart, and A. Thiéry. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Annals of Applied Probability*, 22(6):2320–2356, 2012.

[RC04] C. Robert and G. Casella. *Monte Carlo statistical methods*, volume 319. Springer-Verlag, 2004.

[RGG97] G. Roberts, A. Gelman, and W. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120, 1997.

[RR97] G. Roberts and J. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2(2):13–25, 1997.

[RR98] G. Roberts and J. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[RR01] G. Roberts and J. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[RR04] G. Roberts and J. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[RR12] G. Roberts and J. Rosenthal. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. Technical report, http://probability. ca/jeff/research. html, 2012.

[SC97] L. Saloff-Coste. Lectures on finite Markov chains. *Lectures on probability theory and statistics*, pages 301–413, 1997.

[SFR10] C. Sherlock, P. Fearnhead, and G. Roberts. The random walk Metropolis : linking theory and practice through a case study. *Statistical Science*, 25(2):172–190, 2010.

[Sha03] T. Shardlow. Weak convergence of a numerical method for a stochastic heat equation. *BIT Numerical Mathematics*, 43(1):179–193, 2003.

[She13] C. Sherlock. Optimal scaling of the random walk Metropolis: general criteria for the 0.234 acceptance rule. *Journal of Applied Probability*, 50(1):1–15, 2013.

[SJ89] A. Sinclair and M. Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, 1989.

[SR93] A. Smith and G. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 3–23, 1993.

[SR09] C. Sherlock and G. Roberts. Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, 15(3):774–798, 2009.

[SS12] C. Schwab and A. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inverse Problems*, 28(4):045003, 2012.

[Stu10] A. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19(1):451–559, 2010.

[SVW04] A. Stuart, J. Voss, and P. Wilberg. Conditional path sampling of SDEs and the Langevin MCMC method. *Communications in Mathematical Sciences*, 2(4):685–697, 2004.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 267–288, 1996.

[Tie94] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22(4):1758–1762, 1994.

[Tie98] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, 8(1):1–9, 1998.