Autour de l'algorithme du Langevin : extensions et applications

Nicolas Brosse

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

NNT : 2019SACLX014

ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : École polytechnique

Laboratoire d'accueil : Centre de mathématiques appliquées de Polytechnique, UMR 7641 CNRS

Spécialité de doctorat : Mathématiques appliquées

Nicolas BROSSE

Around the Langevin algorithm in high dimension: extensions and applications

Date de soutenance : 12 Juin 2019 *Lieu de soutenance :* Palaiseau

Après avis des rapporteurs : Andreas EBERLE (Institute for Applied Mathematics, Bonn)
Gabriel STOLTZ (CERMICS)

Jury de soutenance :

Alain DURMUS	(CMLA ENS Paris Saclay – Maître de conférence) Invité
Andreas EBERLE	(Institute for Applied Mathematics, Bonn – Professeur) Rapporteur
Emmanuel GOBET	(CMAP Ecole Polytechnique – Professeur) Président du jury
Éric MOULINES	(CMAP Ecole Polytechnique – Professeur) Directeur de thèse
Yann OLLIVIER	(Facebook Research, Paris – Chercheur) Examinateur
Gabriel STOLTZ	(CERMICS – Professeur) Rapporteur

Fondation mathématique FMJH Jacques Hadamard

CMAP

S'il n'y a pas de solution
C'est qu'il n'y a pas de problème.

Les Shadoks

Acknowledgements
Remerciements

Je remercie Éric Moulines et Alain Durmus pour tout ce qu'ils m'ont appris pendant ces 3 années. Ils m'ont consacré beaucoup de temps et d'énergie, et je leur en suis très reconnaissant.

I thank Andreas Eberle and Gabriel Stoltz for agreeing to be committee members and reviewing my thesis manuscript. Je remercie également Emmanuel Gobet et Yann Ollivier d'avoir accepté d'être membres de mon jury de thèse. L'enseignement d'Emmanuel Gobet à l'Ecole Polytechnique m'a beaucoup apporté, et a sans doute joué une part non négligeable dans mon intérêt pour les méthodes MCMC. The works of Andreas Eberle, Gabriel Stoltz and Yann Ollivier have marked my thesis and inspired me a lot.

I am grateful to all my co-authors: Sean Meyn, Marcelo Pereyra, Sotirios Sabanis, Carlos Riquelme, Sylvain Gelly and Alice Martin. Special thanks to Sean for a warm welcome to the University of Florida, Gainesville, during two weeks. I do not forget also the nice stories of Sotirios about Greece, in a very pleasant Greek restaurant in London. I thank Christophe Andrieu for the invitation to the MCMC workshop in the Newton Institute, Cambridge, in July 2017.

Je remercie l'ensemble du labo pour avoir créé une atmosphère conviviale et agréable: les doctorants, Antoine, Mathilde, Tristan, Jaouad, Geneviève, Belhal, Frédéric, Martin, Kevish, Alice, ... mais aussi l'aide précieuse de Pierre pour l'informatique, et des guides Alexandra, Maud et Nassera dans le labyrinthe des procédures administratives.

I thank Fabio and Ashwani for their great help about the internship at Google Brain Zurich during the autumn 2018. It was a pleasure to meet, work and discuss with all the team members there: Sylvain, Carlos, Neil, Michael, Francesco, Christina, Mario, Ilya, Michael, ...

Je remercie mes amis Olivier, Mathilde, Manu, Laure, Cécile, Elisa, Maxime, Pierre, Julien, ... pour me rappeler régulièrement que les maths c'est une chose, mais qu'il y a tout le reste !

Merci à ma famille pour tout leur soutien: Maman et ses gâteaux, Papa et ses échecs, Sophie et son violon, Alex et Claire, mais aussi les papys (oui, j'ai fini mon rapport), les mamies et leur "tu travailles trop", Antoine, Aline, Didier et Armelle, pour leur bonne humeur !

Contents

Chapter 1

Introduction

In Sections 1.1 to 1.3, we present the general context of this thesis by giving a brief introduction to Markov chains, Bayesian statistics and the unadjusted Langevin algorithm (ULA). Our contributions are divided and introduced in three main topics:

1. extensions of ULA in Section 1.4,
2. applications of ULA in Section 1.5,
3. analysis of Stochastic Gradient Langevin Dynamics (SGLD) in Section 1.6.

1.1 Some preliminaries on Markov chains

Markov chains are a class of stochastic processes commonly used to model many random systems in signal processing and control theory. A Markov chain is a sequence of random variables $(X_k)_{k\in\mathbb{N}}$ defined on an appropriate filtered space $(\mathcal{F}_k)_{k\in\mathbb{N}}$ such that the law of X_{n+1} conditioned on the filtration $\mathcal{F}_n$ is equal to the law of X_{n+1} conditioned on X_n, $\mathbb{P}$-almost surely. Heuristically, a discrete-time stochastic process has the Markov property if the past and future are independent given the present. For the theoretical aspects of Markov chains, we give here three references which each include a very extensive bibliography: [MT09; Dou+18], and [LP17] for discrete-space Markov chains.

Markov chains can appear from the modelling of various situations such as financial time series, storage or queuing models. They can also be built by the practitioner in order to sample from a target probability distribution; in that case, they are called Markov Chains Monte Carlo algorithms (MCMC) and have become increasingly popular these last three decades.

In this thesis, we study Markov chains with values in $\mathbb{R}^d$ endowed with the Borel sigma algebra $\mathcal{B}(\mathbb{R}^d)$. A homogeneous Markov chain $(X_k)_{k\in\mathbb{N}}$ is characterized by its kernel $R : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ which satisfies: 1) for all $x \in \mathbb{R}^d$, $R(x,\cdot)$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$, 2) for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $x \mapsto R(x,\mathsf{A})$ is a measurable function. For any probability measure μ on $\mathcal{B}(\mathbb{R}^d)$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, we denote by $\mu R(\mathsf{A}) = \int_{\mathbb{R}^d} R(x,\mathsf{A})\mu(\mathrm{d}x)$

and for any $k \in \mathbb{N}$, $x \in \mathbb{R}^d$, $R^{k+1}(x, \mathsf{A}) = \int_{\mathbb{R}^d} R(x, \mathrm{d}y) R^k(y, \mathsf{A})$. For μ a probability measure on $\mathcal{B}(\mathbb{R}^d)$, and f a μ-integrable function, $\mu(f) = \int_{\mathbb{R}^d} f(x)\mu(\mathrm{d}x)$. For $x \in \mathbb{R}^d$ and f integrable under $R(x, \cdot)$, we denote by $Rf(x) = \int_{\mathbb{R}^d} R(x, \mathrm{d}y) f(y)$.

A question of major interest is to know if the Markov chain $(X_k)_{k\in\mathbb{N}}$ of kernel R has a (unique) invariant probability measure π satisfying $\pi R = \pi$. A related concept, sometimes easier to check in practice, is the notion of reversibility. A probability measure π is said to be reversible with respect to R, if for all $\mathsf{A}, \mathsf{B} \in \mathcal{B}(\mathbb{R}^d)$, $\int_{\mathsf{A}} \pi(\mathrm{d}x) R(x, \mathsf{B}) = \int_{\mathsf{B}} \pi(\mathrm{d}x) R(x, \mathsf{A})$. Reversibility implies invariance. If π is invariant for R, the next step is the analysis of the convergence rate of $\mu_0 R^k$ to π when $k \to +\infty$ for any initial probability measure μ_0.

In particular, MCMC methods consist in building an appropriate Markov chain, preferably easy to simulate, such that it has a unique invariant probability measure π: the target distribution. The hope is that, for any initial probability measure μ_0, for $n \in \mathbb{N}$ large enough, $\mu_0 R^n$ is approximately equal to π. In that case, $(X_k)_{k\geq n}$ are (correlated) samples approximately drawn from π.

Let us illustrate these different concepts through one of the simplest examples of MCMC algorithms: the Random Walk Metropolis (RWM) algorithm with a Gaussian proposal. Let π be a target probability measure on $\mathcal{B}(\mathbb{R}^d)$ with density with respect to the Lebesgue measure also denoted by π, such that for all $x \in \mathbb{R}^d$, $\pi(x) > 0$. The Markov chain $(X_k)_{k\in\mathbb{N}}$ associated to the RWM algorithm is defined for $\sigma > 0$ and $k \in \mathbb{N}$ by

$$X_0 \text{ drawn from the initial probability measure } \mu_0 \ ,$$
$$X_{k+1} = \begin{cases} X_k + \sigma W_{k+1}, & \text{with probability } \min\left(1, \pi(X_k + \sigma W_{k+1})/\pi(X_k)\right) \ , \\ X_k, & \text{otherwise,} \end{cases}$$

where $(W_k)_{k\in\mathbb{N}}$ is an i.i.d. sequence of d-dimensional standard Gaussian vectors. The kernel R of the RWM algorithm is given for $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$R(x, \mathsf{A}) = \int_{\mathsf{A}} \min\left(1, \frac{\pi(y)}{\pi(x)}\right) \mathrm{e}^{-\|y-x\|^2/(2\sigma^2)} \frac{\mathrm{d}y}{(2\pi\sigma^2)^{d/2}} + \delta_x(\mathsf{A}) \int_{\mathbb{R}^d} \max\left(0, 1 - \frac{\pi(y)}{\pi(x)}\right) \mathrm{e}^{-\|y-x\|^2/(2\sigma^2)} \frac{\mathrm{d}y}{(2\pi\sigma^2)^{d/2}} \ .$$

It is easy to check that π is reversible with respect to R.

Despite the undeniable success of MCMC methods, one of the major obstacles faced by practitioners is to know when the Markov chain reaches convergence, i.e. when $\mu_0 R^n$ is approximately equal to π in an appropriate sense, see e.g. [Gel+14, Sections 11.4 and 11.5]. Therefore, precise non-asymptotic bounds of convergence have a significant value for the practice of MCMC.

1.1.1 Convergence of Markov chains in V-total variation and Wasserstein distances

To measure the convergence of the sequence of probability measures $(\mu_0 R^k)_{k\in\mathbb{N}}$ to π for any initial probability measure μ_0, several distances between probability measures can

be used. Let $V : \mathbb{R}^d \to [1, \infty)$ be a measurable function. We define the V-total variation distance between two probability measures μ and ν as $\|\mu - \nu\|_V = \sup_{|f| \leq V} |\mu(f) - \nu(f)|$. If $V = 1$, then $\|\cdot\|_V$ is the total variation denoted by $\|\cdot\|_{\mathrm{TV}}$.

Convergence of Markov chains has been mainly studied using the V-total variation distance [MT09]. Recently, Wasserstein distance has been more and more advocated as a useful and convenient tool to study the convergence of Markov chains, see for example [Dou+18, Section 20.7] for references on the subject. To define Wasserstein distance, some notations have to be introduced. We say that ζ is a transference plan of μ and ν if it is a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for any Borel set A of $\mathbb{R}^d$, $\zeta(\mathrm{A} \times \mathbb{R}^d) = \mu(\mathrm{A})$ and $\zeta(\mathbb{R}^d \times \mathrm{A}) = \nu(\mathrm{A})$. We denote by $\Pi(\mu, \nu)$ the set of transference plans of μ and ν. Furthermore, we say that a couple of $\mathbb{R}^d$-random variables (X, Y) is a coupling of μ and ν if there exists $\zeta \in \Pi(\mu, \nu)$ such that (X, Y) are distributed according to ζ. For two probability measures μ and ν, we define the Wasserstein distance of order $p \geq 1$ (or p-Wasserstein distance) as

$$W_p(\mu, \nu) = \left(\inf_{\zeta \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, \mathrm{d}\zeta(x, y) \right)^{1/p} .$$

By [Vil09, Theorem 4.1], for all μ, ν probability measure on $\mathbb{R}^d$, there exists a transference plan $\zeta^\star \in \Pi(\mu, \nu)$ such that for any coupling (X, Y) distributed according to $\zeta^\star$, $W_p(\mu, \nu) = \mathbb{E}[\|X - Y\|^p]^{1/p}$.

Based on the V-total variation distance, we introduce here the notion of V-uniform geometric ergodicity [Dou+18, Section 15.2] in order to present a concrete example of convergence bounds, and because some of our results are expressed through this notion. Let $V : \mathbb{R}^d \to [1, \infty)$ be a measurable function. The Markov kernel R is said to be V-uniformly geometrically ergodic is there exist $C > 0$ and $\rho \in [0, 1)$ such that for all $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$,

$$\|\delta_x R^n - \pi\|_V \leq C V(x) \rho^n . \tag{1.1}$$

It is a strong notion of convergence and it can be relaxed in multiple ways. We refer to [MT09; Dou+18] for an extensive presentation of various concepts of convergence for Markov chains. To prove (1.1), operator methods are one possibility [Dou+18, Chapter 18], [HM11]. Two conditions have to be checked. First, there must exist a Doeblin set C in $\mathbb{R}^d$ and $\epsilon > 0$ such that for all $x, x' \in \mathsf{C}$,

$$\|R(x, \cdot) - R(x', \cdot)\|_{\mathrm{TV}} \leq 1 - \epsilon . \tag{1.2}$$

Inside this Doeblin set C, the Markov kernel R should have some uniformity in x, in order to be able to couple two Markov chains of kernel R, starting from two different points. To control the time spent outside this Doeblin set C, we can use a drift condition: there exist $\lambda \in [0, 1)$, $b \in [0, +\infty)$ such that

$$RV \leq \lambda V + b \mathbb{1}_{\mathsf{C}} . \tag{1.3}$$

Under some conditions on C, λ, b, (1.2) and (1.3) imply (1.1).

Wasserstein distance has shown itself to be a useful tool to study the convergence of Markov chains [Dou+18, Chapter 20]. We give a quick overview of its application in the case of the Unadjusted Langevin Algorithm in Section 1.3.

1.1.2 Weak error bounds and a central limit theorem

Once we are able to generate approximate samples $(X_k)_{k\in\{0,\dots,n-1\}}$ from a target probability measure π, and to quantify in some sense their distance to π, we can turn to their utilisation. A popular request is the computation of integrals of specific, π-integrable functions $f : \mathbb{R}^d \to \mathbb{R}^m$ under π:

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(\mathrm{d}x) \approx \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \;, \tag{1.4}$$

where $m \in \mathbb{N}^*$. Common examples cover the mean of π where $f(x) = x$ and $m = d$, and the second order moment of π with $f(x) = xx^{\mathrm{T}}$ and $m = d^2$. Obviously, the empirical average (1.4) is only an approximation of the quantity of interest $\pi(f)$, and we try to quantify as precisely as possible the error. Convergence bounds such as the V-uniform geometric ergodicity (1.1) directly translate into error bounds for (1.4). Let $(X_k)_{k\in\mathbb{N}}$ be a Markov chain of kernel R. If R is V-uniformly geometrically ergodic, we have for any initial probability measure μ_0, any $f : \mathbb{R}^d \to \mathbb{R}$ such that $|f| \leq V$ and any $N, n \in \mathbb{N}^*$,

$$\left|\mathbb{E}\left[\frac{1}{n}\sum_{k=N}^{N+n-1} f(X_k)\right] - \pi(f)\right| \leq C\mu_0(V)\frac{\rho^N}{1-\rho} \;. \tag{1.5}$$

Using the duality formula for the Wasserstein distance, see e.g. [Dou+18, Theorem 20.1.2], a similar control of the error (1.4) can be obtained. However, instead of a condition on the infinity norm of f, the function $f : \mathbb{R}^d \to \mathbb{R}$ is here assumed to be L-Lipschitz. Under this assumption, we get

$$\left|\mathbb{E}\left[\frac{1}{n}\sum_{k=N}^{N+n-1} f(X_k)\right] - \pi(f)\right| \leq \frac{L}{n}\sum_{k=N}^{N+n-1} W_2(\mu_0 R^k, \pi) \;. \tag{1.6}$$

If $W_2(\mu_0 R^k, \pi)$ decreases exponentially fast, i.e. $W_2(\mu_0 R^k, \pi) \leq C_{\mu_0}\rho^k$ for $\rho \in [0, 1)$, we recover (1.5). The upper bounds (1.5) and (1.6) are on the first order moment; the second order moment demands more work but can be handled in a similar way. These bounds are non-asymptotic and are valid for all $n \in \mathbb{N}^*$. However, the constants involved such as ρ and C are often not tight, and the obtained upper bounds are often very large and unusable in practice.

Another fruitful point of view on the approximation (1.4) is possible through the asymptotic $n \to +\infty$. Under appropriate conditions, a central limit theorem (CLT) for $f : \mathbb{R}^d \to \mathbb{R}$ can be derived:

$$n^{-1/2}\left\{\sum_{k=0}^{n-1}\left(f(X_k) - \pi(f)\right)\right\} \underset{n\to+\infty}{\Longrightarrow} \mathcal{N}\left(0, \sigma^2(f)\right) \;, \tag{1.7}$$

where $\mathcal{N}(m, \sigma^2)$ is a real Gaussian random variable of mean m and variance σ^2, and $\Longrightarrow$ denotes the convergence in law on the canonical space, see e.g. [Dou+18, Chapter 21] for a precise formulation. The asymptotic variance $\sigma^2(f)$ can be expressed using various formulas; under some assumptions on R, it can be shown to be equal to

$$\sigma^2(f) = \lim_{n \to +\infty} n^{-1} \mathbb{E}_\pi \left[\left(\sum_{k=0}^{n-1} f(X_k) \right)^2 \right] .$$

Note that a CLT holds for f if R is V-uniformly geometrically ergodic and f is dominated by V, [Dou+18, Theorem 21.2.11]. This asymptotic point of view on (1.4) is a convenient way to build confidence intervals for $\pi(f)$.

1.2 A short presentation of Bayesian statistics

To give a concrete example of an application of MCMC algorithms, we introduce briefly Bayesian statistics. We refer for example to [Gel+14; Rob07; MR07] for many more resources and references on the subject. Note that the field of applications of MCMC is much wider than Bayesian statistics, but the primary goal of this thesis is to develop and extend MCMC algorithms for which Bayesian statistics is a sufficiently vast subject to illustrate our results.

Let $\mathcal{D}$ be some observed data. A common situation is when we observe $N \in \mathbb{N}^*$ pairs (x_i, y_i) for $i \in \{1, \dots, N\}$ where $x_i \in \mathbb{R}^d$ are the covariates associated to the observation y_i which can be continuous or discrete. Here, $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$. To explain the patterns in the observed data $\mathcal{D}$, we assume that there exists a statistical model underlying the observations. Some popular models are the Gaussian regression model for a continuous observation y, and the logistic regression model for a binary y:

$$\text{Gaussian regression: } p\left(y|x,\theta\right) = (2\pi\sigma^2)^{-1/2} \mathrm{e}^{-(y - x^{\mathrm{T}}\theta)^2/(2\sigma^2)} ,$$
$$\text{Logistic regression: } p\left(y|x,\theta\right) = (1 + \mathrm{e}^{-x^{\mathrm{T}}\theta})^{-y} (1 + \mathrm{e}^{x^{\mathrm{T}}\theta})^{y-1} .$$

In both cases, the models are parametric, with a parameter $\theta \in \mathbb{R}^d$. For the Gaussian regression, the variance σ^2 can also be considered as a parameter; here, the variance is assumed to be known. For the logistic regression, the observation y can take only two values 0 or 1. Assuming that the observations are i.i.d. , the probability of observing the data $\mathcal{D}$ given the parameter θ is given by

$$\text{Gaussian regression: } p\left(\mathcal{D}|\theta\right) = (2\pi\sigma^2)^{-N/2} \prod_{i=1}^{N} \mathrm{e}^{-(y_i - x_i^{\mathrm{T}}\theta)^2/(2\sigma^2)} ,$$
$$\text{Logistic regression: } p\left(\mathcal{D}|\theta\right) = \prod_{i=1}^{N} (1 + \mathrm{e}^{-x_i^{\mathrm{T}}\theta})^{-y_i} (1 + \mathrm{e}^{x_i^{\mathrm{T}}\theta})^{y_i - 1} .$$

In the Bayesian paradigm, the parameter θ is a random variable, drawn according to a prior distribution p_0 on $\mathcal{B}(\mathbb{R}^d)$. Under some weak assumptions on the model and the

prior distribution, by Bayes' rule, the posterior distribution is given for all $\theta \in \mathbb{R}^d$ by

$$p\left(\theta|\mathcal{D}\right) = \frac{p\left(\mathcal{D}|\theta\right) p_0(\theta)}{Z} \ , \quad Z = \int_{\mathbb{R}^d} p\left(\mathcal{D}\middle|\tilde{\theta}\right) p_0(\mathrm{d}\tilde{\theta}) \in (0, +\infty) \ . \tag{1.8}$$

Going back to the examples of the Gaussian and logistic regressions, a Gaussian prior for θ is a simple conceivable option for p_0. A practical limitation of Bayesian statistics is the necessity to sample from the posterior distribution $\theta \mapsto p\left(\theta|\mathcal{D}\right)$. Except for conjugate distributions, sampling from the posterior (1.8) is in general a difficult problem, because the normalizing constant Z is unknown. MCMC algorithms enable to target probability distributions with unknown normalizing constants, and are thus particularly adapted to Bayesian statistics. Note that this situation occurs also in other fields such as molecular dynamics and statistical physics, see e.g. [LSR10].

In Chapter 5, we propose an algorithm to compute normalizing constants Z with precise theoretical guarantees. An important application for Bayesian statistics is related to the computation of Bayes factors. In the Gaussian and logistic regressions presented above, we assume that the data $\mathcal{D}$ comes from a specified model; this assumption may be wrong. A reasonable approach consists in suggesting different models and comparing them given the observed data. Consider two different (parametric) statistical model $\mathcal{M}_1$ and $\mathcal{M}_2$, with respective parameters $\theta_1 \in \mathbb{R}^{d_1}$ and $\theta_2 \in \mathbb{R}^{d_2}$. The Bayes factor B_{12} between $\mathcal{M}_1$ and $\mathcal{M}_2$ is defined as

$$B_{12} = \frac{p\left(\mathcal{M}_1|\mathcal{D}\right)}{p\left(\mathcal{M}_2|\mathcal{D}\right)} = \frac{p\left(\mathcal{D}|\mathcal{M}_1\right) p_0(\mathcal{M}_1)}{p\left(\mathcal{D}|\mathcal{M}_2\right) p_0(\mathcal{M}_2)} \ ,$$

where p_0 is a prior distribution on the models considered. Assume that we only look at 2 models with a uniform prior: $p_0(\mathcal{M}_1) = p_0(\mathcal{M}_2) = 1/2$. Besides, we have

$$p\left(\mathcal{D}|\mathcal{M}_1\right) = \int_{\mathbb{R}^{d_1}} p\left(\mathcal{D}|\theta_1, \mathcal{M}_1\right) p\left(\mathrm{d}\theta_1|\mathcal{M}_1\right) \ ,$$

and similarly for $\mathcal{M}_2$. $\theta_1 \mapsto p\left(\theta_1|\mathcal{M}_1\right)$ is a prior distribution on the parameter θ_1 under the model $\mathcal{M}_1$. Therefore, computing the Bayes factor B_{12} requires to evaluate two normalizing constants $p\left(\mathcal{D}|\mathcal{M}_1\right)$ and $p\left(\mathcal{D}|\mathcal{M}_2\right)$ and to calculate their ratio.

For illustrative purposes, we consider the example of a binomial distribution. We compare a model $\mathcal{M}_1$ where the probability of success is $\theta_1 = 1/2$ and another model $\mathcal{M}_2$ where θ_2 is unknown and we take a prior distribution for θ_2 that is uniform on $[0, 1]$. We take a sample $\mathcal{D}$ of 200, and find 115 successes and 85 failures. We have then

$$p\left(\mathcal{D}|\mathcal{M}_1\right) = \binom{200}{115} (1/2)^{200} \approx 0.005956 \ ,$$

$$p\left(\mathcal{D}|\mathcal{M}_2\right) = \int_0^1 \binom{200}{115} \theta_2^{115} (1-\theta_2)^{85} \mathrm{d}\theta_2 = 1/201 \approx 0.004975 \ .$$

The Bayes factor B_{12} is approximately equal to 1.197 which tends to indicate that both models are equally likely to interpret the data.

1.3 The unadjusted Langevin algorithm and avatars

1.3.1 The unadjusted Langevin algorithms

The overdamped Langevin algorithm In this Section, we present the Unadjusted Langevin Algorithm (ULA), also called the Langevin Monte Carlo (LMC) algorithm: an MCMC algorithm which is the main focus of this thesis. Consider π, a target probability measure on $\mathbb{R}^d$ with density w.r.t. the Lebesgue measure given for all $x \in \mathbb{R}^d$ by $\pi(x) = \mathrm{e}^{-U(x)}/Z$, where $U : \mathbb{R}^d \to (-\infty, +\infty]$ is a measurable function and $Z = \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y \in (0, +\infty)$ is an unknown normalizing constant. U is usually referred to as the potential associated with π. Assume for the moment that U is continuously differentiable. Then, the unadjusted Langevin algorithm introduced in [Erm75; Par81] (see also [RT96]) can be used to sample from π. This algorithm is based on the overdamped Langevin stochastic differential equation (SDE) associated with U,

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \;, \tag{1.9}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Under mild assumptions on ∇U, this SDE has a unique strong solution $(Y_t)_{t\geq 0}$ and defines a strong Markovian semigroup $(P_t)_{t\geq 0}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which is ergodic with respect to π. Since simulating exact solutions of (5.7) is in general computationally impossible or very hard, ULA considers the Euler-Maruyama discretization associated with (5.7) to approximate samples from π. Precisely, ULA constructs the discrete-time Markov chain $(X_k)_{k\geq 0}$, started at X_0, given for $k \in \mathbb{N}$ by:

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} W_{k+1} \;, \tag{1.10}$$

where $\gamma > 0$ is the stepsize and $(W_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. standard Gaussian d-dimensional vectors; the process $(X_k)_{k\geq 0}$ is used as approximate samples from π.

If ∇U is globally Lipschitz, the Markov chain (1.10) has a unique invariant probability measure π_γ which is in general different from the target distribution π. Note that if ∇U is not globally Lipschitz, the Markov chain may not have an invariant probability measure and exhibit a transient behaviour, see e.g. [MSH02, Section 6]. Since π_γ is in general different from π, the usual methodology developed to analyze the convergence of Markov chains to their invariant distribution [MT09; Dou+18] is therefore not applicable. To bypass this difficulty, a series of works [Dal17b], [DM17], [DM16], [Dal17a], [DK17] directly compare the Markov chain $(X_k)_{k\in\mathbb{N}}$ to the solution $(Y_t)_{t\geq 0}$ of the continuous-time SDE (1.9) ergodic with respect to π. More precisely, if R denotes the Markov kernel of the ULA algorithm, they study the V-total variation distance and Wasserstein distance between $\mu_0 R^n$ and $\nu_0 P_{n\gamma}$ for $n \in \mathbb{N}$, where μ_0 and ν_0 are two initial probability measures.

Let us briefly sketch the idea of the proof for the Wasserstein distance when the potential U is strongly convex. Let (Y_0, X_0) be drawn according to the optimal coupling for the 2-Wasserstein distance, where $Y_0 \sim \nu_0$ and $X_0 \sim \mu_0$. Consider the pair of random variables (Y_γ, X_1) such that Y_γ is the solution at time $t = \gamma$ of the SDE

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \;,$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion and

$$X_1 = X_0 - \gamma \nabla U(X_0) + \sqrt{2} B_\gamma \; .$$

Then,

$$\begin{aligned} \|Y_\gamma - X_1\|^2 &= \|Y_0 - \gamma \nabla U(Y_0) - X_0 + \gamma \nabla U(X_0)\|^2 + O(\gamma^2) \\ &\leq \rho_\gamma \|Y_0 - X_0\|^2 + O(\gamma^2) \; , \end{aligned}$$

where $\|\cdot\|$ denotes the Euclidian norm and $\rho_\gamma \in [0,1)$, see e.g. [Dal17a, Section 6]. Since the 2-Wasserstein distance is the infimum on all the couplings between Y_γ and X_1 of $\|Y_\gamma - X_1\|^2$, we obtain by recurrence an upper bound on the 2-Wasserstein distance between $\mu_0 R^n$ and $\nu_0 P_{n\gamma}$ for all $n \in \mathbb{N}$. As a particular case, taking $\nu_0 = \pi$, the invariant measure for the semigroup $(P_t)_{t\geq 0}$, we get an upper bound on $W_2(\mu_0 R^n, \pi)$. When U is strongly convex, it is known that $W_2(\mu_0 R^n, \pi) \leq \epsilon$ for a number of iterations $n \gtrsim d/\epsilon^2$, see e.g. [Dal17a, Theorem 1].

Another interpretation consists in viewing the ULA algorithm as an optimisation algorithm on the space of measures endowed with the 2-Wasserstein distance, see [DMM18; Wib18].

In Section 1.1, we highlighted that practitioners are often interested in estimating the integral of specific functions $f : \mathbb{R}^d \to \mathbb{R}$ under the target distribution π: $\pi(f)$, see (1.4). In the special case of ULA, it is possible to go one step further, still by comparing ULA to its continuous-time dynamic counterpart (1.9), but this time, in terms of generators. The generator $\mathcal{L}$ associated to the semigroup $(P_t)_{t\geq 0}$ of the Langevin diffusion (1.9) is defined for any smooth function $f : \mathbb{R}^d \to \mathbb{R}$ compactly supported by

$$\mathcal{L} f = \lim_{t \to 0^+} (P_t f - f)/t = - \langle \nabla U, \nabla f \rangle + \Delta f \; . \tag{1.11}$$

The generator of the discrete-time Markov chain (1.10) of kernel R is defined by $Rf - f$. A simple calculation shows that the two generators are very close. More precisely, let X_1 be the first step of ULA (1.10) starting at $X_0 = x$. A Taylor expansion of f around x gives

$$\mathbb{E}\left[f(X_1)\right] = f(x) + \gamma \mathcal{L} f(x) + O(\gamma^2) \; .$$

Let $\hat{f}$ be a solution of the Poisson equation associated with f, $\mathcal{L}\hat{f} = -(f - \pi(f))$. We obtain heuristically

$$\begin{aligned} \sum_{k=0}^{n-1} \{f(X_k) - \pi(f)\} &= \gamma^{-1} \sum_{k=0}^{n-1} \left\{ \hat{f}(X_k) - \mathbb{E}_{\mathcal{F}_k}\left[\hat{f}(X_{k+1})\right] \right\} + O(n\gamma) \\ &= \gamma^{-1} \left\{ \hat{f}(X_0) - \mathbb{E}_{\mathcal{F}_{n-1}}\left[\hat{f}(X_n)\right] \right\} + \gamma^{-1} \sum_{k=1}^{n-1} \left\{ \hat{f}(X_k) - \mathbb{E}_{\mathcal{F}_{k-1}}\left[\hat{f}(X_k)\right] \right\} + O(n\gamma) \; . \end{aligned}$$

The first term comes from the initial conditions and the second term is a sum of martingale increments. Taking the expectation, we get the equivalent of (1.5):

$$\left| \mathbb{E}\left[\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \right] - \pi(f) \right| \lesssim \frac{1}{n\gamma} + \gamma \; . \tag{1.12}$$

To the best of our knowledge, this methodology was first employed in [TT90]. An example of application is provided in Chapter 4.

The underdamped Langevin algorithm The overdamped Langevin SDE (1.9) can be seen as the limit of the kinetic or underdamped Langevin diffusion $(X_t, V_t)_{t\geq 0}$ on $\mathbb{R}^{2d}$ solution of the SDE:

$$\begin{aligned} \mathrm{d}X_t &= V_t \mathrm{d}t \ , \\ \mathrm{d}V_t &= -\eta V_t \mathrm{d}t - u\nabla U(X_t)\mathrm{d}t + \sqrt{2\eta u}\, \mathrm{d}B_t \ , \end{aligned} \tag{1.13}$$

when the friction coefficient $\eta > 0$ tends to infinity, [Pav14, Section 6.5]. $u > 0$ is the inverse mass and $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. If U is gradient Lipschitz, there exists a unique strong solution $(X_t, V_t)_{t\geq 0}$ for any initial conditions (X_0, V_0). We refer to [EGZ17] for references on the subject and precise rates of convergence using a "sticky" coupling approach. When U is strongly convex, the semigroup associated to the unique strong solution $(X_t, V_t)_{t\geq 0}$ of (1.13) has a unique invariant probability measure π_{kin} of density with respect to the Lebesgue measure proportional to

$$\pi_{\mathrm{kin}}(x, v) \propto \exp\left(-U(x) - \|v\|^2/(2u)\right) \ .$$

An appropriate discretization of (1.13) enables to draw approximate samples $(X_k)_{k\in\mathbb{N}}$ of π, such that $W_2(X_n, \pi) \leq \epsilon$ for $n \gtrsim \sqrt{d}/\epsilon$, see [Che+18; DR18]. The dependence of the number of iterations with respect to d and ϵ is thus better for the kinetic Langevin algorithm compared to ULA. Note that we compare first order discretization schemes: each step takes time linear in d. Second order schemes are studied in [DK19, Section 4] and [DR18, Section 4].

The discretization of the continuous-time SDEs (1.9) and (1.13) enables an easy simulation of the associated MCMC algorithms. However, their invariant distributions are in general different from the target distributions π and π_{kin}. If the desired precision on the approximate samples $(X_k)_{k\in\mathbb{N}}$ is less than $\epsilon > 0$ (in Wasserstein distance for example), the step size γ has to be chosen of the order ϵ^2 or ϵ, which directly reflects on the number of iterations proportional to ϵ^{-2} and ϵ^{-1} respectively.

1.3.2 Adjusted Langevin algorithms

This poor dependence on ϵ, the precision parameter, can be mitigated by modifying the Langevin MCMC algorithms such that their invariant distributions are π and π_{kin}. A simple way to proceed consists in incorporating a Metropolis Hastings step in these algorithms. For ULA, we obtain the Metropolis-Adjusted Langevin Algorithm (MALA)

algorithm, introduced in [RDF78], given for $k \in \mathbb{N}$ by

$$\widetilde{X}_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} W_{k+1} \,,$$
$$X_{k+1} = \begin{cases} \widetilde{X}_{k+1}, & \text{if } U_{k+1} \leq \min\left(1, \mathrm{e}^{-\tau_\gamma(X_k, \widetilde{X}_{k+1})}\right) \\ X_k, & \text{otherwise.} \end{cases}$$
$$\text{where} \quad \tau_\gamma(x,y) = U(y) - U(x) + \frac{\|x - y + \gamma \nabla U(y)\|^2 - \|y - x + \gamma \nabla U(x)\|^2}{4\gamma} \,.$$

Bounds in total variation distance are available for the MALA algorithm. [BH13, Theorem 3.1] shows that the number of iterations for MALA is only proportional to the logarithm of the precision parameter $\epsilon > 0$, i.e. $n \gtrsim \log(1/\epsilon)^p$, where $p \in \mathbb{N}^*$. The main theoretical difficulty in the analysis of MALA lies in the accept/reject step: the probability of rejecting a potential step has to be controlled. In [BH13], the analysis relies on a modified MALA algorithm restrained on a well-chosen compact set where it is possible to control the acceptance ratio uniformly. A comparison between the original MALA algorithm and the modified one gives then the result. A different path is followed by [Dwi+18]: since π is reversible with respect to the Markov kernel of MALA, it is possible to rely on tools based on the conductance of the Markov chain to prove convergence rates. The conductance, also known as the bottleneck ratio or Cheeger constant, is a notion originally defined and applied for discrete-space Markov chains [LP17, Section 7.2]. An extension of this concept, the μ-conductance, that disregards small sets is formulated in [LS90] and an application for general state space Markov chains is given in [LS93, Theorem 1.4]. Based on these founding principles, [Dwi+18] shows that for a strongly convex U, about $d \log(1/\epsilon)$ iterations are sufficient to obtain samples at TV distance at most ϵ from π. Note that a restriction on the initial probability measure μ_0 is required for this result. A recent bound on the Kantorovich or Wasserstein distance (with an appropriate distance function) is formulated in [EM18, Section 2.5] but the dependence on the parameters d and ϵ is intricate.

The MALA algorithm is a particular case of the Hybrid/Hamiltonian Monte Carlo algorithm (HMC). HMC targets the same probability measure as the kinetic Langevin diffusion: $\pi_{\mathrm{HMC}}(x,v) \propto \mathrm{e}^{-H(x,v)}$ on $\mathbb{R}^{2d}$ where $H(x,v) = U(x) + \|v\|^2/2$ is the Hamiltonian function. Note that we are only interested in the first component which is a Markov chain on $\mathbb{R}^d$ with invariant probability measure π. Using the terminology of [BEZ18], one step of the exact HMC takes as inputs an initial position $x \in \mathbb{R}^d$ and a duration parameter $T > 0$, and outputs a final position by taking the following steps

1. Draw an initial velocity $\xi \sim \mathcal{N}(0, \mathrm{Id})$,
2. Run the Hamiltonian dynamics associated to the Hamiltonian function H for a duration T with initial position x and initial velocity ξ,
3. Output the final position of this Hamiltonian dynamics.

In practice, the Hamiltonian dynamics is approximated by a numerical integrator, such as the leap-frog or Störmer-Verlet integrator, that keeps the reversibility and volume-preserving properties of the Hamiltonian flow. A Metropolis Hastings step is added to

remove the bias due to time discretization error; the resulting algorithm is called numerical HMC. This method has been first introduced in [Dua+87] and partially analyzed from a mathematical viewpoint in [Sch99]. The study of HMC has led to numerous research works, see [BEZ18; DMS17; BS18; MV18; TV17; HJ17; MS17; BS17] and references therein. [DMS17] shows the geometric ergodicity of HMC under certain conditions. For a potential U strongly convex outside of a ball, [BEZ18] provides precise rates of convergence in Wasserstein distance, using a coupling approach. Note that the dependence of the bounds with respect to the dimension d for the numerical HMC is however not clear [BEZ18, Remark 2.5]. For strongly convex potentials U and under additional assumptions, [MS17; MV18] establish that numerical HMC can give approximate samples from π with a number of iterations of the order $d^{1/4}$, to be compared with the kinetic Langevin Monte Carlo scaling as $d^{1/2}$.

1.3.3 Remarks

Nonreversible Langevin dynamic The Langevin SDE given in (1.9) defines a semigroup $(P_t)_{t\geq 0}$ which is reversible with respect to the invariant distribution π. The associated generator $\mathcal{L}$, see (1.11), is self-adjoint in $\mathrm{L}^2(\pi)$, i.e. for all $f, g : \mathbb{R}^d \to \mathbb{R}$ twice continuously differentiable and compactly supported, $\int_{\mathbb{R}^d} f\mathcal{L}g\mathrm{d}\pi = \int_{\mathbb{R}^d} g\mathcal{L}f\mathrm{d}\pi$. Nonreversible dynamics that leave the distribution π invariant, have been introduced and analysed in [HHS93; HHS05; DLP16; RS15a; RS15b; WHC14]. They consist in considering the SDE

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + F(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ ,$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is a divergence-free vector field with respect to π, i.e. $\nabla \cdot (\pi F) = 0$. Under this constraint and the non-explosion condition, π is still the invariant distribution of the associated semigroup. It has been shown that nonreversible dynamics (with F different from 0) improve the spectral gap in $\mathrm{L}^2(\pi)$ and decrease the asymptotic variance for the time averages. Note that ULA is nonreversible, the Euler discretization scheme breaks the reversibility of the associated SDE. Several works have studied the discrete-time equivalent of making a reversible dynamic, nonreversible, see e.g. [DPZ17] and references therein. In particular, piecewise deterministic Markov processes are such a class of algorithms; we refer to [DGM18] for references on the subject.

Convexity and beyond In this thesis, we often make the assumption that the potential U is convex or strongly convex. Many widely-used and well-known losses used in statistical learning are convex, which partly justifies this assumption, see e.g. [BV04]. The analysis under this constraint is obviously a first step in a better understanding of the ULA algorithm, and recent works have begun studying the non-convex setting, including [DM17; EGZ17; BEZ18]. Furthermore, a series of works has tackled non-convex optimization problems using discretized diffusions, see e.g. [RRT17; ZLC17; EMS18].

Big data setting In this thesis, we focus on the ULA algorithm because we are mainly interested in sampling from high-dimensional distributions in the big data setting,

where the number of observations (for Bayesian statistics) is very large. In that context, computing the gradient ∇U at each iteration of the Langevin algorithm or HMC is very expensive. Akin to Stochastic Gradient Descent (SGD) in optimization, versions of ULA and kinetic Langevin algorithm have been developed where the gradient of U is estimated by subsampling the data, thus reducing drastically the computational cost of one iteration. Note that this methodology is difficult to transpose to Metropolis Hastings type algorithms such as MALA or HMC, see [Bet15]. Moreover, the kinetic Langevin algorithm does not bring any improvement over the overdamped version when the gradient is noisy, see [Che+18, Section 2.2.1]. All these reasons have led us to concentrate on the ULA algorithm in a first place. Obviously, the different methodologies and contributions introduced below may be adapted to more sophisticated algorithms such as kinetic Langevin and HMC and might be the subjects of future works.

1.4 Extensions of the unadjusted Langevin algorithm

In Part I of this thesis, two limitations of the ULA algorithm defined in (1.10) are addressed. First, ULA is well defined and feasible if the potential U is continuously differentiable on $\mathbb{R}^d$: it cannot be directly applied to a distribution π restricted to a compact convex set. However, many statistical inference problems involve estimating parameters subject to constraints on the parameter space. The MYULA algorithm proposed in [DMP18] enables to draw approximate samples from distributions with compact support by regularizing appropriately the potential U. In Chapter 3, we derive precise bounds of convergence in total variation norm and Wasserstein distance for this algorithm.

Second, when the potential U grows too fast at infinity, i.e. $\|U(x)\| \gtrsim \|x\|^{2+\alpha}$ when $\|x\| \to +\infty$ and with $\alpha > 0$, ULA is unstable and may diverge with positive probability. Inspired by some recent works on discretization of SDEs with superlinear drift coefficients [HJK12; Sab13], we propose a new algorithm in Chapter 4, the tamed ULA, and provide convergence guarantees in V-total variation distance and 2-Wasserstein distance.

1.4.1 Sampling from a distribution with compact support: MYULA

Consider in a Bayesian setting a posterior distribution π with bounded support. Some examples include truncated data problems which arise naturally in failure and survival time studies [KM05], ordinal data models [JA06], constrained Lasso and ridge regressions [Cel+12], Latent Dirichlet Allocation [BNJ03], and non-negative matrix factorization [PBJ14]. Drawing samples from such constrained distributions is a challenging problem that has been investigated in many papers; see [GSL92], [PP14], [LS15], [BEL15], [Hsi+18]. All these works are based on efficient Markov Chain Monte Carlo methods to approximate the posterior distribution; however, with the exception of [BEL15] and [Hsi+18], these methods are not theoretically well understood and do not provide any theoretical guarantees on the estimations delivered.

A modification of ULA has been proposed in [DMP18] to sample from a non-smooth log-concave probability distribution on $\mathbb{R}^d$. This method named MYULA is mainly based

on a regularised version of the target distribution π that enjoys a number of favourable properties that are useful for MCMC simulation. In this study, we analyse the complexity of this algorithm when applied to log-concave distributions constrained to a convex set, with a focus on complexity as the dimension of the state space increases. More precisely, we establish explicit bounds in total variation norm and in Wasserstein distance of order 1 between the iterates of the Markov kernel defined by the algorithm and the target density π.

Note that the Metropolis-Adjusted Langevin Algorithm (MALA) is a viable alternative to ULA to sample from a constrained distribution. However, no precise convergence bounds were available for MALA. Subsequent to this work, [Dwi+18] provided tight convergence bounds for MALA and analyzing again the problem of sampling a constrained distribution under this perspective would be an interesting research work.

Let $\mathsf{K} \subset \mathbb{R}^d$ be a compact convex set such that $\mathrm{B}(0,r) \subset \mathsf{K} \subset \mathrm{B}(0,R)$ and $\iota_{\mathsf{K}} : \mathbb{R}^d \to \{0, +\infty\}$ be the (convex) indicator function of K, defined for $x \in \mathbb{R}^d$ by,

$$\iota_{\mathsf{K}}(x) = \begin{cases} +\infty & \text{if } x \notin \mathsf{K}, \\ 0 & \text{if } x \in \mathsf{K} \ . \end{cases}$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex, gradient Lipschitz, continuously differentiable function. Consider a probability density π associated to a potential $U : \mathbb{R}^d \to (-\infty, +\infty]$ of the form $U = f + \iota_{\mathsf{K}}$. To apply ULA, [DMP18] suggested to regularize U in such a way that

1. the convexity of U is preserved (this property is key to the theoretical analysis of the algorithm),
2. the regularisation of U is continuously differentiable and gradient Lipschitz (this regularity property is key to the algorithm's stability),
3. the resulting approximation is close to π (e.g. in total variation norm).

The tool used to construct such an approximation is the Moreau-Yosida envelope of ι_{K}, $\iota_{\mathsf{K}}^{\lambda} : \mathbb{R}^d \to \mathbb{R}_+$ defined for $x \in \mathbb{R}^d$ by,

$$\iota_{\mathsf{K}}^{\lambda}(x) = \inf_{y \in \mathbb{R}^d} \left(\iota_{\mathsf{K}}(y) + (2\lambda)^{-1} \|x - y\|^2 \right) = (2\lambda)^{-1} \|x - \mathrm{proj}_{\mathsf{K}}(x)\|^2 \ ,$$

where $\lambda > 0$ is a regularization parameter and $\mathrm{proj}_{\mathsf{K}}$ is the projection onto K. $\iota_{\mathsf{K}}^{\lambda}$ is convex and continuously differentiable with gradient given for all $x \in \mathbb{R}^d$ by:

$$\nabla \iota_{\mathsf{K}}^{\lambda}(x) = \lambda^{-1}(x - \mathrm{proj}_{\mathsf{K}}(x)) \ .$$

$\nabla \iota_{\mathsf{K}}^{\lambda}$ is λ^{-1}-Lipschitz. Adding f to $\iota_{\mathsf{K}}^{\lambda}$ leads to the regularization $U^{\lambda} : \mathbb{R}^d \to \mathbb{R}$ of the potential U defined for all $x \in \mathbb{R}^d$ by $U^{\lambda}(x) = f(x) + \iota_{\mathsf{K}}^{\lambda}(x)$. The associated probability measure π^{λ} on $\mathbb{R}^d$ given for all $x \in \mathbb{R}^d$ by

$$\pi^{\lambda}(x) = \mathrm{e}^{-U^{\lambda}(x)} \Big/ \int_{\mathbb{R}^d} \mathrm{e}^{-U^{\lambda}(y)} \mathrm{d}y \ ,$$

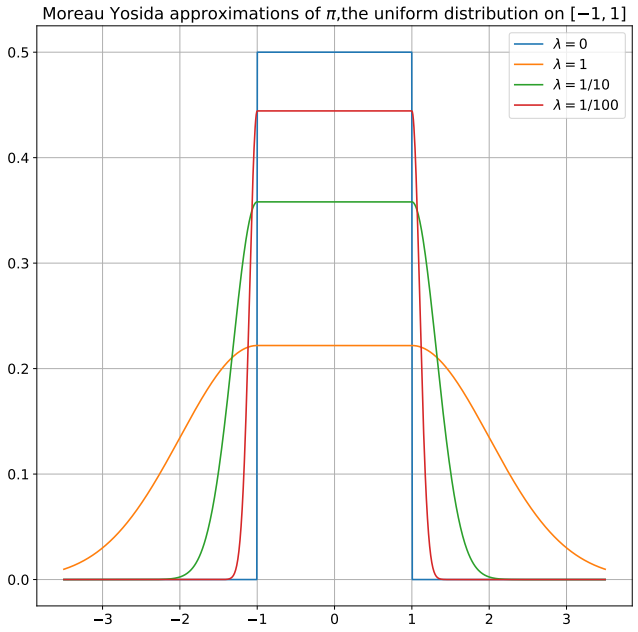

Figure 1.1: Moreau-Yosida approximations π^λ of $\pi = \mathbb{1}_{[-1,1]}/2$ for $\lambda \in \{1, 0.1, 0.01, 0\}$.

is well-defined and log-concave. Furthermore, U^λ is gradient Lipschitz and continuously differentiable with ∇U^λ given for all $x \in \mathbb{R}^d$ by

$$\nabla U^\lambda(x) = -\nabla \log \pi^\lambda(x) = \nabla f(x) + \lambda^{-1}(x - \mathrm{proj}_{\mathsf{K}}(x)) \, .$$

In Figure 1.1, the Moreau-Yosida approximation of the uniform distribution on $\mathsf{K} = [-1, 1]$ is plotted for different values of λ. We observe that as λ decreases, π^λ becomes a better approximation of π.

The algorithm proposed in [DMP18] then proceeds by using the Euler-Maruyama discretization of the Langevin equation associated with U^λ, with π^λ as proxy, to generate approximate samples from π. It uses the Markov chain $(X_k)_{k\in\mathbb{N}}$, started at X_0, given for all $k \in \mathbb{N}$ by

$$X_{k+1} = (1 - \frac{\gamma}{\lambda})X_k - \gamma \nabla f(X_k) + \frac{\gamma}{\lambda} \mathrm{proj}_{\mathsf{K}}(X_k) + \sqrt{2\gamma} W_{k+1} \, , \tag{1.14}$$

where $(W_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. standard Gaussian d-dimensional vectors and $\gamma > 0$ is the stepsize. Note that this algorithm assumes that the projection on K is relatively easy to compute. An example of a trajectory of the MYULA algorithm in 2 dimensions is represented in Figure 1.2 where the target distribution π is the uniform distribution on the rectangle $\mathsf{K} = [0, 5] \times [0, 1]$. Some samples end up outside of K but they are repelled from going to far from the boundary because of the form of the potential $U^\lambda = (2\lambda)^{-1} \|x - \mathrm{proj}_{\mathsf{K}}(x)\|^2$.

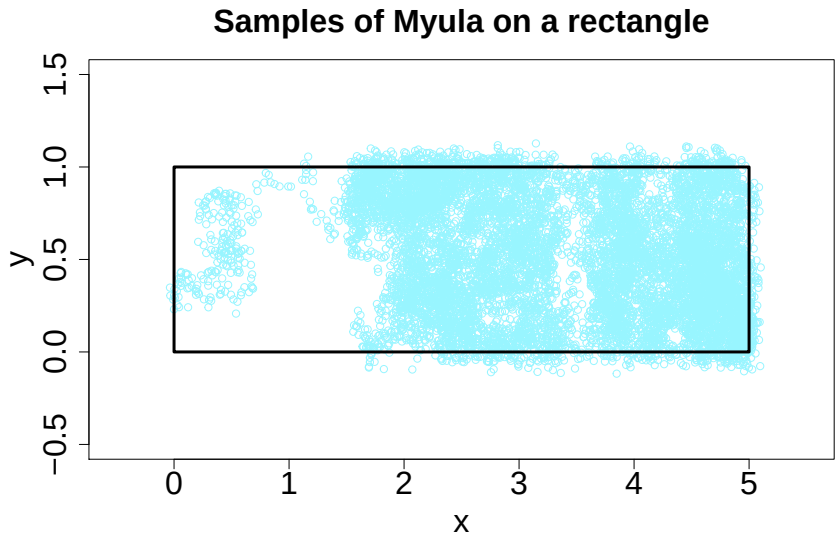

Figure 1.2: A trajectory of the MYULA algorithm targeting the uniform distribution on $\mathsf{K} = [0,5] \times [0,1]$ for $\gamma = 0.01$ and $\lambda = 0.001$.

The kernel of the homogeneous Markov chain defined by (1.14) is given for $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by,

$$R_{\gamma,\lambda}(x,\mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(-(4\gamma)^{-1} \left\| y - x + \gamma \nabla U^{\lambda}(x) \right\|^2\right) \mathrm{d}y \; .$$

Since the target density for the Markov chain (1.14) is the regularized measure π^{λ} and not π, the algorithm is named the Moreau-Yosida regularized Unadjusted Langevin Algorithm (MYULA).

The main result proved in Chapter 3 is that for all $\varepsilon > 0$ and $x \in \mathbb{R}^d$, there exist $\lambda, \gamma_0 > 0$ such that for all $\gamma \in (0, \gamma_0]$,

$$\|\delta_x R^n_{\gamma,\lambda} - \pi\|_{\mathrm{TV}} \leq \varepsilon \quad \text{for} \quad n \gtrsim d^5 \; .$$

An interesting point of this study is the dependence of the regularization parameter λ with respect to the dimension d. We show that λ should be of the order d^{-2} to have a non-trivial bound on $\|\pi^{\lambda} - \pi\|_{\mathrm{TV}}$ when $d \to +\infty$. This dependence may appear severe; geometry and probability measures in high dimension often exhibit counter-intuitive behaviours. For example, most of the volume of a high dimensional ball of radius 1 is contained in a narrow annulus near its surface of thickness of order $1/d$. Similarly, the volume of a high dimensional cube is mostly contained in its corners.

Subsequent to this work, [Hsi+18] suggested a mirrored Langevin Dynamics algorithm to sample from constrained domains in $\mathbb{R}^d$. When π is strongly log-concave, the authors show that it is possible to draw approximate samples from π by iterating at

most about d times an appropriate Markov chain defined on the unconstrained space $\mathbb{R}^d$. However, this general result is only existential and practical algorithms have to be developed on a case-by-case basis. The authors give an explicit application on the simplex. Besides, contrary to what is reported in [Hsi+18, Table 1], our result holds for non-strongly log-concave distributions π and is thus more general than their framework.

1.4.2 Sampling from superquadratic potentials U: TULA

The ULA algorithm is unstable if ∇U is superlinear i.e. $\liminf_{\|x\|\to+\infty} \|\nabla U(x)\| / \|x\| = +\infty$, see [RT96, Theorem 3.2], [MSH02] and [HJK11]. This is illustrated with a particular example in [MSH02, Lemma 6.3] where the SDE (1.9) is considered in one dimension with $U(x) = x^4/4$ along with the associated Euler discretization (1.10). It is shown that for all $\gamma > 0$, if $\mathbb{E}\left[X_0^2\right] \geq 2/\gamma$, then $\lim_{n\to+\infty} \mathbb{E}\left[X_n^2\right] = +\infty$. Moreover, the sample path $(X_n)_{n\in\mathbb{N}}$ diverges to infinity with positive probability.

Until recently, either implicit numerical schemes, e.g. see [MSH02] and [HMS02], or adaptive stepsize schemes, e.g. see [LMS07], were used to address this problem. So called S-ROCK methods were also developed to tackle this issue, see [AL08; AC08]. In the last few years, a new generation of explicit numerical schemes, which are computationally efficient, has been introduced by "taming" appropriately the superlinearly growing drift, see [HJK12] and [Sab13] for more details.

Nonetheless, with the exception of [MSH02], these works focus on the discretization of SDEs with superlinear coefficients in finite time. We aim at extending these techniques to sample from π. To deal with the superlinear nature of ∇U, based on previous studies on the tamed Euler scheme [HJK12], [Sab13], [HJ15], we introduce the tamed ULA (TULA) defined for $k \in \mathbb{N}$ by

$$X_{k+1} = X_k - \gamma \frac{\nabla U(X_k)}{1 + \gamma \|\nabla U(X_k)\|} + \sqrt{2\gamma} W_{k+1} \;,\; X_0 = x_0 \;. \tag{1.15}$$

We denote by R_γ the associated Markov kernel. Note that in Chapter 4, we deal with a more general framework, allowing various ways to tame the superlinear drift.

We provide a simple qualitative comparison of ULA and TULA in dimension 2 with a potential $U(x) = \|x\|^4/4$, $X_0 \sim \mathcal{N}(0, \mathrm{Id})$ and $\gamma = 0.2$. Table 1.1 displays the values of the coordinates of the trajectory of ULA just before divergence. In contrast, the TULA algorithm stays stable and the trajectory centered around 0.

The potential U is assumed to be locally gradient Lipschitz, with a Lipschitz constant growing at most polynomially, i.e. for all $x, y \in \mathbb{R}^d$,

$$\|\nabla U(x) - \nabla U(y)\| \leq L \left\{1 + \|x\|^\ell + \|y\|^\ell\right\} \|x - y\| \;,$$

where $\ell, L \geq 0$. Under an additional weak assumption, we show in Chapter 4 that there exist a function $V : \mathbb{R}^d \to [1, +\infty)$, $\lambda \in (0, 1)$ such that for all γ small enough, $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$,

$$\left\|\delta_x R_\gamma^n - \pi\right\|_V \lesssim n\gamma\lambda^{n\gamma} V(x) + \sqrt{\gamma} \;.$$

x_1	x_2
0.219694	0.543533
−0.557107	0.633944
−3.11446	1.12509
3.91808	−0.0845583
−7.76336	0.392261
85.5875	−3.7498
−125545	5500.15

Table 1.1: Excerpt of the coordinates of ULA before divergence.

In addition of this V-uniform geometric ergodicity of the TULA Markov kernel, an upper bound on the 2-Wasserstein distance is provided:

$$W_2^2(\delta_x R_\gamma^n, \pi) \lesssim n\gamma\lambda^{n\gamma} V(x) + \gamma \,. \tag{1.16}$$

If the Hessian of U, $\nabla^2 U$, is locally β-Hölder, with $\beta \in [0,1]$, (1.16) can be improved as

$$W_2^2(\delta_x R_\gamma^n, \pi) \lesssim n\gamma^{1+\beta}\lambda^{n\gamma} V(x) + \gamma^{1+\beta} \,.$$

Under smoothness assumptions on the potential U, we recover standard bounds with respect to $n \in \mathbb{N}^*$ and $\gamma > 0$ on the weak error, see [MST10, Theorems 5.1, 5.2],

$$\left| \mathbb{E}\left[\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - \pi(f) \right] \right| \lesssim \gamma + \frac{1}{n\gamma} \,,$$

and on the mean square error

$$\mathbb{E}\left[\left(\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - \pi(f) \right)^2 \right] \lesssim \gamma^2 + \frac{1}{n\gamma} \,.$$

The proof technique is a simple adaptation from the method presented in Section 1.3 to show (1.12). In conclusion, TULA is more robust than ULA, with similar theoretical guarantees and should be preferred in practice, especially when the potential U is likely to grow fast at infinity.

1.5 Applications of the unadjusted Langevin algorithm

In Part II, we present a direct application of the ULA algorithm to compute normalizing constants with precise theoretical guarantees. In a second step, we give a new control variates methodology for the ULA, RWM and MALA algorithms, that is directly inspired by the comparison of ULA dynamics to the (overdamped) Langevin diffusion.

1.5.1 Estimating the normalizing constant of log-concave densities

Let $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable convex function such that $Z = \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \mathrm{d}x < +\infty$. Z is the (unknown) normalizing constant of the probability density π associated with the potential U, defined for $x \in \mathbb{R}^d$ by $\pi(x) = Z^{-1}\mathrm{e}^{-U(x)}$. In Chapter 5, we present and discuss a method to estimate Z with polynomial complexity in the dimension d.

Computing the normalizing constant is a challenge which has applications in Bayesian inference and statistical physics in particular. In statistical physics, Z is better known under the name of partition function or free energy [Bal07], [LSR10]. Free energy differences allow to quantify the relative likelihood of different states (microscopic configurations) and are linked to thermodynamic work and heat exchanges. In Bayesian inference, the models can be compared by the computation of the Bayes factor which is the ratio of two normalizing constants (see e.g. [Rob07, chapter 7]). This problem has consequently attracted a wealth of contribution; see for example [CSI00, chapter 5], [MR09], [FW12], [Ard+12], [D+13], [Knu+15], [ZJA15] and, for a more specific molecular simulations flavor, [LSR10]. [CLS12] considers an application originated from computational statistical physics to enhance the sampling of MCMC for mixture Bayesian posterior distributions.

Our algorithm relies on a sequence of Gaussian densities with increasing variances, combined with the precise bounds of [DM16]. Assume without loss of generality that U has a minimum $x^\star = 0$ and $U(x^\star) = 0$. Let $M \in \mathbb{N}^\star$, $\{\sigma_i^2\}_{i=0}^M$ be a positive increasing sequence of real numbers and set $\sigma_M^2 = +\infty$. Consider the sequence of functions $\{U_i\}_{i=0}^M$ defined for all $i \in \{0, \dots, M\}$ and $x \in \mathbb{R}^d$ by

$$U_i(x) = \frac{\|x\|^2}{2\sigma_i^2} + U(x) \;,$$

with the convention $1/\infty = 0$. We define a sequence of probability densities $\{\pi_i\}_{i=0}^M$ for $i \in \{0, \dots, M\}$ and $x \in \mathbb{R}^d$ by

$$\pi_i(x) = Z_i^{-1}\mathrm{e}^{-U_i(x)} \;, \qquad Z_i = \int_{\mathbb{R}^d} \mathrm{e}^{-U_i(y)}\mathrm{d}y \;.$$

By definition, note that $U_M = U$, $Z_M = Z$ and $\pi_M = \pi$. As in the multistage sampling method [GM98, Section 3.3], we use the following decomposition

$$\frac{Z}{Z_0} = \prod_{i=0}^{M-1} \frac{Z_{i+1}}{Z_i} \;.$$

Z_0 is estimated by choosing σ_0^2 small enough so that π_0 is sufficiently close to a Gaussian distribution of mean 0 and covariance $\sigma_0^2 \,\mathrm{Id}$. For $i \in \{0, \dots, M-1\}$, the ratio Z_{i+1}/Z_i may be expressed as

$$\frac{Z_{i+1}}{Z_i} = \int_{\mathbb{R}^d} g_i(x)\pi_i(x)\mathrm{d}x = \pi_i(g_i) \;,$$

where $g_i : \mathbb{R}^d \to \mathbb{R}_+$ is defined for all $x \in \mathbb{R}^d$ by

$$g_i(x) = \exp\left(a_i \|x\|^2\right) \;, \qquad a_i = \frac{1}{2}\left(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_{i+1}^2}\right) \;.$$

The quantity $\pi_i(g_i)$ is estimated by ULA targeting π_i. More precisely, we define M (ULA) Markov chains for $i \in \{0, \dots, M-1\}$ and $k \in \mathbb{N}$ by

$$X_{i,k+1} = X_{i,k} - \gamma_i \nabla U_i(X_{i,k}) + \sqrt{2\gamma_i} W_{i,k+1} \;, \quad X_{i,0} = 0 \;,$$

where $\{(W_{i,k})_{k\in\mathbb{N}^\star}\}_{i=0}^{M-1}$ are independent i.i.d. sequences of standard Gaussian random variables and $\gamma_i > 0$ is the stepsize. For $i \in \{0, \dots, M-1\}$, consider the following estimator of Z_{i+1}/Z_i,

$$\hat{\pi}_i(g_i) = \frac{1}{n_i} \sum_{k=N_i+1}^{N_i+n_i} g_i(X_{i,k}) \;,$$

where $n_i \geq 1$ is the sample size and $N_i \geq 0$ the burn-in period. To simplify the presentation, we assume that U is m-strongly convex. In Chapter 5, we give an explicit choice of the simulation parameters

$$\mathcal{S} = \left\{ M, \{\sigma_i^2\}_{i=0}^{M-1}, \{\gamma_i\}_{i=0}^{M-1}, \{n_i\}_{i=0}^{M-1}, \{N_i\}_{i=0}^{M-1} \right\} \;,$$

such that $\hat{Z}$ the following estimator of Z,

$$\hat{Z} = (2\pi\sigma_0^2)^{d/2} (1 + \sigma_0^2 m)^{-d/2} \left\{ \prod_{i=0}^{M-1} \hat{\pi}_i(g_i) \right\} \;, \tag{1.17}$$

satisfies

$$\mathbb{P}\left(\left| \hat{Z}/Z - 1 \right| > \epsilon \right) \leq \mu \;, \quad \text{for } \mu, \epsilon \in (0,1) \;.$$

The cost of the algorithm defined by $\text{cost} = \sum_{i=0}^{M-1} \{N_i + n_i\}$ is upper bounded by $d^{5/2}$ up to logarithmic factors when U is strongly convex, and ∇U, $\nabla^2 U$ are Lipschitz. Note that in the expression of $\hat{Z}$ (1.17), Z_0 is approximated by $(2\pi\sigma_0^2)^{d/2}(1+\sigma_0^2 m)^{-d/2}$. A more refined decomposition is possible to compute Z_0:

$$Z_0 = (2\pi\sigma_0^2)^{d/2} \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \frac{\mathrm{e}^{-\|x\|^2/(2\sigma_0^2)}}{(2\pi\sigma_0^2)^{d/2}} \mathrm{d}x \;.$$

The integral can be estimated by a classical Monte Carlo method using i.i.d. samples from the Gaussian distribution of covariance matrix $\sigma_0^2 \,\mathrm{Id}$. In Figure 1.3, we display the values of the estimators $\hat{\pi}_i(g_i)$ in y-axis, with respect to the iteration $i \in \{0, \dots, M-1\}$ in x-axis for a Bayesian logistic regression model. 10 independent simulations are run at each phase $i \in \{0, \dots, M-1\}$ to measure the variability of each estimator $\hat{\pi}_i(g_i)$.

1.5.2 Diffusion approximations and control variates for MCMC

In Section 1.1, we underlined the fact that one of the main goals of MCMC methods is to estimate $\pi(f)$ for a specific, π-integrable function $f : \mathbb{R}^d \to \mathbb{R}$, see (1.4). A natural Monte Carlo estimator of $\pi(f)$ is $\hat{\pi}_n(f)$ defined for $n \in \mathbb{N}^*$ by

$$\hat{\pi}_n(f) = \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \;, \tag{1.18}$$

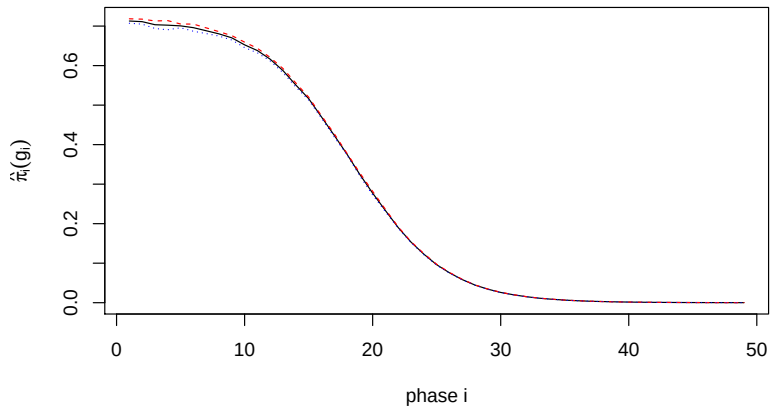

Figure 1.3: Error plot of $\hat{\pi}_i(g_i)$ for $i \in \{0, \dots, M-1\}$ in the example of a logistic regression. The mean of $\hat{\pi}_i(g_i)$ is displayed in black and is spaced apart from the other two curves by the standard deviation of $\hat{\pi}_i(g_i)$.

where $(X_k)_{k\in\mathbb{N}}$ are the samples of a Markov chain targeting π. Reducing the variance of Monte Carlo estimators such as $\hat{\pi}_n(f)$ is a very active research domain: see e.g. [RC04, Chapter 4], [Liu08, Section 2.3], and [RK17, Chapter 5] for an overview of the main methods.

In Chapter 6, we propose a method based on control variates, i.e. π-integrable functions $h = (h_1, \dots, h_p) : \mathbb{R}^d \to \mathbb{R}^p$ satisfying $\pi(h_i) = 0$ for $i \in \{1, \dots, p\}$. Note that for all $\theta \in \mathbb{R}^p$, $\pi(f) = \pi(f + \theta^{\mathrm{T}} h)$, and we try to find $\vartheta \in \mathbb{R}^p$ such that the variance of $\hat{\pi}_n(f + \vartheta^{\mathrm{T}} h)$ is smaller than the variance of $\hat{\pi}_n(f)$.

Unfortunately, explicit, non-asymptotic expressions of the variance of $\hat{\pi}_n(f)$ are often intricate. Our methodology relies on the asymptotic variance of $\hat{\pi}_n(f)$ when $n \to +\infty$. Indeed, under weak conditions [MT09, Chapter 17], the estimator $\hat{\pi}_n(f)$ satisfies for any initial distribution a Central Limit Theorem (CLT)

$$n^{-1/2} \sum_{k=0}^{n-1} (f(X_k) - \pi(f)) \overset{\text{weakly}}{\underset{n\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_{\infty,\mathrm{d}}(f)) \;, \quad \sigma^2_{\infty,\mathrm{d}}(f) = \pi\left((\hat{f}_{\mathrm{d}})^2 - (R\hat{f}_{\mathrm{d}})^2\right) \;, \tag{1.19}$$

where $\mathcal{N}(m, \sigma^2)$ denotes a Gaussian distribution with mean m and variance σ^2, and $\hat{f}_{\mathrm{d}}$ is a solution of the Poisson equation

$$(R - \mathrm{Id})\hat{f}_{\mathrm{d}} = -\{f - \pi(f)\} \;. \tag{1.20}$$

We seek to minimize the asymptotic variance and choose $\theta \in \mathbb{R}^p$ such that $\sigma^2_{\infty,\mathrm{d}}(f + \theta^{\mathrm{T}} h) \leq \sigma^2_{\infty,\mathrm{d}}(f)$.

[Hen97] and [Mey08, Section 11.5] proposed control variates of the form $(R-\mathrm{Id})\theta^{\mathrm{T}}\psi$ where $\psi = (\psi_1, \ldots, \psi_p)$ are known π-integrable functions. The parameter $\theta \in \mathbb{R}^p$ is obtained by minimizing the asymptotic variance

$$\min_{\theta\in\mathbb{R}^p} \sigma^2_{\infty,\mathrm{d}}(f + (R-\mathrm{Id})\theta^{\mathrm{T}}\psi) = \min_{\theta\in\mathbb{R}^p} \pi\left(\left\{\hat{f}_{\mathrm{d}} - \theta^{\mathrm{T}}\psi\right\}^2 - \left\{R(\hat{f}_{\mathrm{d}} - \theta^{\mathrm{T}}\psi)\right\}^2\right) , \qquad (1.21)$$

noting that $(-\theta^{\mathrm{T}}\psi)$ is a solution of the Poisson equation associated to $(R-\mathrm{Id})\theta^{\mathrm{T}}\psi$ and $\hat{f}_{\mathrm{d}}$ is defined in (1.20). The method suggested in [Mey08, Section 11.5] to minimize (1.21) requires estimates of the solution $\hat{f}_{\mathrm{d}}$ of the Poisson equation. Temporal Difference learning is a possible candidate, but this method is complex and suffers from high variance.

[DK12] noticed that if R is reversible w.r.t. π, it is possible to optimize the limiting variance (1.21) without computing explicitly the Poisson solution $\hat{f}_{\mathrm{d}}$. Reversibility plays also an important role in our methodology.

Each of the algorithms in the aforementioned literature requires computation of $R\psi_i$ for each $i \in \{1, \ldots, p\}$, which is in general a computational challenge. In [Hen97; Mey08] this is addressed by restricting to kernels for which $R(x, \cdot)$ has finite support for each x, and in [DK12] the authors restrict mainly to Gibbs samplers in their numerical examples.

In Chapter 6, an alternative class of control variates is used to avoid this computational barrier. This approach follows [AC99] (applications to quantum Monte Carlo calculations) and [MSI13; PMG14] (Bayesian statistics): assume that U is continuously differentiable, and for any twice continuously differentiable function φ, define $\mathcal{L}\varphi$ by

$$\mathcal{L}\varphi = -\langle \nabla U, \nabla\varphi\rangle + \Delta\varphi \, . \qquad (1.22)$$

Note that $\mathcal{L}$ is the generator of the Langevin diffusion given in (1.11). Under mild conditions on φ, it may be shown that $\pi(\mathcal{L}\varphi) = 0$. [MSI13] suggested to use $\mathcal{L}(\theta^{\mathrm{T}}\psi)$ with $\psi = (\psi_1, \ldots, \psi_p)$ as control variates and choose θ by minimizing $\theta \mapsto \pi(\{f - \pi(f) + \mathcal{L}\theta^{\mathrm{T}}\psi\}^2)$. This approach has triggered numerous work, among others [OGC16], [OG16] and [Oat+18] which introduce control functionals; a nonparametric extension of control variates. A drawback of this method stems from the fact that the optimization criterion $\pi(\{f-\pi(f)+\mathcal{L}\theta^{\mathrm{T}}\psi\}^2)$ is only theoretically justified if $(X_k)_{k\in\mathbb{N}}$ is i.i.d. and might significantly differ from the asymptotic variance $\sigma^2_{\infty,\mathrm{d}}(f + \mathcal{L}(\theta^{\mathrm{T}}\psi))$ defined in (6.1).

In Chapter 6, we propose a new method to construct control variates. Analysis and motivation are based on the overdamped Langevin diffusion defined in (1.9). Under smoothness and 'tail' conditions on f and ∇U, the following CLT holds for any initial condition (see [Bha82; CCG12])

$$t^{-1/2}\int_0^t \{f(Y_s) - \pi(f)\}\mathrm{d}s \underset{t\to+\infty}{\overset{\text{weakly}}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_\infty(f)) \, , \quad \sigma^2_\infty(f) = 2\pi\left(\hat{f}\{f - \pi(f)\}\right) \, , \qquad (1.23)$$

where $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ is a solution of the (continuous-time) Poisson equation

$$\mathcal{L}\hat{f} = -\{f - \pi(f)\} \, . \qquad (1.24)$$

We introduce a new class of control variates based on the expression of the asymptotic variance $\sigma^2_\infty(f)$ given in (1.23). Since $\pi(\mathcal{L}(\theta^{\mathrm{T}}\psi)) = 0$ for any $\theta \in \mathbb{R}^d$, we consider the control variate $\mathcal{L}(\theta^*(f)^{\mathrm{T}}\psi)$ where $\theta^*(f)$ is chosen by minimizing

$$\theta \mapsto \sigma^2_\infty(f + \mathcal{L}(\theta^{\mathrm{T}}\psi)) \; . \tag{1.25}$$

Although $\mathcal{L}(\theta^*(f)^{\mathrm{T}}\psi)$ is a control variate for the Langevin diffusion associated with f, the choice of this optimization criterion is motivated by the fact that for some MCMC algorithms, the asymptotic variance $\sigma^2_{\infty,\mathrm{d}}(f)$ defined in (1.19) is (up to a scaling factor) a good approximation of the asymptotic variance of the Langevin diffusion $\sigma^2_\infty(f)$ defined in (1.23). Moreover, the minimization of (1.25) admits a unique solution $\theta^*(f)$, which is in general easy to estimate. It is worthwhile to note that it is not required to know the Poisson solution $\hat{f}$ to minimize (1.25).

1.6 Stochastic Gradient Langevin Dynamics

The ULA algorithm defined in (1.10) requires to compute at each step the gradient of the potential U. However, in Bayesian machine learning, U is proportional to minus the logarithm of the posterior distribution and is the sum of a large number of items. More precisely, let denote by $\mathbf{z} = \{z_i\}_{i=1}^N$ the observations and consider a situation where the target distribution π arises as the posterior in a Bayesian inference problem with prior density $\pi_0(\theta)$ and a large number $N \gg 1$ of i.i.d. observations z_i with likelihoods $p(z_i|\theta)$. In this case, $\pi(\theta) = \pi_0(\theta)\prod_{i=1}^N p(z_i|\theta)$. We denote $U_i(\theta) = -\log(p(z_i|\theta))$ for $i \in \{1,\dots,N\}$, $U_0(\theta) = -\log(\pi_0(\theta))$, $U = \sum_{i=0}^N U_i$. Evaluating the gradient of U, $\nabla U(\theta) = \sum_{i=0}^N \nabla U_i(\theta)$ in $\theta \in \mathbb{R}^d$ is computationally intensive. The cost of one iteration of ULA is Nd which is prohibitively large for massive datasets ($N \gg 1$).

In order to scale up to the big data setting, Welling and Teh [WT11] suggested to replace ∇U with an unbiased estimate $\nabla U_0 + (N/p)\sum_{i\in S}\nabla U_i$ where S is a minibatch of $\{1,\dots,N\}$ with replacement of size p. A single update of the resulting algorithm, Stochastic Gradient Langevin Dynamics (SGLD), is then given for $k \in \mathbb{N}$ by

$$\theta_{k+1} = \theta_k - \gamma\left(\nabla U_0(\theta_k) + \frac{N}{p}\sum_{i\in S_{k+1}}\nabla U_i(\theta_k)\right) + \sqrt{2\gamma}W_{k+1} \; .$$

The idea of using only a fraction of data points to compute an unbiased estimate of the gradient at each iteration comes from Stochastic Gradient Descent (SGD) which is a popular algorithm to minimize the potential U. SGD is very similar to SGLD because it is characterised by the same recursion as SGLD but without Gaussian noise:

$$\theta_{k+1} = \theta_k - \gamma\left(\nabla U_0(\theta_k) + \frac{N}{p}\sum_{i\in S_{k+1}}\nabla U_i(\theta_k)\right) \; .$$

Assuming for simplicity that U has a minimizer $\theta^\star$, we can define a control variates version of SGLD, SGLDFP, see [Dub+16; Che+17], given for $k \in \mathbb{N}$ by

$$\theta_{k+1} = \theta_k - \gamma \left(\nabla U_0(\theta_k) - \nabla U_0(\theta^\star) + \frac{N}{p} \sum_{i \in S_{k+1}} \{\nabla U_i(\theta_k) - \nabla U_i(\theta^\star)\} \right) + \sqrt{2\gamma} W_{k+1} \, .$$

In Chapter 7, we provide insights on the links between SGLD, SGLDFP, ULA and SGD. In our analysis, the algorithms are used with a constant step size and the parameters are set to the standard values used in practice: in particular, $\gamma \approx 1/N$. The ULA, SGD, SGLD and SGLDFP algorithms define homogeneous Markov chains, each of which admits a unique stationary distribution used as a hopefully close proxy of π. Our main contribution is to show that, while the invariant distributions of ULA and SGLDFP become closer to π as the number of data points N increases, on the opposite, the invariant measure of SGLD never comes close to the target distribution π and is in fact very similar to the invariant measure of SGD.

We show that the number of iterations necessary to obtain a sample ε-close from π in Wasserstein distance is the same for ULA and SGLDFP. However for ULA, the cost of one iteration is Nd which is much larger than pd the cost of one iteration for SGLDFP. In other words, to obtain an approximate sample from the target distribution at an accuracy $O(1/\sqrt{N})$ in 2-Wasserstein distance, LMC requires about N operations, in contrast with SGLDFP that needs only a number of operations independent of N.

Summary of our contributions

This thesis is based on the following published articles:

- Brosse, N., Durmus, A., Moulines, É., & Pereyra, M. (2017). *Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo.* Colt Conference. Proceedings of Machine Learning Research, PMLR 65:319-342
- Brosse, N., Durmus, A., Moulines, É. *Normalizing constants of log-concave densities.* Electron. J. Statist. 12 (2018), no. 1, 851–889. doi:10.1214/18-EJS1411.
- Brosse, N., Durmus, A., Moulines, É., & Sabanis, S. *The Tamed Unadjusted Langevin Algorithm.* Stochastic Processes and their Applications (2018), ISSN 0304-4149.
- Brosse, N., Durmus, A., Moulines, É. *The promises and pitfalls of Stochastic Gradient Langevin Dynamics*, NeurIps 2018.

and preprint:

- Brosse, N., Durmus, A., Meyn, S., & Moulines, E. (2018). *Diffusion approximations and control variates for MCMC.* arXiv preprint arXiv:1808.01665.

Chapter 2

Résumé de la thèse

Dans ce Chapitre, nous présentons nos contributions qui sont divisées en 3 grandes parties:

1. les extensions d'ULA en Section 1.4,
2. les applications d'ULA en Section 1.5,
3. une analyse de Stochastic Gradient Langevin Dynamics (SGLD) en Section 1.6.

2.1 Extensions de l'algorithme de Langevin non-ajusté

Dans la partie I de cette thèse, deux limitations de l'algorithme ULA défini en (1.10) sont traitées. Premièrement, ULA est bien défini et réalisable si le potentiel U est continuellement différentiable sur $\mathbb{R}^d$: il ne peut pas être appliqué directement à une distribution π limitée à un ensemble compact convexe. Cependant, de nombreux problèmes d'inférence statistique impliquent l'estimation de paramètres soumis à des contraintes sur l'espace des paramètres. L'algorithme MYULA proposé dans [DMP18] permet de tirer des échantillons approximatifs à partir de distributions avec un support compact en régularisant correctement le potentiel U. Dans le Chapitre 3, nous calculons des limites précises de convergence en variation totale et en distance de Wasserstein pour cet algorithme.

Deuxièmement, quand le potentiel U augmente trop vite à l'infini, i.e. $\|U(x)\| \gtrsim \|x\|^{2+\alpha}$ lorsque $\|x\| \to +\infty$ et avec $\alpha > 0$, ULA est instable et peut diverger avec probabilité non nulle. Inspirés par des travaux récents sur la discrétisation des SDEs avec des coefficients de dérive superlinéaire [HJK12; Sab13], nous proposons un nouvel algorithme dans le Chapitre 4, le "tamed" ULA, et fournissons des garanties de convergence en V-variation totale et en distance de Wasserstein d'ordre 2.

2.1.1 Echantillonnage d'une distribution avec un support compact : MYULA

Considérons dans un cadre bayésien une distribution postérieure π avec un support borné. Voici quelques exemples : les problèmes de données tronquées qui surviennent

naturellement dans les études de temps de survie et d'extinction, [KM05], modèles de données ordinales [JA06], régressions Lasso et régressions ridge [Cel+12], Latent Dirichlet Allocation [BNJ03], et la factorisation matricielle positive [PBJ14]. Tirer des échantillons d'une distribution ainsi contrainte est un problème difficile à résoudre qui a fait l'objet de nombreux articles ; voir [GSL92], [PP14], [LS15], [BEL15], [Hsi+18]. Tous ces travaux sont basés sur l'efficacité de la chaîne de Markov pour estimer la distribution postérieure ; cependant, à l'exception de [BEL15] et [Hsi+18], ces méthodes ne sont pas bien comprises en théorie et ne fournissent aucune garantie théorique sur les estimations fournies.

Une modification d'ULA a été proposée dans [DMP18] pour échantillonner à partir d'une distribution de probabilité log-concave non lisse sur $\mathbb{R}^d$. Cette méthode appelée MYULA est principalement basée sur une version régularisée de la distribution cible π qui jouit d'un certain nombre de propriétés favorables qui sont utiles pour la simulation MCMC. Dans cette étude, nous analysons la complexité de cet algorithme lorsqu'il est appliqué à des distributions log-concave limitées à un ensemble convexe, l'accent étant mis sur la complexité à mesure que la dimension de l'espace d'état augmente. Plus précisément, nous établissons des limites explicites en variation totale et en distance de Wasserstein d'ordre 1 entre les itérations de la chaîne de Markov et la densité cible π.

Il est à noter que l'algorithme de Langevin ajusté par Metropolis (MALA) est une solution alternative à ULA pour échantillonner à partir d'une distribution à support compact. Toutefois, aucune limite de convergence précise n'était disponible pour MALA. Suite à ces travaux, [Dwi+18] a fourni des limites de convergence précises pour MALA et une nouvelle analyse du problème de l'échantillonnage d'une distribution à support compact dans cette perspective serait un travail de recherche intéressant.

Soit $\mathsf{K} \subset \mathbb{R}^d$ un ensemble compact convexe tel que $\mathrm{B}(0,r) \subset \mathsf{K} \subset \mathrm{B}(0,R)$ et $\iota_{\mathsf{K}} : \mathbb{R}^d \to \{0, +\infty\}$ la fonction indicatrice (convexe) de K, définie pour tout $x \in \mathbb{R}^d$ par,

$$\iota_{\mathsf{K}}(x) = \begin{cases} +\infty & \text{si } x \notin \mathsf{K}, \\ 0 & \text{si } x \in \mathsf{K} \ . \end{cases}$$

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction convexe, gradient Lipschitz, continuellement différentiable. Considérons une densité de probabilité π associée à un potentiel $U : \mathbb{R}^d \to (-\infty, +\infty]$ de la forme $U = f + \iota_{\mathsf{K}}$. Pour appliquer ULA, [DMP18] a suggéré de régulariser U de telle sorte que

1. la convexité de U est préservée (cette propriété est la clé de l'analyse théorique de l'algorithme),

2. la régularisation de U est continuellement différentiable et gradient Lipschitz (cette propriété de régularité est la clé de la stabilité de l'algorithme),

3. l'approximation résultante est proche de π (e.g. en variation totale).

L'outil utilisé pour construire une telle approximation est l'enveloppe Moreau-Yosida de ι_{K}, $\iota_{\mathsf{K}}^{\lambda} : \mathbb{R}^d \to \mathbb{R}_+$ définie pour $x \in \mathbb{R}^d$ par,

$$\iota_{\mathsf{K}}^{\lambda}(x) = \inf_{y \in \mathbb{R}^d} \left(\iota_{\mathsf{K}}(y) + (2\lambda)^{-1} \left\| x - y \right\|^2 \right) = (2\lambda)^{-1} \left\| x - \mathrm{proj}_{\mathsf{K}}(x) \right\|^2 \ ,$$

où $\lambda > 0$ est un paramètre de régularisation et $\mathrm{proj}_{\mathsf{K}}$ est la projection sur K. $\iota_{\mathsf{K}}^{\lambda}$ est convexe et continuellement différentiable avec un gradient donné pour tout $x \in \mathbb{R}^d$ par :

$$\nabla \iota_{\mathsf{K}}^{\lambda}(x) = \lambda^{-1}(x - \mathrm{proj}_{\mathsf{K}}(x)) \; .$$

$\nabla \iota_{\mathsf{K}}^{\lambda}$ est λ^{-1}-Lipschitz. Ajouter f à $\iota_{\mathsf{K}}^{\lambda}$ conduit à la régularisation $U^{\lambda} : \mathbb{R}^d \to \mathbb{R}$ du potentiel U défini pour tout $x \in \mathbb{R}^d$ par $U^{\lambda}(x) = f(x) + \iota_{\mathsf{K}}^{\lambda}(x)$. La probabilité associée π^{λ} sur $\mathbb{R}^d$ est donnée pour tout $x \in \mathbb{R}^d$ par

$$\pi^{\lambda}(x) = \mathrm{e}^{-U^{\lambda}(x)} \Big/ \int_{\mathbb{R}^d} \mathrm{e}^{-U^{\lambda}(y)} \mathrm{d}y \;\; ,$$

est bien définie et log-concave. De plus, U^{λ} est gradient Lipschitz et continuellement différentiable avec ∇U^{λ} donné pour tout $x \in \mathbb{R}^d$ par

$$\nabla U^{\lambda}(x) = -\nabla \log \pi^{\lambda}(x) = \nabla f(x) + \lambda^{-1}(x - \mathrm{proj}_{\mathsf{K}}(x)) \; .$$

L'algorithme proposé dans [DMP18] utilise ensuite la discrétisation d'Euler-Maruyama de l'équation de Langevin associée à U^{λ}, avec π^{λ} comme proxy, pour générer des échantillons approximatifs à partir de π. Il utilise la chaîne de Markov $(X_k)_{k\in\mathbb{N}}$, débutée à X_0, donnée pour tout $k \in \mathbb{N}$ par

$$X_{k+1} = (1 - \frac{\gamma}{\lambda})X_k - \gamma \nabla f(X_k) + \frac{\gamma}{\lambda} \mathrm{proj}_{\mathsf{K}}(X_k) + \sqrt{2\gamma} W_{k+1} \; , \tag{2.1}$$

où $(W_k)_{k\in\mathbb{N}}$ est une séquence de vecteurs d-dimensionnels gaussiens standards et $\gamma > 0$ est le pas. Notez que cet algorithme suppose que la projection sur K est relativement facile à calculer.

Le noyau de la chaîne de Markov homogène définie par (2.1) est donné pour $x \in \mathbb{R}^d$ et $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ par,

$$R_{\gamma,\lambda}(x, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(-(4\gamma)^{-1} \left\| y - x + \gamma \nabla U^{\lambda}(x) \right\|^2\right) \mathrm{d}y \; .$$

Puisque la densité cible de la chaîne de Markov (1.14) est la mesure régularisée π^{λ} et non π, l'algorithme est appelé MYULA (Moreau-Yosida Unadjusted Langevin Algorithm).

Le résultat principal prouvé dans le Chapitre 3 est que pour tout $\varepsilon > 0$ et $x \in \mathbb{R}^d$, il existe $\lambda, \gamma_0 > 0$ tel que pour tout $\gamma \in (0, \gamma_0]$,

$$\|\delta_x R_{\gamma,\lambda}^n - \pi\|_{\mathrm{TV}} \leq \varepsilon \quad \text{pour} \quad n \gtrsim d^5 \; .$$

Un point intéressant de cette étude est la dépendance du paramètre de régularisation λ par rapport à la dimension d. Nous montrons que λ devrait être de l'ordre d^{-2} pour avoir une borne non triviale sur $\|\pi^{\lambda} - \pi\|_{\mathrm{TV}}$ quand $d \to +\infty$. Cette dépendance peut sembler importante; La géométrie et les probabilités en haute dimension présentent souvent des comportements contre-intuitifs. Par exemple, la plus grande partie du volume d'une boule de grande dimension de rayon 1 est contenue dans un anneau étroit près

de sa surface d'épaisseur de l'ordre de $1/d$. De même, le volume d'un cube de grande dimension est principalement contenu dans ses coins.

Suite à ce travail, [Hsi+18] a suggéré un algorithme de dynamique de Langevin en miroir pour échantillonner des domaines contraints dans $\mathbb{R}^d$. Lorsque π est fortement log-concave, les auteurs montrent qu'il est possible de tirer des échantillons approximatifs de π en itérant au maximum environ d fois une chaîne de Markov appropriée définie sur l'espace entier $\mathbb{R}^d$. Cependant, ce résultat général n'est qu'existentiel et des algorithmes pratiques doivent être développés au cas par cas. Les auteurs donnent une application explicite sur le simplex. En outre, contrairement à ce qui est rapporté dans [Hsi+18, Tableau 1], notre résultat est valable pour les distributions non fortement log-concaves π et est donc plus général que leur cadre.

2.1.2 Echantillonnage des potentiels superquadratiques U : TULA

L'algorithme ULA est instable si ∇U est super linéaire i.e. $\liminf_{\|x\|\to+\infty} \|\nabla U(x)\| / \|x\| = +\infty$, voir [RT96, Theorem 3.2], [MSH02] et [HJK11]. Ceci est illustré par un exemple particulier dans [MSH02, Lemma 6.3] où la SDE (1.9) est considérée en une dimension avec $U(x) = x^4/4$ avec la discrétisation Euler associée (1.10). Il est montré que pour tout $\gamma > 0$, si $\mathbb{E}\left[X_0^2\right] \geq 2/\gamma$, alors $\lim_{n\to+\infty} \mathbb{E}\left[X_n^2\right] = +\infty$. De plus, $(X_n)_{n\in\mathbb{N}}$ diverge à l'infini avec une probabilité positive.

Jusqu'à récemment, soit des schémas numériques implicites, par exemple voir [MSH02] et [HMS02], soit des schémas pas à pas adaptatifs, par exemple voir [LMS07], étaient utilisés pour aborder ce problème. Des méthodes dites S-ROCK ont également été mises au point pour résoudre ce problème, voir [AL08; AC08]. Au cours des dernières années, une nouvelle génération de schémas numériques explicites, qui sont efficaces sur le plan informatique, a été introduite en "amortissant" de manière appropriée la dérive de croissance superlinéaire, voir [HJK12] et [Sab13] pour plus de détails.

Néanmoins, à l'exception de [MSH02], ces travaux se concentrent sur la discrétisation des SDEs à coefficients superlinéaires en temps fini. Notre objectif est d'étendre ces techniques à l'échantillonnage à partir de π. Pour traiter de la nature super-linéaire de ∇U, nous nous basons sur des études antérieures sur le schéma d'Euler "apprivoisé" [HJK12], [Sab13], [HJ15], et présentons le Tamed ULA (TULA) défini pour $k \in \mathbb{N}$ par

$$X_{k+1} = X_k - \gamma \frac{\nabla U(X_k)}{1 + \gamma \|\nabla U(X_k)\|} + \sqrt{2\gamma} W_{k+1} \ , \ X_0 = x_0 \ . \tag{2.2}$$

Nous désignons par R_γ le noyau de Markov associé. Notez que dans le Chapitre 4, nous traitons d'un cadre plus général, permettant de réduire la dérive super-linéaire de différentes manières.

Le potentiel U est supposé être localement gradient Lipschitz, avec une constante de Lipschitz au plus polynomiale, i.e. pour tout $x, y \in \mathbb{R}^d$,

$$\|\nabla U(x) - \nabla U(y)\| \leq L \left\{1 + \|x\|^\ell + \|y\|^\ell\right\} \|x - y\| \ ,$$

où $\ell, L \geq 0$. Sous une hypothèse faible supplémentaire, nous montrons dans le Chapitre 4 qu'il existe une fonction $V : \mathbb{R}^d \to [1, +\infty)$, $\lambda \in (0, 1)$ telle que pour tout γ assez petit,

$x \in \mathbb{R}^d$ et $n \in \mathbb{N}$,

$$\left\| \delta_x R_\gamma^n - \pi \right\|_V \lesssim n\gamma\lambda^{n\gamma} V(x) + \sqrt{\gamma} \, .$$

En plus de cette ergodicité géométrique uniforme de V du noyau de Markov TULA, une borne supérieure sur la distance en 2-Wasserstein est fournie :

$$W_2^2(\delta_x R_\gamma^n, \pi) \lesssim n\lambda^{n\gamma} V(x) + \gamma \, . \tag{2.3}$$

Si la hessienne de U, $\nabla^2 U$, est localement β-Hölder, avec $\beta \in [0,1]$, (2.3) peut être amélioré:

$$W_2^2(\delta_x R_\gamma^n, \pi) \lesssim n\gamma^{1+\beta}\lambda^{n\gamma} V(x) + \gamma^{1+\beta} \, .$$

En supposant que le potentiel U est suffisamment lisse, nous récupérons les bornes standards par rapport à $n \in \mathbb{N}^*$ et $\gamma > 0$ sur l'erreur faible, voir [MST10, Theorems 5.1, 5.2],

$$\left| \mathbb{E}\left[\frac{\frac{1}{n}\sum_{k=0}; n-1}{f}(X_k) - \pi(f) \right] \right| \lesssim \gamma + \frac{1}{n\gamma} \, ,$$

et sur l'erreur quadratique moyenne

$$\mathbb{E}\left[\left(\frac{\frac{1}{n}\sum_{k=0}; n-1}{f}(X_k) - \pi(f) \right)^2 \right] \lesssim \gamma^2 + \frac{1}{n\gamma} \, .$$

La technique de preuve est une adaptation simple de la méthode présentée dans la Section 1.3 pour montrer (1.12). En conclusion, TULA est plus robuste que ULA, avec des garanties théoriques similaires et devrait être préféré dans la pratique, surtout lorsque le potentiel U est susceptible de croître rapidement à l'infini.

2.2 Applications de l'algorithme de Langevin non-ajusté

Dans la partie II, nous présentons une application directe de l'algorithme ULA pour calculer des constantes de normalisation avec des garanties théoriques précises. Dans une deuxième étape, nous donnons une nouvelle méthodologie de variables de contrôle pour les algorithmes ULA, RWM et MALA, qui s'inspire directement de la comparaison de la dynamique ULA à la diffusion de Langevin.

2.2.1 Estimation de la constante de normalisation de densités log-concaves

Soit $U : \mathbb{R}^d \to \mathbb{R}$ une fonction convexe continuellement différentiable telle que $Z = \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \mathrm{d}x < +\infty$. Z est la constante de normalisation (inconnue) de la densité de probabilité π associée au potentiel U, définie pour tout $x \in \mathbb{R}^d$ par $\pi(x) = Z^{-1}\mathrm{e}^{-U(x)}$. Dans le Chapitre 5, nous présentons et discutons une méthode pour estimer Z avec une complexité polynomiale dans la dimension d.

Le calcul de la constante de normalisation est un défi qui a des applications en inférence bayésienne et en physique statistique en particulier. En physique statistique, Z

est mieux connu sous le nom de fonction de partition ou énergie libre [Bal07], [LSR10]. Les différences d'énergie libre permettent de quantifier la probabilité relative des différents états (configurations microscopiques) et sont liées au travail thermodynamique et aux échanges thermiques. Dans l'inférence bayésienne, les modèles peuvent être comparés par le calcul du facteur de Bayes qui est le rapport de deux constantes de normalisation (voir e.g. [Rob07, chapitre 7]). Ce problème a par conséquent attiré de nombreuses contributions ; voir par exemple [CSI00, chapitre 5], [MR09], [FW12], [Ard+12], [D+13], [Knu+15], [ZJA15] et, pour la simulation moléculaire, [LSR10].

Notre algorithme repose sur une séquence de densités gaussiennes avec des variances croissantes, combinées aux bornes précises de [DM16]. Supposons sans perte de généralité que U a un minimum en $x^\star = 0$ et $U(x^\star) = 0$. Soit $M \in \mathbb{N}^\star$, $\{\sigma_i^2\}_{i=0}^M$ une séquence croissante positive de nombres réels et définissons $\sigma_M^2 = +\infty$. Considérons la séquence de fonctions $\{U_i\}_{i=0}^M$ définie pour tous les $i \in \{0, \dots, M\}$ et $x \in \mathbb{R}^d$ par

$$U_i(x) = \frac{\|x\|^2}{2\sigma_i^2} + U(x) \;,$$

avec la convention $1/\infty = 0$. Nous définissons une séquence de densités de probabilité $\{\pi_i\}_{i=0}^M$ pour $i \in \{0, \dots, M\}$ et $x \in \mathbb{R}^d$ par

$$\pi_i(x) = Z_i^{-1} \mathrm{e}^{-U_i(x)} \;, \qquad Z_i = \int_{\mathbb{R}^d} \mathrm{e}^{-U_i(y)} \mathrm{d}y \;.$$

Par définition, notons que $U_M = U$, $Z_M = Z$ et $\pi_M = \pi$. Comme dans la méthode d'échantillonnage à plusieurs échelles [GM98, Section 3.3], nous utilisons la décomposition suivante

$$\frac{Z}{Z_0} = \prod_{i=0}^{M-1} \frac{Z_{i+1}}{Z_i} \;.$$

Z_0 est estimée en choisissant σ_0^2 assez petit pour que π_0 soit suffisamment proche d'une distribution gaussienne de moyenne 0 et de covariance $\sigma_0^2 \mathrm{Id}$. Pour $i \in \{0, \dots, M-1\}$, le ratio Z_{i+1}/Z_i peut être exprimé comme suit

$$\frac{Z_{i+1}}{Z_i} = \int_{\mathbb{R}^d} g_i(x) \pi_i(x) \mathrm{d}x = \pi_i(g_i) \;,$$

où $g_i : \mathbb{R}^d \to \mathbb{R}_+$ est définie pour tout $x \in \mathbb{R}^d$ par

$$g_i(x) = \exp\left(a_i \|x\|^2\right) \;, \qquad a_i = \frac{1}{2}\left(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_{i+1}^2}\right) \;.$$

La quantité $\pi_i(g_i)$ est estimée par ULA ciblant π_i. Plus précisément, nous définissons M chaînes de Markov basées sur ULA pour $i \in \{0, \dots, M-1\}$ et $k \in \mathbb{N}$ par

$$X_{i,k+1} = X_{i,k} - \gamma_i \nabla U_i(X_{i,k}) + \sqrt{2\gamma_i} W_{i,k+1} \;, \quad X_{i,0} = 0 \;,$$

où $\{(W_{i,k})_{k\in\mathbb{N}^\star}\}_{i=0}^{M-1}$ sont des séquences indépendantes de variables aléatoires gaussiennes et $\gamma_i > 0$ est le pas. Pour $i \in \{0,\dots,M-1\}$, considérons l'estimateur suivant de Z_{i+1}/Z_i,

$$\hat{\pi}_i(g_i) = \frac{1}{n_i}\sum_{k=N_i+1}^{N_i+n_i} g_i(X_{i,k}) \;,$$

où $n_i \geq 1$ est la taille de l'échantillon et $N_i \geq 0$ la "burn-in" période. Pour simplifier la présentation, nous supposons que U est m-fortement convexe. Dans le Chapitre 5, nous donnons un choix explicite des paramètres de simulation

$$\mathcal{S} = \left\{M, \{\sigma_i^2\}_{i=0}^{M-1}, \{\gamma_i\}_{i=0}^{M-1}, \{n_i\}_{i=0}^{M-1}, \{N_i\}_{i=0}^{M-1}\right\} \;,$$

de sorte que $\hat{Z}$ l'estimateur suivant de Z,

$$\hat{Z} = (2\pi\sigma_0^2)^{d/2}(1+\sigma_0^2 m)^{-d/2}\left\{\prod_{i=0}^{M-1}\hat{\pi}_i(g_i)\right\} \;, \tag{2.4}$$

satisfait

$$\mathbb{P}\left(\left|\hat{Z}/Z - 1\right| > \epsilon\right) \leq \mu \;, \quad \text{pour } \mu,\epsilon \in (0,1) \;.$$

Le coût de l'algorithme défini par $\mathrm{cost} = \sum_{i=0}^{M-1}\{N_i + n_i\}$ est borné par $d^{5/2}$ à des facteurs logarithmiques près lorsque U est fortement convexe, et ∇U, $\nabla^2 U$ sont Lipschitz. Notez que dans l'expression $\hat{Z}$ (2.4), Z_0 est approximé par $(2\pi\sigma_0^2)^{d/2}(1+\sigma_0^2 m)^{-d/2}$. Une décomposition plus fine est possible pour calculer Z_0 :

$$Z_0 = (2\pi\sigma_0^2)^{d/2}\int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\frac{\mathrm{e}^{-\|x\|^2/(2\sigma_0^2)}}{(2\pi\sigma_0^2)^{d/2}}\mathrm{d}x \;.$$

L'intégrale peut être estimée par une méthode classique de Monte Carlo en utilisant des échantillons i.i.d. de la distribution gaussienne avec une matrice de covariance $\sigma_0^2\,\mathrm{Id}$. Un point important à souligner est le fait que notre algorithme fournit un choix théoriquement fondé de la séquence de recuit des variances $\{\sigma_i^2\}_{i=0}^{M-1}$.

2.2.2 Limites diffusives et variables de contrôle pour MCMC

Dans la Section 1.1, nous avons souligné le fait que l'un des principaux objectifs des méthodes MCMC est d'estimer $\pi(f)$ pour une fonction spécifique, π-intégrable $f : \mathbb{R}^d \to \mathbb{R}$, voir (1.4). Un estimateur de Monte Carlo naturel de $\pi(f)$ est $\hat{\pi}_n(f)$ défini pour $n \in \mathbb{N}^*$ par

$$\hat{\pi}_n(f) = \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \;, \tag{2.5}$$

où $(X_k)_{k\in\mathbb{N}}$ sont les échantillons d'une chaîne de Markov ciblant π. La réduction de la variance des estimateurs de Monte Carlo tels que $\hat{\pi}_n(f)$ est un domaine de recherche très actif : voir e.g. [RC04, Chapitre 4], [Liu08, Section 2.3] et [RK17, Chapitre 5] pour une présentation des principales méthodes.

Dans le Chapitre 6, nous proposons une méthode basée sur les variables de contrôle, i.e. π-fonctions intégrables $h = (h_1, \dots, h_p) : \mathbb{R}^d \to \mathbb{R}^p$ satisfaisant $\pi(h_i) = 0$ pour $i \in \{1, \dots, p\}$. Notez que pour tout $\theta \in \mathbb{R}^p$, $\pi(f) = \pi(f + \theta^{\mathrm{T}} h)$, et nous essayons de trouver $\vartheta \in \mathbb{R}^p$ tel que la variance de $\hat{\pi}_n(f + \vartheta^{\mathrm{T}} h)$ soit inférieure à celle de $\hat{\pi}_n(f)$.

Malheureusement, les expressions explicites, non-asymptotiques de la variance de $\hat{\pi}_n(f)$ sont souvent complexes. Notre méthodologie repose sur la variance asymptotique de $\hat{\pi}_n(f)$ quand $n \to +\infty$. En effet, sous des hypothèses faibles [MT09, Chapitre 17], l'estimateur $\hat{\pi}_n(f)$ satisfait pour toute distribution initiale un théorème central limite (TCL)

$$n^{-1/2} \sum_{k=0}^{n-1} (f(X_k) - \pi(f)) \overset{\text{weakly}}{\underset{n \to +\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_{\infty,\mathrm{d}}(f)) \;, \quad \sigma^2_{\infty,\mathrm{d}}(f) = \pi\left((\hat{f}_{\mathrm{d}})^2 - (R\hat{f}_{\mathrm{d}})^2 \right) \;, \tag{2.6}$$

où $\mathcal{N}(m, \sigma^2)$ désigne une distribution gaussienne avec moyenne m et variance σ^2, et $\hat{f}_{\mathrm{d}}$ est une solution de l'équation de Poisson

$$(R - \mathrm{Id})\hat{f}_{\mathrm{d}} = -\{f - \pi(f)\} \;. \tag{2.7}$$

Nous cherchons à minimiser la variance asymptotique et choisissons $\theta \in \mathbb{R}^p$ tel que $\sigma^2_{\infty,\mathrm{d}}(f + \theta^{\mathrm{T}} h) \leq \sigma^2_{\infty,\mathrm{d}}(f)$.

[Hen97] et [Mey08, Section 11.5] ont proposé des variables de contrôle de la forme $(R - \mathrm{Id})\theta^{\mathrm{T}}\psi$ où $\psi = (\psi_1, \dots, \psi_p)$ sont des fonctions fixées intégrables sous π. Le paramètre $\theta \in \mathbb{R}^p$ est obtenu en minimisant la variance asymptotique

$$\min_{\theta \in \mathbb{R}^p} \sigma^2_{\infty,\mathrm{d}}(f + (R - \mathrm{Id})\theta^{\mathrm{T}}\psi) = \min_{\theta \in \mathbb{R}^p} \pi\left(\left\{ \hat{f}_{\mathrm{d}} - \theta^{\mathrm{T}}\psi \right\}^2 - \left\{ R(\hat{f}_{\mathrm{d}} - \theta^{\mathrm{T}}\psi) \right\}^2 \right) \;, \tag{2.8}$$

notant que $(-\theta^{\mathrm{T}}\psi)$ est une solution de l'équation de Poisson associée à $(R - \mathrm{Id})\theta^{\mathrm{T}}\psi$ et $\hat{f}_{\mathrm{d}}$ est définie dans (2.7). La méthode suggérée dans [Mey08, Section 11.5] pour minimiser (2.8) nécessite des estimations de la solution $\hat{f}_{\mathrm{d}}$ de l'équation de Poisson. L'apprentissage par différence temporelle est un candidat possible, mais cette méthode est complexe et souffre d'une grande variance.

[DK12] a remarqué que si R est réversible w.r.t. π, il est possible d'optimiser la variance limite (2.8) sans calculer explicitement la solution de Poisson $\hat{f}_{\mathrm{d}}$. La réversibilité joue également un rôle important dans notre méthodologie.

Chacun des algorithmes susmentionnés nécessite le calcul de $R\psi_i$ pour chaque $i \in \{1, \dots, p\}$, ce qui est en général difficile. Dans [Hen97; Mey08] cela est résolu en se limitant aux noyaux pour lesquels $R(x, \cdot)$ a un support fini pour chaque x, et dans [DK12] les auteurs se limitent principalement aux échantillonneurs Gibbs dans leurs exemples numériques.

Dans le Chapitre 6, une classe alternative de variables de contrôle est utilisée pour éviter cette barrière informatique. Cette approche suit [AC99] (applications aux calculs quantiques de Monte Carlo) et [MSI13; PMG14] (statistiques bayésiennes): supposons que U est continuellement différentiable, et pour toute fonction deux fois continuellement différentiable φ, définissons $\mathcal{L}\varphi$ par

$$\mathcal{L}\varphi = -\langle \nabla U, \nabla \varphi \rangle + \Delta \varphi \;. \tag{2.9}$$

Notons que $\mathcal{L}$ est le générateur de la diffusion Langevin donnée dans (1.11). Avec des conditions faibles sur φ, on peut montrer que $\pi(\mathcal{L}\varphi) = 0$. [MSI13] suggère d'utiliser $\mathcal{L}(\theta^{\mathrm{T}}\psi)$ avec $\psi = (\psi_1, \ldots, \psi_p)$ comme variables de contrôle et de choisir θ en minimisant $\theta \mapsto \pi(\{f - \pi(f) + \mathcal{L}\theta^{\mathrm{T}}\psi\}^2)$. Cette approche a déclenché de nombreux travaux, entre autres [OGC16], [OG16] et [Oat+18] qui introduisent des fonctions de contrôle ; une extension non paramétrique des variables de contrôle. Un inconvénient de cette méthode vient du fait que le critère d'optimisation $\pi(\{f - \pi(f) + \mathcal{L}\theta^{\mathrm{T}}\psi\}^2)$ n'est théoriquement justifié que si $(X_k)_{k\in\mathbb{N}}$ est i.i.d. et peut différer significativement de la variance asymptotique $\sigma^2_{\infty,\mathrm{d}}(f + \mathcal{L}(\theta^{\mathrm{T}}\psi))$ définie dans (**??**).

Dans le Chapitre 6, nous proposons une nouvelle méthode pour construire des variables de contrôle. L'analyse est basée sur la diffusion de Langevin définie dans (1.9). Sous des conditions faibles sur f et ∇U, un TCL s'applique pour toute condition initiale (voir [Bha82; CCG12])

$$t^{-1/2}\int_0^t \{f(Y_s) - \pi(f)\}\mathrm{d}s \overset{\text{weakly}}{\underset{t\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_\infty(f)) \;, \quad \sigma^2_\infty(f) = 2\pi\left(\hat{f}\{f - \pi(f)\}\right) \;, \tag{2.10}$$

où $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ est une solution de l'équation de Poisson (en temps continu)

$$\mathcal{L}\hat{f} = -\{f - \pi(f)\} \;. \tag{2.11}$$

Nous introduisons une nouvelle classe de variables de contrôle basée sur l'expression de la variance asymptotique $\sigma^2_\infty(f)$ donnée dans (2.10). Puisque $\pi(\mathcal{L}(\theta^{\mathrm{T}}\psi)) = 0$ pour tout $\theta \in \mathbb{R}^d$, nous considérons la variable de contrôle $\mathcal{L}(\theta^*(f)^{\mathrm{T}}\psi)$ où $\theta^*(f)$ est choisi en minimisant

$$\theta \mapsto \sigma^2_\infty(f + \mathcal{L}(\theta^{\mathrm{T}}\psi)) \;. \tag{2.12}$$

Bien que $\mathcal{L}(\theta^*(f)^{\mathrm{T}}\psi)$ soit une variable de contrôle pour la diffusion Langevin associée à f, le choix de cette option est motivé par le fait que pour certains MCMCs, le critère d'optimisation de la variance asymptotique $\sigma^2_{\infty,\mathrm{d}}(f)$ définie dans (2.6) est (à un facteur multiplicatif près) une bonne approximation de la variance asymptotique de la diffusion de Langevin $\sigma^2_\infty(f)$ définie dans (2.10). De plus, la minimisation de (2.12) admet une solution unique $\theta^*(f)$, qui est en général facile à estimer. Il est intéressant de noter qu'il n'est pas nécessaire de connaître la solution de Poisson $\hat{f}$ pour minimiser (2.12).

2.3 Stochastic Gradient Langevin Dynamics

L'algorithme ULA défini dans (1.10) nécessite de calculer à chaque étape le gradient du potentiel U. Cependant, dans l'apprentissage machine bayésien, U est proportionnel au (moins le) logarithme de la distribution postérieure et est la somme d'un grand nombre d'observations. Plus précisément, dénotons par $\mathbf{z} = \{z_i\}_{i=1}^N$ les observations et considérons une situation où la distribution cible π apparaît comme le postérieur dans un problème d'inférence bayésien avec une densité a priori $\pi_0(\theta)$ et un grand nombre $N \gg 1$ d'observations z_i avec probabilités $p(z_i|\theta)$. Dans ce cas, $\pi(\theta) = \pi_0(\theta)\prod_{i=1}^N p(z_i|\theta)$. Nous dénotons $U_i(\theta) = -\log((p(z_i|\theta))$ pour $i \in \{1, \ldots, N\}$, $U_0(\theta) = -\log(\pi_0(\theta))$, $U =$

$\sum_{i=0}^{N} U_i$. Evaluer le gradient de U, $\nabla U(\theta) = \sum_{i=0}^{N} \nabla U_i(\theta)$ en $\theta \in \mathbb{R}^d$ est un calcul coûteux. Le coût d'une itération d'ULA est de l'ordre de Nd, ce qui est prohibitif pour des ensembles de données massifs ($N \gg 1$).

Afin d'adapter l'algorithme aux grandes données, Welling and Teh [WT11] a suggéré de remplacer ∇U par une estimation non biaisée $\nabla U_0 + (N/p) \sum_{i \in S} \nabla U_i$ où S est un mini-batch avec remplacement de $\{1, \dots, N\}$ de la taille p. Une seule mise à jour de l'algorithme en résultant, Stochastic Gradient Langevin Dynamics (SGLD), est alors donnée pour $k \in \mathbb{N}$ par

$$\theta_{k+1} = \theta_k - \gamma \left(\nabla U_0(\theta_k) + \frac{N}{p} \sum_{i \in S_{k+1}} \nabla U_i(\theta_k) \right) + \sqrt{2\gamma} W_{k+1} \, .$$

L'idée de n'utiliser qu'une fraction des points de données pour calculer une estimation non biaisée du gradient à chaque itération vient de Stochastic Gradient Descent (SGD) qui est un algorithme populaire pour minimiser le potentiel U. SGD est très similaire au SGLD car il se caractérise par la même récursivité que le SGLD mais sans bruit gaussien:

$$\theta_{k+1} = \theta_k - \gamma \left(\nabla U_0(\theta_k) + \frac{N}{p} \sum_{i \in S_{k+1}} \nabla U_i(\theta_k) \right) \, .$$

En supposant pour simplifier que U a un minimiseur $\theta^\star$, nous pouvons définir une version des variables de contrôle pour SGLD, SGLDFP, voir [Dub+16; Che+17], donnée pour $k \in \mathbb{N}$ par

$$\theta_{k+1} = \theta_k - \gamma \left(\nabla U_0(\theta_k) - \nabla U_0(\theta^\star) + \frac{N}{p} \sum_{i \in S_{k+1}} \{ \nabla U_i(\theta_k) - \nabla U_i(\theta^\star) \} \right) + \sqrt{2\gamma} W_{k+1} \, .$$

Dans le Chapitre 7, nous donnons un aperçu des liens entre SGLD, SGLDFP, ULA et SGD. Dans notre analyse, les algorithmes sont utilisés avec une taille de pas constante et les paramètres sont réglés sur les valeurs standards utilisées dans la pratique : en particulier, $\gamma \approx 1/N$. Les algorithmes ULA, SGD, SGLD et SGLDFP définissent des chaînes de Markov homogènes, dont chacune admet une distribution stationnaire unique utilisée comme proxy de π. Notre principale contribution est de montrer que, si les distributions invariantes d'ULA et de SGLDFP se rapprochent de π à mesure que le nombre de points de données N augmente, au contraire, la mesure invariante de SGLD ne s'approche jamais de la distribution cible π et est en fait très similaire à celle de SGD.

Nous montrons que le nombre d'itérations nécessaires pour obtenir un échantillon à distance ε de π en distance de Wasserstein est le même pour ULA et SGLDFP. Cependant, pour ULA, le coût d'une itération est de Nd, ce qui est beaucoup plus élevé que pd le coût d'une itération pour SGLDFP. En d'autres termes, pour obtenir un échantillon approximatif de la distribution cible avec une précision de $O(1/\sqrt{N})$ en distance de 2-Wasserstein, LMC nécessite environ N d'opérations, contrairement au SGLDFP qui ne nécessite qu'un nombre d'opérations indépendant de N.

Part I

Extensions of the unadjusted Langevin algorithm

Chapter 3

Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo

Nicolas Brosse [1], Alain Durmus [2], Éric Moulines [3], Marcelo Pereyra [4]

Abstract

This paper presents a detailed theoretical analysis of the Langevin Monte Carlo sampling algorithm recently introduced in [DMP18] when applied to log-concave probability distributions that are restricted to a convex body K. This method relies on a regularisation procedure involving the Moreau-Yosida envelope of the indicator function associated with K. Explicit convergence bounds in total variation norm and in Wasserstein distance of order 1 are established. In particular, we show that the complexity of this algorithm given a first order oracle is polynomial in the dimension of the state space. Finally, some numerical experiments are presented to compare our method with competing MCMC approaches from the literature.

[1]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France. nicolas.brosse@polytechnique.edu

[2]Ecole Normale Supérieure CMLA 61, Av. du Président Wilson 94235 Cachan Cedex, France Email: alain.durmus@cmla.ens-cachan.fr

[3]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France. eric.moulines@polytechnique.edu

[4]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, U.K. m.pereyra@hw.ac.uk

3.1 Introduction

Many statistical inference problems involve estimating parameters subject to constraints on the parameter space. In a Bayesian setting, these constraints define a posterior distribution π with bounded support. Some examples include truncated data problems which arise naturally in failure and survival time studies [KM05], ordinal data models [JA06], constrained Lasso and ridge regressions [Cel+12], Latent Dirichlet Allocation [BNJ03], and non-negative matrix factorization [PBJ14]. Drawing samples from such constrained distributions is a challenging problem that has been investigated in many papers; see [GSL92], [PP14], [LS15], [BEL15]. All these works are based on efficient Markov Chain Monte Carlo methods to approximate the posterior distribution; however, with the exception of the recent work [BEL15], these methods are not theoretically well understood and do not provide any theoretical guarantees on the estimations delivered.

Recently a new MCMC method has been proposed in [DMP18] to sample from a non-smooth log-concave probability distribution on $\mathbb{R}^d$. This method is mainly based on a carefully designed regularised version of the target distribution π that enjoys a number of favourable properties that are useful for MCMC simulation. In this study, we analyse the complexity of this algorithm when applied to log-concave distributions constrained to a convex set, with a focus on the complexity as the dimension of the state space increases. More precisely, we establish explicit bounds in total variation norm and in Wasserstein distance of order 1 between the iterates of the Markov kernel defined by the algorithm and the target density π.

The paper is organised as follows. Section 3.2.1 introduces the MCMC method of [DMP18]. The main complexity result is stated in Section 3.2.2 and compared to previous works on the subject. The proof of this result is presented in Section 3.3 and Section 3.4. The methodology is then illustrated and compared to other approaches via experiments in Section 3.5. Proofs are finally reported in Section 3.6.

3.2 The Moreau-Yosida Unadjusted Langevin Algorithm (MYULA)

3.2.1 Presentation of MYULA

Let π be a probability measure on $\mathbb{R}^d$ with density w.r.t. the Lebesgue measure given for all $x \in \mathbb{R}^d$ by $\pi(x) = \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y$, where $U : \mathbb{R}^d \to (-\infty, +\infty]$ is a measurable function. In the sequel, U will be referred to as the potential associated with π. Assume for the moment that U is continuously differentiable. Then, the unadjusted Langevin algorithm (ULA) introduced in [Par81] (see also [RT96]) can be used to sample from π. This algorithm is based on the overdamped Langevin stochastic differential equation (SDE) associated with U,

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \,, \tag{3.1}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Under mild assumptions on ∇U, this SDE has a unique strong solution $(Y_t)_{t\geq 0}$ and defines a strong Markovian semigroup

$(P_t)_{t\geq 0}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which is ergodic with respect to π, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-field on $\mathbb{R}^d$. Since simulating exact solutions of (5.7) is in general computationally impossible or very hard, ULA considers the Euler-Maruyama discretization associated with (5.7) to approximate samples from π. Precisely, ULA constructs the discrete-time Markov chain $(X_k)_{k\geq 0}$, started at X_0, given for $k \in \mathbb{N}$ by:

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1} \ ,$$

where $\gamma > 0$ is the stepsize and $(Z_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. standard Gaussian d-dimensional vectors; the process $(X_k)_{k\geq 0}$ is used as approximate samples from π. However, the ULA algorithm cannot be directly applied to a distribution π restricted to a compact convex set. Let $\mathsf{K} \subset \mathbb{R}^d$ be a convex body, i.e. a compact convex set with non-empty interior and $\iota_{\mathsf{K}} : \mathbb{R}^d \to \{0, +\infty\}$ be the (convex) indicator function of K, defined for $x \in \mathbb{R}^d$ by,

$$\iota_{\mathsf{K}}(x) = \begin{cases} +\infty & \text{if } x \notin \mathsf{K}, \\ 0 & \text{if } x \in \mathsf{K} \ . \end{cases}$$

Let $f : \mathbb{R}^d \to \mathbb{R}$. In this paper we consider any probability density π associated to a potential $U : \mathbb{R}^d \to (-\infty, +\infty]$ of the form

$$U = f + \iota_{\mathsf{K}} \ , \tag{3.2}$$

and assume that the function f and the convex body K satisfy the following assumptions. For $x \in \mathbb{R}^d$ and $r > 0$, denote by $\mathrm{B}(x, r)$ the closed ball of center x and radius r: $\mathrm{B}(x, r) = \left\{ y \in \mathbb{R}^d : \|y - x\| \leq r \right\}$.

H1. *(i) f is convex.*

(ii) f is continuously differentiable on $\mathbb{R}^d$ and gradient Lipschitz with Lipschitz constant L_f, i.e. for all $x, y \in \mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\| \ . \tag{3.3}$$

H2. *There exist $r, R > 0$, $r \leq R$, such that,*

$$\mathrm{B}(0, r) \subset \mathsf{K} \subset \mathrm{B}(0, R) \ .$$

To apply ULA, [DMP18] suggested to carefully regularize U in such a way that 1) the convexity of U is preserved (this property is key to the theoretical analysis of the algorithm), 2) the regularisation of U is continuously differentiable and gradient Lipschitz (this regularity property is key to the algorithm's stability), and 3) the resulting approximation is close to π (e.g. in total variation norm). The tool used to construct such an approximation is the Moreau-Yosida envelope of ι_{K}, $\iota_{\mathsf{K}}^{\lambda} : \mathbb{R}^d \to \mathbb{R}_+$ defined for $x \in \mathbb{R}^d$ (see e.g. [RW98, Chapter 1 Section G]) by,

$$\iota_{\mathsf{K}}^{\lambda}(x) = \inf_{y\in\mathbb{R}^d} \left(\iota_{\mathsf{K}}(y) + (2\lambda)^{-1} \|x - y\|^2 \right) = (2\lambda)^{-1} \|x - \mathrm{proj}_{\mathsf{K}}(x)\|^2 \ , \tag{3.4}$$

where $\lambda > 0$ is a regularization parameter and $\mathrm{proj}_{\mathsf{K}}$ is the projection onto K. By [RW98, Example 10.32, Theorem 9.18], the function $\iota_{\mathsf{K}}^{\lambda}$ is convex and continuously differentiable with gradient given for all $x \in \mathbb{R}^d$ by:

$$\nabla \iota_{\mathsf{K}}^{\lambda}(x) = \lambda^{-1}(x - \mathrm{proj}_{\mathsf{K}}(x)) \,. \tag{3.5}$$

Moreover, [RW98, Proposition 12.19] implies that $\iota_{\mathsf{K}}^{\lambda}$ is λ^{-1}-gradient Lipschitz: for all $x, y \in \mathbb{R}^d$,

$$\left\| \nabla \iota_{\mathsf{K}}^{\lambda}(x) - \nabla \iota_{\mathsf{K}}^{\lambda}(y) \right\| \leq \lambda^{-1} \|x - y\| \,. \tag{3.6}$$

Adding f to $\iota_{\mathsf{K}}^{\lambda}$ under **H**1 leads to the regularization $U^{\lambda} : \mathbb{R}^d \to \mathbb{R}$ of the potential U defined for all $x \in \mathbb{R}^d$ by

$$U^{\lambda}(x) = f(x) + \iota_{\mathsf{K}}^{\lambda}(x) \,. \tag{3.7}$$

The following lemma shows that the probability measure π^{λ} on $\mathbb{R}^d$, with density with respect to the Lebesgue measure, also denoted by π^{λ} and given for all $x \in \mathbb{R}^d$ by

$$\pi^{\lambda}(x) = \frac{\mathrm{e}^{-U^{\lambda}(x)}}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^{\lambda}(s)} \mathrm{d}s} \,, \tag{3.8}$$

is well defined. It also shows that U^{λ} has a minimizer $x^{\star} \in \mathbb{R}^d$, a fact that will be used in Section 3.4. Note that the dependence of $x^{\star}$ on λ is implicit.

Lemma 3.1. *Assume **H**1-(i) and **H**2. For all $\lambda > 0$,*

a) U^{λ} has a minimizer $x^{\star} \in \mathbb{R}^d$, i.e. for all $x \in \mathbb{R}^d$, $U^{\lambda}(x) \geq U^{\lambda}(x^{\star})$.

b) $\mathrm{e}^{-U^{\lambda}}$ defines a proper density of a probability measure on $\mathbb{R}^d$, i.e.

$$0 < \int_{\mathbb{R}^d} \mathrm{e}^{-U^{\lambda}(y)} \mathrm{d}y < +\infty \,.$$

Proof. Note that [DMP18, Proposition 1] provides a proof in a more general case. Given the specific form of U^{λ}, a short and self-contained proof can be found in Section 3.6.1. □

Under **H**1, for all $\lambda > 0$, π^{λ} is log-concave and U^{λ} is continuously differentiable by (3.5), with ∇U^{λ} given for all $x \in \mathbb{R}^d$ by

$$\nabla U^{\lambda}(x) = -\nabla \log \pi^{\lambda}(x) = \nabla f(x) + \lambda^{-1}(x - \mathrm{proj}_{\mathsf{K}}(x)) \,. \tag{3.9}$$

In addition, by (3.6), ∇U^{λ} is Lipschitz with constant $L \leq L_f + \lambda^{-1}$. Since U^{λ} is continuously differentiable, ULA is well defined. The algorithm proposed in [DMP18] then proceeds by using the Euler-Maruyama discretization of the Langevin equation associated with U^{λ}, with π^{λ} as proxy, to generate approximate samples from π. Precisely, it uses the Markov chain $(X_k)_{k \in \mathbb{N}}$, started at X_0, given for all $k \in \mathbb{N}$ by

$$X_{k+1} = (1 - \tfrac{\gamma}{\lambda}) X_k - \gamma \nabla f(X_k) + \tfrac{\gamma}{\lambda} \mathrm{proj}_{\mathsf{K}}(X_k) + \sqrt{2\gamma} Z_{k+1} \,, \tag{3.10}$$

where $(Z_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. standard Gaussian d-dimensional vectors and $\gamma > 0$ is the stepsize. Note that one iteration (3.10) requires a projection onto the convex body K and the evaluation of ∇f. The kernel of the homogeneous Markov chain defined by (3.10) is given for $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by,

$$R_\gamma(x, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(-(4\gamma)^{-1} \left\|y - x + \gamma \nabla U^\lambda(x)\right\|^2\right) \mathrm{d}y \;, \tag{3.11}$$

where U^λ is defined in (3.7). Since the target density for the Markov chain (3.10) is the regularized measure π^λ and not π, the algorithm is named the Moreau-Yosida regularized Unadjusted Langevin Algorithm (MYULA).

3.2.2 Context and contributions

The total variation distance between two probability measures μ and ν is defined by $\|\mu - \nu\|_{\mathrm{TV}} = 2\sup_{\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)} |\mu(\mathsf{A}) - \nu(\mathsf{A})|$. Let $\phi, \psi : \mathbb{R}_+ \to \mathbb{R}_+$. Denote by $\phi = \tilde{\mathcal{O}}(\psi)$ or $\phi = \tilde{\Omega}(\psi)$ if there exist $C, c \geq 0$ such that for all $t \in \mathbb{R}_+$ $\phi(t) \leq C\psi(t)(\log t)^c$ or $\phi(t) \geq C\psi(t)(\log t)^c$ respectively. Our main result is the following:

Theorem 3.2. *Assume **H**1 and **H**2. For all $\varepsilon > 0$ and $x \in \mathbb{R}^d$, there exist $\lambda > 0$ and $\gamma \in \left(0, \lambda(1 + L_f^2\lambda^2)^{-1}\right)$ such that,*

$$\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon \quad \textit{for} \quad n = \tilde{\Omega}(d^5) \;,$$

where R_γ is defined in (3.11).

The proof of Theorem 3.2 follows from combining Proposition 3.6 and Proposition 3.4 below. Note that these two results imply explicit bounds between R_γ^n and π for all $n \in \mathbb{N}$ and $\gamma > 0$.

The problem of sampling from a probability measure restricted to a convex compact support has been investigated in several works, mainly in the fields of theoretical computer science and Bayesian statistics. In computer science, a line of works starting with [DF91] has studied the convergence of the ball walk and the hit-and-run algorithm towards the uniform density on a convex body K, or more generally to a log-concave density. The best complexity result is achieved by [LV07, Theorem 2.1] who establishes a mixing time for these two algorithms of order $\tilde{\mathcal{O}}(d^4)$. However, observe that contrary to Theorem 3.2, this result assumes that π is in near-isotropic position, i.e. there exists $C \in \mathbb{R}_+^*$ such that for all $u \in \mathbb{R}^d$, $\|u\| = 1$,

$$C^{-1} \leq \int_{\mathbb{R}^d} \langle u, x\rangle^2 \, \pi(\mathrm{d}x) \leq C \;. \tag{3.12}$$

Note that [LV07, Section 2.5] gives also an algorithm of complexity $\tilde{\mathcal{O}}(d^5)$ which provides an invertible linear map T of $\mathbb{R}^d$ such that the measure π_T defined for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\pi_T(\mathsf{A}) = \pi(T^{-1}(\mathsf{A})) \;,$$

is log-concave and near-isotropic. Also note that, unlike our method, each iteration of the ball walk or the hit-and-run algorithm requires a call to a zero-order oracle, which given $x \in \mathbb{R}^d$, returns the value $U(x)$. MYULA does not require to fulfill the condition (3.12) and is thus dispensed of preprocessing step. However, MYULA needs a first-order oracle which returns the value $\nabla f(x)$ for $x \in \mathbb{R}^d$.

As emphasized in the introdution, probability distributions with convex compact supports or more generally with constrained parameters arise naturally in Bayesian statistics. [GSL92] includes many examples of such problems and suggests to use a Gibbs sampler, see also [RDS04]. [CSI12, Chapter 6] addresses the subject with the additional difficulty of computing normalizing constants. Recently, [PP14] adapted the Hamiltonian Monte Carlo method to sample from a truncated multivariate gaussian, and [LS15] suggested a new approach which consists in mapping the constrained domain to a sphere in an augmented space. However, these methods are not well understood from a theoretical viewpoint, and do not provide any theoretical guarantees for the estimations delivered.

Concerning the ULA algorithm, when U is continuously differentiable, the first explicit convergence bounds have been obtained by [Dal17b], [DM17], [DM16]. In the constrained case $U = f + \iota_{\mathsf{K}}$, [BEL15] suggests a projection step in ULA i.e. to consider the Markov chain $(\tilde{X}_k)_{k\geq 0}$, defined for all $k \in \mathbb{N}$ by

$$\tilde{X}_{k+1} = \mathrm{proj}_{\mathsf{K}}\left(\tilde{X}_k - \gamma \nabla U(\tilde{X}_k) + \sqrt{2\gamma} Z_{k+1}\right) . \tag{3.13}$$

with $\tilde{X}_0 = 0$. This method is referred to as the Projected Langevin Monte Carlo (PLMC) algorithm. As in MYULA, one iteration of PLMC requires a projection onto K and an evaluation of ∇f. Let $\tilde{R}_\gamma$ be the Markov kernel defined by (3.13). [BEL15] proved that for all $\varepsilon > 0$, $\|\delta_0 \tilde{R}_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$ for $n = \tilde{\Omega}(d^7)$ if π is the uniform density on K and $n = \tilde{\Omega}(d^{12})$ if π is a log-concave density. Theorem 3.2 improves these bounds for the MYULA algorithm. Note however that the iterations of PLMC stay within the constraint set K and this property can be useful in some specific problems. Nevertheless, there is a wide range of settings where this property is not particularly beneficial, for example in the case of the computation of volumes discussed in Section 3.5, or in Bayesian model selection where it is necessary to estimate marginal likelihoods.

3.3 Distance between π and π^λ

In this section, we derive bounds between π and π^λ in total variation and in Wasserstein distance (recall that π is associated with a potential of the form (3.2) and π^λ is given by (3.8)). It is shown that the approximation error in both distances can be made arbitrarily small by adjusting the regularisation parameter λ.

The main quantity of interest to analyze the distance between π and π^λ will appear to be the integral of $x \mapsto \mathrm{e}^{-(2\lambda)^{-1}\|x - \mathrm{proj}_{\mathsf{K}}(x)\|^2}$ over $\mathbb{R}^d$. This constant is linked to useful notions borrowed from the field of convex geometry [Kam09, Proposition 3]. Indeed,

Fubini's theorem gives the following equality:

$$\begin{aligned}\int_{\mathbb{R}^d} \mathrm{e}^{-(2\lambda)^{-1}\|x-\mathrm{proj}_{\mathsf{K}}(x)\|^2}\mathrm{d}x &= \int_{\mathbb{R}_+}\int_{\mathbb{R}^d} \mathbb{1}_{[\|x-\mathrm{proj}_{\mathsf{K}}(x)\|,+\infty)}(t)\lambda^{-1}t\mathrm{e}^{-t^2/(2\lambda)}\mathrm{d}x\mathrm{d}t\ , \\ &= \int_{\mathbb{R}_+} \mathrm{Vol}\left(\mathsf{K}+\mathrm{B}(0,t)\right)\lambda^{-1}t\mathrm{e}^{-t^2/(2\lambda)}\mathrm{d}t\ , \end{aligned} \tag{3.14}$$

where $\mathsf{A}+\mathsf{B}$ is the Minkowski sum of $\mathsf{A},\mathsf{B}\subset\mathbb{R}^d$, i.e. $\mathsf{A}+\mathsf{B}=\{x+y : x\in\mathsf{A}, y\in\mathsf{B}\}$, and we have used in the last line that for all $t\in\mathbb{R}_+$, $\mathsf{K}+\mathrm{B}(0,t)=\{x\in\mathbb{R}^d : \|x-\mathrm{proj}_{\mathsf{K}}(x)\|\le t\}$. It turns out that $t\mapsto \mathrm{Vol}\left(\mathsf{K}+\mathrm{B}(0,t)\right)$ on $\mathbb{R}_+$ is a polynomial. More precisely, Steiner's formula states that for all $t\ge 0$,

$$\mathrm{Vol}(\mathsf{K}+\mathrm{B}(0,t)) = \sum_{i=0}^{d} t^i\kappa_i\mathscr{V}_{d-i}(\mathsf{K})\ , \tag{3.15}$$

where $\{\mathscr{V}_i(\mathsf{K})\}_{0\le i\le d}$ are the intrinsic volumes of K, κ_i denotes the volume of the unit ball in $\mathbb{R}^i$, i.e.

$$\kappa_i = \pi^{i/2}/\mathbf{\Gamma}(1+i/2)\ , \tag{3.16}$$

and $\mathbf{\Gamma}:\mathbb{R}_+^*\to\mathbb{R}_+^*$ is the Gamma function. We refer to [Sch13, Chapter 4.2] for this result and an introduction to this topic. Combining (3.14) and (3.15) gives:

$$\int_{\mathbb{R}^d} \mathrm{e}^{-(2\lambda)^{-1}\|x-\mathrm{proj}_{\mathsf{K}}(x)\|^2}\mathrm{d}x = \sum_{i=0}^{d}\mathscr{V}_i(\mathsf{K})(2\pi\lambda)^{(d-i)/2}\ . \tag{3.17}$$

This expression will provide a precise analysis of the distance in total variation and Wasserstein distance between π and π^λ, in particular when π is the uniform density on K. However, in more general cases, an additional assumption on the relation between f and K is necessary to bound the distance between π and π^λ. Under **H**1-(i) and **H**2, f has a minimum x_{K} on K. Define

$$\widetilde{\mathsf{K}} = \{x\in\mathsf{K} \mid \mathrm{B}(x,r)\subset\mathsf{K}\}\ . \tag{3.18}$$

$\widetilde{\mathsf{K}}$ has the following property.

Lemma 3.3. *Assume **H**2. $\widetilde{\mathsf{K}}$ is a non-empty convex compact set.*

Proof. The proof is postponed to Section 3.6.2. □

H3. *(i) There exists $\Delta_1>0$ such that $\exp\left(\inf_{\mathsf{K}^c}(f)-\max_{\mathsf{K}}(f)\right)\ge\Delta_1$.*

(ii) There exists $\Delta_2\ge 0$ such that $0\le f(\mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}}))-f(x_{\mathsf{K}})\le\Delta_2$.

These assumptions are illustrated in Figure 3.1. Under **H**3-(i), the application of Steiner's formula is possible and reveals the precise dependence of the bounds with respect to the intrinsic volumes of K. A complementary view is possible under **H**3-(ii). The obtained bounds are less precise regarding K but more robust with respect to f. Note that if $x_{\mathsf{K}}\in\widetilde{\mathsf{K}}$, Δ_2 can be chosen equal to 0. On the other hand, if f is assumed to be ℓ-Lipschitz inside K, Δ_2 is less than ℓR.

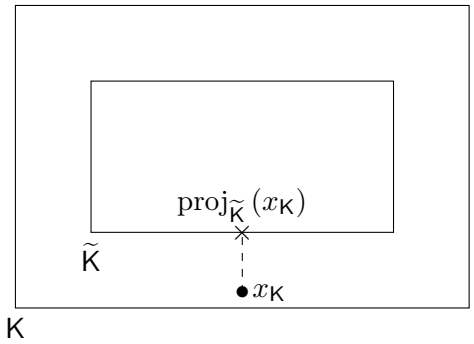

Figure 3.1: Illustration of **H**3

Proposition 3.4. *Assume **H**1-(i) and **H**2.*

*a) Assume **H**3-(i). For all $\lambda > 0$,*

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \le 2\left(1 + \Delta_1 \mathsf{D}(\mathsf{K},\lambda)^{-1}\right)^{-1} , \tag{3.19}$$

where

$$\mathsf{D}(\mathsf{K},\lambda) = (\mathrm{Vol}\,\mathsf{K})^{-1} \sum_{i=0}^{d-1} (2\pi\lambda)^{(d-i)/2} \mathscr{V}_i(\mathsf{K}) , \tag{3.20}$$

and $\mathscr{V}_i(\mathsf{K})$ are defined in (3.15). In addition, for all $\lambda \in (0, (2\pi)^{-1}(r/d)^2)$,

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \le 2^{3/2} \Delta_1^{-1} (\pi\lambda)^{1/2} d r^{-1} . \tag{3.21}$$

*b) Assume **H**3-(ii). For all $\lambda \in (0, 16^{-1}(r/d)^2]$,*

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \le (4/r) \exp\left(4\lambda\,(\Delta_2/r)^2\right) \left\{\sqrt{\lambda}(d + \Delta_2) + (2\lambda\Delta_2)/r\right\} . \tag{3.22}$$

Proof. The proof is postponed to Section 3.6.3. □

In the particular case where $f = 0$ and π is the uniform density on K, Δ_1 equals 1 and the inequality (3.19) is in fact an equality. The dependence of the upper bound in (3.19) w.r.t. to λ, d, r is sharp. Indeed, for the cube C of side c, $\mathsf{D}(\mathsf{C},\lambda)$ can be explicitly computed. [KR97, Theorem 4.2.1] gives for $i \in \{0,\dots,d\}$, $\mathscr{V}_i(\mathsf{C}) = \binom{d}{i} c^i$, which implies:

$$\mathsf{D}(\mathsf{C},\lambda) = \left(1 + c^{-1}\sqrt{2\pi\lambda}\right)^d - 1 ,$$
$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} = 2\left\{1 - \left(1 + c^{-1}\sqrt{2\pi\lambda}\right)^{-d}\right\} , \text{ for } U = \iota_{\mathsf{C}} .$$

For two probability measures μ and ν on $\mathcal{B}(\mathbb{R}^d)$, the Wasserstein distance of order $p \in \mathbb{N}^*$ between μ and ν is defined by

$$W_p(\mu,\nu) = \left(\inf_{\zeta \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \,\mathrm{d}\zeta(x,y)\right)^{1/p} ,$$

where $\Pi(\mu,\nu)$ is the set of transference plans of μ and ν. ζ is a transference plan of μ and ν if it is a probability measure on $(\mathbb{R}^d\times\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d\times\mathbb{R}^d))$ such that for all $\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)$, $\zeta(\mathsf{A}\times\mathbb{R}^d)=\mu(\mathsf{A})$ and $\zeta(\mathbb{R}^d\times\mathsf{A})=\nu(\mathsf{A})$.

Proposition 3.5. *Assume **H**1-(i) and **H**2.*

*a) Assume **H**3-(i). For all $\lambda>0$,*

$$W_1(\pi,\pi^\lambda)\leq \Delta_1^{-1}\mathsf{E}(\mathsf{K},\lambda,R)\ ,$$

where

$$\mathsf{E}(\mathsf{K},\lambda,R)=(\mathrm{Vol}(\mathsf{K}))^{-1}\sum_{i=0}^{d-1}\mathscr{V}_i(\mathsf{K})\,(2\pi\lambda)^{(d-i)/2}\left\{2R+[\lambda(d-i+2)]^{1/2}\right\}\ ,$$

and $\mathscr{V}_i(\mathsf{K})$ are defined in (3.15).

*b) In addition, assuming **H**3-(i), for all $\lambda\in\left(0,(2\pi)^{-1}d^{-2}r^2\right)$,*

$$W_1(\pi,\pi^\lambda)\leq \Delta_1^{-1}(2\pi\lambda)^{1/2}dr^{-1}\left(2R+r\left(3/(2d\pi)\right)^{1/2}\right)\ .$$

*c) Assume **H**3-(ii). For all $\lambda\in\left(0,16^{-1}(r/d)^2\right]$,*

$$W_1(\pi,\pi^\lambda)\leq 4\exp\left(4\lambda\left(\Delta_2/r\right)^2\right)\left\{\sqrt{\lambda}(d+\Delta_2)(R/r)+(2\lambda\Delta_2 R)/r^2+\sqrt{\pi\lambda}\right\}\ .$$

Proof. The proof is postponed to Section 3.6.4. □

Note that the bounds in Wasserstein distance between π and π^λ are roughly similar to those obtained in total variation norm.

3.4 Convergence analysis of MYULA

We now analyse the convergence of the Markov kernel R_γ, given by (3.11), to the target density π^λ defined in (3.8). For $x\in\mathbb{R}^d$ and $n\in\mathbb{N}$, explicit bounds in total variation norm and in Wasserstein distance between $\delta_x R_\gamma^n$ and π^λ are provided in Proposition 3.6 and Proposition 3.7. Because of the regularisation procedure performed in Section 3.2.1, the convergence analysis of MYULA (3.10) is an application of results of [DM17] and [DM16].

3.4.1 Convergence in total variation norm

Define $\omega:\mathbb{R}_+\to\mathbb{R}_+$ for all $r\geq 0$ by

$$\omega(r)=r^2/\left\{2\mathbf{\Phi}^{-1}(3/4)\right\}^2\ , \tag{3.23}$$

where $\mathbf{\Phi}(x)=(2\pi)^{-1/2}\int_{-\infty}^x \mathrm{e}^{-t^2/2}\mathrm{d}t$.

Proposition 3.6. *Assume* ***H****1 and* ***H****2. Let* $\lambda > 0$, L *be the Lipschitz constant of* ∇U^λ *defined in* (3.7) *and* $\bar{\gamma} \in (0, \lambda^{-1}L^{-2})$. *Then for all* $\varepsilon > 0$ *and* $x \in \mathbb{R}^d$, *we get:*

$$\|\delta_x R_\gamma^n - \pi^\lambda\|_{\mathrm{TV}} \leq \varepsilon \;, \tag{3.24}$$

provided that $n > T\gamma^{-1}$ *with*

$$T = \left(\log\{A_2(x)\} - \log(\varepsilon/2)\right) \Big/ \left(-\log(\kappa)\right) \;, \tag{3.25a}$$

$$\gamma \leq \frac{-d + \sqrt{d^2 + (2/3)A_1(x)\varepsilon^2(L^2T)^{-1}}}{2A_1(x)/3} \wedge \bar{\gamma} \;, \tag{3.25b}$$

where

$$A_1(x) = L^2\left(\|x - x^\star\|^2 + 2(d + 8\lambda^{-1}R^2)\mathrm{e}^{\gamma(\lambda^{-1} - \bar{\gamma}L^2)}(\lambda^{-1} - \bar{\gamma}L^2)^{-1}\right) \;,$$

$$\log(\kappa) = -\log(2)(4\lambda)^{-1}\left[\log\left\{\left(1 + \mathrm{e}^{(8\lambda)^{-1}\omega\{\max(1,4R)\}}\right)(1 + \max(1, 4R))\right\} + \log(2)\right]^{-1} \;,$$

$$A_2(x) = 6 + 2^{3/2}\left(d\lambda + 8R^2\right)^{1/2} + 2(A_1(x)/L^2)^{1/2} \;,$$

and $x^\star$ *is a minimizer of* U^λ.

Proof. To apply [DM17, Theorem 21], it is sufficient to check the assumption [DM17, H3], i.e. there exist $\tilde{R} \geq 0$ and $m > 0$ such that for all $x, y \in \mathbb{R}^d$, $\|x - y\| \geq \tilde{R}$,

$$\left\langle \nabla U^\lambda(x) - \nabla U^\lambda(y), x - y \right\rangle \geq m\|x - y\|^2 \;. \tag{3.26}$$

By (3.5) and the Cauchy-Schwarz inequality, we have:

$$\left\langle \nabla \iota_{\mathsf{K}}^\lambda(x) - \nabla \iota_{\mathsf{K}}^\lambda(y), x - y \right\rangle \geq \lambda^{-1}\left(\|x - y\|^2 - 2\left\{\sup_{z \in \mathsf{K}} \|z\|\right\}\|x - y\|\right) \;,$$

which implies under **H**1-(i) and **H**2 that (3.26) holds for $\tilde{R} = 4R$ and $m = (2\lambda)^{-1}$. □

Combining Proposition 3.4 and Proposition 3.6 determines the stepsize γ and the number of samples n to get $\|\delta_{x^\star} R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$. λ is chosen of order $\varepsilon^2 r^2 d^{-2}\Delta_1^2$ under **H** 3-(i) and $\varepsilon^2 r^2 \min(d^{-2}, \Delta_2^{-2})$ under **H**3-(ii). The orders of magnitude of n in d, ε, R, r are reported in Table 3.1, along with the results of [BEL15]. The dependency of n towards Δ_1, Δ_2 is presented in Table 3.2. A detailed table is provided in Table 3.3.

3.4.2 Convergence in Wasserstein distance for strongly convex f

In this section, f is assumed to satisfy an additional assumption.

H4. $f : \mathbb{R}^d \mapsto \mathbb{R}$ *is* m*-strongly convex , i.e. there exists* $m > 0$ *such that for all* $x, y \in \mathbb{R}^d$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + (m/2)\|x - y\|^2 \;. \tag{3.27}$$

Upper bound on n to get $\|\delta_{x^\star} R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$	$d \to +\infty$	$\varepsilon \to 0$	$R \to +\infty$	$r \to 0$
Proposition 3.4 and Proposition 3.6	$\tilde{\mathcal{O}}(d^5)$	$\tilde{\mathcal{O}}(\varepsilon^{-6})$	$\tilde{\mathcal{O}}(R^4)$	$\tilde{\mathcal{O}}(r^{-4})$
[BEL15, Theorem 1] π uniform on K	$\tilde{\mathcal{O}}(d^7)$	$\tilde{\mathcal{O}}(\varepsilon^{-8})$	$\tilde{\mathcal{O}}(R^6)$	$\tilde{\mathcal{O}}(r^{-6})$
[BEL15, Theorem 1] π log concave	$\tilde{\mathcal{O}}(d^{12})$	$\tilde{\mathcal{O}}(\varepsilon^{-12})$	$\tilde{\mathcal{O}}(R^{18})$	$\tilde{\mathcal{O}}(r^{-18})$

Table 3.1: dependency of n on d, ε, R and r to get $\|\delta_{x^\star} R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$

Upper bound on n to get $\|\delta_{x^\star} R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$	$\Delta_1 \to 0$	$\Delta_2 \to +\infty$
Proposition 3.4 and Proposition 3.6	$\tilde{\mathcal{O}}(\Delta_1^{-4})$	$\tilde{\mathcal{O}}(\Delta_2^4)$

Table 3.2: dependency of n on Δ_1 and Δ_2 to get $\|\delta_{x^\star} R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$

	$d \to +\infty$	$\varepsilon \to 0$	$R \to +\infty$	$r \to 0$	$\Delta_1 \to 0$	$\Delta_2 \to +\infty$
L, λ^{-1}	d^2	ε^{-2}	1	r^{-2}	Δ_1^{-2}	Δ_2^2
$A_1(x)$	d^4	ε^{-4}	R^2	r^{-4}	Δ_1^{-4}	Δ_2^4
$-\log(\kappa)$	1	1	R^{-2}	1	1	1
$A_2(x)$	1	ε^{-1}	R	r^{-1}	Δ_1^{-1}	Δ_2
T	1	$\log(\varepsilon^{-1})$	R^2	$\log(r^{-1})$	$\log(\Delta_1^{-1})$	$\log(\Delta_2)$
γ	d^{-5}	ε^6	R^{-2}	r^{-4}	Δ_1^4	Δ_2^{-4}

Table 3.3: dependency of $L, A_1(x), -\log(\kappa), A_2(x), T, \gamma$ on d, ε, R, r, Δ_1 and Δ_2.

Note that under **H**4, U^λ defined in (3.7) is m-strongly convex as well. The following Proposition 3.7 relies on the convergence analysis in Wasserstein distance done in [DM16], which assumes that f is strongly convex. It may be possible to extend the range of validity of these results but this work goes beyond the scope of this paper.

Proposition 3.7. *Assume **H**1 and **H**4. Let $\lambda > 0$, L be the Lipschitz constant of ∇U^λ defined in (3.7) and $\kappa = (2mL)(m+L)^{-1}$. Let $\varepsilon > 0$ and $x \in \mathbb{R}^d$. We have,*

$$W_2(\delta_x R_\gamma^n, \pi^\lambda) \leq \varepsilon \ ,$$

provided that,

$$\gamma \leq \frac{m}{L^2} \left\{ -\frac{13}{12} + \left[\left(\frac{13}{12} \right)^2 + \frac{\varepsilon^2 \kappa^2}{8md} \right]^{1/2} \right\} \wedge \frac{1}{m+L} \ ,$$
$$n \geq 2(\kappa\gamma)^{-1} \left\{ -\log(\varepsilon^2/4) + \log\left(\|x - x^\star\|^2 + d/m \right) \right\} \ .$$

Proof. Assume that $\gamma \in (0, (m+L)^{-1})$. [DM16, Theorem 5] gives for all $n \in \mathbb{N}^\star$:

$$W_2^2(\delta_x R_\gamma^n, \pi^\lambda) \leq 2\left(1 - (\kappa\gamma)/2\right)^n \left\{ \|x - x^\star\|^2 + d/m \right\} + u(\gamma) \ ,$$

Upper bound on n to get $W_1(\delta_{x^\star} R_\gamma^n, \pi) \leq \varepsilon$	$d \to +\infty$	$\varepsilon \to 0$	$R \to +\infty$	$r \to 0$
Proposition 3.5-c) and Proposition 3.7	$\tilde{\mathcal{O}}(d^5)$	$\tilde{\mathcal{O}}(\varepsilon^{-6})$	$\tilde{\mathcal{O}}(R^4)$	$\tilde{\mathcal{O}}(r^{-4})$

Table 3.4: dependency of n on d, ε, R and r to get $W_1(\delta_{x^\star} R_\gamma^n, \pi) \leq \varepsilon$

Upper bound on n to get $W_1(\delta_{x^\star} R_\gamma^n, \pi) \leq \varepsilon$	$\Delta_1 \to 0$	$\Delta_2 \to +\infty$
Proposition 3.5-c) and Proposition 3.7	$\tilde{\mathcal{O}}(\Delta_1^{-4})$	$\tilde{\mathcal{O}}(\Delta_2^4)$

Table 3.5: dependency of n on Δ_1 and Δ_2 to get $W_1(\delta_{x^\star} R_\gamma^n, \pi) \leq \varepsilon$

where,

$$u(\gamma) = 2\kappa^{-1} L^2 d\gamma(\kappa^{-1} + \gamma)\left(2 + \frac{L^2\gamma}{m} + \frac{L^2\gamma^2}{6}\right) .$$

Noting that $\kappa\gamma \leq 1$ and $L^2\gamma^2 \leq 1$, it is then sufficient for γ, n to satisfy,

$$4\kappa^{-2} L^2 d\gamma \left(2 + \frac{1}{6} + \frac{L^2\gamma}{m}\right) \leq \varepsilon^2/2 ,$$
$$2\left(1 - (\kappa\gamma)/2\right)^n \left\{\|x - x^\star\|^2 + d/m\right\} \leq \varepsilon^2/2 ,$$

which concludes the proof. □

Combining Proposition 3.5 and Proposition 3.7 determines the stepsize γ and the number of samples n to get $W_1(\delta_{x^\star} R_\gamma^n, \pi) \leq \varepsilon$. λ is chosen of order $\varepsilon^2 \Delta_1^2 r^2 d^{-2} R^{-2}$ under **H**3-(i) and $\varepsilon^2 r^2 R^{-2} \min(d^{-2}, \Delta_2^{-2})$ under **H**3-(ii). The orders of magnitude of n in $d, \varepsilon, R, r, \Delta_1, \Delta_2$ are reported in Tables 3.4 and 3.5.

3.5 Numerical experiments

In this section we illustrate MYULA with the following three numerical experiments: computation of the volume of a high-dimensional convex set, sampling from a truncated multivariate Gaussian distribution, and Bayesian inference with the constrained LASSO model. We benchmark our results with model-specific specialised algorithms, namely the hit-and-run algorithm [LV06] for set volume computation, the wall HMC (WHMC) [PP14] for truncated Gaussian models, and the auxiliary-variable Gibbs sampler for the Bayesian Lasso model [PC08]. Where relevant we also compare with the Random Walk Metropolis Hastings (RWM) algorithm.

First we consider the computation of the volume of a high-dimensional hypercube. In a manner akin to [CV15a], to apply MYULA to this problem we use an annealing strategy involving truncated Gaussian distributions whose variance is gradually increased at each step $i \in \mathbb{N}$ of the annealing process. Precisely, for $M \in \mathbb{N}^\star$ and $i \in \{0, \ldots, M-1\}$,

the potential U_i (3.2) of the phase i is given for all $x \in \mathbb{R}^d$ by, $U_i(x) = (2\sigma_i^2)^{-1} \|x\|^2 + \iota_{\mathsf{K}}$ where $\mathsf{K} = [-1,1]^d$. Observing that,

$$\frac{\int_{\mathbb{R}^d} \mathrm{e}^{-U_{i+1}(x)} \mathrm{d}x}{\int_{\mathbb{R}^d} \mathrm{e}^{-U_i(x)} \mathrm{d}x} = \pi_i \left(g_i\right) \ , \quad g_i(x) = \mathrm{e}^{2^{-1}(\sigma_i^{-2} - \sigma_{i+1}^{-2})\|x\|^2} \ , \tag{3.28}$$

where π_i is the probability measure associated with U_i, the volume of K is

$$\mathrm{Vol}(\mathsf{K}) = \prod_{i=0}^{M-1} \pi_i(g_i) \int_{\mathbb{R}^d} \mathrm{e}^{-U_0(x)} \ ,$$

where $U_M = \iota_{\mathsf{K}}$. To use MYULA we consider for all $i \in \{0, \ldots, M-1\}$ the potential $U_i^{\lambda_i}$ defined for all $x \in \mathbb{R}^d$ by $U_i^{\lambda_i}(x) = (2\sigma_i^2)^{-1} \|x\|^2 + \iota_{\mathsf{K}}^{\lambda_i}$ where $\iota_{\mathsf{K}}^{\lambda_i}$ is given by (3.4). We choose the step-size γ_i proportional to $1/\{d \max(d, \sigma_i^{-1})\}$ and the regularization parameter λ_i is set equal to $2\gamma_i$. The counterpart of (3.28) is then

$$\frac{\int_{\mathbb{R}^d} \mathrm{e}^{-U_{i+1}^{\lambda_{i+1}}(x)} \mathrm{d}x}{\int_{\mathbb{R}^d} \mathrm{e}^{-U_i^{\lambda_i}(x)} \mathrm{d}x} = \pi_i^{\lambda_i} \left(g_i^{\lambda_i}\right) \ , \quad g_i^{\lambda_i}(x) = \mathrm{e}^{2^{-1}(\sigma_i^{-2} - \sigma_{i+1}^{-2})\|x\|^2 + \iota_{\mathsf{K}}^{\lambda_i} - \iota_{\mathsf{K}}^{\lambda_{i+1}}} \ ,$$

where $\pi_i^{\lambda_i}$ is the probability measure associated with $U_i^{\lambda_i}$, and the volume of K is

$$\mathrm{Vol}(\mathsf{K}) = \prod_{i=0}^{M-1} \pi_i^{\lambda_i}(g_i^{\lambda_i}) \int_{\mathbb{R}^d} \mathrm{e}^{-U_0^{\lambda_0}(x)} \mathrm{d}x \ ,$$

where $U_M^{\lambda_M} = U_M = \iota_{\mathsf{K}}$. We estimate $\pi_i^{\lambda_i}(g_i^{\lambda_i})$ for $i \in \{0, \ldots, M-1\}$ by the empirical averages and we approximate $\int_{\mathbb{R}^d} \mathrm{e}^{-U_0^{\lambda_0}(x)} \mathrm{d}x$ by $(2\pi\sigma_0^2)^{d/2}$.

Figure 3.2 shows the volume estimates (over 10 experiments) obtained with MYULA and the hit-and-run algorithm for a unit hypercube of dimension d ranging from $d = 10$ to $d = 90$ (to simplify visual comparison the estimates are normalised w.r.t. the true volume). Observe that the estimates of MYULA are in agreement with the results of the hit-and-run algorithm, which serves as a benchmark for this problem. The outputs of both algorithms are at similar distances with respect to the true value 1.

The second experiment we consider is the simulation from a d-dimensional truncated Gaussian distribution restricted on a convex set K_d, with mean zero, and covariance matrix Σ with (i,j)th element given by $(\Sigma)_{i,j} = 1/(1 + |i-j|)$. Let $\boldsymbol{\beta} \in \mathbb{R}^d$. The potential U, given by (3.2) and associated with the density $\pi(\boldsymbol{\beta})$, is given by $U(\boldsymbol{\beta}) = (1/2) \langle \boldsymbol{\beta}, \Sigma^{-1}\boldsymbol{\beta} \rangle + \iota_{\mathsf{K}_d}(\boldsymbol{\beta})$. We consider three scenarios of increasing dimension: $d = 2$ with $\mathsf{K}_2 = [0,5] \times [0,1]$, $d = 10$ with $\mathsf{K}_{10} = [0,5] \times [0,0.5]^9$, and $d = 100$ with $\mathsf{K}_{100} = [0,5] \times [0,0.5]^{99}$. We generate 10^6 samples for MYULA, 10^5 samples for WHMC, and 10^6 samples for RWM (in all cases the initial 10% is discarded as burn-in period). Regarding algorithm parameters, we set $\gamma = 1/1000$ and $\lambda = 2\gamma$ for MYULA, and adjust the parameters of RWM and WHMC such that their acceptance rates are approximately 25% and 70%.

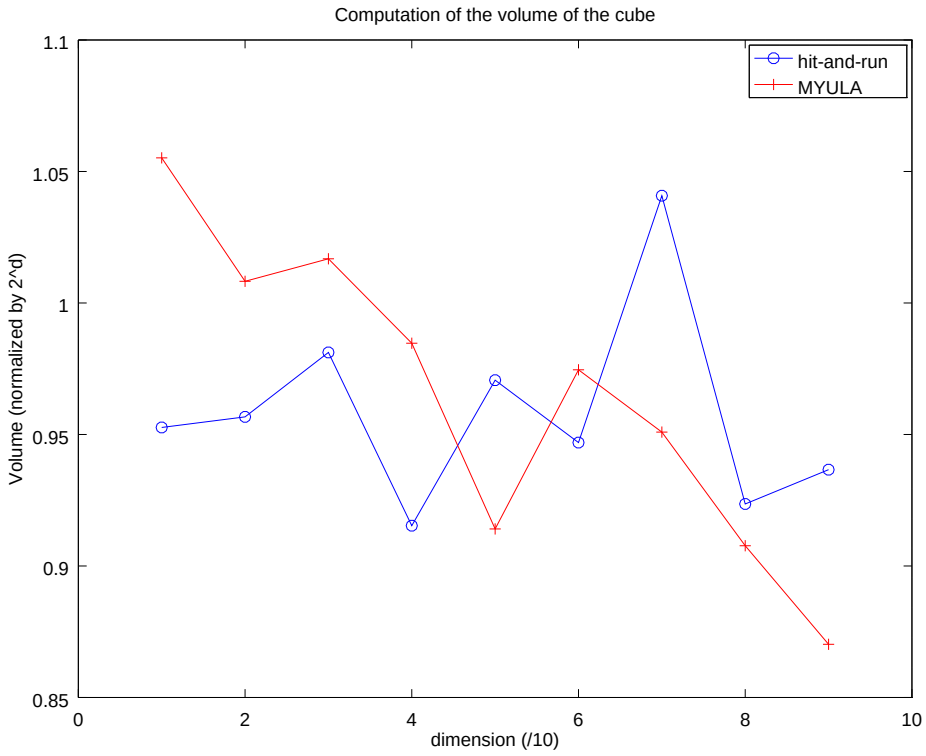

Figure 3.2: Computation of the volume of the cube with MYULA and hit-and-run algorithm.

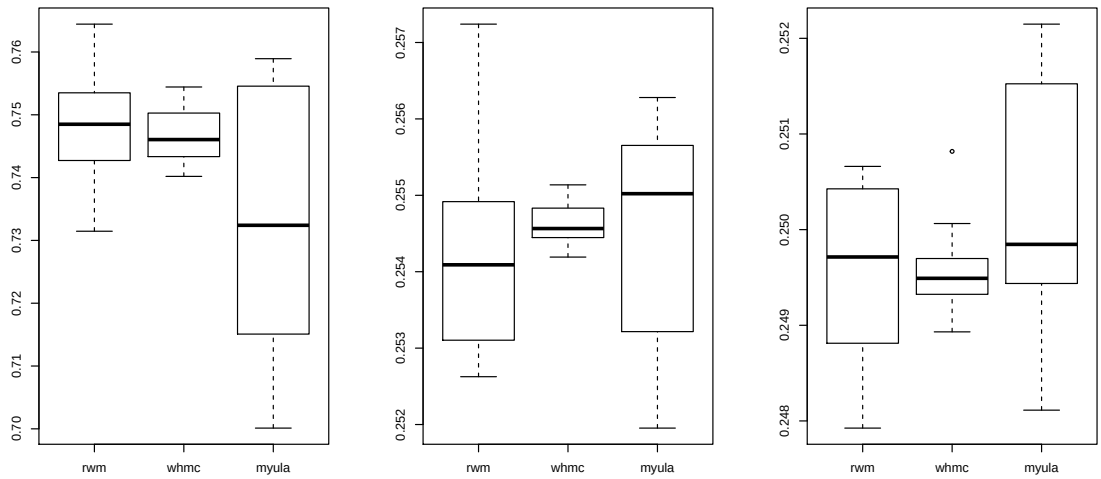

Figure 3.3: Boxplots of $\beta_1, \beta_2, \beta_3$ for the truncated Gaussian variable in dimension 10.

Table 3.6 shows the results obtained with each method for the model $d = 2$, and by performing 100 repetitions to obtain 95% confidence intervals. For this model we also report a solution by a cubature integration [NJ16] which provides a ground truth. Moreover, Figure 3.3 and Figure 3.4 show the results for the first three coordinates of $\boldsymbol{\beta}$ (i.e., $\beta_1, \beta_2, \beta_3$) for $d = 10$ and $d = 100$ respectively. Observe the good performance of MYULA as dimensionality increases, particularly in the challenging case $d = 100$ where it performs comparably to the specialised algorithm WHMC.

Method	Mean	Covariance
Truth	$\begin{bmatrix} 0.790 \\ 0.488 \end{bmatrix}$	$\begin{bmatrix} 0.326 & 0.017 \\ 0.017 & 0.080 \end{bmatrix}$
RWM	$\begin{bmatrix} 0.791 \pm 0.013 \\ 0.486 \pm 0.002 \end{bmatrix}$	$\begin{bmatrix} 0.330 \pm 0.011 & 0.017 \pm 0.002 \\ 0.017 \pm 0.002 & 0.080 \pm 0.0003 \end{bmatrix}$
WHMC	$\begin{bmatrix} 0.789 \pm 0.005 \\ 0.490 \pm 0.005 \end{bmatrix}$	$\begin{bmatrix} 0.324 \pm 0.008 & 0.017 \pm 0.002 \\ 0.017 \pm 0.002 & 0.079 \pm 0.0007 \end{bmatrix}$
MYULA	$\begin{bmatrix} 0.758 \pm 0.052 \\ 0.484 \pm 0.016 \end{bmatrix}$	$\begin{bmatrix} 0.309 \pm 0.038 & 0.017 \pm 0.009 \\ 0.017 \pm 0.009 & 0.088 \pm 0.002 \end{bmatrix}$

Table 3.6: Mean and covariance of $\boldsymbol{\beta}$ in dimension 2 obtained by RWM, WHMC and MYULA.

Finally, we also report an experiment involving the analysis of a real dataset with an ℓ_1-norm constrained Bayesian LASSO model (i.e. least squares regression subject to an ℓ_1-ball constraint). Precisely, the observations $Y = \{Y_1, \dots, Y_n\} \in \mathbb{R}^n$, for $n \geq 1$, are

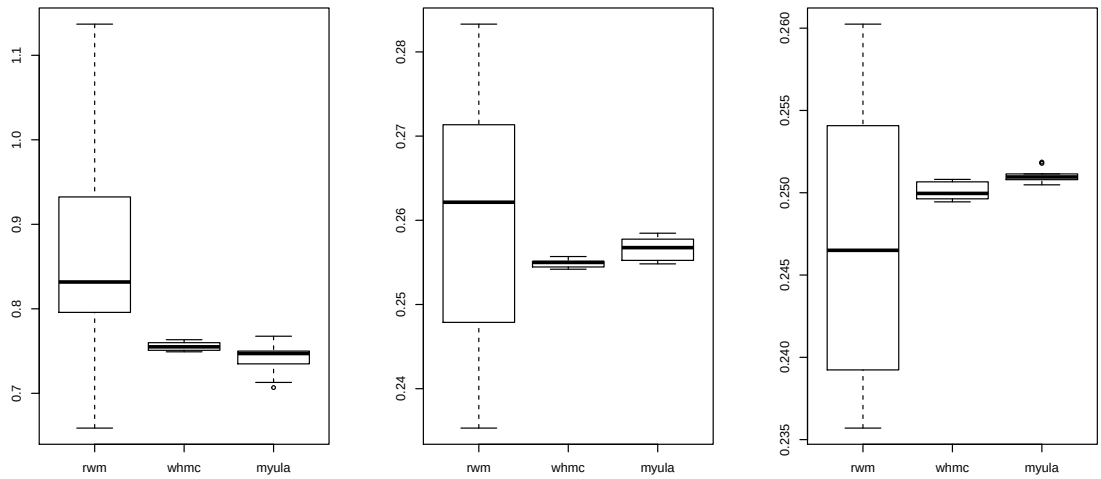

Figure 3.4: Boxplots of $\beta_1, \beta_2, \beta_3$ for the truncated Gaussian variable in dimension 100.

assumed to be distributed from the Gaussian distribution with mean $X\boldsymbol{\beta}$ and covariance matrix $\sigma^2 \mathrm{I}_n$, where $X \in \mathbb{R}^{n\times d}$ is the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^d$ is the regression parameter, $\sigma^2 > 0$ and I_n is the identity matrix of dimension n. The prior on $\boldsymbol{\beta}$ is the uniform distribution over the ℓ_1 ball, $\mathrm{B}(0,s) = \{\boldsymbol{\beta} \in \mathbb{R}^d \quad \|\boldsymbol{\beta}\|_1 \leq s\}$, for $s > 0$, where $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^d |\boldsymbol{\beta}_i|$, $\boldsymbol{\beta}_i$ is the i-th component of $\boldsymbol{\beta}$. The potential U^s, for $s > 0$, associated with the posterior distribution is given for all $\boldsymbol{\beta} \in \mathbb{R}^d$ by $U^s(\boldsymbol{\beta}) = \|Y - X\boldsymbol{\beta}\|^2 + \iota_{\mathrm{B}(0,s)}(\boldsymbol{\beta})$. We consider in our experiment the diabetes data set[1], which consists in $n = 442$ observations and $d = 10$ explanatory variables.

Figure 3.5 shows the "LASSO paths" obtained using MYULA, the WHMC algorithm, and with the specialised Gibbs sampler of [PC08] (these paths are the posterior marginal medians associated with π^s for $s = t\left\|\boldsymbol{\beta}^{\mathrm{OLS}}\right\|_1$, $t \in [0,1]$, and where $\boldsymbol{\beta}^{\mathrm{OLS}}$ is the estimate obtained by the ordinary least square regression). The dot lines represent the confidence interval at level 95%, obtained by performing 100 repetitions. MYULA estimates were obtained by using 10^5 samples (with the initial 10^4 samples discarded as burn-in period) and stepsize $s^{3/2} \times 10^{-5}$. WHMC estimates were obtained by using 10^4 samples (with the initial 10^3 samples discarded as burn-in period), and by adjusting parameters to achieve an acceptance rate of approximately 90%. Finally, the Gibbs sampler is targeting an unconstrained LASSO model with prior $\boldsymbol{\beta} \mapsto (2s)^{-d}\mathrm{e}^{-\|\boldsymbol{\beta}\|_1/s}$, for $s > 0$. The results are comparable for the three algorithms.

[1] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

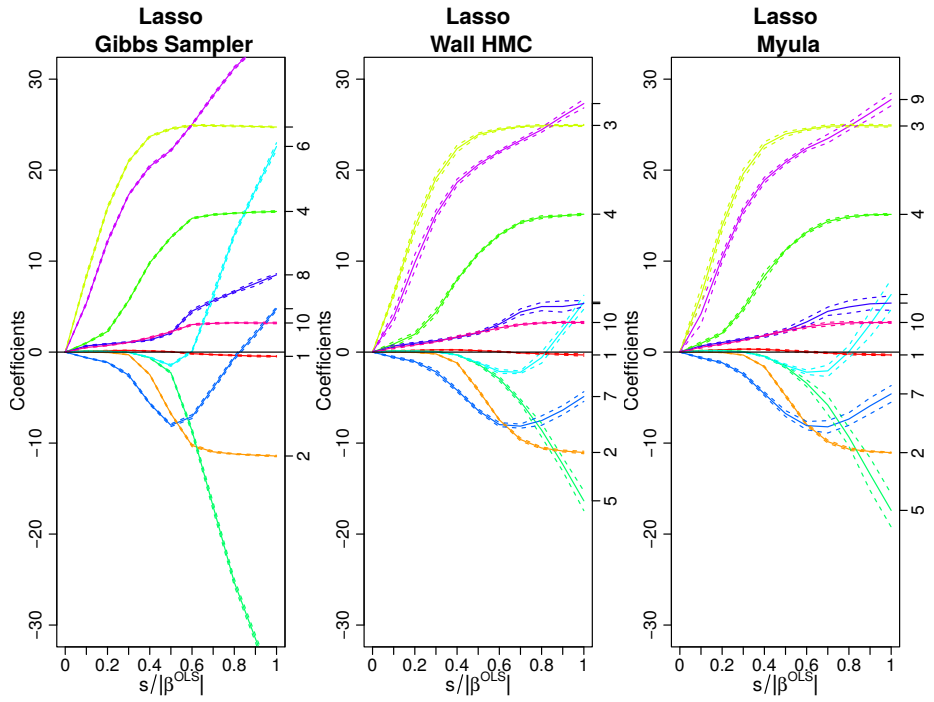

Figure 3.5: Lasso path for the Gibbs sampler, Wall HMC and MYULA algorithms.

3.6 Proofs

3.6.1 Proof of Lemma 3.1

Since f is a (proper) convex function, there exist $a \in \mathbb{R}$, $b \in \mathbb{R}^d$ such that $f(x) \geq a + \langle b, x \rangle$ [Roc15, Theorem 23.4]. By **H**2 and a straightforward calculation, for $\|x\| \geq R + 4\lambda \|b\| + 2\{\lambda(|a| + R\|b\|)\}^{1/2}$, we have,

$$U^{\lambda}(x) \geq (4\lambda)^{-1}(\|x\| - R)^2 ,$$

which concludes the proof.

3.6.2 Proof of Lemma 3.3

Under **H**2, $0 \in \widetilde{\mathsf{K}}$. Let $x_1, x_2 \in \widetilde{\mathsf{K}}$ and $t \in [0,1]$. We have by definition of $\widetilde{\mathsf{K}}$ (3.18) that $\mathrm{B}(tx_1 + (1-t)x_2, r) \subset t\,\mathrm{B}(x_1, r) + (1-t)\,\mathrm{B}(x_2, r) \subset \mathsf{K}$, which implies that $\widetilde{\mathsf{K}}$ is convex.

To show that $\widetilde{\mathsf{K}}$ is closed, it is enough to show that $\widetilde{\mathsf{K}} = \{x \in \mathsf{K} \mid \mathrm{dist}(x, \mathsf{K}^c) \geq r\}$ where $\mathrm{dist}(x, \mathsf{K}^c) = \inf_{y \in \mathsf{K}^c} \|x - y\|$ since $x \mapsto \mathrm{dist}(x, \mathsf{K}^c)$ is Lipschitz continuous. First by definition, we have $\widetilde{\mathsf{K}} \subset \{x \in \mathsf{K} \mid \mathrm{dist}(x, \mathsf{K}^c) \geq r\}$. To show the converse, let $x \in \{y \in \mathsf{K} \mid \mathrm{dist}(y, \mathsf{K}^c) \geq r\}$. Then, $\mathrm{B}_{\mathrm{o}}(x, r) \subset \mathsf{K}$, where $\mathrm{B}_{\mathrm{o}}(x, r) = \left\{y \in \mathbb{R}^d \mid \|y - x\| < r\right\}$, which yields $\mathrm{B}(x, r) \subset \mathsf{K}$ since K is assumed to be closed. This result then concludes the proof by definition of $\widetilde{\mathsf{K}}$.

3.6.3 Proof of Proposition 3.4

a) By a direct calculation, we have:

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} = \int_{\mathbb{R}^d} \left|\pi(x) - \pi^\lambda(x)\right| \mathrm{d}x = 2\left(1 + \left\{\int_{\mathsf{K}^c} \mathrm{e}^{-U^\lambda(x)}\mathrm{d}x\right\}^{-1} \int_{\mathsf{K}} \mathrm{e}^{-f(x)}\mathrm{d}x\right)^{-1} \tag{3.29}$$

$$\leq 2\left(1 + \exp\left(\min_{\mathsf{K}^c}(f) - \max_{\mathsf{K}}(f)\right) A\right)^{-1} . \tag{3.30}$$

where

$$A = \mathrm{Vol}(\mathsf{K}) \Big/ \int_{\mathsf{K}^c} \mathrm{e}^{-(2\lambda)^{-1}\|x - \mathrm{proj}_{\mathsf{K}}(x)\|^2} \mathrm{d}x \ . \tag{3.31}$$

The conclusion follows then from (3.17) and **H**3-(i).

We give two proofs for (3.21), which both consist in lower bounding A. The obtained bounds are identical up to an universal constant. The first one is simpler and was suggested by a referee. The second one is more involved ; however, it has the benefit of establishing the relation between the intrinsic volumes of K and the bound on the total variation norm.

Under **H**2, we have $\mathsf{K} + \mathrm{B}(0,t) \subset (1 + t/r)\mathsf{K}$ and using (3.14),

$$\begin{aligned}
\int_{\mathsf{K}^c} \mathrm{e}^{-(1/2\lambda)\|x - \mathrm{proj}_{\mathsf{K}}(x)\|^2} \mathrm{d}x &\leq \left\{\int_{\mathbb{R}_+} \mathrm{Vol}(\mathsf{K}(1 + t/r))\lambda^{-1} t \mathrm{e}^{-t^2/(2\lambda)} \mathrm{d}t - \mathrm{Vol}(\mathsf{K})\right\} \\
&= \mathrm{Vol}(\mathsf{K}) \left\{\int_{\mathbb{R}_+} (1 + t/r)^d \lambda^{-1} t \mathrm{e}^{-t^2/(2\lambda)} \mathrm{d}t - 1\right\} \\
&= \mathrm{Vol}(\mathsf{K}) \sum_{i=1}^{d} \binom{d}{i} \left(\frac{\sqrt{2\lambda}}{r}\right)^i \mathbf{\Gamma}(1 + i/2) \\
&\leq \mathrm{Vol}(\mathsf{K}) \sum_{i=1}^{d} \left(\frac{\sqrt{2\lambda} d}{r}\right)^i ,
\end{aligned}$$

where the second equality follows from developping $(1 + t/r)^d$, making the change of variable $t \mapsto t^2/(2\lambda)$ and using the Gamma function and the last inequality from $\binom{d}{i}\mathbf{\Gamma}(1 + i/2) \leq d^i$ for $i \in \{1, \ldots, d\}$. For $\lambda \in (0, r^2 d^{-2}/8]$, we get

$$A^{-1} \leq \sum_{i=1}^{d} \left(\frac{\sqrt{2\lambda} d}{r}\right)^i \leq \frac{2\sqrt{2\lambda} d}{r} .$$

Combining it with (3.30) and **H**3-(i) concludes the proof.

For the second proof, it is necessary to introduce first a generalized notion of the intrinsic volumes (3.15), the mixed volumes. Let $\mathcal{K}$ be the class of convex bodies of $\mathbb{R}^d$,

$\mathsf{K}_1, \ldots, \mathsf{K}_m \in \mathcal{K}$ and $\lambda_1, \ldots, \lambda_m \geq 0$. By [Sch13, Theorem 5.1.7], there is a nonnegative symmetric function $\mathcal{V} : (\mathcal{K})^d \to \mathbb{R}_+$, the mixed volume, such that,

$$\mathrm{Vol}(\lambda_1 \mathsf{K}_1 + \ldots + \lambda_m \mathsf{K}_m) = \sum_{i_1, \ldots, i_d = 1}^{m} \lambda_{i_1} \ldots \lambda_{i_d} \mathcal{V}(\mathsf{K}_{i_1}, \ldots, \mathsf{K}_{i_d}) \,. \tag{3.32}$$

Let $m > 1$, $a_1, \ldots, a_m \geq 0$ and $\mathsf{K}_1, \ldots, \mathsf{K}_m, \mathsf{L}$ be $(m+1)$ convex bodies in $\mathbb{R}^d$ such that $\mathsf{K}_1 \subset \mathsf{L}$. By unicity of the coefficients of the polynomial in $\lambda_1, \ldots, \lambda_m$ (3.32) and [Sch13, p.282], we have:

$$\mathcal{V}(a_1 \mathsf{K}_1, \ldots, a_m \mathsf{K}_m) = \left(\prod_{i=1}^{m} a_i \right) \mathcal{V}(\mathsf{K}_1, \ldots, \mathsf{K}_m) \,, \tag{3.33}$$

$$\mathcal{V}(\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_m) \leq \mathcal{V}(\mathsf{L}, \mathsf{K}_2, \ldots, \mathsf{K}_m) \,. \tag{3.34}$$

Denote by B the unity ball of $\mathbb{R}^d$, $\mathsf{B} = \mathrm{B}(0,1)$. Taking $m = 2, \mathsf{K}_1 = \mathsf{K}, \mathsf{K}_2 = \mathsf{B}, \lambda_1 = 1, \lambda_2 = t$ in (3.32), we get:

$$\mathrm{Vol}(\mathsf{K} + \mathrm{B}(0,t)) = \sum_{i=0}^{d} t^i \binom{d}{i} \mathcal{V}(\mathsf{K}[d-i], \mathsf{B}[i]) \,, \tag{3.35}$$

where for a set $A \subset \mathbb{R}^d$, the notation $A[i]$ means A repeated i times: $A[i] = A, \ldots, A$ i times. The quermassintegrals of K are defined for $i \in \{0, \ldots, d\}$ by $\mathscr{W}_i(\mathsf{K}) = \mathcal{V}(\mathsf{K}[d-i], \mathsf{B}[i])$ [Sch13, equation 5.31]. We get then by (3.35) and (3.15),

$$\binom{d}{i} \mathscr{W}_i(\mathsf{K}) = \kappa_i \mathscr{V}_{d-i}(\mathsf{K}) \,, \tag{3.36}$$

where κ_i is given by (3.16).

The proof consists then in identifying an upper bound on $\mathscr{V}_i(\mathsf{K})(\mathrm{Vol}\,\mathsf{K})^{-1}$ for $i \in \{0, \ldots, d\}$. First, the sequence $\{i! \mathscr{V}_i(\mathsf{K})\}_{0 \leq i \leq d}$ is shown to be log-concave, i.e. for $i \in \{1, \ldots, d-1\}$

$$(i! \mathscr{V}_i(\mathsf{K}))^2 \geq (i+1)! \mathscr{V}_{i+1}(\mathsf{K}) (i-1)! \mathscr{V}_{i-1}(\mathsf{K}) \,. \tag{3.37}$$

The Aleksandrov-Fenchel inequality [Sch13, equation 7.66] states, for $i \in \{1, \ldots, d-1\}$,

$$\mathscr{W}_i(\mathsf{K})^2 \geq \mathscr{W}_{i-1}(\mathsf{K}) \mathscr{W}_{i+1}(\mathsf{K}) \,. \tag{3.38}$$

By (3.16), $\kappa_i / \kappa_{i-2} = (2\pi)/i$ and the log convexity of the gamma function, we get for $i \in \{1, \ldots, d-1\}$:

$$\frac{1}{i+1} \frac{\kappa_i}{\kappa_{i+1}} = \frac{1}{i} \frac{\kappa_{i-2}}{\kappa_{i-1}} \leq \frac{1}{i} \frac{\kappa_{i-1}}{\kappa_i} \,. \tag{3.39}$$

Combining (3.39), (3.38) and (3.36) shows (3.37).

The log-concavity of $\{i! \mathscr{V}_i(\mathsf{K})\}_{0 \leq i \leq d}$ gives for $i \in \{0, \ldots, d-1\}$,

$$\frac{\mathscr{V}_i(\mathsf{K})}{\mathscr{V}_{i+1}(\mathsf{K})} \leq \frac{\mathscr{V}_{d-1}(\mathsf{K})}{\mathrm{Vol}(\mathsf{K})} = \frac{d}{2} \frac{\mathscr{W}_1(\mathsf{K})}{\mathscr{W}_0(\mathsf{K})} \,. \tag{3.40}$$

Combining the definition of the quermassintegrals, (3.33), (3.34) and **H**2 give:

$$r\mathscr{W}_1(\mathsf{K}) = \mathcal{V}(\mathsf{K},\dots,\mathsf{K},\mathrm{B}(0,r)) \le \mathcal{V}(\mathsf{K},\dots,\mathsf{K},\mathsf{K}) = \mathscr{W}_0(\mathsf{K}) \; . \tag{3.41}$$

By (3.41) and (3.40), we get:

$$\mathsf{D}(\mathsf{K},\lambda) \le \sum_{i=1}^{d} \left\{ dr^{-1}(\pi\lambda/2)^{1/2} \right\}^i \; , \tag{3.42}$$

where $\mathsf{D}(\mathsf{K},\lambda)$ is defined in (3.20). For all $\lambda \in (0, 2\pi^{-1}(r/d)^2)$, (3.19) gives then,

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \le 2\left\{1 + \exp\left(\min_{\mathsf{K}^c}(f) - \max_{\mathsf{K}}(f)\right)\left(\left\{dr^{-1}(\pi\lambda/2)^{1/2}\right\}^{-1} - 1\right)\right\}^{-1} \; .$$

Using that for all $a, b \in \mathbb{R}_+^*$, $b \ge 2$, $(1 + a(b-1))^{-1} \le b^{-1}/(b^{-1} + a/2)$ and **H**3-(i), we get for $\lambda \in \left(0, 2\pi^{-1}(r/d)^2\right)$

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \le 2^{3/2}(\pi\lambda)^{1/2} dr^{-1}\left\{(2\pi\lambda)^{1/2} dr^{-1} + \Delta_1\right\}^{-1} \; .$$

b) The proof consists in using (3.29) to bound $\|\pi^\lambda - \pi\|_{\mathrm{TV}}$. In the first step we give an upper bound on $\int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(x)}\mathrm{d}x / \int_{\mathsf{K}} \mathrm{e}^{-f(x)}\mathrm{d}x$. By Fubini's theorem, similarly to (3.14) we have

$$\int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(x)}\mathrm{d}x \le \int_{\mathbb{R}_+}\int_{\mathsf{K}+\mathrm{B}(0,t)} \mathrm{e}^{-f(x)}\lambda^{-1}t\mathrm{e}^{-t^2/(2\lambda)}\mathrm{d}x\mathrm{d}t \; . \tag{3.43}$$

Let $t \ge 0$. By definition of $\widetilde{\mathsf{K}}$, using Lemma 3.3 and $\mathsf{K} - \mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}}) + \mathrm{B}(0,t) \subset (1 + t/r)(\mathsf{K} - \mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}}))$, we have

$$\begin{aligned}
\int_{\mathsf{K}+\mathrm{B}(0,t)} \mathrm{e}^{-f(x)}\mathrm{d}x &= \int_{\mathsf{K}-\mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}})+\mathrm{B}(0,t)} \mathrm{e}^{-f(x+\mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}}))}\mathrm{d}x \\
&\le \int_{(1+t/r)(\mathsf{K}-\mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}}))} \mathrm{e}^{-f(x+\mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}}))}\mathrm{d}x \\
&= (1+t/r)^d \int_{\mathsf{K}-\mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}})} \mathrm{e}^{-f((1+t/r)x+\mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}}))}\mathrm{d}x \; .
\end{aligned} \tag{3.44}$$

By **H**1-(i) f is convex and therefore for all $x \in \mathsf{K} - \mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}})$,

$$\begin{aligned}
f((1+t/r)x + \mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}})) &\ge (t/r)\left\{f(x + \mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}})) - f(\mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}}))\right\} + f(x + \mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}})) \\
&\ge -(\Delta_2 t)/r + f(x + \mathrm{proj}_{\widetilde{\mathsf{K}}}(x_{\mathsf{K}})) \; .
\end{aligned}$$

Combining it with (3.43) and (3.44), we get

$$\int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(x)}\mathrm{d}x \le \left(\int_{\mathsf{K}} \mathrm{e}^{-f(x)}\mathrm{d}x\right)\int_{\mathbb{R}_+} (1+t/r)^d \mathrm{e}^{(\Delta_2 t)/r}\lambda^{-1}t\mathrm{e}^{-t^2/(2\lambda)}\mathrm{d}t \; . \tag{3.45}$$

We now bound $B = \int_{\mathsf{K}^c} \mathrm{e}^{-U^\lambda(x)}\mathrm{d}x / \int_{\mathsf{K}} \mathrm{e}^{-f(x)}\mathrm{d}x$. Using (3.45) and an integration by parts, we have

$$\begin{aligned} B &\leq \int_{\mathbb{R}_+} \left\{(1+t/r)^d \mathrm{e}^{(\Delta_2 t)/r} - 1\right\} \lambda^{-1} t \mathrm{e}^{-t^2/(2\lambda)}\mathrm{d}t \\ &\leq \int_{\mathbb{R}_+} (1+t/r)^{d-1} \mathrm{e}^{(\Delta_2 t)/r} r^{-1} \left(d + \Delta_2 + (\Delta_2 t)/r\right) \mathrm{e}^{-t^2/(2\lambda)}\mathrm{d}t \; . \end{aligned}$$

Since for all $t \geq 0$, $(\Delta_2 t)/r - t^2/(2\lambda) \leq -t^2/(4\lambda) + 4\lambda(\Delta_2/r)^2$, it holds

$$B \leq \frac{1}{r} \exp\left(4\lambda \left(\frac{\Delta_2}{r}\right)^2\right) \int_{\mathbb{R}_+} (1+t/r)^{d-1} \left(d + \Delta_2 + (\Delta_2 t)/r\right) \mathrm{e}^{-t^2/(4\lambda)}\mathrm{d}t \; .$$

By developping $(1+t/r)^{d-1}$, using the change of variable $t \mapsto t^2/(4\lambda)$ and the definition of the Gamma function, we have

$$B \leq \frac{2\lambda}{r} \exp\left(4\lambda \left(\frac{\Delta_2}{r}\right)^2\right) \sum_{i=0}^{d-1} \binom{d-1}{i} \left(\frac{2\sqrt{\lambda}}{r}\right)^i \left\{\frac{d+\Delta_2}{2\sqrt{\lambda}} \mathbf{\Gamma}\left(\frac{1+i}{2}\right) + \frac{\Delta_2}{r} \mathbf{\Gamma}\left(1+\frac{i}{2}\right)\right\} \; .$$

Using that for all $i \in \{0, \dots, d-1\}$, $\binom{d-1}{i} \mathbf{\Gamma}(1+i/2) \leq d^i$, we get for $\lambda \in (0, 16^{-1} r^2 d^{-2}]$

$$B \leq \frac{2}{r} \exp\left(4\lambda \left(\frac{\Delta_2}{r}\right)^2\right) \left\{\sqrt{\lambda}(d+\Delta_2) + \frac{2\lambda\Delta_2}{r}\right\} \; ,$$

which combined with (3.29) concludes the proof.

3.6.4 Proof of Proposition 3.5

a) The proof relies on a control of the Wasserstein distance by a weighted total variation. The arguments are similar to those of Proposition 3.4. [Vil09, Theorem 6.15] implies:

$$W_1(\pi, \pi^\lambda) \leq \int_{\mathbb{R}^d} \|x\| \, |\pi(x) - \pi^\lambda(x)| \mathrm{d}x = C + D \; , \tag{3.46}$$

where

$$C = \int_{\mathsf{K}^c} \|x\| \, \pi^\lambda(x)\mathrm{d}x \; , \quad D = \left\{1 - \frac{\int_{\mathsf{K}} \mathrm{e}^{-f}}{\int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda}}\right\} \int_{\mathsf{K}} \|x\| \, \pi(x)\mathrm{d}x \; . \tag{3.47}$$

We bound these two terms separately. First using the same decomposition as in (3.14), $\|x\| \leq R + \|x - \mathrm{proj}_{\mathsf{K}}(x)\|$ and that for all $t \in \mathbb{R}_+$, $\mathsf{K} + \mathrm{B}(0,t) = \{x \in \mathbb{R}^d \, : \, \|x - \mathrm{proj}_{\mathsf{K}}(x)\| \leq t\}$, we get

$$C = \left(\int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda}\right)^{-1} \int_0^{+\infty} \int_{\mathsf{K}^c} \mathrm{e}^{-f(x)} \, \|x\| \, t\lambda^{-1} \mathrm{e}^{-t^2/(2\lambda)} \mathbb{1}_{[\|x - \mathrm{proj}_{\mathsf{K}}(x)\|, +\infty)}(t) \, \mathrm{d}x \, \mathrm{d}t \tag{3.48}$$

$$\leq \mathrm{e}^{\max_{\mathsf{K}}(f) - \min_{\mathsf{K}^c}(f)} \int_0^{+\infty} (R+t)\lambda^{-1} \mathrm{e}^{-t^2/(2\lambda)} \left(\frac{\mathrm{Vol}(\mathsf{K} + \mathrm{B}(0,t)) - \mathrm{Vol}(\mathsf{K})}{\mathrm{Vol}(\mathsf{K})}\right) \mathrm{d}t \; . \tag{3.49}$$

Combining (3.15)-(3.49), **H**3-(i) and using $\mathscr{V}_d(\mathsf{K}) = \mathrm{Vol}(\mathsf{K})$ give

$$C \leq \Delta_1^{-1} \sum_{i=0}^{d-1} \kappa_{d-i} \frac{\mathscr{V}_i(\mathsf{K})}{\mathrm{Vol}(\mathsf{K})} \int_0^{+\infty} (Rt^{d-i+1} + t^{d-i+2}) \lambda^{-1} \mathrm{e}^{-t^2/(2\lambda)} \mathrm{d}t \; . \tag{3.50}$$

Using (3.16), for all $k \geq 0$, $\int_{\mathbb{R}_+} t^k \mathrm{e}^{t^2/(2\lambda)} \mathrm{d}t = (2\lambda)^{(k+1)/2} \mathbf{\Gamma}((k+1)/2)$ and for all $a > 1$, $\mathbf{\Gamma}(a + 1/2) \leq a^{1/2} \mathbf{\Gamma}(a)$ (by log-convexity of the Gamma function), we have

$$C \leq \Delta_1^{-1} \sum_{i=0}^{d-1} \frac{\mathscr{V}_i(\mathsf{K})}{\mathrm{Vol}(\mathsf{K})} (2\pi\lambda)^{(d-i)/2} \left\{ R + [\lambda(d-i+2)]^{1/2} \right\} \; . \tag{3.51}$$

Regarding D defined in (3.47), by **H**2, **H**3-(i), (3.30) and (3.17), we get:

$$D \leq R \Delta_1^{-1} \mathsf{D}(\mathsf{K}, \lambda) \; , \tag{3.52}$$

where $\mathsf{D}(\mathsf{K}, \lambda)$ is defined in (3.20). Combining (3.51) and (3.52) in (3.46) concludes the proof.

b) Using (3.40) and (3.41) in (3.51) gives for all $\lambda \in (0, (2\pi)^{-1} r^2 d^{-2})$

$$\begin{aligned} C &\leq \Delta_1^{-1} \sum_{i=0}^{d-1} \left(\frac{d}{r} \sqrt{\frac{\pi\lambda}{2}} \right)^{d-i} \left\{ R + [\lambda(d-i+2)]^{1/2} \right\} \\ &\leq \Delta_1^{-1} (2\pi\lambda)^{1/2} d r^{-1} \left(R + r \left(\frac{3}{2d\pi} \right)^{1/2} \right) \; . \end{aligned}$$

Finally this bound, (3.52), (3.42) and (3.46) conclude the proof.

c) The proof still relies on the decomposition (3.46), where C and D are defined in (3.47). Eq. (3.48) gives

$$C \leq \int_0^{+\infty} (R+t) t \lambda^{-1} \mathrm{e}^{-t^2/(2\lambda)} \left(\frac{\int_{\mathsf{K}+\mathrm{B}(0,t)} \mathrm{e}^{-f(x)} \mathrm{d}x}{\int_{\mathsf{K}} \mathrm{e}^{-f(x)} \mathrm{d}x} - 1 \right) \mathrm{d}t \; .$$

Under **H**3-(ii), following the steps of Section 3.6.3-b) to upper bound the term

$$\int_{\mathsf{K}+\mathrm{B}(0,t)} \mathrm{e}^{-f(x)} \mathrm{d}x \Big/ \int_{\mathsf{K}} \mathrm{e}^{-f(x)} \mathrm{d}x \; ,$$

we have

$$\begin{aligned} C &\leq \int_0^{+\infty} (R+t) t \lambda^{-1} \mathrm{e}^{-t^2/(2\lambda)} \left((1 + t/r)^d \mathrm{e}^{(t\Delta_2)/r} - 1 \right) \mathrm{d}t \\ &= C_1 + C_2 \; , \end{aligned}$$

where

$$C_1 = R \int_0^{+\infty} t\lambda^{-1} \mathrm{e}^{-t^2/(2\lambda)} \left((1+t/r)^d \mathrm{e}^{(t\Delta_2)/r} - 1 \right) \mathrm{d}t \ ,$$
$$C_2 = \int_0^{+\infty} t^2\lambda^{-1} \mathrm{e}^{-t^2/(2\lambda)} \left((1+t/r)^d \mathrm{e}^{(t\Delta_2)/r} - 1 \right) \mathrm{d}t \ .$$

C_1 is upper bounded in the same way as B in Section 3.6.3-b). Regarding C_2, since for all $t \geq 0$, $(\Delta_2 t)/r - t^2/(2\lambda) \leq -t^2/(4\lambda) + 4\lambda(\Delta_2/r)^2$, developping $(1+t/r)^d$ and using the change of variable $t \mapsto t^2/(4\lambda)$ we get

$$\begin{aligned} C_2 &\leq \mathrm{e}^{4\lambda(\Delta_2/r)^2} \sum_{i=0}^{d} \binom{d}{i} r^{-i} \int_{\mathbb{R}_+} t^{i+2} \lambda^{-1} \mathrm{e}^{-t^2/(4\lambda)} \mathrm{d}t \\ &\leq 4\sqrt{\lambda} \mathrm{e}^{4\lambda(\Delta_2/r)^2} \frac{\sqrt{\pi}}{2} \sum_{i=0}^{d} \binom{d}{i} \left(\frac{2\sqrt{\lambda}}{r} \right)^i \mathbf{\Gamma}\left(\frac{3}{2} + \frac{i}{2} \right) \ . \end{aligned}$$

Using $\binom{d}{i}\mathbf{\Gamma}((3+i)/2) \leq (\sqrt{\pi}/2)d^i$ for $i \in \{0, \ldots, d\}$, we have for $\lambda \in (0, 16^{-1}r^2 d^{-2}]$,

$$\begin{aligned} C_2 &\leq 2\sqrt{\pi\lambda} \mathrm{e}^{4\lambda(\Delta_2/r)^2} \sum_{i=0}^{d} \left(\frac{2\sqrt{\lambda} d}{r} \right)^i \\ &\leq 4\sqrt{\pi\lambda} \mathrm{e}^{4\lambda(\Delta_2/r)^2} \ . \end{aligned}$$

D defined in (3.47) is upper bounded by RB where B is defined in Section 3.6.3-b). Combining the bounds on C_1, C_2, D gives the result.

Acknowledgments

The authors wish to express their thanks to the anonymous referees for several helpful remarks, in particular concerning a simplified proof of Proposition 3.4.

Chapter 4

The Tamed Unadjusted Langevin Algorithm

Nicolas Brosse [1], Alain Durmus [2], Éric Moulines [1] and Sotirios Sabanis [3]

Abstract

In this article, we consider the problem of sampling from a probability measure π having a density on $\mathbb{R}^d$ proportional to $x \mapsto \mathrm{e}^{-U(x)}$. The Euler discretization of the Langevin stochastic differential equation (SDE) is known to be unstable, when the potential U is superlinear. Based on previous works on the taming of superlinear drift coefficients for SDEs, we introduce the Tamed Unadjusted Langevin Algorithm (TULA) and obtain non-asymptotic bounds in V-total variation norm and Wasserstein distance of order 2 between the iterates of TULA and π, as well as weak error bounds. Numerical experiments are presented which support our findings.

4.1 Introduction

The Unadjusted Langevin Algorithm (ULA) first introduced in the physics literature by [Par81] and popularized in the computational statistics community by [Gre83] and [GM94] is a technique to sample complex and high-dimensional probability distributions. This issue has far-reaching consequences in Bayesian statistics and machine learning [And+03], [Cot+13], aggregation of estimators [DT12] and molecular dynamics [LS16]. More precisely, let π be a probability distribution on $\mathbb{R}^d$ which has density (also denoted

[1]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France.
Emails: nicolas.brosse@polytechnique.edu, eric.moulines@polytechnique.edu

[2]Ecole Normale Supérieure CMLA 61, Av. du Président Wilson 94235 Cachan Cedex, France
Email: alain.durmus@cmla.ens-cachan.fr

[3]University of Edinburgh, Scotland, UK. Email: s.sabanis@ed.ac.uk

by π) with respect to the Lebesgue measure given for all $x \in \mathbb{R}^d$ by,

$$\pi(x) = \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y \ , \quad \text{with} \quad \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y < +\infty \ .$$

Assuming that $U : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, the overdamped Langevin stochastic differential equation (SDE) associated with π is given by

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ , \tag{4.1}$$

where $(B_t)_{t \geq 0}$ is a d-dimensional Brownian motion. The discrete time Markov chain associated with the ULA algorithm is obtained by the Euler-Maruyama discretization scheme of the Langevin SDE defined for $k \in \mathbb{N}$ by,

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1} \ , \ X_0 = x_0 \ , \tag{4.2}$$

where $x_0 \in \mathbb{R}^d$, $\gamma > 0$ and $(Z_k)_{k \in \mathbb{N}}$ are i.i.d. standard d-dimensional Gaussian variables. Under adequate assumptions on a globally Lipschitz ∇U, non-asymptotic bounds in total variation and Wasserstein distances between the distribution of $(X_k)_{k \in \mathbb{N}}$ and π can be found in [Dal17b], [DM17], [DM16]. However, the ULA algorithm is unstable if ∇U is superlinear i.e. $\liminf_{\|x\| \to +\infty} \|\nabla U(x)\| / \|x\| = +\infty$, see [RT96, Theorem 3.2], [MSH02] and [HJK11]. This is illustrated with a particular example in [MSH02, Lemma 6.3] where, the SDE (5.7) is considered in one dimension with $U(x) = x^4/4$ along with the associated Euler discretization (4.2) and it is shown that for all $\gamma > 0$, if $\mathbb{E}\left[X_0^2\right] \geq 2/\gamma$, one obtains $\lim_{n \to +\infty} \mathbb{E}\left[X_n^2\right] = +\infty$. Moreover, the sample path $(X_n)_{n \in \mathbb{N}}$ diverges to infinity with positive probability.

Until recently, either implicit numerical schemes, e.g. see [MSH02] and [HMS02], or adaptive stepsize schemes, e.g. see [LMS07], were used to address this problem. However, in the last few years, a new generation of explicit numerical schemes, which are computationally efficient, has been introduced by "taming" appropriately the superlinearly growing drift, see [HJK12] and [Sab13] for more details.

Nonetheless, with the exception of [MSH02], these works focus on the discretization of SDEs with superlinear coefficients in finite time. We aim at extending these techniques to sample from π, the invariant measure of (5.7). To deal with the superlinear nature of ∇U, we introduce a family of drift functions $(G_\gamma)_{\gamma > 0}$ with $G_\gamma : \mathbb{R}^d \to \mathbb{R}^d$ indexed by the step size γ which are close approximations of ∇U in a sense made precise below. Consider then the following Markov chain $(X_k)_{k \in \mathbb{N}}$ defined for all $k \in \mathbb{N}$ by

$$X_{k+1} = X_k - \gamma G_\gamma(X_k) + \sqrt{2\gamma} Z_{k+1} \ , \ X_0 = x_0 \ . \tag{4.3}$$

We suggest two different explicit choices for the family $(G_\gamma)_{\gamma > 0}$ based on previous studies on the tamed Euler scheme [HJK12], [Sab13], [HJ15]. Define for all $\gamma > 0$, $H_\gamma, H_{\gamma,c} : \mathbb{R}^d \to \mathbb{R}^d$ for all $x \in \mathbb{R}^d$ by

$$H_\gamma(x) = \frac{\nabla U(x)}{1 + \gamma \|\nabla U(x)\|} \quad \text{and} \quad H_{\gamma,c}(x) = \left(\frac{\partial_i U(x)}{1 + \gamma |\partial_i U(x)|} \right)_{i \in \{1,\ldots,d\}} \ , \tag{4.4}$$

where $\partial_i U$ is the i^{th}-coordinate of ∇U. The Euler scheme (4.3) with $G_\gamma = H_\gamma$, respectively $G_\gamma = H_{\gamma,c}$, is referred to as the Tamed Unadjusted Langevin Algorithm (TULA), respectively the coordinate-wise Tamed Unadjusted Langevin Algorithm (TULAc).

Another line of work has focused on the Metropolis Adjusted Langevin Algorithm (MALA) that consists in adding a Metropolis-Hastings step to the ULA algorithm. [BH13] provides a detailed analysis of MALA in the case where the drift coefficient is superlinear. Note also that a normalization of the gradient was suggested in [RT96, Section 1.4.3] calling it MALTA (Metropolis Adjusted Langevin Truncated Algorithm) and analyzed in [Atc06] and [BV10].

The article is organized as follows. In Section 4.2, the Markov chain $(X_k)_{k\in\mathbb{N}}$ defined by (4.3) is shown to be V-geometrically ergodic w.r.t. an invariant measure π_γ. Non-asymptotic bounds between the distribution of $(X_k)_{k\in\mathbb{N}}$ and π in total variation and Wasserstein distances are provided, as well as weak error bounds. In Section 4.3, the methodology is illustrated through numerical examples. Finally, proofs of the main results appear in Section 4.4.

Notations

Let $\mathcal{B}(\mathbb{R}^d)$ denote the Borel σ-field of $\mathbb{R}^d$. Moreover, let $\mathrm{L}^1(\mu)$ be the set of μ-integrable functions for μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Further, $\mu(f) = \int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x)$ for an $f \in \mathrm{L}^1(\mu)$. Given a Markov kernel R on $\mathbb{R}^d$, for all $x \in \mathbb{R}^d$ and f integrable under $R(x,\cdot)$, denote by $Rf(x) = \int_{\mathbb{R}^d} f(y)R(x,\mathrm{d}y)$. Let $V : \mathbb{R}^d \to [1,\infty)$ be a measurable function. The V-total variation distance between μ and ν is defined as $\|\mu - \nu\|_V = \sup_{|f|\le V} |\mu(f) - \nu(f)|$. If $V = 1$, then $\|\cdot\|_V$ is the total variation denoted by $\|\cdot\|_{\mathrm{TV}}$. Let μ and ν be two probability measures on a state space Ω with a given σ-algebra. If $\mu \ll \nu$, we denote by $\mathrm{d}\mu/\mathrm{d}\nu$ the Radon-Nikodym derivative of μ w.r.t. ν. In that case, the Kullback-Leibler divergence of μ w.r.t. to ν is defined as

$$\mathrm{KL}(\mu|\nu) = \int_\Omega \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu \,.$$

We say that ζ is a transference plan of μ and ν if it is a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for any Borel set A of $\mathbb{R}^d$, $\zeta(\mathrm{A} \times \mathbb{R}^d) = \mu(\mathrm{A})$ and $\zeta(\mathbb{R}^d \times \mathrm{A}) = \nu(\mathrm{A})$. We denote by $\Pi(\mu,\nu)$ the set of transference plans of μ and ν. Furthermore, we say that a couple of $\mathbb{R}^d$-random variables (X,Y) is a coupling of μ and ν if there exists $\zeta \in \Pi(\mu,\nu)$ such that (X,Y) are distributed according to ζ. For two probability measures μ and ν, we define the Wasserstein distance of order $p \ge 1$ as

$$W_p(\mu,\nu) = \left(\inf_{\zeta\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \,\mathrm{d}\zeta(x,y)\right)^{1/p} .$$

By [Vil09, Theorem 4.1], for all μ,ν probability measure on $\mathbb{R}^d$, there exists a transference plan $\zeta^\star \in \Pi(\mu,\nu)$ such that for any coupling (X,Y) distributed according to $\zeta^\star$, $W_p(\mu,\nu) = \mathbb{E}[\|X-Y\|^p]^{1/p}$.

For $u, v \in \mathbb{R}^d$, define the scalar product $\langle u, v \rangle = \sum_{i=1}^d u_i v_i$ and the Euclidian norm $\|u\| = \langle u, u \rangle^{1/2}$. Denote by $\mathbb{S}^{d-1} = \left\{ u \in \mathbb{R}^d : \|u\| = 1 \right\}$. For $k \in \mathbb{N}$, $m, m' \in \mathbb{N}^*$ and Ω, Ω' two open sets of $\mathbb{R}^m, \mathbb{R}^{m'}$ respectively, denote by $\mathrm{C}^k(\Omega, \Omega')$, the set of k-times continuously differentiable functions. For $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$, denote by ∇f the gradient of f, $\partial_i f$ the i^{th}-coordinate of ∇f, Δf the Laplacian of f and $\nabla^2 f$ the Hessian of f. Define then for $x \in \mathbb{R}^d$, $\|\nabla^2 f(x)\| = \sup_{u \in \mathbb{S}^{d-1}} \|\nabla^2 f(x) u\|$. For $k \in \mathbb{N}$ and $f \in \mathrm{C}^k(\mathbb{R}^d, \mathbb{R})$, denote by $\mathrm{D}^i f$ the i-th derivative of f for $i \in \{0, \ldots, k\}$, i.e. $\mathrm{D}^i f$ is a symmetric i-linear map defined for all $x \in \mathbb{R}^d$ and $j_1, \ldots, j_i \in \{1, \ldots, d\}$ by $\mathrm{D}^i f(x)[e_{j_1}, \ldots, e_{j_i}] = \partial_{j_1 \ldots j_i} f(x)$ where $e_1, \ldots, e_d$ is the canonical basis of $\mathbb{R}^d$. For $x \in \mathbb{R}^d$ and $i \in \{1, \ldots, k\}$, define $\left\|\mathrm{D}^0 f(x)\right\| = |f(x)|$, $\left\|\mathrm{D}^i f(x)\right\| = \sup_{u_1, \ldots, u_i \in \mathbb{S}^{d-1}} \mathrm{D}^i f(x)[u_1, \ldots, u_i]$. Note that $\left\|\mathrm{D}^1 f(x)\right\| = \|\nabla f(x)\|$ and $\left\|\mathrm{D}^2 f(x)\right\| = \|\nabla^2 f(x)\|$. For $m, m' \in \mathbb{N}^*$, define

$$\mathrm{C}_{\mathrm{poly}}(\mathbb{R}^m, \mathbb{R}^{m'}) = \Big\{ f \in \mathrm{C}(\mathbb{R}^m, \mathbb{R}^{m'}) | \exists C_q, q \geq 0, \forall x \in \mathbb{R}^m, \\ \|f(x)\| \leq C_q (1 + \|x\|^q) \Big\} .$$

For all $x \in \mathbb{R}^d$ and $M > 0$, we denote by $\mathrm{B}(x, M)$ (respectively $\overline{B}(x, M)$), the open (respectively closed) ball centered at x of radius M. In the sequel, we take the convention that for $n, p \in \mathbb{N}$, $n < p$ then $\sum_p^n = 0$ and $\prod_p^n = 1$.

4.2 Ergodicity and convergence analysis

In this Section, under appropriate assumptions on ∇U and G_γ, we show that the diffusion process $(Y_t)_{t \geq 0}$ defined by (5.7) and its discretization $(X_k)_{k \in \mathbb{N}}$ defined by (4.3) satisfy a Foster-Lyapunov drift condition and are V-geometrically ergodic, see Proposition 4.1 and Proposition 4.3. Second, for all $k \in \mathbb{N}^*$, non-asymptotic bounds in V-norm between the distribution of X_k and π are established. Our next results give non-asymptotic bounds in Wasserstein distance of order 2, under the additional assumption that U is strongly convex. A summary of our main contributions is given in Table 4.1, where $\lambda \in [0, 1)$. We conclude this part by non-asymptotic bounds on the bias and the variance of the ergodic average $n^{-1} \sum_{k=0}^{n-1} f(X_k)$, $n \in \mathbb{N}^*$, used as an estimator of $\pi(f)$, for $f : \mathbb{R}^d \to \mathbb{R}$ sufficiently smooth.

Henceforth, it is assumed that U is continuously differentiable. Consider the following assumptions on U.

distance	order of the upper bound	assumptions
$\left\|\delta_x R_\gamma^n - \pi\right\|_{V^{1/2}}$	$n\gamma\lambda^{n\gamma} V(x) + \sqrt{\gamma}$	**A**1, **A**2, **H**5 and **H**6
$W_2^2(\delta_x R_\gamma^n, \pi)$	$n\gamma\lambda^{n\gamma} V(x) + \gamma$	**A**1, **A**2, **H**5, **H**6 and **H**7
$W_2^2(\delta_x R_\gamma^n, \pi)$	$n\gamma^{1+\beta}\lambda^{n\gamma} V(x) + \gamma^{1+\beta}$	**A**1, **A**2, **H**6, **H**7 and **H**8

Table 4.1: Summary of the upper bounds on the distances between the distribution of the n^{th} iteration of the Markov chain defined by (4.3) and π.

H5. *There exist* $\ell, L \in \mathbb{R}_+$ *such that for all* $x, y \in \mathbb{R}^d$,

$$\|\nabla U(x) - \nabla U(y)\| \leq L \left\{1 + \|x\|^\ell + \|y\|^\ell\right\} \|x - y\| \ .$$

H6. *i)* $\liminf_{\|x\| \to +\infty} \|\nabla U(x)\| = +\infty$.

ii) $\liminf_{\|x\| \to +\infty} \left\langle \frac{x}{\|x\|}, \frac{\nabla U(x)}{\|\nabla U(x)\|} \right\rangle > 0$.

Note that under **H**6, $\liminf_{\|x\| \to +\infty} U(x) = +\infty$, U has a minimum $x^\star$ and $\nabla U(x^\star) = 0$. Without loss of generality, it is assumed that $x^\star = 0$. It implies under **H**5 that for all $x \in \mathbb{R}^d$,

$$\|\nabla U(x)\| \leq 2L \left\{1 + \|x\|^{\ell+1}\right\} \ . \tag{4.5}$$

Besides, under **H**6-ii), there exists $C \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$, $\langle -\nabla U(x), x \rangle \leq C$. By [MT93, Theorem 2.1], [IW89, Chapter IV, Theorems 2.3, 3.1] and [RT96, Theorem 2.1], (5.7) has a unique strong solution denoted $(Y_t)_{t \geq 0}$. By [KS91, Section 5.4.C, Theorem 4.20], one constructs the associated strongly Markovian semigroup $(P_t)_{t \geq 0}$ given for all $t \geq 0$, $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by $P_t(x, \mathsf{A}) = \mathbb{E}\left[\mathbb{1}_{\mathsf{A}}(Y_t) | Y_0 = x\right]$. Consider the infinitesimal generator $\mathcal{L}$ associated with (5.7) defined for all $h \in \mathrm{C}^2(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ by

$$\mathcal{L}h(x) = -\left\langle \nabla U(x), \nabla h(x) \right\rangle + \Delta h(x) \ , \tag{4.6}$$

and for any $a \in \mathbb{R}_+^*$, define the Lyapunov function $V_a : \mathbb{R}^d \to [1, +\infty)$ for all $x \in \mathbb{R}^d$ by

$$V_a(x) = \exp\left(a(1 + \|x\|^2)^{1/2}\right) \ . \tag{4.7}$$

Foster-Lyapunov conditions enable to control the moments of the diffusion process $(Y_t)_{t \geq 0}$, see e.g. [MT93, Section 6] or [RT96, Theorem 2.2].

Proposition 4.1. *Assume* ***H*** *5,* ***H*** *6 and let* $a \in \mathbb{R}_+^*$. *There exists* $b_a \in \mathbb{R}_+$ *(given explicitly in the proof) such that for all* $x \in \mathbb{R}^d$

$$\mathcal{L}V_a(x) \leq -aV_a(x) + ab_a \tag{4.8}$$

and

$$\sup_{t \geq 0} P_t V_a(x) \leq V_a(x) + b_a \ .$$

Moreover, there exist $C_a \in \mathbb{R}_+$ *and* $\rho_a \in [0, 1)$ *such that for all* $t \in \mathbb{R}_+$ *and probability measures* μ_0, ν_0 *on* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *satisfying* $\mu_0(V_a) + \nu_0(V_a) < +\infty$,

$$\|\mu_0 P_t - \nu_0 P_t\|_{V_a} \leq C_a \rho_a^t \|\mu_0 - \nu_0\|_{V_a} \ , \ \|\mu_0 P_t - \pi\|_{V_a} \leq C_a \rho_a^t \mu_0(V_a) \ . \tag{4.9}$$

Proof. The proof is postponed to Section 4.4.1. □

The Markov chain $(X_k)_{k \in \mathbb{N}}$ defined in (4.3) is a discrete-time approximation of the diffusion $(Y_t)_{t \geq 0}$. To control the total variation and Wasserstein distances of the marginal distributions of $(X_k)_{k \in \mathbb{N}}$ and $(Y_t)_{t \geq 0}$, it is necessary to assume that for $\gamma > 0$ small enough, G_γ and ∇U are close. This is formalized by **A**1. Under the additional assumption **A**2, we obtain the stability and ergodicity of $(X_k)_{k \in \mathbb{N}}$.

A 1. *For all $\gamma > 0$, G_γ is continuous. There exist $\alpha \geq 0$, $C_\alpha < +\infty$ such that for all $\gamma > 0$ and $x \in \mathbb{R}^d$,*

$$\|G_\gamma(x) - \nabla U(x)\| \leq \gamma C_\alpha \left(1 + \|x\|^\alpha\right) .$$

Note that under **H**5, **A** 1 and by (4.5), we have for all $x \in \mathbb{R}^d$

$$\|G_\gamma(x)\| \leq 2L \left\{1 + \|x\|^{\ell+1}\right\} + \gamma C_\alpha \left(1 + \|x\|^\alpha\right) . \tag{4.10}$$

A 2. *For all $\gamma > 0$, $\liminf_{\|x\| \to +\infty} \left\langle \frac{x}{\|x\|}, G_\gamma(x) \right\rangle - \frac{\gamma}{2\|x\|} \|G_\gamma(x)\|^2 > 0$.*

Lemma 4.2. *Assume **H**5 and **H**6. Let $\gamma > 0$ and G_γ be equal to H_γ or $H_{\gamma,c}$ defined in (4.4). Then **A** 1 and **A** 2 are satisfied.*

Proof. The proof is postponed to Section 4.4.2. □

The Markov kernel R_γ associated with (4.3) is given for all $\gamma > 0$, $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$R_\gamma(x, \mathsf{A}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathbb{1}_{\mathsf{A}} \left(x - \gamma G_\gamma(x) + \sqrt{2\gamma} z\right) \mathrm{e}^{-\|z\|^2/2} \mathrm{d}z . \tag{4.11}$$

We then obtain the counterpart of Proposition 4.1 for the Markov chain $(X_k)_{k \in \mathbb{N}}$.

Proposition 4.3. *Assume **H**5, **A** 1, **A** 2 and let $\gamma \in \mathbb{R}_+^*$. There exist $M, \mathfrak{a}, b \in \mathbb{R}_+^*$ (given explicitly in the proof) satisfying for all $x \in \mathbb{R}^d$*

$$R_\gamma V_{\mathfrak{a}}(x) \leq \mathrm{e}^{-\mathfrak{a}^2 \gamma} V_{\mathfrak{a}}(x) + \gamma b \mathbb{1}_{\overline{B}(0,M)}(x) . \tag{4.12}$$

In addition, R_γ has a unique invariant measure π_γ, R_γ is $V_{\mathfrak{a}}$-geometrically ergodic w.r.t. π_γ.

Proof. The proof is postponed to Section 4.4.3. □

Note that a straightforward induction of (4.12) gives for all $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$,

$$R_\gamma^n V_{\mathfrak{a}}(x) \leq \mathrm{e}^{-n\mathfrak{a}^2\gamma} V_{\mathfrak{a}}(x) + \{(b\gamma)(1 - \mathrm{e}^{-n\mathfrak{a}^2\gamma})\}/(1 - \mathrm{e}^{-\mathfrak{a}^2\gamma}) .$$

Using $1 - \mathrm{e}^{-\mathfrak{a}^2\gamma} = \int_0^\gamma \mathfrak{a}^2 \mathrm{e}^{-\mathfrak{a}^2 t} \mathrm{d}t \geq \gamma \mathfrak{a}^2 \mathrm{e}^{-\mathfrak{a}^2\gamma}$, we get for all $n \in \mathbb{N}$

$$R_\gamma^n V_{\mathfrak{a}}(x) \leq \mathrm{e}^{-\mathfrak{a}^2 n\gamma} V_{\mathfrak{a}}(x) + (b/\mathfrak{a}^2) \mathrm{e}^{\mathfrak{a}^2\gamma} . \tag{4.13}$$

In the following result, we compare the discrete and continuous time processes $(X_k)_{k \in \mathbb{N}}$ and $(Y_t)_{t \geq 0}$ using Girsanov's theorem and Pinsker's inequality, see [Dal17b] and [DM17, Theorem 10] for similar arguments.

Theorem 4.4. *Assume* ***H*** *5,* ***H*** *6,* ***A*** *1 and* ***A*** *2. Let* $\gamma_0 > 0$. *There exist* $C > 0$ *and* $\lambda \in (0,1)$ *such that for all* $\gamma \in (0, \gamma_0]$, $x \in \mathbb{R}^d$ *and* $n \in \mathbb{N}$,

$$\left\| \delta_x R_\gamma^n - \pi \right\|_{V_{\text{æ}}^{1/2}} \leq C \left(n\gamma \lambda^{n\gamma} V_{\text{æ}}(x) + \sqrt{\gamma} \right) , \tag{4.14}$$

where æ *is defined in Proposition 4.3 and for all* $\gamma \in (0, \gamma_0]$,

$$\left\| \pi_\gamma - \pi \right\|_{V_{\text{æ}}^{1/2}} \leq C\sqrt{\gamma} . \tag{4.15}$$

Proof. The proof is postponed to Section 4.4.4. □

By adding strong convexity for the potential, one obtains the corresponding bounds for the Wasserstein distance of order 2.

H 7. *U is strongly convex, i.e. there exists $m > 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m \left\| x - y \right\|^2 .$$

By coupling $(Y_t)_{t \geq 0}$ and the linear interpolation of $(X_k)_{k \in \mathbb{N}}$ with the same Brownian motion, the following result is obtained.

Theorem 4.5. *Assume* ***A*** *1,* ***A*** *2,* ***H*** *5,* ***H*** *6 and* ***H*** *7. Let* $\gamma_0 > 0$. *There exist* $C > 0$ *and* $\lambda \in (0,1)$ *such that for all* $x \in \mathbb{R}^d$, $\gamma \in (0, \gamma_0]$ *and* $n \in \mathbb{N}$,

$$W_2^2(\delta_x R_\gamma^n, \pi) \leq C \left(n\gamma \lambda^{n\gamma} V_{\text{æ}}(x) + \gamma \right) , \tag{4.16}$$

where æ *is defined in Proposition 4.3 and for all* $\gamma \in (0, \gamma_0]$,

$$W_2^2(\pi_\gamma, \pi) \leq C\gamma . \tag{4.17}$$

Proof. The proof is postponed to Section 4.4.5. □

If $U \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$ and under the following assumption on $\nabla^2 U$, the bound can be improved.

H 8. *U is twice continuously differentiable and there exist $\nu, L_H \in \mathbb{R}_+$ and $\beta \in [0,1]$ such that for all $x, y \in \mathbb{R}^d$,*

$$\left\| \nabla^2 U(x) - \nabla^2 U(y) \right\| \leq L_H \left\{ 1 + \left\| x \right\|^\nu + \left\| y \right\|^\nu \right\} \left\| x - y \right\|^\beta .$$

It is shown in Section 4.4.5 that **H**8 implies **H**5.

Theorem 4.6. *Assume* ***A*** *1,* ***A*** *2,* ***H*** *6,* ***H*** *7 and* ***H*** *8. Let* $\gamma_0 > 0$. *There exist* $C > 0$ *and* $\lambda \in (0,1)$ *such that for all* $x \in \mathbb{R}^d$, $\gamma \in (0, \gamma_0]$ *and* $n \in \mathbb{N}$,

$$W_2^2(\delta_x R_\gamma^n, \pi) \leq C \left(n\gamma^{1+\beta} \lambda^{n\gamma} V_{\text{æ}}(x) + \gamma^{1+\beta} \right) , \tag{4.18}$$

where æ *is defined in Proposition 4.3 and for all* $\gamma \in (0, \gamma_0]$,

$$W_2^2(\pi_\gamma, \pi) \leq C\gamma^{1+\beta} . \tag{4.19}$$

Proof. The proof is postponed to Section 4.4.5. □

The exponent of γ in (4.16) is improved from 1 to $1+\beta$. In particular, if $\nabla^2 U$ is Lipschitz, $\nu = 0$, $\beta = 1$, and [DM16, Theorem 8] is recovered.

Let $(X_k)_{k\in\mathbb{N}}$ be the Markov chain defined in (4.3). To study the empirical average $(1/n)\sum_{k=0}^{n-1}\{f(X_k) - \pi(f)\}$ for $n \in \mathbb{N}^*$, we follow a method introduced in [MST10] and based on the Poisson equation. For f a π-integrable function, the Poisson equation associated with the generator $\mathcal{L}$ defined in (4.6) is given for all $x \in \mathbb{R}^d$ by

$$\mathcal{L}\phi(x) = -\left(f(x) - \pi(f)\right) , \tag{4.20}$$

where ϕ, if it exists, is a solution of the Poisson equation. This equation has proved to be a useful tool to analyze additive functionals of diffusion processes, see e.g. [CCG12] and references therein. The existence and regularity of a solution of the Poisson equation has been investigated in [GM96], [PV01], [Kop15], [Gor+16]. For that purpose, the following additional assumption on U is introduced.

H9. *$U \in \mathrm{C}^4(\mathbb{R}^d, \mathbb{R})$ and $\left\|\mathrm{D}^i U\right\| \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}_+)$ for $i \in \{1, \ldots, 4\}$.*

Theorem 4.7. *Assume **H**6, **H**9, **A**1 and **A**2. Let $f \in \mathrm{C}^3(\mathbb{R}^d, \mathbb{R})$ be such that $\left\|\mathrm{D}^i f\right\| \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}_+)$ for $i \in \{0, \ldots, 3\}$. Let $\gamma_0 > 0$ and $(X_k)_{k\in\mathbb{N}}$ be the Markov chain defined by (4.3) and starting at $X_0 = 0$. There exists $C > 0$ such that for all $\gamma \in (0, \gamma_0]$ and $n \in \mathbb{N}^*$,*

$$\left|\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \pi(f)\right]\right| \leq C\left(\gamma + \frac{1}{n\gamma}\right) \tag{4.21}$$

and

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \pi(f)\right)^2\right] \leq C\left(\gamma^2 + \frac{1}{n\gamma}\right) . \tag{4.22}$$

Proof. The proof is postponed to Section 4.4.6. □

Note that the standard rates of convergence are recovered, see [MST10, Theorems 5.1, 5.2].

4.3 Numerical examples

We illustrate our theoretical results using three numerical examples.

Multivariate Gaussian variable in high dimension We first consider a multivariate Gaussian variable in dimension $d \in \{100, 1000\}$ of mean 0 and covariance matrix $\Sigma = \mathrm{diag}(1, \ldots, d)$. The potential $U : \mathbb{R}^d \to \mathbb{R}$ defined for all $x \in \mathbb{R}^d$ by $U(x) = (1/2)x^{\mathrm{T}}\Sigma^{-1}x$ is d^{-1}-strongly convex and 1-gradient Lipschitz. The assumptions **H**5, **H**6, **H**7, **H**8 with $\beta = 1$ and **H**9 are thus satisfied. Note that in this case, ULA is

stable and the analysis of [Dal17b], [DM17], [DM16] valid. Nevertheless, implementing TULA and TULAc on this example is still of interest. Indeed, some Bayesian posterior distributions have intricate expressions and identifying the superlinear part in the gradient ∇U may be a difficult task. Within this context, we check the robustness of TULA and TULAc with respect to (globally) Lipschitz ∇U.

We also consider in Section 4.E a badly conditioned multivariate Gaussian variable in dimension $d = 100$ of mean 0 and covariance matrix $\Sigma = \mathrm{diag}(10^{-5}, 1, \ldots, 1)$. In this example, ULA requires a step size of order 10^{-5} to be stable which implies a large number of iterations to obtain relevant results. On the other side, TULA and TULAc are applicable with a step size of order 10^{-2} and within a relatively small number of iterations, valid results for the axes 2 to 100 are obtained.

Double well The potential is defined for all $x \in \mathbb{R}^d$ by $U(x) = (1/4)\|x\|^4 - (1/2)\|x\|^2$. We have $\nabla U(x) = (\|x\|^2 - 1)x$ and $\nabla^2 U(x) = (\|x\|^2 - 1)\,\mathrm{Id} + 2xx^{\mathrm{T}}$. We get $\|\nabla^2 U(x)\| = 3\|x\|^2 - 1$, $\langle x, \nabla U(x)\rangle = \|x\|\,\|\nabla U(x)\|$ for $\|x\| \geq 1$ and

$$\left\|\nabla^2 U(x) - \nabla^2 U(y)\right\| \leq 3\,(\|x\| + \|y\|)\,\|x - y\| \ ,$$

so that **H**5, **H**6, **H**8 with $\beta = 1$ and **H**9 are satisfied.

Ginzburg-Landau model This model of phase transitions in physics [LFR17, Section 6.2] is defined on a three-dimensional $d = p^3$ lattice for $p \in \mathbb{N}^*$ and the potential is given for $x = (x_{ijk})_{i,j,k\in\{1,\ldots,p\}} \in \mathbb{R}^d$ by

$$U(x) = \sum_{i,j,k=1}^{p} \left\{ \frac{1-\tau}{2} x_{ijk}^2 + \frac{\tau\alpha}{2} \left\|\tilde{\nabla} x_{ijk}\right\|^2 + \frac{\tau\lambda}{4} x_{ijk}^4 \right\} \ ,$$

where $\alpha, \lambda, \tau > 0$ and $\tilde{\nabla} x_{ijk} = (x_{i_+jk} - x_{ijk}, x_{ij_+k} - x_{ijk}, x_{ijk_+} - x_{ijk})$ with $i_\pm = i \pm 1 \bmod p$ and similarly for $j_\pm, k_\pm$. In the simulations, p is equal to 10. We have

$$\nabla U(x) = \Big\{ \tau\alpha\,(6x_{ijk} - x_{i_+jk} - x_{ij_+k} - x_{ijk_+} - x_{i_-jk} - x_{ij_-k} - x_{ijk_-}) \\ + (1-\tau)x_{ijk} + \tau\lambda x_{ijk}^3 \Big\}_{i,j,k\in\{1,\ldots,p\}} \ ,$$

and

$$\nabla^2 U(x) = \mathrm{diag}\left(\left(1 - \tau + 6\tau\alpha + 3\tau\lambda x_{ijk}^2 \right)_{i,j,k\in\{1,\ldots,p\}} \right) + M \ ,$$

where $M \in \mathbb{R}^{d\times d}$ is a constant matrix. **H**5, **H**8 with $\beta = 1$ and **H**9 are thus satisfied. Using that $x \mapsto \sum_{i,j,k=1}^{p} \left\|\tilde{\nabla} x_{ijk}\right\|^2$ is convex by composition of convex functions and its gradient evaluated at 0 is 0, we have for all $x \in \mathbb{R}^d$,

$$\langle x, \nabla U(x)\rangle \geq \sum_{i,j,k=1}^{p} \{(1-\tau)x_{ijk}^2 + \tau\lambda x_{ijk}^4\} \ .$$

By Cauchy-Schwarz inequality, $\left\{\sum_{i,j,k=1}^p x_{ijk}^2\right\}^2 \leq d\sum_{i,j,k=1}^p x_{ijk}^4$, and for all $x \in \mathbb{R}^d$, $\|x\|^2 \geq (2\,|1-\tau|\,d)/(\tau\lambda)$, we get $\langle x, \nabla U(x)\rangle \geq \{(\tau\lambda)/2\}\sum_{i,j,k=1}^p x_{ijk}^4$. Besides, we have

$$\|\nabla U(x)\| \leq (|1-\tau| + 12\tau\alpha)\,\|x\| + \tau\lambda \left\|(x_{ijk}^3)_{i,j,k\in\{1,\dots,p\}}\right\| .$$

Let $a, b, c \in \{1,\dots,p\}$ be such that $|x_{abc}| = \max |x_{ijk}|$. We get

$$\|x\| \left\|(x_{ijk}^3)_{i,j,k\in\{1,\dots,p\}}\right\| \leq d x_{abc}^4 \leq d \sum_{i,j,k=1}^p x_{ijk}^4 .$$

Finally, for $\|x\|^2 \geq \max\{1, (2\,|1-\tau|\,d)/(\tau\lambda)\}$, we obtain

$$\|x\|\,\|\nabla U(x)\| \leq \left\{\frac{2d\,|1-\tau|}{\tau\lambda} + \frac{24\alpha d}{\lambda} + 2d\right\}\langle x, \nabla U(x)\rangle ,$$

and **H**6 is satisfied.

We benchmark TULA and TULAc against ULA given by (4.2), MALA and a Random Walk Metropolis-Hastings with a Gaussian proposal (RWM). TMALA (Tamed Metropolis Adjusted Langevin Algorithm) and TMALAc (coordinate-wise Tamed Metropolis Adjusted Langevin Algorithm), the Metropolized versions of TULA and TULAc, are also included in the numerical tests. Their theoretical analysis is similar to the one of MALTA [Atc06, Proposition 2.1].

Since double well and Ginzburg-Landau models are coordinate-wise exchangeable, the results are provided only for their first coordinate. The Markov chains associated with these models are started at $X_0 = 0, (10, 0^{\otimes(d-1)}), (100, 0^{\otimes(d-1)}), (1000, 0^{\otimes(d-1)})$ and for the multivariate Gaussian at a random vector of norm $0, 10, 100, 1000$. For the Gaussian and double well examples, for each initial condition, algorithm, step size $\gamma \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, we run 100 independent Markov chains started at X_0 of 10^6 samples (respectively 10^5) in dimension $d = 100$ (respectively $d = 1000$). For the Ginzburg-Landau model, we run 100 independent Markov chains started at X_0 of 10^5 samples. For each run, we estimate the 1st and 2nd moment for the first and last coordinate, i.e. $\int_{\mathbb{R}^d} x_i\pi(x)\mathrm{d}x$ for $i \in \{1, d\}$, by the empirical average and we compute the boxplots of the errors. For ULA, if the norm of X_k for $k \in \mathbb{N}$ exceeds 10^5, the chain is stopped and for this step size γ the trajectory of ULA is not taken into account. For MALA, RWM, TMALA and TMALAc, if the acceptance ratio is below 0.05, we similarly do not take into account the corresponding trajectories.

For the three examples and for $i \in \{1,\dots,d\}$, $\int_{\mathbb{R}^d} x_i\pi(x)\mathrm{d}x = 0$. By symmetry, for the double well, we have for $i \in \{1,\dots,d\}$ and $r \in \mathbb{R}_+$,

$$\mathbb{E}\left[X_i^2\right] = d^{-1}\int_{\mathbb{R}_+} r^2\nu(r)\mathrm{d}r \Big/ \int_{\mathbb{R}_+}\nu(r)\mathrm{d}r , \quad \nu(r) = r^{d-1}\exp\left\{(r^2/2) - (r^4/4)\right\} .$$

A Random Walk Metropolis run of 10^7 samples gives $\int_{\mathbb{R}^d} x_i^2\pi(x)\mathrm{d}x \approx 0.104 \pm 0.001$ for $d = 100$ and $\int_{\mathbb{R}^d} x_i^2\pi(x)\mathrm{d}x \approx 0.032 \pm 0.001$ for $d = 1000$.

We display boxplots in Figures 4.1 to 4.4. The Python code and all the figures are available at https://github.com/nbrosse/TULA. We remark that TULA, TULAc and to a lesser extent, TMALA and TMALAc, have a stable behavior even with large step sizes and starting far from the origin. This is particularly visible in Figures 4.2 and 4.4 where ULA diverges (i.e. $\liminf_{k\to+\infty} \mathbb{E}\left[\|X_k\|\right] = +\infty$) and MALA does not move even for small step sizes $\gamma = 10^{-3}$. Note however the existence of a bias for ULA, TULA and TULAc in Figure 4.3. Finally, comparison of the results shows that TULAc is preferable to TULA.

Note that other choices are possible for G_γ, depending on the model under study. For example, in the case of the double well, we could "tame" only the superlinear part of ∇U, i.e. consider for all $\gamma > 0$ and $x \in \mathbb{R}^d$,

$$G_\gamma(x) = \frac{\|x\|^2 x}{1 + \gamma \|x\|^2} - x \; . \tag{4.23}$$

A1 is satisfied and we have

$$\left\langle \frac{x}{\|x\|}, G_\gamma(x) \right\rangle - \frac{\gamma}{2\|x\|} \|G_\gamma(x)\|^2 = \frac{\|x\|^3}{1 + \gamma \|x\|^2} \left\{ 1 + \gamma - \frac{\gamma}{2} \frac{\|x\|^2}{1 + \gamma \|x\|^2} \right\} - \|x\| \left\{ 1 + (\gamma/2) \right\} \; ,$$

$$\liminf_{\|x\|\to+\infty} \left\langle \frac{x}{\|x\|^2}, G_\gamma(x) \right\rangle - \frac{\gamma}{2\|x\|^2} \|G_\gamma(x)\|^2 = \frac{\gamma^{-1} - \gamma}{2} \; .$$

A2 is satisfied if and only if $\gamma \in (0, 1)$. It is striking to see that this theoretical threshold is clearly visible on the simulations. The algorithm (4.3) with G_γ defined by (4.23) obtains similar results as TULAc for $\gamma < 1$ but for $\gamma = 1$, the algorithm diverges.

Given the results of the numerical experiments, TULAc should be chosen over ULA to sample from general probability distributions. Indeed, TULAc has similar results as ULA when the step size is small and is more stable when using larger step sizes.

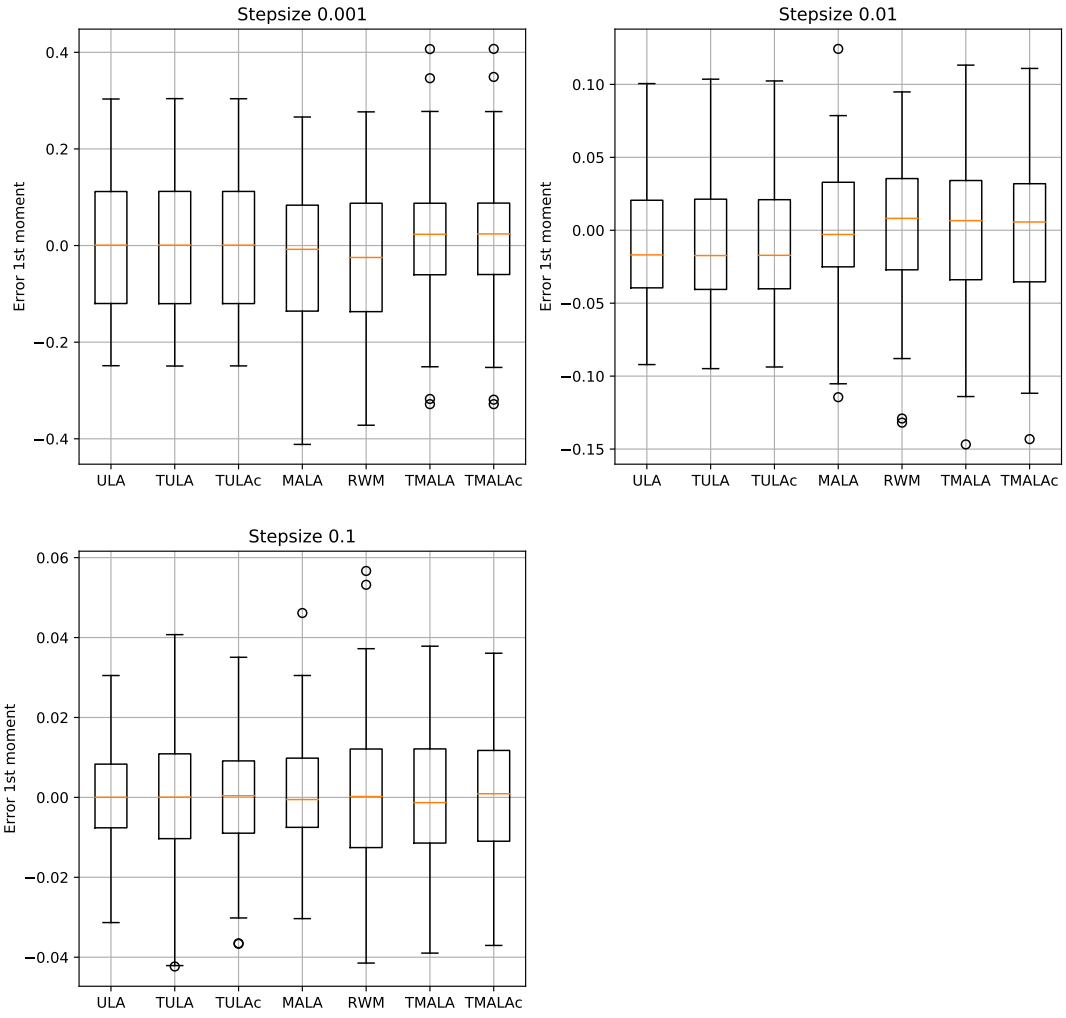

Figure 4.1: Boxplots of the error on the first moment for the multivariate Gaussian (first coordinate) in dimension 1000 starting at 0 for different step sizes.

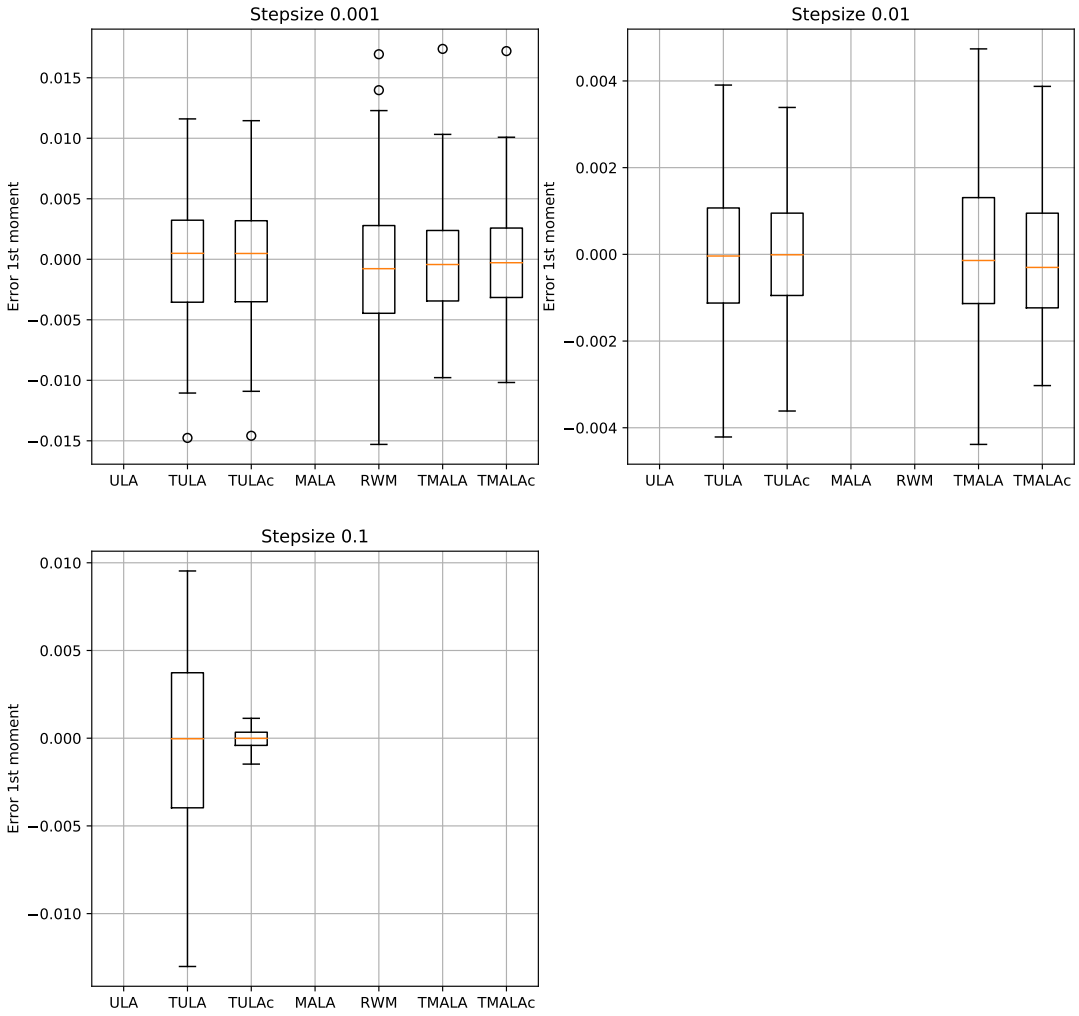

Figure 4.2: Boxplots of the error on the first moment for the double well in dimension 100 starting at $(100, 0^{\otimes 99})$ for different step sizes.

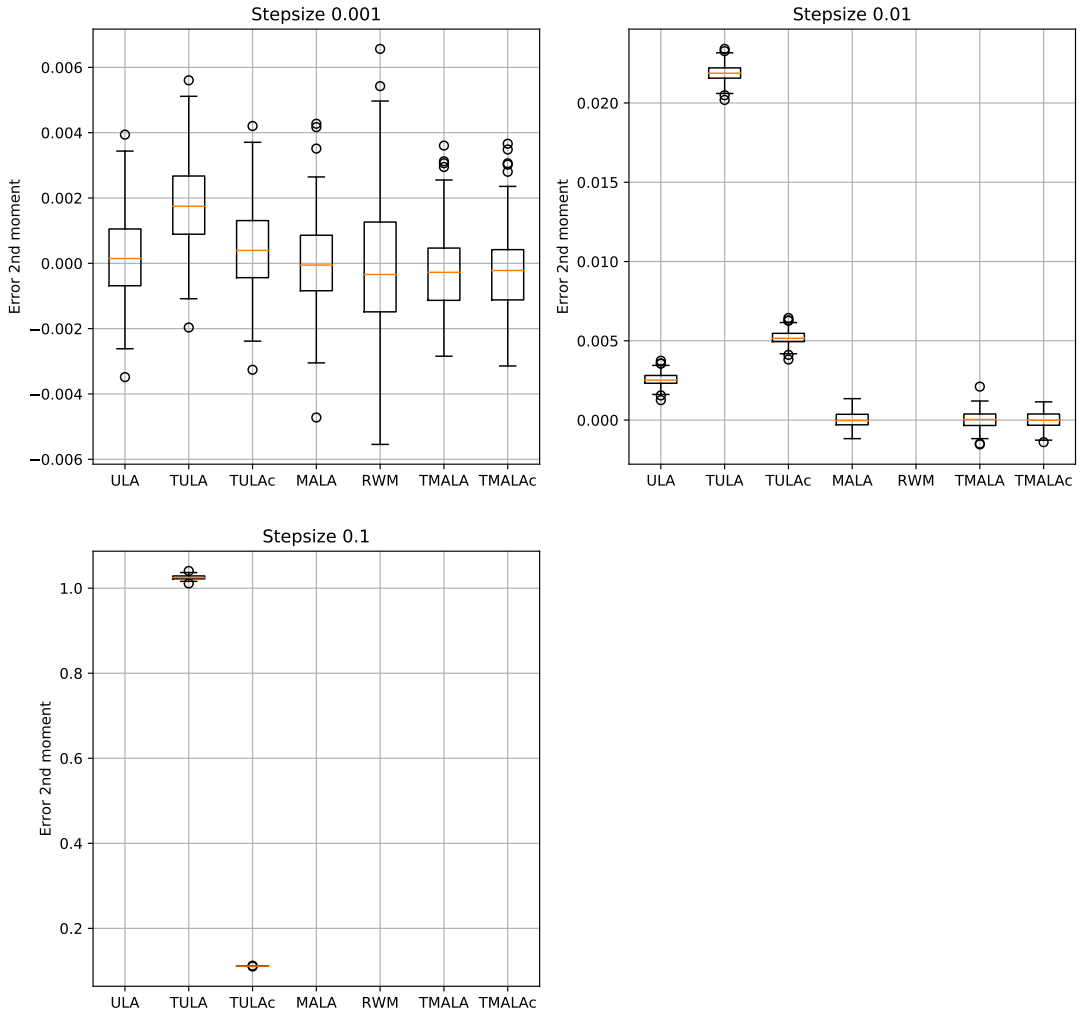

Figure 4.3: Boxplots of the error on the second moment for the double well in dimension 100 starting at 0 for different step sizes.

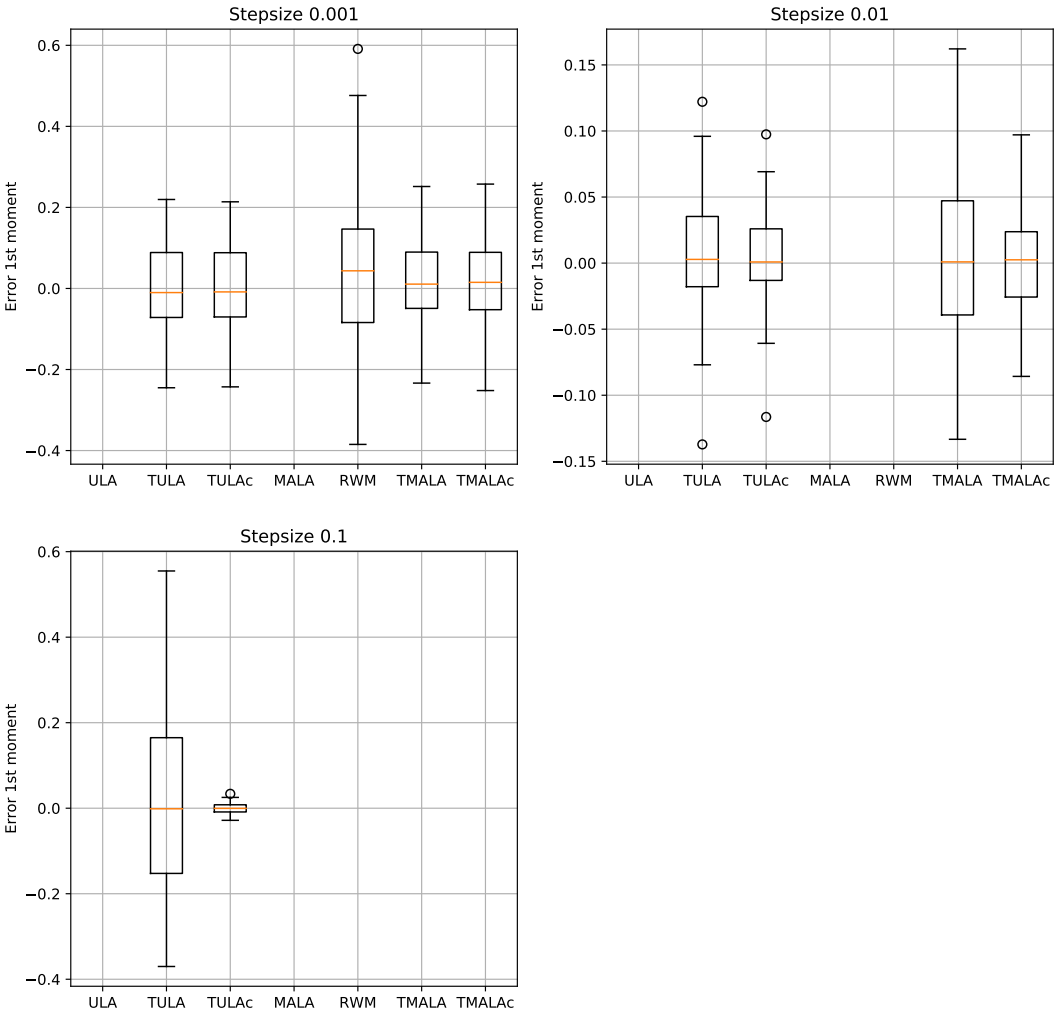

Figure 4.4: Boxplots of the error on the first moment for the Ginzburg-Landau model in dimension 1000 starting at $(100, 0^{\otimes 999})$ for different step sizes.

4.4 Proofs

4.4.1 Proof of Proposition 4.1

We have for all $x \in \mathbb{R}^d$,

$$\frac{\mathcal{L}V_a(x)}{aV_a(x)} = -\left\langle \nabla U(x), \frac{x}{(1+\|x\|^2)^{1/2}} \right\rangle + \frac{a\|x\|^2}{1+\|x\|^2} + \frac{d}{(1+\|x\|^2)^{1/2}} - \frac{\|x\|^2}{(1+\|x\|^2)^{3/2}} . \quad (4.24)$$

By **H**6-ii) and using $s \mapsto s/(1+s^2)^{1/2}$ is non-decreasing for $s \geq 0$, there exist $M_1, \kappa \in \mathbb{R}_+^*$ such that for all $x \in \mathbb{R}^d$, $\|x\| \geq M_1$, $\left\langle \nabla U(x), x(1+\|x\|^2)^{-1/2} \right\rangle \geq \kappa \|\nabla U(x)\|$. By **H**6-i), there exists $M_2 \geq M_1$ such that for all $x \in \mathbb{R}^d$, $\|x\| \geq M_2$, $\|\nabla U(x)\| \geq \kappa^{-1}\{1+a+d(1+M_1^2)^{-1/2}\}$. We then have for all $x \in \mathbb{R}^d$, $\|x\| \geq M_2$, $\mathcal{L}V_a(x) \leq -aV_a(x)$. Define

$$b_a = \exp(a(1+M_2^2)^{1/2})\{2L(1+M_2^{\ell+1}) + a + d\} .$$

Combining (4.5) and (4.24) gives (4.8). By [MT93, Theorem 1.1], we get $P_t V_a(x) \leq \mathrm{e}^{-at}V_a(x) + b_a(1-\mathrm{e}^{-at})$. The second statement is a consequence of [RT96, Theorem 2.2] and [MT93, Theorem 6.1].

4.4.2 Proof of Lemma 4.2

Let $\gamma > 0$. We have for all $x \in \mathbb{R}^d$, $\|H_\gamma(x) - \nabla U(x)\| \leq \gamma \|\nabla U(x)\|^2$ and

$$\|H_{\gamma,c}(x) - \nabla U(x)\| \leq \gamma \left\{ \sum_{i=1}^d (\partial_i U(x))^4 \right\}^{1/2} \leq \gamma \|\nabla U(x)\|^2 .$$

By (4.5), **A**1 is satisfied with $\alpha = 2\ell + 2$. Define for all $x \in \mathbb{R}^d$, $x \neq 0$,

$$A_\gamma(x) = \left\langle \frac{x}{\|x\|}, H_\gamma(x) \right\rangle - \frac{\gamma}{2\|x\|} \|H_\gamma(x)\|^2 .$$

By **H**6-ii), there exist $M_1, \kappa > 0$ such that for all $x \in \mathbb{R}^d$, $\|x\| \geq M_1$, $\langle x, \nabla U(x) \rangle \geq \kappa \|x\| \|\nabla U(x)\|$. We get then for all $x \in \mathbb{R}^d$, $\|x\| \geq M_1$,

$$\begin{aligned} A_\gamma(x) &= \frac{1}{2\|x\|\{1+\gamma\|\nabla U(x)\|\}} \left\{ 2\langle x, \nabla U(x)\rangle - \|\nabla U(x)\| \frac{\gamma\|\nabla U(x)\|}{1+\gamma\|\nabla U(x)\|} \right\} \\ &\geq \frac{\|\nabla U(x)\|}{1+\gamma\|\nabla U(x)\|} \frac{2\kappa\|x\| - 1}{2\|x\|} . \end{aligned}$$

By **H**6-i), there exist $M_2, C > 0$ such that for all $x \in \mathbb{R}^d$, $\|x\| \geq M_2$, $\|\nabla U(x)\| \geq C$. Using that $s \mapsto s(1+\gamma s)^{-1}$ is non-decreasing for $s \geq 0$, we get for all $x \in \mathbb{R}^d$, $\|x\| \geq \max(\kappa^{-1}, M_1, M_2)$, $A_\gamma(x) \geq (\kappa C)/\{2(1+\gamma C)\}$.

Define for all $x \in \mathbb{R}^d$, $x \neq 0$,

$$B_\gamma(x) = \left\langle \frac{x}{\|x\|}, H_{\gamma,c}(x) \right\rangle - \frac{\gamma}{2\|x\|} \|H_{\gamma,c}(x)\|^2 \ .$$

We have for all $x \in \mathbb{R}^d$, $\gamma \|H_{\gamma,c}(x)\| \leq \sqrt{d}$ and for all $x \in \mathbb{R}^d$, $\|x\| \geq M_1$,

$$\left\langle x, \left(\frac{\partial_i U(x)}{1 + \gamma |\partial_i U(x)|} \right)_{i \in \{1,\dots,d\}} \right\rangle \geq \frac{\kappa \|x\| \|\nabla U(x)\|}{1 + \gamma \max_{i \in \{1,\dots,d\}} |\partial_i U(x)|}$$

and

$$\left\| \left(\frac{\partial_i U(x)}{1 + \gamma |\partial_i U(x)|} \right)_{i \in \{1,\dots,d\}} \right\| \leq \frac{\|\nabla U(x)\|}{1 + \gamma \max_{i \in \{1,\dots,d\}} |\partial_i U(x)|} \ .$$

Combining these inequalities, we get for all $x \in \mathbb{R}^d$, $\|x\| \geq \max(\kappa^{-1}\sqrt{d}, M_1)$,

$$B_\gamma(x) \geq \frac{\|\nabla U(x)\|}{1 + \gamma \max_{i \in \{1,\dots,d\}} |\partial_i U(x)|} \frac{1}{2\|x\|} \left\{ 2\kappa \|x\| - \sqrt{d} \right\} \geq \frac{\|\nabla U(x)\|}{1 + \gamma \|\nabla U(x)\|} \frac{\kappa}{2} \ ,$$

and for all $x \in \mathbb{R}^d$, $\|x\| \geq \max(\kappa^{-1}\sqrt{d}, M_1, M_2)$, we get $B_\gamma(x) \geq (\kappa C)/\{2(1+\gamma C)\}$.

4.4.3 Proof of Proposition 4.3

Let $\gamma, a \in \mathbb{R}_+^*$. Note that the function $x \mapsto (1 + \|x\|^2)^{1/2}$ is Lipschitz continuous with Lipschitz constant equal to 1. By the log-Sobolev inequality [BGL14, Proposition 5.5.1], and the Cauchy-Schwarz inequality, we have for all $x \in \mathbb{R}^d$ and $a > 0$

$$\begin{aligned} R_\gamma V_a(x) &\leq \mathrm{e}^{a^2\gamma} \exp\left\{ a \int_{\mathbb{R}^d} (1 + \|y\|^2)^{1/2} R_\gamma(x, \mathrm{d}y) \right\} \\ &\leq \mathrm{e}^{a^2\gamma} \exp\left\{ a \left(1 + \|x - \gamma G_\gamma(x)\|^2 + 2\gamma d \right)^{1/2} \right\} \ . \end{aligned} \tag{4.25}$$

We now bound the term inside the exponential in the right hand side. For all $x \in \mathbb{R}^d$,

$$\|x - \gamma G_\gamma(x)\|^2 = \|x\|^2 - 2\gamma \left(\langle G_\gamma(x), x \rangle - (\gamma/2) \|G_\gamma(x)\|^2 \right) \ . \tag{4.26}$$

By **A** 2, there exist $M_1, \kappa \in \mathbb{R}_+^*$ such that for all $x \in \mathbb{R}^d$, $\|x\| \geq M_1$, $\langle x, G_\gamma(x) \rangle - (\gamma/2)\|G_\gamma(x)\|^2 \geq \kappa \|x\|$. Denote by $M = \max(M_1, 2d\kappa^{-1})$. For all $x \in \mathbb{R}^d$, $\|x\| \geq M$, we have

$$\|x - \gamma G_\gamma(x)\|^2 + 2\gamma d \leq \|x\|^2 - \gamma\kappa \|x\| \ .$$

Using for all $t \in [0,1]$, $(1-t)^{1/2} \leq 1 - t/2$ and $s \mapsto s/(1+s^2)^{1/2}$ is non-decreasing for $s \geq 0$, we have for all $x \in \mathbb{R}^d$, $\|x\| \geq M$,

$$\begin{aligned} \left(1 + \|x - \gamma G_\gamma(x)\|^2 + 2\gamma d \right)^{1/2} &\leq \left(1 + \|x\|^2 \right)^{1/2} \left(1 - \frac{\gamma\kappa \|x\|}{1 + \|x\|^2} \right)^{1/2} \\ &\leq \left(1 + \|x\|^2 \right)^{1/2} - \frac{\gamma\kappa M}{2(1+M^2)^{1/2}} \ . \end{aligned}$$

Plugging this result in (4.25) shows that for all $x \in \mathbb{R}^d$, $\|x\| \geq M$,

$$R_\gamma V_{\text{æ}}(x) \leq \mathrm{e}^{-\text{æ}^2\gamma} V_{\text{æ}}(x) \quad \text{for} \quad \text{æ} = \frac{\kappa M}{4(1+M^2)^{1/2}} \,. \tag{4.27}$$

By (4.10), we have

$$\max_{\|x\| \leq M} \|G_\gamma(x)\| \leq 2L \left\{1 + \|M\|^{\ell+1}\right\} + \gamma C_\alpha \left(1 + \|M\|^\alpha\right) \,.$$

Combining it with (4.25), (4.26), $s \mapsto s/(1+s^2)^{1/2}$ is non-decreasing for $s \geq 0$ and $(1+t_1+t_2)^{1/2} \leq (1+t_1)^{1/2} + t_2/2$ for $t_1 = \|x\|^2$, $t_2 = \gamma^2 \|G_\gamma(x)\|^2 + 2\gamma \|x\| \|G_\gamma(x)\| + 2\gamma d$, we have for all $x \in \mathbb{R}^d$, $\|x\| \leq M$,

$$R_\gamma V_{\text{æ}}(x) \leq \mathrm{e}^{\gamma c} V_{\text{æ}}(x) \,, \tag{4.28}$$

where

$$c = \text{æ}^2 + \text{æ}\Big[M\left\{2L\left\{1 + \|M\|^{\ell+1}\right\} + \gamma C_\alpha\left(1 + \|M\|^\alpha\right)\right\} + \frac{\gamma}{2}\left\{2L\left\{1 + \|M\|^{\ell+1}\right\} + \gamma C_\alpha\left(1 + \|M\|^\alpha\right)\right\}^2 + d\Big] \,.$$

Then, using that for all $t \geq 0$, $1 - \mathrm{e}^{-t} \leq t$, we get for all $x \in \mathbb{R}^d$, $\|x\| \leq M$,

$$R_\gamma V_{\text{æ}}(x) - \mathrm{e}^{-\text{æ}^2\gamma} V_{\text{æ}}(x) \leq \mathrm{e}^{\gamma c}(1 - \mathrm{e}^{-\gamma(\text{æ}^2+c)}) V_{\text{æ}}(x) \leq \gamma \mathrm{e}^{\gamma c}(\text{æ}^2 + c) V_{\text{æ}}(x) \,, \tag{4.29}$$

which combined with (4.27) gives (4.12) with $b = \mathrm{e}^{\gamma c}(\text{æ}^2 + c)\mathrm{e}^{\kappa M/4}$. Finally, using Jensen's inequality and $(s+t)^\varsigma \leq s^\varsigma + t^\varsigma$ for $\varsigma \in (0,1)$, $s, t \geq 0$ in (4.12), by [RT96, Section 3.1], for all $\gamma > 0$, R_γ has a unique invariant probability measure π_γ and R_γ is $V_{\text{æ}}^\varsigma$-geometrically ergodic w.r.t. π_γ.

4.4.4 Proof of Theorem 4.4

The proof is adapted from [DT12, Proposition 2] and [DM17, Theorem 10]. We first state a lemma.

Lemma 4.8. *Assume* ***H***5, ***H***6, ***A***1 *and* ***A***2. *Let* $\gamma_0 > 0$, $p \in \mathbb{N}^*$ *and* ν_0 *be a probability measure on* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. *There exists* $C > 0$ *such that for all* $\gamma \in (0, \gamma_0]$

$$\mathrm{KL}(\nu_0 R_\gamma^p | \nu_0 P_{p\gamma}) \leq C\gamma^2 \int_{\mathbb{R}^d} \sum_{i=0}^{p-1} \left\{\int_{\mathbb{R}^d} V_{\text{æ}}(z) R_\gamma^i(y, \mathrm{d}z)\right\} \nu_0(\mathrm{d}y) \,.$$

Proof. Let $y \in \mathbb{R}^d$ and $\gamma > 0$. Denote by $(Y_t, \overline{Y}_t)_{t \geq 0}$ the unique strong solution of

$$\begin{cases} \mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad , \quad Y_0 = y \,, \\ \mathrm{d}\overline{Y}_t = -G_\gamma\left(\overline{Y}_{\lfloor t/\gamma \rfloor \gamma}\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad , \quad \overline{Y}_0 = y \,, \end{cases} \tag{4.30}$$

and by $(\mathcal{F}_t)_{t\geq 0}$ the filtration associated with $(B_t)_{t\geq 0}$. Denote by μ_p^y and $\overline{\mu}_p^y$ the marginal distributions on $\mathcal{C}([0,p\gamma],\mathbb{R}^d)$ of $(Y_t,\overline{Y}_t)_{t\geq 0}$. By (4.5), (4.10) and Propositions 4.1 and 4.3, we have

$$\mathbb{P}\left(\int_0^{p\gamma} \|\nabla U(Y_t)\|^2 + \left\|G_\gamma\left(Y_{\lfloor t/\gamma\rfloor\gamma}\right)\right\|^2 \mathrm{d}t < +\infty\right) = 1\,,$$
$$\mathbb{P}\left(\int_0^{p\gamma} \left\|\nabla U(\overline{Y}_t)\right\|^2 + \left\|G_\gamma\left(\overline{Y}_{\lfloor t/\gamma\rfloor\gamma}\right)\right\|^2 \mathrm{d}t < +\infty\right) = 1\,.$$

By [LS13, Theorem 7.19], μ_p^y and $\overline{\mu}_p^y$ are equivalent and $\mathbb{P}$-almost surely,

$$\frac{\mathrm{d}\mu_p^y}{\mathrm{d}\overline{\mu}_p^y}((\overline{Y}_t)_{t\in[0,p\gamma]}) = \exp\left(\frac{1}{2}\int_0^{p\gamma}\left\langle -\nabla U(\overline{Y}_s) + G_\gamma\left(\overline{Y}_{\lfloor s/\gamma\rfloor\gamma}\right), \mathrm{d}\overline{Y}_s\right\rangle \right.$$
$$\left. - \frac{1}{4}\int_0^{p\gamma}\left\{\left\|\nabla U(\overline{Y}_s)\right\|^2 - \left\|G_\gamma\left(\overline{Y}_{\lfloor s/\gamma\rfloor\gamma}\right)\right\|^2\right\}\mathrm{d}s\right)\,.$$

We get then

$$\begin{aligned}\mathrm{KL}(\overline{\mu}_p^y|\mu_p^y) &= \mathbb{E}\left[-\log\left\{\frac{\mathrm{d}\mu_p^y}{\mathrm{d}\overline{\mu}_p^y}((\overline{Y}_t)_{t\in[0,p\gamma]})\right\}\right] \\ &= (1/4)\int_0^{p\gamma}\mathbb{E}\left[\left\|\nabla U(\overline{Y}_s) - G_\gamma\left(\overline{Y}_{\lfloor s/\gamma\rfloor\gamma}\right)\right\|^2\right]\mathrm{d}s \\ &= (1/4)\sum_{i=0}^{p-1}\int_{i\gamma}^{(i+1)\gamma}\mathbb{E}\left[\left\|\nabla U(\overline{Y}_s) - G_\gamma(\overline{Y}_{i\gamma})\right\|^2\right]\mathrm{d}s\,.\end{aligned}$$

For $i\in\{0,\dots,p-1\}$ and $s\in[i\gamma,(i+1)\gamma)$, we have $\left\|\nabla U(\overline{Y}_s) - G_\gamma(\overline{Y}_{i\gamma})\right\|^2 \leq 2(A_1+A_2)$ where

$$A_1 = \left\|\nabla U(\overline{Y}_s) - \nabla U(\overline{Y}_{i\gamma})\right\|^2\,,\quad A_2 = \left\|\nabla U(\overline{Y}_{i\gamma}) - G_\gamma(\overline{Y}_{i\gamma})\right\|^2\,.$$

By **A**1, $A_2 \leq \gamma^2 C_\alpha^2\left(1+\left\|\overline{Y}_{i\gamma}\right\|^\alpha\right)^2$ and by **H**5,

$$A_1 \leq L^2\left(1+\left\|\overline{Y}_s\right\|^\ell + \left\|\overline{Y}_{i\gamma}\right\|^\ell\right)^2\left\|\overline{Y}_s - \overline{Y}_{i\gamma}\right\|^2\,. \tag{4.31}$$

On the other hand for $s\in[i\gamma,(i+1)\gamma)$,

$$\left\|\overline{Y}_s - \overline{Y}_{i\gamma}\right\|^2 = (s-i\gamma)^2\left\|G_\gamma(\overline{Y}_{i\gamma})\right\|^2 + 2\left\|B_s - B_{i\gamma}\right\|^2$$
$$- 2^{3/2}(s-i\gamma)\left\langle B_s - B_{i\gamma}, G_\gamma(\overline{Y}_{i\gamma})\right\rangle\,, \tag{4.32}$$
$$\left\|\overline{Y}_s\right\| \leq \left\|\overline{Y}_{i\gamma}\right\| + \gamma\left\|G_\gamma(\overline{Y}_{i\gamma})\right\| + \sqrt{2}\left\|B_s - B_{i\gamma}\right\|\,. \tag{4.33}$$

Define $\mathsf{P}_{\gamma,1}:\mathbb{R}_+\to\mathbb{R}_+$ for all $t\in\mathbb{R}_+$ by

$$\mathsf{P}_{\gamma,1}(t) = (2\pi)^{-d/2}L^2\int_{\mathbb{R}^d}\left[2\|z\|^2 + \gamma\left\{2L(1+t^{\ell+1}) + \gamma C_\alpha(1+t^\alpha)\right\}^2\right]$$
$$\times\left[1+t^\ell + \left\{t+\gamma\left(2L(1+t^{\ell+1}) + \gamma C_\alpha(1+t^\alpha)\right) + \sqrt{2\gamma}\|z\|\right\}^\ell\right]^2 \mathrm{e}^{-\|z\|^2/2}\mathrm{d}z\,. \tag{4.34}$$

By (4.10), (4.31), (4.32) and (4.33), we have for $i \in \{0, \dots, p-1\}$

$$\int_{i\gamma}^{(i+1)\gamma} \mathbb{E}^{\mathcal{F}_{i\gamma}}\left[A_1\right] \mathrm{d}s \leq (\gamma^2/2)\mathsf{P}_{\gamma,1}\left(\left\|\overline{Y}_{i\gamma}\right\|\right)$$

and we get

$$\int_{i\gamma}^{(i+1)\gamma} \mathbb{E}^{\mathcal{F}_{i\gamma}}\left[\left\|\nabla U(\overline{Y}_s) - G_\gamma(\overline{Y}_{i\gamma})\right\|^2\right] \mathrm{d}s \leq \gamma^2 \left\{\mathsf{P}_{\gamma,1}\left(\left\|\overline{Y}_{i\gamma}\right\|\right) + 2\gamma\mathsf{P}_2\left(\left\|\overline{Y}_{i\gamma}\right\|\right)\right\} ,$$

where $\mathsf{P}_2 : \mathbb{R}_+ \to \mathbb{R}_+$ is defined for all $t \in \mathbb{R}_+$ by

$$\mathsf{P}_2(t) = C_\alpha^2 \left(1 + t^\alpha\right)^2 . \tag{4.35}$$

By [Kul97, Theorem 4.1, Chapter 2], we obtain

$$\mathrm{KL}(\delta_y R_\gamma^p | \delta_y P_{p\gamma}) \leq \mathrm{KL}(\overline{\mu}_p^y | \mu_p^y) \leq (\gamma^2/4) \sum_{i=0}^{p-1} \mathbb{E}\left[\mathsf{P}_{\gamma,1}\left(\left\|\overline{Y}_{i\gamma}\right\|\right) + 2\gamma\mathsf{P}_2\left(\left\|\overline{Y}_{i\gamma}\right\|\right)\right] .$$

By (4.34) and (4.35), there exists $C > 0$ such that for all $\gamma \in (0, \gamma_0]$ and $x \in \mathbb{R}^d$, $\mathsf{P}_{\gamma,1}(\|x\|) + 2\gamma\mathsf{P}_2(\|x\|) \leq 4CV_{\text{æ}}(x)$. Combining it with the chain rule for the Kullback-Leibler divergence concludes the proof. □

Proof of Theorem 4.4. Let $\gamma \in (0, \gamma_0]$. By Proposition 4.1, we have for all $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$,

$$\left\|\delta_x R_\gamma^n - \pi\right\|_{V_{\text{æ}}^{1/2}} \leq C_{\text{æ}/2}\rho_{\text{æ}/2}^{n\gamma} V_{\text{æ}}^{1/2}(x) + \left\|\delta_x R_\gamma^n - \delta_x P_{n\gamma}\right\|_{V_{\text{æ}}^{1/2}} .$$

Denote by $k_\gamma = \lceil \gamma^{-1} \rceil$ and by q_γ, r_γ the quotient and the remainder of the Euclidian division of n by k_γ. We have $\|\delta_x R_\gamma^n - \delta_x P_{n\gamma}\|_{V_{\text{æ}}^{1/2}} \leq A + B$ where

$$\begin{aligned} A &= \left\|\delta_x R_\gamma^{q_\gamma k_\gamma} P_{r_\gamma \gamma} - \delta_x R_\gamma^n\right\|_{V_{\text{æ}}^{1/2}} \\ B &= \sum_{i=1}^{q_\gamma} \left\|\delta_x R_\gamma^{(i-1)k_\gamma} P_{(n-(i-1)k_\gamma)\gamma} - \delta_x R_\gamma^{ik_\gamma} P_{(n-ik_\gamma)\gamma}\right\|_{V_{\text{æ}}^{1/2}} \\ &\leq \sum_{i=1}^{q_\gamma} C_{\text{æ}/2}\rho_{\text{æ}/2}^{(n-ik_\gamma)\gamma} \left\|\delta_x R_\gamma^{(i-1)k_\gamma} P_{k_\gamma\gamma} - \delta_x R_\gamma^{ik_\gamma}\right\|_{V_{\text{æ}}^{1/2}} . \end{aligned} \tag{4.36}$$

For $i \in \{1, \dots, q_\gamma\}$ we have by [DM17, Lemma 24],

$$\begin{aligned} \left\|\delta_x R_\gamma^{(i-1)k_\gamma} P_{k_\gamma\gamma} - \delta_x R_\gamma^{ik_\gamma}\right\|_{V_{\text{æ}}^{1/2}}^2 \leq 2 \left\{\delta_x R_\gamma^{(i-1)k_\gamma} P_{k_\gamma\gamma}(V_{\text{æ}}) + \delta_x R_\gamma^{ik_\gamma}(V_{\text{æ}})\right\} \\ \times \mathrm{KL}(\delta_x R_\gamma^{ik_\gamma} | \delta_x R_\gamma^{(i-1)k_\gamma} P_{k_\gamma\gamma}) . \end{aligned} \tag{4.37}$$

By Proposition 4.3, Lemma 4.8 and $k_\gamma \le 1 + \gamma^{-1}$, we have for all $i \in \{1, \dots, q_\gamma\}$

$$\begin{aligned}\mathrm{KL}(\delta_x R_\gamma^{ik_\gamma} | \delta_x R_\gamma^{(i-1)k_\gamma} P_{k_\gamma \gamma}) &\le C\gamma^2 \sum_{j=0}^{k_\gamma - 1} \int_{\mathbb{R}^d} V_{\mathfrak{a}}(z) \delta_x R_\gamma^{(i-1)k_\gamma + j}(\mathrm{d}z) \\ &\le C\gamma^2 (1+\gamma^{-1}) \left\{ \mathrm{e}^{-\mathfrak{a}^2 \gamma k_\gamma (i-1)} V_{\mathfrak{a}}(x) + \frac{b}{\mathfrak{a}^2} \mathrm{e}^{\mathfrak{a}^2 \gamma} \right\} , \quad (4.38)\end{aligned}$$

where C is the constant defined in Lemma 4.8. By Proposition 4.1, we have for $x \in \mathbb{R}^d$, $P_{k_\gamma \gamma} V_{\mathfrak{a}}(x) \le V_{\mathfrak{a}}(x) + b_{\mathfrak{a}}$ and by Proposition 4.3, we get for all $i \in \{1, \dots, q_\gamma\}$

$$\delta_x R_\gamma^{(i-1)k_\gamma} P_{k_\gamma \gamma}(V_{\mathfrak{a}}) + \delta_x R_\gamma^{ik_\gamma}(V_{\mathfrak{a}}) \le 2 \left\{ \mathrm{e}^{-\mathfrak{a}^2 \gamma k_\gamma (i-1)} V_{\mathfrak{a}}(x) + \frac{b}{\mathfrak{a}^2} \mathrm{e}^{\mathfrak{a}^2 \gamma} + b_{\mathfrak{a}} \right\} . \quad (4.39)$$

By (4.36), (4.37), (4.38) and (4.39), we obtain

$$\begin{aligned}B \le 2 C_{\mathfrak{a}/2} C^{1/2} \gamma (1+\gamma^{-1})^{1/2} \\ \times \sum_{i=1}^{q_\gamma} \rho_{\mathfrak{a}/2}^{(q_\gamma - i)\gamma k_\gamma} \left\{ \mathrm{e}^{-(i-1)\gamma k_\gamma \mathfrak{a}^2} V_{\mathfrak{a}}(x) + \left(b_{\mathfrak{a}} + \frac{b}{\mathfrak{a}^2} \mathrm{e}^{\mathfrak{a}^2 \gamma} \right) \right\}\end{aligned}$$

and we get

$$\begin{aligned}B \left\{ 2 C_{\mathfrak{a}/2} C^{1/2} \gamma (1+\gamma^{-1})^{1/2} \right\}^{-1} \le \left(b_{\mathfrak{a}} + \frac{b}{\mathfrak{a}^2} \mathrm{e}^{\mathfrak{a}^2 \gamma} \right) \frac{1}{1 - \rho_{\mathfrak{a}/2}^{k_\gamma \gamma}} \\ + V_{\mathfrak{a}}(x) q_\gamma \max(\rho_{\mathfrak{a}/2}, \mathrm{e}^{-\mathfrak{a}^2})^{(q_\gamma - 1)\gamma k_\gamma} .\end{aligned}$$

Bounding A along the same lines and using $k_\gamma \gamma \ge 1$, we get (4.14). By Proposition 4.3 and taking the limit $n \to +\infty$, we obtain (4.15). □

4.4.5 Proofs of Theorems 4.5 and 4.6

We first state preliminary technical lemmas on the diffusion $(Y_t)_{t \ge 0}$. The proofs are postponed to the Appendix. Define for all $p \in \mathbb{N}^*$ and $k \in \{0, \cdots, p\}$,

$$a_{k,p} = m^{k-p} \prod_{i=k+1}^{p} \left\{ i(d + 2(i-1))(i-k)^{-1} \right\} . \quad (4.40)$$

Lemma 4.9. *Assume **H** 7. Let $p \in \mathbb{N}^*$, $x \in \mathbb{R}^d$ and $(Y_t)_{t \ge 0}$ be the solution of (5.7) started at x. For all $t \ge 0$,*

$$\mathbb{E}\left[\|Y_t\|^{2p} \right] \le a_{0,p} \left(1 - \mathrm{e}^{-2pmt} \right) + \sum_{k=1}^{p} a_{k,p} \mathrm{e}^{-2kmt} \|x\|^{2k} ,$$

where for $k \in \{0, \cdots, p\}$, $a_{k,p}$ is given in (4.40).

Proof. The proof is postponed to Section 4.A. □

Lemma 4.10. *Assume* ***H*** *7 and let* $p \in \mathbb{N}^*$. *We have* $\int_{\mathbb{R}^d} \|y\|^{2p} \pi(\mathrm{d}y) \leq a_{0,p}$.

Proof. By Equation (4.58) and [RT96, Theorem 2.2], $(Y_t)_{t\geq 0}$ the solution of (5.7) is V_p-geometrically ergodic w.r.t. π. Taking the limit $t \to +\infty$ in Lemma 4.9 concludes the proof. □

Let $\gamma > 0$ and under **H**5 set

$$N = \lceil (\ell + 1)/2 \rceil \ . \tag{4.41}$$

Consider $\mathsf{P}_{\gamma,3} : \mathbb{R}_+ \to \mathbb{R}_+$ defined for all $s \in \mathbb{R}_+$ by

$$\mathsf{P}_{\gamma,3}(s) = 2d + 8L^2(1 + s^{\ell+1}) \left\{ \frac{\gamma}{2} \left(2 + \sum_{k=1}^{N} a_{k,N} s^{2k} \right) + N m a_{0,N} \frac{\gamma^2}{3} \right\} \ . \tag{4.42}$$

Lemma 4.11. *Assume* ***H***5 *and* ***H***7. *Let* $x \in \mathbb{R}^d$, $\gamma > 0$ *and* $(Y_t)_{t\geq 0}$ *be the solution of* (5.7) *started at* x. *For all* $t \in [0, \gamma]$, *we have* $\mathbb{E}\left[\|Y_t - x\|^2\right] \leq t\mathsf{P}_{\gamma,3}(\|x\|)$, *where* $\mathsf{P}_{\gamma,3}$ *is defined in* (4.42).

Proof. The proof is postponed to Section 4.B. □

For $p \in \mathbb{N}$ and $\gamma > 0$, define $\mathsf{Q}_{\gamma,p} : \mathbb{R}_+ \to \mathbb{R}_+$ for all $s \in \mathbb{R}_+$ by,

$$\begin{aligned}
\mathsf{Q}_{\gamma,p}(s) = \left\{ \prod_{i=1}^{p} 2i(d + 3i - 2) \right\} \Bigg[& 2d \frac{\gamma^p}{(p+1)!} + 8L^2(1 + s^{\ell+1}) \\
& \times \left\{ (2 + \sum_{k=1}^{N} a_{k,N} s^{2k}) \frac{\gamma^{p+1}}{(p+2)!} + 2N m a_{0,N} \frac{\gamma^{p+2}}{(p+3)!} \right\} \Bigg] \\
+ 2 \sum_{k=1}^{p} & \left\{ \prod_{i=k+1}^{p} 2i(d + 3i - 2) \right\} \left\{ d + 4 + \frac{L^2(1 + s^{\ell+1})^2}{m(k+1)} \right\} \\
& \times \left\{ \left(\sum_{i=1}^{k} a_{i,k} s^{2i} \right) \frac{\gamma^{p-k}}{(p+1-k)!} + 2k m a_{0,k} \frac{\gamma^{p+1-k}}{(p+2-k)!} \right\}
\end{aligned} \tag{4.43}$$

where N is defined in (4.41).

Lemma 4.12. *Assume* ***H***5 *and* ***H***7. *Let* $p \in \mathbb{N}$, $\gamma > 0$, $x \in \mathbb{R}^d$ *and* $(Y_t)_{t\geq 0}$ *be the solution of* (5.7) *started at* x. *For all* $t \in [0, \gamma]$, *we have* $\mathbb{E}\left[\|Y_t\|^{2p} \|Y_t - x\|^2\right] \leq t\mathsf{Q}_{\gamma,p}(\|x\|)$, *where* $\mathsf{Q}_{\gamma,p}$ *is defined in* (4.43).

Proof. The proof is postponed to Section 4.C. □

Lemma 4.13. *Assume* ***H***8.

a) For all $x \in \mathbb{R}^d$, $\|\nabla^2 U(x)\| \leq C_H\{1 + \|x\|^{\nu+\beta}\}$ *where* $C_H = \max(2L_H, \|\nabla^2 U(0)\|)$.

b) For all $x, y \in \mathbb{R}^d$,

$$\left\|\nabla U(x) - \nabla U(y) - \nabla^2 U(y)(x-y)\right\| \leq \frac{2L_H}{1+\beta}\left\{1 + \|x\|^\nu + \|y\|^\nu\right\}\|x-y\|^{1+\beta} .$$

Proof. a) By **H**8, we get for all $x \in \mathbb{R}^d$

$$\begin{aligned}\left\|\nabla^2 U(x)\right\| &\leq \left\|\nabla^2 U(x) - \nabla^2 U(0)\right\| + \left\|\nabla^2 U(0)\right\| \\ &\leq L_H\left\{1 + \|x\|^\nu\right\}\|x\|^\beta + \left\|\nabla^2 U(0)\right\| .\end{aligned}$$

The proof then follows from the upper bound for all $x \in \mathbb{R}^d$, $\|x\|^\beta \leq 1 + \|x\|^{\nu+\beta}$.

b) Let $x, y \in \mathbb{R}^d$. By **H**8,

$$\begin{aligned}&\left\|\nabla U(x) - \nabla U(y) - \nabla^2 U(y)(x-y)\right\| \\ &\qquad \leq \int_0^1 \left\|\nabla^2 U(tx + (1-t)y) - \nabla^2 U(y)\right\| \mathrm{d}t \, \|x-y\| \\ &\qquad \leq L_H \int_0^1 \left\{1 + \|y\|^\nu + \|tx + (1-t)y\|^\nu\right\}\|t(x-y)\|^\beta \, \mathrm{d}t \, \|x-y\| ,\end{aligned}$$

and the proof follows from $\|tx + (1-t)y\|^\nu \leq \|x\|^\nu + \|y\|^\nu$.

□

For all $n \in \mathbb{N}$, we now bound the Wasserstein distance W_2 between π and the distribution of the n^{th} iterate of X_n defined by (4.3). The strategy consists given two initial conditions (x, y), in coupling X_n and $Y_{\gamma n}$ solution of (5.7) at time γn, using the same Brownian motion. Similarly to (4.30), for $\gamma > 0$, consider the unique strong solution $(Y_t, \overline{Y}_t)_{t\geq 0}$ of

$$\begin{cases} \mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad , \quad Y_0 = y \, , \\ \mathrm{d}\overline{Y}_t = -G_\gamma\left(\overline{Y}_{\lfloor t/\gamma \rfloor \gamma}\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad , \quad \overline{Y}_0 = x \, , \end{cases} \tag{4.44}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Note that for $n \in \mathbb{N}$, $\overline{Y}_{n\gamma} = X_n$ and let $(\mathcal{F}_t)_{t\geq 0}$ be the filtration associated with $(B_t)_{t\geq 0}$.

Lemma 4.14. *Assume* ***A*** *1,* ***A*** *2,* ***H*** *5 and* ***H*** *7. Let* $\gamma_0 > 0$. *Define* $(Y_t)_{t\geq 0}$, $(\overline{Y}_t)_{t\geq 0}$ *by* (4.44). *Then there exists* $C > 0$ *such that for all* $n \in \mathbb{N}$ *and* $\gamma \in (0, \gamma_0]$, *almost surely,*

$$\mathbb{E}^{\mathcal{F}_{n\gamma}}\left[\left\|Y_{(n+1)\gamma} - \overline{Y}_{(n+1)\gamma}\right\|^2\right] \leq \mathrm{e}^{-m\gamma}\left\|Y_{n\gamma} - \overline{Y}_{n\gamma}\right\|^2 + C\gamma^2 V_{\text{æ}}(\overline{Y}_{n\gamma}) .$$

Proof. Using the Markov property, we only need to show the result for $n = 0$. Define for $t \in [0, \gamma)$, $\Theta_t = Y_t - \overline{Y}_t$. By Itô's formula, we have for all $t \in [0, \gamma)$,

$$\|\Theta_t\|^2 = \|y - x\|^2 - 2 \int_0^t \langle \Theta_s, \nabla U(Y_s) - G_\gamma(x) \rangle \, \mathrm{d}s \; .$$

By (4.5) and Lemma 4.9, the family of random variables $(\langle \Theta_s, \nabla U(Y_s) - G_\gamma(x) \rangle)_{s \in [0,\gamma)}$ is uniformly integrable. Pathwise continuity implies then for $s \in [0, \gamma)$ the continuity of $s \mapsto \mathbb{E}\left[\langle \Theta_s, \nabla U(Y_s) - G_\gamma(x) \rangle\right]$. Taking the expectation and deriving, we have for $t \in [0, \gamma)$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}\left[\|\Theta_t\|^2\right] &= -2\mathbb{E}\left[\langle \Theta_t, \nabla U(Y_t) - G_\gamma(x) \rangle\right] \\ &= -2\mathbb{E}\left[\left\langle \Theta_t, \nabla U(Y_t) - \nabla U(\overline{Y}_t) \right\rangle\right] - 2A_1 - 2A_2 \\ &\leq -2m\mathbb{E}\left[\|\Theta_t\|^2\right] - 2A_1 - 2A_2 \; , \end{aligned} \tag{4.45}$$

where

$$A_1 = \mathbb{E}\left[\left\langle \Theta_t, \nabla U(\overline{Y}_t) - \nabla U(x) \right\rangle\right] \; , \quad A_2 = \mathbb{E}\left[\langle \Theta_t, \nabla U(x) - G_\gamma(x) \rangle\right] \; . \tag{4.46}$$

Using that $|\langle a, b \rangle| \leq (m/4) \|a\|^2 + m^{-1} \|b\|^2$ for all $a, b \in \mathbb{R}^d$,

$$|A_1| \leq (m/4)\mathbb{E}\left[\|\Theta_t\|^2\right] + m^{-1}\mathbb{E}\left[\left\|\nabla U(\overline{Y}_t) - \nabla U(x)\right\|^2\right] \; .$$

Similarly to the proof of Lemma 4.8, we have $\mathbb{E}\left[\left\|\nabla U(\overline{Y}_t) - \nabla U(x)\right\|^2\right] \leq t\mathsf{P}_{\gamma,1}(\|x\|)$ where $\mathsf{P}_{\gamma,1}$ is defined in (4.34). For A_2, we have

$$|A_2| \leq (m/4)\mathbb{E}\left[\|\Theta_t\|^2\right] + m^{-1} \|\nabla U(x) - \nabla G_\gamma(x)\|^2 \tag{4.47}$$

and $\|\nabla U(x) - G_\gamma(x)\|^2 \leq \gamma^2 \mathsf{P}_2(\|x\|)$ where P_2 is defined in (4.35). We get for $t \in [0, \gamma)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}\left[\|\Theta_t\|^2\right] \leq -m\mathbb{E}\left[\|\Theta_t\|^2\right] + 2m^{-1}\left\{t\mathsf{P}_{\gamma,1}(\|x\|) + \gamma^2 \mathsf{P}_2(\|x\|)\right\} \; .$$

Using Grönwall's lemma and $1 - \mathrm{e}^{-s} \leq s$ for all $s \geq 0$, we obtain

$$\mathbb{E}\left[\left\|Y_\gamma - \overline{Y}_\gamma\right\|^2\right] \leq \mathrm{e}^{-m\gamma} \|y - x\|^2 + m^{-1}\gamma^2 \left\{\mathsf{P}_{\gamma,1}(\|x\|) + 2\gamma \mathsf{P}_2(\|x\|)\right\} \; .$$

Finally, by (4.34) and (4.35), there exists $C > 0$ such that for all $x \in \mathbb{R}^d$, $\mathsf{P}_{\gamma,1}(\|x\|) + 2\gamma \mathsf{P}_2(\|x\|) \leq C m V_{\text{æ}}(x)$. □

Lemma 4.15. *Assume* ***A*** *1,* ***A*** *2,* ***H*** *7 and* ***H*** *8. Let $\gamma_0 > 0$. Define $(Y_t)_{t \geq 0}$, $(\overline{Y}_t)_{t \geq 0}$ by (4.44). Then there exists $C > 0$ such that for all $n \in \mathbb{N}$ and $\gamma \in (0, \gamma_0]$, almost surely,*

$$\begin{aligned} \mathbb{E}^{\mathcal{F}_{n\gamma}}\left[\left\|Y_{(n+1)\gamma} - \overline{Y}_{(n+1)\gamma}\right\|^2\right] \leq \mathrm{e}^{-m\gamma} \left\|Y_{n\gamma} - \overline{Y}_{n\gamma}\right\|^2 \\ + C\gamma^{2+\beta} V_{\text{æ}}(\overline{Y}_{n\gamma}) + C\gamma^3 V_{\text{æ}}(\overline{Y}_{n\gamma}) \; . \end{aligned}$$

Remark 4.16. *The calculations in the proof show that the dependence w.r.t. $\overline{Y}_{n\gamma}$ and $Y_{n\gamma}$ is in fact polynomial but their exact expressions are very involved. For the sake of simplicity, we bound these polynomials by $V_{\mathfrak{E}}$. The same remark applies equally to Lemma 4.14.*

Proof. Note first that by Lemma 4.13-a), **H**8 implies **H**5 with $L = C_H$ and $\ell = \nu + \beta$. By the Markov property, we only need to show the result for $n = 0$. The proof is a refinement of Lemma 4.14 and we use the same notations. We have to improve the bound on A_1 defined in (4.46). We decompose $A_1 = A_{11} + A_{12}$ where

$$A_{11} = \mathbb{E}\left[\left\langle \Theta_t, \nabla U(\overline{Y}_t) - \nabla U(x) - \nabla^2 U(x)(\overline{Y}_t - x)\right\rangle\right] ,$$
$$A_{12} = \mathbb{E}\left[\left\langle \Theta_t, \nabla^2 U(x)(\overline{Y}_t - x)\right\rangle\right] .$$

Using $|\langle a, b\rangle| \le (m/6)\|a\|^2 + \{3/(2m)\}\|b\|^2$ for all $a, b \in \mathbb{R}^d$,

$$|A_{11}| \le \frac{m}{6}\mathbb{E}\left[\|\Theta_t\|^2\right] + \frac{3}{2m}\mathbb{E}\left[\left\|\nabla U(\overline{Y}_t) - \nabla U(x) - \nabla^2 U(x)(\overline{Y}_t - x)\right\|^2\right] . \tag{4.48}$$

By Lemma 4.13-b),

$$\left\|\nabla U(\overline{Y}_t) - \nabla U(x) - \nabla^2 U(x)(\overline{Y}_t - x)\right\|^2 \le \frac{4L_H^2}{(1+\beta)^2}\left(1 + \|x\|^\nu + \left\|\overline{Y}_t\right\|^\nu\right)^2 \left\|\overline{Y}_t - x\right\|^{2(1+\beta)} .$$

Following the proof of Lemma 4.8, using (4.32) and (4.33), we have

$$\mathbb{E}\left[\left\|\nabla U(\overline{Y}_t) - \nabla U(x) - \nabla^2 U(x)(\overline{Y}_t - x)\right\|^2\right] \le t^{1+\beta}\mathsf{P}_{\gamma,4}(\|x\|) . \tag{4.49}$$

where $\mathsf{P}_{\gamma,4} : \mathbb{R}_+ \to \mathbb{R}_+$ is defined for all $s \in \mathbb{R}_+$ by,

$$\mathsf{P}_{\gamma,4}(s) = \frac{4L_H^2}{(1+\beta)^2}\int_{\mathbb{R}^d}\left[\sqrt{2}\|z\| + \sqrt{\gamma}\left\{2L(1+s^{\ell+1}) + \gamma C_\alpha(1+s^\alpha)\right\}\right]^{2(1+\beta)}$$
$$\times\left[1 + s^\nu + \left\{s + \gamma\left(2L(1+s^{\ell+1}) + \gamma C_\alpha(1+s^\alpha)\right) + \sqrt{2\gamma}\|z\|\right\}^\nu\right]^2 \frac{\mathrm{e}^{-\|z\|^2/2}}{(2\pi)^{d/2}}\mathrm{d}z . \tag{4.50}$$

We decompose A_{12} in $A_{12} = A_{121} + A_{122}$ where

$$A_{121} = \mathbb{E}\left[\left\langle \Theta_t, -t\nabla^2 U(x) G_\gamma(x)\right\rangle\right] , \quad A_{122} = \sqrt{2}\mathbb{E}\left[\left\langle \Theta_t, \nabla^2 U(x) B_t\right\rangle\right] .$$

Define $\mathsf{P}_{\gamma,5} : \mathbb{R}_+ \to \mathbb{R}_+$ for $s \in \mathbb{R}_+$ by,

$$\mathsf{P}_{\gamma,5}(s) = C_H^2\left(1 + s^{\nu+\beta}\right)^2\left\{2L(1+s^{\ell+1}) + \gamma C_\alpha(1+s^\alpha)\right\}^2 . \tag{4.51}$$

By Lemma 4.13-a) and (4.10),

$$|A_{121}| \le (m/6)\mathbb{E}\left[\|\Theta_t\|^2\right] + \{3/(2m)\}t^2\mathsf{P}_{\gamma,5}(\|x\|) . \tag{4.52}$$

By Cauchy-Schwarz inequality and Lemma 4.13-a),

$$|A_{122}| = \sqrt{2}\left|\mathbb{E}\left[\left\langle \int_0^t \{\nabla U(Y_s) - \nabla U(y)\}\,\mathrm{d}s, \nabla^2 U(x) B_t \right\rangle\right]\right|$$
$$\leq \sqrt{2dt} C_H (1 + \|x\|^{\nu+\beta}) \mathbb{E}\left[\left\| \int_0^t \{\nabla U(Y_s) - \nabla U(y)\}\,\mathrm{d}s \right\|^2\right]^{1/2} . \tag{4.53}$$

By **H**5, Cauchy-Schwarz inequality and using $(1+\|y\|^\ell + \|Y_s\|^\ell)^2 \leq 3(2+\|y\|^{2\ell} + \|Y_s\|^{2\lceil \ell \rceil})$ for $s \in [0,\gamma)$, we have

$$\mathbb{E}\left[\left\| \int_0^t \{\nabla U(Y_s) - \nabla U(y)\}\,\mathrm{d}s \right\|^2\right] \leq 3tL^2(2+\|y\|^{2\ell}) \int_0^t \mathbb{E}\left[\|Y_s - y\|^2\right] \mathrm{d}s$$
$$+ 3tL^2 \int_0^t \mathbb{E}\left[\|Y_s\|^{2\lceil \ell \rceil} \|Y_s - y\|^2\right] \mathrm{d}s\,.$$

By Lemmas 4.11 and 4.12, we get

$$\mathbb{E}\left[\left\| \int_0^t \{\nabla U(Y_s) - \nabla U(y)\}\,\mathrm{d}s \right\|^2\right] \leq \frac{3t^3 L^2}{2} \left\{\left(2+\|y\|^{2\ell}\right) \mathsf{P}_{\gamma,3}(\|y\|) + \mathsf{Q}_{\gamma,\lceil \ell \rceil}(\|y\|)\right\}\,,$$

where $\mathsf{P}_{\gamma,3}, \mathsf{Q}_{\gamma,\lceil \ell \rceil} \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}_+, \mathbb{R}_+)$ are defined in (4.42) and (4.43). Plugging this result in (4.53), we obtain

$$|A_{122}| \leq t^2 \sqrt{3d} C_H L \left(1 + \|x\|^{\nu+\beta}\right) \left\{\left(2+\|y\|^{2\ell}\right) \mathsf{P}_{\gamma,3}(\|y\|) + \mathsf{Q}_{\gamma,\lceil \ell \rceil}(\|y\|)\right\}^{1/2} . \tag{4.54}$$

Combining (4.48), (4.49), (4.52) and (4.54), we get

$$|A_1| \leq (m/3)\mathbb{E}\left[\|\Theta_t\|^2\right] + \{3/(2m)\}\left\{t^{1+\beta}\mathsf{P}_{\gamma,4}(\|x\|) + t^2 \mathsf{P}_{\gamma,5}(\|x\|)\right\}$$
$$+ t^2\sqrt{3d} C_H L \left(1 + \|x\|^{\nu+\beta}\right) \left\{\left(2+\|y\|^{2\ell}\right) \mathsf{P}_{\gamma,3}(\|y\|) + \mathsf{Q}_{\gamma,\lceil \ell \rceil}(\|y\|)\right\}^{1/2} ,$$

and by (4.47), $|A_2| \leq (m/6)\mathbb{E}\left[\|\Theta_t\|^2\right] + \{3/(2m)\}\gamma^2 \mathsf{P}_2(\|x\|)$, where $\mathsf{P}_2 \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}_+, \mathbb{R}_+)$ is defined in (4.35). Combining these inequalities in (4.45), we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\|\Theta_t\|^2\right] \leq -m\mathbb{E}\left[\|\Theta_t\|^2\right] + 3m^{-1}\left\{\gamma^2 \mathsf{P}_2(\|x\|) + t^{1+\beta}\mathsf{P}_{\gamma,4}(\|x\|) + t^2 \mathsf{P}_{\gamma,5}(\|x\|)\right\}$$
$$+ 2t^2\sqrt{3d} C_H L \left(1 + \|x\|^{\nu+\beta}\right) \left\{\left(2+\|y\|^{2\ell}\right) \mathsf{P}_{\gamma,3}(\|y\|) + \mathsf{Q}_{\gamma,\lceil \ell \rceil}(\|y\|)\right\}^{1/2} .$$

Using Grönwall's lemma and $1 - \mathrm{e}^{-s} \leq s$ for all $s \geq 0$, we obtain

$$\mathbb{E}\left[\left\| Y_\gamma - \overline{Y}_\gamma \right\|^2\right] \leq \mathrm{e}^{-m\gamma} \|y - x\|^2$$
$$+ 3m^{-1}\left\{\gamma^3 \mathsf{P}_2(\|x\|) + \frac{\gamma^{2+\beta}}{2+\beta}\mathsf{P}_{\gamma,4}(\|x\|) + \frac{\gamma^3}{3}\mathsf{P}_{\gamma,5}(\|x\|)\right\}$$
$$+ 2\gamma^3\sqrt{d/3} C_H L \left(1 + \|x\|^{\nu+\beta}\right) \left\{\left(2+\|y\|^{2\ell}\right) \mathsf{P}_{\gamma,3}(\|y\|) + \mathsf{Q}_{\gamma,\lceil \ell \rceil}(\|y\|)\right\}^{1/2} .$$

Finally, by (4.35), (4.42), (4.50), (4.51) and (4.43), there exists $C > 0$ such that for all $x \in \mathbb{R}^d$ and $\gamma \in (0, \gamma_0]$,

$$
\begin{aligned}
&3m^{-1}\left\{\gamma^3 \mathsf{P}_2(\|x\|) + \frac{\gamma^{2+\beta}}{2+\beta}\mathsf{P}_{\gamma,4}(\|x\|) + \frac{\gamma^3}{3}\mathsf{P}_{\gamma,5}(\|x\|)\right\} \leq C\gamma^{2+\beta} V_{\text{æ}}(x) \;, \\
&2\sqrt{d/3}C_H L\left(1 + \|x\|^{\nu+\beta}\right) \leq C^{1/2} V_{\text{æ}}(x)^{1/2} \;, \\
&\left(2 + \|x\|^{2\ell}\right)\mathsf{P}_{\gamma,3}(\|x\|) + \mathsf{Q}_{\gamma,\lceil \ell \rceil}(\|x\|) \leq C V_{\text{æ}}(x) \;.
\end{aligned}
$$

□

Proof of Theorem 4.5. Let $\gamma \in (0, \gamma_0]$. Define $(Y_t)_{t\geq 0}$, $(\overline{Y}_t)_{t\geq 0}$ by (4.44) and $X_n = \overline{Y}_{n\gamma}$ for $n \in \mathbb{N}$. By Lemma 4.14 and Proposition 4.3, we have for all $n \in \mathbb{N}$,

$$
\begin{aligned}
&\mathbb{E}\left[\|Y_{n\gamma} - X_n\|^2\right] \leq \mathrm{e}^{-nm\gamma}\|y - x\|^2 + C\gamma^2 \sum_{k=0}^{n-1} \mathrm{e}^{-m\gamma(n-1-k)}\mathbb{E}\left[V_{\text{æ}}(X_k)\right] \\
&\leq \mathrm{e}^{-nm\gamma}\|y - x\|^2 + \frac{C\gamma^2}{1 - \mathrm{e}^{-m\gamma}}\frac{b}{\text{æ}^2}\mathrm{e}^{\text{æ}^2\gamma} + C\gamma^2 V_{\text{æ}}(x)\sum_{k=0}^{n-1}\mathrm{e}^{-m\gamma(n-1-k)}\mathrm{e}^{-\text{æ}^2\gamma k} \;. \qquad (4.55)
\end{aligned}
$$

Note that

$$
\sum_{k=0}^{n-1}\mathrm{e}^{-m\gamma(n-1-k)}\mathrm{e}^{-\text{æ}^2\gamma k} \leq \frac{n}{1 - \max(\mathrm{e}^{-m}, \mathrm{e}^{-\text{æ}^2})\gamma}
$$

and $1 - s^\gamma \geq -\gamma\log(s)\mathrm{e}^{\gamma\log(s)}$ for $s \in (0,1)$. In eq. (4.55), integrating y with respect to π, for all $n \in \mathbb{N}$, $(Y_{n\gamma}, X_n)$ is a coupling between π and $\delta_x R_\gamma^n$. By Lemma 4.10, we get (4.16). By Proposition 4.3 and [Vil09, Corollary 6.11], we have for all $x \in \mathbb{R}^d$, $\lim_{n\to+\infty} W_2(\delta_x R_\gamma^n, \pi) = W_2(\pi_\gamma, \pi)$ and we obtain (4.17). □

Proof of Theorem 4.6. Let $\gamma \in (0, \gamma_0]$. Define $(Y_t)_{t\geq 0}$, $(\overline{Y}_t)_{t\geq 0}$ by (4.44) and $X_n = \overline{Y}_{n\gamma}$ for $n \in \mathbb{N}$. By Lemma 4.15, we have for all $n \in \mathbb{N}$,

$$
\mathbb{E}\left[\|Y_{n\gamma} - X_n\|^2\right] \leq \mathrm{e}^{-nm\gamma}\|y - x\|^2 + A_n + B_n \;,
$$

where

$$
\begin{aligned}
A_n &= C\gamma^{2+\beta}\sum_{k=0}^{n-1}\mathrm{e}^{-m\gamma(n-1-k)}\mathbb{E}\left[V_{\text{æ}}(X_k)\right] \;, \\
B_n &= C\gamma^3\sum_{k=0}^{n-1}\mathrm{e}^{-m\gamma(n-1-k)}\mathbb{E}\left[V_{\text{æ}}(Y_{k\gamma})\right] \;.
\end{aligned}
$$

Analysis similar to the proof of Theorem 4.5 using Proposition 4.1 instead of Proposition 4.3 for B_n shows then the result. □

4.4.6 Proof of Theorem 4.7

We first state a lemma on the existence and regularity of a solution of the Poisson equation (4.20) which is adapted from [PV01, Theorem 1].

Lemma 4.17. *Assume **H**6 and **H**9. Let $f \in \mathrm{C}^3(\mathbb{R}^d, \mathbb{R})$ be such that $\left\|\mathrm{D}^i f\right\| \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}_+)$ for $i \in \{0, \dots, 3\}$. Then, there exists a solution of the Poisson equation (4.20) $\phi \in \mathrm{C}^4(\mathbb{R}^d, \mathbb{R})$, such that $\left\|\mathrm{D}^i \phi\right\| \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}_+)$ for $i \in \{0, \dots, 4\}$.*

Proof. The proof is postponed to Section 4.D. □

Proof of Theorem 4.7. The proof is adapted from [MST10, Section 5.1] Let $\gamma \in (0, \gamma_0]$. In this Section, C is a positive constant which can change from line to line but does not depend on γ. For $k \in \mathbb{N}$, denote by

$$\delta_{k+1} = X_{k+1} - X_k = -\gamma G_\gamma(X_k) + \sqrt{2\gamma} Z_{k+1} \; .$$

By **H** 6, **H** 9 and Lemma 4.17, there exists a solution to the Poisson equation (4.20) $\phi \in \mathrm{C}^4(\mathbb{R}^d, \mathbb{R})$, such that for all $x \in \mathbb{R}^d$ and $i \in \{0, \dots, 4\}$,

$$\mathcal{L}\phi(x) = -\left(f(x) - \pi(f)\right) \quad \text{and} \quad \left\|\mathrm{D}^i \phi\right\| \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}_+) \; . \tag{4.56}$$

By Taylor's formula, we have for $k \in \mathbb{N}$,

$$\begin{aligned}
\phi(X_{k+1}) = \phi(X_k) + \mathrm{D}\,\phi(X_k)[\delta_{k+1}] + (1/2)\,\mathrm{D}^2\,\phi(X_k)[\delta_{k+1}, \delta_{k+1}] \\
+ (1/6)\,\mathrm{D}^3\,\phi(X_k)[\delta_{k+1}, \delta_{k+1}, \delta_{k+1}] + r_k \; , \\
r_k = (1/6) \int_0^1 (1-s)^3\,\mathrm{D}^4\,\phi(X_k + s\delta_{k+1})[\delta_{k+1}, \delta_{k+1}, \delta_{k+1}, \delta_{k+1}]\mathrm{d}s \; .
\end{aligned}$$

Using the expression of δ_{k+1} and (4.6), we get

$$\begin{aligned}
\phi(X_{k+1}) = \phi(X_k) + \gamma\mathcal{L}\phi(X_k) + \sqrt{2\gamma}\,\mathrm{D}\,\phi(X_k)[Z_{k+1}] \\
+ \gamma\left\{\mathrm{D}^2\,\phi(X_k)[Z_{k+1}, Z_{k+1}] - \Delta\phi(X_k)\right\} + \gamma\,\mathrm{D}\,\phi(X_k)[\nabla U(X_k) - G_\gamma(X_k)] \\
+ (\gamma^2/2)\,\mathrm{D}^2\,\phi(X_k)[G_\gamma(X_k), G_\gamma(X_k)] - \sqrt{2}\gamma^{3/2}\,\mathrm{D}^2\,\phi(X_k)[G_\gamma(X_k), Z_{k+1}] \\
+ (1/6)\,\mathrm{D}^3\,\phi(X_k)[\delta_{k+1}, \delta_{k+1}, \delta_{k+1}] + r_k \; .
\end{aligned}$$

Summing from $k = 0$ to $n-1$ for $n \in \mathbb{N}^\star$, dividing by $n\gamma$, we get

$$\frac{1}{n}\sum_{k=0}^{n-1}(f(X_k) - \pi(f)) = \frac{\phi(X_0) - \phi(X_n)}{n\gamma} + \frac{1}{n\gamma}\left(\sum_{i=0}^{3} M_{i,n} + \sum_{i=0}^{3} S_{i,n}\right) \; ,$$

where

$$M_{0,n} = ((\sqrt{2}\gamma^{3/2})/6) \sum_{k=0}^{n-1} \{2\,\mathrm{D}^3\,\phi(X_k)[Z_{k+1}, Z_{k+1}, Z_{k+1}] + 3\gamma\,\mathrm{D}^3\,\phi(X_k)[G_\gamma(X_k), G_\gamma(X_k), Z_{k+1}]\} ,$$

$$M_{1,n} = \gamma \sum_{k=0}^{n-1} (\mathrm{D}^2\,\phi(X_k)[Z_{k+1}, Z_{k+1}] - \Delta\phi(X_k)) ,$$

$$M_{2,n} = \sqrt{2\gamma} \sum_{k=0}^{n-1} \mathrm{D}\,\phi(X_k)[Z_{k+1}] ,$$

$$M_{3,n} = -\sqrt{2}\gamma^{3/2} \sum_{k=0}^{n-1} \mathrm{D}^2\,\phi(X_k)[G_\gamma(X_k), Z_{k+1}] ,$$

and

$$S_{0,n} = -(\gamma^2/6) \sum_{k=0}^{n-1} \{6\,\mathrm{D}^3\,\phi(X_k)[G_\gamma(X_k), Z_{k+1}, Z_{k+1}] + \gamma\,\mathrm{D}^3\,\phi(X_k)[G_\gamma(X_k), G_\gamma(X_k), G_\gamma(X_k)]\} ,$$

$$S_{1,n} = \gamma \sum_{k=0}^{n-1} \mathrm{D}\,\phi(X_k)[\nabla U(X_k) - G_\gamma(X_k)] ,$$

$$S_{2,n} = (\gamma^2/2) \sum_{k=0}^{n-1} \mathrm{D}^2\,\phi(X_k)[G_\gamma(X_k), G_\gamma(X_k)] ,$$

$$S_{3,n} = \sum_{k=0}^{n-1} r_k .$$

By **A** 1, we calculate for $n \in \mathbb{N}^*$, $|S_{1,n}| \le \gamma^2 C_\alpha \sum_{k=0}^{n-1} \|\mathrm{D}\,\phi(X_k)\|\,(1 + \|X_k\|^\alpha)$. By **H** 9, (4.10) and (4.56), there exist $p, q \ge 1$ and $C_q > 0$ such that the summands of $(M_{i,n})_{n\in\mathbb{N}}$ and $(S_{i,n})_{n\in\mathbb{N}}$ for $i \in \{0, \dots, 3\}$ are dominated by $C_q\,(1 + \|X_k\|^q)\,(1 + \|Z_{k+1}\|^p)$ for $k \in \{0, \dots, n-1\}$. Therefore, by Proposition 4.1, for $i \in \{0, \dots, 3\}$, $(M_{i,n})_{n\in\mathbb{N}}$ are martingales and for $n \in \mathbb{N}^*$, $\mathbb{E}\left[S_{i,n}^2\right] \le Cn^2\gamma^4$,

$$\mathbb{E}\left[M_{0,n}^2\right] \le Cn\gamma^3 , \ \mathbb{E}\left[M_{1,n}^2\right] \le Cn\gamma^2 , \ \mathbb{E}\left[M_{2,n}^2\right] \le Cn\gamma , \ \mathbb{E}\left[M_{3,n}^2\right] \le Cn\gamma^3 ,$$

which yield the result. □

Acknowledgements

This work was supported by the École Polytechnique Data Science Initiative and the Alan Turing Institute under the EPSRC grant EP/N510129/1.

4.A Proof of Lemma 4.9

By **H** 7, (5.7) has a unique strong solution $(Y_t)_{t\geq 0}$ for any initial data $Y_0 = x \in \mathbb{R}^d$. Define for $p \in \mathbb{N}^*$, $V_p : \mathbb{R}^d \to \mathbb{R}_+$ by $V_p(y) = \|y\|^{2p}$ for $y \in \mathbb{R}^d$. We have using **H**7,

$$\mathcal{L}V_p(x) = -2p \|x\|^{2(p-1)} \langle \nabla U(x), x\rangle + 2p(d + 2(p-1)) \|x\|^{2(p-1)} \tag{4.57}$$

$$\leq -2pm \|x\|^{2p} + 2p \|x\|^{2(p-1)} (d + 2(p-1)) . \tag{4.58}$$

Applying [MT93, Theorem 1.1] with $V(x,t) = V_p(x)\mathrm{e}^{2pmt}$, $g_-(t) = 0$ and $g_+(x,t) = 2p(d+2(p-1))V_{p-1}(x)\mathrm{e}^{2pmt}$ for $x \in \mathbb{R}^d$ and $t \geq 0$, we get denoting by $v_p(t,x) = P_t V_p(x)$,

$$v_p(t,x) \leq \mathrm{e}^{-2pmt} V_p(x) + 2p(d + 2(p-1)) \int_0^t \mathrm{e}^{-2pm(t-s)} v_{p-1}(s,x)\mathrm{d}s .$$

A straightforward induction concludes the proof.

4.B Proof of Lemma 4.11

Define $\tilde{V}_x : \mathbb{R}^d \to \mathbb{R}_+$ for all $y \in \mathbb{R}^d$ by $\tilde{V}_x(y) = \|y - x\|^2$. By Lemma 4.9, the process $(\tilde{V}_x(Y_t) - \tilde{V}_x(x) - \int_0^t \mathcal{L}\tilde{V}_x(Y_s)\mathrm{d}s)_{t\geq 0}$, is a $(\mathcal{F}_t)_{t\geq 0}$-martingale. Denote for all $t \geq 0$ and $y \in \mathbb{R}^d$ by $\tilde{v}(t,x) = P_t\tilde{V}_x(x)$. Then we get,

$$\frac{\partial \tilde{v}(t,x)}{\partial t} = P_t \mathcal{L}\tilde{V}_x(x) . \tag{4.59}$$

By **H**7, we have for all $y \in \mathbb{R}^d$,

$$\mathcal{L}\tilde{V}_x(y) = 2\left(-\langle \nabla U(y), y - x\rangle + d\right) \leq 2\left(-m\tilde{V}_x(y) + d - \langle \nabla U(x), y - x\rangle\right) . \tag{4.60}$$

Using (4.59), this inequality and that $\tilde{V}_x$ is nonnegative, we get

$$\frac{\partial \tilde{v}(t,x)}{\partial t} = P_t \mathcal{L}\tilde{V}_x(x) \leq 2\left(d - \int_{\mathbb{R}^d} \langle \nabla U(x), y - x\rangle P_t(x, \mathrm{d}y)\right) . \tag{4.61}$$

Using (4.5) and (5.7), we have

$$\begin{aligned} |\mathbb{E}_x\left[\langle \nabla U(x), Y_t - x\rangle\right]| &\leq \|\nabla U(x)\| \, \|\mathbb{E}_x\left[Y_t - x\right]\| \\ &\leq \|\nabla U(x)\| \left\|\mathbb{E}_x\left[\int_0^t \{\nabla U(Y_s)\} \, \mathrm{d}s\right]\right\| \\ &\leq 2L\left\{1 + \|x\|^{\ell+1}\right\} \int_0^t \mathbb{E}_x\left[\|\nabla U(Y_s)\|\right] \mathrm{d}s . \end{aligned} \tag{4.62}$$

Using (4.5) again,

$$\begin{aligned} \int_0^t \mathbb{E}_x\left[\|\nabla U(Y_s)\|\right] \mathrm{d}s &\leq 2L \int_0^t \mathbb{E}\left[1 + \|Y_s\|^{\ell+1}\right] \mathrm{d}s \\ &\leq 2L\left\{2t + \int_0^t \mathbb{E}\left[\|Y_s\|^{2N}\right] \mathrm{d}s\right\} . \end{aligned} \tag{4.63}$$

Furthermore using that for all $s \geq 0$, $1 - \mathrm{e}^{-s} \leq s$, $s + \mathrm{e}^{-s} - 1 \leq s^2/2$, and Lemma 4.9 we get

$$\int_0^t \mathbb{E}_x\left[\|Y_s\|^{2N}\right] \mathrm{d}s \leq a_{0,N} \frac{2Ntm + \mathrm{e}^{-2Nmt} - 1}{2Nm} + \sum_{k=1}^N a_{k,N} \|x\|^{2k} \frac{1 - \mathrm{e}^{-2mkt}}{2km}$$
$$\leq t^2 N m a_{0,N} + t \sum_{k=1}^N a_{k,N} \|x\|^{2k} \ .$$

Plugging this inequality in (4.63) and (4.62), we get

$$|\mathbb{E}_x\left[\langle \nabla U(x), Y_t - x\rangle\right]| \leq 4L^2(1 + \|x\|^{\ell+1}) \left\{2t + Nma_{0,N}t^2 + t\sum_{k=1}^N a_{k,N}\|x\|^{2k}\right\} \ . \quad (4.64)$$

Using this bound in (4.61) and integrating the inequality gives

$$\tilde{v}(t,x) \leq 2dt + 8L^2(1 + \|x\|^{\ell+1}) \left\{t^2 + Nma_{0,N}(t^3/3) + (t^2/2)\sum_{k=1}^N a_{k,N}\|x\|^{2k}\right\} \ . \quad (4.65)$$

4.C Proof of Lemma 4.12

We show the result by induction on p. The case $p = 0$ follows from (4.65). Suppose $p \geq 1$. Define for $y \in \mathbb{R}^d$, $W_{x,p} : \mathbb{R}^d \to \mathbb{R}_+$ by $W_{x,p}(y) = \|y\|^{2p} \|y - x\|^2$. We have

$$\mathcal{L}W_{x,p}(y) = -2\|y\|^{2p}\langle \nabla U(y), y - x\rangle - (2p)\|y\|^{2(p-1)}\|y - x\|^2 \langle \nabla U(y), y\rangle$$
$$+ 2\|y\|^{2(p-1)}\left\{d\|y\|^2 + 4p\langle y, y - x\rangle + p(d + 2p - 2)\|y - x\|^2\right\} \ .$$

By **H**7, (4.5) and using $|\langle a, b\rangle| \leq \eta\|a\|^2 + (4\eta)^{-1}\|b\|^2$ for all $\eta > 0$, we have

$$\mathcal{L}W_{x,p}(y) \leq \frac{\|y\|^{2p}\|\nabla U(x)\|^2}{2m(p+1)} + 2\|y\|^{2(p-1)}\left\{(d+4)\|y\|^2 + p(d + 3p - 2)\|y - x\|^2\right\}$$
$$\leq \|y\|^{2p}\left\{2(d+4) + \frac{2L^2(1 + \|x\|^{\ell+1})^2}{m(p+1)}\right\} + 2p(d + 3p - 2)\|y - x\|^2\|y\|^{2(p-1)} \ . \quad (4.66)$$

By Lemma 4.9, the process $(W_{x,p}(Y_t) - W_{x,p}(x) - \int_0^t \mathcal{L}W_{x,p}(Y_s)\mathrm{d}s)_{t\geq 0}$ is a $(\mathcal{F}_t)_{t\geq 0}$-martingale. For $x \in \mathbb{R}^d$ and $t \geq 0$, denote by $w_{x,p}(x,t) = P_t W_{x,p}(x)$ and $v_p(x,t) = \mathbb{E}_x\left[\|Y_t\|^{2p}\right]$. Taking the expectation of (4.66) w.r.t. $\delta_x P_t$ and integrating w.r.t. t, we get

$$w_{x,p}(t,x) \leq 2\left\{d + 4 + \frac{L^2(1 + \|x\|^{\ell+1})^2}{m(p+1)}\right\}\int_0^t v_p(s,x)\mathrm{d}s$$
$$+ 2p(d + 3p - 2)\int_0^t w_{x,p-1}(s,x)\mathrm{d}s \ .$$

By Lemma 4.9, $v_p(t,x) \leq 2pma_{0,p}t + \sum_{k=1}^p a_{k,p}\|x\|^{2k}$. A straightforward induction concludes the proof.

4.D Proof of Lemma 4.17

The proof is adapted from [PV01, Theorem 1] and follows the same steps. Define $\bar{f} = f - \pi(f)$. Note that **H**9 implies **H**5. By **H**6, [SV07, Corollary 11.1.5], $(P_t)_{t\geq 0}$ is Feller continuous, which implies that for all $t > 0$, if $(x_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{R}^d$ converging to $x \in \mathbb{R}^d$, then $\delta_{x_n} P_t$ weakly converges to $\delta_x P_t$. Therefore, for all $t > 0$ and $K > 0$, $x \mapsto P_t(f \vee (-K) \wedge K)(x)$ is continuous. By Cauchy-Schwarz and Markov's inequalities, for all $t, K > 0$ and $x \in \mathbb{R}^d$, we have

$$\begin{aligned} |P_t(f \vee (-K) \wedge K)(x) - P_t f(x)| &\leq P_t(|f| \mathbb{1}\{|f| \geq K\})(x) \\ &\leq P_t f^2(x)/K \end{aligned}$$

By Proposition 4.1 and the polynomial growth of f, we get for all $R > 0$,

$$\lim_{K\to+\infty} \sup_{\|x\|\leq R} |P_t(f \vee (-K) \wedge K)(x) - P_t(f)(x)| = 0$$

and therefore $x \mapsto P_t\bar{f}(x)$ is continuous for all $t > 0$.

By (4.57) and [DFG09, Theorem 3.10, Section 4.1], there exist $C, \varsigma > 0$ and $p \in \mathbb{N}$ such that for all $x \in \mathbb{R}^d$ and $N > 0$,

$$\int_N^{+\infty} \left| P_t\bar{f}(x) \right| \mathrm{d}t \leq C\left(1 + \|x\|^p\right) N^{-\varsigma} .$$

Therefore, we may define $\phi(x) = \int_0^{+\infty} P_t\bar{f}(x)\mathrm{d}t$ for all $x \in \mathbb{R}^d$. Denote by $\phi_N = \int_0^N P_t\bar{f}(x)\mathrm{d}t$ for all $N > 0$ and $x \in \mathbb{R}^d$. We have $\lim_{N\to+\infty} \phi_N(x) = \phi(x)$ locally uniformly in x and by continuity of ϕ_N for all $N > 0$, $\phi \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$.

Let $x \in \mathbb{R}^d$ and consider the Dirichlet problem,

$$\mathcal{L}\hat{\phi}(y) = -\bar{f}(y) \quad \text{for} \quad y \in \mathrm{B}(x,1) \quad \text{and} \quad \hat{\phi}(y) = \phi(y) \quad \text{for} \quad y \in \partial\,\mathrm{B}(x,1) ,$$

where $\partial\,\mathrm{B}(x,1) = \overline{B}(x,1) \setminus \mathrm{B}(x,1)$. By [GT15, Lemma 6.10, Theorem 6.17], there exists a solution $\hat{\phi} \in \mathrm{C}^4(\mathrm{B}(x,1),\mathbb{R}) \cap \mathrm{C}(\overline{B}(x,1),\mathbb{R})$. Let $\tilde{x} \in \overline{B}(x,1/2)$. By **H**6, (5.7) has a unique strong solution denoted $(Y_t^{\tilde{x}})_{t\geq 0}$ starting at $Y_0 = \tilde{x}$. Define the stopping time $\tau = \inf\{t \geq 0 : Y_t^{\tilde{x}} \notin \mathrm{B}(x,1)\}$. By [Fri12, Volume I, Chapter 6, Theorem 5.1], we have

$$\hat{\phi}(\tilde{x}) = \mathbb{E}\left[\phi(Y_\tau^{\tilde{x}})\right] + \mathbb{E}\left[\int_0^\tau \bar{f}(Y_t^{\tilde{x}})\mathrm{d}t\right] .$$

For all $N > 0$, we decompose $\phi_N(\tilde{x}) = A_N + B_N$ where

$$A_N = \int_0^N \mathbb{E}\left[\bar{f}(Y_t^{\tilde{x}})\mathbb{1}\{t \leq \tau\}\right] \mathrm{d}t \quad , \quad B_N = \int_0^N \mathbb{E}\left[\bar{f}(Y_t^{\tilde{x}})\mathbb{1}\{t > \tau\}\right] \mathrm{d}t .$$

Since $\mathbb{E}[\tau] < +\infty$ by [Fri12, Volume I, Chapter 6, equation (5.11)],

$$\mathbb{E}\left[\int_0^{+\infty} \left|\bar{f}(Y_t^{\tilde{x}})\right| \mathbb{1}\{t \leq \tau\} \,\mathrm{d}t\right] < +\infty ,$$

and by Fubini's theorem and the dominated convergence theorem, $\lim_{N\to+\infty} A_N = \mathbb{E}\left[\int_0^\tau \bar{f}(Y_t^{\tilde{x}})\mathrm{d}t\right]$. We also have

$$B_N = \mathbb{E}\left[\int_0^{(N-\tau)_+} \bar{f}(Y_{\tau+t}^{\tilde{x}})\mathrm{d}t\right] = \mathbb{E}\left[\phi_{(N-\tau)_+}(Y_\tau^{\tilde{x}})\right] .$$

Since $\mathbb{E}[\tau] < +\infty$, we have $\lim_{N\to+\infty} \phi_{(N-\tau)_+}(Y_\tau^{\tilde{x}}) = \phi(Y_\tau^{\tilde{x}})$ almost surely. Besides, there exist $C, p > 0$ such that $\phi_{\tilde{N}}(Y_\tau^{\tilde{x}}) \leq C(1 + \|x\|^p)$ almost surely and for all $\tilde{N} \geq 0$ because $Y_\tau^{\tilde{x}} \in \overline{B}(x, 1)$ and $\phi_{\tilde{N}}$ converges locally uniformly to ϕ. By the dominated convergence theorem, we get $\lim_{N\to+\infty} B_N = \mathbb{E}\left[\phi(Y_\tau^{\tilde{x}})\right]$. Taking the limit $N \to +\infty$ of $\phi_N(\tilde{x}) = A_N + B_N$, we obtain $\phi(\tilde{x}) = \hat{\phi}(\tilde{x})$.

Finally, by [GT15, Problem 6.1 (a)], we obtain $\left\|\mathrm{D}^i \phi\right\| \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}_+)$ for $i \in \{0, \ldots, 4\}$ which concludes the proof.

4.E Badly conditioned multivariate Gaussian variable

In this example, we consider a badly conditioned multivariate Gaussian variable in dimension $d = 100$, of mean 0 and covariance matrix $\mathrm{diag}(10^{-5}, 1 \ldots, 1)$. We run 100 independent simulations of ULA and TULAc, starting at 0, with a step size $\gamma \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ and a number of iterations equal to 10^6. ULA diverges for all step sizes. We plot the boxplots of the errors for TULAc, for the first and second moment of the first and last coordinate in Figure 4.5. Although the results for the first coordinate are expectedly inaccurate, the results for the last coordinate are valid. In this context, TULAc enables to obtain relevant results for the well-conditioned coordinates within a relatively small number of iterations, which is not possible using ULA.

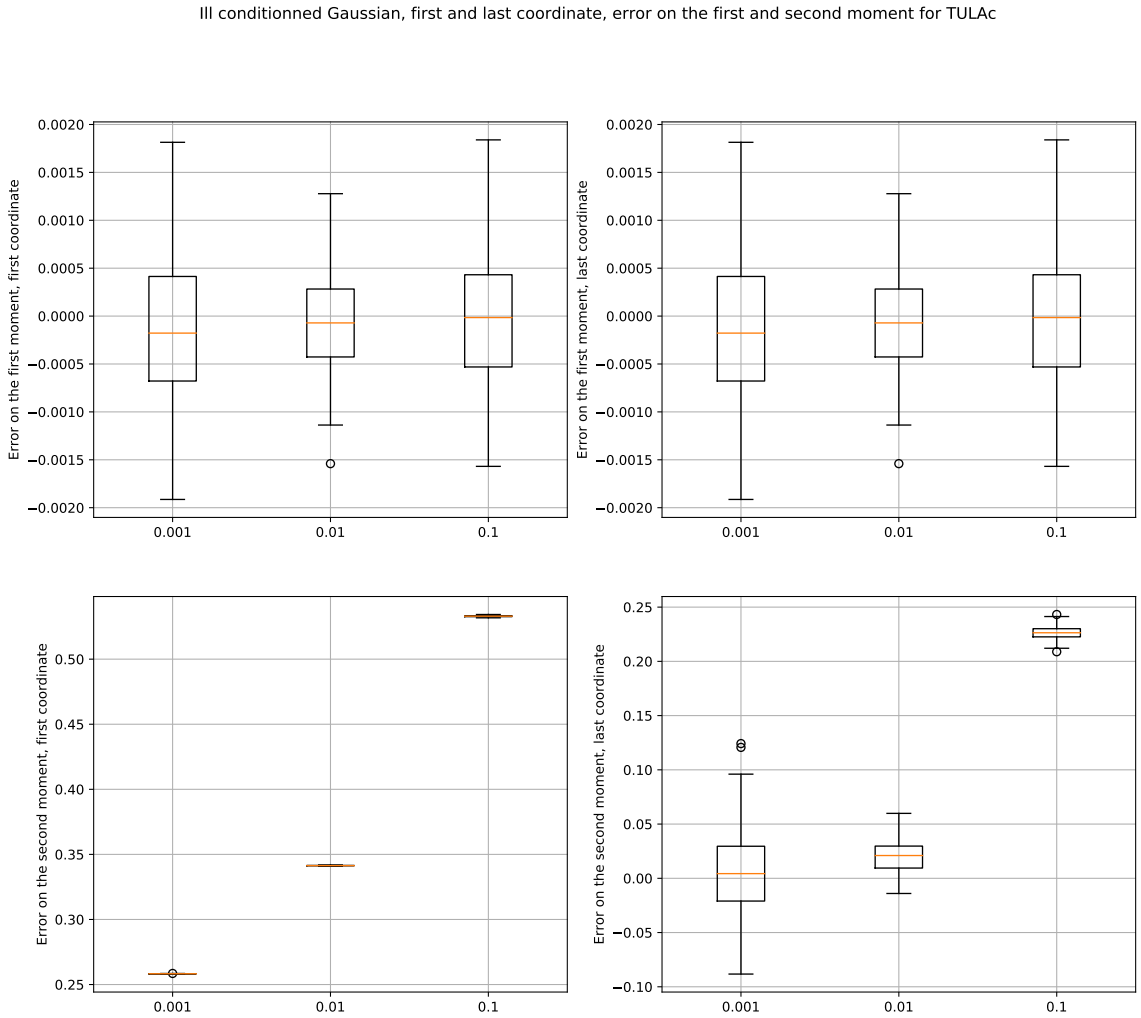

Figure 4.5: Boxplots of the error for TULAc on the first and second moments for the badly conditioned Gaussian variable in dimension 100 starting at 0 for different step sizes.

Part II

Applications of the unadjusted Langevin algorithm

Chapter 5

Normalizing constants of log-concave densities

Nicolas Brosse [1], Alain Durmus [2], Éric Moulines [3]

Abstract

We derive explicit bounds for the computation of normalizing constants Z for log-concave densities $\pi = \mathrm{e}^{-U}/Z$ w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Our approach relies on a Gaussian annealing combined with recent and precise bounds on the Unadjusted Langevin Algorithm [DM16]. Polynomial bounds in the dimension d are obtained with an exponent that depends on the assumptions made on U. The algorithm also provides a theoretically grounded choice of the annealing sequence of variances. A numerical experiment supports our findings. Results of independent interest on the mean squared error of the empirical average of locally Lipschitz functions are established.

5.1 Introduction

Let $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable convex function such that $Z = \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\mathrm{d}x < +\infty$. Z is the normalizing constant of the probability density π associated with the potential U, defined for $x \in \mathbb{R}^d$ by $\pi(x) = Z^{-1}\mathrm{e}^{-U(x)}$. We discuss in this paper a method to estimate Z with polynomial complexity in the dimension d.

Computing the normalizing constant is a challenge which has applications in Bayesian inference and statistical physics in particular. In statistical physics, Z is better known

[1]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France. nicolas.brosse@polytechnique.edu

[2]Ecole Normale Supérieure CMLA 61, Av. du Président Wilson 94235 Cachan Cedex, France Email: alain.durmus@cmla.ens-cachan.fr

[3]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France. eric.moulines@polytechnique.edu

under the name of partition function or free energy [Bal07], [LSR10]. Free energy differences allow to quantify the relative likelihood of different states (microscopic configurations) and are linked to thermodynamic work and heat exchanges. In Bayesian inference, the models can be compared by the computation of the Bayes factor which is the ratio of two normalizing constants (see e.g. [Rob07, chapter 7]). This problem has consequently attracted a wealth of contribution; see for example [CSI00, chapter 5], [MR09], [FW12], [Ard+12], [D+13], [Knu+15], [ZJA15] and, for a more specific molecular simulations flavor, [LSR10]. Compared to the large number of proposed methods to estimate Z, few theoretical guarantees have been obtained on the output of these algorithms; see below for further references and comments. Our algorithm relies on a sequence of Gaussian densities with increasing variances, combined with the precise bounds of [DM16].

The paper is organized as follows. The outline of the algorithm is first described, followed by the assumptions made on U. Our main results are then stated and compared to previous works on the subject. The theoretical analysis of the algorithm is detailed in Section 5.2. In Section 5.3, a numerical experiment is provided to support our theoretical claims. Finally, the proofs are gathered in Section 5.5. In Section 5.4, a result of independent interest on the mean squared error of the empirical average of locally Lipschitz functions is established.

Notations and conventions

Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and f a μ-integrable function, denote by $\mu(f)$ the integral of f w.r.t. μ. We say that ζ is a transference plan of μ and ν if it is a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for all measurable sets A of $\mathbb{R}^d$, $\zeta(\mathrm{A} \times \mathbb{R}^d) = \mu(\mathrm{A})$ and $\zeta(\mathbb{R}^d \times \mathrm{A}) = \nu(\mathrm{A})$. We denote by $\Pi(\mu, \nu)$ the set of transference plans of μ and ν. Furthermore, we say that a couple of $\mathbb{R}^d$-random variables (X, Y) is a coupling of μ and ν if there exists $\zeta \in \Pi(\mu, \nu)$ such that (X, Y) are distributed according to ζ. For two probability measures μ and ν, we define the Wasserstein distance of order $p \geq 1$ as

$$W_p(\mu, \nu) \stackrel{\text{def}}{=} \left(\inf_{\zeta \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} \|x - y\|^p \, \mathrm{d}\zeta(x, y) \right)^{1/p} . \tag{5.1}$$

By [Vil09, Theorem 4.1], for all μ, ν probability measure on $\mathbb{R}^d$, there exists a transference plan $\zeta^\star \in \Pi(\mu, \nu)$ such that the infimum in (5.1) is reached in $\zeta^\star$. $\zeta^\star$ is called an optimal transference plan associated with W_p.

$f : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz function if there exists $C \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $|f(x) - f(y)| \leq C \|x - y\|$. Then we denote

$$\operatorname{Lip} f = \sup\{|f(x) - f(y)| \, \|x - y\|^{-1} \mid x, y \in \mathbb{R}^d, x \neq y\} .$$

For $k \in \mathbb{N}$, $\mathcal{C}^k(\mathbb{R}^d)$ denotes the set of k-continuously differentiable functions $\mathbb{R}^d \to \mathbb{R}$, with the convention that $\mathcal{C}^0(\mathbb{R}^d)$ is the set of continuous functions. Let $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$, $n, m \in \mathbb{N}^*$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ be a twice continuously differentiable function. Denote by

∇F and $\nabla^2 F$ the Jacobian and the Hessian of F respectively. For $m = 1$, the Laplacian is defined by $\Delta F = \operatorname{Tr} \nabla^2 F$ where Tr is the trace operator. In the sequel, we take the convention that for $n, p \in \mathbb{N}$, $n < p$ then $\sum_p^n = 0$ and $\prod_p^n = 1$. By convention, $\inf\{\emptyset\} = +\infty$, $\sup\{\emptyset\} = -\infty$ and for $j > i$ in $\mathbb{Z}$, $\{j, \dots, i\} = \emptyset$. For a finite set E, $|E|$ denotes the cardinality of E. For $a, b \in \mathbb{R}$, $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Let $\psi, \phi : \mathbb{R}_+^\star \to \mathbb{R}_+^\star$. We write $\psi = \tilde{\mathcal{O}}(\phi)$ if there exists $t_0 > 0$, $C, c > 0$ such that $\psi(t) \leq C\phi(t) \left|\log t\right|^c$ for all $t \in (0, t_0]$. Denote by $\mathrm{B}(x, r) = \left\{y \in \mathbb{R}^d : \|y - x\| \leq r\right\}$.

Presentation of the algorithm

Since $Z < +\infty$ and U is convex, by [Bra+14, Lemma 2.2.1], there exist constants $\rho_1 > 0$ and $\rho_2 \in \mathbb{R}$ such that $U(x) \geq \rho_1 \|x\| - \rho_2$. Therefore, by continuity, U has a minimum $x^\star$. Without loss of generality, it is assumed in the sequel that $x^\star = 0$ and $U(x^\star) = 0$.

Let $M \in \mathbb{N}^\star$, $\{\sigma_i^2\}_{i=0}^M$ be a positive increasing sequence of real numbers and set $\sigma_M^2 = +\infty$. Consider the sequence of functions $\{U_i\}_{i=0}^M$ defined for all $i \in \{0, \dots, M\}$ and $x \in \mathbb{R}^d$ by

$$U_i(x) = \frac{\|x\|^2}{2\sigma_i^2} + U(x) \;, \tag{5.2}$$

with the convention $1/\infty = 0$. We define a sequence of probability densities $\{\pi_i\}_{i=0}^M$ for $i \in \{0, \dots, M\}$ and $x \in \mathbb{R}^d$ by

$$\pi_i(x) = Z_i^{-1} \mathrm{e}^{-U_i(x)} \;, \qquad Z_i = \int_{\mathbb{R}^d} \mathrm{e}^{-U_i(y)} \mathrm{d}y \;. \tag{5.3}$$

The dependence of Z_i in σ_i^2 is implicit. By definition, note that $U_M = U$, $Z_M = Z$ and $\pi_M = \pi$. As in the multistage sampling method [GM98, Section 3.3], we use the following decomposition

$$\frac{Z}{Z_0} = \prod_{i=0}^{M-1} \frac{Z_{i+1}}{Z_i} \;. \tag{5.4}$$

Z_0 is estimated by choosing σ_0^2 small enough so that π_0 is sufficiently close to a Gaussian distribution of mean 0 and covariance $\sigma_0^2 \operatorname{Id}$. For $i \in \{0, \dots, M-1\}$, the ratio Z_{i+1}/Z_i may be expressed as

$$\frac{Z_{i+1}}{Z_i} = \int_{\mathbb{R}^d} g_i(x) \pi_i(x) \mathrm{d}x = \pi_i(g_i) \;, \tag{5.5}$$

where $g_i : \mathbb{R}^d \to \mathbb{R}_+$ is defined for all $x \in \mathbb{R}^d$ by

$$g_i(x) = \exp\left(a_i \|x\|^2\right) \;, \qquad a_i = \frac{1}{2}\left(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_{i+1}^2}\right) \;. \tag{5.6}$$

The quantity $\pi_i(g_i)$ is estimated by the Unadjusted Langevin Algorithm (ULA) targeting π_i. Introduced in [Erm75] and [Par81] (see also [RT96]), the ULA algorithm can be described as follows. For $i \in \{0, \dots, M-1\}$, the (overdamped) Langevin stochastic differential equation (SDE) is given by

$$\mathrm{d}Y_{i,t} = -\nabla U_i(Y_{i,t})\mathrm{d}t + \sqrt{2}\mathrm{d}B_{i,t} \;, \quad Y_{i,0} = 0 \;, \tag{5.7}$$

where $\{(B_{i,t})_{t\geq 0}\}_{i=0}^{M-1}$ are independent d-dimensional Brownian motions. The sampling method is based on the Euler discretization of the Langevin diffusion, which defines a discrete-time Markov chain, for $i \in \{0, \dots, M-1\}$ and $k \in \mathbb{N}$

$$X_{i,k+1} = X_{i,k} - \gamma_i \nabla U_i(X_{i,k}) + \sqrt{2\gamma_i} W_{i,k+1} \,, \quad X_{i,0} = 0 \,, \tag{5.8}$$

where $\{(W_{i,k})_{k\in\mathbb{N}^\star}\}_{i=0}^{M-1}$ are independent i.i.d. sequences of standard Gaussian random variables and $\gamma_i > 0$ is the stepsize. For $i \in \{0, \dots, M-1\}$, consider the following estimator of Z_{i+1}/Z_i,

$$\hat{\pi}_i(g_i) = \frac{1}{n_i} \sum_{k=N_i+1}^{N_i+n_i} g_i(X_{i,k}) \,, \tag{5.9}$$

where $n_i \geq 1$ is the sample size and $N_i \geq 0$ the burn-in period. We introduce the following assumptions on U.

H 10. *$U : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and L-gradient Lipschitz, i.e. there exists $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla U(x) - \nabla U(y)\| \leq L \|x - y\| \,. \tag{5.10}$$

H 11 (m)**.** *$U : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and satisfies for all $x, y \in \mathbb{R}^d$,*

$$U(y) \geq U(x) + \langle \nabla U(x), y - x \rangle + (m/2) \|x - y\|^2 \,. \tag{5.11}$$

H 12. *The function U is three times continuously differentiable and there exists $\tilde{L} \geq 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\left\| \nabla^2 U(x) - \nabla^2 U(y) \right\| \leq \tilde{L} \|x - y\| \,. \tag{5.12}$$

The strongly convex case (**H** 11(m) with $m > 0$) is considered in Section 5.2.1 and the convex case (**H** 11(m) with $m = 0$) is dealt with in Section 5.2.2. Assuming **H** 10 and **H** 11(m) for $m \geq 0$, for $i \in \{0, \dots, M\}$, U_i defined in (5.2) is L_i-gradient Lipschitz and m_i-strongly convex if $m_i > 0$ (and convex if $m_i = 0$) where

$$L_i = L + \frac{1}{\sigma_i^2} \,, \quad m_i = m + \frac{1}{\sigma_i^2} \,. \tag{5.13}$$

Define also the following useful quantities,

$$\kappa = \frac{2mL}{m+L} \,, \quad \kappa_i = \frac{2m_i L_i}{m_i + L_i} \,. \tag{5.14}$$

H 12 enables to have tighter bounds on the mean squared error of $\hat{\pi}_i(g_i)$ defined in (5.9). Under **H** 12, for all $i \in \{0, \dots, M\}$, U_i satisfies (5.12) with $\tilde{L}$. Finally, since $Z < +\infty$ and by [Bra+14, Lemma 2.2.1], there exist $\rho_1 > 0$ and $\rho_2 \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$,

$$U(x) \geq \rho_1 \|x\| - \rho_2 \,. \tag{5.15}$$

Denote by $\mathcal{S}$ the set of simulation parameters,

$$\mathcal{S} = \left\{ M, \{\sigma_i^2\}_{i=0}^{M-1}, \{\gamma_i\}_{i=0}^{M-1}, \{n_i\}_{i=0}^{M-1}, \{N_i\}_{i=0}^{M-1} \right\} , \tag{5.16}$$

and by $\hat{Z}$ the following estimator of Z,

$$\hat{Z} = (2\pi\sigma_0^2)^{d/2}(1+\sigma_0^2 m)^{-d/2} \left\{ \prod_{i=0}^{M-1} \hat{\pi}_i(g_i) \right\} , \tag{5.17}$$

where $\hat{\pi}_i(g_i)$ is defined in (5.9). The dependence of $\hat{Z}$ in $\mathcal{S}$ is implicit. Note that $\hat{Z}$ is a biased estimator of Z because Z_0 is approximated by $(2\pi\sigma_0^2)^{d/2}(1+\sigma_0^2 m)^{-d/2}$. We define the cost of the algorithm by the total number of iterations performed by the M Markov chains $(X_{i,n})_{n\geq 0}$ for $i \in \{0, \dots, M-1\}$, i.e.

$$\text{cost} = \sum_{i=0}^{M-1} \{N_i + n_i\} . \tag{5.18}$$

Observe that each step of the Markov chain takes time linear in d. We state below a simplified version of our results; explicit bounds are given in Theorems 5.5, 5.6, 5.12 and 5.13.

Theorem 5.1. *Assume* ***H***10, ***H***11(m) *for* $m \geq 0$. *Let* $\mu, \epsilon \in (0,1)$. *There exists an explicit choice of the simulation parameters* $\mathcal{S}$ *such that the estimator* $\hat{Z}$ *defined in* (5.17) *satisfies*

$$\mathbb{P}\left(\left| \hat{Z}/Z - 1 \right| > \epsilon \right) \leq \mu . \tag{5.19}$$

Moreover, the cost *of the algorithm* (5.18) *is upper-bounded by,*

	H10,***H***11(m) *for* $m > 0$
cost	$\frac{L^3}{\mu^2 m^3} \log(d) d^3 \times \tilde{\mathcal{O}}(\epsilon^{-4})$
	H10,***H***11(m) *for* $m > 0$,***H***12
cost	$\left(\frac{\tilde{L}}{\mu^{3/2} m^{3/2}} + \frac{L^2}{\mu^{3/2} m^2} \right) \log(d) d^{5/2} \times \tilde{\mathcal{O}}(\epsilon^{-3})$
	H10,***H***11(m) *for* $m \geq 0$
cost	$\frac{L^2}{\mu^2 \rho_1^4} (d+\rho_2)^4 \log(d) d^3 \times \tilde{\mathcal{O}}(\epsilon^{-4})$
	H10,***H***11(m) *for* $m \geq 0$,***H***12
cost	$\left(\frac{L^2}{\mu^{3/2} \rho_1^4} + \frac{\tilde{L}}{\mu^{3/2}(d+\rho_2)\rho_1^3} \right) (d+\rho_2)^4 \log(d) d^{5/2} \times \tilde{\mathcal{O}}(\epsilon^{-3})$

By the median trick (see e.g. [JVV86, Lemma 6.1] or [NP09]), the dependence in μ of the cost can be reduced to a logarithmic factor, see Corollaries 5.7 and 5.14.

It is interesting to compare these complexity bounds with previously reported results. In [MDJ06] and [Bes+14] (see also [Del04]), the authors propose to use sequential Monte Carlo (SMC) samplers to estimate the normalizing constant Z of a probability distribution π. In [Bes+14], π is supported on a compact set K included in $\mathbb{R}^d$ and

satisfies for $x = (x_1, \dots, x_d) \in \mathsf{K}$, $\pi(x) = Z^{-1} \prod_{i=1}^{d} \exp(g(x_i))$. [Bes+14, Theorem 3.2] states that there exists an estimator $\hat{Z}$ of Z such that $\lim_{d\to+\infty} \mathbb{E}[|\hat{Z}/Z - 1|^2] = C/N$ where N is the number of particles and C depends on g and on the parameters of the SMC (choice of the Markov kernel and of the annealing schedule). With our definition (5.18), the computational cost of the SMC algorithm is $\mathcal{O}(Nd)$ (there are d phases and N particles for each phase). To obtain an estimator $\hat{Z}$ satisfying (5.19) implies a cost of $d\mu^{-1}\mathcal{O}(\epsilon^{-2})$. However, the product form of the density π is restrictive, the result is only asymptotic in d and the state space is assumed to be compact. [Del+16] combines SMC with a multilevel approach and [Jas+16] establishes results on a multilevel particle filter.

[Hub15] deals with the case where $\pi(x) = \exp(-\beta H(x))/Z(\beta)$ where $x \in \Omega$, a finite state space, $\beta \geq 0$ and $H(x) \in \{0, \dots, n\}$. These distributions known as Gibbs distributions include in particular the Ising model. To compute $Z(\beta)$, [Hub15] relies on an annealing process on the parameter β, starting from $Z(0)$. Let $q = \log(Z(0))/\log(Z(\beta))$. [Hub15, Theorem 1.1] states that there exists an estimator $\hat{Z}(\beta)$ of $Z(\beta)$ such that (5.19) is satisfied with $\mu = 1/4$ and $q\log(n)\tilde{\mathcal{O}}(\epsilon^{-2})$ draws from the Gibbs distribution.

Our complexity results can also be related to the computation of the volume of a convex body K (compact convex set with non-empty interior) on $\mathbb{R}^d$. This problem has attracted a lot of attention in the field of computer science, starting with the breakthrough of [DF91] until the most recent results of [CV15b]. Define for $x \in \mathbb{R}^d$, $\pi(x) = 1_{\mathsf{K}}(x)/\operatorname{Vol}(\mathsf{K})$. Under the assumptions $\mathrm{B}(0,1) \subset \mathsf{K}$ and $\int_{\mathbb{R}^d} \|x\|^2 \pi(x)\mathrm{d}x = \mathcal{O}(d)$, [CV15b, Theorem 1.1] states that there exists an estimator $\hat{Z}$ of $Z = \operatorname{Vol}(\mathsf{K})$ such that (5.19) is satisfied with $\mu = 1/4$ and a cost of $\log(d)d^3\tilde{\mathcal{O}}(\epsilon^{-2})$.

Nonequilibrium methods have been recently developed and studied in order to compute free energy differences or Bayes factors, see [Jar97] and [LSR10, Chapter 4]. They are based on an inhomogeneous diffusion evolving (for example) from $t = 0$ to $t = 1$ such that π_0 and π_1 are the stationary distributions respectively for $t = 0$ and $t = 1$. Recently, [ARW16] provided an asymptotic and non-asymptotic analysis of the bias and variance for estimators associated with this methodology. The main aim of this paper is to obtain polynomial complexity and inspection of their results suggests a cost of order d^{15} at most to compute an estimator $\hat{Z}$ satisfying (5.19). However, this cost may be due to the strategy of proofs.

Multistage sampling type algorithms are widely used and known under different names: multistage sampling [VC72], (extended) bridge sampling [GM98], annealed importance sampling (AIS) [Nea01], thermodynamic integration [OPG16], power posterior [BFH12]. For the stability and accuracy of the method, the choice of the parameters (in our case $\{\sigma_i^2\}_{i=0}^{M-1}$) is crucial and is known to be difficult. Indeed, the issue has been pointed out in several articles under the names of tuning tempered transitions [BFH12], temperature placement [FHW14], annealing sequence [Bes+14, Sections 3.2.1, 4.1], temperature ladder [OPG16, Section 3.3.2], effects of grid size [D+13], cooling schedule [CV15b]. An approach based on large deviation has also been suggested to derive an optimal choice of the annealing schedule, see [DD09]. In Sections 5.2.1 and 5.2.2, we explicitly define the sequence $\{\sigma_i^2\}_{i=0}^{M-1}$.

5.2 Theoretical analysis of the algorithm

In this Section, we analyse the algorithm outlined in Section 5.1. The strongly convex and convex cases are considered in Sections 5.2.1 and 5.2.2, respectively. The choice of the simulation parameters $\mathcal{S}$ explicitly depends on the (strong) convexity of U. Throughout this Section, we assume that $L > m$; note that if $L = m$, π is a Gaussian density and Z is known. For $M \in \mathbb{N}^\star$ and $i \in \{0, \dots, M-1\}$, we first provide an upper bound on the mean squared error MSE_i of $\hat{\pi}_i(g_i)$ defined by

$$\mathrm{MSE}_i = \mathbb{E}\left[\{\hat{\pi}_i(g_i) - \pi_i(g_i)\}^2\right] , \tag{5.20}$$

where $\pi_i(g_i)$ and $\hat{\pi}_i(g_i)$ are given by (5.5) and (5.9) respectively. The MSE_i can be decomposed as a sum of the squared bias and variance,

$$\mathrm{MSE}_i = \{\mathbb{E}[\hat{\pi}_i(g_i)] - \pi_i(g_i)\}^2 + \mathrm{Var}\left[\hat{\pi}_i(g_i)\right] . \tag{5.21}$$

Propositions 5.2 and 5.3 give upper bounds on the squared bias and Proposition 5.4 on the variance. The results are based on the non-asymptotic bounds of the Wasserstein distance for a strongly convex potential obtained in [DM16] (see also [Dal17b], [DM17]). We introduce the following conditions on the stepsize γ_i used in the Euler discretization and the variance σ_{i+1}^2

$$\gamma_i \in \left(0, \frac{1}{m + L + 2/\sigma_i^2}\right] , \quad \sigma_{i+1}^2 \leq 2(d+4)\left(\frac{2d+7}{\sigma_i^2} - m\right)_+^{-1} , \tag{5.22}$$

where by convention $1/0 = +\infty$. Note that the condition on σ_{i+1}^2 is equivalent to $a_i \in [0, m_i/\{4(d+4)\} \wedge (2\sigma_i^2)^{-1}]$ where a_i is defined in (5.6) and m_i in (5.13). Assuming that γ_i and σ_{i+1}^2 satisfy (5.22), we define the positive quantities

$$C_{i,0} = \exp\left(\frac{4a_i(d+2)}{\kappa_i - 8a_i}\right) , \quad C_{i,1} = 2d\frac{1 - 8a_i\gamma_i}{\kappa_i - 8a_i} , \quad C_{i,2} = 4\frac{d}{m_i} , \tag{5.23}$$

where m_i, L_i and κ_i are defined in (5.13) and (5.14), respectively. Denote by,

$$A_{i,0} = 2L_i^2\kappa_i^{-1}d , \tag{5.24}$$

$$A_{i,1} = 2dL_i^2 + dL_i^4(\kappa_i^{-1} + (m_i + L_i)^{-1})(m_i^{-1} + 6^{-1}(m_i + L_i)^{-1}) . \tag{5.25}$$

Proposition 5.2. *Assume* ***H****10 and* ***H****11(m) for some* $m \geq 0$. *For* $N_i \in \mathbb{N}$, $n_i \in \mathbb{N}^*$ *and* γ_i, σ_{i+1}^2 *satisfying* (5.22), *we have*

$$\{\mathbb{E}[\hat{\pi}_i(g_i)] - \pi_i(g_i)\}^2 \leq 4a_i^2(C_{i,2} + C_{i,0}C_{i,1}) \times \left\{\frac{4d}{n_i m_i \kappa_i \gamma_i}\exp\left(-N_i\frac{\kappa_i\gamma_i}{2}\right) + 2\kappa_i^{-1}(A_{i,0}\gamma_i + A_{i,1}\gamma_i^2)\right\} .$$

Proof. The proof is postponed to Section 5.5.1. □

The first term $\exp(-N_i\kappa_i\gamma_i/2)$ is the exponential forgetting of the initial condition and the second term proportional to γ_i is the stationary term. The squared bias can thus be controlled by adjusting the parameters γ_i, n_i and N_i. If U satisfies **H**12, the bound on the squared bias can be improved. Define,

$$B_{i,0} = d\left(2L_i^2 + \kappa_i^{-1}\{(d\tilde{L}^2)/3 + 4L_i^4/(3m_i)\}\right) , \tag{5.26}$$

$$B_{i,1} = dL_i^4\left(\kappa_i^{-1} + \{6(m_i + L_i)\}^{-1} + m_i^{-1}\right) . \tag{5.27}$$

Proposition 5.3. *Assume **H**10, **H**11(m) for some $m \geq 0$, and **H**12. For $N_i \in \mathbb{N}$, $n_i \in \mathbb{N}^*$ and γ_i, σ_{i+1}^2 satisfying (5.22), we have*

$$\{\mathbb{E}[\hat{\pi}_i(g_i)] - \pi_i(g_i)\}^2 \leq 4a_i^2(C_{i,2} + C_{i,0}C_{i,1}) \times \left\{\frac{4d}{n_i m_i \kappa_i \gamma_i}\exp\left(-N_i\frac{\kappa_i\gamma_i}{2}\right) + 2\kappa_i^{-1}(B_{i,0}\gamma_i^2 + B_{i,1}\gamma_i^3)\right\} .$$

Proof. The proof is postponed to Section 5.5.1. □

Note that the leading term is of order γ_i^2 instead of γ_i. We consider now the variance term in (5.21).

Proposition 5.4. *Assume **H**10 and **H**11(m) for some $m \geq 0$. For $N_i \in \mathbb{N}$, $n_i \in \mathbb{N}^*$ and γ_i, σ_{i+1}^2 satisfying (5.22), we have*

$$\mathrm{Var}\left[\hat{\pi}_i(g_i)\right] \leq \frac{32a_i^2 C_{i,0}C_{i,1}}{\kappa_i^2 n_i \gamma_i}\left(1 + \frac{2}{\kappa_i n_i \gamma_i}\right) .$$

Proof. The proof is postponed to Section 5.5.1. □

5.2.1 Strongly convex potential U

Theorem 5.5. *Assume **H**10 and **H**11(m) for $m > 0$ and let $\mu, \epsilon \in (0,1)$. There exists an explicit choice of the simulation parameters $\mathcal{S}$ (5.16) such that the estimator $\hat{Z}$ defined in (5.17) satisfies with probability at least $1-\mu$*

$$(1-\epsilon)Z \leq \hat{Z} \leq (1+\epsilon)Z ,$$

and the cost *(5.18) of the algorithm is upper-bounded by*

$$\text{cost} \leq \left(\frac{6272C}{\epsilon^2\mu} + \log(5Cd^2)\right)\frac{(1088C)^2 d^2(d+4)}{\epsilon^2\mu}\left(\frac{m+L}{2m}\right)^3 (C+3) , \tag{5.28}$$

with

$$C = \left\lceil \frac{1}{\log(2)}\log\left(d\left(d+\frac{7}{2}\right)\left(\frac{L}{m}-1\right)\frac{1}{\log(1+\epsilon/3)}\right)\right\rceil . \tag{5.29}$$

Proof. The proof is postponed to Section 5.5.3. □

Theorem 5.6. *Assume* ***H*** *10,* ***H*** *11(m) for* $m > 0$, ***H*** *12 and let* $\mu, \epsilon \in (0,1)$. *There exists an explicit choice of the simulation parameters* $\mathcal{S}$ (5.16) *such that the estimator* $\hat{Z}$ *defined in* (5.17) *satisfies with probability at least* $1-\mu$

$$(1-\epsilon)Z \leq \hat{Z} \leq (1+\epsilon)Z \ ,$$

and the cost (5.18) *of the algorithm is upper-bounded by*

$$\text{cost} \leq \left(\frac{6272C}{\epsilon^2\mu} + \log(5Cd^2)\right)\sqrt{\frac{7}{3}}\frac{512Cd^{3/2}}{\epsilon\sqrt{\mu}}(d+4)(C+3) \times \left\{\tilde{L}\frac{2^{3/2}}{m^{3/2}} + \sqrt{10}\left(\frac{m+L}{2m}\right)^2\right\} \ , \quad (5.30)$$

with C *defined in* (5.29).

Proof. The proof is postponed to Section 5.5.3. □

The dependence of the upper bound with respect to d is improved from d^3 to $d^{5/2}$. Using the median trick (see e.g. [JVV86, Lemma 6.1] or [NP09]), we have the following corollary,

Corollary 5.7. *Let* $\epsilon, \tilde{\mu} \in (0,1)$. *Repeat* $2\lceil 4\log(\tilde{\mu}^{-1})\rceil + 1$ *times the algorithm of Theorems 5.5 and 5.6 with* $\mu = 1/4$ *and denote by* $\hat{Z}$ *the median of the output values. We have with probability at least* $1-\tilde{\mu}$,

$$(1-\epsilon)Z \leq \hat{Z} \leq (1+\epsilon)Z \ .$$

Proof. The proof is postponed to Section 5.5.3. □

The proof of Theorems 5.5 and 5.6 and Corollary 5.7 rely on several lemmas which are stated below. These lemmas explain how the simulation parameters $\mathcal{S}$ must be chosen. The details of the proofs are gathered in Section 5.5.3. Set

$$\sigma_0^2 = \{2\log(1+\epsilon/3)\}/\{d(L-m)\} \ . \quad (5.31)$$

This choice of σ_0^2 is justified by the following result,

Lemma 5.8. *Under* ***H*** *10 and* ***H*** *11(m) for* $m \geq 0$, *we have*

$$Z_0 \leq (2\pi\sigma_0^2)^{d/2}/(1+\sigma_0^2 m)^{d/2} \leq Z_0\,(1+\epsilon/3) \ . \quad (5.32)$$

Proof. The proof is postponed to Section 5.5.3. □

Given a choice of $\mathcal{S}$, define the event

$$\mathrm{A}_{\mathcal{S},\epsilon} = \left\{\left|\prod_{i=0}^{M-1}\hat{\pi}_i(g_i) - \prod_{i=0}^{M-1}\pi_i(g_i)\right| \leq \prod_{i=0}^{M-1}\pi_i(g_i)\frac{\epsilon}{2}\right\} \ . \quad (5.33)$$

On $\mathrm{A}_{\mathcal{S},\epsilon}$, using Lemma 5.8, (5.4) and (5.17), we have:

$$Z\,(1-\epsilon/2) \leq \hat{Z} \leq Z\,(1+\epsilon)\ .$$

It remains to choose $\mathcal{S}$ to minimize approximately the cost defined in (5.18) under the constraint $\mathbb{P}(\mathrm{A}_{\mathcal{S},\epsilon}) \geq 1-\mu$. We define the positive increasing sequence $\{\sigma_i^2\}_{i=0}^{M-1}$ recursively, starting from $i=0$. For $i \in \mathbb{N}$, set

$$\sigma_{i+1}^2 = \varsigma_s(\sigma_i^2)\ , \tag{5.34}$$

where $\varsigma_s : \mathbb{R}_+^\star \to \mathbb{R}$ is defined for $t \in (0, (2d+7)/m)$ by

$$\varsigma_s(t) = \left(\frac{1}{t} - \frac{m + (2^{k(t)+1}\sigma_0^2)^{-1}}{2(d+4)}\right)^{-1}\ , \quad k(t) = \left\lfloor \frac{\log(t/\sigma_0^2)}{\log(2)} \right\rfloor \tag{5.35}$$

and $\varsigma_s(t) = +\infty$ otherwise. The subscript s in ς_s stresses that this choice is valid for the strongly convex case and will be different for the convex case. With this choice of $(\sigma_i^2)_{i\geq 0}$, the number of phases M is defined by

$$M = \inf\left\{i \geq 1 : \sigma_{i-1}^2 \geq (2d+7)/m\right\}\ . \tag{5.36}$$

By (5.35), for $t \in [\sigma_0^2, (2d+7)/m)$, $\varsigma_s(t) \geq t(4d+16)/(4d+15)$, which implies $M < +\infty$. With this definition of ς_s, for $i \in \{0,\dots,M-2\}$, we have

$$a_i = \frac{1}{2}\left(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_{i+1}^2}\right) = \frac{m + (2^{k+1}\sigma_0^2)^{-1}}{4(d+4)}\ , \quad \text{if} \quad 2^k\sigma_0^2 \leq \sigma_i^2 < 2^{k+1}\sigma_0^2\ , \tag{5.37}$$

and $a_{M-1} = (2\sigma_{M-1}^2)^{-1}$. Define $\mathcal{I}_k \subset \mathbb{N}$ for $k \in \mathbb{N}$ and $K \in \mathbb{N}$ by,

$$\mathcal{I}_k = \left\{i \in \{0,\dots,M-2\} : 2^k\sigma_0^2 \leq \sigma_i^2 < 2^{k+1}\sigma_0^2\right\}\ , \tag{5.38}$$

$$K = \inf\{k \geq 0 : \mathcal{I}_k = \emptyset\} < +\infty\ . \tag{5.39}$$

The number of phases M and variances $\{\sigma_i^2\}_{i=0}^{M-1}$ being defined, we now proceed with the choice of the stepsize γ_i, the number of samples n_i and the burn-in period N_i for $i \in \{0,\dots,M-1\}$.

Lemma 5.9. *Set $\eta = (\epsilon\sqrt{\mu})/8$. Assume that there exists a choice of the simulation parameters $\{N_i\}_{i=0}^{M-1}$, $\{n_i\}_{i=0}^{M-1}$ and $\{\gamma_i\}_{i=0}^{M-1}$ satisfying,*

i) For all $k \in \{0,\dots,K-1\}$, $i \in \mathcal{I}_k$,

$$|\mathbb{E}\left[\hat{\pi}_i(g_i)\right] - \pi_i(g_i)| \leq \frac{\eta}{K\,|\mathcal{I}_k|}, \quad \mathrm{Var}\left[\hat{\pi}_i(g_i)\right] \leq \frac{\eta^2}{K\,|\mathcal{I}_k|}\ ,$$

ii) $|\mathbb{E}\left[\hat{\pi}_{M-1}(g_{M-1})\right] - \pi_{M-1}(g_{M-1})| \leq \eta, \quad \mathrm{Var}\left[\hat{\pi}_{M-1}(g_{M-1})\right] \leq \eta^2\ ,$

where $\pi_i(g_i)$ is defined in (5.5) and $\hat{\pi}_i(g_i)$ in (5.9). Then $\mathbb{P}(\mathrm{A}_{\mathcal{S},\epsilon}) \geq 1-\mu$, where $\mathrm{A}_{\mathcal{S},\epsilon}$ is defined in (5.33).

Proof. The proof is postponed to Section 5.5.2 □

To show the existence of γ_i, n_i, N_i satisfying the conditions of Lemma 5.9 , we apply Propositions 5.2 to 5.4 for each $i \in \{0, \dots, M-1\}$. We then have the following lemmas,

Lemma 5.10. *Set $\eta = (\epsilon\sqrt{\mu})/8$. Assume **H**10, **H**11(m) for $m > 0$ and,*

i) for all $k \in \{0, \dots, K-1\}$, $i \in \mathcal{I}_k$,

$$\gamma_i \leq \frac{1}{2285} \frac{\eta^2 \kappa_i^2 \sigma_i^4 m_i}{K^2 d^2 L_i^2} \leq \frac{1}{m_i + L_i} , \tag{5.40}$$

$$n_i \geq \frac{196K}{\eta^2} \frac{\sqrt{m_i}}{\kappa_i \sigma_i} \frac{1}{\kappa_i \gamma_i} , \tag{5.41}$$

$$N_i \geq 2(\kappa_i \gamma_i)^{-1} \log\left(5Kd^2\right) , \tag{5.42}$$

ii)

$$\gamma_{M-1} \leq 40^{-1} \eta^2 L_{M-1}^{-2} m_{M-1} \leq (m_{M-1} + L_{M-1})^{-1} , \tag{5.43}$$

$$n_{M-1} \geq 19(\kappa_{M-1}\gamma_{M-1})^{-1} \eta^{-2} , \tag{5.44}$$

$$N_{M-1} \geq (\kappa_{M-1}\gamma_{M-1})^{-1} . \tag{5.45}$$

Then, the conditions i)-ii) of Lemma 5.9 are satisfied.

Proof. The proof is postponed to Section 5.5.3. □

We have a similar result under the additional assumption **H**12.

Lemma 5.11. *Set $\eta = (\epsilon\sqrt{\mu})/8$. Assume **H**10, **H**11(m) for $m > 0$, **H**12 and,*

i) for all $k \in \{0, \dots, K-1\}$, $i \in \mathcal{I}_k$,

$$\gamma_i \leq \sqrt{\frac{3}{7}} \frac{\eta \kappa_i m_i^{1/2} \sigma_i^2}{8Kd} \left(d\tilde{L}^2 + 10 L_i^4 m_i^{-1}\right)^{-1/2} \leq \frac{1}{m_i + L_i} , \tag{5.46}$$

and n_i, N_i as in (5.41), (5.42),

ii)

$$\gamma_{M-1} \leq \sqrt{\frac{3}{7}} \frac{\eta \kappa_{M-1} m_{M-1}^{-1/2}}{4} \left(d\tilde{L}^2 + 10 L_{M-1}^4 m_{M-1}^{-1}\right)^{-1/2} \leq \frac{1}{m_{M-1} + L_{M-1}} , \tag{5.47}$$

and n_{M-1}, N_{M-1} as in (5.44), (5.45).

Then, the conditions i)-ii) of Lemma 5.9 are satisfied.

Proof. The proof is postponed to Section 5.A.2. □

5.2.2 Convex potential U

We now consider the convex case. The annealing process on the variances $\{\sigma_i^2\}_{i=0}^{M-1}$ is different from the strongly convex case and is defined in (5.54). In particular, the stopping criteria for the annealing process is distinct from the case where U is strongly convex and relies on a truncation argument. More precisely, a concentration theorem for log-concave functions [Per16, Theorem 3.1] states that for $\alpha \in (0,1)$,

$$\int_{\mathbb{R}^d} \mathbb{1}_{\{U \geq d(\tau_\alpha+1)\}}(x)\pi(x)\mathrm{d}x \leq \alpha \ , \quad \tau_\alpha = \left(\frac{16\log(3/\alpha)}{d}\right)^{1/2} \ .$$

Let $\epsilon \in (0,1)$, $\tau = \tau_{\epsilon/2}$ and $D = \rho_1^{-1}\{d(\tau+1)+\rho_2\}$. By (5.15), we have

$$\int_{\mathbb{R}^d} \mathbb{1}_{\mathrm{B}(0,D)}(x)\pi(x)\mathrm{d}x \geq 1-\epsilon/2 \ . \tag{5.48}$$

Given a choice of M and σ_{M-1}^2, define $\bar{g}_{M-1} : \mathbb{R}^d \to \mathbb{R}_+$ for all $x \in \mathbb{R}^d$ by

$$\bar{g}_{M-1}(x) = \exp\left\{\frac{1}{2\sigma_{M-1}^2}(\|x\|^2 \wedge D^2)\right\} \ , \tag{5.49}$$

and J by,

$$J = \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\mathrm{d}x \Big/ \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)-(\|x\|^2-D^2)_+/(2\sigma_{M-1}^2)}\mathrm{d}x \ .$$

Note that $Z/Z_{M-1} = J \times \pi_{M-1}(\bar{g}_{M-1})$ and by (5.48),

$$J(1-\epsilon/2) \leq 1 \leq J \ . \tag{5.50}$$

On the event $\mathrm{A}_{\mathcal{S},\epsilon}$ defined in (5.33) with g_{M-1} replaced by $\bar{g}_{M-1}$ and by (5.32) (with $m=0$), (5.50), we get

$$Z\,(1-\epsilon/2)^2 \leq \hat{Z} \leq Z\,(1+\epsilon) \ ,$$

where $\hat{Z}$ is defined in (5.17) with g_{M-1} replaced by $\bar{g}_{M-1}$. We now state our results in the convex case.

Theorem 5.12. *Assume **H**10, **H**11(m) for $m \geq 0$. Let $\epsilon, \mu \in (0,1)$. There exists an explicit choice of the simulation parameters $\mathcal{S}$ (5.16) such that the estimator $\hat{Z}$ defined in (5.17) (with g_{M-1} replaced by $\bar{g}_{M-1}$ defined in (5.49)) satisfies with probability at least $1-\mu$*

$$(1-\epsilon)Z \leq \hat{Z} \leq (1+\epsilon)Z \ ,$$

and the cost (5.18) *of the algorithm is upper-bounded by*

$$\mathrm{cost} \leq \left(\frac{17728C}{\epsilon^2\mu} + \log(Cd^2)\right)\frac{(487C)^2d^2(d+4)}{\epsilon^2\mu} \times \left(C + \frac{6L\{d(\tau+1)+\rho_2\}^2}{\rho_1^2} + \frac{8L^2\{d(\tau+1)+\rho_2\}^4}{3\rho_1^4}\right) \ , \tag{5.51}$$

where ρ_1, ρ_2 *are defined in* (5.15), $\tau = 4d^{-1/2}\{\log(6/\epsilon)\}^{1/2}$ *and*

$$C = \left\lceil \frac{1}{\log(2)} \log\left(\frac{dL\{d(\tau+1)+\rho_2\}^2}{2\rho_1^2 \log(1+\epsilon/3)} \right) \right\rceil . \tag{5.52}$$

Theorem 5.13. *Assume* ***H***10, ***H***11(m) *for* $m \geq 0$ *and* ***H***12. *Let* $\epsilon, \mu \in (0,1)$. *There exists an explicit choice of the simulation parameters* $\mathcal{S}$ (5.16) *such that the estimator* $\hat{Z}$ *defined in* (5.17) *(with* g_{M-1} *replaced by* $\bar{g}_{M-1}$ *defined in* (5.49)*) satisfies with probability at least* $1-\mu$

$$(1-\epsilon)Z \leq \hat{Z} \leq (1+\epsilon)Z ,$$

and the cost (5.18) *of the algorithm is upper-bounded by*

$$\begin{aligned} \text{cost} \leq 2474 \left(\frac{17728C}{\epsilon^2\mu} + \log(Cd^2) \right) \frac{(C+1)d(d+4)}{\epsilon\sqrt{\mu}} \Bigg\{ & \frac{8L^2\{d(\tau+1)+\rho_2\}^4}{3\rho_1^4} \\ & + \frac{d^{1/2}\tilde{L}\{d(\tau+1)+\rho_2\}^3}{\sqrt{10}\rho_1^3} \max\left(\frac{5\rho_1}{d(\tau+1)+\rho_2}, \left(\frac{5}{9} + \frac{\rho_1^2}{\{d(\tau+1)+\rho_2\}^2 L} \right)^2 \right) \\ & + \frac{6L\{d(\tau+1)+\rho_2\}^2}{\rho_1^2} + C \Bigg\} , \end{aligned} \tag{5.53}$$

where ρ_1, ρ_2, C *are defined in* (5.15), (5.52) *respectively and* $\tau = 4d^{-1/2}\{\log(6/\epsilon)\}^{1/2}$.

Corollary 5.14. *Let* $\epsilon, \tilde{\mu} \in (0,1)$. *Repeat* $2\lceil 4\log(\tilde{\mu}^{-1})\rceil + 1$ *times the algorithm of Theorems* 5.12 *and* 5.13 *with* $\mu = 1/4$ *and denote by* $\hat{Z}$ *the median of the output values. We have with probability at least* $1-\tilde{\mu}$,

$$(1-\epsilon)Z \leq \hat{Z} \leq (1+\epsilon)Z .$$

The proofs follow the same arguments as Theorems 5.5 and 5.6 and corollary 5.7 and are detailed in the appendix Section 5.B.3.

Note that $\bar{g}_{M-1}$ (5.49) is a $\operatorname{Lip}\bar{g}_{M-1}$-Lipschitz function where,

$$\operatorname{Lip}\bar{g}_{M-1} = \frac{D}{\sigma_{M-1}^2} \exp\left(\frac{D^2}{2\sigma_{M-1}^2} \right) .$$

The results of Section 5.4 give an upper bound on MSE_{M-1} which is polynomial in the parameters if σ_{M-1}^2 is approximately equal to D^2. For $i \in \mathbb{N}^\star$, we define $(\sigma_i^2)_{i\geq 0}$ recursively. Set σ_0^2 as in (5.31) and

$$\sigma_{i+1}^2 = \varsigma_c(\sigma_i^2) , \tag{5.54}$$

where $\varsigma_c : \mathbb{R}_+^\star \to \mathbb{R}$ is defined for $t \in (0, D^2)$ by,

$$\varsigma_c(t) = \left(\frac{1}{t} - \frac{1}{2(d+4)(2^{k(t)+1}\sigma_0^2)} \right)^{-1} , \quad k(t) = \left\lfloor \frac{\log(t/\sigma_0^2)}{\log(2)} \right\rfloor , \tag{5.55}$$

and $\varsigma_c(t) = +\infty$ otherwise. Define M in this Section by,

$$M = \inf \left\{ i \geq 1 : \sigma_{i-1}^2 \geq D^2 \right\} . \tag{5.56}$$

By (5.55), for $t \in [\sigma_0^2, D^2)$, $\varsigma_c(t) \geq \{(4d+16)/(4d+15)\}\, t$, which implies $M < +\infty$. The following lemmas are the counterparts of Lemmas 5.10 and 5.11. They specify the choice of $\{\gamma_i\}_{i=0}^{M-1}$, $\{n_i\}_{i=0}^{M-1}$, $\{N_i\}_{i=0}^{M-1}$ to satisfy the conditions of Lemma 5.9.

Lemma 5.15. *Set $\eta = (\epsilon\sqrt{\mu})/8$. Assume **H**10, **H**11(m) for $m \geq 0$ and,*

i) for all $k \in \{0, \ldots, K-1\}$, $i \in \mathcal{I}_k$,

$$\gamma_i \leq \frac{1}{462} \frac{\eta^2 L_i^{-2} \sigma_i^{-2}}{K^2 d^2} \leq \frac{1}{m_i + L_i} , \tag{5.57}$$

$$n_i \geq \frac{453K}{\eta^2} \frac{1}{\kappa_i \gamma_i} , \tag{5.58}$$

$$N_i \geq 2(\kappa_i \gamma_i)^{-1} \log\left(Kd^2\right) , \tag{5.59}$$

ii)

$$\gamma_{M-1} \leq (1/26)\eta^2 d^{-1} L_{M-1}^{-2} \kappa_{M-1} \leq (m_{M-1} + L_{M-1})^{-1} , \tag{5.60}$$

$$n_{M-1} \geq 29\eta^{-2} (\kappa_{M-1}\gamma_{M-1})^{-1} , \tag{5.61}$$

$$N_{M-1} \geq 2(\kappa_{M-1}\gamma_{M-1})^{-1} \log(d) . \tag{5.62}$$

Then, the conditions i)-ii) of Lemma 5.9 are satisfied, with g_{M-1} replaced by $\bar{g}_{M-1}$.

Proof. The proof is postponed to Section 5.B.1. □

We have a similar result under the additional assumption **H**12.

Lemma 5.16. *Set $\eta = (\epsilon\sqrt{\mu})/8$. Assume **H**10, **H**11(m) for $m \geq 0$, **H**12 and,*

i) for all $k \in \{0, \ldots, K-1\}$, $i \in \mathcal{I}_k$,

$$\gamma_i \leq \sqrt{\frac{3}{7}} \frac{\eta \sigma_i^{-1}}{8Kd} \left(d\tilde{L}^2 + 10 L_i^4 \sigma_i^2 \right)^{-1/2} \leq \frac{1}{m_i + L_i} , \tag{5.63}$$

n_i, N_i as in (5.58), (5.59) and,

ii)

$$\gamma_{M-1} \leq \sqrt{\frac{3}{8\mathrm{e}}} \frac{\eta \kappa_{M-1} \sigma_{M-1}}{\sqrt{d}} \left(d\tilde{L}^2 + 10 L_{M-1}^4 \sigma_{M-1}^2 \right)^{-1/2} \leq \frac{1}{m_{M-1} + L_{M-1}} , \tag{5.64}$$

n_{M-1}, N_{M-1} as in (5.61), (5.62).

Then, the conditions i)-ii) of Lemma 5.9 are satisfied, with g_{M-1} replaced by $\bar{g}_{M-1}$.

Proof. The proof is postponed to Section 5.B.2. □

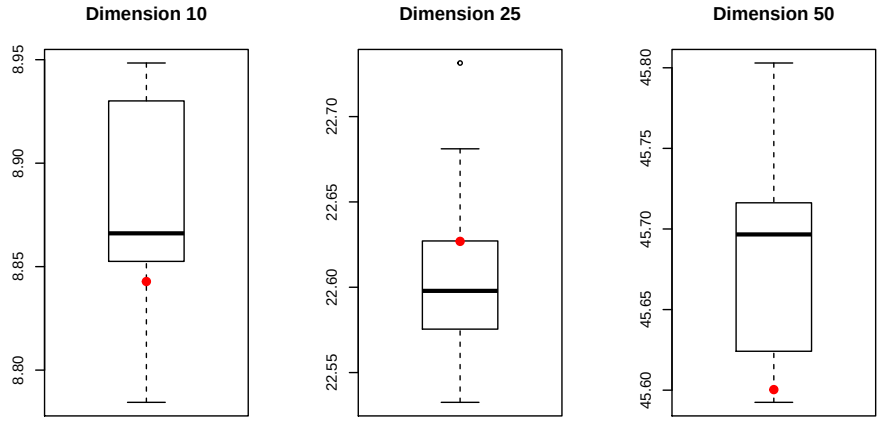

Figure 5.1: Boxplots of the logarithm of the normalizing constants of a multivariate Gaussian distribution in dimension $d \in \{10, 25, 50\}$.

5.3 Numerical experiments

For the following numerical experiments, the code and data are available at `https://github.com/nbrosse/normalizingconstant`. We first experiment our algorithm to compute the logarithm of the normalizing constant of a multivariate Gaussian distribution in dimension $d \in \{10, 25, 50\}$, of mean 0 and inverse covariance matrix $\mathrm{diag}(2, 1^{\otimes(d-1)})$. We set $\epsilon = \mu = 0.1$. The number of phases M of the algorithm and the variances $\{\sigma_i^2\}_{i=0}^{M-1}$ are chosen according to the formulas (5.34) and (5.36). For each phase of the algorithm, the step size γ_i is set equal to $10^{-2}(m_i + L_i)^{-1}$, the burn-in period N_i to 10^4 and the number of samples n_i to 10^5 where m_i, L_i are defined in (5.13). We carry out 10 independent runs of the algorithm and compute the boxplots in Figure 5.1. The true values of the logarithm of the normalizing constants are known and displayed by the red points in Figure 5.1.

We illustrate then our methodology to compute Bayesian model evidence; see [FW12] and the references therein. Let $y \in \mathbb{R}^p$ be a vector of observations and $\mathcal{M}_1, \dots, \mathcal{M}_l$ be a collection of competing models. Let $\{p(\mathcal{M}_i)\}_{i=1}^l$ be a prior distribution on the collection of models. For $i \in \{0, \dots, l\}$, denote by $p(y|\theta^{(\mathcal{M}_i)}, \mathcal{M}_i)$ the likelihood of the model $\mathcal{M}_i$. The dominating measure is implicitly considered to be the Lebesgue measure on $\mathbb{R}^p$. Similarly, for $i \in \{0, \dots, l\}$, denote by $p(\theta^{(\mathcal{M}_i)}|\mathcal{M}_i)$ the prior density on the parameters $\theta^{(\mathcal{M}_i)}$ under the model $\mathcal{M}_i$ where the dominating measure is implicitly considered to be the Lebesgue measure on $\mathbb{R}^{d(\mathcal{M}_i)}$. The posterior distribution of interest is then for $i \in \{0, \dots, l\}$,

$$p(\theta^{(\mathcal{M}_i)}, \mathcal{M}_i|y) \propto p(y|\theta^{(\mathcal{M}_i)}, \mathcal{M}_i) p(\theta^{(\mathcal{M}_i)}|\mathcal{M}_i) p(\mathcal{M}_i)$$

The posterior distribution conditional on model $\mathcal{M}_i$ can also be considered

$$p(\theta^{(\mathcal{M}_i)}|\mathcal{M}_i, y) \propto p(y|\theta^{(\mathcal{M}_i)}, \mathcal{M}_i)p(\theta^{(\mathcal{M}_i)}|\mathcal{M}_i) \tag{5.65}$$

For $i \in \{0, \dots, l\}$, the evidence $p(y|\mathcal{M}_i)$ of the model $\mathcal{M}_i$ is defined by the normalizing constant for the posterior distribution (5.65)

$$p(y|\mathcal{M}_i) = \int_{\mathbb{R}^{d(\mathcal{M}_i)}} p(y|\theta^{(\mathcal{M}_i)}, \mathcal{M}_i)p(\theta^{(\mathcal{M}_i)}|\mathcal{M}_i)\mathrm{d}\theta^{(\mathcal{M}_i)} \ .$$

The Bayes factor BF_{12} between two models $\mathcal{M}_i$ and $\mathcal{M}_j$ is then defined by the ratio of evidences [Rob07, Section 7.2.2], $BF_{ij} = p(y|\mathcal{M}_i)/p(y|\mathcal{M}_j)$. In the following experiments, we estimate the log evidence $\log(p(y|\mathcal{M}_i))$. For ease of notation, the dependence on the model $\mathcal{M}$ of the parameters θ and the dimension d of the state space is implicit in the sequel.

Define $\ell^{(\mathcal{M})} : \mathbb{R}^d \to \mathbb{R}$ by $\ell^{(\mathcal{M})}(\theta) = -\log(p(y|\theta, \mathcal{M})p(\theta|\mathcal{M}))$ for $\theta \in \mathbb{R}^d$. In the examples we consider, $\ell^{(\mathcal{M})}$ satisfies **H**10, **H**11, **H**12 and has a unique minimum $\theta_\star^{(\mathcal{M})}$. Define then $U^{(\mathcal{M})} : \mathbb{R}^d \to \mathbb{R}_+$ by $U^{(\mathcal{M})}(\theta) = \ell^{(\mathcal{M})}(\theta + \theta_\star^{(\mathcal{M})}) - \ell^{(\mathcal{M})}(\theta_\star^{(\mathcal{M})})$ for $\theta \in \mathbb{R}^d$. The algorithm described in Section 5.2 can be applied to $U^{(\mathcal{M})}$. For each example, two different models will be considered and $U^{(\mathcal{M})}$ will be written as $U^{(k)}$ for $k = 1, 2$.

The numerical experiments are carried out on a Gaussian linear and logistic regression following the experimental setup of [FW12, Section 4], which is now considered as a classical benchmark.

Linear regression The linear regression is conducted on $p = 42$ specimens of radiata pine [WW59]. The response variable $y \in \mathbb{R}^p$ is the maximum compression strength parallel to the grain. The explanatory variables are $x \in \mathbb{R}^p$ the density and $z \in \mathbb{R}^p$ the density adjusted for resin content. x and z are centered. The covariates of the first model $\mathcal{M}_1$, $X^{(1)} \in \mathbb{R}^{p\times 2}$, are composed of an intercept and x, while the covariates of the second model $\mathcal{M}_2$, $X^{(2)} \in \mathbb{R}^{p\times 2}$, are composed of an intercept and z. For $k = 1, 2$, the likelihood is defined by,

$$p(y|\theta, \mathcal{M}_k) = \left(\frac{\lambda}{2\pi}\right)^{d/2} \exp\left(-(\lambda/2)\left\|y - X^{(k)}\theta\right\|^2\right) \ ,$$

where $\lambda = 10^{-5}$. For the two models, the parameter θ follows the same Gaussian prior of mean $(3000, 185)$ and inverse covariance matrix $\lambda Q_0 = \lambda \,\mathrm{diag}(0.06, 6)$ where diag denotes a diagonal matrix. These values are taken from [FW12, section 4.1]. For $k = 1, 2$, $U^{(k)}$ is $m^{(k)}$-strictly convex and $L^{(k)}$-gradient Lipschitz, where $m^{(k)}$ (resp. $L^{(k)}$) is the minimal (resp. maximal) eigenvalue of $\lambda([X^{(k)}]^{\mathrm{T}}X^{(k)} + Q_0)$. We set $\epsilon = \mu = 0.1$. The number of phases M of the algorithm and the variances $\{\sigma_i^2\}_{i=0}^{M-1}$ are chosen accordingly to the formulas (5.34) and (5.36). For each phase, the step size γ_i is set equal to $10^{-2}(\kappa_i\sigma_i^2 m_i)/(dL_i^2)$, the burn-in period N_i to $10^3(\kappa_i\gamma_i)^{-1}$ and the number of samples n_i to $10^4 m_i^{1/2}/(\kappa_i^2\sigma_i\gamma_i)$ where m_i, L_i, κ_i are defined in (5.13) and (5.14). The experiments are repeated 10 times and the boxplots for each model $\mathcal{M}$ are plotted in Figure 5.2. Note

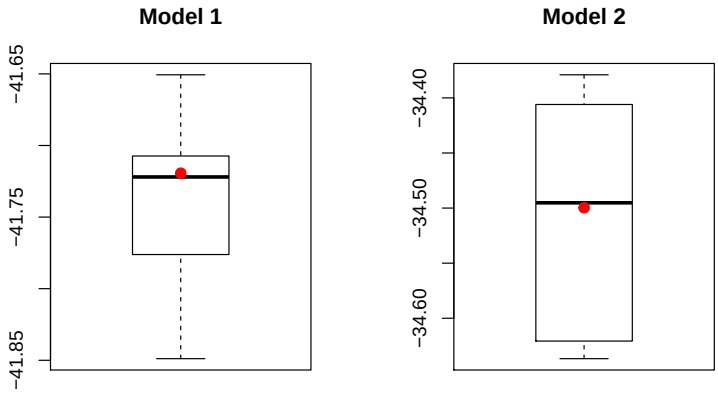

Figure 5.2: Boxplots of the log evidence for the two models on the Gaussian regression.

that for this Gaussian model, the log evidence is known and displayed by the red points in Figure 5.2.

With the same parameters for the algorithm, we run 10 independent runs at each phase to measure the variability of each estimator $\hat{\pi}_i(g_i)$ defined in (5.9). The result is plotted in Figure 5.3 for the model $\mathcal{M}_1$.

Logistic regression The logistic regression is performed on the Pima Indians dataset[1]. In this case, $y \in \{0,1\}^p$ is a vector of diabetes indicators for $p = 532$ Pima Indian women and the potential predictors for diabetes are: number of pregnancies NP $\in \mathbb{R}^p$, plasma glucose concentration PGC $\in \mathbb{R}^p$, diastolic blood pressure BP $\in \mathbb{R}^p$, triceps skin fold thickness TST $\in \mathbb{R}^p$, body mass index BMI $\in \mathbb{R}^p$, diabetes pedigree function DP $\in \mathbb{R}^p$ and age AGE $\in \mathbb{R}^p$. These variates are centered and standardized. The covariates of the first model $\mathcal{M}_1$ are $X^{(1)} = (\text{intercept}, \text{NP}, \text{PGC}, \text{BMI}, \text{DP}) \in \mathbb{R}^{p\times 5}$ and the covariates of the second model $\mathcal{M}_2$ are $X^{(2)} = (\text{intercept}, \text{NP}, \text{PGC}, \text{BMI}, \text{DP}, \text{AGE}) \in \mathbb{R}^{p\times 6}$, where intercept is the intercept of the regressions. The likelihood is defined for $k = 1, 2$ by,

$$p(y|\theta, \mathcal{M}_k) = \exp\left(\sum_{i=1}^{p}\left\{y_i\theta^{\mathrm{T}}X_i^{(k)} - \log\left(1 + \mathrm{e}^{\theta^{\mathrm{T}}X_i^{(k)}}\right)\right\}\right) ,$$

where $X_i^{(k)}$ denotes the i^{th} row of $X^{(k)}$. For the two models, the prior on θ is Gaussian, of mean 0 and inverse covariance matrix $\tau\,\mathrm{Id}$ where $\tau = 0.01$. For $k = 1, 2$, $U^{(k)}$ is τ-strongly convex and $L^{(k)}$-gradient Lipschitz, where $L^{(k)} = \lambda_{\max}([X^{(k)}]^{\mathrm{T}}X^{(k)})/4 + \tau$ and $\lambda_{\max}([X^{(k)}]^{\mathrm{T}}X^{(k)})$ is the maximal eigenvalue of $[X^{(k)}]^{\mathrm{T}}X^{(k)}$. We set $\epsilon = \mu = 0.1$. The algorithm to estimate $\log(p(y|\mathcal{M}))$ described in Section 5.2.1 is applied with the following

[1] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

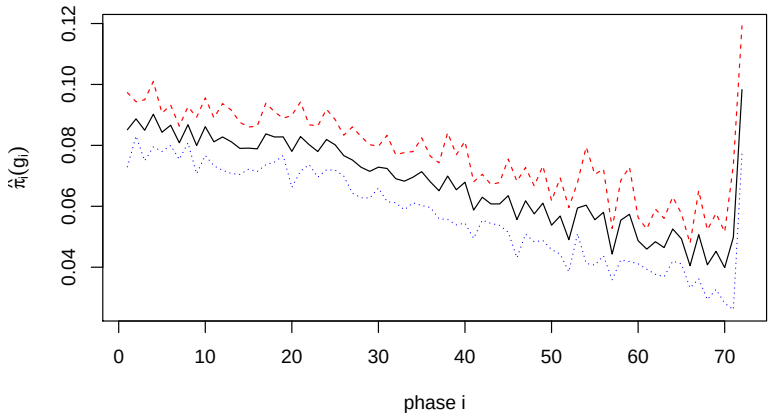

Figure 5.3: Error plot of $\hat{\pi}_i(g_i)$ for $i \in \{0, \dots, M-1\}$ in the example of the Gaussian regression (model $\mathcal{M}_1$). The mean of $\hat{\pi}_i(g_i)$ is displayed in black and is spaced apart from the other two curves by the standard deviation of $\hat{\pi}_i(g_i)$.

modifications. The number of phases is decreased and the recurrence for the variances $\{\sigma_i^2\}_{i=0}^{M-1}$ is thus redefined by $\sigma_{i+1}^2 = \varsigma_s^5(\sigma_i^2)$ as long as the stopping condition (5.36) is not fulfilled. For $i \in \{1, \dots, 30\}$, the burn-in period N_i is set equal to 10^4, the number of samples n_i to 10^6 and the step size γ_i to $10^{-2}(m_i + L_i)^{-1}$ where m_i, L_i are defined in (5.13); for $i > 30$, the number of samples n_i is set equal to 10^5 and the step size γ_i to $10^{-1}(m_i+L_i)^{-1}$. We compare our results with different methods reviewed in [FW12] and implemented in [Wys11]. These are the Laplace method (L), Laplace at the Maximum a Posteriori (L-MAP), Chib's method (C) Annealed Importance Sampling (AIS) and Power Posterior (PP). The experiments are repeated 10 times and the boxplots for each model $\mathcal{M}$ and each method are plotted in Figure 5.4. The comparison is difficult because of the absence of ground truth; nevertheless, we observe that AIS has a high variance and our methodology AV seems to have a bias.

With the same parameters for the algorithm, we run 10 independent runs at each phase to measure the variability of each estimator $\hat{\pi}_i(g_i)$ defined in (5.9) and display the result in Figure 5.5 for the model $\mathcal{M}_1$. We observe that in the initial phase of the algorithm, $\hat{\pi}_i(g_i)$ is high and decreases gradually.

Mixture of Gaussian distributions The final example we address is a Bayesian analysis of a finite mixture of Gaussian distributions, see [MDJ06, Section 4.2] and we aim at estimating the log evidence of the posterior distribution. Note that this model does not fit into our assumptions because the potential U is not continuously differentiable on $\mathbb{R}^d$ and neither convex. Nevertheless, we experiment heuristically our

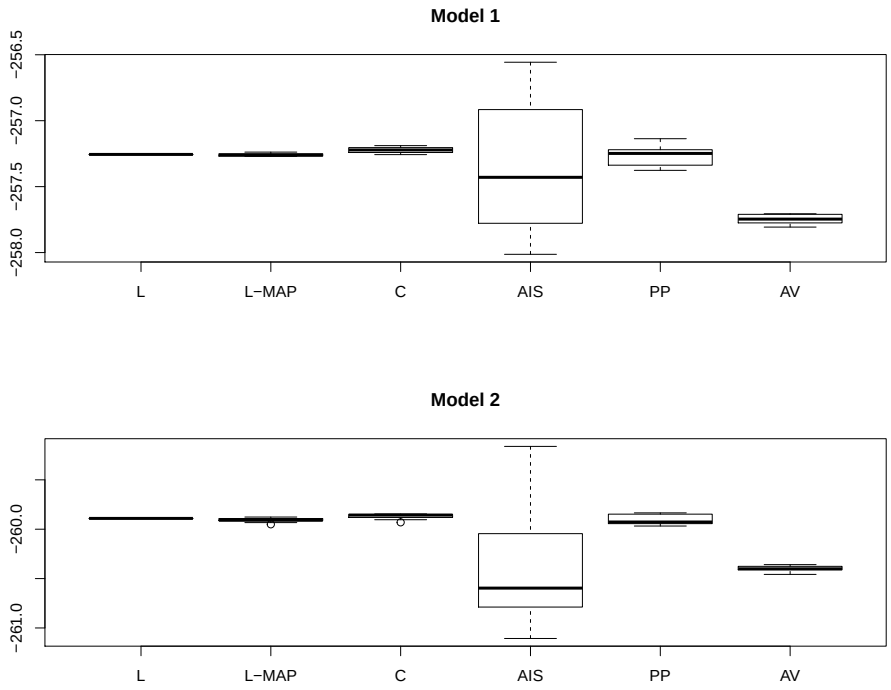

Figure 5.4: Boxplots of the log evidence for the two models on the logistic regression. The methods are the Laplace method (L), Laplace at the Maximum a Posteriori (L-MAP), Chib's method (C), Annealed Importance Sampling (AIS), Power Posterior (PP) and our method (AV).

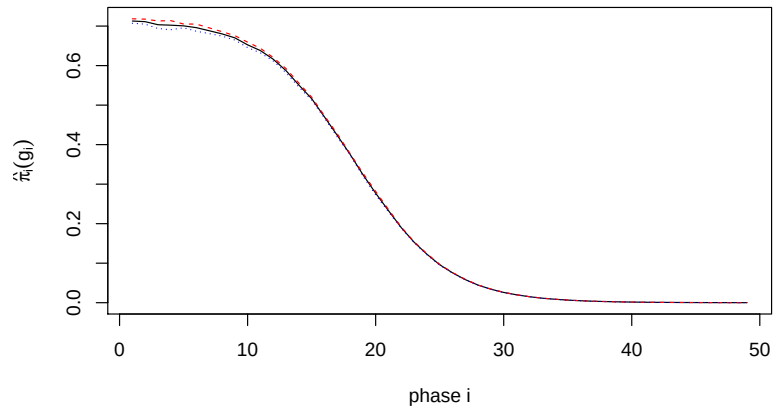

Figure 5.5: Error plot of $\hat{\pi}_i(g_i)$ for $i \in \{0, \dots, M-1\}$ in the example of the logistic regression (model $\mathcal{M}_1$). The mean of $\hat{\pi}_i(g_i)$ is displayed in black and is spaced apart from the other two curves by the standard deviation of $\hat{\pi}_i(g_i)$.

algorithm on a close model given by its likelihood

$$p(y|\{\theta_j\}_{j=1}^4) = \prod_{i=1}^{p} \left[\frac{1}{4} \left(\frac{\lambda}{2\pi} \right)^2 \left\{ \sum_{j=1}^{4} \exp\left(-(\lambda/2)(y_i - \theta_j)^2 \right) \right\} \right]$$

for $y = (y_1, \dots, y_p) \in \mathbb{R}^p$ a vector of observations. The prior distributions are set following the recommendations of [MDJ06, Section 4.2.1] and [RG97]. For $j \in \{1, \dots, 4\}$, θ_j is drawn from a Gaussian distribution of mean $\xi = 1.35$ and inverse variance $\varsigma = 7.6 \times 10^{-3}$. λ is set equal to 0.03. The observations $y \in \mathbb{R}^{100}$ are 100 simulated data points from an equally weighted mixture of four Gaussian densities with means $(-3, 0, 3, 6)$ and standard deviations 0.55, taken from [JHS05]. Define for $\theta = (\theta_1, \dots, \theta_4) \in \mathbb{R}^4$, $\ell : \mathbb{R}^4 \to \mathbb{R}$ by $\ell(\theta) = -\log(p(y|\theta)p(\theta))$. The `optim` function of R [R C18] gives a local minimum at $\theta^* \approx (1.76562^{\otimes 4})$. Define then the potential $U : \mathbb{R}^4 \to \mathbb{R}$ for $\theta \in \mathbb{R}^4$ by $U(\theta) = \ell(\theta + \theta^*) - \ell(\theta^*)$. Set $\epsilon = \mu = 0.1$, $m = \varsigma$ and $L = 1$. Similarly to the logistic regression, to decrease the running time of the algorithm, the recurrence for the variances $\{\sigma_i^2\}_{i=0}^{M-1}$ is defined by $\sigma_{i+1}^2 = \varsigma_s^5(\sigma_i^2)$ as long as the stopping condition (5.36) is not fulfilled. For each phase, the step size γ_i is set equal to $10^{-1}(\kappa_i \sigma_i^2 m_i)/(dL_i^2)$, the burn-in period N_i to 10^4 and the number of samples n_i to 10^5 where m_i, L_i, κ_i are defined in (5.13) and (5.14). For comparison purposes, we run the same algorithm using the Metropolis Adjusted Langevin Algorithm (MALA) instead of ULA to estimate $\hat{\pi}_i(g_i)$ at each phase. The step size γ_i is set equal to $(\kappa_i \sigma_i^2 m_i)/(dL_i^2)$ and the number of samples n_i to 10^6. The experiments are repeated 10 times. The boxplot is plotted in Figure 5.6 and the red point indicates the mean of our algorithm using MALA.

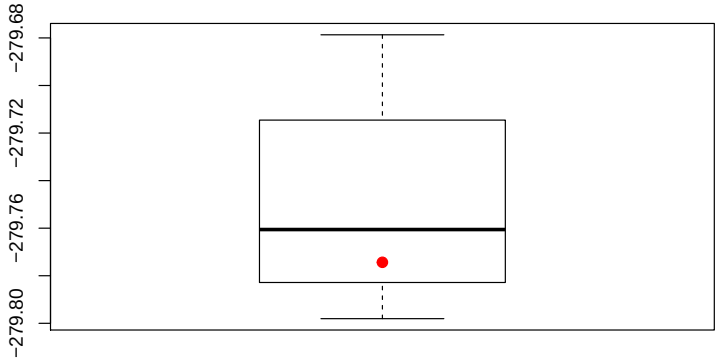

Figure 5.6: Boxplot of the log evidence for the mixture of Gaussian distributions.

5.4 Mean squared error for locally Lipschitz functions

In this Section, we extend the results of [DM16, Section 3] to locally Lipschitz functions. This Section is of independent interest and only Propositions 5.17 and 5.20 are used in **??**. Let $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. Consider the target distribution π with density $x \mapsto \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y$ w.r.t. the Lebesgue measure. We deal with the problem of estimating $\int_{\mathbb{R}^d} f(x) \mathrm{d}\pi(x)$ for locally Lipschitz $f : \mathbb{R}^d \to \mathbb{R}$ by the ULA algorithm defined for $k \in \mathbb{N}$ by,

$$X_{k+1} = X_k - \gamma_{k+1} \nabla U(X_k) + \sqrt{2\gamma_{k+1}} Z_{k+1} \,, \tag{5.66}$$

where $(Z_k)_{k \geq 1}$ is an i.i.d. sequence of d-dimensional Gaussian vectors with zero mean, identity covariance and $(\gamma_k)_{k \geq 1}$ is a sequence of positive step sizes, which can either be held constant or be chosen to decrease to 0. For $n, p \in \mathbb{N}$, denote by

$$\Gamma_{n,p} \stackrel{\text{def}}{=} \sum_{k=n}^{p} \gamma_k \,, \qquad \Gamma_n = \Gamma_{1,n} \,, \tag{5.67}$$

and consider the Markov kernel R_γ given for all $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ by

$$R_\gamma(x, \mathrm{A}) = \int_{\mathrm{A}} (4\pi\gamma)^{-d/2} \exp\left(-(4\gamma)^{-1} \left\| y - x + \gamma \nabla U(x) \right\|^2\right) \mathrm{d}y \,. \tag{5.68}$$

Define

$$Q_\gamma^{n,p} = R_{\gamma_n} \cdots R_{\gamma_p} \,, \qquad Q_\gamma^n = Q_\gamma^{1,n} \,, \tag{5.69}$$

with the convention that for $n, p \geq 0$, $n < p$, $Q_\gamma^{p,n}$ and $Q_\gamma^{0,0}$ are the identity operator.

For all initial distribution μ_0 on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $\mathbb{P}_{\mu_0}$ and $\mathbb{E}_{\mu_0}$ denote the probability and the expectation respectively associated with the sequence of Markov kernels (5.68) and the initial distribution μ_0 on the canonical space $((\mathbb{R}^d)^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^d)^{\otimes\mathbb{N}})$ and $(X_k)_{k\in\mathbb{N}}$ denotes the canonical process. Let $f : \mathbb{R}^d \to \mathbb{R}$ and consider the following assumption,

L 1. *1. There exists $L_f : \mathbb{R}^d \to [0, +\infty)$ a continuous function such that for all $x, y \in \mathbb{R}^d$, $|f(y) - f(x)| \leq \|y - x\| \max\{L_f(x), L_f(y)\}$.*

2. There exist $\epsilon > 0$, $C_\pi > 0$ and continuous functions $C_Q, C_{Q,\epsilon} : \mathbb{R}^d \to [0, +\infty)$ such that for all $x \in \mathbb{R}^d$,

$$\pi(L_f^2) \leq C_\pi \;, \quad \sup_{p\geq n\geq 0} \delta_x Q_\gamma^{n,p}\left(L_f^2\right) \leq C_Q(x) \;, \tag{5.70}$$
$$\sup_{p\geq n\geq 0} \delta_x Q_\gamma^{n,p}\left(L_f^{2(1+\epsilon)}\right) \leq C_{Q,\epsilon}(x)$$

Under **L** 1, we study the approximation of $\int_{\mathbb{R}^d} f(y)\pi(\mathrm{d}y)$ by the weighted average estimator

$$\hat{\pi}_n^N(f) = \sum_{k=N+1}^{N+n} \omega_{k,n}^N f(X_k) \;, \quad \omega_{k,n}^N = \gamma_{k+1}\Gamma_{N+2,N+n+1}^{-1} \;, \tag{5.71}$$

where $N \geq 0$ is the length of the burn-in period and $n \geq 1$ is the number of samples. The Mean Squared Error (MSE) of $\hat{\pi}_n^N(f)$ is defined by:

$$\mathrm{MSE}_f(x, N, n) = \mathbb{E}_x\left[\left\{\hat{\pi}_n^N(f) - \pi(f)\right\}^2\right] \;, \tag{5.72}$$

and can be decomposed as,

$$\mathrm{MSE}_f(x, N, n) = \left\{\mathbb{E}_x[\hat{\pi}_n^N(f)] - \pi(f)\right\}^2 + \mathrm{Var}_x\left[\hat{\pi}_n^N(f)\right] \;. \tag{5.73}$$

The analysis of $\mathrm{MSE}_f(x, N, n)$ is similar to [DM16, Section 3]. First, the squared bias in (5.73) is bounded. Denote by,

$$A_0 = 2L^2\kappa^{-1}d \;, \tag{5.74}$$
$$A_1 = 2dL^2 + dL^4(\kappa^{-1} + (m+L)^{-1})(m^{-1} + 6^{-1}(m+L)^{-1}) \;, \tag{5.75}$$
$$B_0 = d\left(2L^2 + \kappa^{-1}\{d\tilde{L}^2/3 + 4L^4/(3m)\}\right) \;, \tag{5.76}$$
$$B_1 = dL^4\left(\kappa^{-1} + \{6(m+L)\}^{-1} + m^{-1}\right) \;, \tag{5.77}$$

where κ is given by (5.14). Define then for $n \in \mathbb{N}^\star$,

$$u_n^{(1)}(\gamma) = \prod_{k=1}^{n}(1 - \kappa\gamma_k/2) \;, \tag{5.78}$$
$$u_n^{(2)}(\gamma) = \sum_{i=1}^{n}\left(A_0\gamma_i^2 + A_1\gamma_i^3\right)\prod_{k=i+1}^{n}(1 - \kappa\gamma_k/2) \;, \tag{5.79}$$
$$u_n^{(3)}(\gamma) = \sum_{i=1}^{n}\left(B_0\gamma_i^3 + B_1\gamma_i^4\right)\prod_{k=i+1}^{n}(1 - \kappa\gamma_k/2) \;. \tag{5.80}$$

Proposition 5.17. *Assume **H**10 and **H**11(m) for $m > 0$. Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfying **L** 1. Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 1/(m+L)$. Let $x^\star$ be the unique minimizer of U. Let $(X_n)_{n\geq 0}$ be given by (5.66) and started at $x \in \mathbb{R}^d$. Then for all $N \geq 0$, $n \geq 1$:*

$$\left\{\mathbb{E}_x[\hat{\pi}_n^N(f)] - \pi(f)\right\}^2 \leq \{C_\pi + C_Q(x)\} \times \sum_{k=N+1}^{N+n} \omega_{k,n}^N \left\{2(\|x - x^\star\|^2 + d/m)u_k^{(1)}(\gamma) + w_k(\gamma)\right\} , \quad (5.81)$$

*where $u_n^{(1)}(\gamma)$ is given in (5.78) and $w_n(\gamma)$ is equal to $u_n^{(2)}(\gamma)$ defined by (5.79) and to $u_n^{(3)}(\gamma)$, defined by (5.80), if **H**12 holds.*

Proof. For all $k \in \{N+1, \ldots, N+n\}$, let ξ_k be the optimal transference plan between $\delta_x Q_\gamma^k$ and π for W_2. By the Jensen and the Cauchy-Schwarz inequalities, and **L** 1, we have:

$$\begin{aligned}\left(\mathbb{E}_x[\hat{\pi}_n^N(f)] - \pi(f)\right)^2 &= \left(\sum_{k=N+1}^{N+n} \omega_{k,n}^N \int_{\mathbb{R}^d\times\mathbb{R}^d} \{f(z) - f(y)\}\xi_k(\mathrm{d}z, \mathrm{d}y)\right)^2 \\ &\leq \sum_{k=N+1}^{N+n} \omega_{k,n}^N \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|z-y\| \max\{L_f(z), L_f(y)\}\, \xi_k(\mathrm{d}z, \mathrm{d}y)\right)^2 \\ &\leq \{C_\pi + C_Q(x)\} \sum_{k=N+1}^{N+n} \omega_{k,n}^N \int_{\mathbb{R}^d\times\mathbb{R}^d} \|z-y\|^2\, \xi_k(\mathrm{d}z, \mathrm{d}y) .\end{aligned}$$

The proof follows from [DM16, Theorems 5 and 8]. □

To deal with the variance term in (5.73), we adapt the proof of [JO10, Theorem 2] to our setting, where f is only locally Lipschitz and the Markov chain (5.66) is inhomogeneous. It is based on the Gaussian Poincaré inequality [BLM13, Theorem 3.20]. Let $Z = (Z_1, \ldots, Z_d)$ be a Gaussian vector with identity covariance matrix and $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function. Recall that by Rademacher's Theorem [EG15, Theorem 3.2], a locally Lipschitz function is almost everywhere differentiable w.r.t. Lebesgue measure on $\mathbb{R}^d$. The Gaussian Poincaré inequality states that $\mathrm{Var}[f(Z)] \leq \mathbb{E}[\|\nabla f(Z)\|^2]$. Noticing that for all $x \in \mathbb{R}^d$, $R_\gamma(x, \cdot)$ defined in (5.68) is a Gaussian distribution with mean $x - \gamma\nabla U(x)$ and covariance matrix $2\gamma\,\mathrm{I}_d$, the Gaussian Poincaré inequality implies:

$$0 \leq \int R_\gamma(x, \mathrm{d}y)\,\{f(y) - R_\gamma f(x)\}^2 \leq 2\gamma \int R_\gamma(x, \mathrm{d}y)\,\|\nabla f(y)\|^2 . \quad (5.82)$$

First consider the following decomposition of $\hat{\pi}_n^N(f) - \mathbb{E}_x[\hat{\pi}_n^N(f)]$ as the sum of martingale increments,

$$\hat{\pi}_n^N(f) - \mathbb{E}_x[\hat{\pi}_n^N(f)] = \sum_{k=N}^{N+n-1} \left\{\mathbb{E}_x^{\mathcal{G}_{k+1}}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x^{\mathcal{G}_k}\left[\hat{\pi}_n^N(f)\right]\right\} + \mathbb{E}_x^{\mathcal{G}_N}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x[\hat{\pi}_n^N(f)] ,$$

where $(\mathcal{G}_n)_{n\geq 0}$ is the natural filtration associated with the Markov chain $(X_n)_{n\geq 0}$. This implies that the variance may be decomposed as the following sum

$$\mathrm{Var}_x\left[\hat{\pi}_n^N(f)\right] = \sum_{k=N}^{N+n-1} \mathbb{E}_x\left[\left(\mathbb{E}_x^{\mathcal{G}_{k+1}}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x^{\mathcal{G}_k}\left[\hat{\pi}_n^N(f)\right]\right)^2\right] + \mathbb{E}_x\left[\left(\mathbb{E}_x^{\mathcal{G}_N}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x[\hat{\pi}_n^N(f)]\right)^2\right] . \quad (5.83)$$

Because $\hat{\pi}_n^N(f)$ is an additive functional, the martingale increment $\mathbb{E}_x^{\mathcal{G}_{k+1}}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x^{\mathcal{G}_k}\left[\hat{\pi}_n^N(f)\right]$ has a simple expression. For $k = N+n, \ldots, N+1$, define backward in time the function

$$\Phi_{n,k}^N : x_k \mapsto \omega_{k,n}^N f(x_k) + R_{\gamma_{k+1}}\Phi_{n,k+1}^N(x_k) , \quad (5.84)$$

with the convention $\Phi_{n,N+n+1}^N = 0$. Denote finally

$$\Psi_n^N : x_N \mapsto R_{\gamma_{N+1}}\Phi_{n,N+1}^N(x_N) . \quad (5.85)$$

Note that for $k \in \{N, \ldots, N+n-1\}$, by the Markov property,

$$\Phi_{n,k+1}^N(X_{k+1}) - R_{\gamma_{k+1}}\Phi_{n,k+1}^N(X_k) = \mathbb{E}_x^{\mathcal{G}_{k+1}}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x^{\mathcal{G}_k}\left[\hat{\pi}_n^N(f)\right] , \quad (5.86)$$

and $\Psi_n^N(X_N) = \mathbb{E}_x^{\mathcal{G}_N}\left[\hat{\pi}_n^N(f)\right]$. With these notations, (5.83) may be equivalently expressed as

$$\mathrm{Var}_x\left[\hat{\pi}_n^N(f)\right] = \sum_{k=N}^{N+n-1} \mathbb{E}_x\left[R_{\gamma_{k+1}}\left\{\Phi_{n,k+1}^N(\cdot) - R_{\gamma_{k+1}}\Phi_{n,k+1}^N(X_k)\right\}^2(X_k)\right] + \mathrm{Var}_x\left[\Psi_n^N(X_N)\right] . \quad (5.87)$$

Now for $k = N+n, \ldots, N+1$, we will use the Gaussian Poincaré inequality (5.82) to the sequence of function $\Phi_{n,k}^N$. It is required to prove that $\Phi_{n,k}^N$ is locally Lipschitz (see Lemma 5.18). For the variance of $\Psi_n^N(X_N)$, similar arguments apply using Lemma 5.19.

Lemma 5.18. *Assume **H**10, **H**11(m) for $m > 0$ and let $f : \mathbb{R}^d \to \mathbb{R}$ satisfying **L**1. Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 2/(m+L)$. Then for all $\ell \geq n \geq 0$, $Q_\gamma^{n,\ell} f$ is locally Lipschitz and differentiable for almost all $x \in \mathbb{R}^d$. Its gradient is bounded by,*

$$\left\|\nabla Q_\gamma^{n,\ell} f(x)\right\| \leq \prod_{k=n}^{\ell} (1-\kappa\gamma_k)^{1/2} (\delta_x Q_\gamma^{n,\ell} L_f^2)^{1/2} . \quad (5.88)$$

Proof. Let $\xi_{x,y}$ be the optimal transference plan between $\delta_x Q_\gamma^{n,\ell}$ and $\delta_y Q_\gamma^{n,\ell}$ for W_2. By Rademacher's Theorem [EG15, Theorem 3.2], $\nabla Q_\gamma^{n,\ell} f(x)$ exists for almost all $x \in \mathbb{R}^d$.

For such x, using Cauchy-Schwarz's inequality and [DM16, Theorem 4], we have

$$\begin{aligned}
\left\|\nabla Q_\gamma^{n,\ell} f(x)\right\| &= \sup_{\|u\|\le 1} \lim_{t\to 0} \left|\left(Q_\gamma^{n,\ell} f(x+tu) - Q_\gamma^{n,\ell} f(x)\right)/t\right| \\
&= \sup_{\|u\|\le 1} \lim_{t\to 0} \left| t^{-1} \int_{\mathbb{R}^d\times\mathbb{R}^d} \{f(z_2) - f(z_1)\}\, \xi_{x,x+tu}(\mathrm{d}z_1, \mathrm{d}z_2)\right| \\
&\le \sup_{\|u\|\le 1} \liminf_{t\to 0} t^{-1} W_2(\delta_x Q_\gamma^{n,\ell}, \delta_{x+tu} Q_\gamma^{n,\ell}) \\
&\qquad\qquad \times \left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} (L_f^2(z_1)\vee L_f^2(z_2))\xi_{x,x+tu}(\mathrm{d}z_1,\mathrm{d}z_2)\right\}^{1/2} \\
&\le \sup_{\|u\|\le 1} \liminf_{t\to 0} \prod_{k=n}^{\ell} (1-\kappa\gamma_k)^{1/2} \left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} (L_f^2(z_1)\vee L_f^2(z_2))\xi_{x,x+tu}(\mathrm{d}z_1,\mathrm{d}z_2)\right\}^{1/2} .
\end{aligned}$$

It is then sufficient to prove that,

$$\lim_{y\to x} \int_{\mathbb{R}^d\times\mathbb{R}^d} L_f^2(z_1)\vee L_f^2(z_2)\xi_{x,y}(\mathrm{d}z_1,\mathrm{d}z_2) = \int_{\mathbb{R}^d} L_f^2(z_1)\delta_x Q_\gamma^{n,\ell}(\mathrm{d}z_1) .$$

Let $\varepsilon, \eta, \mathsf{R} > 0$ and $y \in \mathbb{R}^d$. Since $a\vee b - a = (b-a)_+$, we have

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} (L_f^2(z_2) - L_f^2(z_1))_+ \xi_{x,y}(\mathrm{d}z_1,\mathrm{d}z_2) = E_1(y) + E_2(y) + E_3(y)$$

where,

$$\begin{aligned}
E_1(y) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(L_f^2(z_2) - L_f^2(z_1)\right)_+ \mathbb{1}_{\{\|z_1\|+\|z_2\|\ge 2\mathsf{R}\}} \xi_{x,y}(\mathrm{d}z_1,\mathrm{d}z_2) , \\
E_2(y) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(L_f^2(z_2) - L_f^2(z_1)\right)_+ \mathbb{1}_{\{\|z_1\|+\|z_2\|\le 2\mathsf{R}\}} \mathbb{1}_{\{\|z_1-z_2\|\le\eta\}} \xi_{x,y}(\mathrm{d}z_1,\mathrm{d}z_2) , \\
E_3(y) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(L_f^2(z_2) - L_f^2(z_1)\right)_+ \mathbb{1}_{\{\|z_1\|+\|z_2\|\le 2\mathsf{R}\}} \mathbb{1}_{\{\|z_1-z_2\|\ge\eta\}} \xi_{x,y}(\mathrm{d}z_1,\mathrm{d}z_2) .
\end{aligned}$$

Hölder's inequality gives for $p, q > 1$, $1/p + 1/q = 1$,

$$\begin{aligned}
E_1(y) \le &\left(\int_{\mathbb{R}^d} L_f^{2q}(z_2)\delta_y Q_\gamma^{n,\ell}(\mathrm{d}z_2)\right)^{1/q} \\
&\times \left(\int_{\mathbb{R}^d} \mathbb{1}_{\{\|z_1\|\ge\mathsf{R}\}} \delta_x Q_\gamma^{n,\ell}(\mathrm{d}z_1) + \int_{\mathbb{R}^d} \mathbb{1}_{\{\|z_2\|\ge\mathsf{R}\}} \delta_y Q_\gamma^{n,\ell}(\mathrm{d}z_2)\right)^{1/p} .
\end{aligned}$$

Under **L**1-2, the first term on the right hand side is dominated by a constant for q small enough, and the second term tends to 0 for R large enough, uniformly for y in a compact neighborhood of x by [DM16, Theorem 3] and

$$\int_{\mathbb{R}^d} \mathbb{1}_{\{\|z_2\|\ge\mathsf{R}\}} \delta_y Q_\gamma^{n,\ell}(\mathrm{d}z_2) \le \mathsf{R}^{-2} \int_{\mathbb{R}^d} \|z_2\|^2\, \delta_y Q_\gamma^{n,\ell}(\mathrm{d}z_2) .$$

We can then choose R such that $E_1(y) \leq \varepsilon/3$. We consider now $E_2(y)$. L_f^2 is a continuous function, uniformly continuous on a compact set and we can then choose η such that $E_2(y) \leq \varepsilon/3$. We finally consider $E_3(y)$. By Markov's inequality and $\lim_{y\to x} W_2^2(\delta_x Q_\gamma^{n,\ell}, \delta_y Q_\gamma^{n,\ell}) = 0$, there exists a compact neighborhood $\mathcal{V}(x)$ of x such that $y \in \mathcal{V}(x)$ implies $E_3(y) \leq \varepsilon/3$.

□

Lemma 5.19. *Assume **H**10 and **H**11(m) for $m > 0$. Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 2/(m+L)$ and $N \geq 0$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be such that $Q_\gamma^{k+1,N} f$ is locally Lipschitz for $k \in \{1, \dots, N\}$. Then for all $x \in \mathbb{R}^d$,*

$$\int_{\mathbb{R}^d} Q_\gamma^N(x, \mathrm{d}y) \left\{ f(y) - Q_\gamma^N f(x) \right\}^2 \leq 2 \sum_{k=1}^N \gamma_k \int_{\mathbb{R}^d} Q_\gamma^k(x, \mathrm{d}y) \left\| \nabla Q_\gamma^{k+1,N} f(y) \right\|^2 .$$

Proof. Using $\mathbb{E}_x^{\mathcal{G}_k}[f(X_N)] = Q_\gamma^{k+1,N} f(X_k)$, we get

$$\begin{aligned}
\mathrm{Var}_x[f(X_N)] &= \sum_{k=1}^N \mathbb{E}_x \left[\mathbb{E}_x^{\mathcal{G}_{k-1}} \left[\left(\mathbb{E}_x^{\mathcal{G}_k}[f(X_N)] - \mathbb{E}_x^{\mathcal{G}_{k-1}}[f(X_N)] \right)^2 \right] \right] \\
&= \sum_{k=1}^N \mathbb{E}_x \left[R_{\gamma_k} \left\{ Q_\gamma^{k+1,N} f(\cdot) - R_{\gamma_k} Q_\gamma^{k+1,N} f(X_{k-1}) \right\}^2 (X_{k-1}) \right] .
\end{aligned}$$

Eq. (5.82) implies that

$$\mathrm{Var}_x[f(X_N)] \leq 2 \sum_{k=1}^N \gamma_k \int_{\mathbb{R}^d} Q_\gamma^k(x, \mathrm{d}y) \left\| \nabla Q_\gamma^{k+1,N} f(y) \right\|^2 ,$$

which concludes the proof. □

Proposition 5.20. *Assume **H**10 and **H**11(m) for $m > 0$. Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfying **L**1 and $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 2/(m+L)$. Then for all $N \geq 0$, $n \geq 1$, we get*

$$\mathrm{Var}_x \left[\hat{\pi}_n^N(f) \right] \leq \frac{8 C_Q(x)}{\kappa^2 \Gamma_{N+2,N+n+1}} \left\{ 1 + \Gamma_{N+2,N+n+1}^{-1} \left(\kappa^{-1} + \frac{2}{m+L} \right) \right\} . \tag{5.89}$$

Proof. For $k \in \{N, \dots, N+n-1\}$ and for all $y, x \in \mathbb{R}^d$, we have

$$\begin{aligned}
\left| \Phi_{n,k+1}^N(y) - \Phi_{n,k+1}^N(x) \right| = \Big| & \omega_{k+1,n}^N \{ f(y) - f(x) \} \\
& + \sum_{i=k+2}^{N+n} \omega_{i,n}^N \left\{ Q_\gamma^{k+2,i} f(y) - Q_\gamma^{k+2,i} f(x) \right\} \Big| .
\end{aligned} \tag{5.90}$$

By Lemma 5.18, $\Phi^N_{n,k+1}$ is locally Lipschitz and for almost all $x \in \mathbb{R}^d$,

$$\left\|\nabla \Phi^N_{n,k+1}(x)\right\| \leq \sum_{i=k+1}^{N+n} \omega^N_{i,n} \left\{ \prod_{\ell=k+2}^{i} (1-\kappa\gamma_\ell)^{1/2} \right\} (\delta_x Q_\gamma^{k+2,i} L_f^2)^{1/2} .$$

For $k \in \{N, \ldots, N+n-1\}$ and $x \in \mathbb{R}^d$, we have by (5.82) and the Cauchy-Schwarz inequality,

$$R_{\gamma_{k+1}} \left\{ \Phi^N_{n,k+1}(\cdot) - R_{\gamma_{k+1}} \Phi^N_{n,k+1}(x) \right\}^2 (x) \leq 2\gamma_{k+1} \Omega^N_{k,n} \left\{ \sum_{i=k+1}^{N+n} \omega^N_{i,n} \prod_{\ell=k+2}^{i} (1-\kappa\gamma_\ell)^{1/2} (\delta_x Q_\gamma^{k+1,i} L_f^2) \right\} ,$$

where,

$$\Omega^N_{k,n} = \sum_{i=k+1}^{N+n} \omega^N_{i,n} \prod_{\ell=k+2}^{i} (1-\kappa\gamma_\ell)^{1/2} . \tag{5.91}$$

By **L**1-2, we get for $k \in \{N, \ldots, N+n-1\}$

$$\mathbb{E}_x \left[R_{\gamma_{k+1}} \left\{ \Phi^N_{n,k+1}(\cdot) - R_{\gamma_{k+1}} \Phi^N_{n,k+1}(X_k) \right\}^2 (X_k) \right] \leq 2\gamma_{k+1} \Omega^N_{k,n} \left\{ \sum_{i=k+1}^{N+n} \omega^N_{i,n} \prod_{\ell=k+2}^{i} (1-\kappa\gamma_\ell)^{1/2} (\delta_x Q_\gamma^{i} L_f^2) \right\} \leq 2\gamma_{k+1} C_Q(x) (\Omega^N_{k,n})^2 .$$

Using $(1-t)^{1/2} \leq (1-t/2)$ for $t \in [0,1]$, we have

$$\Omega^N_{k,n} \leq (\kappa \Gamma_{N+2,N+n+1}/2)^{-1} . \tag{5.92}$$

Using this inequality, we get

$$\sum_{k=N}^{N+n-1} \mathbb{E}_x \left[R_{\gamma_{k+1}} \left\{ \Phi^N_{n,k+1}(\cdot) - R_{\gamma_{k+1}} \Phi^N_{n,k+1}(X_k) \right\}^2 (X_k) \right] \leq 8 C_Q(x) \Gamma_{N+1,N+n} / (\kappa \Gamma_{N+2,N+n+1})^2 . \tag{5.93}$$

We now bound $\mathrm{Var}_x \left[\Psi^N_n(X_N) \right]$. Since for all $x \in \mathbb{R}^d$, we have

$$\Psi^N_n(x) = \sum_{i=N+1}^{N+n} \omega^N_{i,n} Q_\gamma^{N+1,i} f(x) ,$$

by Lemma 5.18, $Q_\gamma^{k+1,N} \Psi^N_n$ is locally Lipschitz for $k \in \{1, \ldots, N\}$ with for almost all $x \in \mathbb{R}^d$,

$$\left\| \nabla Q_\gamma^{k+1,N} \Psi^N_n(x) \right\| \leq \sum_{i=N+1}^{N+n} \omega^N_{i,n} \prod_{\ell=k+1}^{i} (1-\kappa\gamma_\ell)^{1/2} \left(\delta_x Q_\gamma^{k+1,i} L_f^2 \right)^{1/2} .$$

Isolating the term $\prod_{\ell=k+1}^{N}(1-\kappa\gamma_\ell)^{1/2}$ and since $(1-\kappa\gamma_{N+1})^{1/2} \leq 1$, the Cauchy-Schwarz inequality implies

$$\left\|\nabla Q_\gamma^{k+1,N}\Psi_n^N(x)\right\|^2 \leq \left\{\prod_{\ell=k+1}^{N}(1-\kappa\gamma_\ell)\right\} \times \Omega_{N,n}^N \sum_{i=N+1}^{N+n}\omega_{i,n}^N \prod_{\ell=N+1}^{i}(1-\kappa\gamma_\ell)^{1/2}\delta_x Q_\gamma^{k+1,i} L_f^2 \,.$$

Plugging this inequality in Lemma 5.19, using **L**1-2, $\sum_{k=1}^{N}\gamma_k\prod_{i=k+1}^{N}(1-\kappa\gamma_i) \leq \kappa^{-1}$ and (5.92), we get

$$\mathrm{Var}_x\left[\Psi_n^N(X_N)\right] \leq 2\kappa^{-1}C_Q(x)(\kappa/2)^{-2}\Gamma_{N+2,N+n+1}^{-2} \,. \tag{5.94}$$

Combining (5.93) and (5.94) in (5.83) concludes the proof. □

5.5 Proofs

5.5.1 Proofs of propositions 5.2, 5.3, 5.4

We assume in this Section that **H**10 and **H**11(m) for some $m \geq 0$ hold. The proofs rely on the results given in Section 5.4, Propositions 5.17 and 5.20 which establish bounds on the mean squared error for locally Lipschitz functions. For $i \in \{0,\dots,M-1\}$, $\sigma_i^2 > 0$ and $\gamma_i > 0$, consider the Markov chain $(X_{i,n})_{n\geq 0}$ (5.8) and its associated Markov kernel R_i defined for all $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ by

$$R_i(x,\mathrm{A}) = \int_{\mathrm{A}}(4\pi\gamma_i)^{-d/2}\exp\left(-(4\gamma_i)^{-1}\left\|y-x+\gamma_i\nabla U_i(x)\right\|^2\right)\mathrm{d}y \,. \tag{5.95}$$

Under **H**10 and **H**11(m) for $m \geq 0$, [Nes13, Theorems 2.1.12, 2.1.9] show the following useful inequalities for all $x, y \in \mathbb{R}^d$,

$$\langle\nabla U_i(y)-\nabla U_i(x), y-x\rangle \geq \frac{\kappa_i}{2}\left\|y-x\right\|^2 + \frac{1}{m_i+L_i}\left\|\nabla U_i(y)-\nabla U_i(x)\right\|^2 \,, \tag{5.96}$$

$$\langle\nabla U_i(y)-\nabla U_i(x), y-x\rangle \geq m_i\left\|y-x\right\|^2 \,, \tag{5.97}$$

where L_i, m_i are defined in (5.13) and κ_i in (5.14). We then check **L**1 for g_i, where $g_i : \mathbb{R}^d \to \mathbb{R}$ is defined in (5.6). Note that g_i is continuously differentiable and for $x \in \mathbb{R}^d$, $\nabla g_i(x) = 2a_i x \mathrm{e}^{a_i\|x\|^2}$. Define $L_{g_i} : \mathbb{R}^d \to \mathbb{R}_+$ for $x \in \mathbb{R}^d$ by,

$$L_{g_i}(x) = 2a_i\left\|x\right\|\mathrm{e}^{a_i\|x\|^2} \tag{5.98}$$

We have for all $x, y \in \mathbb{R}^d$:

$$|g_i(y)-g_i(x)| = \left|\int_0^1\langle\nabla g_i(ty+(1-t)x), y-x\rangle\,\mathrm{d}t\right| \leq \left\|y-x\right\|\max(L_{g_i}(x), L_{g_i}(y)) \,, \tag{5.99}$$

which implies that **L**1-1 holds for g_i. The following Lemmas 5.21 and 5.22 enable to check **L**1-2 for g_i.

Lemma 5.21. *Assume **H**10 and **H**11(m) for $m \geq 0$. For all $\sigma_i^2 \in (0, +\infty)$, $n \in \mathbb{N}$, $\gamma_i \in (0, 2/(m_i + L_i)]$, $a_i \in [0, \kappa_i/8 \wedge (2\sigma_i^2)^{-1})$ and $x \in \mathbb{R}^d$, we have:*

$$\sup_{n\in\mathbb{N}} R_i^n \left(L_{g_i}^2\right)(x) \leq 4a_i^2 g_i^2(x) C_{i,0} \left\{\|x\|^2 + C_{i,1}\right\} ,$$

where L_{g_i} is defined in (5.98) and $C_{i,0}, C_{i,1}$ in (5.23).

Proof. In the proof, the subscript i is not specified for ease of notation. Let $\gamma \in (0, 2/(m + L)]$. Note that for all $\alpha \in [0, (4\gamma)^{-1})$, we have

$$\begin{aligned} R_\gamma(\mathrm{e}^{\alpha\|\cdot\|^2})(x) &= \frac{\mathrm{e}^{-(4\gamma)^{-1}\|x-\gamma\nabla U(x)\|^2}}{(4\pi\gamma)^{d/2}} \int_{\mathbb{R}^d} \mathrm{e}^{(\alpha-(4\gamma)^{-1})\|y\|^2 + (2\gamma)^{-1}\langle y, x-\gamma\nabla U(x)\rangle} \mathrm{d}y \\ &= \phi(x) , \end{aligned}$$

where $\phi(x) = (1-4\gamma\alpha)^{-d/2} \exp\{(\alpha/(1-4\alpha\gamma)) \|x - \gamma\nabla U(x)\|^2\}$. By the Leibniz integral rule and (5.96), we obtain:

$$\begin{aligned} &R_\gamma(\|\cdot\|^2 \mathrm{e}^{\alpha\|\cdot\|^2})(x) = \partial_\alpha R_\gamma(\mathrm{e}^{\alpha\|\cdot\|^2})(x) \\ &= (1-4\gamma\alpha)^{-d/2-1} \left\{2\gamma d + \frac{\|x-\gamma\nabla U(x)\|^2}{1-4\alpha\gamma}\right\} \exp\left(\frac{\alpha}{1-4\alpha\gamma} \|x-\gamma\nabla U(x)\|^2\right) \\ &\leq (1-4\gamma\alpha)^{-d/2-1} \left\{2\gamma d + \frac{1-\kappa\gamma}{1-4\alpha\gamma} \|x\|^2\right\} \exp\left(\frac{\alpha(1-\kappa\gamma)}{1-4\alpha\gamma} \|x\|^2\right) . \end{aligned}$$

Let $a \in [0, \kappa/8)$. Since $a < (4\gamma)^{-1}$, by a straightforward induction we have

$$\begin{aligned} \delta_x R_\gamma^p(\|\cdot\|^2 \mathrm{e}^{2a\|\cdot\|^2}) &\leq (1-4\gamma\alpha_0)^{-d/2-1} \exp\left(\alpha_p \|x\|^2\right) \\ &\quad \times \sum_{\ell=0}^{p-1} 2\gamma d \alpha_\ell \alpha_0^{-1} \left\{\prod_{k=1}^{\ell} (1-4\gamma\alpha_k)^{-d/2-1}\right\} \left\{\prod_{k=\ell+1}^{p-1} (1-4\gamma\alpha_k)^{-d/2}\right\} \\ &\quad + (1-4\gamma\alpha_0)^{-d/2-1} \left\{\prod_{k=1}^{p-1} (1-4\gamma\alpha_k)^{-d/2-1}\right\} \alpha_p \alpha_0^{-1} \|x\|^2 \exp\left(\alpha_p \|x\|^2\right) \\ &\leq \frac{1}{\alpha_0} \exp\left(\alpha_p \|x\|^2\right) \left\{\prod_{k=0}^{p-1} (1-4\gamma\alpha_k)^{-d/2-1}\right\} \left\{\alpha_p \|x\|^2 + 2d\gamma \sum_{\ell=0}^{p-1} \alpha_\ell\right\} , \end{aligned} \tag{5.100}$$

where $(\alpha_\ell)_{\ell\in\mathbb{N}}$ is the decreasing sequence defined for $\ell \geq 1$ by:

$$\alpha_0 = 2a, \quad \alpha_\ell = \alpha_{\ell-1} \frac{(1-\kappa\gamma)}{1-4\alpha_{\ell-1}\gamma} . \tag{5.101}$$

We now bound the right-hand-side of (5.100). First, by using the following inequality,

$$\log(1-4\gamma\alpha) = -4\alpha\int_0^\gamma (1-4\alpha t)^{-1}\mathrm{d}t \geq -4\alpha\gamma(1-4\alpha\gamma)^{-1}\,,$$

we have:

$$\begin{aligned}\prod_{k=0}^{p-1}(1-4\gamma\alpha_k)^{-d/2-1} &= \exp\left(-\left(\frac{d}{2}+1\right)\sum_{k=0}^{p-1}\log(1-4\alpha_k\gamma)\right)\\ &\leq \exp\left(\left(\frac{d}{2}+1\right)\frac{4\gamma}{1-\kappa\gamma}\sum_{k=0}^{p-1}\alpha_k\frac{1-\kappa\gamma}{1-4\alpha_k\gamma}\right)\,. \end{aligned} \tag{5.102}$$

Second, by a straightforward induction we get for all $\ell \geq 0$, $\alpha_\ell \leq 2a\{(1-\kappa\gamma)(1-8a\gamma)^{-1}\}^\ell$. Using (5.101) and this result implies:

$$\sum_{k=0}^{p-1}\alpha_k\frac{1-\kappa\gamma}{1-4\alpha_k\gamma} = \sum_{k=1}^{p}\alpha_k \leq 2a\frac{1-\kappa\gamma}{\kappa\gamma-8a\gamma}\,, \qquad \sum_{\ell=0}^{p-1}\alpha_\ell \leq 2a\frac{1-8a\gamma}{\kappa\gamma-8a\gamma}\,.$$

Combining these inequalities and (5.102) in (5.100) concludes the proof. □

Lemma 5.22. *Assume* ***H*** *10 and* ***H*** *11(m) for $m \geq 0$. For all $\sigma_i^2 \in (0,+\infty)$ and $a_i \in [0, m_i/\{4(d+4)\} \wedge (2\sigma_i^2)^{-1}]$, we have*

$$\pi(L_{g_i}^2) \leq 4a_i^2 C_{i,2}\,,$$

where $C_{i,2}$ is defined in (5.23).

Proof. In the proof, the subscript i is not specified for ease of notations. Recall that the generator of the Langevin diffusion (5.7) associated to U is defined for any f in $\mathcal{C}^2(\mathbb{R}^d)$ by

$$\mathcal{L}f = -\langle\nabla U, \nabla f\rangle + \Delta f\,.$$

In particular, for $f(x) = \|x\|^2 \mathrm{e}^{2a\|x\|^2}$ and $x \in \mathbb{R}^d$, we have

$$\begin{aligned}\nabla f(x) &= 2(1+2a\|x\|^2)x\mathrm{e}^{2a\|x\|^2}\,,\\ \Delta f(x) &= \mathrm{e}^{2a\|x\|^2}\left\{16a^2\|x\|^4 + 4a(d+4)\|x\|^2 + 2d\right\}\,.\end{aligned}$$

Using (5.97) and $\nabla U(0) = 0$, we get

$$\mathcal{L}(\|\cdot\|^2\mathrm{e}^{2a\|\cdot\|^2})(x) \leq \mathrm{e}^{2a\|x\|^2}\left\{2d + 2\left(2a(d+4) - m\right)\|x\|^2 + 4a\left(4a-m\right)\|x\|^4\right\}\,.$$

Using that $a \in [0, m/(4(d+4))]$, we have $2a(4a-m) \leq -(8/5)am$. Then an elementary study of $t \mapsto \mathrm{e}^{2at}\left\{2d + 4a\left(4a-m\right)t^2\right\}$ on $\mathbb{R}_+$ shows that:

$$\sup_{x\in\mathbb{R}^d}\mathrm{e}^{2a\|x\|^2}\left\{2d + 4a\left(4a-m\right)\|x\|^4\right\} \leq 4d\,.$$

Therefore we get using $2(2a(d+4)-m) \leq -m$,

$$\mathcal{L}(\|\cdot\|^2 \, \mathrm{e}^{2a\|\cdot\|^2})(x) \leq -m \, \|x\|^2 \, \mathrm{e}^{2a\|x\|^2} + 4d \; .$$

Finally applying [MT93, Theorem 4.3-(ii)] shows the result. □

Proofs of Propositions 5.2 and 5.3. Lemmas 5.21 and 5.22 and proposition 5.17 prove the result. □

Proof of Proposition 5.4. The proof follows from Lemma 5.21 and proposition 5.20. □

5.5.2 Proof of Lemma 5.9

The case $K = 0$ being straightforward, assume $K \in \mathbb{N}^\star$. Using Markov's inequality, we have

$$\mathbb{P}(\mathrm{A}^{\mathrm{c}}_{\mathcal{S},\epsilon}) \leq \frac{4}{\epsilon^2} \frac{\mathbb{E}\left[\left(\prod_{i=0}^{M-1} \hat{\pi}_i(g_i) - \prod_{i=0}^{M-1} \pi_i(g_i)\right)^2\right]}{\left(\prod_{i=0}^{M-1} \pi_i(g_i)\right)^2} \; . \tag{5.103}$$

Since $\hat{\pi}_i(g_i)$ for $i \in \{0, \ldots, M-1\}$ are independent, we get

$$\frac{\mathbb{E}\left[\left(\prod_{i=0}^{M-1} \hat{\pi}_i(g_i) - \prod_{i=0}^{M-1} \pi_i(g_i)\right)^2\right]}{\left(\prod_{i=0}^{M-1} \pi_i(g_i)\right)^2} = F_1^2(F_2 - 1) + (F_1 - 1)^2 \; , \tag{5.104}$$

where

$$F_1 = \prod_{i=0}^{M-1} \mathbb{E}\left[\hat{\pi}_i(g_i)\right] / \pi_i(g_i) \; , \quad F_2 = \prod_{i=0}^{M-1} \mathbb{E}\left[\{\hat{\pi}_i(g_i)\}^2\right] / \mathbb{E}^2\left[\hat{\pi}_i(g_i)\right] \; .$$

In addition, since $\{0, \ldots, M-2\} = \cup_{k=0}^{K-1} \mathcal{I}_k$, we can consider the following decomposition

$$\begin{aligned} F_1 &= \prod_{k=0}^{K-1} \prod_{i \in \mathcal{I}_k} \left(1 + \frac{\mathbb{E}\left[\hat{\pi}_i(g_i)\right] - \pi_i(g_i)}{\pi_i(g_i)}\right) \\ &\qquad \times \left(1 + \frac{\mathbb{E}\left[\hat{\pi}_{M-1}(g_{M-1})\right] - \pi_{M-1}(g_{M-1})}{\pi_{M-1}(g_{M-1})}\right) \; , \\ F_2 &= \prod_{k=0}^{K-1} \prod_{i \in \mathcal{I}_k} \left(1 + \frac{\mathrm{Var}\left[\hat{\pi}_i(g_i)\right]}{\mathbb{E}\left[\hat{\pi}_i(g_i)\right]^2}\right) \left(1 + \frac{\mathrm{Var}\left[\hat{\pi}_{M-1}(g_{M-1})\right]}{\mathbb{E}\left[\hat{\pi}_{M-1}(g_{M-1})\right]^2}\right) \; . \end{aligned}$$

We now bound F_1, F_2 separately. Using $1 + t \leq \exp(t)$ for $t \in \mathbb{R}$ with $t = \eta/(K\,|\mathcal{I}_k|)$ and leaving the term $i = M-1$ out, we get by conditions i)-ii)

$$F_1 \leq (1+\eta) \exp(\eta) \; . \tag{5.105}$$

Since $\hat{\pi}_i(g_i) \geq 1$, we have $\mathrm{Var}\left[\hat{\pi}_i(g_i)\right] / \mathbb{E}\left[\hat{\pi}_i(g_i)\right]^2 \leq \eta^2 / K\left|\mathcal{I}_k\right|$. Therefore using $1+t \leq \exp(t)$ for $t \in \mathbb{R}$ with $t = \eta^2/(K\left|\mathcal{I}_k\right|)$ leaving the term $i = M-1$ out, we obtain by conditions i)-ii)

$$F_2 \leq \left(1+\eta^2\right)\exp(\eta^2)\ . \tag{5.106}$$

By combining (5.103), (5.104), (5.105) and (5.106), we get:

$$(\epsilon^2/4)\mathbb{P}(\mathrm{A}^{\mathrm{c}}_{\mathcal{S},\epsilon}) \leq (1+\eta)^2 \mathrm{e}^{2\eta}\left((1+\eta^2)\mathrm{e}^{\eta^2}-1\right) + ((1+\eta)\mathrm{e}^{\eta}-1)^2\ .$$

With $\eta \leq 1/8$ and $\mathrm{e}^t - 1 \leq t\mathrm{e}^t$ for $t \geq 0$, we have $(\epsilon^2/4)\mathbb{P}(\mathrm{A}^{\mathrm{c}}_{\mathcal{S},\epsilon}) \leq 9\eta^2$.

5.5.3 Proofs of Section 5.2.1

We preface the proofs by a technical lemma which gathers useful bounds and inequalities. We recall that in this Section the number of phases M is defined by (5.36)

$$M = \inf\left\{i \geq 1 : \sigma_{i-1}^2 \geq (2d+7)/m\right\}\ .$$

Lemma 5.23. *Assume **H**10 and **H**11(m) for $m > 0$. Let $\{\sigma_i^2\}_{i=0}^{M-1}$ defined by (5.34) for σ_0^2 given in (5.31) and M in (5.36).*

1. *$K \leq \lceil (1/\log(2))\log\{(2d+7)/(m\sigma_0^2)\}\rceil$ where K is defined in (5.39).*
2. *For all $k \in \{0,\dots,K-1\}$ and $i \in \mathcal{I}_k$, $2^{k+1}\sigma_0^2 a_i \left|\mathcal{I}_k\right| \leq 1$, where a_i is defined in (5.37) and $\mathcal{I}_k$ in (5.38).*
3. *For all $i \in \{0,\dots,M-1\}$ and $\gamma_i \leq 1/(m_i+L_i)$, there exist $\alpha_i \in [4,14]$ and $\beta_i \in [1,10]$ such that $C_{i,2}+C_{i,0}C_{i,1} = \alpha_i d m_i^{-1}$ and $C_{i,0}C_{i,1} = \beta_i d \kappa_i^{-1}$ where $C_{i,0}, C_{i,1}, C_{i,2}$ and κ_i are given in (5.23) and (5.14) respectively.*
4. *For all $i \in \{0,\dots,M-1\}$, $0 < A_{i,1} \leq 4dL_i^4\kappa_i^{-1}m_i^{-1}$, where L_i, m_i and κ_i are given in (5.13) and (5.14) respectively.*
5. *For all $i \in \{0,\dots,M-2\}$, $\kappa_i\sigma_i^2 \leq 4d+16$.*
6. *For all $i \in \{0,\dots,M-1\}$, $\sqrt{m_i}/(\kappa_i\sigma_i) \leq 1$*
7. *For all $i \in \{0,\dots,M-1\}$,*

$$\frac{m_i+L_i}{2m_i} \leq \frac{m+L}{2m}\ , \quad \frac{L_i^2}{\kappa_i^3\sigma_i^4 m_i} \leq \left(\frac{m+L}{2m}\right)^3\ .$$

8. *For all $k \in \{0,\dots,K-1\}$ and $i \in \mathcal{I}_k$,*

$$\kappa_i^{-2}m_i^{-1/2}\sigma_i^{-2} \leq \frac{(2^{k+1}\sigma_0^2)^{3/2}}{(1+m2^k\sigma_0^2)^{5/2}}\ , \quad \frac{L_i^2 m_i^{-1/2}}{\kappa_i^2\sigma_i^2 m_i^{1/2}} \leq \left(\frac{m+L}{2m}\right)^2 \frac{1}{1+m2^k\sigma_0^2}\ .$$

Proof. 1. By (5.36) and (5.38),

$$K \leq \inf\left\{k \geq 0 : 2^k \sigma_0^2 \geq \frac{2d+7}{m}\right\} = \left\lceil \frac{1}{\log(2)} \log\left(\frac{2d+7}{m\sigma_0^2}\right)\right\rceil \ .$$

2. Let $k \in \{0, \ldots, K-1\}$. Denote by $i_0 = \inf \mathcal{I}_k$. By (5.37) and (5.38),

$$|\mathcal{I}_k|\, a_i = \sum_{i \in \mathcal{I}_k} a_i \leq \frac{1}{2\sigma_{i_0}^2} \leq \frac{1}{2^{k+1}\sigma_0^2} \ ,$$

and the proof follows.

3. Let $k \in \{0, \ldots, K-1\}$ and $i \in \mathcal{I}_k$, Since $m_i \geq m + (2^{k+1}\sigma_0^2)^{-1}$, $a_i \leq m_i/\{4(d+4)\}$. Therefore using in addition that $\gamma_i \leq 1/(m_i + L_i)$, we have $\kappa_i - 8a_i \geq \kappa_i(d+2)/(d+4)$ and $1 - 8a_i\gamma_i \geq (d+3)/(d+4)$. The definition of $C_{i,0}, C_{i,1}, C_{i,2}$ (5.23) completes the proof.

4. The upper bound is a straightforward consequence of $(1 + \kappa_i(m_i + L_i)^{-1})(1 + 6^{-1}m_i(m_i + L_i)^{-1}) \leq 2$.

5. The bound follows using that $\sigma_{M-2}^2 \leq (2d+7)/m$ by (5.36) and the sequence $\{\kappa_i\sigma_i^2\}_{i=0}^{M-2}$ is non-decreasing since

$$\kappa_i\sigma_i^2 = 2\left\{1 + \frac{mL\sigma_i^2 - 1/\sigma_i^2}{m + L + 2/\sigma_i^2}\right\} \ ,$$

and $\{\sigma_i^2\}_{i=0}^{M-2}$ is non-decreasing.

6. The proof is a direct consequence of the fact that the sequence $i \mapsto \sqrt{m_i}/(\kappa_i\sigma_i)$ is non-increasing since $m < L$, $\{\sigma_i^2\}_{i=0}^{M-2}$ is non-decreasing and

$$\frac{\sqrt{m_i}}{\kappa_i\sigma_i} = \frac{1}{2}\frac{1}{\sqrt{1 + m\sigma_i^2}}\left\{1 + \frac{1 + m\sigma_i^2}{1 + L\sigma_i^2}\right\} \ .$$

7. Using that $\{\sigma_i^2\}_{i=0}^{M-2}$ is non-decreasing and

$$\frac{m_i + L_i}{2m_i} = \frac{2 + (m+L)\sigma_i^2}{2 + 2m\sigma_i^2} \ , \quad \frac{L_i^2}{\kappa_i^3\sigma_i^4 m_i} \leq \left(\frac{(m+L)\sigma_i^2 + 2}{(2m)\sigma_i^2 + 2}\right)^3 \ ,$$

concludes the proof.

8. Let $k \in \{0, \ldots, K-1\}$ and $i \in \mathcal{I}_k$. Since $\kappa_i \geq m_i$ and $2^k\sigma_0^2 \leq \sigma_i^2 \leq 2^{k+1}\sigma_0^2$, we have

$$\kappa_i^{-2} m_i^{-1/2} \sigma_i^{-2} \leq \frac{\sigma_i^3}{(m_i\sigma_i^2)^{5/2}} \leq \frac{(2^{k+1}\sigma_0^2)^{3/2}}{(1 + m2^k\sigma_0^2)^{5/2}} \ ,$$

and

$$\frac{L_i^2}{\kappa_i^2} \leq \left(\frac{m+L}{2m}\right)^2 \ , \quad \frac{1}{m_i\sigma_i^2} \leq \frac{1}{1 + m2^k\sigma_0^2} \ .$$

□

Proof of Lemma 5.8

Because U satisfies **H**10, **H**11(m) for $m \geq 0$ and $U(0) = 0$, $\nabla U(0) = 0$, we have:

$$\exp(-(L/2)\,\|x\|^2) \leq \exp(-U(x)) \leq \exp(-(m/2)\,\|x\|^2)\,,$$

which implies by integration that,

$$(2\pi\sigma_0^2)^{d/2}/(1+\sigma_0^2 L)^{d/2} \leq Z_0 \leq (2\pi\sigma_0^2)^{d/2}/(1+\sigma_0^2 m)^{d/2}\,,$$

where $Z_0 = \int_{\mathbb{R}^d} \mathrm{e}^{-U_0}$ and U_0 is defined in (5.2). The proof follows from the expression of σ_0^2 and the bound,

$$\left(\frac{1+L\sigma_0^2}{1+m\sigma_0^2}\right)^{d/2} \leq \exp\left(\frac{d}{2}\sigma_0^2(L-m)\right)\,.$$

Proof of Lemma 5.10

Let $k \in \{0,\ldots,K-1\}$ and $i \in \mathcal{I}_k$. Assume that $\gamma_i \leq (m_i + L_i)^{-1}$. By Proposition 5.2, Proposition 5.4, Lemma 5.23-2 and $\sigma_i^2 \leq 2^{k+1}\sigma_0^2$, to check condition-i) of Lemma 5.9, it is then sufficient for γ_i, n_i, N_i to satisfy,

$$\frac{4d}{n_i m_i \kappa_i \gamma_i} \exp\left(-N_i \frac{\kappa_i \gamma_i}{2}\right) + 2\kappa_i^{-1}\left(A_{i,0}\gamma_i + A_{i,1}\gamma_i^2\right) \leq \frac{\eta^2}{4K^2}\frac{\sigma_i^4}{C_{i,2} + C_{i,0}C_{i,1}}\,, \tag{5.107}$$

$$\frac{32 a_i C_{i,0} C_{i,1}}{\kappa_i^2 n_i \gamma_i}\left(1 + \frac{2}{\kappa_i n_i \gamma_i}\right) \leq \frac{\sigma_i^2 \eta^2}{K}\,. \tag{5.108}$$

By (5.24), Lemma 5.23-3 and Lemma 5.23-4, there exist $\alpha_i \in [4,14]$ and $\beta_i \in [1,10]$ such that these two inequalities hold if γ_i, n_i, N_i satisfy

$$2L_i^2\kappa_i^{-1} d\gamma_i + 4dL_i^4\kappa_i^{-1}m_i^{-1}\gamma_i^2 \leq \frac{\eta^2}{16K^2}\frac{\kappa_i m_i \sigma_i^4}{\alpha_i d}\,, \tag{5.109}$$

$$\frac{1}{n_i}\left(1 + \frac{2}{\kappa_i n_i \gamma_i}\right) \leq \frac{\eta^2 \sigma_i^2 \kappa_i^3 \gamma_i}{32 K a_i \beta_i d}\,, \tag{5.110}$$

$$N_i \geq \bar{N}_i = -2(\kappa_i\gamma_i)^{-1}\log\left(\frac{\eta^2\sigma_i^4 m_i^2 n_i \kappa_i \gamma_i}{32K^2\alpha_i d^2}\right)\,. \tag{5.111}$$

These inequalities are shown to be true successively for γ_i, n_i and N_i chosen as in the statement of the Lemma. Denote by $\bar{\gamma}_i$ and $\bar{n}_i^{-1}$ the positive roots associated to (5.109) and (5.110) seen as equalities and given by

$$\bar{\gamma}_i = 4^{-1}L_i^{-2}m_i\left(-1 + \sqrt{1 + \frac{\eta^2\kappa_i^2\sigma_i^4}{4\alpha_i K^2 d^2}}\right)\,, \tag{5.112}$$

$$\bar{n}_i^{-1} = 4^{-1}\kappa_i\gamma_i\left(-1 + \sqrt{1 + \frac{\eta^2\kappa_i^2\sigma_i^2}{4K a_i \beta_i d}}\right)\,. \tag{5.113}$$

Note that for (5.109) and (5.110) to hold, it suffices that $\gamma_i \leq \bar{\gamma}_i$ and $n_i \geq \bar{n}_i$. We now lower bound $\bar{\gamma}_i$ and upper bound $\bar{n}_i$.

Using that $\sqrt{1+t} \geq 1+2^{-1}t(1+t)^{-1/2}$ for $t = \eta^2\kappa_i^2\sigma_i^4/(4\alpha_i K^2 d^2)$ and $(\eta^2\kappa_i^2\sigma_i^4)/(4\alpha_i K^2 d^2) \leq 25$ by $\alpha_i \geq 4$ and Lemma 5.23-5, concludes that if (5.40) holds then $\gamma_i \leq \bar{\gamma}_i$. The fact that $\gamma_i \leq (m_i + L_i)^{-1}$ can be checked by simple algebra.

First, by (5.37) and the definition of $\mathcal{I}_k$, $a_i \leq m_i/\{4(d+4)\}$,

$$\bar{n}_i^{-1} \geq 4^{-1}\kappa_i\gamma_i \left(-1 + \sqrt{1 + \frac{\eta^2\kappa_i^2\sigma_i^2(d+4)}{Km_i\beta_i d}} \right) .$$

Then using that $\sqrt{1+t} \geq 1+2^{-1}t(1+t)^{-1/2}$ for $t = \eta^2\kappa_i^2\sigma_i^2(d+4)/(Km_i\beta_i d)$ and $\beta_i \geq 1$ concludes that if (5.41) holds then $n_i \geq \bar{n}_i$. Finally, we have by (5.41), $(n_i\kappa_i\gamma_i)^{-1} \leq \eta^2\kappa_i\sigma_i/(196\sqrt{m_i}K)$, which gives with $\kappa_i \geq m_i$,

$$\begin{aligned} \bar{N}_i &\leq 2(\kappa_i\gamma_i)^{-1} \log\left\{ \frac{64\alpha_i}{196} K d^2 (1 + m\sigma_i^2)^{-3/2} \right\} \\ &\leq 2(\kappa_i\gamma_i)^{-1} \log\left(5Kd^2\right) , \end{aligned}$$

which concludes that (5.42) implies (5.111).

The same reasoning applies to check condition-ii) of Lemma 5.9. The details are gathered in the appendix Section 5.A.1.

Proof of Theorems 5.5 and 5.6 and Corollary 5.7

For $i \in \{0, \ldots, M-1\}$, set γ_i, n_i, N_i such that (5.40), (5.41), (5.42), (5.43), (5.44) and (5.45) are equalities. By (5.18), we consider the following decomposition for the cost $= A + B$ where $A = \sum_{i=0}^{M-2}\{N_i + n_i\}$ and $B = n_{M-1} + N_{M-1}$. We bound A and B separately.

First Lemma 5.23-6 implies that for all $i \in \{0, \ldots, M-2\}$, $n_i \leq (196K)/(\eta^2\kappa_i\gamma_i)$ and therefore using Lemma 5.23-7

$$A \leq \left(\frac{196K}{\eta^2} + 2\log(5Kd^2) \right) \frac{2285K^2d^2}{\eta^2} \left(\frac{m+L}{2m} \right)^3 (M-1) . \tag{5.114}$$

We now give a bound on $M-1$. Define

$$K_{\mathrm{int}} = \sup\left\{ k \geq 1 : m2^k\sigma_0^2 \leq 1 \right\} \wedge K \leq \left\lfloor -\frac{\log(m\sigma_0^2)}{\log(2)} \right\rfloor . \tag{5.115}$$

By Lemma 5.23-2 and (5.37), we have

$$\frac{M-1}{4(d+4)} = \sum_{k=0}^{K-1} \frac{|\mathcal{I}_k|}{4(d+4)} \leq K_{\mathrm{int}} + 2 . \tag{5.116}$$

Note that $K_{\text{int}} \le K \le C$ by (5.115) and Lemma 5.23-1. Combining (5.114), Lemma 5.23-7 and (5.116), we get

$$\sum_{i=0}^{M-2} N_i + n_i \le \left(\frac{98K}{\eta^2} + \log(5Kd^2)\right) \frac{4570K^2d^2}{\eta^2} \left(\frac{m+L}{2m}\right)^3 4(d+4)(C+2) \,. \quad (5.117)$$

Regarding the term $i = M-1$, we have

$$n_{M-1} + N_{M-1} \le \left(\frac{19}{\eta^2} + 1\right) \frac{40}{\eta^2} \frac{m+L}{2m} \frac{L}{m} \,. \quad (5.118)$$

Replacing η by $(\epsilon\sqrt{\mu})/8$ and combining (5.117) and (5.118) gives (5.28).

Assume **H**12. We now prove Theorem 5.6 and use Lemma 5.11 instead of Lemma 5.10. For $i \in \{0, \dots, M-1\}$, set γ_i, n_i, N_i such that (5.46), (5.41), (5.42), (5.47), (5.44) and (5.45) are equalities. By (5.18), we have the decomposition cost $= A + B$ where $A = \sum_{i=0}^{M-2}\{N_i + n_i\}$ and $B = n_{M-1} + N_{M-1}$. Lemma 5.23-6 implies that for all $i \in \{0, \dots, M-2\}$, $n_i \le (196K)/(\eta^2\kappa_i\gamma_i)$, and using that for $a, b \ge 0$, $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, we have

$$A \le \left(\frac{196K}{\eta^2} + 2\log(5Kd^2)\right) \sqrt{\frac{7}{3}} \frac{8Kd}{\eta} \sum_{i=0}^{M-2} \frac{d^{1/2}\tilde{L} + \sqrt{10}L_i^2 m_i^{-1/2}}{\kappa_i^2\sigma_i^2 m_i^{1/2}} \,.$$

Then, by Lemma 5.23-2 and Lemma 5.23-8, and splitting the sum in two parts $k \le K_{\text{int}}$ and $k > K_{\text{int}}$,

$$\begin{aligned}
\sum_{i=0}^{M-2} \frac{1}{\kappa_i^2\sigma_i^2 m_i^{1/2}} &\le \sum_{k=0}^{K-1} \sum_{i\in\mathcal{I}_k} \frac{(2^{k+1}\sigma_0^2)^{3/2}}{(1+m2^k\sigma_0^2)^{5/2}} \\
&\le 4(d+4)\frac{2^{3/2}}{m^{3/2}} \sum_{k=0}^{K-1} \frac{(m2^k\sigma_0^2)^{3/2}}{(1+m2^k\sigma_0^2)^{7/2}} \\
&\le 4(d+4)\frac{2^{3/2}}{m^{3/2}} \left(K_{\text{int}} + \sum_{k=K_{\text{int}}+1}^{K-1} (m2^k\sigma_0^2)^{-2}\right) \\
&\le 4(d+4)\frac{2^{3/2}}{m^{3/2}} \left(K_{\text{int}} + \frac{4}{3}\right) \,.
\end{aligned}$$

We have similarly by Lemma 5.23-2 and Lemma 5.23-8,

$$\sum_{i=0}^{M-2} \frac{L_i^2 m_i^{-1/2}}{\kappa_i^2\sigma_i^2 m_i^{1/2}} \le 4(d+4)\left(\frac{m+L}{2m}\right)^2 (K_{\text{int}}+2) \,.$$

Combining these inequalities with

$$n_{M-1} + N_{M-1} \le \left(\frac{19}{\eta^2} + 1\right) \sqrt{\frac{7}{3}} \frac{4}{\eta} \left\{\frac{d^{1/2}\tilde{L}}{m^{3/2}} + \sqrt{10}\left(\frac{m+L}{2m}\right)^2\right\} \,, \quad (5.119)$$

and replacing η by $(\epsilon\sqrt{\mu})/8$ establish (5.30).

Proof of Corollary 5.7. Let $N = \lceil 4\log(\tilde{\mu}^{-1}) \rceil$ and $(\hat{Z}_i)_{i\in\{1,\dots,2N+1\}}$ be $2N+1$ independent outputs of the algorithms of Theorems 5.5 and 5.6 with $\mu = 1/4$, sorted by increasing order. Denote by $\hat{Z} = \hat{Z}_{N+1}$ the median of $(\hat{Z}_i)_{i\in\{1,\dots,2N+1\}}$. In addition, define the independent Bernoulli random variables $(W_i)_{i\in\{1,\dots,2N+1\}}$ by

$$W_i = \mathbb{1}_{\mathsf{A}_i} \ , \text{ where } \mathsf{A}_i = \left\{ \left| \hat{Z}_i/Z - 1 \right| \geq \epsilon \right\} \ .$$

Since $\hat{Z}$ is the median of $(\hat{Z}_i)_{i\in\{1,\dots,2N+1\}}$, we have

$$\mathbb{P}\left(\left| \hat{Z}/Z - 1 \right| > \epsilon \right) \leq \mathbb{P}\left(\sum_{i=1}^{2N+1} W_i \geq N+1 \right) \ .$$

In addition since $\mathbb{P}(W_i = 1) \leq 1/4$, we have by [PS76, Corollary 5.2]

$$\mathbb{P}\left(\sum_{i=1}^{2N+1} W_i \geq N+1 \right) \leq \mathbb{P}\left(\sum_{i=1}^{2N+1} \tilde{W}_i \geq N+1 \right) \ ,$$

where $(\tilde{W}_i)_{i\in\{1,\dots,2N+1\}}$ are i.i.d. Bernoulli random variables with parameter $1/4$. Then by Hoeffding's inequality [BLM13, Theorem 2.8] and using for all $t \geq 1$, $8(t/2+3/4)^2/\{t(2t+1)\} \geq 1$, we get

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{2N+1} \tilde{W}_i \geq N+1 \right) &\leq \mathbb{P}\left(\sum_{i=1}^{2N+1} \tilde{W}_i - (1/4)(2N+1) \geq N/2 + 3/4 \right) \\
&\leq \exp\left(\frac{-2(N/2+3/4)^2}{2N+1} \right) \\
&\leq \exp(-N/4) \ ,
\end{aligned}$$

which concludes the proof. □

Acknowledgements

This work was supported by the Ecole Polytechnique Data Science Initiative. The authors thank the anonymous referees for useful suggestions which improved the manuscript.

5.A Additional proofs of Section 5.2.1

5.A.1 Proof of Lemma 5.10

In this Section, the proof for the case $i = M-1$ of Lemma 5.10 is dealt with. Note that $a_{M-1} = (2\sigma_{M-1}^2)^{-1}$. By Propositions 5.2 and 5.4, to check condition-ii) of Lemma 5.9,

it is then sufficient for $\gamma_{M-1}, n_{M-1}, N_{M-1}$ to satisfy,

$$\frac{4d}{n_{M-1}m_{M-1}\kappa_{M-1}\gamma_{M-1}} \exp\left(-N_{M-1}\frac{\kappa_{M-1}\gamma_{M-1}}{2}\right)$$
$$+ 2\kappa_{M-1}^{-1}\left(A_{M-1,0}\gamma_{M-1} + A_{M-1,1}\gamma_{M-1}^2\right) \leq \frac{\eta^2\sigma_{M-1}^4}{C_{M-1,2} + C_{M-1,0}C_{M-1,1}} , \tag{5.120}$$
$$\frac{8C_{M-1,0}C_{M-1,1}}{\kappa_{M-1}^2 n_{M-1}\gamma_{M-1}}\left(1 + \frac{2}{\kappa_{M-1}n_{M-1}\gamma_{M-1}}\right) \leq \sigma_{M-1}^4\eta^2 . \tag{5.121}$$

Then (5.120) and (5.121) are satisfied if,

$$2L_{M-1}^2\kappa_{M-1}^{-1}d\gamma_{M-1} + 4dL_{M-1}^4\kappa_{M-1}^{-1}m_{M-1}^{-1}\gamma_{M-1}^2 \leq \frac{\eta^2}{4}\frac{\kappa_{M-1}m_{M-1}\sigma_{M-1}^4}{\alpha_{M-1}d} , \tag{5.122}$$
$$\frac{1}{n_{M-1}}\left(1 + \frac{2}{\kappa_{M-1}n_{M-1}\gamma_{M-1}}\right) \leq \frac{\eta^2\sigma_{M-1}^4\kappa_{M-1}^3\gamma_{M-1}}{8\beta_{M-1}d} , \tag{5.123}$$
$$- 2(\kappa_{M-1}\gamma_{M-1})^{-1}\log\left(\frac{\eta^2\sigma_{M-1}^4 m_{M-1}^2 n_{M-1}\kappa_{M-1}\gamma_{M-1}}{8\alpha_{M-1}d^2}\right) = \bar{N}_{M-1} \leq N_{M-1} . \tag{5.124}$$

Denote by $\bar{\gamma}_{M-1}$ and $\bar{n}_{M-1}^{-1}$ the positive roots associated to (5.122) and (5.123) seen as equalities. We have:

$$\bar{\gamma}_{M-1} = 4^{-1}L_{M-1}^{-2}m_{M-1}\left(-1 + \sqrt{1 + \frac{\eta^2\kappa_{M-1}^2\sigma_{M-1}^4}{\alpha_{M-1}d^2}}\right) , \tag{5.125}$$
$$\bar{n}_{M-1}^{-1} = 4^{-1}\kappa_{M-1}\gamma_{M-1}\left(-1 + \sqrt{1 + \frac{\eta^2\kappa_{M-1}^2\sigma_{M-1}^4}{\beta_{M-1}d}}\right) . \tag{5.126}$$

Note that for (5.122) and (5.123) to hold, it suffices that $\gamma_{M-1} \leq \bar{\gamma}_{M-1}$ and $n_{M-1} \geq \bar{n}_{M-1}$. We now lower bound $\bar{\gamma}_{M-1}$ and upper bound $\bar{n}_{M-1}$.

Using that $t \geq 0$, $\sqrt{1+t} \geq 1 + 2^{-1}t(1+t)^{-1/2}$ for $t = (\eta^2\kappa_{M-1}^2\sigma_{M-1}^4)/(\alpha_{M-1}d^2)$ and $\kappa_{M-1}\sigma_{M-1}^2 d^{-1} \geq 2$, $\alpha_{M-1} \geq 4$ concludes that if (5.43) holds then $\gamma_{M-1} \leq \bar{\gamma}_{M-1}$. The fact that $\gamma_{M-1} \leq (m_{M-1} + L_{M-1})^{-1}$ can be checked by simple algebra.

Then using that $\sqrt{1+t} \geq 1 + 2^{-1}t(1+t)^{-1/2}$ for $t = (\eta^2\kappa_{M-1}^2\sigma_{M-1}^4)/(\beta_{M-1}d)$ and $\kappa_{M-1}\sigma_{M-1}^2 \geq 10$, $\beta_{M-1} \geq 1$ concludes that if (5.44) holds then $n_{M-1} \geq \bar{n}_{M-1}$. Finally, by (5.44), we get

$$\bar{N}_{M-1} \leq (\kappa_{M-1}\gamma_{M-1})^{-1}\log(7/3) ,$$

which concludes that (5.45) implies (5.124).

5.A.2 Proof of Lemma 5.11

Let $k \in \{0, \dots, K-1\}$ and $i \in \mathcal{I}_k$. Assume that $\gamma_i \leq (m_i + L_i)^{-1}$. The proof of Lemma 5.10 only needs to be slightly adapted. More precisely, Proposition 5.3 is applied instead of Proposition 5.2. By (5.26) and (5.27), we have

$$B_{i,0} \leq 3^{-1}d\kappa_i^{-1}(d\tilde{L}^2 + 10L_i^4 m_i^{-1}) , \quad B_{i,1} \leq (25/12)dL_i^4 m_i^{-1} . \tag{5.127}$$

It is sufficient for γ_i, n_i, N_i to satisfy (5.107) and (5.108) with $A_{i,0}\gamma_i + A_{i,1}\gamma_i^2$ replaced by $B_{i,0}\gamma_i^2 + B_{i,1}\gamma_i^3$. The counterpart of (5.109) is then

$$\frac{1}{3\kappa_i}\left(d\tilde{L}^2 + 10L_i^4 m_i^{-1}\right)\gamma_i^2 + \frac{25}{12}L_i^4 m_i^{-1}\gamma_i^3 \leq \frac{\eta^2\kappa_i m_i \sigma_i^4}{16K^2\alpha_i d^2} \,. \tag{5.128}$$

Since $\gamma_i \leq 1/(m_i + L_i)$ and $\kappa_i \leq L_i$, we have

$$(3\kappa_i)^{-1}\left(d\tilde{L}^2 + 10L_i^4 m_i^{-1}\right) \geq (25/12)L_i^4 m_i^{-1}\gamma_i \,,$$

which establishes that if (5.46) holds, then (5.128) is satisfied. $\gamma_i \leq (m_i + L_i)^{-1}$ can be checked by simple algebra. For $i = M-1$, the conclusion follows from $m_{M-1}\sigma_{M-1}^2 d^{-1} \geq 2$ because $\sigma_{M-1}^2 \geq (2d+7)/m$.

5.B Additional proofs of Section 5.2.2

First, we state a technical lemma that gathers useful bounds. We recall that M is defined in this Section by (5.56),

$$M = \inf\left\{i \geq 1 : \sigma_{i-1}^2 \geq D^2\right\} \,.$$

Lemma 5.24. *Assume* ***H****10 and* ***H****11(m) for $m \geq 0$. Let $\{\sigma_i^2\}_{i=0}^{M-1}$ defined by (5.54) for σ_0^2 given in (5.31) and M in (5.56).*

1. *$K \leq \left\lceil (1/\log(2))\log\left(\sigma_0^{-2}\rho^{-2}d^2(\tau+1)^2\right)\right\rceil$ where K is defined in (5.39).*
2. *For $k \in \{0,\dots,K-1\}$ and $i \in \mathcal{I}_k$, $2^{k+1}\sigma_0^2 a_i\,|\mathcal{I}_k| \leq 1$ where a_i is defined in (5.37) (with $m = 0$) and $\mathcal{I}_k$ in (5.38). As a consequence, $|\mathcal{I}_k| \leq 4(d+4)$.*
3. *For $i \in \{0,\dots,M-1\}$, $\kappa_i\sigma_i^2 \in [1,2]$.*
4. *$\sigma_{M-1}^2 \in [D^2, (10/9)D^2]$.*
5. *For all $i \in \{0,\dots,M-1\}$ and $\gamma_i \leq 1/(m_i + L_i)$, there exist $\alpha_i \in [4,14]$ and $\beta_i \in [1,10]$ such that $C_{i,2} + C_{i,0}C_{i,1} = \alpha_i d m_i^{-1}$ and $C_{i,0}C_{i,1} = \beta_i d\kappa_i^{-1}$ where $C_{i,0}, C_{i,1}, C_{i,2}$ and κ_i are given in (5.23) and (5.14) respectively.*
6. *For all $i \in \{0,\dots,M-1\}$, $0 < A_{i,1} \leq 4dL_i^4\kappa_i^{-1}m_i^{-1}$, where L_i, m_i and κ_i are given in (5.13) and (5.14) respectively.*

Proof. The proofs of 1,2,5,6 are identical to the ones of Lemma 5.23.

3. $\kappa_i\sigma_i^2 = (2L_i)/(m_i + L_i)$.
4. By definition of M, $\sigma_{M-2}^2 \leq D^2$ and $a_{M-2} \leq \sigma_{M-2}^{-2}/\{4(d+4)\}$. By (5.6), we get:

$$\sigma_{M-1}^{-2} = \sigma_{M-2}^{-2} - 2a_{M-2} \geq \sigma_{M-2}^{-2}\left(1 - \frac{1}{2(d+4)}\right)$$

that is $\sigma_{M-1}^2 \leq (10/9)\sigma_{M-2}^2 \leq (10/9)D^2$.

□

5.B.1 Proof of Lemma 5.15

Let $k \in \{0, \dots, K-1\}$ and $i \in \mathcal{I}_k$. Assume that $\gamma_i \leq (m_i + L_i)^{-1}$. The proof follows the same lines as the one in Section 5.5.3. By Lemma 5.24-5 and Lemma 5.24-6, to check condition-i) of Lemma 5.9, it suffices that $\gamma_i \leq \bar{\gamma}_i$, $n_i \geq \bar{n}_i$ and N_i satisfies (5.111), where $\bar{\gamma}_i$ is defined in (5.112) and $\bar{n}_i$ in (5.113).

Using that $\sqrt{1+t} \geq 1+2^{-1}t(1+t)^{-1/2}$ for $t = (\eta^2\kappa_i^2\sigma_i^4)/(4\alpha_i K^2 d^2)$ and by Lemma 5.24-3, concludes that if (5.57) holds then $\gamma_i \leq \bar{\gamma}_i$. $\gamma_i \leq (m_i + L_i)^{-1}$ can be checked by simple algebra.

By (5.37) (with $m = 0$) and the definition of $\mathcal{I}_k$, $a_i \leq \sigma_i^{-2}/\{4(d+4)\}$,

$$\bar{n}_i^{-1} \geq 4^{-1}\kappa_i\gamma_i \left(-1 + \sqrt{1 + \frac{\eta^2\kappa_i^2\sigma_i^4(d+4)}{K\beta_i d}} \right) .$$

Using that $\sqrt{1+t} \geq 1+2^{-1}t(1+t)^{-1/2}$ for $t = (\eta^2\kappa_i^2\sigma_i^4)/(4\alpha_i K^2 d^2)$ and by Lemma 5.24-3, concludes that if (5.58) holds then $n_i \geq \bar{n}_i$. Finally, by (5.58), if (5.59) holds, (5.111) is satisfied.

The case $i = M-1$ is different because $\bar{g}_{M-1}$ is Lipschitz. Assume $\gamma_{M-1} \leq (m_{M-1} + L_{M-1})^{-1}$. [DM16, section 2.1] entails that condition-ii) of Lemma 5.9 is satisfied if

$$\begin{aligned} \operatorname{Lip} \bar{g}_{M-1}{}^2 \Big\{ \frac{4d}{n_{M-1}m_{M-1}\kappa_{M-1}\gamma_{M-1}} \exp\left(-N_{M-1}\frac{\kappa_{M-1}\gamma_{M-1}}{2}\right) \\ + 2\kappa_{M-1}^{-1}\left(A_{M-1,0}\gamma_{M-1} + A_{M-1,1}\gamma_{M-1}^2\right) \Big\} \leq \eta^2 , \qquad (5.129) \\ \frac{8 \operatorname{Lip} \bar{g}_{M-1}{}^2}{\kappa_{M-1}^2 n_{M-1}\gamma_{M-1}} \left\{ 1 + \frac{2}{n_{M-1}\kappa_{M-1}\gamma_{M-1}} \right\} \leq \eta^2 . \end{aligned}$$

Using $\operatorname{Lip} \bar{g}_{M-1}{}^2 \leq (\sigma_{M-1}^{-2}\mathrm{e})$ and (5.24), Lemma 5.24-6 for $i = M-1$, it is sufficient for $\gamma_{M-1}, n_{M-1}, n_{M-1}$ to satisfy

$$2L_{M-1}^2\kappa_{M-1}^{-1}d\gamma_{M-1} + 4dL_{M-1}^4\kappa_{M-1}^{-1}m_{M-1}^{-1}\gamma_{M-1}^2 \leq (4\mathrm{e})^{-1}\kappa_{M-1}\eta^2\sigma_{M-1}^2 , \qquad (5.130)$$

$$n_{M-1}^{-1}\left(1 + 2(\kappa_{M-1}\gamma_{M-1}n_{M-1})^{-1}\right) \leq \frac{\eta^2\kappa_{M-1}^2\sigma_{Mc-1}^2\gamma_{M-1}}{8\mathrm{e}} , \qquad (5.131)$$

$$-2\frac{\log\left((8\mathrm{e}d)^{-1}n_{M-1}\kappa_{M-1}\gamma_{M-1}\eta^2\right)}{\kappa_{M-1}\gamma_{M-1}} = \bar{N}_{M-1} \leq N_{M-1} . \qquad (5.132)$$

Denote by $\bar{\gamma}_{M-1}, \bar{n}_{M-1}^{-1}$ the roots of (5.130), (5.131) seen as equalities. We have

$$\bar{\gamma}_{M-1} = 4^{-1}L_{M-1}^{-2}\sigma_{Mc-1}^{-2}\left\{ -1 + \sqrt{1 + \frac{\eta^2\kappa_{M-1}^2\sigma_{M-1}^4}{\mathrm{e}d}} \right\} , \qquad (5.133)$$

$$\bar{n}_{M-1}^{-1} = 4^{-1}\kappa_{M-1}\gamma_{M-1}\left\{ -1 + \sqrt{1 + \mathrm{e}^{-1}\eta^2\kappa_{M-1}\sigma_{M-1}^2} \right\} . \qquad (5.134)$$

Using that $\sqrt{1+t} \geq 1+2^{-1}t(1+t)^{-1/2}$ for $t = (\eta^2\kappa_{M-1}^2\sigma_{M-1}^4)/(\mathrm{e}d)$ and by Lemma 5.24-3, concludes that if (5.60) holds then $\gamma_{M-1} \leq \bar{\gamma}_{M-1}$. $\gamma_{M-1} \leq (m_{M-1}+L_{M-1})^{-1}$ can be checked by simple algebra.

Using that $\sqrt{1+t} \geq 1+2^{-1}t(1+t)^{-1/2}$ for $t = \mathrm{e}^{-1}\eta^2\kappa_{M-1}\sigma_{M-1}^2$ and by Lemma 5.24-3, concludes that if (5.61) holds then $n_{M-1} \geq \bar{n}_{M-1}$.

Finally by (5.61), if (5.62) holds, (5.132) is satisfied.

5.B.2 Proof of Lemma 5.16

The proof is identical to the one of Lemma 5.11. For $k \in \{0,\dots,K-1\}$ and $i \in \mathcal{I}_k$, it is sufficient for γ_i to satisfy (5.63) by (5.46) and Lemma 5.24-3.

Regarding the case $i = M-1$, assuming that $\gamma_{M-1} \leq (m_{M-1}+L_{M-1})^{-1}$, it is sufficient for $\gamma_{M-1}, n_{M-1}, N_{M-1}$ to satisfy (5.129) with $A_{M-1,0}\gamma_{M-1} + A_{M-1,1}\gamma_{M-1}^2$ replaced by $B_{M-1,0}\gamma_{M-1}^2 + B_{M-1,1}\gamma_{M-1}^3$. The counterpart of (5.130) is then,

$$\frac{1}{3\kappa_{M-1}}\left(d\tilde{L}^2 + 10L_{M-1}^4 m_{M-1}^{-1}\right)\gamma_{M-1}^2 + \frac{25}{12}\frac{L_{M-1}^4}{m_{M-1}}\gamma_{M-1}^3 \leq \frac{\eta^2\kappa_{M-1}\sigma_{M-1}^2}{4\mathrm{e}d} \; .$$

This concludes the proof with the same argument as in Section 5.A.2.

5.B.3 Proof of Theorems 5.12 and 5.13 and corollary 5.14

For $i \in \{0,\dots,M-1\}$, set γ_i, n_i, N_i such that (5.57), (5.58), (5.59), (5.60), (5.61) and (5.62) are equalities. By (5.18), we have

$$\mathrm{cost} = \left(\frac{453K}{\eta^2} + 2\log(Kd^2)\right)\frac{462K^2d^2}{\eta^2}\sum_{i=0}^{M-2}\kappa_i^{-1}L_i^2\sigma_i^2 + n_{M-1} + N_{M-1} \; .$$

Note that for $i \in \{0,\dots,M-2\}$,

$$\kappa_i^{-1}L_i^2\sigma_i^2 = 1 + (3/2)L\sigma_i^2 + (L^2/2)\sigma_i^4 \; .$$

By Lemma 5.24-2, for $k \in \{0,\dots,K-1\}$, $|\mathcal{I}_k| \leq 4(d+4)$ and for $i \in \mathcal{I}_k$, $\sigma_i^2 \leq 2^{k+1}\sigma_0^2$. We then have

$$\begin{aligned}\sum_{i=0}^{M-2}\frac{L_i^2\sigma_i^2}{\kappa_i} &\leq 4(d+4)\sum_{k=0}^{K-1}\left\{1 + \frac{3L}{2}2^{k+1}\sigma_0^2 + \frac{L^2}{2}(2^{k+1}\sigma_0^2)^2\right\} \\ &\leq 4(d+4)\left\{K + 3L(2^K\sigma_0^2) + \frac{2L^2}{3}(2^K\sigma_0^2)^2\right\} \; .\end{aligned}$$

By (5.56) and the definition of K, (5.39), $2^K\sigma_0^2 \leq 2D^2$. The expressions of $\gamma_{M-1}, n_{M-1}, N_{M-1}$ give

$$n_{M-1} + N_{M-1} = \left(\frac{29}{\eta^2} + 2\log(d)\right)\frac{26d}{\eta^2}\kappa_{M-1}^{-2}L_{M-1}^2 \; ,$$

with $\kappa_{M-1}^{-2}L_{M-1}^2 = (1+2^{-1}L\sigma_{M-1}^2)^2$. By Lemma 5.24-4, we then have

$$n_{M-1}+N_{M-1} \leq \left(\frac{29}{\eta^2}+2\log(d)\right)\frac{26d}{\eta^2}\left(1+\frac{5L}{9}D^2\right)^2 , \tag{5.135}$$

and (5.51) is established.

Assume **H**12. We now prove Theorem 5.13 and use Lemma 5.11 instead of Lemma 5.10. For $i \in \{0,\dots,M-1\}$, set γ_i, n_i, N_i such that (5.63), (5.58), (5.59), (5.64), (5.61) and (5.62) are equalities. By (5.18) and using that for $a,b \geq 0$, $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$, we have

$$\text{cost} \leq \left(\frac{453K}{\eta^2}+2\log(Kd^2)\right)\sqrt{\frac{7}{3}}\frac{8Kd}{\eta}\sum_{i=0}^{M-2}\frac{\sigma_i}{\kappa_i}\left(d^{1/2}\tilde{L}+\sqrt{10}L_i^2\sigma_i\right)+n_{M-1}+N_{M-1} .$$

For $k \in \{0,\dots,K-1\}$ and $i \in \mathcal{I}_k$, note that

$$\kappa_i^{-1}\sigma_i \leq \sigma_i^3 , \quad \kappa_i^{-1}\sigma_i^2L_i^2 = 1+\frac{3L}{2}\sigma_i^2+\frac{L^2}{2}\sigma_i^4 .$$

Using for $k \in \{0,\dots,K-1\}$, $|\mathcal{I}_k| \leq 4(d+4)$ by Lemma 5.24-2 and for $i \in \mathcal{I}_k$, $\sigma_i^2 \leq 2^{k+1}\sigma_0^2$, we get

$$\begin{aligned}\sum_{i=0}^{M-2}\frac{\sigma_i}{\kappa_i}\left(d^{1/2}\tilde{L}+\sqrt{10}L_i^2\sigma_i\right) &\leq 4(d+4)\sum_{k=0}^{K-1}\left\{d^{1/2}\tilde{L}(2^{k+1}\sigma_0^2)^{3/2}\right.\\ &\qquad \left.+\sqrt{10}\left(1+\frac{3L}{2}(2^{k+1}\sigma_0^2)+\frac{L^2}{2}(2^{k+1}\sigma_0^2)^2\right)\right\}\\ &\leq 4(d+4)\left\{5d^{1/2}\tilde{L}D^3+\sqrt{10}\left(K+6LD^2+\frac{8L^2}{3}D^4\right)\right\} ,\end{aligned}$$

with $2^K\sigma_0^2 \leq 2D^2$. The expressions of $\gamma_{M-1}, n_{M-1}, N_{M-1}$ give

$$n_{M-1}+N_{M-1} \leq \left(2\log(d)+\frac{29}{\eta^2}\right)\sqrt{\frac{8\mathrm{e}}{3}}\frac{\sqrt{d}}{\eta}\frac{d^{1/2}\tilde{L}+\sqrt{10}L_{M-1}^2\sigma_{M-1}}{\kappa_{M-1}^2\sigma_{M-1}} .$$

By Lemma 5.24-4, $\sigma_{M-1}^2 \in [D^2,(10/9)D^2]$. We get then

$$\kappa_{M-1}^{-2}\sigma_{M-1}^{-1} = \frac{\left(1+(L/2)\sigma_{M-1}^2\right)^2}{L^2\sigma_{M-1}} \leq \frac{1}{DL^2}\left(1+\frac{5L}{9}D^2\right)^2 ,$$

and,

$$\kappa_{M-1}^{-2}L_{M-1}^2 = \left(1+\frac{L}{2}\sigma_{M-1}^2\right)^2 \leq \left(1+\frac{5L}{9}D^2\right)^2 ,$$

which gives,

$$n_{M-1}+N_{M-1} \leq \left(2\log(d)+\frac{29}{\eta^2}\right)\sqrt{\frac{8\mathrm{e}}{3}}\frac{\sqrt{d}}{\eta}\left(1+\frac{5L}{9}D^2\right)^2\left(\frac{d^{1/2}\tilde{L}}{DL^2}+\sqrt{10}\right) . \tag{5.136}$$

(5.53) is established. The proof of Corollary 5.14 is the same as the one of Corollary 5.7.

Chapter 6

Diffusion approximations and control variates for MCMC

Nicolas Brosse [1], Alain Durmus [2], Sean Meyn [3] and Éric Moulines [1]

Abstract

A new methodology is presented for the construction of control variates to reduce the variance of additive functionals of Markov Chain Monte Carlo (MCMC) samplers. Our control variates are defined as linear combinations of functions whose coefficients are obtained by minimizing a proxy for the asymptotic variance. The construction is theoretically justified by two new results. We first show that the asymptotic variances of some well-known MCMC algorithms, including the Random Walk Metropolis and the (Metropolis) Unadjusted/Adjusted Langevin Algorithm, are close to the asymptotic variance of the Langevin diffusion. Second, we provide an explicit representation of the optimal coefficients minimizing the asymptotic variance of the Langevin diffusion. Several examples of Bayesian inference problems demonstrate that the corresponding reduction in the variance is significant, and that in some cases it can be dramatic.

6.1 Introduction

Let $U : \mathbb{R}^d \to \mathbb{R}$ be a measurable function on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $\int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \mathrm{d}x < \infty$. This function is associated to a probability measure π on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ defined for all $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$ by $\pi(\mathrm{A}) \stackrel{\text{def}}{=} \int_{\mathrm{A}} \mathrm{e}^{-U(x)} \mathrm{d}x / \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \mathrm{d}x$. We are interested in approximating $\pi(f) \stackrel{\text{def}}{=} \int f(x) \pi(\mathrm{d}x)$, where f is a π-integrable function. The classical Monte Carlo solution to

[1]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France.
Emails: nicolas.brosse@polytechnique.edu, eric.moulines@polytechnique.edu

[2]Ecole Normale Supérieure CMLA 61, Av. du Président Wilson 94235 Cachan Cedex, France
Email: alain.durmus@cmla.ens-cachan.fr

[3]University of Florida, Department of Electrical and Computer Engineering, Gainesville, Florida.
Email: meyn@ece.ufl.edu

this problem is to simulate i.i.d. random variables $(X_k)_{k\in\mathbb{N}}$ with distribution π, and then to estimate $\pi(f)$ by the sample mean

$$\hat{\pi}_n(f) = n^{-1}\sum_{i=0}^{n-1} f(X_i)\ . \tag{6.1}$$

In most applications, sampling from π is not an option. Markov Chain Monte Carlo (MCMC) methods amount to sample a Markov chain $(X_k)_{k\in\mathbb{N}}$ from a Markov kernel R with (unique) invariant distribution π. Under weak additional conditions [MT09, Chapter 17], the estimator $\hat{\pi}_n(f)$ defined by (6.1) satisfies for any initial distribution a Central Limit Theorem (CLT)

$$n^{-1/2}\sum_{k=0}^{n-1}(f(X_k)-\pi(f)) \underset{n\to+\infty}{\overset{\text{weakly}}{\Longrightarrow}} \mathcal{N}(0,\sigma^2_{\infty,\mathrm{d}}(f))\ , \quad \sigma^2_{\infty,\mathrm{d}}(f) = \pi\left((\hat{f}_{\mathrm{d}})^2 - (R\hat{f}_{\mathrm{d}})^2\right)\ , \tag{6.2}$$

where $\mathcal{N}(m,\sigma^2)$ denotes a Gaussian distribution with mean m and variance σ^2, and $\hat{f}_{\mathrm{d}}$ is a solution of the Poisson equation

$$(R-\mathrm{Id})\hat{f}_{\mathrm{d}} = -\{f-\pi(f)\}\ . \tag{6.3}$$

Reducing the variance of Monte Carlo estimators is a very active research domain: see e.g. [RC04, Chapter 4], [Liu08, Section 2.3], and [RK17, Chapter 5] for an overview of the main methods. In this paper, we use control variates, i.e. π-integrable functions $h = (h_1,\dots,h_p) : \mathbb{R}^d \to \mathbb{R}^p$ satisfying $\pi(h_i) = 0$ for $i \in \{1,\dots,p\}$ and then choose $\theta\in\mathbb{R}^p$ such that $\sigma^2_{\infty,\mathrm{d}}(f+\theta^{\mathrm{T}}h) \le \sigma^2_{\infty,\mathrm{d}}(f)$. [Hen97] and [Mey08, Section 11.5] proposed control variates of the form $(R-\mathrm{Id})\theta^{\mathrm{T}}\psi$ where $\psi = (\psi_1,\dots,\psi_p)$ are known π-integrable functions. The parameter $\theta\in\mathbb{R}^p$ is obtained by minimizing the asymptotic variance

$$\min_{\theta\in\mathbb{R}^p}\sigma^2_{\infty,\mathrm{d}}(f+(R-\mathrm{Id})\theta^{\mathrm{T}}\psi) = \min_{\theta\in\mathbb{R}^p}\pi\left(\left\{\hat{f}_{\mathrm{d}}-\theta^{\mathrm{T}}\psi\right\}^2 - \left\{R(\hat{f}_{\mathrm{d}}-\theta^{\mathrm{T}}\psi)\right\}^2\right)\ , \tag{6.4}$$

noting that $(-\theta^{\mathrm{T}}\psi)$ is a solution of the Poisson equation associated to $(R-\mathrm{Id})\theta^{\mathrm{T}}\psi$ and $\hat{f}_{\mathrm{d}}$ is defined in (6.3). The method suggested in [Mey08, Section 11.5] to minimize (6.4) requires estimates of the solution $\hat{f}_{\mathrm{d}}$ of the Poisson equation. Temporal Difference learning is a possible candidate, but this method is complex and suffers from high variance.

[DK12] noticed that if R is reversible w.r.t. π, it is possible to optimize the limiting variance (6.4) without computing explicitly the Poisson solution $\hat{f}_{\mathrm{d}}$. Reversibility will play an important role in this paper as well.

Each of the algorithms in the aforementioned literature requires computation of $R\psi_i$ for each $i\in\{1,\dots,p\}$, which is in general a computational challenge. In [Hen97; Mey08] this is addressed by restricting to kernels for which $R(x,\cdot)$ has finite support for each x, and in [DK12] the authors restrict mainly to Gibbs samplers in their numerical examples.

In this paper an alternative class of control variates is used to avoid this computational barrier. This approach follows [AC99] (applications to quantum Monte Carlo

calculations) and [MSI13; PMG14] (Bayesian statistics): assume that U is continuously differentiable, and for any twice continuously differentiable function φ, define $\mathscr{L}\varphi$ by

$$\mathscr{L}\varphi = -\langle \nabla U, \nabla \varphi \rangle + \Delta \varphi \; . \tag{6.5}$$

Under mild conditions on φ, it may be shown that $\pi(\mathscr{L}\varphi) = 0$. [MSI13] suggested to use $\mathscr{L}(\theta^{\mathrm{T}}\psi)$ with $\psi = (\psi_1, \ldots, \psi_p)$ as control variates and choose θ by minimizing $\theta \mapsto \pi(\{f - \pi(f) + \mathscr{L}\theta^{\mathrm{T}}\psi\}^2)$. This approach has triggered numerous work, among others [OGC16], [OG16] and [Oat+18] which introduce control functionals; a nonparametric extension of control variates. A drawback of this method stems from the fact that the optimization criterion $\pi(\{f - \pi(f) + \mathscr{L}\theta^{\mathrm{T}}\psi\}^2)$ is only theoretically justified if $(X_k)_{k\in\mathbb{N}}$ is i.i.d. and might significantly differ from the asymptotic variance $\sigma^2_{\infty,\mathrm{d}}(f + \mathscr{L}(\theta^{\mathrm{T}}\psi))$ defined in (6.1).

In this paper, we propose a new method to construct control variates. Analysis and motivation are based on the overdamped Langevin diffusion defined for $t \geq 0$ by

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \; , \tag{6.6}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. If ∇U is Lipschitz, the Stochastic Differential Equation (SDE) (6.6) has a unique strong solution $(Y_t)_{t\geq 0}$ for every initial condition $x \in \mathbb{R}^d$. Denote by $(P_t)_{t\geq 0}$ the semigroup associated to the SDE (6.6) defined by $P_t f(x) = \mathbb{E}\left[f(Y_t)\right]$ where f is bounded measurable and $(Y_t)_{t\geq 0}$ is a solution of (6.6) started at x. Under mild additional conditions (see e.g. [RT96]), π is invariant for the semigroup $(P_t)_{t\geq 0}$, i.e. $\pi P_t = \pi$ for all $t \geq 0$. In addition, under smoothness and 'tail' conditions on f and ∇U, the following CLT holds for any initial condition (see [Bha82; CCG12])

$$t^{-1/2}\int_0^t \{f(Y_s) - \pi(f)\}\mathrm{d}s \overset{\text{weakly}}{\underset{t\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_\infty(f)) \; , \quad \sigma^2_\infty(f) = 2\pi\left(\hat{f}\{f - \pi(f)\}\right) \; , \tag{6.7}$$

where $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ is a solution of the (continuous-time) Poisson equation

$$\mathscr{L}\hat{f} = -\{f - \pi(f)\} \; . \tag{6.8}$$

The main contribution of this paper is the introduction of a new class of control variates based on the expression of the asymptotic variance $\sigma^2_\infty(f)$ given in (6.7). Since $\pi(\mathscr{L}(\theta^{\mathrm{T}}\psi)) = 0$ for any $\theta \in \mathbb{R}^d$, we consider the control variate $\mathscr{L}(\theta^*(f)^{\mathrm{T}}\psi)$ where $\theta^*(f)$ is chosen by minimizing

$$\theta \mapsto \sigma^2_\infty(f + \mathscr{L}(\theta^{\mathrm{T}}\psi)) \; . \tag{6.9}$$

Although $\mathscr{L}(\theta^*(f)^{\mathrm{T}}\psi)$ is a control variate for the Langevin diffusion associated with f, the choice of this optimization criterion is motivated by the fact that for some MCMC algorithms, the asymptotic variance $\sigma^2_{\infty,\mathrm{d}}(f)$ defined in (6.2) is (up to a scaling factor) a good approximation of the asymptotic variance of the Langevin diffusion $\sigma^2_\infty(f)$ defined in (6.7). Moreover, the minimization of (6.9) admits a unique solution $\theta^*(f)$, which is

in general easy to estimate. It is worthwhile to note that it is not required to know the Poisson solution $\hat{f}$ to minimize (6.9).

The construction of control variates for MCMC and the related problem of approximating solutions of Poisson equations are very active fields of research. It is impossible to give credit for all the contributions undertaken in this area; see [DK12], [PMG14] and references therein for further background.

Amongst recent studies on this subject, [MV15] approximate directly the solution $\hat{f}_{\mathrm{d}}$ of the Poisson equation by subdividing the state space. Close to the methodology presented in the present paper, [MV17] uses the scaling limit of the RWM algorithm when the dimension d of the state space $\mathbb{R}^d$ goes to infinity to implement a control variates based on a solution of the Poisson equation for the Langevin diffusion. This approach uses a strong assumption on the stationary distribution which is assumed to be in product form. It is difficult to predict the performance of this methodology when this assumption is not met. Based on [LS16, Section 3.4.2], [RS17] addresses the control variates design for diffusions with possibly unknown invariant probability measures. Concerning the link between $\sigma^2_{\infty,\mathrm{d}}(f)$ and $\sigma^2_{\infty}(f)$, an analogous result associated to the error estimates for the Green-Kubo formula can be found in [LS16, Theorem 5.6]. It was initially derived in [LMS16]. An analogous study [DPZ17] uses the comparison between the discrete Markov chain and the Langevin diffusion limit. Note that in this paper, we explicit check the assumptions introduced to prove the geometric ergodicity of the Markov chain on $\mathbb{R}^d$.

The remainder of the paper is organized as follows. In Section 6.2, we present our methodology to compute the minimizer $\theta^*(f)$ of (6.9) and the construction of control variates for some MCMC algorithms. In Section 6.3, we state our main result which guarantees that the asymptotic variance $\sigma^2_{\infty,\mathrm{d}}(f)$ defined in (6.2) and associated with a given MCMC method is close (up to a scaling factor) to the asymptotic variance of the Langevin diffusion $\sigma^2_{\infty}(f)$ defined in (6.7). We provide a CLT and we show that under appropriate conditions on U, the Unadjusted Langevin Algorithm (ULA) fits the framework of our methodology. In Section 6.4, a Monte Carlo experiment illustrating the performance of our method is presented. In Section 6.5, we establish conditions under which the results of Sections 6.2 and 6.3 can be applied to the Random Walk Metropolis (RWM) and the Metropolis Adjusted Langevin Algorithm (MALA). The proofs are postponed to Section 6.6 and to the Appendix.

Notation

Let $\mathcal{B}(\mathbb{R}^d)$ denote the Borel σ-field of $\mathbb{R}^d$. Moreover, let $\mathrm{L}^1(\mu)$ be the set of μ-integrable functions for μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Further, $\mu(f) = \int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x)$ for an $f \in \mathrm{L}^1(\mu)$. Given a Markov kernel R on $\mathbb{R}^d$, for all $x \in \mathbb{R}^d$ and f integrable under $R(x,\cdot)$, denote by $Rf(x) = \int_{\mathbb{R}^d} f(y)R(x,\mathrm{d}y)$. Let $V : \mathbb{R}^d \to [1,\infty)$ be a measurable function. The V-total variation distance between two probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is defined as $\|\mu - \nu\|_V = \sup_{|f|\leq V} |\mu(f) - \nu(f)|$. If $V = 1$, then $\|\cdot\|_V$ is the total variation denoted by $\|\cdot\|_{\mathrm{TV}}$. For a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, define $\|f\|_V = \sup_{x\in\mathbb{R}^d} |f(x)|/V(x)$.

For $u, v \in \mathbb{R}^d$, define the scalar product $\langle u, v\rangle = \sum_{i=1}^d u_i v_i$ and the Euclidian norm $\|u\| = \langle u, u\rangle^{1/2}$. Denote by $\mathbb{S}(\mathbb{R}^d) = \left\{u \in \mathbb{R}^d : \|u\| = 1\right\}$. For $a, b \in \mathbb{R}$, denote by $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$ and $a_+ = a \vee 0$. For $a \in \mathbb{R}_+$, $\lfloor a \rfloor$ and $\lceil a \rceil$ denote respectively the floor and ceil functions evaluated in a. We take the convention that for $n, p \in \mathbb{N}$, $n < p$ then $\sum_p^n = 0$, $\prod_p^n = 1$ and $\{p, \ldots, n\} = \emptyset$. Define for $t \in \mathbb{R}$, $\Phi(t) = (2\pi)^{-1/2} \int_{-\infty}^t \mathrm{e}^{-r^2/2} \mathrm{d}r$ and $\bar{\Phi}(t) = 1 - \Phi(t)$.

For $k \in \mathbb{N}$, $m, m' \in \mathbb{N}^*$ and Ω, Ω' two open sets of $\mathbb{R}^m, \mathbb{R}^{m'}$ respectively, denote by $\mathrm{C}^k(\Omega, \Omega')$, the set of k-times continuously differentiable functions. For $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$, denote by ∇f the gradient of f and by Δf the Laplacian of f. For $k \in \mathbb{N}$ and $f \in \mathrm{C}^k(\mathbb{R}^d, \mathbb{R})$, denote by $\mathrm{D}^i f$ the i-th order differential of f for $i \in \{0, \ldots, k\}$. For $x \in \mathbb{R}^d$ and $i \in \{1, \ldots, k\}$, define $\left\|\mathrm{D}^0 f(x)\right\| = |f(x)|$, $\left\|\mathrm{D}^i f(x)\right\| = \sup_{u_1, \ldots, u_i \in \mathbb{S}(\mathbb{R}^d)} \mathrm{D}^i f(x)[u_1, \ldots, u_i]$. For $k, p \in \mathbb{N}$ and $f \in \mathrm{C}^k(\mathbb{R}^d, \mathbb{R})$, define the norm

$$\|f\|_{k,p} = \sup_{x \in \mathbb{R}^d,\, i \in \{0,\ldots,k\}} \left\|\mathrm{D}^i f(x)\right\| / (1 + \|x\|^p) \,,$$

and $\mathrm{C}^k_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}) = \left\{f \in \mathrm{C}^k(\mathbb{R}^d, \mathbb{R}) : \inf_{p \in \mathbb{N}} \|f\|_{k,p} < +\infty\right\}$.

6.2 Langevin-based control variates for MCMC methods

Before introducing our new methodology based on the Langevin diffusion (6.6), we need to briefly recall some of its properties; this requires 'tail' and regularity assumptions on U. Let $k \geq 2$.

H13 (k). *$U \in \mathrm{C}^k_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and there exist $\upsilon > 0$ and $M_\upsilon \geq 0$ such that for all $x \in \mathbb{R}^d$, $\|x\| \geq M_\upsilon$, $\langle \nabla U(x), x\rangle \geq \upsilon \|x\|$.*

Proposition 6.1. *Assume **H**13(k) for $k \geq 2$.*

(i) The semigroup $(P_t)_{t \geq 0}$ associated to (6.6) admits π as its unique invariant probability measure and for all $p \in \mathbb{N}$, $\int_{\mathbb{R}^d} \|x\|^p \pi(\mathrm{d}x) < +\infty$.

(ii) For any initial condition Y_0 and $f \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, the solution $(Y_t)_{t \geq 0}$ of the Langevin diffusion (6.6) satisfies the CLT (6.7).

(iii) For all $f \in \mathrm{C}^{k-1}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, there exists $\hat{f} \in \mathrm{C}^k_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ such that $\mathscr{L}\hat{f} = \pi(f) - f$, where $\mathscr{L}$ is the generator of the Langevin diffusion defined in (6.5). For all $p \in \mathbb{N}$, there exist $C \geq 0$, $q \in \mathbb{N}$ such that for all $f \in \mathrm{C}^{k-1}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\|\hat{f}\|_{k,q} \leq C \|f\|_{k-1,p}$.

(iv) For all $f, g \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$,

$$\pi\left(f(-\mathscr{L})g\right) = \pi\left(g(-\mathscr{L})f\right) = \pi\left(\langle \nabla f, \nabla g\rangle\right). \tag{6.10}$$

Proof. All these results are classical. A sketch of proof together with relevant references is postponed to Section 6.C.1. □

Proposition 6.1-(iii) ensures the existence and regularity of a solution of the Poisson equation (6.8) for any $f \in \mathrm{C}^{k-1}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $k \in \mathbb{N}$, $k \geq 2$. Proposition 6.1-(iv) is a classical "carré du champ" identity, see for example [BGL14, Section 1.6.2, formula 1.6.3]. It means in particular that the generator $\mathscr{L}$ is (formally) self-adjoint in $\mathrm{L}^2(\pi)$ which plays a key role in the construction of our control variates.

A straightforward consequence of (6.10) (setting $f = 1$) is that for any function $g \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\pi(\mathscr{L}g) = 0$. This suggests taking as a class of control variates for π the family of functions $\{\mathscr{L}(\theta^{\mathrm{T}}\psi) : \theta \in \mathbb{R}^p\}$, where $\psi = (\psi_1, \ldots, \psi_p) : \mathbb{R}^d \to \mathbb{R}^p$, $p \in \mathbb{N}^*$, is a fixed sieve of functions such that for all $i \in \{1, \ldots, p\}$, $\psi_i \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. Let $f \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$; by Proposition 6.1-(ii), for all $\theta \in \mathbb{R}^p$,

$$t^{-1/2} \int_0^t \{(f + \mathscr{L}(\theta^{\mathrm{T}}\psi))(Y_s) - \pi(f)\} \mathrm{d}s \overset{\text{weakly}}{\underset{t\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_\infty(f + \mathscr{L}(\theta^{\mathrm{T}}\psi))) \, ,$$

and an appropriate choice for the parameter $\theta \in \mathbb{R}^p$ is given by a minimizer of $\theta \mapsto \sigma^2_\infty(f + \mathscr{L}(\theta^{\mathrm{T}}\psi))$ defined in (6.7). We now show that this minimization problem has a unique solution which can be computed explicitly.

By Proposition 6.1-(iii), for any $f \in \mathrm{C}^1_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, the Poisson equation $\mathscr{L}\hat{f} = -\{f - \pi(f)\}$ has a solution $\hat{f} \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. Then, for all $\theta \in \mathbb{R}^p$, $\hat{f}_\theta = \hat{f} - \theta^{\mathrm{T}}\psi \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ is a solution of the Poisson equation $\mathscr{L}\hat{f}_\theta = -\{f_\theta - \pi(f_\theta)\}$, where $f_\theta = f + \mathscr{L}(\theta^{\mathrm{T}}\psi)$. Using the expression (6.7) of the asymptotic variance and $\pi(\mathscr{L}(\theta^{\mathrm{T}}\psi)) = 0$, we get for all $\theta \in \mathbb{R}^p$

$$\sigma^2_\infty(f_\theta) = 2\pi \left((\hat{f} - \theta^{\mathrm{T}}\psi) \left\{ f - \pi(f) + \mathscr{L}(\theta^{\mathrm{T}}\psi) \right\} \right) \, . \tag{6.11}$$

Now by Proposition 6.1-(iv) and since $\mathscr{L}\hat{f} = \pi(f) - f$, we obtain

$$\pi(\hat{f}\mathscr{L}\psi) = \pi(\{\pi(f) - f\}\psi) \, .$$

Plugging this identity in (6.11) and using Proposition 6.1-(iv) imply for all $\theta \in \mathbb{R}^p$,

$$\sigma^2_\infty(f + \mathscr{L}(\theta^{\mathrm{T}}\psi)) = 2\theta^{\mathrm{T}} H \theta - 4\theta^{\mathrm{T}} \pi\left(\psi \left\{ f - \pi(f) \right\}\right) + 2\pi \left(\hat{f} \left\{ f - \pi(f) \right\} \right) \, ,$$

where $H \in \mathbb{R}^{p \times p}$ is given for any $i, j \in \{1, \ldots, p\}$ by

$$H_{ij} = \pi(\langle \nabla\psi_i, \nabla\psi_j \rangle) \, . \tag{6.12}$$

Therefore, $\theta \mapsto \sigma^2_\infty(f + \mathscr{L}(\theta^{\mathrm{T}}\psi))$ is a quadratic function and has a unique minimizer if and only if H is symmetric positive definite and this minimizer is given by

$$\theta^*(f) = H^{-1} \pi\left(\psi \{ f - \pi(f) \}\right) \, . \tag{6.13}$$

Note that H is by definition a symmetric semi-positive definite matrix. It is easily seen that if $(1, \psi_1, \ldots, \psi_p)$ is linearly independent in $\mathrm{C}(\mathbb{R}^d, \mathbb{R})$, then H is full rank and the minimizer of $\sigma^2_\infty(f + \mathscr{L}(\theta^{\mathrm{T}}\psi))$ is given by (6.13).

To sum up, constructing control variates of the form $\mathscr{L}(\theta^{\mathrm{T}}\psi)$ for the Langevin diffusion is straightforward and the optimal parameter $\theta^*(f)$ minimizing the asymptotic variance has an explicit expression (6.13) that does not involve the (usually unknown) solution $\hat{f}$ of the Poisson equation $\mathscr{L}\hat{f} = \pi(f) - f$.

Implications for MCMC The continuous-time setting has mainly theoretical interest. The main contribution of this paper is to show that the optimal control variate for the diffusion remains nearly optimal for many classes of discrete-time MCMC algorithms.

One example is the Markov kernel associated with the Unadjusted Langevin Algorithm (ULA). A diffusion approximation is to be expected since the ULA algorithm is the Euler discretization scheme associated to the Langevin SDE (6.6): $X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1}$, where $\gamma > 0$ is the step size and $(Z_k)_{k\in\mathbb{N}}$ is an i.i.d. sequence of standard Gaussian d-dimensional random vectors. The idea of using the Markov chain $(X_k)_{k\in\mathbb{N}}$ to sample approximately from π has been first introduced in the physics literature by [Par81] and popularized in the computational statistics community by [Gre83] and [GM94]. Other examples include the Metropolis Adjusted Langevin Algorithm (MALA) algorithm, and the Random Walk Metropolis algorithm (RWM).

Each of these MCMC algorithms define a family of Markov kernels $\{R_\gamma, \gamma \in (0, \bar{\gamma}]\}$, indexed by the step-size parameter $\gamma \in (0, \bar{\gamma}]$, for $\bar{\gamma} > 0$. For any initial distribution ξ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\gamma \in (0, \bar{\gamma}]$, denote by $\mathbb{P}_{\xi,\gamma}$ and $\mathbb{E}_{\xi,\gamma}$ the probability and the expectation respectively on the canonical space of a Markov chain with initial distribution ξ and of transition kernel R_γ. By convention, we set $\mathbb{E}_{x,\gamma} = \mathbb{E}_{\delta_x,\gamma}$ for all $x \in \mathbb{R}^d$. We denote by $(X_k)_{k\geq 0}$ the canonical process. Under $\mathbb{P}_{\xi,\gamma}$, $(X_k)_{k\geq 0}$ is a Markov chain with initial distribution ξ and Markov kernel R_γ. The following assumptions are imposed here. General criteria to justify (I)–(III) are postponed to Section 6.3.

(I) For each $\gamma \in (0, \bar{\gamma}]$, R_γ is a positive Harris Markov kernel with invariant distribution π_γ satisfying $\pi_\gamma(|f|) < \infty$ for any $f \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$.

(II) For any initial condition X_0, each $f \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $\gamma \in (0, \bar{\gamma}]$,

$$\sqrt{n}(\hat{\pi}_n(f) - \pi_\gamma(f)) \underset{n\to+\infty}{\overset{\text{weakly}}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_{\infty,\gamma}(f)) \tag{6.14}$$

where $\hat{\pi}_n(f)$ is defined by (6.1), and $\sigma^2_{\infty,\gamma}(f) \geq 0$ is the asymptotic variance associated with f defined by (6.2) relatively to R_γ.

(III) For any functions f, g sufficiently smooth and satisfying growth conditions,

$$\sigma^2_{\infty,\gamma}(f + \gamma g) = \gamma^{-1}\sigma^2_\infty(f) + o(\gamma^{-1}) \tag{6.15}$$

$$\pi_\gamma(f) = \pi(f) + O(\gamma) \,, \tag{6.16}$$

where $\sigma^2_\infty(f)$ is defined in (6.7) and for $\gamma \downarrow 0^+$.

The standard conditions (I)–(II) are in particular satisfied if R_γ is V-uniformly geometrically ergodic, see e.g. [MT09]. Let $V : \mathbb{R}^d \to [1, +\infty)$ be a measurable function. We say that R_γ, $\gamma \in (0, \bar{\gamma}]$ is V-uniformly geometrically ergodic if it admits an invariant probability measure π_γ such that $\pi_\gamma(V) < +\infty$ and there exist $C \geq 0$ and $\rho \in [0, 1)$ such that for any probability measure ξ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $n \in \mathbb{N}$,

$$\|\xi R^n_\gamma - \pi_\gamma\|_V \leq C\xi(V)\rho^n \,.$$

The approximation result (III) requires more sophisticated arguments given in Section 6.3.

Based on (I)–(III) and (6.13), the estimator of $\pi(f)$ we suggest is given for $n \in \mathbb{N}^*$ by

$$\pi_n^{\mathrm{CV}}(f) = \frac{1}{n}\sum_{k=0}^{n-1}\left\{f(X_k) + \mathscr{L}\left(\hat{\theta}_n^*(f)^{\mathrm{T}}\psi(X_k)\right)\right\} , \tag{6.17}$$

and $\hat{\theta}_n^*(f)$ is an estimator of $\theta^*(f)$ defined in (6.13) given by

$$\hat{\theta}_n^*(f) = H_n^+\left[\frac{1}{n}\sum_{k=0}^{n-1}\psi(X_k)\left\{f(X_k) - \hat{\pi}_n(f)\right\}\right] , \tag{6.18}$$

where H_n^+ is the Moore-Penrose pseudoinverse of $H_n \in \mathbb{R}^{p\times p}$ defined for all $i, j \in \{1, \ldots, p\}$ by

$$(H_n)_{ij} = \frac{1}{n}\sum_{k=0}^{n-1}\langle\nabla\psi_i(X_k), \nabla\psi_j(X_k)\rangle \ . \tag{6.19}$$

We sketch informally the arguments required to justify (6.17). Since, under (I), for any $\gamma \in (0, \bar{\gamma}]$, the Markov kernel R_γ is positive Harris, by the strong law of large numbers, $\hat{\pi}_n(\psi\{f - \hat{\pi}_n(f)\})$ and H_n converge $\mathbb{P}_{\xi,\gamma}$-almost surely for any initial probability measure ξ to $\pi_\gamma(\{f - \pi_\gamma(f)\}\psi)$ and H_γ where

$$(H_\gamma)_{ij} = \pi_\gamma\left(\langle\nabla\psi_i, \nabla\psi_j\rangle\right) \ , \quad i, j \in \{1, \ldots, p\} \ . \tag{6.20}$$

If $(1, \psi_1, \ldots, \psi_p)$ is linearly independent in $\mathrm{C}(\mathbb{R}^d, \mathbb{R})$ and π_γ admits a positive density w.r.t. the Lebesgue measure, H_γ is a symmetric positive definite matrix. Hence, the sequence $(\hat{\theta}_n^*(f))_{n\geq 0}$ converges $\mathbb{P}_{\xi,\gamma}$-almost surely to

$$\theta_\gamma^*(f) = H_\gamma^{-1}\pi_\gamma\left\{(f - \pi_\gamma(f))\psi\right\} \ . \tag{6.21}$$

Under (II), using standard arguments, see Proposition 6.15, the following central limit theorem holds

$$\sqrt{n\gamma}\left\{\pi_n^{\mathrm{CV}}(f) - \pi_\gamma(f + \mathscr{L}(\theta_\gamma^*(f)^{\mathrm{T}}\psi))\right\} \overset{\mathbb{P}_{\xi,\gamma}-\text{weakly}}{\underset{n\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \gamma\sigma_{\infty,\gamma}^2(f + \mathscr{L}(\theta_\gamma^*(f)^{\mathrm{T}}\psi))) \ . \tag{6.22}$$

Moreover, under (III), since $\theta_\gamma^*(f) = \theta^*(f) + O(\gamma)$, we get that

$$\begin{aligned}\gamma\sigma_{\infty,\gamma}^2(f + \mathscr{L}(\theta_\gamma^*(f)^{\mathrm{T}}\psi)) &= \gamma\sigma_{\infty,\gamma}^2\left(f + \mathscr{L}(\theta^*(f)^{\mathrm{T}}\psi) + O(\gamma)\sum_{i=1}^{p}\mathscr{L}\psi_i\right) \\ &= \sigma_\infty^2(f + \mathscr{L}(\theta^*(f)^{\mathrm{T}}\psi)) + o(1) \ ,\end{aligned} \tag{6.23}$$

for $\gamma \downarrow 0^+$. Therefore for any $\gamma \in (0, \bar{\gamma}]$, we get

$$\sqrt{n\gamma}\left\{\pi_n^{\mathrm{CV}}(f) - \pi_\gamma(f + \mathscr{L}(\theta_\gamma^*(f)^{\mathrm{T}}\psi))\right\} \overset{\mathbb{P}_{\xi,\gamma}-\text{weakly}}{\underset{n\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma_\infty^2(f + \mathscr{L}(\theta^*(f)^{\mathrm{T}}\psi)) + o(1)) \ , \tag{6.24}$$

showing that the optimal control variate for the Langevin diffusion $\mathscr{L}(\theta^*(f)^{\mathrm{T}}\psi)$ is asymptotically optimal as $\gamma \downarrow 0^+$ for the considered MCMC algorithm. Note that (6.24) also displays the existence of a bias term $\pi_\gamma(f + \mathscr{L}(\theta^*_\gamma(f)^{\mathrm{T}}\psi)) - \pi(f)$ which vanishes when $\pi_\gamma = \pi$. As shown in Section 6.3, we may get rid of the bias term by letting the step size γ depend on the number of samples n.

6.3 Asymptotic expansion for the asymptotic variance of MCMC algorithms

In this Section, we justify (III). Let $\bar{\gamma} > 0$, $V : \mathbb{R}^d \to [1, +\infty)$ and $k \in \mathbb{N}$. Consider the following assumptions:

A 3 $(V, \bar{\gamma})$**.** *There exist* $\lambda \in [0, 1)$*,* $b < +\infty$ *and* $c > 0$ *such that*

$$\sup_{x \in \mathbb{R}^d} \{\exp(c\,\|x\|)/V(x)\} < +\infty \quad \textit{and} \quad R_\gamma V \leq \lambda^\gamma V + \gamma b\,, \quad \textit{for all } \gamma \in (0, \bar{\gamma}]. \tag{6.25}$$

Moreover, there exists $\varepsilon \in (0, 1]$ *such that for all* $\gamma \in (0, \bar{\gamma}]$ *and* $x, x' \in \{V \leq M\}$*,*

$$\|R_\gamma^{\lceil 1/\gamma \rceil}(x, \cdot) - R_\gamma^{\lceil 1/\gamma \rceil}(x', \cdot)\|_{\mathrm{TV}} \leq 1 - \varepsilon\,, \tag{6.26}$$

where

$$M > \left(\frac{4b\lambda^{-\bar{\gamma}}}{\log(1/\lambda)} - 1 \right) \vee 1\,. \tag{6.27}$$

A4 $(\bar{\gamma}, k)$**.** *There exist* $\alpha \geq 3/2$ *and a family of operators* $(\mathscr{A}_\gamma)_{\gamma \in (0, \bar{\gamma}]}$ *with* $\mathscr{A}_\gamma : \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}) \to \mathrm{C}^i_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *for* $i \in \{0, \ldots, k\}$*, such that for all* $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *and* $\gamma \in (0, \bar{\gamma}]$*,*

$$R_\gamma \varphi = \varphi + \gamma \mathscr{L}\varphi + \gamma^\alpha \mathscr{A}_\gamma \varphi\,.$$

For all $p \in \mathbb{N}$*, there exist* $C \geq 0$ *and* $q \in \mathbb{N}$ *such that for all* $i \in \{0, \ldots, k\}$*,* $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *and* $\gamma \in (0, \bar{\gamma}]$*,*

$$\|\mathscr{A}_\gamma \varphi\|_{i,q} \leq C\,\|\varphi\|_{4+i,p}\,.$$

For any $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *and* $x \in \mathbb{R}^d$*,* $\lim_{\gamma \downarrow 0^+} \mathscr{A}_\gamma \varphi(x)$ *exists (this limit is denoted* $\mathscr{A}_0 \varphi(x)$*).*

We show below and in Section 6.5 that these conditions are satisfied for the Metropolis Adjusted / Unadjusted Langevin Algorithm (MALA and ULA) algorithms (in which case γ is the stepsize in the Euler discretization of the Langevin diffusion) and also by the Random Walk Metropolis algorithm (RWM) (in which case γ is the variance of the increment distribution). In the following result, we establish the V-uniform geometric ergodicity of R_γ for $\gamma \in (0, \bar{\gamma}]$.

Lemma 6.2. *Let $V : \mathbb{R}^d \to [1, +\infty)$ and $\bar{\gamma} > 0$. Assume $\boldsymbol{A}3(V, \bar{\gamma})$. For all $\gamma \in (0, \bar{\gamma}]$, R_γ has a unique invariant measure π_γ. There exist $C > 0$ and $\rho \in (0, 1)$ such that for all $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$,*

$$\|\delta_x R_\gamma^n - \pi_\gamma\|_V \leq C \rho^{n\gamma} V(x) \; . \tag{6.28}$$

Proof. The proof is postponed to Section 6.6.1. □

Note that under $\mathbf{A}3(V, \bar{\gamma})$, iterating the drift condition (6.25), using Lemma 6.2 and $1 - \lambda^\gamma \geq \gamma \ln(1/\lambda) \lambda^\gamma$, we obtain for all $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}^*$, $x \in \mathbb{R}^d$,

$$R_\gamma^n V(x) \leq \lambda^{n\gamma} V(x) + \frac{b \lambda^{-\bar{\gamma}}}{\ln(1/\lambda)} \quad \text{and} \quad \pi_\gamma(V) \leq \frac{b \lambda^{-\bar{\gamma}}}{\ln(1/\lambda)} \; . \tag{6.29}$$

We next give an upper bound on the bias between π_γ and π, i.e. $|\pi_\gamma(\varphi) - \pi(\varphi)|$ for φ smooth enough and $\pi_\gamma \neq \pi$.

Proposition 6.3. *Assume $\boldsymbol{H}13(4)$. Let $V : \mathbb{R}^d \to [1, +\infty)$, $\bar{\gamma} > 0$ and assume $\boldsymbol{A}3(V, \bar{\gamma})$ and $\boldsymbol{A}4(\bar{\gamma}, 0)$. For all $p \in \mathbb{N}$, there exists $C \geq 0$ such that for all $\varphi \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $\gamma \in (0, \bar{\gamma}]$,*

$$|\pi_\gamma(\varphi) - \pi(\varphi)| \leq C \, \|\varphi\|_{3,p} \, \gamma^{\alpha - 1} \; . \tag{6.30}$$

Proof. The proof is postponed to Section 6.6.2. □

We now state the main theorem of this Section.

Theorem 6.4. *Let $V : \mathbb{R}^d \to [1, +\infty)$, $\bar{\gamma} > 0$. Assume $\boldsymbol{H}13(7)$, $\boldsymbol{A}3(V, \bar{\gamma})$ and $\boldsymbol{A}4(\bar{\gamma}, 3)$. Then, for all $p \in \mathbb{N}$, there exists $C \geq 0$ such that for all $f \in \mathrm{C}^6_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$, and $n \in \mathbb{N}^*$*

$$\left| \frac{\gamma}{n} \mathbb{E}_{x,\gamma} \left[\left(\sum_{k=0}^{n-1} \{ f(X_k) - \pi_\gamma(f) \} \right)^2 \right] - \sigma_\infty^2(f) \right| \leq C \, \|f\|_{6,p}^2 \left\{ \gamma^{(\alpha-1) \wedge 1} + \frac{V(x)}{n\gamma} \right\} \; , \tag{6.31}$$

where $\sigma_\infty^2(f)$ is defined in (6.7).

Proof. The proof is postponed to Section 6.6.3. □

Bias and confidence intervals In the CLT given in (6.24), the bias $\pi_\gamma(f + \mathscr{L}(\theta_\gamma^*(f)^{\mathrm{T}} \psi)) - \pi(f)$ is different from 0, except if $\pi_\gamma = \pi$. To obtain asymptotically valid confidence intervals for $\pi(f)$, we let the step size γ depend on the total number of iterations n.

Let $(\gamma_n)_{n \in \mathbb{N}^*}$ be a positive sequence and $\pi_{n,\gamma_n}^{\mathrm{CV}}(f)$ be defined in (6.17) where $(X_k)_{k \in \mathbb{N}}$ is associated to the kernel R_{γ_n}. We show that, for an appropriate sequence $(\gamma_n)_{n \in \mathbb{N}^*}$, $\pi_{n,\gamma_n}^{\mathrm{CV}}(f)$ targets $\pi(f)$ and a CLT holds with an asymptotic variance equal to $\sigma_\infty^2(f + \mathscr{L}(\theta^*(f)^{\mathrm{T}} \psi))$. The optimal control variates for the Langevin diffusion $\mathscr{L}(\theta^*(f)^{\mathrm{T}} \psi)$ is then also optimal for the MCMC algorithm of kernel R_{γ_n} in the limit $n \to +\infty$.

Theorem 6.5. *Let $V : \mathbb{R}^d \to [1, +\infty)$, $\bar{\gamma} > 0$. Assume **H** 13(10), **A** 3$(V, \bar{\gamma})$, and **A** 4$(\bar{\gamma}, 6)$. Let $f \in \mathrm{C}^9_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\psi = (\psi_1, \dots, \psi_p) : \mathbb{R}^d \to \mathbb{R}^p$, $p \in \mathbb{N}^*$, be a fixed sieve of functions such that $(1, \psi_1, \dots, \psi_p)$ is linearly independent in $\mathrm{C}(\mathbb{R}^d, \mathbb{R})$ and for all $i \in \{1, \dots, p\}$, $\psi_i \in \mathrm{C}^{11}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. Let $(\gamma_n)_{n \in \mathbb{N}^*}$ be a positive sequence satisfying $\lim_{n \to +\infty} (n\gamma_n)^{-1} + \gamma_n = 0$, $\hat{f}$ be a solution of the Poisson equation $\mathscr{L}\hat{f} = \pi(f) - f$, $\theta^*(f)$ be defined in (6.13) and ξ be a probability measure such that $\xi(V) < +\infty$. Then,*

(i) if $\pi(\mathscr{A}_0(\hat{f} - \theta^*(f)^{\mathrm{T}}\psi)) \lim_{n \to +\infty} n^{1/2}\gamma_n^{\alpha - 1/2} = 0$,

$$n^{1/2}\gamma_n^{1/2} \left\{ \pi^{\mathrm{CV}}_{n,\gamma_n}(f) - \pi(f) \right\} \overset{\mathbb{P}_{\xi,\gamma_n}-\textit{weakly}}{\underset{n \to +\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_\infty(f + \mathscr{L}(\theta^*(f)^{\mathrm{T}}\psi))) \, ,$$

(ii) if $\lim_{n \to +\infty} n^{1/2}\gamma_n^{\alpha - 1/2} = \gamma_\infty \in [0, +\infty)$,

$$n^{1/2}\gamma_n^{1/2} \left\{ \pi^{\mathrm{CV}}_{n,\gamma_n}(f) - \pi(f) \right\} \overset{\mathbb{P}_{\xi,\gamma_n}-\textit{weakly}}{\underset{n \to +\infty}{\Longrightarrow}} \mathcal{N}(\gamma_\infty \pi(\mathscr{A}_0(\hat{f} - \theta^*(f)^{\mathrm{T}}\psi)), \sigma^2_\infty(f + \mathscr{L}(\theta^*(f)^{\mathrm{T}}\psi))) \, ,$$

(iii) if $\pi(\mathscr{A}_0(\hat{f} - \theta^*(f)^{\mathrm{T}}\psi)) \liminf_{n \to +\infty} n^{1/2}\gamma_n^{\alpha - 1/2} = +\infty$,

$$\gamma_n^{1-\alpha} \left\{ \pi^{\mathrm{CV}}_{n,\gamma_n}(f) - \pi(f) \right\} \overset{\mathbb{P}_{\xi,\gamma_n}-\textit{weakly}}{\underset{n \to +\infty}{\Longrightarrow}} \pi\left(\mathscr{A}_0(\hat{f} - \theta^*(f)^{\mathrm{T}}\psi) \right) \quad ,$$

where $\pi^{\mathrm{CV}}_{n,\gamma_n}(f)$ and $\sigma^2_\infty(f)$ are defined in (6.17) and (6.7), respectively.

Proof. The proof is postponed to Section 6.B. □

Note that if the invariant distribution of R_γ is π for all $\gamma \in (0, \bar{\gamma}]$ (e.g. the case of MALA or RWM), we have under $\mathbf{A}4(\bar{\gamma}, 0)$ and by the dominated convergence theorem, $\pi(\mathscr{A}_0(\hat{f} - \theta^*(f)^{\mathrm{T}}\psi)) = 0$ and (i) always holds. For the ULA algorithm (see below), $\alpha = 2$ and setting $\gamma_n = n^{-a}$ for $a \in (0, 1)$, we have:

(i) $a \in (1/3, 1)$, a CLT without bias term,

(ii) $a = 1/3$, a CLT with a bias term,

(iii) $a \in (0, 1/3)$, a LGN.

The ULA algorithm The Markov kernel R^{ULA}_γ associated to the ULA algorithm is given for $\gamma > 0$, $x \in \mathbb{R}^d$ and $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$R^{\mathrm{ULA}}_\gamma(x, \mathrm{A}) = (4\pi\gamma)^{-d/2} \int_{\mathrm{A}} \exp\left(-(4\gamma)^{-1} \left\| y - x + \gamma \nabla U(x) \right\|^2 \right) \mathrm{d}y \, . \tag{6.32}$$

Based on the results of [DM17] and [DM16], the following lemmas enable to check $\mathbf{A}\,3(V, \bar{\gamma})$ and $\mathbf{A}4(\bar{\gamma}, k)$, $k \in \mathbb{N}$, for the ULA algorithm. Analysis of the MALA and RWM algorithms is postponed to Section 6.5. Consider the following assumptions on U.

H14. *$U : \mathbb{R}^d \to \mathbb{R}$ is gradient Lipschitz, i.e. there exists $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $\|\nabla U(x) - \nabla U(y)\| \leq L \|x - y\|$.*

H15. *There exist $\nu > 0$, $\alpha \in (1, 2]$, a minimizer $x^\star \in \arg\min U$ and $M_\nu \geq 0$ such that for all $x \in \mathbb{R}^d$, $\|x - x^\star\| \geq M_\nu$, $\langle \nabla U(x), x - x^\star \rangle \geq \nu \|x - x^\star\|^\alpha$.*

H16. *U is convex and admits a minimizer $x^\star \in \arg\min U$.*

For simplicity we have assumed that ∇U is Lipschitz but following [Bro+18], this assumption can be relaxed. Note that under **H**16, by [Bra+14, Lemma 2.2.1], there exist $\eta > 0$ and $M_\eta \geq 0$ such that for all $x \in \mathbb{R}^d$, $\|x - x^\star\| \geq M_\eta$, $\langle \nabla U(x), x - x^\star \rangle \geq U(x) - U(x^\star) \geq \eta \|x - x^\star\|$ where $x^\star \in \arg\min_{\mathbb{R}^d} U$. Let $k \in \mathbb{N}$, $k \geq 2$. Note that if $U \in \mathrm{C}^k_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and U satisfies **H**14 and **H**15 or **H**16, then **H**13(k) holds.

Lemma 6.6. *(i) Assume **H**14. R_γ^{ULA} satisfies the Doeblin condition (6.26).*

*(ii) Assume **H**14 and **H**15 or **H**16. Then **A**3(V, L^{-1}) is satisfied where V is defined for $x \in \mathbb{R}^d$ by,*

$$V(x) = \begin{cases} \exp(U(x)/2) & \text{under } \boldsymbol{H}15 \\ \exp\left((\eta/4)(1 + \|x - x^\star\|^2)^{1/2}\right) & \text{under } \boldsymbol{H}16. \end{cases}$$

Proof. The proof is postponed to Section 6.C.2. □

To establish **A**$4(\bar\gamma, 6)$, let $i \in \{0, \dots, 6\}$, $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\bar\gamma > 0$, $\gamma \in [0, \bar\gamma]$ and $x \in \mathbb{R}^d$. Using $X_1 = X_0 - \gamma \nabla U(X_0) + \sqrt{2\gamma} Z_1$ where Z_1 is an i.i.d. standard d-dimensional Gaussian vector, we get

$$\begin{aligned} &\varphi(X_1) = \varphi(x) - \gamma \langle \nabla U(x), \nabla \varphi(x) \rangle + \sqrt{2\gamma} \langle \nabla \varphi(x), Z \rangle + \gamma \, \mathrm{D}^2 \varphi(x)[Z^{\otimes 2}] \\ &- \sqrt{2} \gamma^{3/2} \, \mathrm{D}^2 \varphi(x)[\nabla U(x), Z] + (\gamma^2/2) \, \mathrm{D}^2 \varphi(x)[\nabla U(x)^{\otimes 2}] - \gamma^2 \, \mathrm{D}^3 \varphi(x)[\nabla U(x), Z^{\otimes 2}] \\ &- (1/6) \gamma^3 \, \mathrm{D}^3 \varphi(x)[\nabla U(x)^{\otimes 3}] + 2^{-1/2} \gamma^{5/2} \, \mathrm{D}^3 \varphi(x)[\nabla U(x)^{\otimes 2}, Z] \\ &+ \frac{\sqrt{2}}{3} \gamma^{3/2} \, \mathrm{D}^3 \varphi(x)[Z^{\otimes 3}] + \frac{1}{6} \int_0^1 (1 - t)^3 \, \mathrm{D}^4 \varphi(x + t(X_1 - x))[(X_1 - x)^{\otimes 4}] \mathrm{d}t \, . \end{aligned} \tag{6.33}$$

Taking the expectation in (6.33), we obtain $R_\gamma \varphi(x) = \varphi(x) + \gamma \mathscr{L} \varphi(x) + \gamma^2 \mathscr{A}_\gamma^{\mathrm{ULA}} \varphi(x)$ where,

$$\begin{aligned} &\mathscr{A}_\gamma^{\mathrm{ULA}} \varphi(x) = \frac{1}{2} \mathrm{D}^2 \varphi(x)[\nabla U(x)^{\otimes 2}] - \frac{1}{6} \gamma \, \mathrm{D}^3 \varphi(x)[\nabla U(x)^{\otimes 3}] - \mathbb{E}\left[\mathrm{D}^3 \varphi(x)[\nabla U(x), Z^{\otimes 2}]\right] \\ &+ \frac{1}{6} \int_0^1 (1 - t)^3 \mathbb{E}\left[\mathrm{D}^4 \varphi(x - t\gamma \nabla U(x) + t\sqrt{2\gamma} Z)[(-\sqrt{\gamma} \nabla U(x) + \sqrt{2} Z)^{\otimes 4}]\right] \mathrm{d}t \, . \end{aligned} \tag{6.34}$$

Taking the limit $\gamma \downarrow 0^+$ in (6.34), we get for any $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $x \in \mathbb{R}^d$, $\lim_{\gamma \downarrow 0^+} \mathscr{A}_\gamma^{\mathrm{ULA}} \varphi(x) = \mathscr{A}_0^{\mathrm{ULA}} \varphi(x)$ where

$$\mathscr{A}_0^{\mathrm{ULA}} \varphi(x) = \frac{1}{2} \mathrm{D}^2 \varphi(x)[\nabla U(x)^{\otimes 2}] - \mathbb{E}\left[\mathrm{D}^3 \varphi(x)[\nabla U(x), Z^{\otimes 2}]\right] + \frac{1}{6} \mathbb{E}\left[\mathrm{D}^4 \varphi(x)[Z^{\otimes 4}]\right] \, . \tag{6.35}$$

Summarizing this discussion, it is easy to show that

Lemma 6.7. *Assume that* $U \in \mathrm{C}^7_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$.

(i) For all $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *and* $i \in \{0, \dots, 6\}$, $\mathscr{A}^{\mathrm{ULA}}_\gamma \varphi \in \mathrm{C}^i_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *for* $\gamma > 0$ *and for any* $\bar{\gamma} > 0$ *and* $p \in \mathbb{N}$, *there exist* $C \geq 0$, $q \in \mathbb{N}$ *such that for all* $i \in \{0, \dots, 6\}$, $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *and* $\gamma \in (0, \bar{\gamma}]$, $\|\mathscr{A}^{\mathrm{ULA}}_\gamma \varphi\|_{i,q} \leq C \|\varphi\|_{4+i,p}$.

(ii) For any $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *and* $x \in \mathbb{R}^d$, $\lim_{\gamma \downarrow 0^+} \mathscr{A}^{\mathrm{ULA}}_\gamma \varphi(x) = \mathscr{A}^{\mathrm{ULA}}_0 \varphi(x)$.

If $U \in \mathrm{C}^7_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, under **H**14 and **H**15 or **H**16, by Lemma 6.6 and Lemma 6.7, **A** 3(V, L^{-1}) and **A**4$(L^{-1}, 6)$ with $\alpha = 2$ are satisfied; Theorem 6.4 and Theorem 6.5 then hold.

6.4 Numerical experiments

We illustrate the proposed control variates method on Bayesian logistic and probit regressions, see [Gel+14, Chapter 16], [MR07, Chapter 4]. The examples and the data sets are taken from [PMG14]. The code used to run the experiments is available at `https://github.com/nbrosse/controlvariates`. Let $\mathsf{Y} = (\mathsf{Y}_1, \dots, \mathsf{Y}_n) \in \{0, 1\}^N$ be a vector of binary response variables, $x \in \mathbb{R}^d$ be the regression coefficients, and $\mathsf{X} \in \mathbb{R}^{N \times d}$ be a design matrix. The log-likelihood for the logistic and probit regressions are given respectively by

$$\ell_{\log}(\mathsf{Y}|x, \mathsf{X}) = \sum_{i=1}^N \left\{ \mathsf{Y}_i \mathsf{X}_i^{\mathrm{T}} x - \ln\left(1 + \mathrm{e}^{\mathsf{X}_i^{\mathrm{T}} x}\right) \right\} ,$$
$$\ell_{\mathrm{pro}}(\mathsf{Y}|x, \mathsf{X}) = \sum_{i=1}^N \left\{ \mathsf{Y}_i \ln(\Phi(\mathsf{X}_i^{\mathrm{T}} x)) + (1 - \mathsf{Y}_i) \ln(\Phi(-\mathsf{X}_i^{\mathrm{T}} x)) \right\} ,$$

where $\mathsf{X}_i^{\mathrm{T}}$ is the i^{th} row of X for $i \in \{1, \dots, N\}$. For both models, a Gaussian prior of mean 0 and variance $\varsigma^2 \,\mathrm{Id}$ is assumed for x where $\varsigma^2 = 100$. The posterior probability distributions $\pi_{\log}$ and π_{pro} for the logistic and probit regressions are proportional for all $x \in \mathbb{R}^d$ to

$$\pi_{\log}(x|\mathsf{Y}, \mathsf{X}) \propto \exp\left(-U_{\log}(x)\right) \quad \text{with} \quad U_{\log}(x) = -\ell_{\log}(\mathsf{Y}|x, \mathsf{X}) + (2\varsigma^2)^{-1} \|x\|^2 ,$$
$$\pi_{\mathrm{pro}}(x|\mathsf{Y}, \mathsf{X}) \propto \exp\left(-U_{\mathrm{pro}}(x)\right) \quad \text{with} \quad U_{\mathrm{pro}}(x) = -\ell_{\mathrm{pro}}(\mathsf{Y}|x, \mathsf{X}) + (2\varsigma^2)^{-1} \|x\|^2 .$$

In the following lemma, we check the assumptions on $U_{\log}$ and U_{pro} in order to apply Theorem 6.4 and Theorem 6.5 for the ULA, MALA and RWM algorithms. Note that **H**17 and **H**18 are two additional conditions on U given in Section 6.5, introduced to check **A**3$(\exp(U/2), \bar{\gamma})$ and **A**4$(\bar{\gamma}, 6)$, for the RWM algorithm.

Lemma 6.8. $U_{\log}$ *and* U_{pro} *satisfy* ***H****13(k) for any* $k \in \mathbb{N}^*$, ***H****14*, ***H****16*, ***H****17 and* ***H****18*.

Proof. The proof is postponed to Section 6.C.3. □

Following [PMG14, Section 2.1], we compare two bases for the construction of a control variate, based on first and second degree polynomials. Define $\psi^{\mathrm{1st}} = (\psi_1^{\mathrm{1st}}, \ldots, \psi_d^{\mathrm{1st}})$ given for $i \in \{1, \ldots, d\}$ and $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ by $\psi_i^{\mathrm{1st}}(x) = x_i$ and $\psi^{\mathrm{2nd}} = (\psi_1^{\mathrm{2nd}}, \ldots, \psi_{d(d+3)/2}^{\mathrm{2nd}})$ given for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ by

$$\psi_k^{\mathrm{2nd}}(x) = x_k \quad \text{for } k \in \{1, \ldots, d\} \ , \quad \psi_{k+d}^{\mathrm{2nd}}(x) = x_k^2 \quad \text{for } k \in \{1, \ldots, d\} \ ,$$
$$\psi_k^{\mathrm{2nd}}(x) = x_i x_j \quad \text{for } k = 2d + (j-1)(d - j/2) + (i-j) \text{ and all } 1 \leq j < i \leq d \ .$$

ψ^{1st} and ψ^{2nd} are in $\mathrm{C}^\infty_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and are linearly independent in $\mathrm{C}(\mathbb{R}^d, \mathbb{R})$. The estimators associated to ψ^{1st} and ψ^{2nd} are referred to as CV-1 and CV-2, respectively.

For the ULA, MALA and RWM algorithms, we make a run of $n = 10^6$ samples with a burn-in period of 10^5 samples, started at the mode of the posterior. The step size is set equal to 10^{-2} for ULA and to 5×10^{-2} for MALA and RWM, the acceptance ratio in the stationary regime being close to 0.23 for RWM and 0.57 for MALA, see [RGG97; RR98]. We consider $2d$ scalar test functions $\{f_k\}_{k=1}^{2d}$ defined for all $x \in \mathbb{R}^d$ and $k \in \{1, \ldots, d\}$ by $f_k(x) = x_k$ and $f_{k+d}(x) = x_k^2$. For $k \in \{1, \ldots, 2d\}$, we compute the empirical average $\hat{\pi}_n(f_k)$ and the control variate estimator $\pi_n^{\mathrm{CV}}(f_k)$ defined in (6.1) and (6.17) respectively. For comparison purposes, the zero-variance estimators of [PMG14] using the same bases of functions ψ^{1st}, ψ^{2nd} are also computed and are referred to as ZV-1 for ψ^{1st} and ZV-2 for ψ^{2nd}. We run 100 independent Markov chains for ULA, MALA, RWM algorithms. The boxplots for the logistic example are displayed in Figure 6.2 for x_1 and x_1^2. Note the impressive decrease in the variance using the control variates for each algorithm ULA, MALA and RWM. It is worthwhile to note that for ULA, the bias $|\pi(f) - \pi_\gamma(f)|$ is reduced dramatically using the CV-2 estimator. It can be explained by the fact that for n large enough, $\theta_n^*(f)^{\mathrm{T}} \psi^{\mathrm{2nd}}$ is an efficient approximation of the solution $\hat{f}$ of the Poisson equation $\mathscr{L}\hat{f} = -(f - \pi(f))$. We then get

$$\lim_{n \to +\infty} \pi_{n,\gamma_n}^{\mathrm{CV}}(f) \approx \pi_\gamma(f) + \pi_\gamma\left(\mathscr{L}(\theta_\gamma^*(f)^{\mathrm{T}} \psi^{\mathrm{2nd}})\right) \approx \pi_\gamma(f) - \pi_\gamma\left(f - \pi(f)\right) = \pi(f)$$

where $\pi_{n,\gamma_n}^{\mathrm{CV}}(f)$ is defined in (6.17).

To have a more quantitative estimate of the variance reduction, we compute for each algorithm and test function $f \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, the spectral estimator $\hat{\sigma}_n^2(f)$ of the asymptotic variance with a Tukey-Hanning window, see [FJ10], given by

$$\hat{\sigma}_n^2(f) = \sum_{k=-(\lfloor n^{1/2} \rfloor - 1)}^{\lfloor n^{1/2} \rfloor - 1} \left\{ \frac{1}{2} + \frac{1}{2} \cos\left(\frac{\pi |k|}{\lfloor n^{1/2} \rfloor} \right) \right\} \omega_n^f(|k|) \ , \tag{6.36}$$
$$\omega_n^f(k) = \frac{1}{n} \sum_{s=0}^{n-1-k} \left\{ f(X_s) - \hat{\pi}_n(f) \right\} \left\{ f(X_{s+k}) - \hat{\pi}_n(f) \right\} \ .$$

We compute the average of these estimators $\hat{\sigma}_n^2(f)$ over the 100 independent runs of the Markov chains and the values for the logistic regression are given in Table 6.1. The Variance Reduction Factor (VRF) is defined as the ratio of the asymptotic variances

obtained by the ordinary empirical average and the control variate (or zero-variance) estimator. We again observe the considerable decrease of the asymptotic variances using control variates. In this example, our approach produces slightly larger VRFs compared to the zero-variance estimators. We obtain similar results for the probit regression; see Section 6.D.

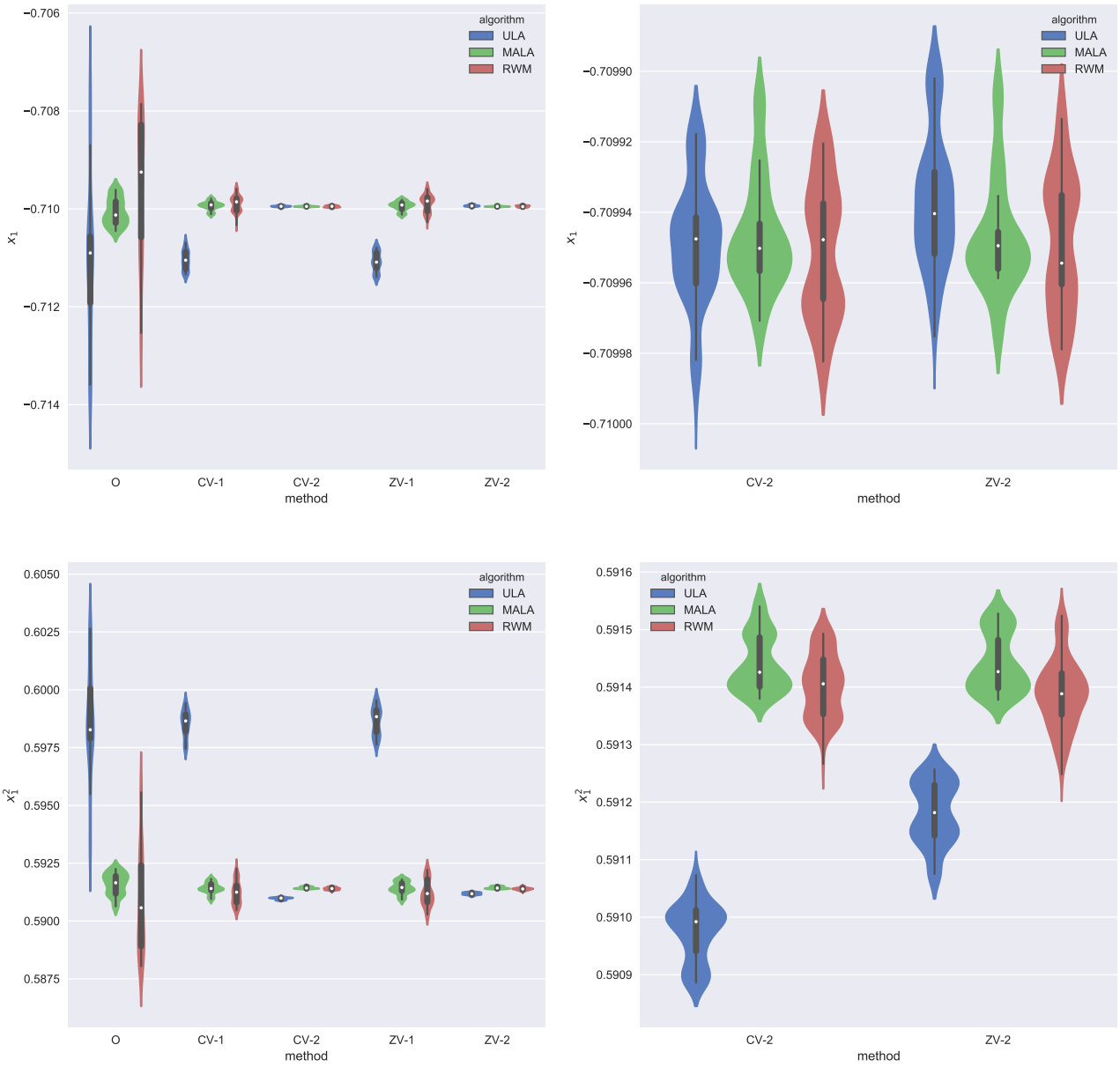

Figure 6.1: Boxplots of x_1, x_1^2 using the ULA, MALA and RWM algorithms for the logistic regression. The compared estimators are the ordinary empirical average (O), our estimator with a control variate (6.17) using first (CV-1) or second (CV-2) order polynomials for ψ, and the zero-variance estimators of [PMG14] using a first (ZV-1) or second (ZV-2) order polynomial bases. The plots in the second column are close-ups for CV-2 and ZV-2.

Table 6.1: Estimates of the asymptotic variances for ULA, MALA and RWM and each parameter x_i, x_i^2 for $i \in \{1, \ldots, d\}$, and of the variance reduction factor (VRF) on the example of the logistic regression.

		MCMC	CV-1-MCMC		CV-2-MCMC		ZV-1-MCMC		ZV-2-MCMC	
		Variance	VRF	Variance	VRF	Variance	VRF	Variance	VRF	Variance
x_1	ULA	2	33	0.061	3.2e+03	0.00062	33	0.061	3e+03	0.00066
	MALA	0.41	33	0.012	2.6e+03	0.00016	30	0.014	2.5e+03	0.00017
	RWM	1.3	33	0.039	2.6e+03	0.00049	32	0.04	2.7e+03	0.00048
x_2	ULA	10	57	0.18	8.1e+03	0.0013	53	0.19	7.4e+03	0.0014
	MALA	2.5	59	0.042	7.7e+03	0.00032	54	0.046	7.3e+03	0.00034
	RWM	5.6	52	0.11	5.6e+03	0.001	50	0.11	5.6e+03	0.001
x_2	ULA	10	56	0.18	7.3e+03	0.0014	52	0.19	6.7e+03	0.0015
	MALA	2.4	58	0.041	6.8e+03	0.00035	52	0.045	6.5e+03	0.00037
	RWM	5.6	45	0.13	5.1e+03	0.0011	42	0.13	5.1e+03	0.0011
x_4	ULA	13	26	0.5	3.9e+03	0.0033	22	0.59	3.4e+03	0.0038
	MALA	3.1	25	0.12	3.6e+03	0.00087	21	0.14	3.3e+03	0.00095
	RWM	7.5	19	0.4	2.5e+03	0.003	18	0.43	2.4e+03	0.0031
x_1^2	ULA	4.6	10	0.46	5.5e+02	0.0084	9.3	0.49	4.8e+02	0.0095
	MALA	0.98	9.6	0.1	4.6e+02	0.0021	8.6	0.11	4.2e+02	0.0023
	RWM	3	8.3	0.36	4.3e+02	0.0069	8	0.37	4.3e+02	0.0069
x_2^2	ULA	29	11	2.6	5.2e+02	0.055	10	2.8	4.7e+02	0.062
	MALA	7	11	0.64	5.2e+02	0.013	10	0.68	4.8e+02	0.014
	RWM	16	9.1	1.8	4.4e+02	0.037	8.8	1.8	4.3e+02	0.037
x_3^2	ULA	46	11	4.1	6.7e+02	0.069	10	4.5	5.9e+02	0.079
	MALA	11	11	0.97	6e+02	0.018	10	1	5.6e+02	0.019
	RWM	26	9	2.9	4.3e+02	0.061	8.6	3.1	4.2e+02	0.062
x_4^2	ULA	5.1e+02	14	37	8.2e+02	0.62	12	43	6.9e+02	0.73
	MALA	1.2e+02	14	9	7.9e+02	0.15	12	10	7.1e+02	0.17
	RWM	2.9e+02	11	27	5.8e+02	0.51	10	29	5.6e+02	0.53

6.5 The RWM and MALA algorithms

In this Section, we establish the assumptions of Theorem 6.4 and Theorem 6.5 for the RWM and MALA algorithms. For $\gamma > 0$, the Markov kernel R_γ^{RWM} of the RWM algorithm with a Gaussian proposal of mean 0 and variance 2γ is given for $x \in \mathbb{R}^d$ and $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$R_\gamma^{\mathrm{RWM}}(x,\mathrm{A}) = \int_{\mathrm{A}} \exp\left(-(4\gamma)^{-1}\left\|y-x\right\|^2\right) \min(1, \mathrm{e}^{-\tau^{\mathrm{RWM}}(x,y)}) \frac{\mathrm{d}y}{(4\pi\gamma)^{d/2}} + \delta_x(\mathrm{A}) \left\{1 - \int_{\mathbb{R}^d} \exp\left(-(4\gamma)^{-1}\left\|y-x\right\|^2\right) \min(1, \mathrm{e}^{-\tau^{\mathrm{RWM}}(x,y)}) \frac{\mathrm{d}y}{(4\pi\gamma)^{d/2}}\right\} , \tag{6.37}$$

$$\tau^{\mathrm{RWM}}(x,y) = U(y) - U(x) . \tag{6.38}$$

For all $x \in \mathbb{R}^d$ and $\gamma > 0$, define the acceptance region

$$\mathbf{A}_{x,\gamma}^{\mathrm{RWM}} = \left\{z \in \mathbb{R}^d : \tau^{\mathrm{RWM}}(x, x + \sqrt{2\gamma} z) \leq 0\right\} \tag{6.39}$$

and denote by $\partial\mathbf{A}_{x,\gamma}^{\mathrm{RWM}}$ the boundaries of the connected components of $\mathbf{A}_{x,\gamma}^{\mathrm{RWM}}$.

H17. *For all $x \in \mathbb{R}^d$ and $\gamma > 0$, $\partial\mathbf{A}_{x,\gamma}^{\mathrm{RWM}}$ is a Lebesgue null set.*

Set for all $\gamma > 0$,

$$\mathscr{A}_\gamma^{\mathrm{RWM}} = (R_\gamma^{\mathrm{RWM}} - \mathrm{Id} - \gamma\mathscr{L})/\gamma^{3/2} . \tag{6.40}$$

If $U \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, define for any $\varphi \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $x, z \in \mathbb{R}^d$,

$$\begin{aligned} \mathscr{A}_0^{\mathrm{RWM}}\varphi(x) &= -\sqrt{2}\mathbb{E}\left[\langle \nabla U(x), Z\rangle_+ \mathrm{D}^2\varphi(x)[Z^{\otimes 2}] + \zeta_0(x,Z)\langle \nabla\varphi(x), Z\rangle\right] , \\ \zeta_0(x,z) &= \left\{\mathrm{D}^2 U(x)[z^{\otimes 2}] + \langle \nabla U(x), z\rangle^2\right\} \mathbb{1}_{\langle \nabla U(x), z\rangle > 0} \\ &\quad + \left(\mathrm{D}^2 U(x)[z^{\otimes 2}]\right)_+ \mathbb{1}_{\langle \nabla U(x), z\rangle = 0} , \end{aligned} \tag{6.41}$$

where Z is a standard d-dimensional Gaussian vector.

Lemma 6.9. *(i) Assume that $U \in \mathrm{C}^7_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and **H**17. For all $i \in \{0, \dots, 6\}$ and $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\mathscr{A}_\gamma^{\mathrm{RWM}}\varphi \in \mathrm{C}^i_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ for $\gamma > 0$ and for any $\bar\gamma > 0$ and $p \in \mathbb{N}$, there exist $C \geq 0$, $q \in \mathbb{N}$ such that for all $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $\gamma \in (0, \bar\gamma]$, $\|\mathscr{A}_\gamma^{\mathrm{RWM}}\varphi\|_{0,q} \leq C\|\varphi\|_{4,p}$.*

(ii) Assume that $U \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. For any $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $x \in \mathbb{R}^d$, $\lim_{\gamma \downarrow 0^+} \mathscr{A}_\gamma^{\mathrm{RWM}}\varphi(x) = \mathscr{A}_0^{\mathrm{RWM}}\varphi(x)$.

Proof. The proof is postponed to Section 6.E.1. □

We now proceed to check the drift condition (6.25). For that purpose, consider the following additional assumption on U.

H18. *There exist $\chi, M > 0$ such that for all $x \in \mathbb{R}^d$, $\|x\| \geq M$,*

$$\|\nabla U(x)\| \geq \chi^{-1} \ , \quad \left\|\mathrm{D}^3 U(x)\right\| \leq \chi \left\|\mathrm{D}^2 U(x)\right\| \ , \quad \left\|\mathrm{D}^2 U(x)\right\| \leq \chi \|\nabla U(x)\|$$

and $\lim_{\|x\|\to+\infty} \left\|\mathrm{D}^2 U(x)\right\| / \|\nabla U(x)\|^2 = 0$.

Lemma 6.10. *Assume that $U \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and **H**18. There exists $\bar{\gamma} > 0$ such that for all $\gamma \in (0, \bar{\gamma}]$, R_γ^{RWM} satisfies the drift condition (6.25) with $V = \exp(U/2)$.*

Proof. The proof is postponed to Section 6.E.2. □

We now consider the MALA algorithm. The Markov kernel R_γ^{MALA} of the MALA algorithm, see [RT96], is given for $\gamma > 0$, $x \in \mathbb{R}^d$, and $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$R_\gamma^{\mathrm{MALA}}(x, \mathrm{A}) = \int_{\mathrm{A}} R_\gamma^{\mathrm{ULA}}(x, \mathrm{d}y) \min(1, \mathrm{e}^{-\tau_\gamma^{\mathrm{MALA}}(x,y)}) + \delta_x(\mathrm{A}) \left\{1 - \int_{\mathbb{R}^d} R_\gamma^{\mathrm{ULA}}(x, \mathrm{d}y) \min(1, \mathrm{e}^{-\tau_\gamma^{\mathrm{MALA}}(x,y)})\right\} \ , \tag{6.42}$$

$$\tau_\gamma^{\mathrm{MALA}}(x, y) = U(y) - U(x) + \frac{\|x - y + \gamma \nabla U(y)\|^2 - \|y - x + \gamma \nabla U(x)\|^2}{4\gamma} \ . \tag{6.43}$$

For all $x \in \mathbb{R}^d$ and $\gamma > 0$, define the acceptance region

$$\mathbf{A}_{x,\gamma}^{\mathrm{MALA}} = \left\{z \in \mathbb{R}^d : \tau_\gamma^{\mathrm{MALA}}(x, x - \gamma \nabla U(x) + \sqrt{2\gamma} z) \leq 0\right\}$$

and denote by $\partial \mathbf{A}_{x,\gamma}^{\mathrm{MALA}}$ the boundaries of the connected components of $\mathbf{A}_{x,\gamma}^{\mathrm{MALA}}$.

H19. *For all $x \in \mathbb{R}^d$ and $\gamma > 0$, $\partial \mathbf{A}_{x,\gamma}^{\mathrm{MALA}}$ is a Lebesgue null set.*

Under this assumption, the following Lemma shows that $\mathbf{A}4(\bar{\gamma}, k)$ is satisfied for the MALA algorithm for any $\bar{\gamma} > 0$ and $k \in \mathbb{N}$. Set for all $\gamma > 0$,

$$\mathscr{A}_\gamma^{\mathrm{MALA}} = (R_\gamma^{\mathrm{MALA}} - \mathrm{Id} - \gamma \mathscr{L})/\gamma^2 \ , \tag{6.44}$$

and define for all $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $x, z \in \mathbb{R}^d$,

$$\begin{aligned} \mathscr{A}_0^{\mathrm{MALA}} \varphi(x) &= \mathscr{A}_0^{\mathrm{ULA}} \varphi(x) - \sqrt{2} \mathbb{E}\left[\max(0, \xi_0(x, Z)) \left\langle \nabla \varphi(x), Z \right\rangle\right] \ , \\ \xi_0(x, z) &= -(\sqrt{2}/6) \, \mathrm{D}^3 U(x)[z^{\otimes 3}] + 2^{-1/2} \left\langle \nabla U(x), \mathrm{D}^2 U(x)[z] \right\rangle \ , \end{aligned} \tag{6.45}$$

where Z is a standard d-dimensional Gaussian vector and $\mathscr{A}_0^{\mathrm{ULA}} \varphi$ is given in (6.35).

Lemma 6.11. *(i) Assume that $U \in \mathrm{C}^{10}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and **H**19. For all $i \in \{0, \dots, 6\}$ and $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\mathscr{A}_\gamma^{\mathrm{MALA}} \varphi \in \mathrm{C}^i_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ for $\gamma > 0$ and for any $\bar{\gamma} > 0$ and $p \in \mathbb{N}$, there exist $C \geq 0$, $q \in \mathbb{N}$ such that for all $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $i \in \{0, \dots, 6\}$ and $\gamma \in (0, \bar{\gamma}]$, $\|\mathscr{A}_\gamma^{\mathrm{MALA}} \varphi\|_{i,q} \leq C \|\varphi\|_{4+i,p}$.*

(ii) Assume that $U \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. For any $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $x \in \mathbb{R}^d$, $\lim_{\gamma \downarrow 0^+} \mathscr{A}_\gamma^{\mathrm{MALA}} \varphi(x) = \mathscr{A}_0^{\mathrm{MALA}} \varphi(x)$.

Proof. The proof is postponed to Section 6.C.4. □

6.6 Proofs

6.6.1 Proof of Lemma 6.2

By (6.25), the drift condition $\mathrm{D}_g(V, \lambda^\gamma, \gamma b)$ is satisfied. By (6.26) and (6.27), the set $\{V \leq M\}$ is an $(\lceil 1/\gamma \rceil, 1-\varepsilon)$-Doeblin set and by the choice of M, see (6.27), $\lambda^\gamma + (2\gamma b)/(1+M) < 1$. A direct application of [Dou+18, Theorem 18.4.3] concludes the proof.

6.6.2 Proof of Proposition 6.3

Under **H**13(4), by Proposition 6.1-(iii) with $k=4$, there exist $q_1 \in \mathbb{N}$ and $C \geq 0$ such that for all $\varphi \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\hat{\varphi} \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ where $\mathscr{L}\hat{\varphi} = \pi(\varphi) - \varphi$ and $\|\hat{\varphi}\|_{4,q_1} \leq C\|\varphi\|_{3,p}$. Under $\mathbf{A}4(\bar{\gamma}, 0)$, we have for all $\gamma \in (0, \bar{\gamma}]$,

$$R_\gamma \hat{\varphi} = \hat{\varphi} + \gamma \mathscr{L}\hat{\varphi} + \gamma^\alpha \mathscr{A}_\gamma \hat{\varphi} \,. \tag{6.46}$$

By Proposition 6.1-(iii), $\mathbf{A}4(\bar{\gamma}, 0)$ and (6.29), there exist $q_2 \in \mathbb{N}$ and $C \geq 0$ such that for all $\gamma \in (0, \bar{\gamma}]$,

$$|\pi_\gamma(\mathscr{A}_\gamma \hat{\varphi})| \leq \pi_\gamma(|\mathscr{A}_\gamma \hat{\varphi}|) \leq C \, \|\mathscr{A}_\gamma \hat{\varphi}\|_{0,q_2} \leq C \, \|\hat{\varphi}\|_{4,q_1} \leq C \, \|\varphi\|_{3,p} \,. \tag{6.47}$$

Integrating (6.46) w.r.t. π_γ and using $\mathscr{L}\hat{\varphi} = \pi(\varphi) - \varphi$, we obtain that $\pi_\gamma(\varphi) - \pi(\varphi) = \gamma^{\alpha-1}\pi_\gamma(\mathscr{A}_\gamma \hat{\varphi})$. Combining this result with (6.47) concludes the proof.

6.6.3 Proof of Theorem 6.4

The proof is divided into two parts. In the first part, we derive some elementary bounds on the first and second order moments of the estimator $\hat{\pi}_n(f)$ defined in (6.1) and where $(X_k)_{k\in\mathbb{N}}$ is a Markov chain of kernel R_γ, see Lemma 6.12 below. The arguments are based solely on the study of R_γ and rely on $\mathbf{A}3(V, \bar{\gamma})$ and Lemma 6.2. We also provide an upper bound on $\mathbb{E}_{x,\gamma}\left[\{f(X_1) - R_\gamma f(x)\}^2\right]$ for $f \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, see Lemma 6.13. In a second part, we compare the discrete-time Markov chain $(X_k)_{k\in\mathbb{N}}$ and the Langevin diffusion $(Y_t)_{t\geq 0}$, see Lemma 6.14. The proof of Theorem 6.4 is then derived by a bootstrap argument based on Lemma 6.14. In the sequel, C is a non-negative constant independent of $\gamma > 0$ which may take different values at each appearance.

Note that if (6.25) holds, we obtain by Jensen's inequality and using $\sqrt{a+b} - \sqrt{a} \leq b/(2\sqrt{a})$ for all $a, b > 0$,

$$R_\gamma V^{1/2} \leq (\lambda^\gamma V + \gamma b)^{1/2} \leq \lambda^{\gamma/2} V^{1/2} + \gamma b \lambda^{-\gamma/2}/2 \,. \tag{6.48}$$

Lemma 6.12. *Let $V : \mathbb{R}^d \to [1, +\infty)$ and $\bar{\gamma} > 0$. Assume $\boldsymbol{A}3(V, \bar{\gamma})$. There exists $C > 0$ such that for all $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$, $n \in \mathbb{N}^*$ and $f \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$,*

$$\left| \mathbb{E}_{x,\gamma} \left[\sum_{k=0}^{n-1} \{f(X_k) - \pi_\gamma(f)\} \right] \right| \leq C\gamma^{-1} \|f\|_{V^{1/2}} V^{1/2}(x) \,, \tag{6.49}$$

$$\mathbb{E}_{x,\gamma} \left[\left(\sum_{k=0}^{n-1} \{f(X_k) - \pi_\gamma(f)\} \right)^2 \right] \leq C\gamma^{-2} \|f\|^2_{V^{1/2}} \left\{ n + \gamma^{-1} V(x) \right\} \,. \tag{6.50}$$

Proof. By (6.48), $\mathbf{A}3(V^{1/2}, \bar{\gamma})$ is satisfied and using Lemma 6.2 with $V^{1/2}$, and $1 - \rho^\gamma \geq \gamma \ln(1/\rho)\rho^\gamma$, we obtain for all $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$,

$$\sum_{k=0}^{+\infty} \left| R_\gamma^k \{f - \pi_\gamma(f)\}(x) \right| \leq C\gamma^{-1} \|f\|_{V^{1/2}} V^{1/2}(x) \,, \tag{6.51}$$

which gives (6.49). For all $\gamma \in (0, \bar{\gamma}]$, define $\hat{f}_\gamma : \mathbb{R}^d \to \mathbb{R}$ by $\hat{f}_\gamma(x) = \sum_{k=0}^{+\infty} R_\gamma^k \{f - \pi_\gamma(f)\}(x)$, which is a solution of the Poisson equation, $(R_\gamma - \mathrm{Id})\hat{f}_\gamma = -\{f - \pi_\gamma(f)\}$, see [MT09, Section 17.4.1]. We get for all $n \in \mathbb{N}^*$,

$$\sum_{k=0}^{n-1} \{f(X_k) - \pi_\gamma(f)\} = \hat{f}_\gamma(X_0) - \hat{f}_\gamma(X_n) + \sum_{k=0}^{n-1} \{\hat{f}_\gamma(X_{k+1}) - R_\gamma \hat{f}_\gamma(X_k)\} \,. \tag{6.52}$$

By (6.51), for all $x \in \mathbb{R}^d$,

$$\hat{f}_\gamma^2(x) \leq C\gamma^{-2} \|f\|^2_{V^{1/2}} V(x) \tag{6.53}$$

and $(\sum_{k=0}^{n-1} \{\hat{f}_\gamma(X_{k+1}) - R_\gamma \hat{f}_\gamma(X_k)\})_{n \in \mathbb{N}}$ is a square integrable martingale under $\mathbb{P}_{x,\gamma}$ for all $x \in \mathbb{R}^d$. By (6.52), we have

$$\begin{aligned} \mathbb{E}_{x,\gamma} \left[\left(\sum_{k=0}^{n-1} \{f(X_k) - \pi_\gamma(f)\} \right)^2 \right] &\leq 2\mathbb{E}_{x,\gamma} \left[\left(\hat{f}_\gamma(X_0) - \hat{f}_\gamma(X_n) \right)^2 \right] \\ &\quad + 2\mathbb{E}_{x,\gamma} \left[\sum_{k=0}^{n-1} \{\hat{f}_\gamma(X_{k+1}) - R_\gamma \hat{f}_\gamma(X_k)\}^2 \right] \,. \end{aligned} \tag{6.54}$$

Set $g_\gamma(x) = \mathbb{E}_{x,\gamma} \left[\{\hat{f}_\gamma(X_1) - R_\gamma \hat{f}_\gamma(x)\}^2 \right]$. By (6.29) and (6.53), for all $x \in \mathbb{R}^d$, $g_\gamma(x) \leq C\gamma^{-2} \|f\|^2_{V^{1/2}} V(x)$ and $\pi_\gamma(g_\gamma) \leq C\gamma^{-2} \|f\|^2_{V^{1/2}}$. By (6.51) for $f = g_\gamma$, we obtain

$$\left| \sum_{k=0}^{n-1} \{\mathbb{E}_{x,\gamma} [g_\gamma(X_k)] - \pi_\gamma(g_\gamma)\} \right| \leq C\gamma^{-3} \|f\|^2_{V^{1/2}} V(x) \,.$$

Combining this result with (6.54), we obtain (6.50). □

Lemma 6.13. *Let $\bar{\gamma} > 0$ and $k \in \mathbb{N}$. Assume that $U \in \mathrm{C}^{k+1}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $\mathbf{A}4(\bar{\gamma}, k)$. For any $p \in \mathbb{N}$, there exist $q \in \mathbb{N}$ and $C \geq 0$ such that for all $\gamma \in (0, \bar{\gamma}]$ and $f \in \mathrm{C}^{k+4}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\|\tilde{f}_\gamma\|_{k,q} \leq C\gamma \|f\|^2_{k+4,p}$ where $\tilde{f}_\gamma : \mathbb{R}^d \to \mathbb{R}$ is defined for all $x \in \mathbb{R}^d$ by*

$$\tilde{f}_\gamma(x) = \mathbb{E}_{x,\gamma}\left[\{f(X_1) - R_\gamma f(x)\}^2\right] .$$

Proof. Let $p \in \mathbb{N}$. By $\mathbf{A}4(\bar{\gamma}, k)$, for all $f \in \mathrm{C}^{k+4}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$,

$$\begin{aligned}
\tilde{f}_\gamma(x) &= \mathbb{E}_{x,\gamma}\left[\{f(X_1) - f(x) - \gamma \mathscr{L} f(x) - \gamma^\alpha \mathscr{A}_\gamma f(x)\}^2\right] \\
&= \mathbb{E}_{x,\gamma}\left[\{f(X_1) - f(x)\}^2\right] + \gamma^2 \left\{\mathscr{L} f(x) + \gamma^{\alpha-1} \mathscr{A}_\gamma f(x)\right\}^2 \\
&\qquad - 2\gamma \left\{\mathscr{L} f(x) + \gamma^{\alpha-1} \mathscr{A}_\gamma f(x)\right\} \mathbb{E}_{x,\gamma}\left[f(X_1) - f(x)\right] . \qquad (6.55)
\end{aligned}$$

Besides, for all $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}_{x,\gamma}\left[\{f(X_1) - f(x)\}^2\right] &= \mathbb{E}_{x,\gamma}\left[f^2(X_1)\right] + f^2(x) - 2f(x)\mathbb{E}_{x,\gamma}\left[f(X_1)\right] \\
&= \gamma \mathscr{L}(f^2)(x) + \gamma^\alpha \mathscr{A}_\gamma(f^2)(x) - 2\gamma f(x) \mathscr{L} f(x) - 2\gamma^\alpha f(x) \mathscr{A}_\gamma f(x) \\
&= \gamma \left\{2 \left\|\nabla f(x)\right\|^2 + \gamma^{\alpha-1} \left(\mathscr{A}_\gamma(f^2)(x) - 2f(x) \mathscr{A}_\gamma f(x)\right)\right\} \qquad (6.56)
\end{aligned}$$

and $\mathbb{E}_{x,\gamma}\left[f(X_1) - f(x)\right] = \gamma \mathscr{L} f(x) + \gamma^\alpha \mathscr{A}_\gamma f(x)$. Then, combining (6.55) and (6.56), under $\mathbf{A}4(\bar{\gamma}, k)$, $\tilde{f}_\gamma \in \mathrm{C}^k_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and there exist $q \in \mathbb{N}$ and $C \geq 0$ such that $\|\tilde{f}_\gamma\|_{k,q} \leq C\gamma \|f\|^2_{k+4,p}$. □

Lemma 6.14. *Let $V : \mathbb{R}^d \to [1, +\infty)$ and $\bar{\gamma} > 0$. Assume $\boldsymbol{H}13(4)$, $\mathbf{A}3(V, \bar{\gamma})$ and $\mathbf{A}4(\bar{\gamma}, 0)$. Then, for all $p \in \mathbb{N}$, there exists $C > 0$ such that for all $f \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and $n \in \mathbb{N}^*$,*

$$\begin{aligned}
&\left| \mathbb{E}_{x,\gamma}\left[\frac{1}{n}\left(\sum_{k=0}^{n-1} \{f(X_k) - \pi_\gamma(f)\}\right)^2\right] - \frac{\sigma^2_\infty(f)}{\gamma} \right| \leq C \Bigg\{ \|f\|^2_{3,p} \gamma^{(\alpha-2)\wedge 0} + \frac{\|f\|^2_{3,p} V(x)}{n\gamma^2} \\
&+ A_1^f(x,n,\gamma) + \frac{\|f\|^2_{3,p} V^{1/2}(x) \gamma^{(\alpha/2-1)\wedge 0}}{n^{1/2}\gamma} + A_1^f(x,n,\gamma)^{1/2} \|f\|_{3,p} \left(\gamma^{-1/2} + \frac{V^{1/2}(x)}{n^{1/2}\gamma}\right) \Bigg\} , \qquad (6.57)
\end{aligned}$$

where $\sigma^2_\infty(f)$ is defined in (6.7) and

$$A_1^f(x,n,\gamma) = \frac{\gamma^{2(\alpha-1)}}{n} \mathbb{E}_{x,\gamma}\left[\left(\sum_{k=0}^{n-1} \left\{\mathscr{A}_\gamma \hat{f}(X_k) - \gamma^{1-\alpha}\left(\pi_\gamma(f) - \pi(f)\right)\right\}\right)^2\right] , \qquad (6.58)$$

$\alpha, \mathscr{A}_\gamma$ are given in $\mathbf{A}4(\bar{\gamma}, 0)$, and $\hat{f}$ is a solution of $\mathscr{L}\hat{f} = -\{f - \pi(f)\}$. Moreover,

$$A_1^f(x,n,\gamma) \leq C\gamma^{2(\alpha-2)} \|f\|^2_{3,p} \{1 + V(x)/(n\gamma)\} . \qquad (6.59)$$

Proof. Under **H**13(4) and by Proposition 6.1-(iii), let $\hat{f} \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ be a solution of the Poisson equation $\mathscr{L}\hat{f} = -\{f - \pi(f)\}$. Under **A**4$(\bar{\gamma}, 0)$, we get for all $\gamma \in (0, \bar{\gamma}]$,

$$R_\gamma \hat{f} = \hat{f} + \gamma \mathscr{L}\hat{f} + \gamma^\alpha \mathscr{A}_\gamma \hat{f} \; . \tag{6.60}$$

Since $\mathscr{L}\hat{f} = -\{f - \pi(f)\}$, we have for all $n \in \mathbb{N}^*$ and $\gamma \in (0, \bar{\gamma}]$,

$$\begin{aligned} \sum_{k=0}^{n-1} \{f(X_k) - \pi_\gamma(f)\} = \frac{\hat{f}(X_0) - \hat{f}(X_n)}{\gamma} + \frac{1}{\gamma} \sum_{k=0}^{n-1} \left\{ \hat{f}(X_{k+1}) - R_\gamma \hat{f}(X_k) \right\} \\ + \gamma^{\alpha-1} \sum_{k=0}^{n-1} \left\{ \mathscr{A}_\gamma \hat{f}(X_k) - \gamma^{1-\alpha} \left(\pi_\gamma(f) - \pi(f) \right) \right\} \; . \end{aligned} \tag{6.61}$$

Consider the following decomposition

$$\mathbb{E}_{x,\gamma} \left[\frac{1}{n} \left(\sum_{k=0}^{n-1} \{f(X_k) - \pi_\gamma(f)\} \right)^2 \right] = \sum_{i=1}^{4} A_i^f(x, n, \gamma) \; , \tag{6.62}$$

where $A_1^f(x, n, \gamma)$ is given in (6.58),

$$A_2^f(x, n, \gamma) = \mathbb{E}_{x,\gamma} \left[(\hat{f}(X_0) - \hat{f}(X_n))^2 / (n\gamma^2) \right] \; , \tag{6.63}$$

$$A_3^f(x, n, \gamma) = \mathbb{E}_{x,\gamma} \left[\frac{1}{n\gamma^2} \left(\sum_{k=0}^{n-1} \hat{f}(X_{k+1}) - R_\gamma \hat{f}(X_k) \right)^2 \right] \; ,$$

and by Cauchy-Schwarz inequality,

$$(1/2) \left| A_4^f(x, n, \gamma) \right| \leq \sum_{1 \leq i < j \leq 3} A_i^f(x, n, \gamma)^{1/2} A_j^f(x, n, \gamma)^{1/2} \; . \tag{6.64}$$

We show below that $\max_{i \in \{1,\ldots,4\}} \left| A_i^f(x, n, \gamma) \right| < +\infty$ for any $f \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. By Proposition 6.1-(iii), there exists $q_1 \in \mathbb{N}$ such that $\|\hat{f}\|_{4,q_1} \leq C \|f\|_{3,p}$ and combining it with **A**3$(V, \bar{\gamma})$ and (6.29), we obtain for all $x \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}^*$,

$$A_2^f(x, n, \gamma) \leq C \, \|f\|_{3,p}^2 \, V(x) / (n\gamma^2) \; . \tag{6.65}$$

For all $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$, set $g_\gamma(x) = \mathbb{E}_{x,\gamma} \left[\left\{ \hat{f}(X_1) - R_\gamma \hat{f}(x) \right\}^2 \right]$. By Proposition 6.1-(iii) and Lemma 6.13 with $k = 0$, $g_\gamma \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and for all $\gamma \in (0, \bar{\gamma}]$, $\|g_\gamma\|_V \leq C\gamma \|f\|_{3,p}^2$. Since $(\sum_{k=0}^{n-1} \hat{f}(X_{k+1}) - R_\gamma \hat{f}(X_k))_{k \in \mathbb{N}}$ is a $\mathbb{P}_{x,\gamma}$-square integrable martingale, for all $x \in \mathbb{R}^d$, $n \in \mathbb{N}^*$ and $\gamma \in (0, \bar{\gamma}]$, we have by the Markov property

$$A_3^f(x, n, \gamma) = \gamma^{-2} \mathbb{E}_{x,\gamma} \left[\frac{1}{n} \sum_{k=0}^{n-1} g_\gamma(X_k) \right] \; . \tag{6.66}$$

By Lemma 6.12, eq. (6.49), we have for all $x \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}^*$,

$$\left| \mathbb{E}_{x,\gamma} \left[\frac{1}{n} \sum_{k=0}^{n-1} g_\gamma(X_k) \right] - \pi_\gamma(g_\gamma) \right| \leq C \left\| g_\gamma \right\|_V \frac{V(x)}{n\gamma} \leq C \left\| f \right\|_{3,p}^2 \frac{V(x)}{n} \,. \tag{6.67}$$

We now show that $\pi_\gamma(g_\gamma)$ is approximately equal to $\gamma\sigma_\infty^2(f)$. Observe that

$$\begin{aligned} \pi_\gamma(g_\gamma) &= \mathbb{E}_{\pi_\gamma,\gamma} \left[\left\{ \hat{f}(X_1) - R_\gamma \hat{f}(X_0) \right\}^2 \right] \\ &= \mathbb{E}_{\pi_\gamma,\gamma} \left[\left\{ \hat{f}(X_1) - \hat{f}(X_0) \right\}^2 \right] - \mathbb{E}_{\pi_\gamma,\gamma} \left[\left\{ \hat{f}(X_0) - R_\gamma \hat{f}(X_0) \right\}^2 \right] \,. \end{aligned} \tag{6.68}$$

We have by (6.60)

$$\mathbb{E}_{\pi_\gamma,\gamma} \left[\left\{ \hat{f}(X_1) - \hat{f}(X_0) \right\}^2 \right] = 2\mathbb{E}_{\pi_\gamma,\gamma} \left[\hat{f}(X_0) \left\{ \hat{f}(X_0) - R_\gamma \hat{f}(X_0) \right\} \right] \tag{6.69}$$

$$= -2\gamma\pi_\gamma(\hat{f}\mathscr{L}\hat{f}) - 2\gamma^\alpha \pi_\gamma(\hat{f}\mathscr{A}_\gamma \hat{f}) \,. \tag{6.70}$$

In the next step, we consider separately the cases $\pi_\gamma = \pi$ and $\pi_\gamma \neq \pi$.

- If $\pi = \pi_\gamma$, $-\pi_\gamma(\hat{f}\mathscr{L}\hat{f}) = (1/2)\sigma_\infty^2(f)$.
- If $\pi_\gamma \neq \pi$, $(-\mathscr{L}\hat{f})\hat{f} \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and by Proposition 6.3, for all $\gamma \in (0, \bar{\gamma}]$,

$$\left| \pi_\gamma(\hat{f}\mathscr{L}\hat{f}) - \pi(\hat{f}\mathscr{L}\hat{f}) \right| \leq C \left\| f \right\|_{3,p}^2 \gamma^{\alpha-1} \,.$$

In both cases, using $\mathbf{A}4(\bar{\gamma}, 0)$, (6.29) and $\left|\pi_\gamma(\hat{f}\mathscr{A}_\gamma\hat{f})\right| \leq C\|f\|_{3,p}^2$, (6.70) becomes

$$\left| \mathbb{E}_{\pi_\gamma,\gamma} \left[\left\{ \hat{f}(X_1) - \hat{f}(X_0) \right\}^2 \right] - \gamma\sigma_\infty^2(f) \right| \leq C \left\| f \right\|_{3,p}^2 \gamma^\alpha \,. \tag{6.71}$$

By $\mathbf{A}4(\bar{\gamma}, 0)$, (6.29) and (6.60), $\mathbb{E}_{\pi_\gamma,\gamma} \left[\left\{ \hat{f}(X_0) - R_\gamma \hat{f}(X_0) \right\}^2 \right] \leq C\|f\|_{3,p}^2 \gamma^2$. Combining this result, (6.68) and (6.71),

$$\left| \pi_\gamma(g_\gamma) - \gamma\sigma_\infty^2(f) \right| \leq C \left\| f \right\|_{3,p}^2 \gamma^{\alpha\wedge 2} \,. \tag{6.72}$$

Combining it with (6.67), for all $x \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}^*$,

$$\left| A_3^f(x, n, \gamma) - \frac{\sigma_\infty^2(f)}{\gamma} \right| \leq C \left\| f \right\|_{3,p}^2 \left\{ \gamma^{(\alpha-2)\wedge 0} + \frac{V(x)}{n\gamma^2} \right\} \,. \tag{6.73}$$

Combining (6.62), (6.64), (6.65) and (6.73) give (6.57).

For any $\gamma \in (0, \bar{\gamma}]$, by (6.60), $\pi_\gamma(\mathscr{A}_\gamma \hat{f}) = \gamma^{1-\alpha}\{\pi_\gamma(f) - \pi(f)\}$. Hence, by Lemma 6.12 and $\mathbf{A}4(\bar{\gamma}, 0)$, there exists $q_3 \in \mathbb{N}$ such that for all $x \in \mathbb{R}^d$ and $n \in \mathbb{N}^*$,

$$\begin{aligned} A_1^f(x, n, \gamma) &\leq C\gamma^{2(\alpha-2)} \left\| \mathscr{A}_\gamma \hat{f} \right\|_{V^{1/2}}^2 \{1 + V(x)/(n\gamma)\} \\ &\leq C\gamma^{2(\alpha-2)} \left\| \mathscr{A}_\gamma \hat{f} \right\|_{0,q_3}^2 \{1 + V(x)/(n\gamma)\} \\ &\leq C\gamma^{2(\alpha-2)} \|f\|_{3,p}^2 \{1 + V(x)/(n\gamma)\} \ , \end{aligned} \tag{6.74}$$

which gives (6.59). Finally, for any $f \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, there exists $p_f \in \mathbb{N}$ such that $\|f\|_{3,p_f} < +\infty$, and by (6.64), (6.65), (6.73) and (6.74), $\max_{i \in \{1,\ldots,4\}} \left| A_i^f(x, n, \gamma) \right| < +\infty$. □

Proof of Theorem 6.4. To get the result, we use a bootstrap argument based on Lemma 6.14. Let $\hat{f} \in \mathrm{C}^7_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ be given by Proposition 6.1-(iii). We first apply Lemma 6.14 to the function $\mathscr{A}_\gamma \hat{f} \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. Note that by Proposition 6.1-(iii), $\mathbf{A}4(\bar{\gamma}, 3)$ and (6.59), there exist $q_1, q_2 \in \mathbb{N}$ such that for all $\gamma \in (0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and $n \in \mathbb{N}^*$,

$$\begin{aligned} \left\| \mathscr{A}_\gamma \hat{f} \right\|_{3,q_1} &\leq C \left\| \hat{f} \right\|_{7,q_2} \leq C \|f\|_{6,p} \ , \\ A_1^{\mathscr{A}_\gamma \hat{f}}(x, n, \gamma) &\leq C\gamma^{2(\alpha-2)} \|f\|_{6,p}^2 \{1 + V(x)/(n\gamma)\} \ . \end{aligned}$$

By Lemma 6.14 applied to the function $\mathscr{A}_\gamma \hat{f}$, (6.57) and using $\pi_\gamma(\mathscr{A}_\gamma \hat{f}) = \gamma^{1-\alpha}\{\pi_\gamma(f) - \pi(f)\}$, we obtain for all $x \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}^*$

$$\begin{aligned} A_1^f(x, n, \gamma) &= \frac{\gamma^{2(\alpha-1)}}{n} \mathbb{E}_{x,\gamma} \left[\left(\sum_{k=0}^{n-1} \left\{ \mathscr{A}_\gamma \hat{f}(X_k) - \gamma^{1-\alpha} \left(\pi_\gamma(f) - \pi(f) \right) \right\} \right)^2 \right] \\ &\leq C \|f\|_{6,p}^2 \gamma^{2(\alpha-1)} \left\{ \gamma^{-1} + \frac{V^{1/2}(x)\gamma^{(\alpha/2-1)\wedge 0}}{n^{1/2}\gamma} + \frac{V(x)}{n\gamma^2} \right\} . \end{aligned} \tag{6.75}$$

Combining Lemma 6.14 applied to the function $f \in \mathrm{C}^6_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, (6.57) and the upper bound (6.75) give

$$\begin{aligned} \left| \mathbb{E}_{x,\gamma} \left[\frac{1}{n} \left(\sum_{k=0}^{n-1} \{f(X_k) - \pi_\gamma(f)\} \right)^2 \right] - \frac{\sigma_\infty^2(f)}{\gamma} \right| &\leq C \|f\|_{6,p}^2 \left\{ \gamma^{(\alpha-2)\wedge 0} \right. \\ &+ \frac{V^{1/2}(x)\gamma^{(\alpha/2-1)\wedge 0}}{n^{1/2}\gamma} + \frac{V^{1/4}(x)\gamma^{(\alpha/4-1/2)\wedge 0}}{n^{1/4}\gamma^{3/2-\alpha}} \left(\gamma^{-1/2} + \frac{V^{1/2}(x)}{n^{1/2}\gamma} \right) + \left. \frac{V(x)}{n\gamma^2} \right\} . \end{aligned} \tag{6.76}$$

Note that by Young's inequality, we get for all $x \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}]$ and $n \in \mathbb{N}^*$,

$$\frac{V^{1/4}(x)}{n^{1/4}\gamma^{1/2}}\gamma^{(5\alpha/4-2)\wedge(\alpha-3/2)} \leq \frac{1}{4}\frac{V(x)}{n\gamma^2} + \frac{3}{4}\gamma^{(5\alpha/3-8/3)\wedge(4\alpha/3-2)} ,$$
$$2\frac{V^{1/2}(x)\gamma^{(\alpha/2-1)\wedge 0}}{n^{1/2}\gamma} \leq \frac{V(x)}{n\gamma^2} + \gamma^{(\alpha-2)\wedge 0} ,$$
$$\frac{V(x)^{3/4}}{n^{3/4}\gamma^{3/2}}\gamma^{(5\alpha/4-3/2)\wedge(\alpha-1)} \leq \frac{4}{3}\frac{V(x)}{n\gamma^2} + \frac{1}{4}\gamma^{(5\alpha-6)\wedge 4(\alpha-1)} .$$

Combining it with the fact that for all $\alpha \geq 3/2$ and $\gamma \in (0, \bar{\gamma}]$,

$$\gamma^{(5\alpha/3-8/3)\wedge(4\alpha-2)} + \gamma^{(5\alpha-6)\wedge 4(\alpha-1)} \leq C\gamma^{(\alpha-2)\wedge 0} ,$$

concludes the proof. □

Acknowledgements

This work was supported by the École Polytechnique Data Science Initiative.

6.A Strong Law of Large Numbers and Central Limit Theorem for the control variates estimator

Proposition 6.15. *Let $V : \mathbb{R}^d \to [1, +\infty)$, $\bar{\gamma} > 0$. Assume **A**3$(V, \bar{\gamma})$ and that π_γ admits a positive density w.r.t. the Lebesgue measure for all $\gamma \in (0, \bar{\gamma}]$. Let $f \in \mathrm{C}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\psi = (\psi_1, \ldots, \psi_p) : \mathbb{R}^d \to \mathbb{R}^p$, $p \in \mathbb{N}^*$ be a fixed sieve of functions such that $(1, \psi_1, \ldots, \psi_p)$ is linearly independent in $\mathrm{C}(\mathbb{R}^d, \mathbb{R})$ and for all $i \in \{1, \ldots, p\}$, $\psi_i \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. Then, for any initial probability measure ξ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$*

$$\lim_{n\to+\infty} \hat{\theta}^*_n(f) = \theta^*_\gamma(f) , \quad \mathbb{P}_{\xi,\gamma} - a.s , \tag{6.77}$$

where $\hat{\theta}^*_n(f)$ and $\theta^*_\gamma(f)$ are defined in (6.18) and (6.21) respectively. Moreover, the following CLT holds for $\pi^{\mathrm{CV}}_n(f)$ defined in (6.17),

$$\sqrt{n}\left\{\pi^{\mathrm{CV}}_n(f) - \pi_\gamma(f + \mathscr{L}(\theta^*_\gamma(f)^{\mathrm{T}}\psi))\right\} \overset{\mathbb{P}_{\xi,\gamma}-weakly}{\underset{n\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_{\infty,\gamma}(f + \mathscr{L}(\theta^*_\gamma(f)^{\mathrm{T}}\psi))) , \tag{6.78}$$

where $\sigma^2_{\infty,\gamma}(f + \mathscr{L}(\theta^*_\gamma(f)^{\mathrm{T}}\psi))$ is defined in (6.14).

Proof. By [Dou+18, Proposition 5.2.14], $\hat{\pi}_n(\psi\{f - \hat{\pi}_n(f)\})$ and H_n converges $\mathbb{P}_{\xi,\gamma}$-almost surely to $\pi_\gamma(\{f - \pi_\gamma(f)\}\psi)$ and H_γ where H_γ is a symmetric positive definite matrix defined in (6.20), and we obtain (6.77). Denote $W_{\gamma,n} \in \mathbb{R}^{p+1}$ for $n \in \mathbb{N}^*$ and $\gamma \in (0, \bar{\gamma}]$ by

$$W_{\gamma,n} = \sqrt{n}\left(\hat{\pi}_n(f) - \pi_\gamma(f), \hat{\pi}_n(\mathscr{L}\psi) - \pi_\gamma(\mathscr{L}\psi)\right) .$$

By [Dou+18, Proposition 21.1.3 and Theorem 21.2.11], $((1,\theta)^{\mathrm{T}} W_{\gamma,n})_{n\in\mathbb{N}^*}$ converges $\mathbb{P}_{\xi,\gamma}$-weakly, for every $\theta \in \mathbb{R}^p$ and any initial probability measure ξ, to a one-dimensional Gaussian variable of mean 0 and variance $\sigma^2_{\infty,\gamma}(f + \mathscr{L}(\theta^{\mathrm{T}}\psi))$. By the Cramér-Wold theorem, $(W_{\gamma,n})_{n\in\mathbb{N}^*}$ converges $\mathbb{P}_{\xi,\gamma}$-weakly to a $(p+1)$-dimensional Gaussian vector W_γ for any initial probability measure ξ, of mean 0 and covariance matrix

$$\pi_\gamma \left((\hat{f}_\gamma, \widehat{\mathscr{L}\psi}_\gamma)(\hat{f}_\gamma, \widehat{\mathscr{L}\psi}_\gamma)^{\mathrm{T}} - (R_\gamma \hat{f}_\gamma, R_\gamma \widehat{\mathscr{L}\psi}_\gamma)(R_\gamma \hat{f}_\gamma, R_\gamma \widehat{\mathscr{L}\psi}_\gamma)^{\mathrm{T}} \right) ,$$

where $\hat{f}_\gamma$ and $\widehat{\mathscr{L}\psi}_\gamma$ are solutions of the Poisson equations

$$(R_\gamma - \mathrm{Id})\hat{f}_\gamma = -(f - \pi_\gamma(f)) \quad , \quad (R_\gamma - \mathrm{Id})\widehat{\mathscr{L}\psi}_\gamma = -(\mathscr{L}\psi - \pi_\gamma(\mathscr{L}\psi)) \ .$$

By Slutsky's theorem, $(\hat{\theta}^*_n(f), W_{\gamma,n})_{n\in\mathbb{N}^*}$ converges $\mathbb{P}_{\xi,\gamma}$-weakly to $(\theta^*_\gamma(f), W_\gamma)$ and we obtain (6.78). □

Note that for the MALA and RWM algorithms, $\pi_\gamma = \pi$ and π has a positive density w.r.t. Leb (where Leb denotes the Lebesgue measure on $\mathbb{R}^d$). For ULA, since R^{ULA}_γ is Leb-irreducible for $\gamma > 0$, Leb is absolutely continuous w.r.t. π_γ. Indeed, π_γ is a maximal irreducibility measure and then Leb $\ll \pi_\gamma$.

6.B Law of Large Numbers and Central Limit Theorem for a step size γ_n function of the number of samples n

In this Section, we move away from the formalism of the canonical space to construct iteratively an array of Markov chains on the same filtered probability space, which allows us to give a precise meaning to the convergence in law of Theorem 6.5. Note first that every homogeneous Markov chain $(X_k)_{k\in\mathbb{N}}$ with values in $\mathbb{R}^d$ can be represented as a random iterative sequence, i.e. $X_{k+1} = F(X_k, \zeta_{k+1})$, where $(\zeta_k)_{k\in\mathbb{N}^*}$ is an i.i.d. sequence of uniform random variables on $[0,1]$, X_0 is independent of $(\zeta_k)_{k\in\mathbb{N}^*}$ and F is a measurable function. See for example [Dou+18, Section 1.3.2] for a proof for $\mathbb{R}$-valued Markov chains, which can be extended to any Polish space by Kuratowski's theorem [BS78, Corollary 7.16.1].

Let $(\zeta_k)_{k\in\mathbb{N}^*}$ be an i.i.d. sequence of uniform random variables on $[0,1]$ and Ξ be a random variable distributed according to the initial probability measure ξ, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ defined for all $k \in \mathbb{N}$ by $\mathcal{F}_k = \sigma(\Xi, \zeta_1, \ldots, \zeta_k)$. Let $(\gamma_n)_{n\in\mathbb{N}^*}$ be a positive sequence. By the preceding discussion, for all $n \in \mathbb{N}^*$, there exists a Borel measurable function $F_{\gamma_n} : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ such that the process $(X^n_k)_{k\in\mathbb{N}}$ defined for all $k \in \mathbb{N}$ by

$$X^n_{k+1} = F_{\gamma_n}(X^n_k, \zeta_{k+1}) \quad \text{and} \quad X^n_0 = \Xi \, , \tag{6.79}$$

is a Markov chain on $(\Omega, (\mathcal{F}_k)_{k\in\mathbb{N}})$ associated with the Markov kernel R_{γ_n}.

In the sequel, C is a non-negative constant independent of $n \in \mathbb{N}^*$ which may take different values at each appearance. We first derive a Law of Large Numbers for the array

$\{(X_k^n)_{k\in\{0,\dots,n-1\}},\ n\in\mathbb{N}\}$ in Lemma 6.16. As an application, we show in Lemma 6.17 that $\hat{\theta}_n^*(f)$ converges in probability to $\theta^*(f)$ for a smooth $f:\mathbb{R}^d\to\mathbb{R}$, where $\hat{\theta}_n^*(f)$ and $\theta^*(f)$ are defined in (6.18) and (6.13), relatively to $(X_k^n)_{k\in\mathbb{N}}$. A Central Limit Theorem is provided in Proposition 6.18. Combining these results, we obtain the proof of Theorem 6.5.

Lemma 6.16. *Let $V:\mathbb{R}^d\to[1,+\infty)$, $\bar{\gamma}>0$. Assume **H**13(4), **A**3($V,\bar{\gamma}$), and **A**4($\bar{\gamma},0$). Let $\{(X_k^n)_{k\in\{0,\dots,n-1\}},\ n\in\mathbb{N}\}$ be defined in (6.79) and assume that $\xi(V)<+\infty$. Let $f\in\mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d,\mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}^*}$ be a positive sequence such that $\gamma_n\le\bar{\gamma}$ for all $n\in\mathbb{N}^*$, and $\lim_{n\to+\infty}(n\gamma_n)^{-1}+\gamma_n=0$. Then,*

$$n^{-1}\sum_{k=0}^{n-1}\{f(X_k^n)-\pi(f)\}\underset{n\to+\infty}{\overset{\mathbb{P}}{\Longrightarrow}}0\,.$$

Proof. Let $f\in\mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d,\mathbb{R})$ and $p\in\mathbb{N}$ such that $\|f\|_{3,p}<+\infty$. By Proposition 6.1-(iii), there exists $\hat{f}\in\mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d,\mathbb{R})$ such that $\mathscr{L}\hat{f}=-(f-\pi(f))$. By (6.61), we have for all $n\in\mathbb{N}^*$,

$$n^{-1}\sum_{k=0}^{n-1}\{f(X_k^n)-\pi(f)\}=\sum_{i=1}^{4}T_i^f(n)\,,$$

where

$$\begin{aligned}
T_1^f(n)&=(n\gamma_n)^{-1}\left\{\hat{f}(X_0^n)-\hat{f}(X_n^n)\right\}\,,\\
T_2^f(n)&=(n\gamma_n)^{-1}\sum_{k=0}^{n-1}\left\{\hat{f}(X_{k+1}^n)-R_{\gamma_n}\hat{f}(X_k^n)\right\}\,,\\
T_3^f(n)&=n^{-1}\gamma_n^{\alpha-1}\sum_{k=0}^{n-1}\mathscr{A}_{\gamma_n}\hat{f}(X_k^n)\,.
\end{aligned}$$

By (6.29), $\lim_{n\to+\infty}\mathbb{E}\left[\left(T_1^f(n)\right)^2\right]=0$. Set $g_{\gamma_n}(x)=\mathbb{E}\left[\left(\hat{f}(X_1^n)-R_{\gamma_n}\hat{f}(x)\right)^2\right]$. By Lemma 6.13 with $k=0$, $g_{\gamma_n}\in\mathrm{C}^0_{\mathrm{poly}}(\mathbb{R}^d,\mathbb{R})$ and there exists $q_1\in\mathbb{N}$ such that for all $n\in\mathbb{N}^*$, $\|g_{\gamma_n}\|_{0,q_1}\le C\gamma_n\|f\|_{3,p}^2$. By the Markov property and (6.29), we obtain for all $n\in\mathbb{N}^*$,

$$\mathbb{E}\left[\left(T_2^f(n)\right)^2\right]=(n\gamma_n)^{-2}\sum_{k=0}^{n-1}\mathbb{E}\left[g_{\gamma_n}(X_k^n)\right]\le C(n\gamma_n)^{-2}n\,\|g_{\gamma_n}\|_{0,q_1}\le C(n\gamma_n)^{-1}\,\|f\|_{3,p}^2\,,$$

and $\lim_{n\to+\infty}\mathbb{E}\left[\left(T_2^f(n)\right)^2\right]=0$ by assumption on $(\gamma_n)_{n\in\mathbb{N}^*}$. By **A**4($\bar{\gamma},0$), there exists q_2 such that for all $n\in\mathbb{N}^*$, $\|\mathscr{A}_{\gamma_n}\hat{f}\|_{0,q_2}\le C\|f\|_{3,p}$ and we get

$$\mathbb{E}\left[\left(T_3^f(n)\right)^2\right]\le n^{-1}\gamma_n^{2(\alpha-1)}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(\mathscr{A}_{\gamma_n}\hat{f}(X_k^n)\right)^2\right]\le C\gamma_n^{2(\alpha-1)}\,\|f\|_{3,p}^2\,,$$

and $\lim_{n\to+\infty} \mathbb{E}\left[\left(T_3^f(n)\right)^2\right] = 0$ by assumption on $(\gamma_n)_{n\in\mathbb{N}^*}$, which concludes the proof. □

Lemma 6.17. *Let $V : \mathbb{R}^d \to [1, +\infty)$, $\bar{\gamma} > 0$. Assume **H**13(4), **A**3$(V, \bar{\gamma})$, and **A**4$(\bar{\gamma}, 0)$. Let $\{(X_k^n)_{k\in\{0,\dots,n-1\}},\ n \in \mathbb{N}\}$ be defined in (6.79) and assume that $\xi(V) < +\infty$. Let $f \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\psi = (\psi_1, \dots, \psi_p) : \mathbb{R}^d \to \mathbb{R}^p$, $p \in \mathbb{N}^*$ be a fixed sieve of functions such that $(1, \psi_1, \dots, \psi_p)$ is linearly independent in $\mathrm{C}(\mathbb{R}^d, \mathbb{R})$ and for all $i \in \{1, \dots, p\}$, $\psi_i \in \mathrm{C}^5_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$. Let $(\gamma_n)_{n\in\mathbb{N}^*}$ be a positive sequence such that $\gamma_n \le \bar{\gamma}$ for all $n \in \mathbb{N}^*$, $\lim_{n\to+\infty}(n\gamma_n)^{-1} + \gamma_n = 0$. Then,*

$$\hat{\theta}_n^*(f) - \theta^*(f) \underset{n\to+\infty}{\overset{\mathbb{P}}{\Longrightarrow}} 0\,,$$

where $\hat{\theta}_n^*(f)$ and $\theta^*(f)$ are defined in (6.18) and (6.13), respectively, relatively to $(X_k^n)_{k\in\mathbb{N}}$.

Proof. For all $n \in \mathbb{N}^*$,

$$\hat{\theta}_n^*(f) - \theta^*(f) = \left(H_n^+ - H^{-1}\right) \hat{\pi}_n\left(\psi(f - \hat{\pi}_n(f))\right) + H^{-1}\left\{\hat{\pi}_n\left(\psi(f - \hat{\pi}_n(f))\right) - \pi\left(\psi(f - \pi(f))\right)\right\}\,.$$

By Lemma 6.16,

$$\hat{\pi}_n\left(\psi(f - \hat{\pi}_n(f))\right) - \pi\left(\psi(f - \pi(f))\right) \underset{n\to+\infty}{\overset{\mathbb{P}}{\Longrightarrow}} 0\,,$$

and it is enough to show that

$$H_n^+ - H^{-1} \underset{n\to+\infty}{\overset{\mathbb{P}}{\Longrightarrow}} 0\,,$$

to conclude the proof. Let $\epsilon > 0$ and consider the following decomposition:

$$\left\{\left\|H_n^+ - H^{-1}\right\| \ge \epsilon\right\} = \left\{\left\|H_n^+ - H^{-1}\right\| \ge \epsilon\right\} \cap \left\{\left\|H^{-1}\right\| \|H_n - H\| \le 1/2\right\} \cup \left\{\left\|H_n^+ - H^{-1}\right\| \ge \epsilon\right\} \cap \left\{\left\|H^{-1}\right\| \|H_n - H\| > 1/2\right\}\,,$$

where $\|\cdot\|$ denotes the operator norm. Since by Lemma 6.16,

$$H_n - H \underset{n\to+\infty}{\overset{\mathbb{P}}{\Longrightarrow}} 0\,,$$

we obtain

$$\mathbb{P}\left(\left\{\left\|H_n^+ - H^{-1}\right\| \ge \epsilon\right\} \cap \left\{\left\|H^{-1}\right\| \|H_n - H\| > 1/2\right\}\right) \le \mathbb{P}\left(\left\|H^{-1}\right\| \|H_n - H\| > 1/2\right) \underset{n\to+\infty}{\longrightarrow} 0\,.$$

By [Dou+18, Corollary 22.A.6], on the event $\{\|H^{-1}\| \|H_n - H\| \leq 1/2\}$,

$$\left\|H_n^+ - H^{-1}\right\| = \left\|H_n^{-1} - H^{-1}\right\| \leq \frac{\|H^{-1}\|^2 \|H_n - H\|}{1 - \|H^{-1}\| \|H_n - H\|} \leq 2 \left\|H^{-1}\right\|^2 \|H_n - H\| \ ,$$

and,

$$\mathbb{P}\left(\left\{\left\|H_n^+ - H^{-1}\right\| \geq \epsilon\right\} \cap \left\{\left\|H^{-1}\right\| \|H_n - H\| \leq 1/2\right\}\right) \leq \mathbb{P}\left(2 \left\|H^{-1}\right\|^2 \|H_n - H\| \geq \epsilon\right) \underset{n\to+\infty}{\longrightarrow} 0 \ ,$$

which gives the result. □

Proposition 6.18. *Let $V : \mathbb{R}^d \to [1, +\infty)$, $\bar{\gamma} > 0$. Assume **H**13(10), **A**3($V, \bar{\gamma}$), and **A**4($\bar{\gamma}, 6$). Let $\{(X_k^n)_{k\in\{0,\dots,n-1\}},\ n \in \mathbb{N}\}$ be defined in (6.79) and assume that $\xi(V) < +\infty$. Let $f \in \mathrm{C}^9_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}^*}$ be a positive sequence satisfying $\lim_{n\to+\infty}(n\gamma_n)^{-1} + \gamma_n = 0$ and $\hat{f}$ be a solution of the Poisson equation $\mathscr{L}\hat{f} = \pi(f) - f$. Then,*

(i) if $\pi(\mathscr{A}_0\hat{f}) \lim_{n\to+\infty} n^{1/2}\gamma_n^{\alpha-1/2} = 0$,

$$n^{-1/2}\gamma_n^{1/2} \sum_{k=0}^{n-1} \{f(X_k^n) - \pi(f)\} \overset{\mathbb{P}-\text{weakly}}{\underset{n\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma_\infty^2(f)) \ ,$$

(ii) if $\lim_{n\to+\infty} n^{1/2}\gamma_n^{\alpha-1/2} = \gamma_\infty \in [0, +\infty)$,

$$n^{-1/2}\gamma_n^{1/2} \sum_{k=0}^{n-1} \{f(X_k^n) - \pi(f)\} \overset{\mathbb{P}-\text{weakly}}{\underset{n\to+\infty}{\Longrightarrow}} \mathcal{N}(\gamma_\infty \pi(\mathscr{A}_0\hat{f}), \sigma_\infty^2(f)) \ ,$$

(iii) if $\pi(\mathscr{A}_0\hat{f}) \liminf_{n\to+\infty} n^{1/2}\gamma_n^{\alpha-1/2} = +\infty$,

$$\gamma_n^{1-\alpha} \sum_{k=0}^{n-1} \{f(X_k^n) - \pi(f)\} \overset{\mathbb{P}}{\underset{n\to+\infty}{\Longrightarrow}} \pi(\mathscr{A}_0\hat{f}) \quad ,$$

where $\sigma_\infty^2(f)$ is defined in (6.7).

Note that if the invariant distribution of R_γ is π for all $\gamma \in (0, \bar{\gamma}]$ (e.g. the case of MALA or RWM), we have under **A**4($\bar{\gamma}, 0$) and by the dominated convergence theorem, $\pi(\mathscr{A}_0\hat{f}) = 0$.

Proof. Let $f \in \mathrm{C}^9_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $p \in \mathbb{N}$ such that $\|f\|_{9,p} < +\infty$. By Proposition 6.1-(iii), there exists $\hat{f} \in \mathrm{C}^{10}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ such that $\mathscr{L}\hat{f} = -(f - \pi(f))$. By (6.61), we have for all $n \in \mathbb{N}^*$,

$$n^{-1/2}\gamma_n^{1/2} \sum_{k=0}^{n-1} \{f(X_k^n) - \pi(f)\} = \sum_{i=1}^{4} B_i^f(n) \ ,$$

where

$$B_1^f(n) = (n\gamma_n)^{-1/2} \left\{ \hat{f}(X_0^n) - \hat{f}(X_n^n) \right\} ,$$
$$B_2^f(n) = (n\gamma_n)^{-1/2} \sum_{k=0}^{n-1} \left\{ \hat{f}(X_{k+1}^n) - R_{\gamma_n} \hat{f}(X_k^n) \right\} ,$$
$$B_3^f(n) = n^{-1/2} \gamma_n^{1/2} \gamma_n^{\alpha-1} \sum_{k=0}^{n-1} \left\{ \mathscr{A}_{\gamma_n} \hat{f}(X_k^n) - \pi_{\gamma_n}(\mathscr{A}_{\gamma_n} \hat{f}) \right\} ,$$
$$B_4^f(n) = n^{1/2} \gamma_n^{1/2} \gamma_n^{\alpha-1} \pi_{\gamma_n}(\mathscr{A}_{\gamma_n} \hat{f}) .$$

We show in the sequel that $B_1^f(n)$ and $B_3^f(n)$ are remainder terms that converge in probability to 0 as $n \to +\infty$. $B_2^f(n)$ converges in law to a Gaussian random variable of mean 0 and variance $\sigma_\infty^2(f)$. $B_4^f(n)$ is the bias term.

By (6.29), $\lim_{n\to+\infty} \mathbb{E}\left[\left|B_1^f(n)\right|\right] = 0$ and then $B_1^f(n)$ converges in probability to 0 as $n \to +\infty$. By $\mathbf{A}4(\bar{\gamma}, 6)$, $\mathscr{A}_{\gamma_n} \hat{f} \in \mathrm{C}^6_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and there exists $q_1 \in \mathbb{N}$ such that for all $n \in \mathbb{N}^*$, $\|\mathscr{A}_{\gamma_n} \hat{f}\|_{6,q_1} \le C\|f\|_{9,p}$. We obtain by Theorem 6.4 and using that $\xi(V) < +\infty$,

$$\mathbb{E}\left[\left(B_3^f(n)\right)^2\right] \le C \left\|\mathscr{A}_{\gamma_n} \hat{f}\right\|_{6,p}^2 \gamma_n^{2(\alpha-1)} \left\{1 + \frac{1}{n\gamma_n}\right\} ,$$

and $\lim_{n\to+\infty} \mathbb{E}\left[\left(B_3^f(n)\right)^2\right] = 0$.

We now consider $B_2^f(n)$ for $n \in \mathbb{N}^*$. For $k \in \{0, \ldots, n-1\}$, denote by

$$\theta_{k+1,n} = (n\gamma_n)^{-1/2} \left\{ \hat{f}(X_{k+1}^n) - R_{\gamma_n} \hat{f}(X_k^n) \right\} .$$

By [HH14, Corollary 3.1, Chapter 3], $B_2^f(n)$ converges in law to a Gaussian random variable of mean 0 and variance $\sigma_\infty^2(f)$ if

$$\sum_{k=0}^{n-1} \mathbb{E}\left[\theta_{k+1,n}^2 \middle| \mathcal{F}_k\right] \underset{n\to+\infty}{\overset{\mathbb{P}}{\Longrightarrow}} \sigma_\infty^2(f) , \tag{6.80}$$
$$\sum_{k=0}^{n-1} \mathbb{E}\left[\theta_{k+1,n}^4 \middle| \mathcal{F}_k\right] \underset{n\to+\infty}{\overset{\mathbb{P}}{\Longrightarrow}} 0 . \tag{6.81}$$

Set $g_{\gamma_n}(x) = \mathbb{E}\left[\left(\hat{f}(X_1^n) - R_{\gamma_n} \hat{f}(x)\right)^2\right]$. By Lemma 6.13 with $k = 6$, $g_{\gamma_n} \in \mathrm{C}^6_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and there exists $q_2 \in \mathbb{N}$ such that for all $n \in \mathbb{N}^*$

$$\|g_{\gamma_n}\|_{6,q_2} \le C\gamma_n \|f\|_{9,p}^2 . \tag{6.82}$$

By Proposition 6.1-(iii), for all $n \in \mathbb{N}^*$, there exists $\hat{g}_{\gamma_n} \in \mathrm{C}^7_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ such that $\mathscr{L}\hat{g}_{\gamma_n} = -(g_{\gamma_n} - \pi_{\gamma_n}(g_{\gamma_n}))$. By the Markov property and (6.61), we have for all $n \in \mathbb{N}^*$

$$\sum_{k=0}^{n-1} \mathbb{E}\left[\theta_{k+1,n}^2 \middle| \mathcal{F}_k\right] = \frac{1}{n\gamma_n} \sum_{k=0}^{n-1} g_{\gamma_n}(X_k^n) = B_{21}^f(n) + B_{22}^f(n) + B_{23}^f(n) + B_{24}^f(n) ,$$

where

$$B_{21}^f(n) = \gamma_n^{-1}\pi_{\gamma_n}(g_{\gamma_n}) \ ,$$
$$B_{22}^f(n) = (n\gamma_n^2)^{-1}\left\{\hat{g}_{\gamma_n}(X_0^n) - \hat{g}_{\gamma_n}(X_n^n)\right\} \ ,$$
$$B_{23}^f(n) = (n\gamma_n^2)^{-1}\sum_{k=0}^{n-1}\left\{\hat{g}_{\gamma_n}(X_{k+1}^n) - R_{\gamma_n}\hat{g}_{\gamma_n}(X_k^n)\right\} \ ,$$
$$B_{24}^f(n) = \frac{1}{n\gamma_n}\gamma_n^{\alpha-1}\sum_{k=0}^{n-1}\left\{\mathscr{A}_{\gamma_n}\hat{g}_{\gamma_n}(X_k^n) - \gamma_n^{1-\alpha}\left(\pi_{\gamma_n}(\hat{g}_{\gamma_n}) - \pi(\hat{g}_{\gamma_n})\right)\right\} \ .$$

By (6.29),

$$\mathbb{E}\left[\left|B_{22}^f(n)\right|\right] \le C(n\gamma_n^2)^{-1}\left\|g_{\gamma_n}\right\|_{6,q_2} \le C(n\gamma_n)^{-1}\left\|f\right\|_{9,p}^2 \ ,$$

$\lim_{n\to+\infty}\mathbb{E}\left[\left|B_{22}^f(n)\right|\right] = 0$, and $B_{22}^f(n)$ converges in probability to 0 as $n \to +\infty$. By (6.29), (6.82), the Markov property and Lemma 6.13 with $k = 0$, we get for all $n \in \mathbb{N}^*$

$$\mathbb{E}\left[\left(B_{23}^f(n)\right)^2\right] = \frac{1}{n\gamma_n^4}\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(\hat{g}_{\gamma_n}(X_{k+1}^n) - R_{\gamma_n}\hat{g}_{\gamma_n}(X_k^n)\right)^2\right]$$
$$\le C\frac{1}{n\gamma_n^4}\gamma_n\left\|g_{\gamma_n}\right\|_{6,q_2}^2 \le C\frac{1}{n\gamma_n}\left\|f\right\|_{9,p}^4 \ ,$$

and $\lim_{n\to+\infty}\mathbb{E}\left[\left(B_{23}^f(n)\right)^2\right] = 0$. We can decompose $B_{24}^f(n)$ as,

$$\mathbb{E}\left[\left(B_{24}^f(n)\right)^2\right] = \frac{\gamma_n^{2(\alpha-1)}}{n\gamma_n^2}\mathbb{E}\left[\frac{1}{n}\left(\sum_{k=0}^{n-1}\left\{\mathscr{A}_{\gamma_n}\hat{g}_{\gamma_n}(X_k^n) - \gamma_n^{1-\alpha}\left(\pi_{\gamma_n}(\hat{g}_{\gamma_n}) - \pi(\hat{g}_{\gamma_n})\right)\right\}\right)^2\right]$$
$$= \frac{1}{n\gamma_n^3}B_{241}^f(n) \tag{6.83}$$

where,

$$B_{241}^f(n) = \gamma_n^{2(\alpha-1)}\gamma_n\mathbb{E}\left[\frac{1}{n}\left(\sum_{k=0}^{n-1}\left\{\mathscr{A}_{\gamma_n}\hat{g}_{\gamma_n}(X_k^n) - \gamma_n^{1-\alpha}\left(\pi_{\gamma_n}(\hat{g}_{\gamma_n}) - \pi(\hat{g}_{\gamma_n})\right)\right\}\right)^2\right] \ .$$

By $\mathbf{A}\,4(\bar{\gamma}, 6)$ and Proposition 6.1-(iii), there exists $q_3 \in \mathbb{N}$ such that for all $n \in \mathbb{N}^*$, $\|\mathscr{A}_{\gamma_n}\hat{g}_{\gamma_n}\|_{3,q_3} \le C\|g_{\gamma_n}\|_{6,q_2}$. By Lemma 6.14, (6.59), (6.82) and using that $\xi(V) < +\infty$,

$$B_{241}^f(n) \le C\gamma_n\gamma_n^{2(\alpha-2)}\left\|\mathscr{A}_{\gamma_n}\hat{g}_{\gamma_n}\right\|_{3,q_3}^2\left\{1 + 1/(n\gamma_n)\right\} \le C\left\|g_{\gamma_n}\right\|_{6,q_2}^2\left\{1 + 1/(n\gamma_n)\right\}$$
$$\le C\gamma_n^2\left\|f\right\|_{9,p}^4\left\{1 + 1/(n\gamma_n)\right\} \ .$$

Combining it with (6.83), we obtain $\lim_{n\to+\infty}\mathbb{E}\left[\left(B_{24}^f(n)\right)^2\right] = 0$. For $B_{21}^f(n)$, we have by (6.72) and (6.82) for all $n \in \mathbb{N}^*$,

$$\left|\gamma_n^{-1}\pi_{\gamma_n}(g_{\gamma_n}) - \sigma_\infty^2(f)\right| \le C\left\|g_{\gamma_n}\right\|_{3,q_2}^2\gamma_n^{\alpha\wedge 2}\gamma_n^{-1} \le C\left\|f\right\|_{9,p}^4\gamma_n\gamma_n^{\alpha\wedge 2} \ ,$$

and $\lim_{n\to+\infty} \gamma_n^{-1}\pi_{\gamma_n}(g_{\gamma_n}) = \sigma_\infty^2(f)$. This gives (6.80). For (6.81), we have for $k \in \{0,\dots,n-1\}$

$$\mathbb{E}\left[\theta_{k+1,n}^4 \middle| \mathcal{F}_k\right] = (n\gamma_n)^{-2} h_{\gamma_n}(X_k^n)$$

where $h_{\gamma_n}(x) = \mathbb{E}\left[\left\{\hat{f}(X_1^n) - R_{\gamma_n}\hat{f}(x)\right\}^4\right]$. By Cauchy-Schwarz inequality and Lemma 6.13, there exists $q_4 \in \mathbb{N}$ such that for all $n \in \mathbb{N}^*$, $\|h_{\gamma_n}\|_{0,q_4} \le C\gamma_n^2 \|f\|_{3,p}^4$. By (6.29), we obtain for all $n \in \mathbb{N}^*$

$$\mathbb{E}\left[\left|\sum_{k=0}^{n-1} \mathbb{E}\left[\theta_{k+1,n}^4 \middle| \mathcal{F}_k\right]\right|\right] \le C n^{-1} \|f\|_{3,p}^4$$

and (6.81) is satisfied.

For $B_4^f(n)$, we only have to show that $\lim_{n\to+\infty} \pi_{\gamma_n}(\mathscr{A}_{\gamma_n}\hat{f}) = \pi(\mathscr{A}_0\hat{f})$. Using Proposition 6.3, we have for all $n \in \mathbb{N}^*$,

$$\left|\pi_{\gamma_n}(\mathscr{A}_{\gamma_n}\hat{f}) - \pi(\mathscr{A}_{\gamma_n}\hat{f})\right| \le C \left\|\mathscr{A}_{\gamma_n}\hat{f}\right\|_{3,q_1} \gamma_n^{\alpha-1} \le C \|f\|_{9,p}\, \gamma_n^{\alpha-1} .$$

Combining it with $\mathbf{A}\,4(\bar{\gamma},6)$ and the dominated convergence theorem, we obtain the result, which concludes the proof. □

Proof of Theorem 6.5. We consider the case $\lim_{n\to+\infty} n^{1/2}\gamma_n^{\alpha-1/2} = \gamma_\infty \in [0,+\infty)$, and we denote by $\mu_f^{\mathrm{CV}} = \gamma_\infty \pi(\mathscr{A}_0(\hat{f} - \theta^*(f)^{\mathrm{T}}\psi) \in [0,+\infty)$. The case

$$\pi(\mathscr{A}_0(\hat{f} - \theta^*(f)^{\mathrm{T}}\psi)) \liminf_{n\to+\infty} n^{1/2}\gamma_n^{\alpha-1/2} = +\infty$$

can be handled in a similar way. Denote $W_n \in \mathbb{R}^{p+1}$ for $n \in \mathbb{N}^*$ by

$$W_n = n^{1/2}\gamma_n^{1/2}\left(\hat{\pi}_n(f) - \pi(f), \hat{\pi}_n(\mathscr{L}\psi)\right) .$$

By Proposition 6.18, $((1,\theta)^{\mathrm{T}}W_n)_{n\in\mathbb{N}^*}$ converges $\mathbb{P}$-weakly, for every $\theta \in \mathbb{R}^p$, to a one-dimensional Gaussian variable of mean μ_f^{CV} and variance $\sigma_\infty^2(f + \mathscr{L}(\theta^{\mathrm{T}}\psi))$. By the Cramér-Wold theorem, $(W_n)_{n\in\mathbb{N}^*}$ converges $\mathbb{P}$-weakly to a $(p+1)$-dimensional Gaussian vector W of mean $\gamma_\infty\left(\pi(\mathscr{A}_0\hat{f}), -\pi(\mathscr{A}_0\psi)\right)$ and covariance matrix

$$2\pi\left((\hat{f},-\psi)(-\mathscr{L})(\hat{f},-\psi)^{\mathrm{T}}\right) .$$

By Lemma 6.17 and Slutsky's theorem, $(\hat{\theta}_n^*(f), W_n)_{n\in\mathbb{N}^*}$ converges $\mathbb{P}$-weakly to $(\theta^*(f), W)$, which concludes the proof. □

6.C Additional proofs

6.C.1 Proof of Proposition 6.1

(i) By **H** 13(2), [RT96, Theorems 2.1], π is the unique stationary distribution of the semigroup $(P_t)_{t\geq 0}$ associated to (6.6). In addition, by [RT96, Theorems 2.1] and [Bak+08, Corollary 1.6], $(P_t)_{t\geq 0}$ is V-uniformly geometrically ergodic w.r.t. π with $V(x) = \exp\{(\upsilon/4)(1 + \|x - x^\star\|^2)^{1/2}\}$.

(ii) is given by [GM96, Theorem 4.4] using that $(P_t)_{t\geq 0}$ is V-uniformly geometrically ergodic; see also see [Bha82] and [CCG12].

(iii) follows from [PV01, Theorem 1].

(iv) Let $f, g \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $M > 0$. We split $\pi(f(-\mathscr{L})g)$ into $I_1 + I_2$ where

$$I_1 = \int_{[-M,M]^d} f(x)(-\mathscr{L})g(x)\pi(\mathrm{d}x) \;, \quad I_2 = \int_{([-M,M]^d)^{\mathrm{c}}} f(x)(-\mathscr{L})g(x)\pi(\mathrm{d}x) \;.$$

By the dominated convergence theorem, $\lim_{M\to+\infty} I_2 = 0$. For all $i \in \{1, \ldots, d\}$, $a \in \mathbb{R}$ and $x \in \mathbb{R}^d$, denote by $x^a_{-i} = (x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_d)$ and by $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. By integrations by parts,

$$\begin{aligned}
I_1 &= \int_{[-M,M]^d} f(x) \langle \nabla U(x), \nabla g(x) \rangle \, \pi(\mathrm{d}x) + \int_{[-M,M]^d} f(x)(-\Delta g(x))\pi(\mathrm{d}x) \\
&= \int_{[-M,M]^d} f(x) \langle \nabla U(x), \nabla g(x) \rangle \, \pi(\mathrm{d}x) \\
&+ \sum_{i=1}^{d} \int_{[-M,M]^{d-1}} \left\{ f(x^{-M}_{-i}) \frac{\partial g}{\partial x_i}(x^{-M}_{-i})\pi(x^{-M}_{-i}) - f(x^{M}_{-i}) \frac{\partial g}{\partial x_i}(x^{M}_{-i})\pi(x^{M}_{-i}) \right\} \mathrm{d}x_{-i} \\
&+ \int_{[-M,M]^d} \{ \langle \nabla f(x), \nabla g(x) \rangle - f(x) \langle \nabla U(x), \nabla g(x) \rangle \} \, \pi(\mathrm{d}x)
\end{aligned}$$

and $\lim_{M\to+\infty} I_1 = \pi(\langle \nabla f, \nabla g \rangle)$ which concludes the proof.

6.C.2 Proof of Lemma 6.6

By [DM16, Theorem 32], we have for $x, y \in \mathbb{R}^d$ and $\gamma > 0$

$$\|\delta_x (R^{\mathrm{ULA}}_\gamma)^{\lceil 1/\gamma \rceil} - \delta_y (R^{\mathrm{ULA}}_\gamma)^{\lceil 1/\gamma \rceil}\|_{\mathrm{TV}} \leq 1 - 2\Phi\left(- \frac{\|x - y\|}{2\Xi^{1/2}_{\lceil 1/\gamma \rceil}} \right)$$

where

$$\Xi_{\lceil 1/\gamma \rceil} = \sum_{i=1}^{\lceil 1/\gamma \rceil} (2\gamma) \prod_{j=1}^{i} (1 + \gamma L)^{-2} = \frac{1 - \exp\left(-2 \lceil 1/\gamma \rceil \ln(1 + \gamma L)\right)}{L + (\gamma L^2)/2} \;,$$

which gives (i). The assertion (ii) follows from [DM17, Proposition 8] and [DM17, Proposition 13].

6.C.3 Proof of Lemma 6.8

We have for all $x \in \mathbb{R}^d$

$$\nabla U_{\log}(x) = -\mathsf{X}^{\mathrm{T}}\mathsf{Y} + \sum_{i=1}^{N} \mathsf{X}_i/(1+\mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x}) + x/\varsigma^2 \;,$$
$$\mathrm{D}^2 U_{\log}(x) = \sum_{i=1}^{N} \frac{\mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x}}{\left(1+\mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x}\right)^2} \mathsf{X}_i\mathsf{X}_i^{\mathrm{T}} + \mathrm{Id}\,/\varsigma^2 \;,$$
$$\mathrm{D}^3 U_{\log}(x) = \sum_{i=1}^{N} \frac{\mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x}}{\left(1+\mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x}\right)^2} \left\{2\frac{\mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x}}{1+\mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x}} - 1\right\} \mathsf{X}_i^{\otimes 3} \;.$$

Using for all $i \in \{1,\dots,N\}$ and $x \in \mathbb{R}^d$ that $0 < \mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x}/(1+\mathrm{e}^{-\mathsf{X}_i^{\mathrm{T}}x})^2 \leq 1/4$, $U_{\log}$ is strongly convex, gradient Lipschitz and satisfies **H**14, **H**16, **H**13(k) for all $k \in \mathbb{N}^*$, **H**17 and **H**18.

For U_{pro}, define $h : \mathbb{R} \to \mathbb{R}_-$ for all $t \in \mathbb{R}$ by $h(t) = \ln(\Phi(t))$. We have for all $t \in \mathbb{R}$,

$$h'(t) = \frac{\Phi'(t)}{\Phi(t)} \quad , \quad h''(t) = -\frac{\Phi'(t)}{\Phi(t)}\left\{t + \frac{\Phi'(t)}{\Phi(t)}\right\} \;,$$
$$h^{(3)}(t) = \frac{\Phi'(t)}{\Phi(t)}\left\{2\left(\frac{\Phi'(t)}{\Phi(t)}\right)^2 + 3t\frac{\Phi'(t)}{\Phi(t)} + t^2 - 1\right\}$$

and for all $x \in \mathbb{R}^d$

$$\nabla U_{\mathrm{pro}}(x) = \sum_{i=1}^{N} \left\{(1-\mathsf{Y}_i)h'(-\mathsf{X}_i^{\mathrm{T}}x) - \mathsf{Y}_i h'(\mathsf{X}_i^{\mathrm{T}}x)\right\}\mathsf{X}_i + x/\varsigma^2 \;,$$
$$\mathrm{D}^2 U_{\mathrm{pro}}(x) = \sum_{i=1}^{N} \left\{-(1-\mathsf{Y}_i)h''(-\mathsf{X}_i^{\mathrm{T}}x) - \mathsf{Y}_i h''(\mathsf{X}_i^{\mathrm{T}}x)\right\}\mathsf{X}_i\mathsf{X}_i^{\mathrm{T}} + \mathrm{Id}\,/\varsigma^2 \;,$$
$$\mathrm{D}^3 U_{\mathrm{pro}}(x) = \sum_{i=1}^{N} \left\{(1-\mathsf{Y}_i)h^{(3)}(-\mathsf{X}_i^{\mathrm{T}}x) - \mathsf{Y}_i h^{(3)}(\mathsf{X}_i^{\mathrm{T}}x)\right\}\mathsf{X}_i^{\otimes 3} \;.$$

By an integration by parts, we have for all $t < 0$

$$t + \frac{\Phi'(t)}{\Phi(t)} = -\frac{t}{\Phi(t)}\int_{-\infty}^{t} \frac{\mathrm{e}^{-s^2/2}}{\sqrt{2\pi s^2}}\mathrm{d}s$$

and $t + \Phi'(t)/\Phi(t) \geq 0$ for all $t \in \mathbb{R}$. Let $t < 0$ and $s = -t > 0$. We have $\Phi(t) = \bar{\Phi}(s) = \mathrm{erfc}(s/\sqrt{2})/2$ where $\mathrm{erfc} : \mathbb{R} \to \mathbb{R}_+$ is the complementary error function defined for all $u \in \mathbb{R}$ by $\mathrm{erfc}(u) = (2/\sqrt{\pi})\int_u^{+\infty} \mathrm{e}^{-v^2}\mathrm{d}v$. By [GR14, Section 8.25, formula 8.254], we have the following asymptotic expansion for $s \to +\infty$

$$\bar{\Phi}(s) = \frac{\mathrm{e}^{-s^2/2}}{\sqrt{2\pi}s}\left(1 - s^{-2} + 3s^{-4} + O(s^{-6})\right) \;.$$

Using that $\Phi'(t) = (2\pi)^{-1/2}\mathrm{e}^{-t^2/2}$ for all $t \in \mathbb{R}$, we get asymptotically for $t \to -\infty$ and $s = -t \to +\infty$,

$$\Phi'(t)/\Phi(t) = s\left(1 + s^{-2} - 2s^{-4} + O(s^{-6})\right) \tag{6.84}$$

and $\lim_{t\to-\infty} h''(t) = -1$. There exists then $C > 0$ such that for all $t \in \mathbb{R}$, $-C \leq h''(t) \leq 0$. U_{pro} is then strongly convex, gradient Lipschitz and satisfies **H**14, **H**16, **H**13(k) for all $k \in \mathbb{N}$ and **H**17. By (6.84), we have for $t \to -\infty$ and $s = -t \to +\infty$, $h^{(3)}(t) = O(s^{-1})$. U_{pro} satisfies then **H**18.

6.C.4 Proof of Lemma 6.11

The proof is adapted from [FHS15, Lemma 1]. Let $i \in \{0, \dots, 6\}$, $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\bar{\gamma} > 0$, $\gamma \in [0, \bar{\gamma}]$ and $x, y \in \mathbb{R}^d$. Note that $\tau_\gamma^{\mathrm{MALA}}(x, y)$ defined in (6.43) may be expressed as

$$\begin{aligned}\tau_\gamma^{\mathrm{MALA}}(x, y) = U(y) - U(x) - (1/2)\left\langle y - x, \nabla U(x) + \nabla U(y)\right\rangle \\ + (\gamma/4)\left\{\|\nabla U(y)\|^2 - \|\nabla U(x)\|^2\right\} .\end{aligned} \tag{6.85}$$

A Taylor expansion of U and ∇U around x yields

$$\begin{aligned}U(y) - U(x) = \langle \nabla U(x), y - x\rangle + (1/2)\,\mathrm{D}^2 U(x)[(y-x)^{\otimes 2}] + (1/6)\,\mathrm{D}^3 U(x)[(y-x)^{\otimes 3}] \\ + (1/6)\int_0^1 (1-t)^3\,\mathrm{D}^4 U((1-t)x + ty)[(y-x)^{\otimes 4}]\mathrm{d}t ,\end{aligned} \tag{6.86}$$

$$\begin{aligned}\nabla U(y) = \nabla U(x) + \mathrm{D}^2 U(x)[y - x] + (1/2)\,\mathrm{D}^3 U(x)[(y-x)^{\otimes 2}] \\ + (1/2)\int_0^1 (1-t)^2\,\mathrm{D}^4 U((1-t)x + ty)[(y-x)^{\otimes 3}]\mathrm{d}t .\end{aligned} \tag{6.87}$$

Substituting (6.86) and (6.87) into (6.85), we obtain for $z \in \mathbb{R}^d$, $\tau_\gamma^{\mathrm{MALA}}(x, x - \gamma\nabla U(x) + \sqrt{2\gamma}z) = \gamma^{3/2}\xi_\gamma(x, z)$ where ξ_γ is defined for all $x, z \in \mathbb{R}^d$ and $\gamma \in [0, \bar{\gamma}]$ by

$$\begin{aligned}\xi_\gamma(x, z) &= -(1/12)\,\mathrm{D}^3 U(x)[(-\sqrt{\gamma}\nabla U(x) + \sqrt{2}z)^{\otimes 3}] \\ &- (\sqrt{\gamma}/12)\int_0^1 (1-t)^2(1+2t)\,\mathrm{D}^4 U(x - t\gamma\nabla U(x) + t\sqrt{2\gamma}z)[(-\sqrt{\gamma}\nabla U(x) + \sqrt{2}z)^{\otimes 4}]\mathrm{d}t \\ &+ (1/2)\left\langle \nabla U(x), \int_0^1 \mathrm{D}^2 U(x - t\gamma\nabla U(x) + t\sqrt{2\gamma}z)[-\sqrt{\gamma}\nabla U(x) + \sqrt{2}z]\mathrm{d}t\right\rangle \\ &+ (\sqrt{\gamma}/4)\left\|\int_0^1 \mathrm{D}^2 U(x - t\gamma\nabla U(x) + t\sqrt{2\gamma}z)[-\sqrt{\gamma}\nabla U(x) + \sqrt{2}z]\mathrm{d}t\right\|^2 .\end{aligned}$$

Note that by the dominated convergence theorem, for all $x, z \in \mathbb{R}^d$, $\lim_{\gamma\to 0}\xi_\gamma(x, z) = \xi_0(x, z)$ where ξ_0 is defined in (6.45). By (6.42), we get

$$\begin{aligned}R_\gamma^{\mathrm{MALA}}\varphi(x) - \varphi(x) = \mathbb{E}\left[\varphi(x - \gamma\nabla U(x) + \sqrt{2\gamma}Z) - \varphi(x)\right] \\ + \mathbb{E}\left[\left(\mathrm{e}^{-\gamma^{3/2}\xi_\gamma(x,Z)_+} - 1\right)\left\{\varphi(x - \gamma\nabla U(x) + \sqrt{2\gamma}Z) - \varphi(x)\right\}\right] ,\end{aligned} \tag{6.88}$$

where Z is an i.i.d. standard d-dimensional Gaussian variable. Combining (6.88) with the Taylor expansion (6.33), we get (i) with $\mathscr{A}_\gamma^{\mathrm{MALA}} : \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}) \to \mathrm{C}^{i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ given for all $\varphi \in \mathrm{C}^{4+i}_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $x \in \mathbb{R}^d$ and $\gamma \in (0, \bar{\gamma}]$ by

$$\mathscr{A}_\gamma^{\mathrm{MALA}}\varphi(x) = \mathscr{A}_\gamma^{\mathrm{ULA}}\varphi(x) + \mathbb{E}\Big[\gamma^{-3/2}\left\{1 - \mathrm{e}^{-\gamma^{3/2}\max(0,\xi_\gamma(x,Z))}\right\} \times \left\{\int_0^1 \left\langle \nabla\varphi(x - t\gamma\nabla U(x) + t\sqrt{2\gamma}Z), \sqrt{\gamma}\nabla U(x) - \sqrt{2}Z\right\rangle \mathrm{d}t\right\}\Big] \quad (6.89)$$

and $\mathscr{A}_\gamma^{\mathrm{ULA}}$ given in (6.34). The assertion (ii) follows from taking the limit $\gamma \downarrow 0^+$ in (6.89) and the dominated convergence theorem.

6.D Numerical experiments - additional results

We provide additional plots for the logistic regression, see Figure 6.2 and Figure 6.3, and the results for the Bayesian probit regression presented in Section 6.4, see Table 6.2, Figure 6.4 and Figure 6.5. The parameters are set to the same values as for the Bayesian logistic regression. The results are similar to the results obtained for the Bayesian logistic regression.

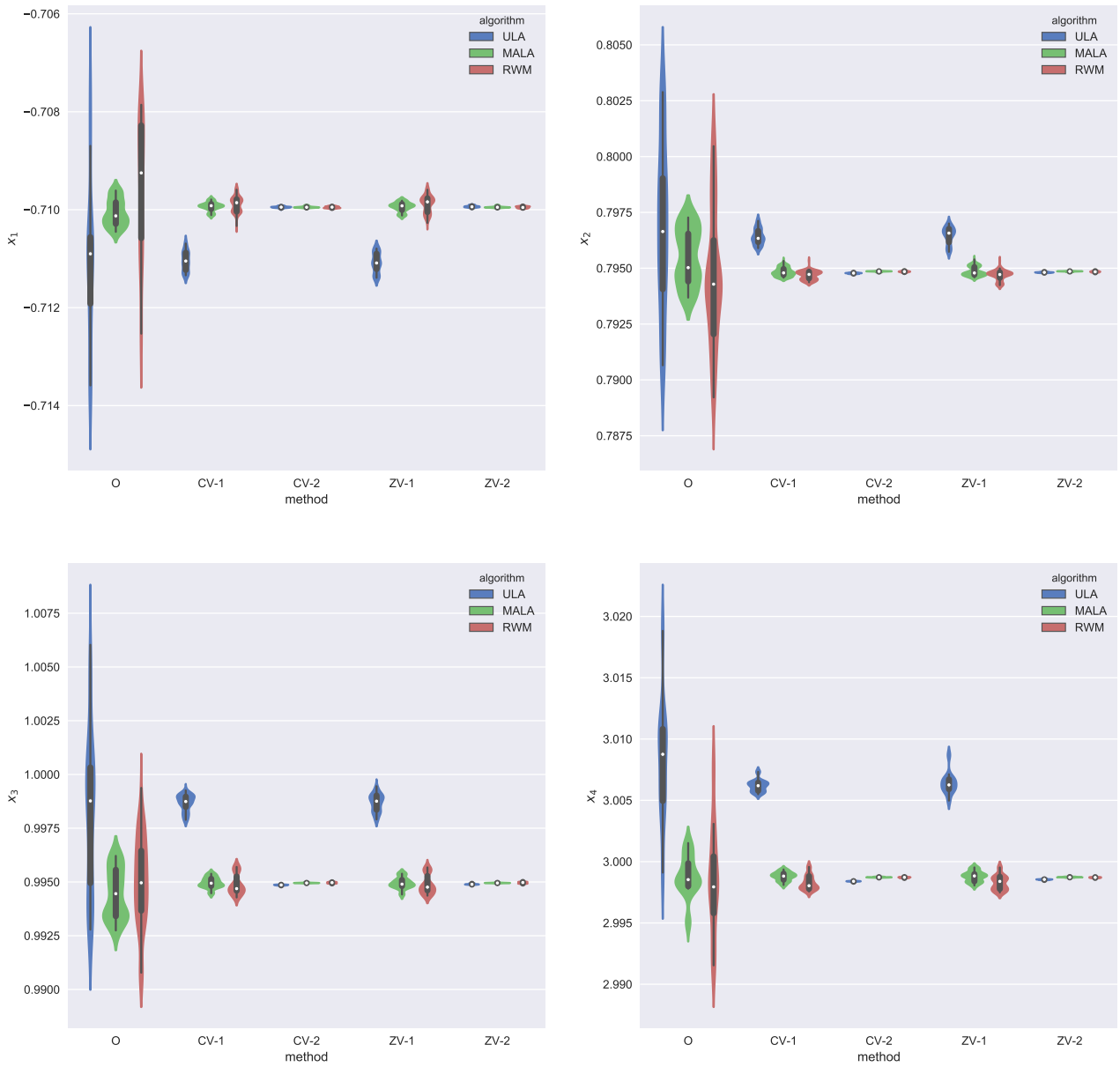

Figure 6.2: Boxplots of x_1, x_2, x_3, x_4 using the ULA, MALA and RWM algorithms for the logistic regression. The compared estimators are the ordinary empirical average (O), our estimator with a control variate (6.17) using first (CV-1) or second (CV-2) order polynomials for ψ, and the zero-variance estimator of [PMG14] using a first (ZV-1) or second (ZV-2) order polynomial basis.

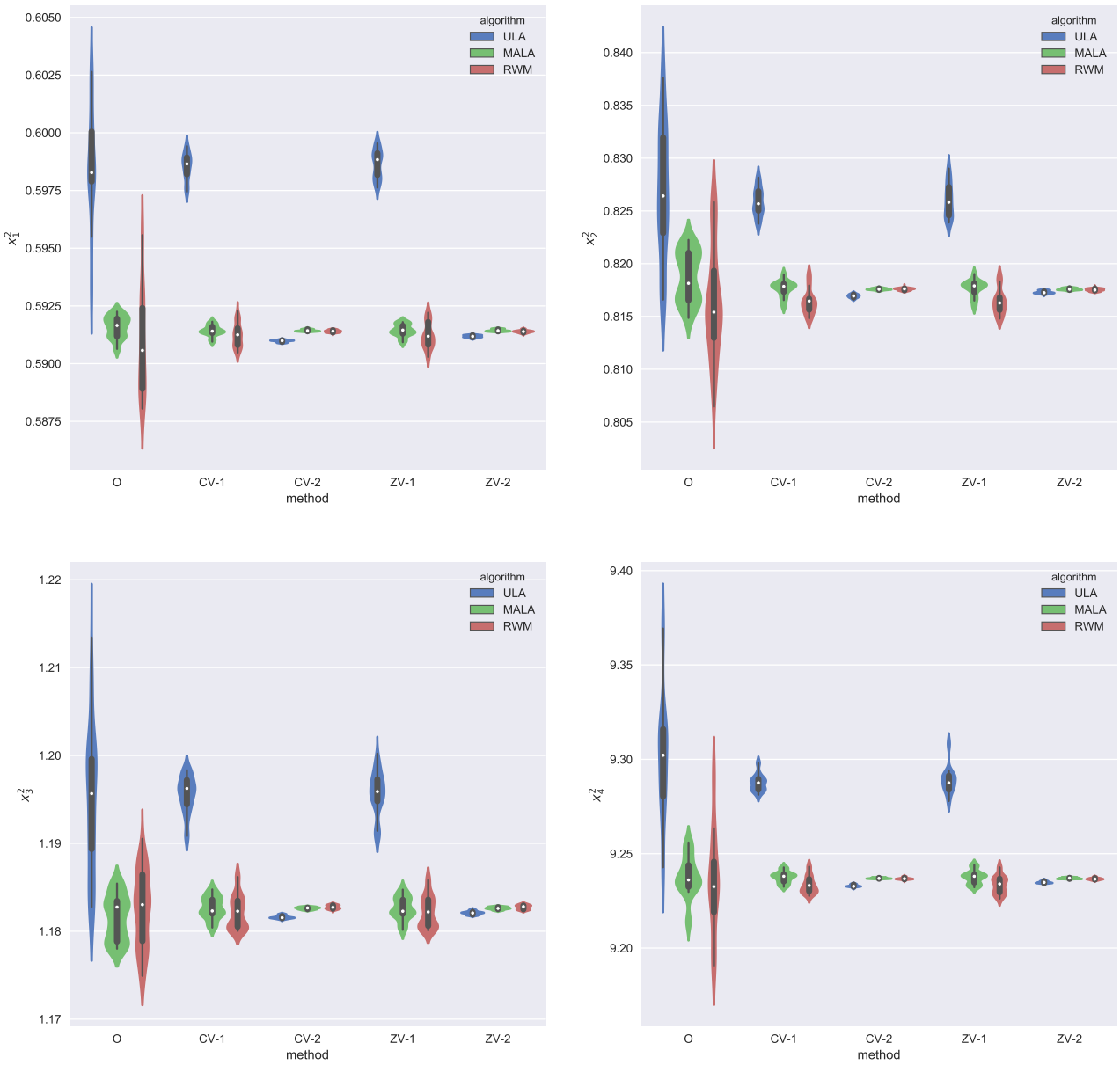

Figure 6.3: Boxplots of $x_1^2, x_2^2, x_3^2, x_4^2$ using the ULA, MALA and RWM algorithms for the logistic regression. The compared estimators are the ordinary empirical average (O), our estimator with a control variate (6.17) using first (CV-1) or second (CV-2) order polynomials for ψ, and the zero-variance estimator of [PMG14] using a first (ZV-1) or second (ZV-2) order polynomial basis.

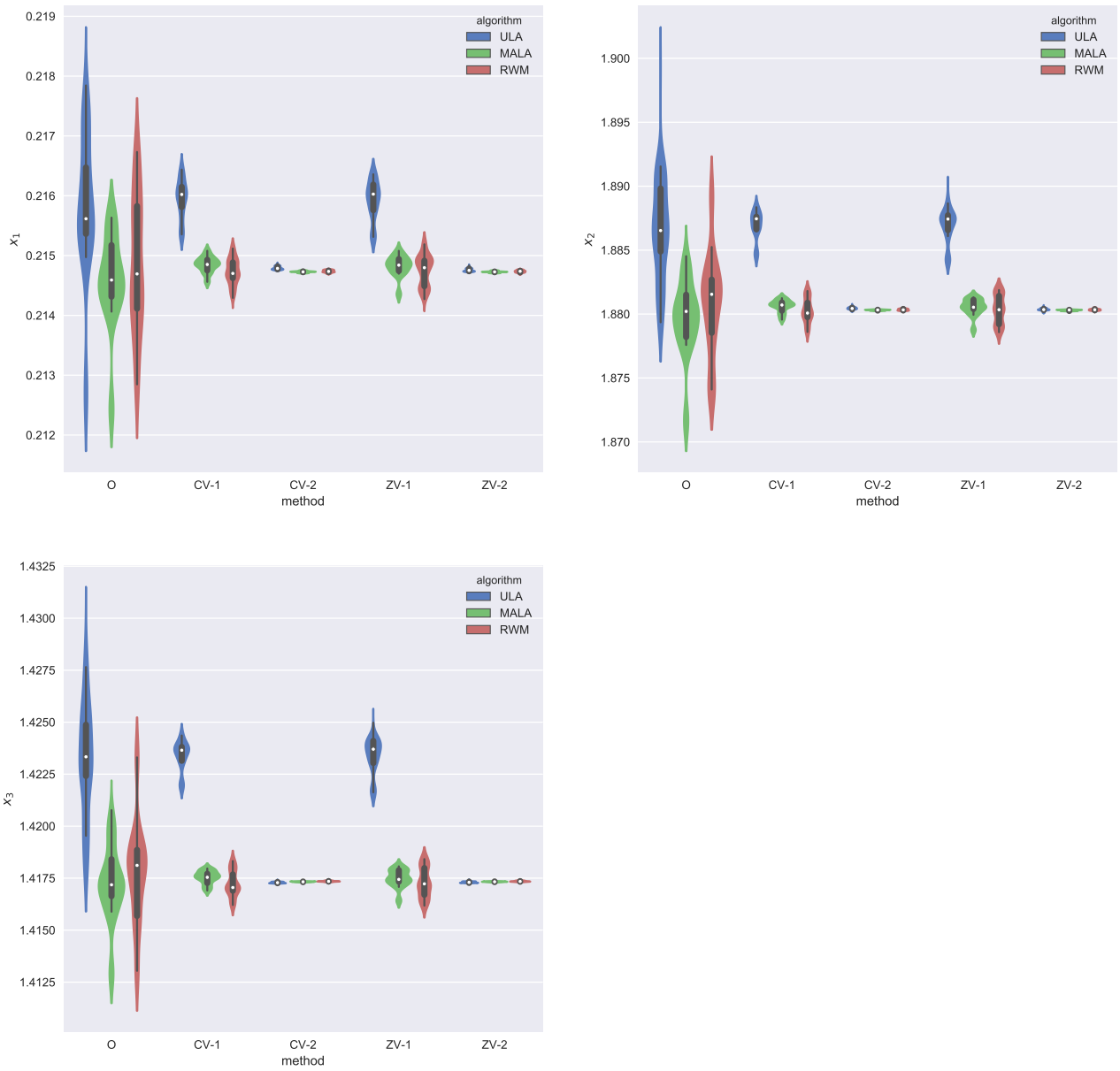

Figure 6.4: Boxplots of x_1, x_2, x_3 using the ULA, MALA and RWM algorithms for the probit regression. The compared estimators are the ordinary empirical average (O), our estimator with a control variate (6.17) using first (CV-1) or second (CV-2) order polynomials for ψ, and the zero-variance estimator of [PMG14] using a first (ZV-1) or second (ZV-2) order polynomial basis.

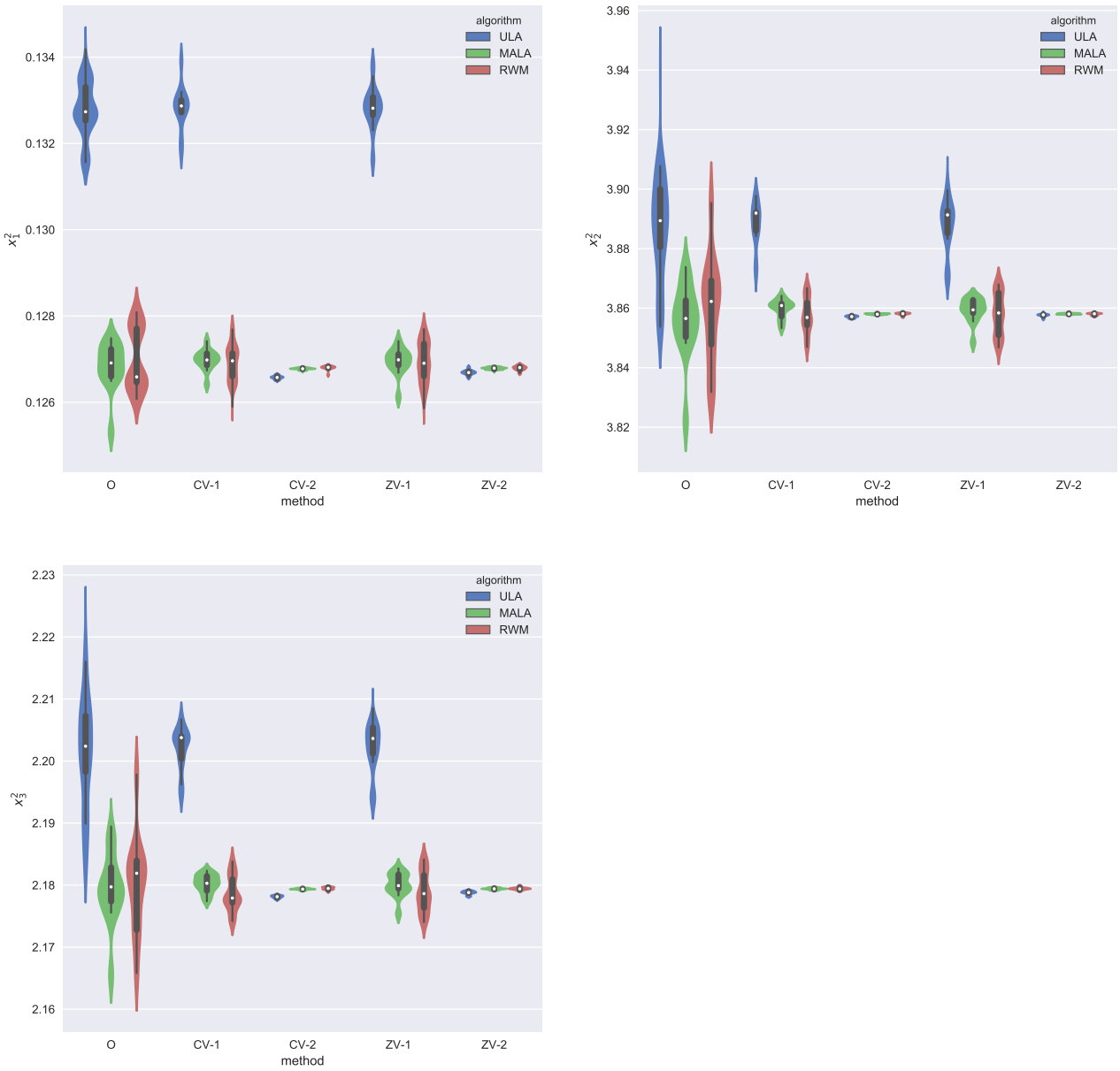

Figure 6.5: Boxplots of x_1^2, x_2^2, x_3^2 using the ULA, MALA and RWM algorithms for the probit regression. The compared estimators are the ordinary empirical average (O), our estimator with a control variate (6.17) using first (CV-1) or second (CV-2) order polynomials for ψ, and the zero-variance estimator of [PMG14] using a first (ZV-1) or second (ZV-2) order polynomial basis.

Table 6.2: Estimates of the asymptotic variances for ULA, MALA and RWM and each parameter x_i, x_i^2 for $i \in \{1, \dots, d\}$, and of the variance reduction factor (VRF) on the example of the probit regression.

		MCMC	CV-1-MCMC		CV-2-MCMC		ZV-1-MCMC		ZV-2-MCMC	
		Variance	VRF	Variance	VRF	Variance	VRF	Variance	VRF	Variance
x_1	ULA	2.1	24	0.089	2.9e+03	0.00073	20	0.11	2.7e+03	0.00078
	MALA	0.41	22	0.019	2.7e+03	0.00015	18	0.023	2.6e+03	0.00016
	RWM	1.2	23	0.05	2.2e+03	0.00054	21	0.056	2.2e+03	0.00053
x_2	ULA	27	24	1.1	2.8e+03	0.0099	18	1.5	2.4e+03	0.011
	MALA	6.4	24	0.27	2.9e+03	0.0022	19	0.34	2.6e+03	0.0025
	RWM	13	18	0.72	1.8e+03	0.0073	16	0.81	1.8e+03	0.0075
x_3	ULA	11	24	0.47	6.7e+03	0.0017	18	0.62	6.3e+03	0.0018
	MALA	2.6	23	0.11	7e+03	0.00037	18	0.14	6.8e+03	0.00038
	RWM	5.5	18	0.3	4.3e+03	0.0013	16	0.34	4.3e+03	0.0013
x_1^2	ULA	0.75	3.5	0.22	1.6e+02	0.0048	2.8	0.26	1.3e+02	0.0057
	MALA	0.15	3.5	0.043	1.5e+02	0.001	2.8	0.053	1.3e+02	0.0011
	RWM	0.43	2.6	0.16	1.2e+02	0.0035	2.4	0.18	1.2e+02	0.0037
x_2^2	ULA	4.7e+02	9.3	51	1.4e+03	0.33	7.5	63	1.2e+03	0.4
	MALA	1.1e+02	9.1	12	1.5e+03	0.073	7.6	14	1.3e+03	0.085
	RWM	2.2e+02	7.7	29	1e+03	0.22	6.9	33	9.8e+02	0.23
x_3^2	ULA	1.1e+02	9.8	11	9.7e+02	0.11	7.9	14	7.9e+02	0.14
	MALA	24	9.7	2.5	9.8e+02	0.025	8.1	3	8.5e+02	0.029
	RWM	52	7.9	6.7	6.1e+02	0.086	7.1	7.4	5.9e+02	0.088

6.E Additional proofs on the diffusion approximation of RWM

In this Section, we give the proofs of Lemma 6.9 and Lemma 6.10. These results deal with the diffusion approximation for the Random Walk Metropolis algorithm.

In the sequel, C is a positive constant which can change from line to line but does not depend on γ. For $M \in \mathbb{R}^{d\times d}$, denote by $\|M\|_{\mathrm{F}}$ the Frobenius norm of M. For a set $A \subset \mathbb{R}^d$, define by $A^{\mathrm{c}} = \mathbb{R}^d \setminus A$. For all $x \in \mathbb{R}^{\tilde{d}}$ and $M > 0$, we denote by $\mathrm{B}_{\tilde{d}}(x, M)$ (respectively $\overline{\mathrm{B}}_{\tilde{d}}(x, M)$), the open (respectively close) ball centered at x of radius M. When the dimension d of the state space $\mathbb{R}^d$ is unambiguous, they are respectively denoted by $\mathrm{B}(x, M)$ and $\overline{B}(x, M)$.

6.E.1 Proof of Lemma 6.9

The proof is adapted from [FHS15, Lemma 1]. Let $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $\bar{\gamma} > 0$, $\gamma \in [0, \bar{\gamma}]$, $x \in \mathbb{R}^d$ and Z be an i.i.d. standard d-dimensional Gaussian variable. By a Taylor expansion, we obtain

$$\tau^{\mathrm{RWM}}(x, x+\sqrt{2\gamma}Z) = \sqrt{2\gamma}\,\langle \nabla U(x), Z\rangle + (2\gamma)\int_0^1 (1-t)\,\mathrm{D}^2 U(x+t\sqrt{2\gamma}Z)[Z^{\otimes 2}]\mathrm{d}t \quad (6.90)$$

where τ^{RWM} is defined in (6.38). Define $\zeta_\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for all $x, z \in \mathbb{R}$ and $\gamma \in (0, \bar{\gamma}]$ by,

$$\zeta_\gamma(x, z) = \left[1 - \exp\left\{-\tau^{\mathrm{RWM}}(x, x + \sqrt{2\gamma}z)_+\right\} - \sqrt{2\gamma}\,\langle \nabla U(x), z\rangle_+\right] \Big/ \gamma\,. \quad (6.91)$$

ζ_γ is continuous. Note that for all $x, y \in \mathbb{R}^d$,

$$\tau^{\mathrm{RWM}}(x, y)_+ - (1/2)\left\{\tau^{\mathrm{RWM}}(x, y)_+\right\}^2 \le 1 - \mathrm{e}^{-\tau^{\mathrm{RWM}}(x,y)_+} \le \tau^{\mathrm{RWM}}(x, y)_+\,. \quad (6.92)$$

By (6.90), there exists $p_1 \ge 0$ such that for all $\gamma \in (0, \bar{\gamma}]$ and $x, z \in \mathbb{R}^d$

$$\left|\tau^{\mathrm{RWM}}(x, x + \sqrt{2\gamma}z)_+ - \sqrt{2\gamma}\,\langle \nabla U(x), z\rangle_+\right| \le C\gamma(1 + \|x\|^{p_1} + \|z\|^{p_1})\,.$$

Combining it with (6.92), there exists $p_2 \ge 0$ such that for all $\gamma \in (0, \bar{\gamma}]$, $x, z \in \mathbb{R}^d$, $|\zeta_\gamma(x, z)| \le C(1 + \|x\|^{p_2} + \|z\|^{p_2})$. By (6.37), we have

$$\begin{aligned} R^{\mathrm{RWM}}_\gamma\varphi(x) - \varphi(x) &= \mathbb{E}\left[\varphi(x + \sqrt{2\gamma}Z) - \varphi(x)\right] \\ &\quad + \mathbb{E}\left[\left(\mathrm{e}^{-\tau^{\mathrm{RWM}}(x,Y)_+} - 1\right)\left\{\varphi(x + \sqrt{2\gamma}Z) - \varphi(x)\right\}\right] \end{aligned} \quad (6.93)$$

and using a Taylor expansion again, we get for all $z \in \mathbb{R}^d$,

$$\begin{aligned} \varphi(x + \sqrt{2\gamma}z) - \varphi(x) &= \sqrt{2\gamma}\,\langle \nabla\varphi(x), z\rangle + (2\gamma)\int_0^1 (1-t)\,\mathrm{D}^2\varphi(x + t\sqrt{2\gamma}z)[z^{\otimes 2}]\mathrm{d}t \\ &= \sqrt{2\gamma}\,\langle \nabla\varphi(x), z\rangle + \gamma\,\mathrm{D}^2\varphi(x)[z^{\otimes 2}] + (\sqrt{2}/3)\gamma^{3/2}\,\mathrm{D}^3\varphi(x)[z^{\otimes 3}] \\ &\quad + (2/3)\gamma^2\int_0^1 (1-t)^3\,\mathrm{D}^4\varphi(x + t\sqrt{2\gamma}z)[z^{\otimes 4}]\mathrm{d}t\,. \end{aligned} \quad (6.94)$$

Note that (6.40) is equivalent to $R_\gamma^{\mathrm{RWM}} = \mathrm{Id} + \gamma \mathscr{L} + \gamma^{3/2} \mathscr{A}_\gamma^{\mathrm{RWM}}$ for $\gamma > 0$. By (6.93), (6.94) and using for all $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\langle \nabla U(x), Z \rangle_+ \langle \nabla \varphi(x), Z \rangle \right] = (1/2) \langle \nabla U(x), \nabla \varphi(x) \rangle \ ,$$

$\mathscr{A}_\gamma^{\mathrm{RWM}} : \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}) \to \mathrm{C}^0_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ is given for $\varphi \in \mathrm{C}^4_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$, $x \in \mathbb{R}^d$ and $\gamma \in (0, \bar{\gamma}]$ by

$$\begin{aligned} \mathscr{A}_\gamma^{\mathrm{RWM}} \varphi(x) = -\mathbb{E}\Big[\int_0^1 (1-t)\, \mathrm{D}^2 \varphi(x + t\sqrt{2\gamma} Z)[Z^{\otimes 2}] \mathrm{d}t \left\{ 2^{3/2} \langle \nabla U(x), Z \rangle_+ + 2\sqrt{\gamma} \zeta_\gamma(x, Z) \right\} \\ + \sqrt{2} \zeta_\gamma(x, Z) \langle \nabla \varphi(x), Z \rangle - (2/3)\sqrt{\gamma} \int_0^1 (1-t)^3 \, \mathrm{D}^4 \varphi(x + t\sqrt{2\gamma} Z)[Z^{\otimes 4}] \mathrm{d}t \Big] \ , \end{aligned} \tag{6.95}$$

which gives (i).

For the assertion (ii), for any $x, z \in \mathbb{R}^d$, distinguishing the cases where $\langle \nabla U(x), z \rangle$ is positive, zero, or negative, respectively, we obtain taking the limit $\gamma \downarrow 0^+$ in (6.91) and using (6.90), $\lim_{\gamma \downarrow 0^+} \zeta_\gamma(x, z) = \zeta_0(x, z)$ where ζ_0 is defined in (6.41). By the dominated convergence theorem, taking the limit $\gamma \downarrow 0^+$ in (6.95), we get (ii).

6.E.2 Proof of Lemma 6.10

We first state two technical lemmas. Define $G : \mathbb{R}_+ \to [0, 1]$ for all $t \geq 0$ by

$$G(t) = 1/2 + 2\mathrm{e}^{t^2/2} \bar{\Phi}(t) - \mathrm{e}^{2t^2} \bar{\Phi}(2t) \ . \tag{6.96}$$

Lemma 6.19. *There exists $t_0 > 0$ such that for all $t \in [0, t_0]$, $G(t) \leq 1 - (t^2/2)$ and the function G is non-increasing.*

Proof. We have for all $t \geq 0$,

$$G'(t) = 2t\mathrm{e}^{t^2/2} \left\{ \bar{\Phi}(t) - 2\mathrm{e}^{(3t^2)/2} \bar{\Phi}(2t) \right\} \tag{6.97}$$

and $G'(0) = 0$, $G''(0) = -1$ so there exists $t_0 > 0$ such that for all $t \in [0, t_0]$, $G(t) \leq 1 - (t^2/2)$, which is the first statement of the lemma. Regarding the second statement, by an integration by parts, we have for all $s > 0$

$$\bar{\Phi}(s) = \frac{\mathrm{e}^{-s^2/2}}{\sqrt{2\pi} s} - \frac{1}{\sqrt{2\pi}} \int_s^{+\infty} \frac{\mathrm{e}^{-u^2/2}}{u^2} \mathrm{d}u$$

and using a change of variables $u = v + t$, we get for all $t > 0$

$$\bar{\Phi}(t) - 2\mathrm{e}^{(3t^2)/2} \bar{\Phi}(2t) = \int_t^{+\infty} \left\{ \frac{2\mathrm{e}^{t(t-v)}}{(v+t)^2} - \frac{1}{v^2} \right\} \frac{\mathrm{e}^{-v^2/2}}{\sqrt{2\pi}} \mathrm{d}v \ .$$

We now show that $\bar{\Phi}(t) - 2\mathrm{e}^{(3t^2)/2}\bar{\Phi}(2t) \leq 0$ for all $t \geq 0$ which will finish the proof using (6.97). We distinguish the case $t \geq 0.4$ and $t \in [0, 0.4]$. For $t \geq 0.4$, define $h_t : [t, +\infty) \to \mathbb{R}$ given for all $v \geq t$ by

$$h_t(v) = 2\ln(1 + t/v) - \ln(2) - t^2 + vt \; .$$

We show in the sequel that $h_t(v) \geq 0$ for all $v \geq t \geq 0.4$, which implies $\bar{\Phi}(t) - 2\mathrm{e}^{(3t^2)/2}\bar{\Phi}(2t) \leq 0$ for all $t \geq 0.4$. We have for all $v \geq t$

$$h_t'(v) = t\{-2/\{v(t+v)\} + 1\}$$

and h_t is decreasing on $[t, v_{\min} \vee t]$ and increasing on $[v_{\min} \vee t, +\infty)$ where $v_{\min} = (-t + \sqrt{t^2+8})/2$. Note that $v_{\min} \geq t$ is equivalent to $t \leq 1$ and for all $t \geq 1$, $h_t(t) = \ln(2) > 0$. Define $\ell : (0, 1] \to \mathbb{R}$ given for all $t \in (0, 1]$ by

$$\begin{aligned}\ell(t) = h_t(v_{\min}) &= 2\ln\left(\frac{\sqrt{t^2+8}+t}{\sqrt{t^2+8}-t}\right) - \ln(2) + (t/2)\left(-3t + \sqrt{t^2+8}\right) \\ &= 5\ln(2) - 4\ln\left(-t + \sqrt{t^2+8}\right) + (t/2)\left(-3t + \sqrt{t^2+8}\right) \; .\end{aligned}$$

We have for all $t \in (0, 1]$

$$\ell'(t) = -3t + \sqrt{t^2+8} \geq 0 \; ,$$

ℓ is non-decreasing and $\ell(0.4) > 0$, which implies that for all $t \in [0.4, 1]$ and $v \geq t$, $h_t(v) \geq 0$. Therefore, $G'(t) \leq 0$ for all $t \geq 0.4$.

For $t \in [0, 0.4]$, we use the following lower and upper bounds by [CCM11, Theorems 1 and 2] for all $s \geq 0$

$$\frac{\sqrt{\mathrm{e}}}{3\sqrt{\pi}}\mathrm{e}^{-(3/4)s^2} \leq \bar{\Phi}(s) \leq (1/2)\mathrm{e}^{-s^2/2}$$

and we get for all $t \in [0, 0.4]$

$$2\mathrm{e}^{(3t^2)/2}\bar{\Phi}(2t) - \bar{\Phi}(t) \geq \mathrm{e}^{-t^2/2}\left\{\frac{2\sqrt{\mathrm{e}}}{3\sqrt{\pi}}\mathrm{e}^{-t^2} - \frac{1}{2}\right\} \; .$$

The right hand side is decreasing on $[0, 0.4]$ and positive because $(2\sqrt{\mathrm{e}}\mathrm{e}^{-(0.4)^2})/(3\sqrt{\pi}) - (1/2) \geq 0.02$, which implies that $G'(t) \leq 0$ for all $t \in [0, 0.4]$. □

Lemma 6.20. *Assume that* $U \in \mathrm{C}^3_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ *and* ***H*** *18. Let* $x \in \mathbb{R}^d$, $\|x\| \geq M$ *and* $K > 0$. *For all* $\gamma > 0$ *and* $z \in \overline{B}(0, K)$, *we have*

$$\left\|\mathrm{D}^2 U(x + \sqrt{2\gamma}z)\right\| \leq \left\|\mathrm{D}^2 U(x)\right\| \{1 + C(K)\} \;\; \textit{where} \; C(K) = (C\chi K)^{1/2}\gamma^{1/4}\mathrm{e}^{C\chi\sqrt{\gamma}K/2} \; .$$

Proof. Let $z \in \overline{B}(0, K)$. Define $f : [0, 1] \to \mathbb{R}^{d\times d}$ by $f(t) = \mathrm{D}^2 U(x + t\sqrt{2\gamma}z) - \mathrm{D}^2 U(x)$ for $t \in [0, 1]$. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|f(t)\|_{\mathrm{F}}^2 = \left\langle f(t), \mathrm{D}^3 U(x + t\sqrt{2\gamma}z) \cdot \sqrt{2\gamma}z \right\rangle_{\mathrm{F}}$$

where for $i, j \in \{1, \dots, d\}$

$$\left(\mathrm{D}^3 U(x + t\sqrt{2\gamma}z) \cdot \sqrt{2\gamma}z\right)_{ij} = \sum_{k=1}^{d} \partial_{ijk} U(x + t\sqrt{2\gamma}z)\sqrt{2\gamma}z_k \, .$$

Using the equivalence of norms in finite dimension and **H**18, we get

$$\left|\frac{\mathrm{d}}{\mathrm{d}t} \|f(t)\|_{\mathrm{F}}^2\right| \leq C \|f(t)\|_{\mathrm{F}} \left\|\mathrm{D}^3 U(x + t\sqrt{2\gamma}z)\right\| \sqrt{2\gamma} \|z\| \leq C\chi \left(\|f(t)\|_{\mathrm{F}}^2 + \left\|\mathrm{D}^2 U(x)\right\|^2\right) \sqrt{\gamma} \|z\|$$

which gives by Grönwall's inequality,

$$\|f(1)\|^2 \leq \left\|\mathrm{D}^2 U(x)\right\|^2 \left(\mathrm{e}^{C\chi\sqrt{\gamma}\|z\|} - 1\right) \, .$$

Using $(\mathrm{e}^s - 1)^{1/2} \leq \sqrt{s}\mathrm{e}^{s/2}$ for all $s \geq 0$, we get the result. □

We now proceed to the proof of Lemma 6.10. Note that we have for all $x \in \mathbb{R}^d$ and $\gamma > 0$

$$\begin{aligned} \frac{R_\gamma^{\mathrm{RWM}} V(x)}{V(x)} &= \int_{\mathbf{A}_{x,\gamma}^{\mathrm{RWM}}} \sqrt{\frac{\pi(x)}{\pi(x + \sqrt{2\gamma}z)}} \frac{\mathrm{e}^{-\|z\|^2/2}}{(2\pi)^{d/2}} \mathrm{d}z \\ &\quad + \int_{(\mathbf{A}_{x,\gamma}^{\mathrm{RWM}})^{\mathrm{c}}} \left\{1 + \sqrt{\frac{\pi(x + \sqrt{2\gamma}z)}{\pi(x)}} - \frac{\pi(x + \sqrt{2\gamma}z)}{\pi(x)}\right\} \frac{\mathrm{e}^{-\|z\|^2/2}}{(2\pi)^{d/2}} \mathrm{d}z \end{aligned} \tag{6.98}$$

where $\mathbf{A}_{x,\gamma}^{\mathrm{RWM}}$ is defined in (6.39).

Intuition behind the proof Before giving the proof of Lemma 6.10, we sketch here the analysis of a simple case in one dimension where $U(x) = a|x|$ (with a proper regularization near 0), $a > 0$ and let $x > 0$ be large enough. By (6.98), we get

$$\begin{aligned} \frac{R_\gamma^{\mathrm{RWM}} V(x)}{V(x)} &\approx \int_0^{+\infty} \mathrm{e}^{-a\sqrt{\gamma/2}z} \frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}} \mathrm{d}z + \int_0^{+\infty} \left\{1 + \mathrm{e}^{-a\sqrt{\gamma/2}z} - \mathrm{e}^{-a\sqrt{2\gamma}z/2}\right\} \frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}} \mathrm{d}z \\ &= (1/2) + 2\mathrm{e}^{a^2\gamma/4} \bar{\Phi}(\sqrt{\gamma/2}a) - \mathrm{e}^{a^2\gamma} \bar{\Phi}(\sqrt{2\gamma}a) \\ &= G(a\sqrt{\gamma/2}) \approx 1 - (\gamma a^2)/4 + O(\gamma^{3/2}a^3) \end{aligned}$$

and the expected contraction in $1 - C\gamma$. The proof below is devoted to make this intuition rigorous and the main steps are a localization argument, a comparison to the one dimensional case and an upper bound on the remainder terms.

In the sequel, let $x \in \mathbb{R}^d$, $\|x\| \geq M$ where M is given by **H**18.

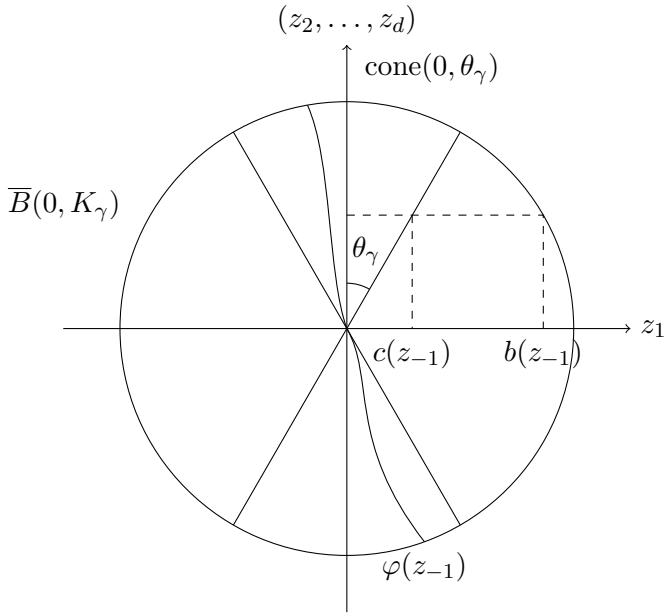

Figure 6.6: Figure illustrating the definitions of $\mathrm{cone}(0, \theta_\gamma)$, $b(z_{-1})$, $c(z_{-1})$ and $\varphi(z_{-1})$.

Step 1: restriction to $\overline{B}(0, K_\gamma)$. Define for all $\gamma > 0$

$$K_\gamma = \{8 \log((1/\gamma) \vee 1) + 2d \log(2)\}^{1/2} . \tag{6.99}$$

Let Z be an i.i.d. standard d-dimensional Gaussian variable. By Markov's inequality and (6.99), we have

$$\mathbb{P}\left(\|Z\| \geq K_\gamma\right) \leq \mathrm{e}^{-K_\gamma^2/4} \mathbb{E}\left[\mathrm{e}^{\|Z\|^2/4}\right] \leq \exp\left(-\frac{K_\gamma^2}{4} + \frac{d}{2}\log(2)\right) \leq \gamma^2 . \tag{6.100}$$

Using $\pi(x)/\pi(x+\sqrt{2\gamma}z) \leq 1$ for $z \in \mathbf{A}_{x,\gamma}^{\mathrm{RWM}}$, $1+\sqrt{\pi(x+\sqrt{2\gamma}z)/\pi(x)}-\pi(x+\sqrt{2\gamma}z)/\pi(x) \leq 5/4$ for $z \in (\mathbf{A}_{x,\gamma}^{\mathrm{RWM}})^{\mathrm{c}}$, (6.98) and (6.100), we get

$$\begin{aligned}\frac{R_\gamma^{\mathrm{RWM}} V(x)}{V(x)} &\leq (5/4)\gamma^2 + \int_{\mathbf{A}_{x,\gamma}^{\mathrm{RWM}}} \mathbb{1}_{\overline{B}(0,K_\gamma)}(z) \sqrt{\frac{\pi(x)}{\pi(x+\sqrt{2\gamma}z)}} \frac{\mathrm{e}^{-\|z\|^2/2}}{(2\pi)^{d/2}} \mathrm{d}z \\ &+ \int_{(\mathbf{A}_{x,\gamma}^{\mathrm{RWM}})^{\mathrm{c}}} \mathbb{1}_{\overline{B}(0,K_\gamma)}(z) \left\{1 + \sqrt{\frac{\pi(x+\sqrt{2\gamma}z)}{\pi(x)}} - \frac{\pi(x+\sqrt{2\gamma}z)}{\pi(x)}\right\} \frac{\mathrm{e}^{-\|z\|^2/2}}{(2\pi)^{d/2}} \mathrm{d}z . \end{aligned} \tag{6.101}$$

Step 2: splitting $\overline{B}(0, K_\gamma)$ into $\overline{B}(0, K_\gamma) \cap \mathbf{A}_{x,\gamma}^{\mathrm{RWM}}$ and $\overline{B}(0, K_\gamma) \cap (\mathbf{A}_{x,\gamma}^{\mathrm{RWM}})^{\mathbf{c}}$. In this paragraph, we introduce several geometric quantities illustrated with Figure 6.6. Define $\bar{\gamma} > 0$ by

$$\max\left\{(C\chi K_{\bar{\gamma}})^{1/2} \bar{\gamma}^{1/4} \exp(C\chi \bar{\gamma}^{1/2} K_{\bar{\gamma}}/2),\ (3/2)\sqrt{2\bar{\gamma}} K_{\bar{\gamma}} \chi\right\} = 1/2 , \tag{6.102}$$

where C is the positive constant given in Lemma 6.20. Denote by

$$C_1 = (C\chi K_{\bar\gamma})^{1/2}\bar\gamma^{1/4}\exp(C\chi\bar\gamma^{1/2}K_{\bar\gamma}/2) \in [0,1/2] \ . \tag{6.103}$$

Let $e_1(x) = \nabla U(x)/\left\|\nabla U(x)\right\|$ and consider the decomposition $z = (z_1,\dots,z_d)$ of z in an orthonormal basis $(e_1(x), e_2(x),\dots,e_d(x))$ of $\mathbb{R}^d$. For all $z\in\mathbb{R}^d$, denote by $z_{-1} = (z_2,\dots,z_d)\in\mathbb{R}^{d-1}$. For all $\gamma\in(0,\bar\gamma]$, define $\theta_\gamma\in[0,\pi/4]$ by

$$\tan\theta_\gamma = 2\sqrt{2\gamma}K_\gamma\frac{\left\|\mathrm{D}^2U(x)\right\|}{\left\|\nabla U(x)\right\|}(1+C_1)\in[0,1] \ . \tag{6.104}$$

Denote by

$$\mathrm{cone}(0,\theta_\gamma) = \left\{z\in\mathbb{R}^d : |z_1|\le(\tan\theta_\gamma)\left\|z_{-1}\right\|\right\} \ .$$

Define $b, c : \overline{\mathrm{B}}_{d-1}(0,K_\gamma)\to\mathbb{R}_+$ for all $z_{-1}\in\overline{\mathrm{B}}_{d-1}(0,K_\gamma)$ by

$$b(z_{-1}) = (K_\gamma^2 - \|z_{-1}\|^2)^{1/2} \quad\text{and}\quad c(z_{-1}) = (\tan\theta_\gamma)\left\|z_{-1}\right\| \ . \tag{6.105}$$

By Lemma 6.20 with $K = K_\gamma$, we have for all $z\in\overline{B}(0,K_\gamma)$

$$\left\|\mathrm{D}^2U(x+\sqrt{2\gamma}z)\right\| \le \left\|\mathrm{D}^2U(x)\right\|(1+C_1) \ . \tag{6.106}$$

where C_1 is given in (6.103). By Taylor's theorem, we have for all $z\in\overline{B}(0,K_\gamma)$

$$U(x+\sqrt{2\gamma}z) - U(x) = \sqrt{2\gamma}\left\|\nabla U(x)\right\|z_1 + 2\mathsf{r}_\gamma(z) \tag{6.107}$$

where $\mathsf{r}_\gamma : \overline{B}(0,K_\gamma)\to\mathbb{R}$ is defined for all $z\in\overline{B}(0,K_\gamma)$ by

$$\mathsf{r}_\gamma(z) = \gamma\int_0^1(1-t)\,\mathrm{D}^2U(x+t\sqrt{2\gamma}z)[z^{\otimes 2}]\mathrm{d}t \ . \tag{6.108}$$

By (6.104), (6.106) and (6.108), we have for all $z\in\overline{B}(0,K_\gamma)\cap\mathrm{cone}(0,\theta_\gamma)^{\mathrm{c}}$

$$\begin{aligned}4\mathsf{r}_\gamma(z) &\le 2\gamma K_\gamma\left\|\mathrm{D}^2U(x)\right\|(1+C_1)\left(|z_1|+\|z_{-1}\|\right)\\ &\le \sqrt{2\gamma}\left\|\nabla U(x)\right\|(1/2)\tan\theta_\gamma\left\{1+(\tan\theta_\gamma)^{-1}\right\}|z_1| \le \sqrt{2\gamma}\left\|\nabla U(x)\right\||z_1| \ .\end{aligned} \tag{6.109}$$

By (6.107) and (6.109), we obtain for all $z\in\overline{B}(0,K_\gamma)\cap\mathrm{cone}(0,\theta_\gamma)^{\mathrm{c}}$, $z\neq 0$,

$$\left\{U(x+\sqrt{2\gamma}z) - U(x)\right\}z_1 > 0 \ . \tag{6.110}$$

Moreover, by **H**18 and (6.106), we have for all $z\in\overline{B}(0,K_\gamma)$

$$\begin{aligned}&\left\langle e_1(x), \nabla U(x+\sqrt{2\gamma}z)\right\rangle - \left\|\nabla U(x)\right\| = \sqrt{2\gamma}\int_0^1\mathrm{D}^2U(x+t\sqrt{2\gamma}z)[z,e_1(x)]\mathrm{d}t \ ,\\ &\left|\left\langle e_1(x), \nabla U(x+\sqrt{2\gamma}z)\right\rangle - \left\|\nabla U(x)\right\|\right| \le \sqrt{2\gamma}(1+C_1)\chi K_\gamma\left\|\nabla U(x)\right\|\end{aligned}$$

and $\langle e_1(x), \nabla U(x + \sqrt{2\gamma}z)\rangle > 0$. By a version of the implicit function theorem given in Section 6.E.3, there exists $\varphi : \overline{\mathrm{B}}_{d-1}(0, K_\gamma) \to \mathbb{R}$ continuous such that for all $\gamma \in (0, \bar{\gamma}]$,

$$\left\{z \in \overline{B}(0, K_\gamma) : U(x + \sqrt{2\gamma}z) = U(x)\right\} = \left\{(\varphi(z_{-1}), z_{-1}) : z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)\right\} . \tag{6.111}$$

Combining (6.110) and (6.111), we obtain for all $\gamma \in (0, \bar{\gamma}]$,

$$\mathbf{A}_{x,\gamma}^{\mathrm{RWM}} \cap \overline{B}(0, K_\gamma) = \left\{z \in \overline{B}(0, K_\gamma) : z_1 \leq \varphi(z_{-1})\right\} , \tag{6.112}$$

$$(\mathbf{A}_{x,\gamma}^{\mathrm{RWM}})^{\mathrm{c}} \cap \overline{B}(0, K_\gamma) = \left\{z \in \overline{B}(0, K_\gamma) : z_1 \geq \varphi(z_{-1})\right\} , \tag{6.113}$$

and for all $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)$, $|\varphi(z_{-1})| \leq c(z_{-1})$. These properties and definitions are summarized in Figure 6.6.

Step 3: intermediate upper bound on $R_\gamma^{\mathrm{RWM}} V(x)/V(x)$. Using (6.101) and the definitions of b and φ, see (6.105), (6.111), (6.112) and (6.113), we have

$$\frac{R_\gamma^{\mathrm{RWM}} V(x)}{V(x)} \leq (5/4)\gamma^2 + \int_{z_{-1} \in \overline{\mathrm{B}}_{d-1}(0,K_\gamma)} g_\gamma(z_{-1}) \frac{\mathrm{e}^{-\|z_{-1}\|^2/2}}{(2\pi)^{(d-1)/2}} \mathrm{d}z_{-1} \tag{6.114}$$

where $g_\gamma : \overline{\mathrm{B}}_{d-1}(0, K_\gamma) \to \mathbb{R}_+$ is defined for all $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)$ by

$$\begin{aligned} g_\gamma(z_{-1}) = &\int_{-b(z_{-1})}^{(\varphi(z_{-1}) \vee -b(z_{-1})) \wedge b(z_{-1})} \sqrt{\frac{\pi(x)}{\pi(x + \sqrt{2\gamma}z)}} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \\ &+ \int_{(\varphi(z_{-1}) \vee -b(z_{-1})) \wedge b(z_{-1})}^{b(z_{-1})} \left\{1 + \sqrt{\frac{\pi(x + \sqrt{2\gamma}z)}{\pi(x)}} - \frac{\pi(x + \sqrt{2\gamma}z)}{\pi(x)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 . \end{aligned}$$

For all $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)$, we decompose $g_\gamma(z_{-1})$ in $g_\gamma(z_{-1}) = A_1(z_{-1}) + A_2(z_{-1})$ where $A_1(z_{-1})$ and $A_2(z_{-1})$ are defined by

$$\begin{aligned} A_1(z_{-1}) = &\int_{-b(z_{-1})}^{(\varphi(z_{-1}) \vee -b(z_{-1})) \wedge 0} \sqrt{\frac{\pi(x)}{\pi(x + \sqrt{2\gamma}z)}} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \\ &+ \int_{(\varphi(z_{-1}) \vee -b(z_{-1})) \wedge 0}^{0} \left\{1 + \sqrt{\frac{\pi(x + \sqrt{2\gamma}z)}{\pi(x)}} - \frac{\pi(x + \sqrt{2\gamma}z)}{\pi(x)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 , \end{aligned} \tag{6.115}$$

$$\begin{aligned} A_2(z_{-1}) = &\int_{0}^{(\varphi(z_{-1}) \vee 0) \wedge b(z_{-1})} \sqrt{\frac{\pi(x)}{\pi(x + \sqrt{2\gamma}z)}} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \\ &+ \int_{(\varphi(z_{-1}) \vee 0) \wedge b(z_{-1})}^{b(z_{-1})} \left\{1 + \sqrt{\frac{\pi(x + \sqrt{2\gamma}z)}{\pi(x)}} - \frac{\pi(x + \sqrt{2\gamma}z)}{\pi(x)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 . \end{aligned} \tag{6.116}$$

Combining it with (6.114), we obtain

$$\frac{R_\gamma^{\mathrm{RWM}} V(x)}{V(x)} \le (5/4)\gamma^2 + \int_{z_{-1}\in\overline{\mathrm{B}}_{d-1}(0,K_\gamma)} \{A_1(z_{-1}) + A_2(z_{-1})\} \frac{\mathrm{e}^{-\|z_{-1}\|^2/2}}{(2\pi)^{(d-1)/2}} \mathrm{d}z_{-1} \ . \quad (6.117)$$

By (6.107) and (6.115), we have for all $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)$

$$A_1(z_{-1}) = A_{11}(z_{-1}) + A_{12}(z_{-1}) + A_{13}(z_{-1}) + A_{14}(z_{-1}) \quad (6.118)$$

where

$$\begin{aligned}
A_{11}(z_{-1}) &= \int_{-b(z_{-1})}^{0} \mathrm{e}^{\sqrt{\gamma/2}\|\nabla U(x)\| z_1} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ , \\
A_{12}(z_{-1}) &= \int_{-b(z_{-1})}^{-b(z_{-1})\vee -c(z_{-1})} \mathrm{e}^{\sqrt{\gamma/2}\|\nabla U(x)\| z_1 + \mathsf{r}_\gamma(z)} \left\{1 - \mathrm{e}^{-\mathsf{r}_\gamma(z)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ , \\
A_{13}(z_{-1}) &= \int_{-b(z_{-1})\vee -c(z_{-1})}^{(\varphi(z_{-1})\vee -b(z_{-1}))\wedge 0} \mathrm{e}^{\sqrt{\gamma/2}\|\nabla U(x)\| z_1 + \mathsf{r}_\gamma(z)} \left\{1 - \mathrm{e}^{-\mathsf{r}_\gamma(z)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ , \\
A_{14}(z_{-1}) &= \int_{(\varphi(z_{-1})\vee -b(z_{-1}))\wedge 0}^{0} \left\{1 + \sqrt{\frac{\pi(x+\sqrt{2\gamma}z)}{\pi(x)}} - \frac{\pi(x+\sqrt{2\gamma}z)}{\pi(x)} \right. \\
&\qquad \left. - \mathrm{e}^{\sqrt{\gamma/2}\|\nabla U(x)\| z_1} \right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ .
\end{aligned}$$

By (6.107) and (6.116), we have for all $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)$

$$\begin{aligned}
A_2(z_{-1}) &= A_{21}(z_{-1}) + A_{22}(z_{-1}) + A_{23}(z_{-1}) + A_{24}(z_{-1}) + A_{25}(z_{-1}) \\
&\quad + \int_0^{(\varphi(z_{-1})\vee 0)\wedge b(z_{-1})} \left\{ \sqrt{\frac{\pi(x)}{\pi(x+\sqrt{2\gamma}z)}} - 1 - \mathrm{e}^{-\sqrt{\gamma/2}\|\nabla U(x)\| z_1} \right. \\
&\qquad \left. + \mathrm{e}^{-\sqrt{2\gamma}\|\nabla U(x)\| z_1} \right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \quad (6.119)
\end{aligned}$$

where

$$\begin{aligned}
A_{21}(z_{-1}) &= \int_0^{b(z_{-1})} \left\{1 + \mathrm{e}^{-\sqrt{\gamma/2}\|\nabla U(x)\| z_1} - \mathrm{e}^{-\sqrt{2\gamma}\|\nabla U(x)\| z_1}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ , \\
A_{22}(z_{-1}) &= \int_{(\varphi(z_{-1})\vee 0)\wedge b(z_{-1})}^{c(z_{-1})\wedge b(z_{-1})} \mathrm{e}^{-\sqrt{\gamma/2}\|\nabla U(x)\| z_1 - \mathsf{r}_\gamma(z)} \left\{1 - \mathrm{e}^{\mathsf{r}_\gamma(z)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ , \\
A_{23}(z_{-1}) &= \int_{(\varphi(z_{-1})\vee 0)\wedge b(z_{-1})}^{c(z_{-1})\wedge b(z_{-1})} \mathrm{e}^{-\sqrt{2\gamma}\|\nabla U(x)\| z_1} \left\{1 - \mathrm{e}^{-2\mathsf{r}_\gamma(z)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ , \\
A_{24}(z_{-1}) &= \int_{c(z_{-1})\wedge b(z_{-1})}^{b(z_{-1})} \mathrm{e}^{-\sqrt{\gamma/2}\|\nabla U(x)\| z_1 - \mathsf{r}_\gamma(z)} \left\{1 - \mathrm{e}^{\mathsf{r}_\gamma(z)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ , \\
A_{25}(z_{-1}) &= \int_{c(z_{-1})\wedge b(z_{-1})}^{b(z_{-1})} \mathrm{e}^{-\sqrt{2\gamma}\|\nabla U(x)\| z_1} \left\{1 - \mathrm{e}^{-2\mathsf{r}_\gamma(z)}\right\} \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \ .
\end{aligned}$$

By (6.112), $\{\pi(x)/\pi(x+\sqrt{2\gamma}z)\}^{1/2} \leq 1$ for all $z_1 \in [0, \varphi(z_{-1}) \vee 0]$. Hence, the last term in the right hand side of (6.119) is nonpositive and we get

$$A_2(z_{-1}) \leq A_{21}(z_{-1}) + A_{22}(z_{-1}) + A_{23}(z_{-1}) + A_{24}(z_{-1}) + A_{25}(z_{-1}) \; . \tag{6.120}$$

Combining (6.118) and (6.120), we obtain for all $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)$

$$\begin{aligned} A_1(z_{-1}) + A_2(z_{-1}) \leq A_{11}(z_{-1}) + A_{21}(z_{-1}) + A_{12}(z_{-1}) + A_{13}(z_{-1}) + A_{14}(z_{-1}) \\ + A_{22}(z_{-1}) + A_{23}(z_{-1}) + A_{24}(z_{-1}) + A_{25}(z_{-1}) \; . \end{aligned} \tag{6.121}$$

Step 4: upper bound on $A_1(z_{-1}) + A_2(z_{-1})$. We upper bound each term in the right hand side of (6.121) and we first consider the terms $A_{11} + A_{21}$. Define $a : (0, \bar{\gamma}] \times \mathbb{R}^d \to \mathbb{R}_+$ for all $\tilde{\gamma} \in (0, \bar{\gamma}]$ and $\tilde{x} \in \mathbb{R}^d$, $\|\tilde{x}\| \geq M$ by

$$a(\tilde{\gamma}, \tilde{x}) = \sqrt{\tilde{\gamma}/2}\, \|\nabla U(\tilde{x})\| \; . \tag{6.122}$$

We have for all $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)$,

$$A_{11}(z_{-1}) + A_{21}(z_{-1}) \leq G(a(\gamma, x)) \tag{6.123}$$

where G is defined in (6.96).

We now consider the remainder terms $A_{12}(z_{-1}), A_{13}(z_{-1}), A_{14}(z_{-1}), A_{22}(z_{-1}), A_{23}(z_{-1})$, $A_{24}(z_{-1})$ and $A_{25}(z_{-1})$ in (6.121). Let $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)$. By definition of $c(z_{-1})$, see (6.105), we have for all $z_1 \in [-b(z_{-1}), -c(z_{-1}) \vee -b(z_{-1})]$, $z \notin \mathrm{cone}(0, \theta_\gamma)$, and by (6.109)

$$\sqrt{\gamma/2}\, \|\nabla U(x)\|\, z_1 + \mathsf{r}_\gamma(z) \leq (1/2)\sqrt{\gamma/2}\, \|\nabla U(x)\|\, z_1 \; .$$

Combining it with $1 - \mathrm{e}^s \leq |s|$ for all $s \in \mathbb{R}$, (6.106) and (6.108), we get

$$A_{12}(z_{-1}) \leq C \int_{-b(z_{-1})}^{-c(z_{-1}) \vee -b(z_{-1})} \mathrm{e}^{(1/2)\sqrt{\gamma/2}\|\nabla U(x)\| z_1} \gamma \left\| \mathrm{D}^2 U(x) \right\| \|z\|^2 \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \; .$$

Considering the upper bound $\|z\|^2 \leq K_\gamma^2$ or the decomposition $\|z\|^2 = z_1^2 + \|z_{-1}\|^2$, we obtain

$$A_{12}(z_{-1}) \leq C\gamma \left\| \mathrm{D}^2 U(x) \right\| \min \left\{ K_\gamma^2 \mathrm{e}^{a(\gamma,x)^2/8} \bar{\Phi}(a(\gamma, x)/2), (\|z_{-1}\|^2 + 1) \right\}$$

where $a(\gamma, x)$ is defined in (6.122), and using for all $t > 0$, $\mathrm{e}^{t^2/8} \bar{\Phi}(t/2) \leq \sqrt{2}/(\sqrt{\pi} t)$, we get

$$A_{12}(z_{-1}) \leq C \min \left(\sqrt{\gamma} K_\gamma^2 \frac{\left\| \mathrm{D}^2 U(x) \right\|}{\|\nabla U(x)\|}, (\|z_{-1}\|^2 + 1) \frac{\left\| \mathrm{D}^2 U(x) \right\|}{\|\nabla U(x)\|^2} a(\gamma, x)^2 \right) \; . \tag{6.124}$$

Similarly, we have the same upper bound (6.124) for $A_{24}(z_{-1})$ and $A_{25}(z_{-1})$.

Using for all $s \in \mathbb{R}$, $1-\mathrm{e}^s \leq \min(1,|s|)$, $\pi(x)/\pi(x+\sqrt{2\gamma}z) \leq 1$ for $z \in \mathbf{A}_{x,\gamma}^{\mathrm{RWM}}$, (6.104), (6.105), (6.106), (6.107), (6.108) and (6.112), we have for all $z_{-1} \in \overline{\mathrm{B}}_{d-1}(0,K_\gamma)$,

$$\begin{aligned} A_{13}(z_{-1}) &\leq \int_{-b(z_{-1})\vee -c(z_{-1})}^{(\varphi(z_{-1})\vee -b(z_{-1}))\wedge 0} \min(1,|\mathsf{r}_\gamma(z)|) \frac{\mathrm{e}^{-z_1^2/2}}{(2\pi)^{1/2}} \mathrm{d}z_1 \\ &\leq c(z_{-1}) \min\left(1, C \left\|\mathrm{D}^2 U(x)\right\| \gamma K_\gamma^2\right) \\ &\leq C\sqrt{\gamma} K_\gamma^2 \frac{\left\|\mathrm{D}^2 U(x)\right\|}{\|\nabla U(x)\|} \min\left(1, C \left\|\mathrm{D}^2 U(x)\right\| \gamma K_\gamma^2\right) \\ &\leq C \min\left(\sqrt{\gamma} K_\gamma^2 \frac{\left\|\mathrm{D}^2 U(x)\right\|}{\|\nabla U(x)\|}, \sqrt{\gamma} K_\gamma^4 \frac{\left\|\mathrm{D}^2 U(x)\right\|^2}{\|\nabla U(x)\|^3} a(\gamma,x)^2 \right) . \end{aligned} \tag{6.125}$$

where $a(\gamma,x)$ is defined in (6.122). Similarly, we have the same upper bound (6.125) for $A_{22}(z_{-1})$ and $A_{23}(z_{-1})$.

Concerning $A_{14}(z_{-1})$, note first that by definition of $\varphi(z_{-1})$, see (6.111), (6.112), (6.113), and (6.107), (6.108) we have for all $z_1 \in [(\varphi(z_{-1}) \vee -b(z_{-1})) \wedge 0, 0]$

$$2\mathsf{r}_\gamma(z) \geq \left| \sqrt{2\gamma} \|\nabla U(x)\| z_1 \right| . \tag{6.126}$$

Using $1-\mathrm{e}^s \leq |s|$ for all $s \in \mathbb{R}$, $\sqrt{\pi(x+\sqrt{2\gamma}z)/\pi(x)} \leq 1$ for all $z \in (\mathbf{A}_{x,\gamma}^{\mathrm{RWM}})^{\mathrm{c}}$, (6.106), (6.108) and (6.126), we obtain

$$\begin{aligned} &\left\{1-\mathrm{e}^{\sqrt{\gamma/2}\|\nabla U(x)\| z_1}\right\} + \sqrt{\frac{\pi(x+\sqrt{2\gamma}z)}{\pi(x)}} \left\{1-\mathrm{e}^{-\sqrt{\gamma/2}\|\nabla U(x)\| z_1 - \mathsf{r}_\gamma(z)}\right\} \\ &\qquad \leq \min\left(1, \sqrt{\gamma/2}\|\nabla U(x)\| |z_1|\right) + \min\left(1, \left|\sqrt{\gamma/2}\|\nabla U(x)\| z_1 + \mathsf{r}_\gamma(z)\right|\right) \\ &\qquad \leq C \min\left(1, \gamma \left\|\mathrm{D}^2 U(x)\right\| K_\gamma^2\right) . \end{aligned}$$

By (6.104), (6.105) and using $|\varphi(z_{-1})| \leq c(z_{-1})$, we obtain

$$A_{14}(z_{-1}) \leq C \min\left(\sqrt{\gamma} K_\gamma^2 \frac{\left\|\mathrm{D}^2 U(x)\right\|}{\|\nabla U(x)\|}, \sqrt{\gamma} K_\gamma^4 \frac{\left\|\mathrm{D}^2 U(x)\right\|^2}{\|\nabla U(x)\|^3} a(\gamma,x)^2 \right) \tag{6.127}$$

where $a(\gamma,x)$ is defined in (6.122).

Step 5: conclusion. Let $\epsilon = (1/4)\min(1,t_0^2)$ where t_0 is defined in Lemma 6.19. Let $\tilde{\gamma} > 0$ be defined by $C\sqrt{\tilde{\gamma}} K_{\tilde{\gamma}}^2 \chi \max\left(1, K_{\tilde{\gamma}}^2 \chi^2\right) = \epsilon$ where C is the maximum of the positive constants given in (6.124), (6.125) and (6.127). Define then $\bar{\gamma}_1 = \bar{\gamma} \wedge \tilde{\gamma} \wedge t_0^2 \wedge \min(1,\chi^2/2)/10$ where $\bar{\gamma}$ is given in (6.102). By **H**18, there exists $\widetilde{M} \geq M$ such that for all $x \in \mathbb{R}^d$, $\|x\| \geq \widetilde{M}$, $Cd \left\|\mathrm{D}^2 U(x)\right\| / \|\nabla U(x)\|^2 \leq \epsilon$, where C is given in (6.124).

By (6.124), (6.125) and (6.127), we have for all $x \in \mathbb{R}^d$, $\|x\| \geq \widetilde{M}$ and $\gamma \in (0, \bar{\gamma}_1]$

$$\int_{z_{-1} \in \overline{\mathrm{B}}_{d-1}(x, K_\gamma)} \{A_{12}(z_{-1}) + A_{13}(z_{-1}) + A_{14}(z_{-1}) + A_{22}(z_{-1}) + A_{23}(z_{-1}) + A_{24}(z_{-1}) + A_{25}(z_{-1})\} \frac{\mathrm{e}^{-\|z_{-1}\|^2/2}}{(2\pi)^{(d-1)/2}} \mathrm{d}z_{-1} \leq \min(\epsilon, \epsilon a(\gamma, x)^2) \tag{6.128}$$

where $a(\gamma, x)$ is defined in (6.122). We consider now two cases:

- if $a(\gamma, x) > t_0$, by (6.121), (6.123), (6.128) and Lemma 6.19, for all $x \in \mathbb{R}^d$, $\|x\| \geq \widetilde{M}$, $\gamma \in (0, \bar{\gamma}_1]$

$$\int_{z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)} \{A_1(z_{-1}) + A_2(z_{-1})\} \frac{\mathrm{e}^{-\|z_{-1}\|^2/2}}{(2\pi)^{(d-1)/2}} \mathrm{d}z_{-1} \leq 1 - (t_0^2/2) + \epsilon \leq 1 - (t_0^2/4) \leq 1 - (1/4)\gamma \, .$$

- if $a(\gamma, x) \in (0, t_0]$, by (6.121), (6.123), (6.128), Lemma 6.19 and **H**18, for all $x \in \mathbb{R}^d$, $\|x\| \geq \widetilde{M}$, $\gamma \in (0, \bar{\gamma}_1]$,

$$\int_{z_{-1} \in \overline{\mathrm{B}}_{d-1}(0, K_\gamma)} \{A_1(z_{-1}) + A_2(z_{-1})\} \frac{\mathrm{e}^{-\|z_{-1}\|^2/2}}{(2\pi)^{(d-1)/2}} \mathrm{d}z_{-1} \leq 1 - (1/2 - \epsilon) a(\gamma, x)^2 \leq 1 - \frac{\gamma \|\nabla U(x)\|^2}{8} \leq 1 - \frac{\chi^{-2}\gamma}{8} \, .$$

Combining it with (6.117), we obtain for all $x \in \mathbb{R}^d$, $\|x\| \geq \widetilde{M}$, $\gamma \in (0, \bar{\gamma}_1]$,

$$R_\gamma^{\mathrm{RWM}} V(x)/V(x) \leq 1 - \min(1, \chi^{-2}/2)\gamma/8 \, .$$

Besides, denote by

$$A = \sup_{y, \|y\| \leq \widetilde{M}} \left\{ \frac{\mathscr{L} V(y)}{V(y)} + \bar{\gamma}_1^{1/2} \frac{\mathscr{A}_\gamma^{\mathrm{RWM}} V(y)}{V(y)} \right\} \, .$$

By Lemma 6.9, we have for all $x \in \mathbb{R}^d$, $\|x\| \leq \widetilde{M}$, $\gamma \in (0, \bar{\gamma}_1]$, $R_\gamma^{\mathrm{RWM}} V(x)/V(x) \leq 1 + \gamma A$. We get then for all $x \in \mathbb{R}^d$, $\gamma \in (0, \bar{\gamma}_1]$,

$$R_\gamma^{\mathrm{RWM}} V(x) \leq \left(1 - \frac{\min(1, \chi^{-2}/2)\gamma}{8}\right) V(x) + \gamma \left(A + \frac{\min(1, \chi^{-2}/2)}{8}\right) V(x) \mathbb{1}\left\{\|x\| \leq \widetilde{M}\right\}$$

which concludes the proof.

6.E.3 A version of the implicit function theorem

The following proposition is taken from [Apo69, Theorem 7.21] and [Bor13, Theorem 6].

Proposition 6.21. *Let* K *be a compact metric space and* $f : \mathbb{R} \times \mathsf{K} \to \mathbb{R}$ *be a continuous function. Assume that there exist* $M \geq m > 0$ *such that for all* $z \in \mathsf{K}$, $x, y \in \mathbb{R}$, $x \neq y$,

$$m \leq \frac{f(x,z) - f(y,z)}{x - y} \leq M \ . \tag{6.129}$$

Then, there exists a unique continuous function $\xi : \mathsf{K} \to \mathbb{R}$ *satisfying for all* $z \in \mathsf{K}$, $f(\xi(z), z) = 0$.

Proof. Denote by $\mathrm{C}(\mathsf{K})$ the set of real continuous functions on K. By standard arguments, $\mathrm{C}(\mathsf{K})$ is complete under the uniform norm defined for all $g_1, g_2 \in \mathrm{C}(\mathsf{K})$ by $\|g_1 - g_2\|_\infty = \sup_{z \in \mathsf{K}} \|g_1(z) - g_2(z)\|$. Define $\psi : \mathrm{C}(\mathsf{K}) \to \mathrm{C}(\mathsf{K})$ for all $g \in \mathrm{C}(\mathsf{K})$ and $z \in \mathsf{K}$ by

$$\psi(g)(z) = g(z) - (1/M) f(g(z), z) \ .$$

By (6.129), we have for all $g, h \in \mathrm{C}(\mathsf{K})$ and $z \in \mathsf{K}$,

$$|\psi(g)(z) - \psi(h)(z)| \leq \{1 - (m/M)\} \, |g(z) - h(z)|$$

and $\|\psi(g) - \psi(h)\|_\infty \leq \{1 - (m/M)\} \|g - h\|_\infty$. ψ is a contraction on $\mathrm{C}(\mathsf{K})$ and has a unique fixed point ξ in $\mathrm{C}(\mathsf{K})$ which satisfies $f(\xi(z), z) = 0$ for all $z \in \mathsf{K}$. □

Part III

Stochastic Gradient Langevin Dynamics

Chapter 7

The promises and pitfalls of Stochastic Gradient Langevin Dynamics

Nicolas Brosse [1], Alain Durmus [2] and Éric Moulines [1]

Abstract

Stochastic Gradient Langevin Dynamics (SGLD) has emerged as a key MCMC algorithm for Bayesian learning from large scale datasets. While SGLD with decreasing step sizes converges weakly to the posterior distribution, the algorithm is often used with a constant step size in practice and has demonstrated successes in machine learning tasks. The current practice is to set the step size inversely proportional to N where N is the number of training samples. As N becomes large, we show that the SGLD algorithm has an invariant probability measure which significantly departs from the target posterior and behaves like Stochastic Gradient Descent (SGD). This difference is inherently due to the high variance of the stochastic gradients. Several strategies have been suggested to reduce this effect; among them, SGLD Fixed Point (SGLDFP) uses carefully designed control variates to reduce the variance of the stochastic gradients. We show that SGLDFP gives approximate samples from the posterior distribution, with an accuracy comparable to the Langevin Monte Carlo (LMC) algorithm for a computational cost sublinear in the number of data points. We provide a detailed analysis of the Wasserstein distances between LMC, SGLD, SGLDFP and SGD and explicit expressions of the means and covariance matrices of their invariant distributions. Our findings are supported by some numerical experiments.

[1]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France.
Emails: nicolas.brosse@polytechnique.edu, eric.moulines@polytechnique.edu

[2]Ecole Normale Supérieure CMLA 61, Av. du Président Wilson 94235 Cachan Cedex, France
Email: alain.durmus@cmla.ens-cachan.fr

7.1 Introduction

Most MCMC algorithms have not been designed to process huge sample sizes, a typical setting in machine learning. As a result, many classical MCMC methods fail in this context, because the mixing time becomes prohibitively long and the cost per iteration increases proportionally to the number of training samples N. The computational cost in standard Metropolis-Hastings algorithm comes from 1) the computation of the proposals, 2) the acceptance/rejection step. Several approaches to solve these issues have been recently proposed in machine learning and computational statistics.

Among them, the stochastic gradient langevin dynamics (SGLD) algorithm, introduced in [WT11], is a popular choice. This method is based on the Langevin Monte Carlo (LMC) algorithm proposed in [Gre83; GM94]. Standard versions of LMC require to compute the gradient of the log-posterior at the current fit of the parameter, but avoid the accept/reject step. The LMC algorithm is a discretization of a continuous-time process, the overdamped Langevin diffusion, which leaves invariant the target distribution π. To further reduce the computational cost, SGLD uses unbiased estimators of the gradient of the log-posterior based on subsampling. This method has triggered a huge number of works among others [ABW12; KCW14; ASW14; CDC15; CFG14; Din+14; MCF15; Dub+16; BDH17] and have been successfully applied to a range of state of the art machine learning problems [PT13; LAW16].

The properties of SGLD with decreasing step sizes have been studied in [TTV16]. The two key findings in this work are that 1) the SGLD algorithm converges weakly to the target distribution π, 2) the optimal rate of convergence to equilibrium scales as $n^{-1/3}$ where n is the number of iterations, see [TTV16, Section 5]. However, in most of the applications, constant rather than decreasing step sizes are used, see [ABW12; CFG14; Has+17; Li+16; SN14; VZT16]. A natural question for the practical design of SGLD is the choice of the minibatch size. This size controls on the one hand the computational complexity of the algorithm per iteration and on the other hand the variance of the gradient estimator. Non-asymptotic bounds in Wasserstein distance between the marginal distribution of the SGLD iterates and the target distribution π have been established in [Dal17a; DK17]. These results highlight the cost of using stochastic gradients and show that, for a given precision ϵ in Wasserstein distance, the computational cost of the plain SGLD algorithm does not improve over the LMC algorithm; Nagapetyan et al. [Nag+17] reports also similar results on the mean square error.

It has been suggested to use control variates to reduce the high variance of the stochastic gradients. For strongly log-concave models, Nagapetyan et al. [Nag+17] and Baker et al. [Bak+17] use the mode of the posterior distribution as a reference point and introduce the SGLDFP (Stochastic Gradient Langevin Dynamics Fixed Point) algorithm. Nagapetyan et al. [Nag+17] and Baker et al. [Bak+17] provide upper bounds on the mean square error and the Wasserstein distance between the marginal distribution of the iterates of SGLDFP and the posterior distribution. In addition, Nagapetyan et al. [Nag+17] and Baker et al. [Bak+17] show that the overall cost remains sublinear in the number of individual data points, up to a preprocessing step. Other control variates

methodologies are provided for non-concave models in the form of SAGA-Langevin Dynamics and SVRG-Langevin Dynamics [Dub+16; Che+17], albeit a detailed analysis in Wasserstein distance of these algorithms is only available for strongly log-concave models [Cha+18].

In this paper, we provide further insights on the links between SGLD, SGLDFP, LMC and SGD (Stochastic Gradient Descent). In our analysis, the algorithms are used with a constant step size and the parameters are set to the standard values used in practice [ABW12; CFG14; Has+17; Li+16; SN14; VZT16]. The LMC, SGLD and SGLDFP algorithms define homogeneous Markov chains, each of which admits a unique stationary distribution used as a hopefully close proxy of π. The main contribution of this paper is to show that, while the invariant distributions of LMC and SGLDFP become closer to π as the number of data points increases, on the opposite, the invariant measure of SGLD never comes close to the target distribution π and is in fact very similar to the invariant measure of SGD.

In Section 7.3.1, we give an upper bound in Wasserstein distance of order 2 between the marginal distribution of the iterates of LMC and the Langevin diffusion, SGLDFP and LMC, and SGLD and SGD. We provide a lower bound on the Wasserstein distance between the marginal distribution of the iterates of SGLDFP and SGLD. In Section 7.3.2, we give a comparison of the means and covariance matrices of the invariant distributions of LMC, SGLDFP and SGLD with those of the target distribution π. Our claims are supported by numerical experiments in Section 7.4.

7.2 Preliminaries

Denote by $\mathbf{z} = \{z_i\}_{i=1}^N$ the observations. We are interested in situations where the target distribution π arises as the posterior in a Bayesian inference problem with prior density $\pi_0(\theta)$ and a large number $N \gg 1$ of i.i.d. observations z_i with likelihoods $p(z_i|\theta)$. In this case, $\pi(\theta) = \pi_0(\theta) \prod_{i=1}^N p(z_i|\theta)$. We denote $U_i(\theta) = -\log(p(z_i|\theta))$ for $i \in \{1, \dots, N\}$, $U_0(\theta) = -\log(\pi_0(\theta))$, $U = \sum_{i=0}^N U_i$.

Under mild conditions, π is the unique invariant probability measure of the Langevin Stochastic Differential Equation (SDE):

$$\mathrm{d}\theta_t = -\nabla U(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \,, \tag{7.1}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Based on this observation, Langevin Monte Carlo (LMC) is an MCMC algorithm that enables to sample (approximately) from π using an Euler discretization of the Langevin SDE:

$$\theta_{k+1} = \theta_k - \gamma \nabla U(\theta_k) + \sqrt{2\gamma} Z_{k+1} \,, \tag{7.2}$$

where $\gamma > 0$ is a constant step size and $(Z_k)_{k\geq 1}$ is a sequence of i.i.d. standard d-dimensional Gaussian vectors. Discovered and popularised in the seminal works [Gre83; GM94; RT96], LMC has recently received renewed attention [Dal17b; DM17; DM16; DK17]. However, the cost of one iteration is Nd which is prohibitively large for massive

datasets. In order to scale up to the big data setting, Welling and Teh [WT11] suggested to replace ∇U with an unbiased estimate $\nabla U_0 + (N/p)\sum_{i\in S}\nabla U_i$ where S is a minibatch of $\{1,\dots,N\}$ with replacement of size p. A single update of SGLD is then given for $k \in \mathbb{N}$ by

$$\theta_{k+1} = \theta_k - \gamma\left(\nabla U_0(\theta_k) + \frac{N}{p}\sum_{i\in S_{k+1}}\nabla U_i(\theta_k)\right) + \sqrt{2\gamma}Z_{k+1}\,. \tag{7.3}$$

The idea of using only a fraction of data points to compute an unbiased estimate of the gradient at each iteration comes from Stochastic Gradient Descent (SGD) which is a popular algorithm to minimize the potential U. SGD is very similar to SGLD because it is characterised by the same recursion as SGLD but without Gaussian noise:

$$\theta_{k+1} = \theta_k - \gamma\left(\nabla U_0(\theta_k) + \frac{N}{p}\sum_{i\in S_{k+1}}\nabla U_i(\theta_k)\right)\,. \tag{7.4}$$

Assuming for simplicity that U has a minimizer $\theta^\star$, we can define a control variates version of SGLD, SGLDFP, see [Dub+16; Che+17], given for $k \in \mathbb{N}$ by

$$\theta_{k+1} = \theta_k - \gamma\left(\nabla U_0(\theta_k) - \nabla U_0(\theta^\star) + \frac{N}{p}\sum_{i\in S_{k+1}}\{\nabla U_i(\theta_k) - \nabla U_i(\theta^\star)\}\right) + \sqrt{2\gamma}Z_{k+1}\,. \tag{7.5}$$

It is worth mentioning that the objectives of the different algorithms presented so far are distinct. On the one hand, LMC, SGLD and SGDLFP are MCMC methods used to obtain approximate samples from the posterior distribution π. On the other hand, SGD is a stochastic optimization algorithm used to find an estimate of the mode $\theta^\star$ of the posterior distribution. In this paper, we focus on the fixed step-size SGLD algorithm and assess its ability to reliably sample from π. For that purpose and to quantify precisely the relation between LMC, SGLD, SGDFP and SGD, we make for simplicity the following assumptions on U.

H 20. *For all $i \in \{0,\dots,N\}$, U_i is four times continuously differentiable and for all $j \in \{2,3,4\}$, $\sup_{\theta\in\mathbb{R}^d}\left\|\mathrm{D}^j U_i(\theta)\right\| \le \tilde{L}$. In particular for all $i \in \{0,\dots,N\}$, U_i is $\tilde{L}$-gradient Lipschitz, i.e. for all $\theta_1,\theta_2 \in \mathbb{R}^d$, $\|\nabla U_i(\theta_1) - \nabla U_i(\theta_2)\| \le \tilde{L}\,\|\theta_1 - \theta_2\|$.*

H 21. *U is m-strongly convex, i.e. for all $\theta_1,\theta_2 \in \mathbb{R}^d$, $\langle \nabla U(\theta_1) - \nabla U(\theta_2), \theta_1 - \theta_2\rangle \ge m\,\|\theta_1 - \theta_2\|^2$.*

H 22. *For all $i \in \{0,\dots,N\}$, U_i is convex.*

Note that under **H**20, U is four times continuously differentiable and for $j \in \{2,3,4\}$, $\sup_{\theta\in\mathbb{R}^d}\left\|\mathrm{D}^j U(\theta)\right\| \le L$, with $L = (N+1)\tilde{L}$ and where

$$\left\|\mathrm{D}^j U(\theta)\right\| = \sup_{\|u_1\|\le 1,\dots,\|u_j\|\le 1}\mathrm{D}^j U(\theta)[u_1,\dots,u_j]\,.$$

In particular, U is L-gradient Lipschitz and $m \leq L$. Furthermore, under **H**21, U has a unique minimizer $\theta^\star$. In this paper, we focus on the asymptotic $N \to +\infty$,. We assume that $\liminf_{N\to+\infty} N^{-1}m > 0$, which is a common assumption for the analysis of SGLD and SGLDFP [Bak+17; Cha+18]. In practice [ABW12; CFG14; Has+17; Li+16; SN14; VZT16], γ is of order $1/N$ and we adopt this convention in this article.

For a practical implementation of SGLDFP, an estimator $\hat{\theta}$ of $\theta^\star$ is necessary. The theoretical analysis and the bounds remain unchanged if, instead of considering SGLDFP centered w.r.t. $\theta^\star$, we study SGLDFP centered w.r.t. $\hat{\theta}$ satisfying $\mathbb{E}[\|\hat{\theta} - \theta^\star\|^2] = O(1/N)$. Such an estimator $\hat{\theta}$ can be computed using for example SGD with decreasing step sizes, see [Nem+09, eq.(2.8)] and [Bak+17, Section 3.4], for a computational cost linear in N.

7.3 Results

7.3.1 Analysis in Wasserstein distance

Before presenting the results, some notations and elements of Markov chain theory have to be introduced. Denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures with finite second moment and by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-algebra of $\mathbb{R}^d$. For $\lambda, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, define the Wasserstein distance of order 2 by

$$\mathrm{W}_2(\lambda, \nu) = \inf_{\xi \in \Pi(\lambda,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta - \vartheta\|^2 \, \xi(\mathrm{d}\theta, \mathrm{d}\vartheta) \right)^{1/2} ,$$

where $\Pi(\lambda, \nu)$ is the set of probability measures ξ on $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d)$ satisfying for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\xi(\mathsf{A} \times \mathbb{R}^d)) = \lambda(\mathsf{A})$ and $\xi(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})$.

A Markov kernel R on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ is a mapping $R : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ satisfying the following conditions: (i) for every $\theta \in \mathbb{R}^d$, $R(\theta, \cdot) : \mathsf{A} \mapsto R(\theta, \mathsf{A})$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$ (ii) for every $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $R(\cdot, A) : \theta \mapsto R(\theta, A)$ is a measurable function. For any probability measure λ on $\mathcal{B}(\mathbb{R}^d)$, we define λR for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by $\lambda R(\mathsf{A}) = \int_{\mathbb{R}^d} \lambda(\mathrm{d}\theta) R(\theta, \mathsf{A})$. For all $k \in \mathbb{N}^*$, we define the Markov kernel R^k recursively by $R^1 = R$ and for all $\theta \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $R^{k+1}(\theta, \mathsf{A}) = \int_{\mathbb{R}^d} R^k(\theta, \mathrm{d}\vartheta) R(\vartheta, \mathsf{A})$. A probability measure $\bar{\pi}$ is invariant for R if $\bar{\pi} R = \bar{\pi}$.

The LMC, SGLD, SGD and SGLDFP algorithms defined respectively by (7.2), (7.3), (7.4) and (7.5) are homogeneous Markov chains with Markov kernels denoted $R_{\mathrm{LMC}}, R_{\mathrm{SGLD}}, R_{\mathrm{SGD}}$, and R_{FP}. To avoid overloading the notations, the dependence on γ and N is implicit.

Lemma 7.1. *Assume **H**20, **H**21 and **H**22. For any step size $\gamma \in (0, 2/L)$, R_{SGLD} (respectively $R_{\mathrm{LMC}}, R_{\mathrm{SGD}}, R_{\mathrm{FP}}$) has a unique invariant measure $\pi_{\mathrm{SGLD}} \in \mathcal{P}_2(\mathbb{R}^d)$ (respectively $\pi_{\mathrm{LMC}}, \pi_{\mathrm{SGD}}, \pi_{\mathrm{FP}}$). In addition, for all $\gamma \in (0, 1/L]$, $\theta \in \mathbb{R}^d$ and $k \in \mathbb{N}$,*

$$\mathrm{W}_2^2(R_{\mathrm{SGLD}}^k(\theta, \cdot), \pi_{\mathrm{SGLD}}) \leq (1 - m\gamma)^k \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \, \pi_{\mathrm{SGLD}}(\mathrm{d}\vartheta)$$

and the same inequality holds for LMC, SGD and SGLDFP.

Proof. The proof is postponed to Section 7.A.1. □

Under **H** 20, (7.1) has a unique strong solution $(\theta_t)_{t\geq 0}$ for every initial condition $\theta_0 \in \mathbb{R}^d$ [KS91, Chapter 5, Theorems 2.5 and 2.9]. Denote by $(P_t)_{t\geq 0}$ the semigroup of the Langevin diffusion defined for all $\theta_0 \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by $P_t(\theta_0, \mathsf{A}) = \mathbb{P}(\theta_t \in \mathsf{A})$.

Theorem 7.2. *Assume **H**20, **H**21 and **H**22. For all $\gamma \in (0, 1/L]$, $\lambda, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $n \in \mathbb{N}$, we have the following upper-bounds in Wasserstein distance between*

i) LMC and SGLDFP,

$$\mathrm{W}_2^2(\lambda R^n_{\mathrm{LMC}}, \mu R^n_{\mathrm{FP}}) \leq (1 - m\gamma)^n \, \mathrm{W}_2^2(\lambda, \mu) + \frac{2L^2\gamma d}{pm^2} + \frac{L^2\gamma^2}{p} n(1 - m\gamma)^{n-1} \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \, \mu(\mathrm{d}\vartheta) \, ,$$

ii) the Langevin diffusion and LMC,

$$\mathrm{W}_2^2(\lambda R^n_{\mathrm{LMC}}, \mu P_{n\gamma}) \leq 2 \left(1 - \frac{mL\gamma}{m+L}\right)^n \mathrm{W}_2^2(\lambda, \mu) + d\gamma \frac{m+L}{2m} \left(3 + \frac{L}{m}\right) \left(\frac{13}{6} + \frac{L}{m}\right) + n\mathrm{e}^{-(m/2)\gamma(n-1)} L^3\gamma^3 \left(1 + \frac{m+L}{2m}\right) \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \, \mu(\mathrm{d}\vartheta) \, ,$$

iii) SGLD and SGD,

$$\mathrm{W}_2^2(\lambda R^n_{\mathrm{SGLD}}, \mu R^n_{\mathrm{SGD}}) \leq (1 - m\gamma)^n \, \mathrm{W}_2^2(\lambda, \mu) + (2d)/m \, .$$

Proof. The proof is postponed to Section 7.A.2. □

The two last terms in the right hand side of Theorem 7.2-i) come from the approximation by minibatch of the gradient ∇U. Theorem 7.2-ii) is a direct application of [DM16, Theorem 5]. Finally, $2d/m$ in Theorem 7.2-iii) originates from the addition of the Gaussian noise $\sqrt{2\gamma}Z$ at each step of the algorithm.

Corollary 7.3. *Assume **H**20, **H**21 and **H**22. Set $\gamma = \eta/N$ with $\eta \in (0, 1/(2\tilde{L})]$ and assume that $\liminf_{N\to\infty} mN^{-1} > 0$. Then,*

i) for all $n \in \mathbb{N}$, we get $\mathrm{W}_2(R^n_{\mathrm{LMC}}(\theta^\star, \cdot), R^n_{\mathrm{FP}}(\theta^\star, \cdot)) = \sqrt{d\eta}\, O(N^{-1/2})$ and $\mathrm{W}_2(\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}) = \sqrt{d\eta}\, O(N^{-1/2})$, $\mathrm{W}_2(\pi_{\mathrm{LMC}}, \pi) = \sqrt{d\eta}\, O(N^{-1/2})$.

ii) for all $n \in \mathbb{N}$, we get $\mathrm{W}_2(R^n_{\mathrm{SGLD}}(\theta^\star, \cdot), R^n_{\mathrm{SGD}}(\theta^\star, \cdot)) = \sqrt{d}\, O(N^{-1/2})$ and $\mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi_{\mathrm{SGD}}) = \sqrt{d}\, O(N^{-1/2})$.

Theorem 7.2 implies that the number of iterations necessary to obtain a sample ε-close from π in Wasserstein distance is the same for LMC and SGLDFP. However for LMC, the cost of one iteration is Nd which is larger than pd the cost of one iteration for

SGLDFP. In other words, to obtain an approximate sample from the target distribution at an accuracy $O(1/\sqrt{N})$ in 2-Wasserstein distance, LMC requires $\Theta(N)$ operations, in contrast with SGLDFP that needs only $\Theta(1)$ operations.

We show in the sequel that $\mathrm{W}_2(\pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}) = \Omega(1)$ when $N \to +\infty$ in the case of a Bayesian linear regression, where for two sequences $(u_N)_{N\geq 1}$, $(v_N)_{N\geq 1}$, $u_N = \Omega(v_N)$ if $\liminf_{N\to+\infty} u_N/v_N > 0$. The dataset is $\mathbf{z} = \{(y_i, x_i)\}_{i=1}^N$ where $y_i \in \mathbb{R}$ is the response variable and $x_i \in \mathbb{R}^d$ are the covariates. Set $\mathbf{y} = (y_1, \dots, y_N) \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{N\times d}$ the matrix of covariates such that the i^{th} row of $\mathbf{X}$ is x_i. Let $\sigma_y^2, \sigma_\theta^2 > 0$. For $i \in \{1, \dots, N\}$, the conditional distribution of y_i given x_i is Gaussian with mean $x_i^{\mathrm{T}}\theta$ and variance σ_y^2. The prior $\pi_0(\theta)$ is a normal distribution of mean 0 and variance $\sigma_\theta^2\,\mathrm{Id}$. The posterior distribution π is then proportional to $\pi(\theta) \propto \exp\left(-(1/2)(\theta-\theta^\star)^{\mathrm{T}}\Sigma(\theta-\theta^\star)\right)$ where

$$\Sigma = \mathrm{Id}\,/\sigma_\theta^2 + \mathbf{X}^{\mathrm{T}}\mathbf{X}/\sigma_y^2 \quad \text{and} \quad \theta^\star = \Sigma^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y})/\sigma_y^2\,.$$

We assume that $\mathbf{X}^{\mathrm{T}}\mathbf{X} \succeq m\,\mathrm{Id}$, with $\liminf_{N\to+\infty} m/N > 0$. Let S be a minibatch of $\{1, \dots, N\}$ with replacement of size p. Define

$$\nabla U_0(\theta) + (N/p)\sum_{i\in S}\nabla U_i(\theta) = \Sigma(\theta-\theta^\star) + \rho(S)(\theta-\theta^\star) + \xi(S)$$

where

$$\rho(S) = \frac{\mathrm{Id}}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S} x_i x_i^{\mathrm{T}} - \Sigma\,,\ \xi(S) = \frac{\theta^\star}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S}\left(x_i^{\mathrm{T}}\theta^\star - y_i\right)x_i\,. \tag{7.6}$$

$\rho(S)(\theta-\theta^\star)$ is the multiplicative part of the noise in the stochastic gradient, and $\xi(S)$ the additive part that does not depend on θ. The additive part of the stochastic gradient for SGLDFP disappears since

$$\nabla U_0(\theta) - \nabla U_0(\theta^\star) + (N/p)\sum_{i\in S}\{\nabla U_i(\theta) - \nabla U_i(\theta^\star)\} = \Sigma(\theta-\theta^\star) + \rho(S)(\theta-\theta^\star)\,.$$

In this setting, the following theorem shows that the Wasserstein distances between the marginal distribution of the iterates of SGLD and SGLDFP, and π_{SGLD} and π, is of order $\Omega(1)$ when $N \to +\infty$. This is in sharp contrast with the results of Corollary 7.3 where the Wasserstein distances tend to 0 as $N \to +\infty$ at a rate $N^{-1/2}$. For simplicity, we state the result for $d = 1$.

Theorem 7.4. *Consider the case of the Bayesian linear regression in dimension* 1.

i) For all $\gamma \in (0, \Sigma^{-1}\{1 + N/(p\sum_{i=1}^N x_i^2)\}^{-1}]$ *and* $n \in \mathbb{N}^*$,

$$\left(\frac{1-\mu}{1-\mu^n}\right)^{1/2}\mathrm{W}_2(R^n_{\mathrm{SGLD}}(\theta^\star,\cdot), R^n_{\mathrm{FP}}(\theta^\star,\cdot)) \geq \left\{2\gamma + \frac{\gamma^2 N}{p}\sum_{i=1}^N\left(\frac{(x_i\theta^\star - y_i)x_i}{\sigma_y^2} + \frac{\theta^\star}{N\sigma_\theta^2}\right)^2\right\}^{1/2} - \sqrt{2\gamma}\,,$$

where $\mu \in (0, 1-\gamma\Sigma]$.

ii) Set $\gamma = \eta/N$ *with* $\eta \in (0, \liminf_{N\to+\infty} N\Sigma^{-1}\{1 + N/(p\sum_{i=1}^N x_i^2)\}^{-1}]$ *and assume that* $\liminf_{N\to+\infty} N^{-1}\sum_{i=1}^N x_i^2 > 0$. *We have* $\mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi) = \Omega(1)$.

Proof. The proof is postponed to Section 7.A.3. □

The study in Wasserstein distance emphasizes the different behaviors of the LMC, SGLDFP, SGLD and SGD algorithms. When $N \to \infty$ and $\lim_{N\to+\infty} m/N > 0$, the marginal distributions of the k^{th} iterates of the LMC and SGLDFP algorithm are very close to the Langevin diffusion and their invariant probability measures π_{LMC} and π_{FP} are similar to the posterior distribution of interest π. In contrast, the marginal distributions of the k^{th} iterates of SGLD and SGD are analogous and their invariant probability measures π_{SGLD} and π_{SGD} are very different from π when $N \to +\infty$.

Note that to fix the asymptotic bias of SGLD, other strategies can be considered: choosing a step size $\gamma \propto N^{-\beta}$ where $\beta > 1$ and/or increasing the batch size $p \propto N^{\alpha}$ where $\alpha \in [0, 1]$. Using the Wasserstein (of order 2) bounds of SGLD w.r.t. the target distribution π, see e.g. [DK17, Theorem 3], $\alpha + \beta$ should be equal to 2 to guarantee the ε-accuracy in Wasserstein distance of SGLD for a cost proportional to N (up to logarithmic terms), independently of the choice of α and β.

7.3.2 Mean and covariance matrix of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$

We now establish an expansion of the mean and second moments of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD} as $N \to +\infty$, and compare them. We first give an expansion of the mean and second moments of π as $N \to +\infty$.

Proposition 7.5. *Assume* ***H****20 and* ***H****21 and that* $\liminf_{N\to+\infty} N^{-1}m > 0$. *Then,*

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2}\pi(\mathrm{d}\theta) = \nabla^2 U(\theta^\star)^{-1} + O_{N\to+\infty}(N^{-3/2}) \,,$$

$$\int_{\mathbb{R}^d} \theta\,\pi(\mathrm{d}\theta) - \theta^\star = -(1/2)\nabla^2 U(\theta^\star)^{-1}\,\mathrm{D}^3 U(\theta^\star)[\nabla^2 U(\theta^\star)^{-1}] + O_{N\to+\infty}(N^{-3/2}) \,.$$

Proof. The proof is postponed to Section 7.B.1. □

Contrary to the Bayesian linear regression where the covariance matrices can be explicitly computed, see Section 7.C, only approximate expressions are available in the general case. For that purpose, we consider two types of asymptotics. For LMC and SGLDFP, we assume that $\lim_{N\to+\infty} m/N > 0$, $\gamma = \eta/N$, for $\eta > 0$, and we develop an asymptotic when $N \to +\infty$. Combining Proposition 7.5 and Theorem 7.6 , we show that the biases and covariance matrices of π_{LMC} and π_{FP} are of order $\Theta(1/N)$ with remainder terms of the form $O(N^{-3/2})$, where for two sequences $(u_N)_{N\geq 1}$, $(v_N)_{N\geq 1}$, $u = \Theta(v)$ if $0 < \liminf_{N\to+\infty} u_N/v_N \leq \limsup_{N\to+\infty} u_N/v_N < +\infty$.

Regarding SGD and SGLD, we do not have such concentration properties when $N \to +\infty$ because of the high variance of the stochastic gradients. The biases and covariance matrices of SGLD and SGD are of order $\Theta(1)$ when $N \to +\infty$. To obtain approximate expressions of these quantities, we set $\gamma = \eta/N$ where $\eta > 0$ is the step

size for the gradient descent over the normalized potential U/N. Assuming that m is proportional to N and $N \geq 1/\eta$, we show by combining Proposition 7.5 and Theorem 7.7 that the biases and covariance matrices of SGLD and SGD are of order $\Theta(\eta)$ with remainder terms of the form $O(\eta^{3/2})$ when $\eta \to 0$.

Before giving the results associated to $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD}, we need to introduce some notations. For any matrices $A_1, A_2 \in \mathbb{R}^{d\times d}$, we denote by $A_1 \otimes A_2$ the Kronecker product defined on $\mathbb{R}^{d\times d}$ by $A_1 \otimes A_2 : Q \mapsto A_1 Q A_2$ and $A^{\otimes 2} = A \otimes A$. Besides, for all $\theta_1 \in \mathbb{R}^d$ and $\theta_2 \in \mathbb{R}^d$, we denote by $\theta_1 \otimes \theta_2 \in \mathbb{R}^{d\times d}$ the tensor product of θ_1 and θ_2. For any matrix $A \in \mathbb{R}^{d\times d}$, $\mathrm{Tr}(A)$ is the trace of A.

Define $\mathrm{K} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ for all $A \in \mathbb{R}^{d\times d}$ by

$$\mathrm{K}(A) = \frac{N}{p} \sum_{i=1}^{N} \left(\nabla^2 U_i(\theta^\star) - \frac{1}{N} \sum_{j=1}^{N} \nabla^2 U_j(\theta^\star) \right)^{\otimes 2} A \,, \tag{7.7}$$

and H and $\mathrm{G} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ by

$$\mathrm{H} = \nabla^2 U(\theta^\star) \otimes \mathrm{Id} + \mathrm{Id} \otimes \nabla^2 U(\theta^\star) - \gamma \nabla^2 U(\theta^\star) \otimes \nabla^2 U(\theta^\star) \,, \tag{7.8}$$

$$\mathrm{G} = \nabla^2 U(\theta^\star) \otimes \mathrm{Id} + \mathrm{Id} \otimes \nabla^2 U(\theta^\star) - \gamma (\nabla^2 U(\theta^\star) \otimes \nabla^2 U(\theta^\star) + \mathrm{K}) \,. \tag{7.9}$$

K, H and G can be interpreted as perturbations of $\nabla^2 U(\theta^\star)^{\otimes 2}$ and $\nabla^2 U(\theta^\star)$, respectively, due to the noise of the stochastic gradients. It can be shown, see Section 7.B.2, that for γ small enough, H and G are invertible.

Theorem 7.6. *Assume **H**20, **H**21 and **H**22. Set $\gamma = \eta/N$ and assume that $\liminf_{N\to+\infty} N^{-1} m > 0$. There exists an (explicit) η_0 independent of N such that for all $\eta \in (0, \eta_0)$,*

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\mathrm{LMC}}(\mathrm{d}\theta) = \mathrm{H}^{-1}(2\,\mathrm{Id}) + O_{N\to+\infty}(N^{-3/2}) \,, \tag{7.10}$$

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\mathrm{FP}}(\mathrm{d}\theta) = \mathrm{G}^{-1}(2\,\mathrm{Id}) + O_{N\to+\infty}(N^{-3/2}) \,, \tag{7.11}$$

and

$$\int_{\mathbb{R}^d} \theta \pi_{\mathrm{LMC}}(\mathrm{d}\theta) - \theta^\star = -\nabla^2 U(\theta^\star)^{-1}\, \mathrm{D}^3 U(\theta^\star)[\mathrm{H}^{-1}\,\mathrm{Id}] + O_{N\to+\infty}(N^{-3/2}) \,,$$

$$\int_{\mathbb{R}^d} \theta \pi_{\mathrm{FP}}(\mathrm{d}\theta) - \theta^\star = -\nabla^2 U(\theta^\star)^{-1}\, \mathrm{D}^3 U(\theta^\star)[\mathrm{G}^{-1}\,\mathrm{Id}] + O_{N\to+\infty}(N^{-3/2}) \,.$$

Proof. The proof is postponed to Section 7.B.2. □

Theorem 7.7. *Assume **H**20, **H**21 and **H**22. Set $\gamma = \eta/N$ and assume that $\liminf_{N\to+\infty} N^{-1} m > 0$. There exists an (explicit) η_0 independent of N such that for all $\eta \in (0, \eta_0)$ and $N \geq 1/\eta$,*

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\mathrm{SGLD}}(\mathrm{d}\theta) = \mathrm{G}^{-1} \{2\,\mathrm{Id} + (\eta/p)\,\mathrm{M}\} + O_{\eta\to 0}(\eta^{3/2}) \,, \tag{7.12}$$

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\mathrm{SGD}}(\mathrm{d}\theta) = (\eta/p)\,\mathrm{G}^{-1}\,\mathrm{M} + O_{\eta\to 0}(\eta^{3/2}) \,, \tag{7.13}$$

and

$$\int_{\mathbb{R}^d} \theta \pi_{\mathrm{SGLD}}(\mathrm{d}\theta) - \theta^\star = -(1/2)\nabla^2 U(\theta^\star)^{-1}\,\mathrm{D}^3\,U(\theta^\star)[\mathrm{G}^{-1}\,\{2\,\mathrm{Id} + (\eta/p)\,\mathrm{M}\}] + O_{\eta\to 0}(\eta^{3/2})\;,$$
$$\int_{\mathbb{R}^d} \theta \pi_{\mathrm{SGD}}(\mathrm{d}\theta) - \theta^\star = -(\eta/2p)\nabla^2 U(\theta^\star)^{-1}\,\mathrm{D}^3\,U(\theta^\star)[\mathrm{G}^{-1}\,\mathrm{M}] + O_{\eta\to 0}(\eta^{3/2})\;,$$

where

$$\mathrm{M} = \sum_{i=1}^N \left(\nabla U_i(\theta^\star) - \frac{1}{N}\sum_{j=1}^N \nabla U_j(\theta^\star) \right)^{\otimes 2}\;, \tag{7.14}$$

and G *is defined in* (7.9).

Proof. The proof is postponed to Section 7.B.2. □

Note that this result implies that the mean and the covariance matrix of π_{SGLD} and π_{SGD} stay lower bounded by a positive constant for any $\eta > 0$ as $N \to +\infty$. Figure 7.1 illustrates the results of Theorem 7.6 and Theorem 7.7 in the asymptotic $N \to +\infty$.

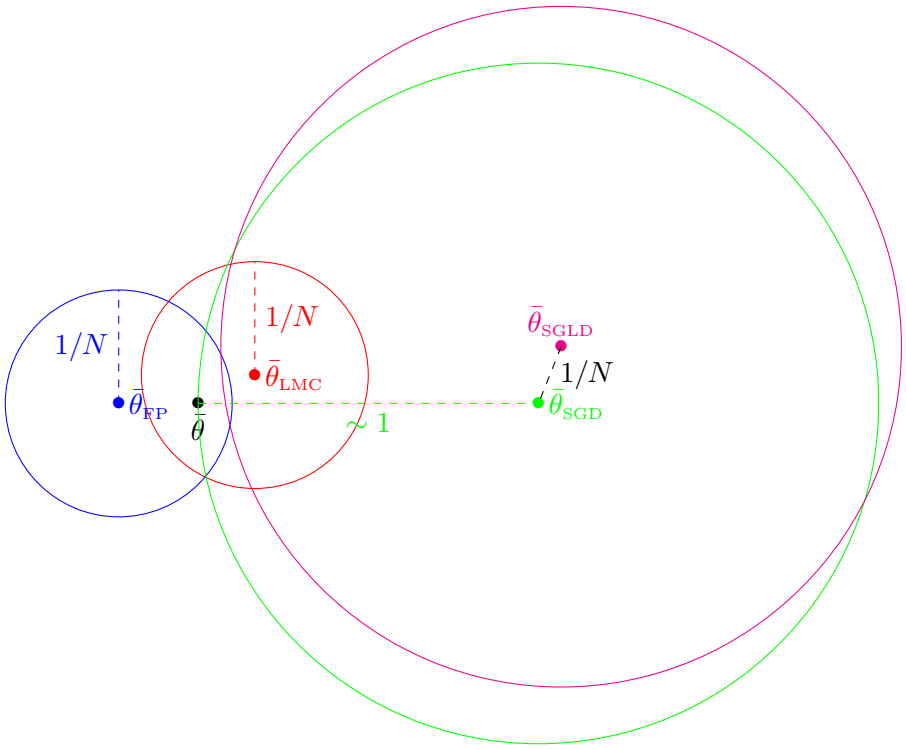

Figure 7.1: Illustration of Proposition 7.5, Theorem 7.6 and Theorem 7.7 in the asymptotic $N \to +\infty$. $\bar{\theta}$, $\bar{\theta}_{\text{SGD}}$, $\bar{\theta}_{\text{LMC}}$, $\bar{\theta}_{\text{FP}}$ and $\bar{\theta}_{\text{SGLD}}$ are the means under the stationary distributions π, π_{SGD}, π_{LMC}, π_{FP} and π_{SGLD}, respectively. The associated circles indicate the order of magnitude of the covariance matrix. While LMC and SGLDFP concentrate to the posterior mean $\bar{\theta}$ with a covariance matrix of the order $1/N$, SGLD and SGD are at a distance of order ~ 1 of $\bar{\theta}$ and do not concentrate as $N \to +\infty$.

7.4 Numerical experiments

Simulated data For illustrative purposes, we consider a Bayesian logistic regression in dimension $d = 2$. We simulate $N = 10^5$ covariates $\{x_i\}_{i=1}^N$ drawn from a standard 2-dimensional Gaussian distribution and we denote by $\mathbf{X} \in \mathbb{R}^{N \times d}$ the matrix of covariates such that the i^{th} row of $\mathbf{X}$ is x_i. Our Bayesian regression model is specified by a Gaussian prior of mean 0 and identity covariance matrix, and a likelihood given for $y_i \in \{0, 1\}$ by $p(y_i|x_i, \theta) = (1 + \mathrm{e}^{-x_i^{\mathrm{T}}\theta})^{-y_i}(1 + \mathrm{e}^{x_i^{\mathrm{T}}\theta})^{y_i - 1}$. We simulate N observations $\{y_i\}_{i=1}^N$ under this model. In this setting, **H**20 and **H**22 are satisfied, and **H**21 holds if the state space is compact.

To illustrate the results of Section 7.3.2, we consider 10 regularly spaced values of N between 10^2 and 10^5 and we truncate the dataset accordingly. We compute an estimator $\hat{\theta}$ of $\theta^\star$ using SGD [Ped+11] combined with the BFGS algorithm [J+01]. For the LMC, SGLDFP, SGLD and SGD algorithms, the step size γ is set equal to $(1 + \delta/4)^{-1}$ where δ is the largest eigenvalue of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. We start the algorithms at $\theta_0 = \hat{\theta}$ and run $n = 1/\gamma$ iterations where the first 10% samples are discarded as a burn-in period.

We estimate the means and covariance matrices of $\pi_{\text{LMC}}, \pi_{\text{FP}}, \pi_{\text{SGLD}}$ and π_{SGD} by their empirical averages $\bar{\theta}_n = (1/n)\sum_{k=0}^{n-1}\theta_k$ and $\{1/(n-1)\}\sum_{k=0}^{n-1}(\theta_k - \bar{\theta}_n)^{\otimes 2}$. We plot the mean and the trace of the covariance matrices for the different algorithms, averaged over 100 independent trajectories, in Figure 7.2 and Figure 7.3 in logarithmic scale.

The slope for LMC and SGLDFP is -1 which confirms the convergence of $\left\|\bar{\theta}_n - \theta^\star\right\|$ to 0 at a rate N^{-1}. On the other hand, we can observe that $\left\|\bar{\theta}_n - \theta^\star\right\|$ converges to a constant for SGD and SGLD.

Covertype dataset We then illustrate our results on the covertype dataset[1] with a Bayesian logistic regression model. The prior is a standard multivariate Gaussian distribution. Given the size of the dataset and the dimension of the problem, LMC requires high computational resources and is not included in the simulations. We truncate the training dataset at $N \in \{10^3, 10^4, 10^5\}$. For all algorithms, the step size γ is set equal to $1/N$ and the trajectories are started at $\hat{\theta}$, an estimator of $\theta^\star$, computed using SGD combined with the BFGS algorithm.

We empirically check that the variance of the stochastic gradients scale as N^2 for SGD and SGLD, and as N for SGLDFP. We compute the empirical variance estimator of the gradients, take the mean over the dimension and display the result in a logarithmic plot in Figure 7.4. The slopes are 2 for SGD and SGLD, and 1 for SGLDFP.

On the test dataset, we also evaluate the negative loglikelihood of the three algorithms for different values of $N \in \{10^3, 10^4, 10^5\}$, as a function of the number of iterations. The plots are shown in Figure 7.5. We note that for large N, SGLD and SGD give very similar results that are below the performance of SGLDFP.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/covtype.libsvm.binary.scale.bz2

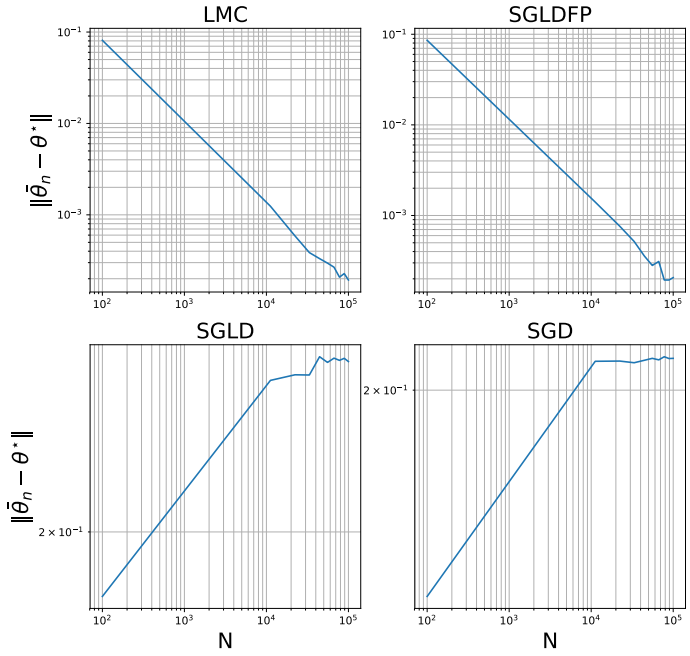

Figure 7.2: Distance to $\theta^\star$, $\left\|\bar{\theta}_n - \theta^\star\right\|$ for LMC, SGLDFP, SGLD and SGD, function of N, in logarithmic scale.

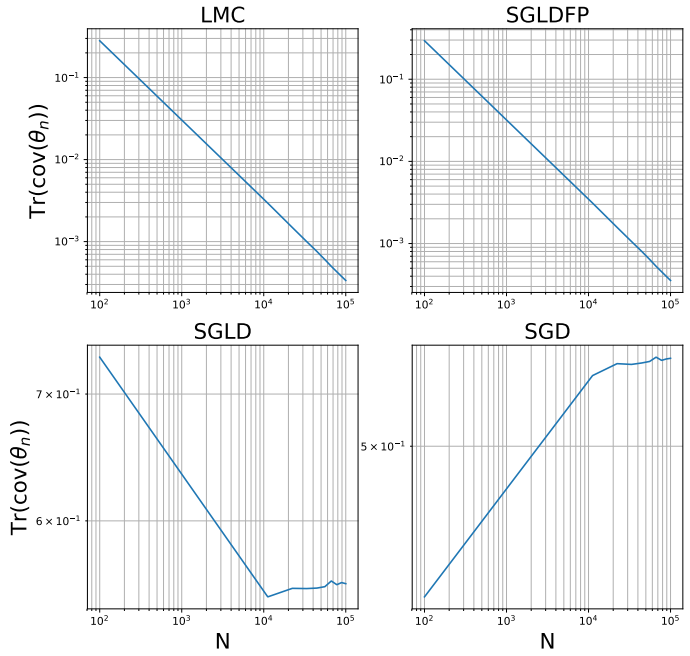

Figure 7.3: Trace of the covariance matrices for LMC, SGLDFP, SGLD and SGD, function of N, in logarithmic scale.

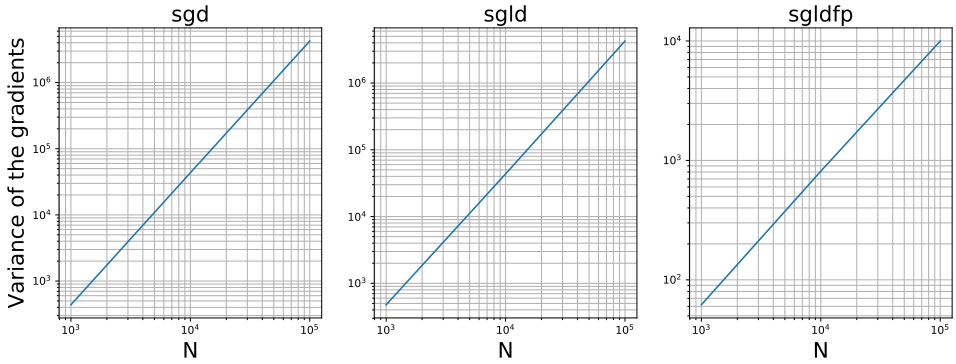

Figure 7.4: Variance of the stochastic gradients of SGLD, SGLDFP and SGD function of N, in logarithmic scale.

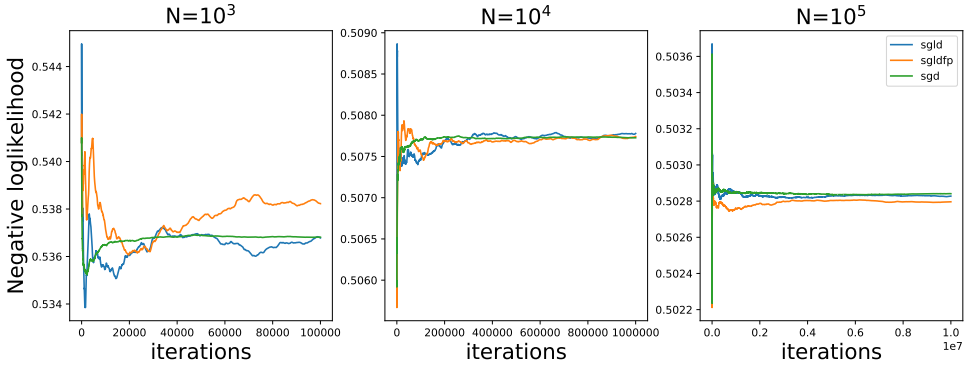

Figure 7.5: Negative loglikelihood on the test dataset for SGLD, SGLDFP and SGD function of the number of iterations for different values of $N \in \{10^3, 10^4, 10^5\}$.

7.A Proofs of Section 7.3.1

7.A.1 Proof of Lemma 7.1

The convergence in Wasserstein distance is classically done via a standard synchronous coupling [DDB17, Proposition 2]. We prove the statement for SGLD; the adaptation for LMC, SGLDFP and SGD is immediate. Let $\gamma \in (0, 2/L)$ and $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By [Vil09, Theorem 4.1], there exists a couple of random variables $(\theta_0^{(1)}, \theta_0^{(2)})$ such that $\mathrm{W}_2^2(\lambda_1, \lambda_2) = \mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|^2\right]$. Let $(\theta_k^{(1)}, \theta_k^{(2)})_{k\in\mathbb{N}}$ be the SGLD iterates starting from $\theta_0^{(1)}$ and $\theta_0^{(2)}$ respectively and driven by the same noise, i.e. for all $k \in \mathbb{N}$,

$$\begin{cases} \theta_{k+1}^{(1)} &= \theta_k^{(1)} - \gamma \left\{ \nabla U_0(\theta_k^{(1)}) + (N/p) \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(1)}) \right\} + \sqrt{2\gamma} Z_{k+1} \;, \\ \theta_{k+1}^{(2)} &= \theta_k^{(2)} - \gamma \left\{ \nabla U_0(\theta_k^{(2)}) + (N/p) \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(2)}) \right\} + \sqrt{2\gamma} Z_{k+1} \;, \end{cases}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of standard Gaussian variables and $(S_k)_{k\geq 1}$ an i.i.d. sequence of subsamples of $\{1, \dots, N\}$ of size p. Denote by $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k^{(1)}, \theta_k^{(2)})_{k\in\mathbb{N}}$. We have for $k \in \mathbb{N}$,

$$\begin{aligned} &\left\|\theta_{k+1}^{(1)} - \theta_{k+1}^{(2)}\right\|^2 = \\ &\left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|^2 + \gamma^2 \left\| \nabla U_0(\theta_k^{(1)}) + \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(1)}) - \nabla U_0(\theta_k^{(2)}) - \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(2)}) \right\|^2 \\ &- 2\gamma \left\langle \theta_k^{(1)} - \theta_k^{(2)}, \nabla U_0(\theta_k^{(1)}) + \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(1)}) - \nabla U_0(\theta_k^{(2)}) - \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(2)}) \right\rangle \;. \end{aligned}$$

By **H** 20 and **H** 22, $\theta \mapsto \nabla U_0(\theta) + (N/p) \sum_{i\in S} \nabla U_i(\theta)$ is $\mathbb{P}$-a.s L-co-coercive [ZM96]. Taking the conditional expectation w.r.t. $\mathcal{F}_k$, we obtain

$$\mathbb{E}\left[\left\|\theta_{k+1}^{(1)} - \theta_{k+1}^{(2)}\right\|^2 \middle| \mathcal{F}_k\right] \leq \left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|^2 - 2\gamma \left\{1 - (\gamma L)/2\right\} \left\langle \theta_k^{(1)} - \theta_k^{(2)}, \nabla U(\theta_k^{(1)}) - \nabla U(\theta_k^{(2)}) \right\rangle \;,$$

and by **H**21

$$\mathbb{E}\left[\left\|\theta_{k+1}^{(1)} - \theta_{k+1}^{(2)}\right\|^2 \middle| \mathcal{F}_k\right] \leq \left\{1 - 2m\gamma(1 - (\gamma L)/2)\right\} \left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|^2 \;.$$

Since for all $k \geq 0$, $(\theta_k^{(1)}, \theta_k^{(2)})$ belongs to $\Pi(\lambda_1 R_{\mathrm{SGLD}}^k, \lambda_2 R_{\mathrm{SGLD}}^k)$, we get by a straightforward induction

$$\mathrm{W}_2^2(\lambda_1 R_{\mathrm{SGLD}}^k, \lambda_2 R_{\mathrm{SGLD}}^k) \leq \mathbb{E}\left[\left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|^2\right] \leq \left\{1 - 2m\gamma(1 - (\gamma L)/2)\right\}^k \mathrm{W}_2^2(\lambda_1, \lambda_2) \;. \tag{7.15}$$

By **H**20, $\lambda_1 R_{\mathrm{SGLD}} \in \mathcal{P}_2(\mathbb{R}^d)$ and taking $\lambda_2 = \lambda_1 R_{\mathrm{SGLD}}$, we get $\sum_{k=0}^{+\infty} \mathrm{W}_2^2(\lambda_1 R_{\mathrm{SGLD}}^k, \lambda_1 R_{\mathrm{SGLD}}^{k+1}) < +\infty$.By [Vil09, Theorem 6.16], $\mathcal{P}_2(\mathbb{R}^d)$ endowed with W_2 is a Polish space. $(\lambda_1 R_{\mathrm{SGLD}}^k)_{k\geq 0}$

is a Cauchy sequence and converges to a limit $\pi^{\lambda_1}_{\mathrm{SGLD}} \in \mathcal{P}_2(\mathbb{R}^d)$. The limit $\pi^{\lambda_1}_{\mathrm{SGLD}}$ does not depend on λ_1 because, given $\lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$, by the triangle inequality

$$\mathrm{W}_2(\pi^{\lambda_1}_{\mathrm{SGLD}}, \pi^{\lambda_2}_{\mathrm{SGLD}}) \leq \mathrm{W}_2(\pi^{\lambda_1}_{\mathrm{SGLD}}, \lambda_1 R^k_{\mathrm{SGLD}}) + \mathrm{W}_2(\lambda_1 R^k_{\mathrm{SGLD}}, \lambda_2 R^k_{\mathrm{SGLD}}) + \mathrm{W}_2(\pi^{\lambda_2}_{\mathrm{SGLD}}, \lambda_2 R^k_{\mathrm{SGLD}}) \ .$$

Taking the limit $k \to +\infty$, we get $\mathrm{W}_2(\pi^{\lambda_1}_{\mathrm{SGLD}}, \pi^{\lambda_2}_{\mathrm{SGLD}}) = 0$. The limit is thus the same for all initial distributions and is denoted π_{SGLD}. π_{SGLD} is invariant for R_{SGLD} since we have for all $k \in \mathbb{N}^*$,

$$\mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi_{\mathrm{SGLD}} R_{\mathrm{SGLD}}) \leq \mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi_{\mathrm{SGLD}} R^k_{\mathrm{SGLD}}) + \mathrm{W}_2(\pi_{\mathrm{SGLD}} R_{\mathrm{SGLD}}, \pi_{\mathrm{SGLD}} R^k_{\mathrm{SGLD}}) \ .$$

Taking the limit $k \to +\infty$, we obtain $\mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi_{\mathrm{SGLD}} R_{\mathrm{SGLD}}) = 0$. Using (7.15), π_{SGLD} is the unique invariant probability measure for R_{SGLD}.

7.A.2 Proof of Theorem 7.2

Proof of i). Let $\gamma \in (0, 1/L]$ and $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By [Vil09, Theorem 4.1], there exists a couple of random variables (θ_0, ϑ_0) such that $\mathrm{W}_2^2(\lambda_1, \lambda_2) = \mathbb{E}\left[\|\theta_0 - \vartheta_0\|^2\right]$. Let $(\theta_k, \vartheta_k)_{k\in\mathbb{N}}$ be the LMC and SGLDFP iterates starting from θ_0 and ϑ_0 respectively and driven by the same noise, i.e. for all $k \in \mathbb{N}$,

$$\begin{cases} \theta_{k+1} &= \theta_k - \gamma \nabla U(\theta_k) + \sqrt{2\gamma} Z_{k+1} \ , \\ \vartheta_{k+1} &= \vartheta_k - \gamma \left(\nabla U_0(\vartheta_k) - \nabla U_0(\theta^\star) + (N/p) \sum_{i \in S_{k+1}} \{\nabla U_i(\vartheta_k) - \nabla U_i(\theta^\star)\} \right) + \sqrt{2\gamma} Z_{k+1} \ , \end{cases}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of standard Gaussian variables and $(S_k)_{k\geq 1}$ an i.i.d. sequence of subsamples with replacement of $\{1, \dots, N\}$ of size p. Denote by $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k, \vartheta_k)_{k\in\mathbb{N}}$. We have for $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_{k+1} - \vartheta_{k+1}\|^2 \middle| \mathcal{F}_k\right] = \|\theta_k - \vartheta_k\|^2 - 2\gamma \langle \theta_k - \vartheta_k, \nabla U(\theta_k) - \nabla U(\vartheta_k) \rangle + \gamma^2 A \quad (7.16)$$

where

$$\begin{aligned} A &= \mathbb{E}\left[\left\| \nabla U(\theta_k) - \left(\nabla U_0(\vartheta_k) - \nabla U_0(\theta^\star) + (N/p) \sum_{i \in S_{k+1}} \{\nabla U_i(\vartheta_k) - \nabla U_i(\theta^\star)\} \right) \right\|^2 \middle| \mathcal{F}_k \right] \\ &= A_1 + A_2 \ , \\ A_1 &= \|\nabla U(\theta_k) - \nabla U(\vartheta_k)\|^2 \ , \\ A_2 &= \mathbb{E}\left[\left\| \nabla U(\vartheta_k) - \left(\nabla U_0(\vartheta_k) - \nabla U_0(\theta^\star) + (N/p) \sum_{i \in S_{k+1}} \{\nabla U_i(\vartheta_k) - \nabla U_i(\theta^\star)\} \right) \right\|^2 \middle| \mathcal{F}_k \right] \ . \end{aligned}$$

Denote by W the random variable equal to $\nabla U_i(\vartheta_k) - \nabla U_i(\theta^\star) - (1/N) \sum_{j=1}^N \{\nabla U_j(\vartheta_k) - \nabla U_j(\theta^\star)\}$ for $i \in \{1, \dots, N\}$ with probability $1/N$. By **H**20 and using the fact that the subsamples $(S_k)_{k\geq 1}$ are drawn with replacement, we obtain

$$A_2 = (N^2/p) \mathbb{E}\left[\|W\|^2 | \mathcal{F}_k\right] \leq (L^2/p) \|\vartheta_k - \theta^\star\|^2 \ .$$

Combining it with (7.16), and using the L-co-coercivity of ∇U under **H**20 and **H**21, we get

$$\mathbb{E}\left[\|\theta_{k+1} - \vartheta_{k+1}\|^2 \middle| \mathcal{F}_k\right] \leq (1 - m\gamma)\,\|\theta_k - \vartheta_k\|^2 + \{(L^2\gamma^2)/p\}\,\|\vartheta_k - \theta^\star\|^2 \ .$$

Iterating and using Lemma 7.8-i), we have for $n \in \mathbb{N}$

$$\begin{aligned}
&\mathrm{W}_2^2(\lambda_1 R_{\mathrm{LMC}}^n, \lambda_2 R_{\mathrm{FP}}^n) \leq \mathbb{E}\left[\|\theta_n - \vartheta_n\|^2\right] \\
&\leq (1 - m\gamma)^n\,\mathrm{W}_2^2(\lambda_1, \lambda_2) + \frac{L^2\gamma^2}{p}\sum_{k=0}^{n-1}(1 - m\gamma)^{n-1-k}\mathbb{E}\left[\|\vartheta_k - \theta^\star\|^2\right] \\
&\leq (1 - m\gamma)^n\,\mathrm{W}_2^2(\lambda_1, \lambda_2) + \frac{L^2\gamma^2}{p} n(1 - m\gamma)^{n-1}\int_{\mathbb{R}^d}\|\vartheta - \theta^\star\|^2\,\lambda_2(\mathrm{d}\vartheta) + \frac{2L^2\gamma d}{pm^2} \ .
\end{aligned}$$

Proof of ii). Denote by $\kappa = (2mL)/(m + L)$. By **H**20, **H**21 and [DM16, Theorem 5], we have for all $n \in \mathbb{N}$,

$$\begin{aligned}
\mathrm{W}_2^2(\lambda_1 P_{n\gamma}, \lambda_2 R_{\mathrm{LMC}}^n) \leq 2\,(1 - \kappa\gamma/2)^n\,\mathrm{W}_2^2(\lambda_1, \lambda_2) + \frac{2L^2\gamma}{\kappa}(\kappa^{-1} + \gamma)\left(2d + \frac{dL^2\gamma^2}{6}\right) \\
+ L^4\gamma^3(\kappa^{-1} + \gamma)\sum_{k=1}^{n}\delta_k\,\{1 - \kappa\gamma/2\}^{n-k}
\end{aligned}$$

where for all $k \in \{1, \ldots, n\}$,

$$\delta_k \leq \mathrm{e}^{-2m(k-1)\gamma}\int_{\mathbb{R}^d}\|\vartheta - \theta^\star\|^2\,\lambda_1(\mathrm{d}\vartheta) + d/m \ .$$

We get the result by straightforward simplifications and using $\gamma \leq 1/L$.

Proof of iii). Let $\gamma \in (0, 1/L]$ and $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By [Vil09, Thereom 4.1], there exists a couple of random variables (θ_0, ϑ_0) such that $\mathrm{W}_2^2(\lambda_1, \lambda_2) = \mathbb{E}\left[\|\theta_0 - \vartheta_0\|^2\right]$. Let $(\theta_k, \vartheta_k)_{k\in\mathbb{N}}$ be the SGLD and SGD iterates starting from θ_0 and ϑ_0 respectively and driven by the same noise, i.e. for all $k \in \mathbb{N}$,

$$\begin{cases}
\theta_{k+1} &= \theta_k - \gamma\left(\nabla U_0(\theta_k) + (N/p)\sum_{i\in S_{k+1}}\nabla U_i(\theta_k)\right) + \sqrt{2\gamma}Z_{k+1} \ , \\
\vartheta_{k+1} &= \vartheta_k - \gamma\left(\nabla U_0(\vartheta_k) + (N/p)\sum_{i\in S_{k+1}}\nabla U_i(\vartheta_k)\right) \ ,
\end{cases}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of standard Gaussian variables and $(S_k)_{k\geq 1}$ an i.i.d. sequence of subsamples with replacement of $\{1, \ldots, N\}$ of size p. Denote by $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k, \vartheta_k)_{k\in\mathbb{N}}$. We have for $k \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_{k+1} - \vartheta_{k+1}\|^2 \middle| \mathcal{F}_k\right] = \|\theta_k - \vartheta_k\|^2 - 2\gamma\,\langle\theta_k - \vartheta_k, \nabla U(\theta_k) - \nabla U(\vartheta_k)\rangle + 2\gamma d \\
&+ \gamma^2\mathbb{E}\left[\left\|\nabla U_0(\theta_k) + (N/p)\sum_{i\in S_{k+1}}\nabla U_i(\theta_k) - \nabla U_0(\vartheta_k) - (N/p)\sum_{i\in S_{k+1}}\nabla U_i(\vartheta_k)\right\|^2 \middle| \mathcal{F}_k\right] \ .
\end{aligned}$$

By **H**20 and **H**22, $\theta \mapsto \nabla U_0(\theta) + (N/p)\sum_{i\in S}\nabla U_i(\theta)$ is $\mathbb{P}$-a.s L-co-coercive and we obtain

$$\mathbb{E}\left[\left\|\theta_{k+1} - \vartheta_{k+1}\right\|^2 \middle| \mathcal{F}_k\right] \leq \left\{1 - 2m\gamma(1-\gamma L/2)\right\}\left\|\theta_k - \vartheta_k\right\|^2 + 2\gamma d \,,$$

which concludes the proof by a straightforward induction.

7.A.3 Proof of Theorem 7.4

Proof of i). Let $\gamma \in \left(0, \Sigma^{-1}\{1 + N/(p\sum_{i=1}^N x_i^2)\}^{-1}\right]$, $(\theta_k)_{k\in\mathbb{N}}$ be the iterates of SGLD (7.3) started at $\theta^\star$ and $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the associated filtration. For all $k \in \mathbb{N}$, $\mathbb{E}\left[\theta_k\right] = \theta^\star$. The variance of θ_k satisfies the following recursion for $k \in \mathbb{N}$

$$\begin{aligned}
&\mathbb{E}\left[(\theta_{k+1} - \theta^\star)^2 \middle| \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\{\theta_k - \theta^\star - \gamma\left(\Sigma(\theta_k - \theta^\star) + \rho(S_{k+1})(\theta_k - \theta^\star) + \xi(S_{k+1})\right) + \sqrt{2\gamma}Z_{k+1}\right\}^2 \middle| \mathcal{F}_k\right] \\
&= \mu(\theta_k - \theta^\star)^2 + 2\gamma + \gamma^2 A \,,
\end{aligned}$$

where

$$\mu = \mathbb{E}\left[\left\{1 - \gamma\left(\frac{1}{\sigma_\theta^2} + \frac{N}{\sigma_y^2 p}\sum_{i\in S} x_i^2\right)\right\}^2\right] \,, \quad A = \mathbb{E}\left[\left\{\frac{\theta^\star}{\sigma_\theta^2} + \frac{N}{\sigma_y^2 p}\sum_{i\in S}\left(x_i\theta^\star - y_i\right)x_i\right\}^2\right] \,.$$

We have for μ,

$$\begin{aligned}
\mu &= 1 - 2\gamma\Sigma + \gamma^2\mathbb{E}\left[\left\{\frac{N}{\sigma_y^2 p}\sum_{i\in S} x_i^2 - \frac{1}{\sigma_y^2}\sum_{i=1}^N x_i^2\right\}^2\right] + \gamma^2\Sigma^2 \\
&= 1 - 2\gamma\Sigma + \gamma^2\left\{\Sigma^2 + \frac{N}{\sigma_y^4 p}\sum_{i=1}^N\left(x_i^2 - \frac{1}{N}\sum_{j=1}^N x_j^2\right)\right\} \leq 1 - \gamma\Sigma \,,
\end{aligned}$$

and for A,

$$A = \frac{N}{p}\sum_{i=1}^N\left\{\frac{\left(x_i\theta^\star - y_i\right)x_i}{\sigma_y^2} + \frac{\theta^\star}{N\sigma_\theta^2}\right\}^2 \,.$$

By a straightforward induction, we obtain that the variance of the n^{th} iterate of SGLD started at $\theta^\star$ is for $n \in \mathbb{N}^*$

$$\int_{\mathbb{R}}(\theta - \theta^\star)^2 R^n_{\text{SGLD}}(\theta^\star, \mathrm{d}\theta) = \frac{1-\mu^n}{1-\mu}2\gamma + \frac{1-\mu^n}{1-\mu}\frac{N\gamma^2}{p}\sum_{i=1}^N\left\{\frac{\left(x_i\theta^\star - y_i\right)x_i}{\sigma_y^2} + \frac{\theta^\star}{N\sigma_\theta^2}\right\}^2 \,.$$

For SGLDFP, the additive part of the noise in the stochastic gradient disappears and we obtain similarly for $n \in \mathbb{N}^*$

$$\int_{\mathbb{R}}(\theta - \theta^\star)^2 R^n_{\text{FP}}(\theta^\star, \mathrm{d}\theta) = \frac{1-\mu^n}{1-\mu}2\gamma \,.$$

To conclude, we use that for two probability measures with given mean and covariance matrices, the Wasserstein distance between the two Gaussians with these respective parameters is a lower bound for the Wasserstein distance between the two measures [Gel, Theorem 2.1].

The proof of ii) is straightforward.

7.B Proofs of Section 7.3.2

7.B.1 Proof of Proposition 7.5

Let θ be distributed according to π. By **H** 21, for all $\vartheta \in \mathbb{R}^d$, $U(\vartheta) \geq U(\theta^\star) + (m/2)\|\vartheta - \theta^\star\|^2$ and $\mathbb{E}[\nabla U(\theta)] = 0$. By a Taylor expansion of ∇U around $\theta^\star$, we obtain

$$0 = \mathbb{E}[\nabla U(\theta)] = \nabla^2 U(\theta^\star)(\mathbb{E}[\theta] - \theta^\star) + (1/2)\,\mathrm{D}^3 U(\theta^\star)\left[\mathbb{E}\left[(\theta - \theta^\star)^{\otimes 2}\right]\right] + \mathbb{E}[\mathcal{R}_1(\theta)] \ ,$$

where by **H**20, $\mathcal{R}_1 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\vartheta \in \mathbb{R}^d} \left\{ \|\mathcal{R}_1(\vartheta)\| / \|\vartheta - \theta^\star\|^3 \right\} \leq L/6 \ . \tag{7.17}$$

Rearranging the terms, we get

$$\mathbb{E}[\theta] - \theta^\star = -(1/2)\nabla^2 U(\theta^\star)^{-1}\,\mathrm{D}^3 U(\theta^\star)\left[\mathbb{E}\left[(\theta - \theta^\star)^{\otimes 2}\right]\right] - \nabla^2 U(\theta^\star)^{-1}\mathbb{E}[\mathcal{R}_1(\theta)] \ .$$

To estimate the covariance matrix of π around $\theta^\star$, we start again from the Taylor expansion of ∇U around $\theta^\star$ and we obtain

$$\mathbb{E}\left[\nabla U(\theta)^{\otimes 2}\right] = \mathbb{E}\left[\left(\nabla^2 U(\theta^\star)(\theta - \theta^\star) + \mathcal{R}_2(\theta)\right)^{\otimes 2}\right] = \nabla^2 U(\theta^\star)^{\otimes 2}\mathbb{E}\left[(\theta - \theta^\star)^{\otimes 2}\right] + \mathbb{E}[\mathcal{R}_3(\theta)] \tag{7.18}$$

where by **H**20, $\mathcal{R}_2 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\vartheta \in \mathbb{R}^d} \left\{ \|\mathcal{R}_2(\vartheta)\| / \|\vartheta - \theta^\star\|^2 \right\} \leq L/2 \ , \tag{7.19}$$

and $\mathcal{R}_3 : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is defined for all $\vartheta \in \mathbb{R}^d$ by

$$\mathcal{R}_3(\vartheta) = \nabla^2 U(\theta^\star)(\vartheta - \theta^\star) \otimes \mathcal{R}_2(\vartheta) + \mathcal{R}_2(\vartheta) \otimes \nabla^2 U(\theta^\star)(\vartheta - \theta^\star) + \mathcal{R}_2(\vartheta)^{\otimes 2} \ . \tag{7.20}$$

$\mathbb{E}\left[\nabla U(\theta)^{\otimes 2}\right]$ is the Fisher information matrix and by a Taylor expansion of $\nabla^2 U$ around $\theta^\star$ and an integration by parts,

$$\mathbb{E}\left[\nabla U(\theta)^{\otimes 2}\right] = \mathbb{E}\left[\nabla^2 U(\theta)\right] = \nabla^2 U(\theta^\star) + \mathbb{E}[\mathcal{R}_4(\theta)]$$

where by **H**20, $\mathcal{R}_4 : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ satisfies

$$\sup_{\vartheta \in \mathbb{R}^d} \{\|\mathcal{R}_4(\vartheta)\| / \|\vartheta - \theta^\star\|\} \leq L \ . \tag{7.21}$$

Combining this result, (7.17), (7.18), (7.19), (7.20), (7.21) and $\mathbb{E}[\|\theta - \theta^\star\|^4] \leq d(d+2)/m^2$ by [Bro+18, Lemma 9] conclude the proof.

7.B.2 Proofs of Theorem 7.6 and Theorem 7.7

First note that under **H**20, **H**21 and **H**22, there exists $r \in [0, L/(\sqrt{p}m)]$ such that

$$\mathrm{K} \preceq r^2 (\nabla^2 U(\theta^\star))^{\otimes 2} \,, \tag{7.22}$$

i.e. for all $A \in \mathbb{R}^{d\times d}$,

$$\mathrm{Tr}(A^{\mathrm{T}}\,\mathrm{K}(A)) \leq r^2\,\mathrm{Tr}(A^{\mathrm{T}}(\nabla^2 U(\theta^\star))^{\otimes 2} A) \,,$$

and where K is defined in (7.7). In addition, if $\liminf_{N\to+\infty} N^{-1}m > 0$, r can be chosen independently of N.

Moreover, for all $\gamma \in (0, 2/L)$, H defined in (7.8), is invertible and for all $\gamma \in (0, 2/\{(1+r^2)L\})$, G defined in (7.9), is invertible. Indeed,

$$\begin{aligned}
\mathrm{H} &= \nabla^2 U(\theta^\star) \otimes \left(\mathrm{Id} - \frac{\gamma}{2}\nabla^2 U(\theta^\star)\right) + \left(\mathrm{Id} - \frac{\gamma}{2}\nabla^2 U(\theta^\star)\right) \otimes \nabla^2 U(\theta^\star) \succ 0 \,, \\
\mathrm{G} &\succeq \nabla^2 U(\theta^\star) \otimes \mathrm{Id} + \mathrm{Id} \otimes \nabla^2 U(\theta^\star) - \gamma(1+r^2)\nabla^2 U(\theta^\star) \otimes \nabla^2 U(\theta^\star) \\
&\succeq \nabla^2 U(\theta^\star) \otimes \left(\mathrm{Id} - \frac{\gamma(1+r^2)}{2}\nabla^2 U(\theta^\star)\right) + \left(\mathrm{Id} - \frac{\gamma(1+r^2)}{2}\nabla^2 U(\theta^\star)\right) \otimes \nabla^2 U(\theta^\star) \succ 0 \,.
\end{aligned}$$

For simplicity of notation, in this Section, we use $\epsilon(\theta)$ to denote the difference between the stochastic and the exact gradients at $\theta \in \mathbb{R}^d$. More precisely, ϵ is the null function for LMC and is defined for $\theta \in \mathbb{R}^d$ by

$$\epsilon(\theta) = \frac{N}{p}\sum_{i\in S} \nabla U_i(\theta) - \sum_{j=1}^{N} \nabla U_j(\theta) \quad \text{for SGLD and SGD,} \tag{7.23}$$

$$\epsilon(\theta) = \nabla U_0(\theta) - \nabla U_0(\theta^\star) + \frac{N}{p}\sum_{i\in S}\{\nabla U_i(\theta) - \nabla U_i(\theta^\star)\} - \nabla U(\theta) \quad \text{for SGLDFP,} \tag{7.24}$$

where S is a random subsample of $\{1, \dots, N\}$ with replacement of size $p \in \mathbb{N}^*$. In this setting, the update equation for LMC, SGLD and SGLDFP is given for $k \in \mathbb{N}$ by

$$\theta_{k+1} = \theta_k - (\nabla U(\theta_k) + \epsilon_{k+1}(\theta_k)) + \sqrt{2\gamma} Z_{k+1} \,, \tag{7.25}$$

where $(Z_k)_{k\geq 1}$ is a sequence of i.i.d. standard d-dimensional Gaussian variables and the sequence of vector fields $(\epsilon_k)_{k\geq 1}$ is associated to a sequence $(S_k)_{k\geq 1}$ of i.i.d. random subsample of $\{1, \dots, N\}$ with replacement of size $p \in \mathbb{N}^*$. We also denote by $\bar{\pi} \in \mathcal{P}_2(\mathbb{R}^d)$ the invariant probability measure of LMC, SGLDFP or SGLD.

Control of the moments of order 2 and 4 of LMC, SGLDFP and SGLD

Lemma 7.8. *Assume **H**20, **H**21 and **H**22.*

i) For all initial distribution $\lambda \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma \in (0, 1/L]$ *and* $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \leq (1 - m\gamma)^k \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \lambda(\mathrm{d}\vartheta) + (2d)/m$$

where $(\theta_k)_{k\in\mathbb{N}}$ *are the iterates of SGLDFP* (7.5) *or LMC* (7.2).

ii) For all initial distribution $\lambda \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma \in (0, 1/(2L)]$ *and* $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \leq (1 - m\gamma)^k \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \lambda(\mathrm{d}\vartheta) + \frac{2d}{m} + \frac{2\gamma N}{mp} \sum_{i=1}^N \left\| \nabla U_i(\theta^\star) - \frac{1}{N} \sum_{j=1}^N \nabla U_j(\theta^\star) \right\|^2$$

where $(\theta_k)_{k\in\mathbb{N}}$ *are the iterates of SGLD* (7.3).

Proof. i). We prove the result for SGLDFP, the case of LMC is identical. Let $\gamma \in (0, 1/L]$, $(\theta_k)_{k\in\mathbb{N}}$ be the iterates of SGLDFP and $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k)_{k\in\mathbb{N}}$. By (7.5), we have for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_{k+1} - \theta^\star\|^2 \middle| \mathcal{F}_k\right] = \|\theta_k - \theta^\star\|^2 - 2\gamma \langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\rangle + 2\gamma d + \gamma^2 \mathbb{E}\left[\left\| \nabla U_0(\theta_k) - \nabla U_0(\theta^\star) + \frac{N}{p} \sum_{i\in S_{k+1}} \{\nabla U_i(\theta_k) - \nabla U_i(\theta^\star)\} \right\|^2 \middle| \mathcal{F}_k\right]$$

By **H**20 and **H**22, $\theta \mapsto \nabla U_0(\theta) - \nabla U_0(\theta^\star) + (N/p)\sum_{i\in S}\{\nabla U_i(\theta) - \nabla U_i(\theta^\star)\}$ is $\mathbb{P}$-a.s L-co-coercive and we obtain

$$\mathbb{E}\left[\|\theta_{k+1} - \theta^\star\|^2 \middle| \mathcal{F}_k\right] \leq \{1 - 2m\gamma(1 - \gamma L/2)\} \|\theta_k - \theta^\star\|^2 + 2\gamma d \,.$$

A straightforward induction concludes the proof.

ii). Let $\gamma \in (0, 1/(2L)]$, $(\theta_k)_{k\in\mathbb{N}}$ be the iterates of SGLD and $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k)_{k\in\mathbb{N}}$. By (7.3), we have for all $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1} - \theta^\star\|^2 \middle| \mathcal{F}_k\right] &= \|\theta_k - \theta^\star\|^2 - 2\gamma \langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\rangle + 2\gamma d \\
&\quad + \gamma^2 \mathbb{E}\left[\left\| \nabla U_0(\theta_k) + \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k) \right\|^2 \middle| \mathcal{F}_k\right] \\
&\leq \|\theta_k - \theta^\star\|^2 - 2\gamma \langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\rangle + 2\gamma d \\
&\quad + 2\gamma^2 \mathbb{E}\left[\left\| \nabla U_0(\theta_k) - \nabla U_0(\theta^\star) + \frac{N}{p} \sum_{i\in S_{k+1}} \{\nabla U_i(\theta_k) - \nabla U_i(\theta^\star)\} \right\|^2 \middle| \mathcal{F}_k\right] \\
&\quad + 2\gamma^2 \mathbb{E}\left[\left\| \nabla U_0(\theta^\star) + \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta^\star) \right\|^2 \middle| \mathcal{F}_k\right] \,.
\end{aligned}$$

By **H**20 and **H**22, $\theta \mapsto \nabla U_0(\theta) + (N/p) \sum_{i \in S} \nabla U_i(\theta)$ is $\mathbb{P}$-a.s L-co-coercive and we obtain

$$\mathbb{E}\left[\|\theta_{k+1} - \theta^\star\|^2 \middle| \mathcal{F}_k\right] \leq \{1 - 2m\gamma(1-\gamma L)\} \|\theta_k - \theta^\star\|^2 + 2\gamma d + \frac{2\gamma^2 N}{p} \sum_{i=1}^N \left\| \nabla U_i(\theta^\star) - \frac{1}{N} \sum_{j=1}^N \nabla U_j(\theta^\star) \right\|^2 .$$

A straightforward induction concludes the proof. □

Lemma 7.9. *Assume* ***H***20, ***H***21 *and* ***H***22. *For all initial distribution* $\lambda \in \mathcal{P}_4(\mathbb{R}^d)$, $\gamma \in (0, 1/\{12(L \vee 1)\}]$ *and* $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|\theta_k - \theta^\star\|^4\right] &\leq (1-2m\gamma)^k \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^4 \lambda(\mathrm{d}\vartheta) \\ &+ \left\{12\gamma^2 \mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right] + 2\gamma(2d+1)\right\} k(1-m\gamma)^{k-1} \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \lambda(\mathrm{d}\vartheta) \\ &+ \left\{\frac{2d+1}{m} + \frac{6\gamma}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right]\right\}^2 \\ &+ \frac{2\gamma d(2+d)}{m} + \frac{4\gamma^3}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^4\right] + \frac{4\gamma^2(d+2)}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right] .\end{aligned}$$

where $(\theta_k)_{k\in\mathbb{N}}$ *are the iterates of LMC* (7.2), *SGLD* (7.3) *or SGLDFP* (7.5).

Proof. Let $\gamma \in (0, 1/\{12(L \vee 1)\}]$, $(\theta_k)_{k\in\mathbb{N}}$ be the iterates of LMC (7.2), SGLD (7.3) or SGLDFP (7.5) and $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be the associated filtration. By developing the square, we have

$$\begin{aligned}\|\theta_1 - \theta^\star\|^4 = \Big(&\|\theta_0 - \theta^\star\|^2 + 2\gamma \|Z_1\|^2 + \gamma^2 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^2 \\ &- 2\gamma \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), \theta_0 - \theta^\star \rangle + \sqrt{2\gamma} \langle \theta_0 - \theta^\star, Z_1 \rangle - (2\gamma)^{3/2} \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), Z_1 \rangle \Big)^2 ,\end{aligned}$$

and taking the conditional expectation w.r.t. $\mathcal{F}_0$,

$$\begin{aligned}\mathbb{E}\left[\|\theta_1 - \theta^\star\|^4 \middle| \mathcal{F}_0\right] = \mathbb{E}\Big[&\|\theta_0 - \theta^\star\|^4 + 4\gamma^2 \|Z_1\|^4 + \gamma^4 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^4 \\ &+ 4\gamma^2 \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), \theta_0 - \theta^\star \rangle^2 + 2\gamma \langle \theta_0 - \theta^\star, Z_1 \rangle^2 + (2\gamma)^3 \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), Z_1 \rangle^2 \\ &+ 4\gamma \|Z_1\|^2 \|\theta_0 - \theta^\star\|^2 + 2\gamma^2 \|\theta_0 - \theta^\star\|^2 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^2 \\ &- 4\gamma \|\theta_0 - \theta^\star\|^2 \langle \nabla U(\theta_0), \theta_0 - \theta^\star \rangle + 4\gamma^3 \|Z_1\|^2 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^2 \\ &- 8\gamma^2 \|Z_1\|^2 \langle \nabla U(\theta_0), \theta_0 - \theta^\star \rangle - 4\gamma^3 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^2 \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), \theta_0 - \theta^\star \rangle \\ &- 8\gamma^2 \langle \theta_0 - \theta^\star, Z_1 \rangle \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), Z_1 \rangle \,|\mathcal{F}_0\Big] .\end{aligned}$$

By **H**20 and **H**22, $\theta \mapsto \nabla U(\theta) + \epsilon_1(\theta)$ is $\mathbb{P}$-a.s L-co-coercive and we have for all $\theta \in \mathbb{R}^d$, $\mathbb{P}$-a.s ,

$$\begin{aligned}\|\nabla U(\theta) + \epsilon_1(\theta) - \epsilon_1(\theta^\star)\|^2 &\leq L \langle \theta - \theta^\star, \nabla U(\theta) + \epsilon_1(\theta) - \epsilon_1(\theta^\star) \rangle , \\ \|\nabla U(\theta) + \epsilon_1(\theta) - \epsilon_1(\theta^\star)\|^4 &\leq L^2 \|\theta - \theta^\star\|^2 \langle \theta - \theta^\star, \nabla U(\theta) + \epsilon_1(\theta) - \epsilon_1(\theta^\star) \rangle .\end{aligned}$$

Combining it with $\mathbb{E}\left[\|Z_1\|^4\right] = d(2+d)$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_1 - \theta^\star\|^4 \middle| \mathcal{F}_0, S_1\right] \leq \|\theta_0 - \theta^\star\|^4 - 4\gamma(1 - 3\gamma L - 2\gamma^3 L^2)\|\theta_0 - \theta^\star\|^2 \\
&\times \langle \theta_0 - \theta^\star, \nabla U(\theta_0) + \epsilon_1(\theta_0) - \epsilon_1(\theta^\star)\rangle + (12\gamma^2\|\epsilon_1(\theta^\star)\|^2 + 2\gamma(2d+1))\|\theta_0 - \theta^\star\|^2 \\
&+ 4\gamma^2 d(2+d) + 8\gamma^4\|\epsilon_1(\theta^\star)\|^4 + 8\gamma^3(d+2)\|\epsilon_1(\theta^\star)\|^2 \\
&- 8(d+1)\gamma^2(1-2\gamma L)\langle \theta_0 - \theta^\star, \nabla U(\theta_0) + \epsilon_1(\theta_0) - \epsilon_1(\theta^\star)\rangle \ .
\end{aligned}$$

By **H**21 and using $\gamma \leq 1/\{12(L \vee 1)\}$, we get

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_1 - \theta^\star\|^4 \middle| \mathcal{F}_0\right] \leq (1 - 2m\gamma)\|\theta_0 - \theta^\star\|^4 + \left\{12\gamma^2\mathbb{E}\left[\|\epsilon_1(\theta^\star)\|^2\right] + 2\gamma(2d+1)\right\}\|\theta_0 - \theta^\star\|^2 \\
&+ 4\gamma^2 d(2+d) + 8\gamma^4\mathbb{E}\left[\|\epsilon_1(\theta^\star)\|^4\right] + 8\gamma^3(d+2)\mathbb{E}\left[\|\epsilon_1(\theta^\star)\|^2\right] \ .
\end{aligned}$$

By a straightforward induction, we have for all $n \in \mathbb{N}$

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_n - \theta^\star\|^4\right] \leq (1 - 2m\gamma)^n \mathbb{E}\left[\|\theta_0 - \theta^\star\|^4\right] \\
&+ \left\{12\gamma^2\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right] + 2\gamma(2d+1)\right\}\sum_{k=0}^{n-1}(1-2m\gamma)^{n-1-k}\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \\
&+ (2m\gamma)^{-1}\left\{4\gamma^2 d(2+d) + 8\gamma^4\mathbb{E}\left[\|\epsilon(\theta^\star)\|^4\right] + 8\gamma^3(d+2)\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right]\right\}
\end{aligned}$$

and by Lemma 7.8,

$$\begin{aligned}
\mathbb{E}\left[\|\theta_n - \theta^\star\|^4\right] &\leq (1-2m\gamma)^n \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^4 \lambda(\mathrm{d}\vartheta) \\
&+ \left\{12\gamma^2\mathbb{E}\left[\|\epsilon_1(\theta^\star)\|^2\right] + 2\gamma(2d+1)\right\} n(1-m\gamma)^{n-1}\int_{\mathbb{R}^d}\|\vartheta - \theta^\star\|^2\lambda(\mathrm{d}\vartheta) \\
&+ \left\{\frac{2d+1}{m} + \frac{6\gamma}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right]\right\}^2 \\
&+ \frac{2\gamma d(2+d)}{m} + \frac{4\gamma^3}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^4\right] + \frac{4\gamma^2(d+2)}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right] \ .
\end{aligned}$$

□

Thanks to this lemma, we obtain the following corollary. The upper bound for SGD is given by [DDB17, Lemma 13].

Corollary 7.10. *Assume* ***H***20, ***H***21 *and* ***H***22.

i) Let $\gamma = \eta/N$ *with* $\eta \in (0, 1/\{24(\tilde{L} \vee 1)\}]$ *and assume that* $\liminf_{N\to+\infty} N^{-1}m > 0$. *Then,*

$$\begin{aligned}
\int_{\mathbb{R}^d}\|\theta - \theta^\star\|^4 \pi_{\mathrm{LMC}}(\mathrm{d}\theta) &= d^2 O_{N\to+\infty}(N^{-2}) \ , \\
\int_{\mathbb{R}^d}\|\theta - \theta^\star\|^4 \pi_{\mathrm{FP}}(\mathrm{d}\theta) &= d^2 O_{N\to+\infty}(N^{-2}) \ .
\end{aligned}$$

ii) Let $\gamma = \eta/N$ *with* $\eta \in (0, 1/\{24(\tilde{L} \vee 1)\}]$ *and assume that* $\liminf_{N\to+\infty} N^{-1}m > 0$ *and that* $N \geq 1/\eta$. *Then,*

$$\int_{\mathbb{R}^d} \|\theta - \theta^\star\|^4 \, \pi_{\mathrm{SGLD}}(\mathrm{d}\theta) = d^2 O_{\eta\to 0}(\eta^2) \; , \quad \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^4 \, \pi_{\mathrm{SGD}}(\mathrm{d}\theta) = d^2 O_{\eta\to 0}(\eta^2) \; .$$

Proofs of Theorem 7.6 and Theorem 7.7

Denote by

$$\eta_0 = \inf_{N\geq 1} \left\{ \frac{N}{12(L \vee 1)} \wedge \frac{2N}{(1+r^2)L} \right\} > 0 \; , \tag{7.26}$$

and set $\gamma = \eta/N$ with $\eta \in (0, \eta_0)$. Let $\delta \in \{0, 1\}$ be equal to 1 for LMC, SGLDFP and SGLD and 0 for SGD. Let θ_0 be distributed according to $\bar{\pi}$. By (7.25) and using a Taylor expansion around $\theta^\star$ for ∇U, we obtain

$$\theta_1 - \theta^\star = \theta_0 - \theta^\star - \gamma \left(\nabla^2 U(\theta^\star)(\theta_0 - \theta^\star) + \mathcal{R}_1(\theta_0) + \epsilon_1(\theta_0) \right) + \delta\sqrt{2\gamma} Z_1 \; ,$$

where by **H**20, $\mathcal{R}_1 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\theta\in\mathbb{R}^d} \left\{ \|\mathcal{R}_1(\theta)\| \, / \, \|\theta - \theta^\star\|^2 \right\} \leq L/2 \; . \tag{7.27}$$

Taking the tensor product and the expectation, and using that $\theta_0, \epsilon_1, Z_1$ are mutually independent, we obtain

$$\begin{aligned} \mathrm{H}\,\mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right] &= 2\delta\,\mathrm{Id} + \gamma\mathbb{E}\left[\epsilon_1(\theta_0)^{\otimes 2}\right] + \mathbb{E}\left[\mathcal{R}_1(\theta_0) \otimes \{\theta_0 - \theta^\star\} + \{\theta_0 - \theta^\star\} \otimes \mathcal{R}_1(\theta_0)\right] \\ &+ \gamma\mathbb{E}\left[\mathcal{R}_1(\theta_0)^{\otimes 2} + \{\nabla^2 U(\theta^\star)(\theta_0 - \theta^\star)\} \otimes \mathcal{R}_1(\theta_0) + \mathcal{R}_1(\theta_0) \otimes \nabla^2 U(\theta^\star)(\theta_0 - \theta^\star)\right] \; . \end{aligned} \tag{7.28}$$

For LMC, ϵ_1 is the null function and by Corollary 7.10-i), (7.27) and (7.28), we obtain (7.10). Regarding SGLDFP, SGLD and SGD, by a Taylor expansion of ϵ_1 around $\theta^\star$, we get for all $\theta \in \mathbb{R}^d$, $\mathbb{P}$-a.s ,

$$\epsilon_1(\theta) = \epsilon_1(\theta^\star) + \nabla\epsilon_1(\theta^\star)(\theta - \theta^\star) + \mathcal{R}_2(\theta)$$

where by **H**20,$\mathcal{R}_2 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\theta\in\mathbb{R}^d} \left\{ \|\mathcal{R}_2(\theta)\| \, / \, \|\theta - \theta^\star\|^2 \right\} \leq L/2 \; . \tag{7.29}$$

Therefore, taking the tensor product and the expectation, we obtain

$$\mathbb{E}\left[\epsilon_1(\theta_0)^{\otimes 2}\right] = \mathbb{E}\left[\epsilon_1(\theta^\star)^{\otimes 2}\right] + (\nabla\epsilon_1(\theta^\star))^{\otimes 2}\,\mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right] + \mathbb{E}\left[\mathcal{R}_3(\theta_0)\right] \tag{7.30}$$

where $\mathcal{R}_3 : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is defined for all $\theta \in \mathbb{R}^d$, $\mathbb{P}$-a.s ,

$$\begin{aligned} \mathcal{R}_3(\theta) &= \epsilon_1(\theta^\star) \otimes \{\nabla\epsilon_1(\theta^\star)(\theta - \theta^\star)\} + \{\nabla\epsilon_1(\theta^\star)(\theta - \theta^\star)\} \otimes \epsilon_1(\theta^\star) \\ &+ \{\epsilon_1(\theta^\star) + \nabla\epsilon_1(\theta^\star)(\theta - \theta^\star)\} \otimes \mathcal{R}_2(\theta) + \mathcal{R}_2(\theta) \otimes \{\epsilon_1(\theta^\star) + \nabla\epsilon_1(\theta^\star)(\theta - \theta^\star)\} + \mathcal{R}_2^{\otimes 2}(\theta) \; . \end{aligned} \tag{7.31}$$

Note that $\mathrm{K} = \mathbb{E}\left[(\nabla\epsilon_1(\theta^\star))^{\otimes 2}\right]$. For SGLDFP, $\epsilon_1(\theta^\star) = 0$ a.s By Corollary 7.10-i), (7.27), (7.28), (7.29), (7.30) and (7.31), we obtain (7.11).

Regarding SGLD and SGD, we have $\mathbb{E}\left[\epsilon_1(\theta^\star)^{\otimes 2}\right] = (N/p)\,\mathrm{M}$ where M is defined in (7.14). By Corollary 7.10-ii), (7.27), (7.28), (7.29), (7.30) and (7.31), we obtain (7.12) and (7.13).

For the mean of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD}, by a Taylor expansion around $\theta^\star$ for ∇U of order 3, we obtain

$$\theta_1 - \theta^\star = \theta_0 - \theta^\star - \gamma\left(\nabla^2 U(\theta^\star)(\theta_0 - \theta^\star) + (1/2)\,\mathrm{D}^3 U(\theta^\star)(\theta_0 - \theta^\star)^{\otimes 2} + \mathcal{R}_4(\theta_0) + \epsilon_1(\theta_0)\right) + \delta\sqrt{2\gamma}Z_1\,,$$

where by **H**20, $\mathcal{R}_4 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\theta\in\mathbb{R}^d}\left\{\|\mathcal{R}_4(\theta)\| \,/\, \|\theta - \theta^\star\|^3\right\} \le L/6\,. \tag{7.32}$$

Taking the expectation and using that θ_1 is distributed according to $\bar{\pi}$, we get

$$\mathbb{E}\left[\theta_0\right] - \theta^\star = -(1/2)\nabla^2 U(\theta^\star)\,\mathrm{D}^3 U(\theta^\star)[\mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right]] - \nabla^2 U(\theta^\star)^{-1}\mathbb{E}\left[\mathcal{R}_4(\theta_0)\right]\,.$$

(7.10), (7.11), (7.12),(7.13), (7.32) and Corollary 7.10 conclude the proof.

7.C Means and covariance matrices of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD} in the Bayesian linear regression

In this Section, we provide explicit expressions of the covariance matrices of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD} in the context of the Bayesian linear regression. In this setting, the algorithms are without bias, i.e.

$$\int_{\mathbb{R}^d}\theta\pi_{\mathrm{LMC}}(\mathrm{d}\theta) = \int_{\mathbb{R}^d}\theta\pi_{\mathrm{FP}}(\mathrm{d}\theta) = \int_{\mathbb{R}^d}\theta\pi_{\mathrm{SGLD}}(\mathrm{d}\theta) = \int_{\mathbb{R}^d}\theta\pi_{\mathrm{SGD}}(\mathrm{d}\theta) = \int_{\mathbb{R}^d}\theta\pi(\mathrm{d}\theta) = \theta^\star\,. \tag{7.33}$$

Before giving the expressions of the variances in Theorem 7.11, we define $\mathrm{T} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ for all $A \in \mathbb{R}^{d\times d}$ by

$$\mathrm{T}(A) = \mathbb{E}\left[\left(\frac{\mathrm{Id}}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S} x_i x_i^{\mathrm{T}} - \Sigma\right)^{\otimes 2} A\right] = \frac{N}{p}\sum_{i=1}^{N}\left(\frac{x_i x_i^{\mathrm{T}}}{\sigma_y^2} + \frac{\mathrm{Id}}{N\sigma_\theta^2} - \frac{\Sigma}{N}\right)^{\otimes 2} A\,, \tag{7.34}$$

where S is a random subsample of $\{1,\dots,N\}$ with replacement of size $p \in \mathbb{N}^*$. Note that, in this setting, $\tilde{L} = \max_{i\in\{1,\dots,N\}}\|x_i\|^2$ and m is the smallest eigenvalue of Σ. There exists $r \in [0, L/(\sqrt{p}m)]$ such that

$$\mathrm{T} \preceq r^2\Sigma^{\otimes 2} \tag{7.35}$$

i.e. for all $A \in \mathbb{R}^{d\times d}$, $\mathrm{Tr}(A^{\mathrm{T}}\,\mathrm{T}\cdot A) \le r^2\,\mathrm{Tr}(A^{\mathrm{T}}\Sigma^{\otimes 2}A)$. Assuming that $\liminf_{N\to+\infty} N^{-1}m > 0$, r can be chosen independently of N.

Theorem 7.11. *Consider the case of the Bayesian linear regression. We have for all* $\gamma \in (0, 2/L)$

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\mathrm{LMC}}(\mathrm{d}\theta) = \left(\mathrm{Id} \otimes \Sigma + \Sigma \otimes \mathrm{Id} - \gamma \Sigma \otimes \Sigma\right)^{-1} (2\,\mathrm{Id}) \,,$$

and for all $\gamma \in \left(0, 2/\{(1+r^2)L\}\right)$,

$$\begin{aligned}
\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\mathrm{FP}}(\mathrm{d}\theta) &= \left\{\mathrm{Id} \otimes \Sigma + \Sigma \otimes \mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})\right\}^{-1} (2\,\mathrm{Id}) \,, \\
\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\mathrm{SGLD}}(\mathrm{d}\theta) &= \left\{\mathrm{Id} \otimes \Sigma + \Sigma \otimes \mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})\right\}^{-1} \\
&\qquad \cdot \left\{ 2\,\mathrm{Id} + \frac{\gamma N}{p} \sum_{i=1}^{N} \left(\frac{(x_i^{\mathrm{T}} \theta^\star - y_i) x_i}{\sigma_y^2} + \frac{\theta^\star}{\sigma_\theta^2} \right)^{\otimes 2} \right\} \,, \\
\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\mathrm{SGD}}(\mathrm{d}\theta) &= \left\{\mathrm{Id} \otimes \Sigma + \Sigma \otimes \mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})\right\}^{-1} \\
&\qquad \cdot \frac{\gamma N}{p} \sum_{i=1}^{N} \left(\frac{(x_i^{\mathrm{T}} \theta^\star - y_i) x_i}{\sigma_y^2} + \frac{\theta^\star}{\sigma_\theta^2} \right)^{\otimes 2} \,.
\end{aligned}$$

Proof. We prove the result for SGLD, the adaptation to the other algorithms is immediate. Let $\gamma \in \left(0, 2/\{(1+r^2)L\}\right)$, θ_0 be distributed according to π_{SGLD} and θ_1 be given by (7.3). By definition of π_{SGLD}, θ_1 is distributed according to π_{SGLD}. We have

$$\begin{aligned}
\mathbb{E}\left[(\theta_1 - \theta^\star)^{\otimes 2}\right] = \mathbb{E}\Bigg[\Bigg[\Bigg\{ \mathrm{Id} - \gamma \left(\frac{\mathrm{Id}}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2} \sum_{i \in S_1} x_i x_i^{\mathrm{T}} \right) \Bigg\} (\theta_0 - \theta^\star) \\
- \gamma \left(\frac{\theta^\star}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2} \sum_{i \in S_1} (x_i^{\mathrm{T}} \theta^\star - y_i) x_i \right) + \sqrt{2\gamma} Z_1 \Bigg]^{\otimes 2} \Bigg] \,.
\end{aligned}$$

Using that θ_0, S_1, Z_1 are mutually independent, we obtain

$$\begin{aligned}
\Bigg\{ \mathrm{Id} \otimes \Sigma + \Sigma \otimes \mathrm{Id} - \gamma \mathbb{E}\left[\left(\frac{\mathrm{Id}}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2} \sum_{i \in S_1} x_i x_i^{\mathrm{T}} \right)^{\otimes 2} \right] \Bigg\} \mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right] \\
= 2\,\mathrm{Id} + \gamma \mathbb{E}\left[\left(\frac{\theta^\star}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2} \sum_{i \in S_1} (x_i^{\mathrm{T}} \theta^\star - y_i) x_i \right)^{\otimes 2} \right]
\end{aligned}$$

and

$$\begin{aligned}
\left\{\mathrm{Id} \otimes \Sigma + \Sigma \otimes \mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})\right\} \mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right] \\
= 2\,\mathrm{Id} + \frac{\gamma N}{p} \sum_{i=1}^{N} \left(\frac{(x_i^{\mathrm{T}} \theta^\star - y_i) x_i}{\sigma_y^2} + \frac{\theta^\star}{\sigma_\theta^2} \right)^{\otimes 2} \,.
\end{aligned}$$

On $\mathbb{R}^{d\times d}$ equipped with the Hilbert-Schmidt inner product, $\mathrm{Id}\otimes\Sigma+\Sigma\otimes\mathrm{Id}-\gamma(\Sigma^{\otimes 2}+\mathrm{T})$ is a positive definite operator. Indeed, by (7.35),

$$\begin{aligned}\mathrm{Id}\otimes\Sigma+\Sigma\otimes\mathrm{Id}-\gamma(\Sigma^{\otimes 2}+\mathrm{T}) &\succeq \mathrm{Id}\otimes\Sigma+\Sigma\otimes\mathrm{Id}-\gamma(1+r^2)\Sigma^{\otimes 2}\\ &= \left(\mathrm{Id}-\gamma\frac{1+r^2}{2}\Sigma\right)\otimes\Sigma+\Sigma\otimes\left(\mathrm{Id}-\gamma\frac{1+r^2}{2}\Sigma\right)\succ 0\end{aligned}$$

for $\gamma\in\big(0,2/\{(1+r^2)L\}\big)$. $\mathrm{Id}\otimes\Sigma+\Sigma\otimes\mathrm{Id}-\gamma(\Sigma^{\otimes 2}+\mathrm{T})$ is thus invertible, which concludes the proof. □

The covariance matrices make clearly visible the different origins of the noise. The Gaussian noise is responsible of the term $2\,\mathrm{Id}$, while the multiplicative and additive parts of the stochastic gradient (see (7.6)) are related to the operator T and to the term

$$\frac{\gamma N}{p}\sum_{i=1}^{N}\left(\frac{(x_i^{\mathrm{T}}\theta^\star-y_i)x_i}{\sigma_y^2}+\frac{\theta^\star}{\sigma_\theta^2}\right)^{\otimes 2} \tag{7.36}$$

respectively.

Denote by

$$\eta_1=\inf_{N\geq 1}\left\{\frac{2N}{L}\wedge\frac{2N}{(1+r^2)L}\right\}>0\,. \tag{7.37}$$

Corollary 7.12. *Consider the case of the Bayesian linear regression. Set $\gamma=\eta/N$ with $\eta\in(0,\eta_1)$ and assume that $\liminf_{N\to+\infty}N^{-1}m>0$.*

$$\int_{\mathbb{R}^d}\|\theta-\theta^\star\|^2\,\pi_{\mathrm{LMC}}(\mathrm{d}\theta)=d\Theta_{N\to+\infty}(N^{-1})\,,\quad \int_{\mathbb{R}^d}\|\theta-\theta^\star\|^2\,\pi_{\mathrm{FP}}(\mathrm{d}\theta)=d\Theta_{N\to+\infty}(N^{-1})\,,$$
$$\int_{\mathbb{R}^d}\|\theta-\theta^\star\|^2\,\pi_{\mathrm{SGLD}}(\mathrm{d}\theta)=\eta d\Theta_{N\to+\infty}(1)\,,\quad \int_{\mathbb{R}^d}\|\theta-\theta^\star\|^2\,\pi_{\mathrm{SGD}}(\mathrm{d}\theta)=\eta d\Theta_{N\to+\infty}(1)\,.$$

Recall that, according to the Bernstein-von Mises theorem, the variance of π is of the order d/N when N is large. The corollary confirms that π_{SGLD} is very far from π when the constant step size γ is chosen proportional to $1/N$.

Bibliography

[ABW12] Sungjin Ahn, Anoop Korattikara Balan, and Max Welling. "Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring". In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. 2012.

[AC08] A. Abdulle and S. Cirilli. "S-ROCK: Chebyshev Methods for Stiff Stochastic Differential Equations". In: *SIAM Journal on Scientific Computing* 30.2 (2008), pp. 997–1014.

[AC99] Roland Assaraf and Michel Caffarel. "Zero-variance principle for Monte Carlo algorithms". In: *Physical review letters* 83.23 (1999), p. 4682.

[AL08] A. Abdulle and T. Li. "S-ROCK methods for stiff Ito SDEs". In: *Commun. Math. Sci.* 6.4 (Dec. 2008), pp. 845–868.

[And+03] C. Andrieu, N. De Freitas, A. Doucet, and M. I Jordan. "An introduction to MCMC for machine learning". In: *Machine learning* 50.1-2 (2003), pp. 5–43.

[Apo69] Tom M Apostol. *Calculus: Multi Variable Calculus and Linear Algebra, with Applications to Differential Equations and Probability*. John Wiley & Sons, 1969.

[Ard+12] David Ardia, Nalan Baştürk, Lennart Hoogerheide, and Herman K Van Dijk. "A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood". In: *Computational Statistics & Data Analysis* 56.11 (2012), pp. 3398–3414.

[ARW16] C. Andrieu, J. Ridgway, and N. Whiteley. "Sampling normalizing constants in high dimensions using inhomogeneous diffusions". In: *ArXiv e-prints* (Dec. 2016).

[ASW14] Sungjin Ahn, Babak Shahbaba, and Max Welling. "Distributed Stochastic Gradient MCMC". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 1044–1052.

[Atc06] Yves F. Atchadé. "An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift". In: *Methodology and Computing in Applied Probability* 8.2 (June 2006), pp. 235–254. ISSN: 1573-7713.

[Bak+08] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. "A simple proof of the Poincaré inequality for a large class of probability measures." eng. In: *Electronic Communications in Probability [electronic only]* 13 (2008), pp. 60–66.

[Bak+17] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. "Control Variates for Stochastic Gradient MCMC". In: *ArXiv e-prints 1706.05439* (June 2017).

[Bal07] Roger Balian. *From microphysics to macrophysics: methods and applications of statistical physics*. Vol. 1. Springer Science & Business Media, 2007.

[BDH17] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. "On Markov chain Monte Carlo methods for tall data". In: *Journal of Machine Learning Research* 18.47 (2017), pp. 1–43.

[BEL15] S. Bubeck, R Eldan, and J. Lehec. "Finite-time Analysis of Projected Langevin Monte Carlo". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 1243–1251.

[Bes+14] Alexandros Beskos, Dan O. Crisan, Ajay Jasra, and Nick Whiteley. "Error bounds and normalising constants for sequential Monte Carlo samplers in high dimensions". In: *Adv. in Appl. Probab.* 46.1 (Mar. 2014), pp. 279–306.

[Bet15] Michael Betancourt. "The Fundamental Incompatibility of Scalable Hamiltonian Monte Carlo and Naive Data Subsampling". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 533–540.

[BEZ18] Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. "Coupling and Convergence for Hamiltonian Monte Carlo". In: *arXiv e-prints*, arXiv:1805.00452 (May 2018), arXiv:1805.00452.

[BFH12] Gundula Behrens, Nial Friel, and Merrilee Hurn. "Tuning tempered transitions". In: *Statistics and Computing* 22.1 (2012), pp. 65–78.

[BGL14] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552. ISBN: 978-3-319-00226-2; 978-3-319-00227-9.

[BH13] N. Bou-Rabee and M. Hairer. "Nonasymptotic mixing of the MALA algorithm". In: *IMA Journal of Numerical Analysis* 33.1 (2013), pp. 80–110.

[Bha82] R. N. Bhattacharya. "On Classical Limit Theorems for Diffusions". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 44.1 (1982), pp. 47–71. ISSN: 0581572X.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford university press, 2013.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[Bor13] KC Border. "Notes on the Implicit Function Theorem". In: 2013.

[Bra+14] Silouanos Brazitikos, Apostolos Giannopoulos, Petros Valettas, and Beatrice-Helen Vritsiou. *Geometry of isotropic convex bodies*. Vol. 196. American Mathematical Society Providence, 2014.

[Bro+18] Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. "The tamed unadjusted Langevin algorithm". In: *Stochastic Processes and their Applications* (2018). ISSN: 0304-4149.

[BS17] Nawaf Bou-Rabee and Jesús María Sanz-Serna. "Randomized Hamiltonian Monte Carlo". In: *Ann. Appl. Probab.* 27.4 (Aug. 2017), pp. 2159–2194.

[BS18] Nawaf Bou-Rabee and J. M. Sanz-Serna. "Geometric integrators and the Hamiltonian Monte Carlo method". In: *Acta Numerica* 27 (2018), pp. 113–206.

[BS78] D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control*. Vol. 139. Mathematics in Science and Engineering. The discrete time case. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1978, pp. xiii+323. ISBN: 0-12-093260-1.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BV10] Nawaf Bou-Rabee and Eric Vanden-Eijnden. "Pathwise accuracy and ergodicity of metropolized integrators for SDEs". In: *Communications on Pure and Applied Mathematics* 63.5 (2010), pp. 655–696. ISSN: 1097-0312.

[CCG12] Patrick Cattiaux, Djalil Chafaı, and Arnaud Guillin. "Central limit theorems for additive functionals of ergodic Markov diffusions processes". In: *ALEA* 9.2 (2012), pp. 337–382.

[CCM11] S. H. Chang, P. C. Cosman, and L. B. Milstein. "Chernoff-Type Bounds for the Gaussian Error Function". In: *IEEE Transactions on Communications* 59.11 (Nov. 2011), pp. 2939–2944. ISSN: 0090-6778.

[CDC15] C. Chen, N. Ding, and L. Carin. "On the Convergence of Stochastic Gradient MCMC Algorithms with High-Order Integrators". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 2278–2286.

[Cel+12] Gilles Celeux, Mohammed El Anbari, Jean-Michel Marin, and Christian P. Robert. "Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informative Situation". In: *Bayesian Anal.* 7.2 (June 2012), pp. 477–502.

[CFG14] Tianqi Chen, Emily Fox, and Carlos Guestrin. "Stochastic gradient hamiltonian Monte Carlo". In: *Proceedings of the 31st International Conference on Machine Learning*. 2014, pp. 1683–1691.

[Cha+18] N. S. Chatterji, N. Flammarion, Y.-A. Ma, P. L. Bartlett, and M. I. Jordan. "On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo". In: *ArXiv e-prints 1802.05431* (Feb. 2018).

[Che+17] C. Chen, W. Wang, Y. Zhang, Q. Su, and L. Carin. "A Convergence Analysis for A Class of Practical Variance-Reduction Stochastic Gradient MCMC". In: *ArXiv e-prints 1709.01180* (Sept. 2017).

[Che+18] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. "Underdamped Langevin MCMC: A non-asymptotic analysis". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 300–323.

[CLS12] Nicolas Chopin, Tony Lelièvre, and Gabriel Stoltz. "Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors". In: *Statistics and Computing* 22.4 (July 2012), pp. 897–916. ISSN: 1573-1375.

[Cot+13] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. "MCMC methods for functions: modifying old algorithms to make them faster". In: *Statist. Sci.* 28.3 (2013), pp. 424–446. ISSN: 0883-4237.

[CSI00] MH Chen, QM Shao, and JG Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer, New York, 2000.

[CSI12] Ming-Hui Chen, Qi-Man Shao, and Joseph G Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media, 2012.

[CV15a] Ben Cousins and Santosh Vempala. *Computation of the volume of convex bodies*. June 2015.

[CV15b] Benjamin Cousins and Santosh Vempala. "Bypassing KLS: Gaussian cooling and an $O^*(n^3)$ volume algorithm". In: *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*. ACM. 2015, pp. 539–548.

[D+13] Ritabrata Dutta, Jayanta K Ghosh, et al. "Bayes model selection with path sampling: factor models and other examples". In: *Statistical Science* 28.1 (2013), pp. 95–115.

[Dal17a] Arnak Dalalyan. "Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, July 2017, pp. 678–689.

[Dal17b] Arnak S. Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676. ISSN: 1467-9868.

[DD09] Thomas Dean and Paul Dupuis. "Splitting for rare event simulation: A large deviation approach to design and analysis". In: *Stochastic Processes and their Applications* 119.2 (2009), pp. 562–587. ISSN: 0304-4149.

[DDB17] A. Dieuleveut, A. Durmus, and F. Bach. "Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains". In: *ArXiv e-prints* (July 2017).

[Del+16] P. Del Moral, A. Jasra, K. Law, and Y. Zhou. "Multilevel sequential Monte Carlo samplers for normalizing constants". In: *ArXiv e-prints* (Mar. 2016).

[Del04] P. Del Moral. *Feynman-Kac formulae*. Probability and its Applications (New York). Genealogical and interacting particle systems with applications. Springer-Verlag, New York, 2004, pp. xviii+555. ISBN: 0-387-20268-4.

[DF91] Martin Dyer and Alan Frieze. "Computing the volume of convex bodies: a case where randomness provably helps". In: *Probabilistic combinatorics and its applications* 44 (1991), pp. 123–170.

[DFG09] Randal Douc, Gersende Fort, and Arnaud Guillin. "Subgeometric rates of convergence of f-ergodic strong Markov processes". In: *Stochastic Processes and their Applications* 119.3 (2009), pp. 897–923. ISSN: 0304-4149.

[DGM18] Alain Durmus, Arnaud Guillin, and Pierre Monmarché. "Geometric ergodicity of the bouncy particle sampler". In: *arXiv e-prints*, arXiv:1807.05401 (July 2018), arXiv:1807.05401.

[Din+14] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, and Hartmut Neven. "Bayesian Sampling Using Stochastic Gradient Thermostats". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 3203–3211.

[DK12] P. Dellaportas and I. Kontoyiannis. "Control variates for estimation based on reversible Markov chain Monte Carlo samplers". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.1 (2012). ISSN: 1467-9868.

[DK17] A. S. Dalalyan and A. G. Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". In: *ArXiv e-prints 1710.00095* (Sept. 2017).

[DK19] Arnak S. Dalalyan and Avetik Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". In: *Stochastic Processes and their Applications* (2019). ISSN: 0304-4149.

[DLP16] A. B. Duncan, T. Lelièvre, and G. A. Pavliotis. "Variance Reduction Using Nonreversible Langevin Samplers". In: *Journal of Statistical Physics* 163.3 (May 2016), pp. 457–491. ISSN: 1572-9613.

[DM16] A. Durmus and E. Moulines. "High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm". In: *ArXiv e-prints* (May 2016).

[DM17] Alain Durmus and Éric Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". In: *Ann. Appl. Probab.* 27.3 (June 2017), pp. 1551–1587.

[DMM18] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. "Analysis of Langevin Monte Carlo via convex optimization". In: *arXiv e-prints*, arXiv:1802.09188 (Feb. 2018), arXiv:1802.09188.

[DMP18] A. Durmus, É. Moulines, and M. Pereyra. "Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau". In: *SIAM Journal on Imaging Sciences* 11.1 (2018), pp. 473–506.

[DMS17] Alain Durmus, Eric Moulines, and Eero Saksman. "On the convergence of Hamiltonian Monte Carlo". In: *arXiv e-prints*, arXiv:1705.00166 (Apr. 2017), arXiv:1705.00166.

[Dou+18] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Ed. by Springer. Springer Series in Operations Research and Financial Engineering. Springer, Dec. 2018. ISBN: 978-3319977034.

[DPZ17] A. B. Duncan, G. A. Pavliotis, and K. C. Zygalakis. "Nonreversible Langevin Samplers: Splitting Schemes, Analysis and Implementation". In: *ArXiv e-prints* (Jan. 2017).

[DR18] Arnak S. Dalalyan and Lionel Riou-Durand. *On sampling from a log-concave density using kinetic Langevin diffusions*. Submitted 1807.09382. arXiv, July 2018.

[DT12] A. S. Dalalyan and A. B. Tsybakov. "Sparse regression learning by aggregation and Langevin Monte-Carlo". In: *J. Comput. System Sci.* 78.5 (2012), pp. 1423–1443.

[Dua+87] Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (1987), pp. 216–222. ISSN: 0370-2693.

[Dub+16] Kumar Avinava Dubey, Sashank J. Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. "Variance Reduction in Stochastic Gradient Langevin Dynamics". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., 2016, pp. 1154–1162.

[Dwi+18] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. "Log-concave sampling: Metropolis-Hastings algorithms are fast!" In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 793–797.

[EG15] Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015.

[EGZ17] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. "Couplings and quantitative contraction rates for Langevin dynamics". In: *arXiv e-prints*, arXiv:1703.01617 (Mar. 2017), arXiv:1703.01617.

[EM18] Andreas Eberle and Mateusz B. Majka. "Quantitative contraction rates for Markov chains on general state spaces". In: *arXiv e-prints*, arXiv:1808.07033 (Aug. 2018), arXiv:1808.07033.

[EMS18] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. "Global Non-convex Optimization with Discretized Diffusions". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 9671–9680.

[Erm75] D. L Ermak. "A computer simulation of charged particles in solution. I. Technique and equilibrium properties". In: *The Journal of Chemical Physics* 62.10 (1975), pp. 4189–4196.

[FHS15] Max Fathi, Ahmed-Amine Homman, and Gabriel Stoltz. "Error analysis of the transport properties of Metropolized schemes". In: *ESAIM: Proc.* 48 (2015), pp. 341–363.

[FHW14] Nial Friel, Merrilee Hurn, and Jason Wyse. "Improving power posterior estimation of statistical evidence". In: *Statistics and Computing* 24.5 (2014), pp. 709–723. ISSN: 1573-1375.

[FJ10] James M. Flegal and Galin L. Jones. "Batch means and spectral variance estimators in Markov chain Monte Carlo". In: *Ann. Statist.* 38.2 (Apr. 2010), pp. 1034–1070.

[Fri12] Avner Friedman. *Stochastic differential equations and applications*. Courier Corporation, 2012.

[FW12] Nial Friel and Jason Wyse. "Estimating the evidence–a review". In: *Statistica Neerlandica* 66.3 (2012), pp. 288–308.

[Gel] Matthias Gelbrich. "On a Formula for the L2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces". In: *Mathematische Nachrichten* 147.1 (), pp. 185–203.

[Gel+14] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[GM94] U. Grenander and M. I. Miller. "Representations of knowledge in complex systems". In: *J. Roy. Statist. Soc. Ser. B* 56.4 (1994). With discussion and a reply by the authors, pp. 549–603. ISSN: 0035-9246.

[GM96] Peter W. Glynn and Sean P. Meyn. "A Liapounov bound for solutions of the Poisson equation". In: *Ann. Probab.* 24.2 (Apr. 1996), pp. 916–931.

[GM98] Andrew Gelman and Xiao-Li Meng. "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling". In: *Statistical science* (1998), pp. 163–185.

[Gor+16] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. "Measuring Sample Quality with Diffusions". In: *ArXiv e-prints* (Nov. 2016).

[GR14] Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 2014.

[Gre83] U. Grenander. "Tutorial in pattern theory". Division of Applied Mathematics, Brown University, Providence. 1983.

[GSL92] A. E. Gelfand, A. F. Smith, and T.-M. Lee. "Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling". In: *Journal of the American Statistical Association* 87.418 (1992), pp. 523–532.

[GT15] David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*. springer, 2015.

[Has+17] Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. "Distributed Bayesian Learning with Stochastic Natural Gradient Expectation Propagation and the Posterior Server". In: *Journal of Machine Learning Research* 18.106 (2017), pp. 1–37.

[Hen97] Shane G Henderson. "Variance reduction via an approximating markov process". Available at http://people.orie.cornell. edu/shane/pubs/thesis.pdf. PhD thesis. Department of Operations Research, Stanford University, 1997.

[HH14] Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

[HHS05] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. "Accelerating diffusions". In: *Ann. Appl. Probab.* 15.2 (May 2005), pp. 1433–1444.

[HHS93] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. "Accelerating Gaussian Diffusions". In: *Ann. Appl. Probab.* 3.3 (Aug. 1993), pp. 897–913.

[HJ15] Martin Hutzenthaler and Arnulf Jentzen. *Numerical approximations of stochastic differential equations with non-globally Lipschitz continuous coefficients*. Vol. 236. 1112. American Mathematical Society, 2015.

[HJ17] Jeremy Heng and Pierre E. Jacob. "Unbiased Hamiltonian Monte Carlo with couplings". In: *arXiv e-prints*, arXiv:1709.00404 (Sept. 2017), arXiv:1709.00404.

[HJK11] Martin Hutzenthaler, Arnulf Jentzen, and Peter E. Kloeden. "Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 467.2130 (2011), pp. 1563–1576. ISSN: 1364-5021.

[HJK12] Martin Hutzenthaler, Arnulf Jentzen, and Peter E. Kloeden. "Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients". In: *Ann. Appl. Probab.* 22.4 (Aug. 2012), pp. 1611–1641.

[HM11] M. Hairer and J. C. Mattingly. "Yet another look at Harris' ergodic theorem for Markov chains". In: *Seminar on Stochastic Analysis, Random Fields and Applications VI*. Vol. 63. Progr. Probab. Birkhäuser/Springer Basel AG, Basel, 2011, pp. 109–117.

[HMS02] Desmond J. Higham, Xuerong Mao, and Andrew M. Stuart. "Strong Convergence of Euler-Type Methods for Nonlinear Stochastic Differential Equations". In: *SIAM Journal on Numerical Analysis* 40.3 (2002), pp. 1041–1063.

[Hsi+18] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. "Mirrored Langevin Dynamics". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 2878–2887.

[Hub15] Mark Huber. "Approximation algorithms for the normalizing constant of Gibbs distributions". In: *Ann. Appl. Probab.* 25.2 (Apr. 2015), pp. 974–985.

[IW89] N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes*. North-Holland Mathematical Library. Elsevier Science, 1989.

[J+01] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001.

[JA06] Valen E Johnson and James H Albert. *Ordinal data modeling*. Springer Science & Business Media, 2006.

[Jar97] C. Jarzynski. "Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach". In: *Physical Review E* 56.5 (1997), p. 5018.

[Jas+16] Ajay Jasra, Kengo Kamatani, Prince Peprah Osei, and Yan Zhou. "Multilevel particle filters: normalizing constant estimation". In: *Statistics and Computing* (2016), pp. 1–14. ISSN: 1573-1375.

[JHS05] A. Jasra, C. C. Holmes, and D. A. Stephens. "Markov chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling". In: *Statist. Sci.* 20.1 (Feb. 2005), pp. 50–67.

[JO10] A. Joulin and Y. Ollivier. "Curvature, concentration and error estimates for Markov chain Monte Carlo". In: *Ann. Probab.* 38.6 (Nov. 2010), pp. 2418–2442.

[JVV86] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. "Random generation of combinatorial structures from a uniform distribution". In: *Theoretical Computer Science* 43 (1986), pp. 169–188. ISSN: 0304-3975.

[Kam09] Jürgen Kampf. "On weighted parallel volumes". In: *Beiträge Algebra Geom* 50.2 (2009), pp. 495–519.

[KCW14] Anoop Korattikara, Yutian Chen, and Max Welling. "Austerity in MCMC Land: cutting the Metropolis-hastings Budget". In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, 2014, pp. I-181–I-189.

[KM05] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[Knu+15] Kevin H. Knuth, Michael Habeck, Nabin K. Malakar, Asim M. Mubeen, and Ben Placek. "Bayesian evidence and model selection". In: *Digital Signal Processing* 47 (2015). Special Issue in Honour of William J. (Bill) Fitzgerald, pp. 50–67. ISSN: 1051-2004.

[Kop15] Marie Kopec. "Weak backward error analysis for overdamped Langevin processes". In: *IMA Journal of Numerical Analysis* 35.2 (2015), pp. 583–614.

[KR97] Daniel A Klain and Gian-Carlo Rota. *Introduction to geometric probability*. Cambridge University Press, 1997.

[KS91] I. Karatzas and S.E. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer New York, 1991. ISBN: 9780387976556.

[Kul97] S. Kullback. *Information theory and statistics*. Reprint of the second (1968) edition. Dover Publications, Inc., Mineola, NY, 1997, pp. xvi+399. ISBN: 0-486-69684-7.

[LAW16] Wenzhe Li, Sungjin Ahn, and Max Welling. "Scalable MCMC for mixed membership stochastic blockmodels". In: *Artificial Intelligence and Statistics*. 2016, pp. 723–731.

[LFR17] S. Livingstone, M. F. Faulkner, and G. O. Roberts. "Kinetic energy choice in Hamiltonian/hybrid Monte Carlo". In: *ArXiv e-prints* (June 2017).

[Li+16] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. "Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI'16. Phoenix, Arizona: AAAI Press, 2016, pp. 1788–1794.

[Liu08] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[LMS07] H. Lamba, J. C. Mattingly, and A. M. Stuart. "An adaptive Euler–Maruyama scheme for SDEs: convergence and stability". In: *IMA Journal of Numerical Analysis* 27.3 (2007), pp. 479–506.

[LMS16] Benedict Leimkuhler, Charles Matthews, and Gabriel Stoltz. "The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics". In: *IMA Journal of Numerical Analysis* 36.1 (2016), p. 13.

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[LS13] Robert Liptser and Albert N Shiryaev. *Statistics of random Processes: I. general Theory*. Vol. 5. Springer Science & Business Media, 2013.

[LS15] S. Lan and B. Shahbaba. "Sampling constrained probability distributions using Spherical Augmentation". In: *ArXiv e-prints* (June 2015).

[LS16] Tony Lelièvre and Gabriel Stoltz. "Partial differential equations and stochastic methods in molecular dynamics". In: *Acta Numerica* 25 (2016), pp. 681–880.

[LS90] L. Lovasz and M. Simonovits. "The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume". In: *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*. Oct. 1990, 346–354 vol. 1.

[LS93] L. Lovász and M. Simonovits. "Random walks in a convex body and an improved volume algorithm". In: *Random structures & algorithms* 4.4 (1993), pp. 359–412.

[LSR10] Tony Lelièvre, Gabriel Stoltz, and Mathias Rousset. *Free energy computations: A mathematical perspective*. World Scientific, 2010.

[LV06] L. Lovász and S. Vempala. "Hit-and-Run from a Corner". In: *SIAM Journal on Computing* 35.4 (2006), pp. 985–1005.

[LV07] László Lovász and Santosh Vempala. "The Geometry of Logconcave Functions and Sampling Algorithms". In: *Random Struct. Algorithms* 30.3 (May 2007), pp. 307–358. ISSN: 1042-9832.

[MCF15] Yi-An Ma, Tianqi Chen, and Emily Fox. "A Complete Recipe for Stochastic Gradient MCMC". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 2917–2925.

[MDJ06] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. "Sequential Monte Carlo Samplers". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 68.3 (2006), pp. 411–436. ISSN: 13697412, 14679868.

[Mey08] Sean Meyn. *Control techniques for complex networks*. Cambridge University Press, 2008.

[MR07] Jean-Michel Marin and Christian Robert. *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Science & Business Media, 2007.

[MR09] Jean-Michel Marin and Christian P Robert. "Importance sampling methods for Bayesian discrimination between embedded models". In: *arXiv preprint arXiv:0910.2325* (2009).

[MS17] Oren Mangoubi and Aaron Smith. "Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions". In: *arXiv e-prints*, arXiv:1708.07114 (Aug. 2017), arXiv:1708.07114.

[MSH02] J. C. Mattingly, A. M. Stuart, and D. J. Higham. "Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise". In: *Stochastic Process. Appl.* 101.2 (2002), pp. 185–232. ISSN: 0304-4149.

[MSI13] Antonietta Mira, Reza Solgi, and Daniele Imparato. "Zero variance Markov chain Monte Carlo for Bayesian estimators". In: *Statistics and Computing* 23.5 (2013), pp. 653–662. ISSN: 1573-1375.

[MST10] Jonathan C. Mattingly, Andrew M. Stuart, and M. V. Tretyakov. "Convergence of Numerical Time-Averaging and Stationary Measures via Poisson Equations". In: *SIAM Journal on Numerical Analysis* 48.2 (2010), pp. 552–577.

[MT09] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. 2nd. New York, NY, USA: Cambridge University Press, 2009. ISBN: 0521731828, 9780521731829.

[MT93] Sean P. Meyn and R. L. Tweedie. "Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes". In: *Adv. in Appl. Probab.* 25.3 (1993), pp. 518–548. ISSN: 0001-8678.

[MV15] A. Mijatovic and J. Vogrinc. "On the Poisson equation for Metropolis-Hastings chains". In: *ArXiv e-prints* (Nov. 2015).

[MV17] A. Mijatović and J. Vogrinc. "Asymptotic variance for Random Walk Metropolis chains in high dimensions: logarithmic growth via the Poisson equation". In: *ArXiv e-prints* (July 2017).

[MV18] Oren Mangoubi and Nisheeth Vishnoi. "Dimensionally Tight Bounds for Second-Order Hamiltonian Monte Carlo". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 6027–6037.

[Nag+17] T. Nagapetyan, A. B. Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. Zygalakis. "The True Cost of Stochastic Gradient Langevin Dynamics". In: *ArXiv e-prints 1706.02692* (June 2017).

[Nea01] Radford M. Neal. "Annealed importance sampling". In: *Statistics and Computing* 11.2 (2001), pp. 125–139. ISSN: 1573-1375.

[Nem+09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609.

[Nes13] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.

[NJ16] Balasubramanian Narasimhan and Steven G. Johnson. *cubature: Adaptive Multivariate Integration over Hypercubes*. R package version 1.3-6. 2016.

[NP09] Wojciech Niemiro and Piotr Pokarowski. "Fixed precision MCMC estimation by median of products of averages". In: *J. Appl. Probab.* 46.2 (June 2009), pp. 309–329.

[Oat+18] C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. "Convergence Rates for a Class of Estimators Based on Stein's Method". In: *Accepted in Bernoulli* (Mar. 2018).

[OG16] Chris Oates and Mark Girolami. "Control Functionals for Quasi-Monte Carlo Integration". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by Arthur Gretton and Christian C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, May 2016, pp. 56–65.

[OGC16] Chris J. Oates, Mark Girolami, and Nicolas Chopin. "Control functionals for Monte Carlo integration". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2016), n/a–n/a. ISSN: 1467-9868.

[OPG16] Chris J. Oates, Theodore Papamarkou, and Mark Girolami. "The Controlled Thermodynamic Integral for Bayesian Model Evidence Evaluation". In: *Journal of the American Statistical Association* 111.514 (2016), pp. 634–645.

[Par81] G. Parisi. "Correlation functions and computer simulations". In: *Nuclear Physics B* 180 (1981), pp. 378–384.

[Pav14] Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*. Vol. 60. Springer, 2014.

[PBJ14] John Paisley, David M Blei, and Michael I Jordan. "Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference". In: *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC, 2014, pp. 205–224.

[PC08] T. Park and G. Casella. "The Bayesian lasso". In: *J. Amer. Statist. Assoc.* 103.482 (2008), pp. 681–686. ISSN: 0162-1459.

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[Per16] Marcelo Pereyra. "Maximum-a-posteriori estimation with Bayesian confidence regions". In: *arXiv preprint arXiv:1602.08590* (2016).

[PMG14] Theodore Papamarkou, Antonietta Mira, and Mark Girolami. "Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms". In: *Bayesian Anal.* 9.1 (Mar. 2014), pp. 97–128.

[PP14] Ari Pakman and Liam Paninski. "Exact hamiltonian monte carlo for truncated multivariate gaussians". In: *Journal of Computational and Graphical Statistics* 23.2 (2014), pp. 518–542.

[PS76] F. Proschan and J. Sethuraman. "Stochastic comparisons of order statistics from heterogeneous populations, with applications in reliability". In: *Journal of Multivariate Analysis* 6.4 (1976), pp. 608–616. ISSN: 0047-259X.

[PT13] Sam Patterson and Yee Whye Teh. "Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., 2013, pp. 3102–3110.

[PV01] E. Pardoux and Yu. Veretennikov. "On the Poisson Equation and Diffusion Approximation. I". In: *Ann. Probab.* 29.3 (July 2001), pp. 1061–1085.

[R C18] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018.

[RC04] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Second. Springer Texts in Statistics. Springer-Verlag, New York, 2004, pp. xxx+645. ISBN: 0-387-21239-6.

[RDF78] P. J. Rossky, J. D. Doll, and H. L. Friedman. "Brownian dynamics as smart Monte Carlo simulation". In: *The Journal of Chemical Physics* 69.10 (1978), pp. 4628–4633.

[RDS04] Gabriel Rodriguez-Yam, Richard A Davis, and Louis L Scharf. "Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression". In: *Unpublished manuscript* (2004).

[RG97] Sylvia. Richardson and Peter J. Green. "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.4 (1997), pp. 731–792. ISSN: 1467-9868.

[RGG97] G. O. Roberts, A. Gelman, and W. R. Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *Ann. Applied Prob.* 7 (1997), pp. 110–120.

[RK17] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*. Wiley Series in Probability and Statistics. Third edition [of MR0624270]. John Wiley & Sons, Inc., Hoboken, NJ, 2017, pp. xvii+414. ISBN: 978-1-118-63216-1.

[Rob07] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[Roc15] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[RR98] Gareth O. Roberts and Jeffrey S. Rosenthal. "Optimal Scaling of Discrete Approximations to Langevin Diffusions". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 60.1 (1998), pp. 255–268. ISSN: 13697412, 14679868.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. "Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, July 2017, pp. 1674–1703.

[RS15a] Luc Rey-Bellet and Konstantinos Spiliopoulos. "Irreversible Langevin samplers and variance reduction: a large deviations approach". In: *Nonlinearity* 28.7 (May 2015), pp. 2081–2103.

[RS15b] Luc Rey-Bellet and Konstantinos Spiliopoulos. "Variance reduction for irreversible Langevin samplers and diffusion on graphs". In: *Electron. Commun. Probab.* 20 (2015), 16 pp.

[RS17] J. Roussel and G. Soltz. "A perturbative approach to control variates in molecular dynamics". In: *ArXiv e-prints* (Dec. 2017).

[RT96] G. O. Roberts and R. L. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* 2.4 (1996), pp. 341–363. ISSN: 1350-7265.

[RW98] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Vol. 317. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1998, pp. xiv+733. ISBN: 3-540-62772-3.

[Sab13] Sotirios Sabanis. "A note on tamed Euler approximations". In: *Electron. Commun. Probab.* 18 (2013), 10 pp.

[Sch13] Rolf Schneider. *Convex bodies: the Brunn–Minkowski theory*. 151. Cambridge University Press, 2013.

[Sch99] Christof Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules*. 1999.

[SN14] Issei Sato and Hiroshi Nakagawa. "Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Ito Process". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 982–990.

[SV07] Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes*. Springer, 2007.

[TT90] Denis Talay and Luciano Tubaro. "Expansion of the global error for numerical schemes solving stochastic differential equations". In: *Stochastic Analysis and Applications* 8.4 (1990), pp. 483–509.

[TTV16] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. "Consistency and fluctuations for stochastic gradient Langevin dynamics". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 193–225.

[TV17] Yin Tat Lee and Santosh S. Vempala. "Convergence Rate of Riemannian Hamiltonian Monte Carlo and Faster Polytope Volume Computation". In: *arXiv e-prints*, arXiv:1710.06261 (Oct. 2017), arXiv:1710.06261.

[VC72] J. P. Valleau and D. N. Card. "Monte Carlo Estimation of the Free Energy by Multistage Sampling". In: *The Journal of Chemical Physics* 57.12 (1972), pp. 5457–5462.

[Vil09] C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Berlin: Springer, 2009. ISBN: 978-3-540-71049-3.

[VZT16] Sebastian J. Vollmer, Konstantinos C. Zygalakis, and Yee Whye Teh. "Exploration of the (Non-)Asymptotic Bias and Variance of Stochastic Gradient Langevin Dynamics". In: *Journal of Machine Learning Research* 17.159 (2016), pp. 1–48.

[WHC14] Sheng-Jhih Wu, Chii-Ruey Hwang, and Moody T. Chu. "Attaining the Optimal Gaussian Diffusion Acceleration". In: *Journal of Statistical Physics* 155.3 (May 2014), pp. 571–590. ISSN: 1572-9613.

[Wib18] Andre Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 2093–3027.

[WT11] Max Welling and Yee Whye Teh. "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 681–688. ISBN: 978-1-4503-0619-5.

[WW59] Evan James Williams and EJ Williams. *Regression analysis*. Vol. 14. wiley New York, 1959.

[Wys11] Jason Wyse. *Estimating the statistical evidence - a review*. 2011.

[ZJA15] Yan Zhou, Adam M Johansen, and John AD Aston. "Towards automatic model comparison: an adaptive sequential Monte Carlo approach". In: *Journal of Computational and Graphical Statistics* just-accepted (2015).

[ZLC17] Yuchen Zhang, Percy Liang, and Moses Charikar. "A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. Amsterdam, Netherlands: PMLR, July 2017, pp. 1980–2022.

[ZM96] D. L. Zhu and P. Marcotte. "Co-Coercivity and Its Role In the Convergence of Iterative Schemes For Solving Variational Inequalities". In: *SIAM J. on Optimization* 6.3 (Mar. 1996), pp. 714–726. ISSN: 1052-6234.

List of Figures

List of Tables

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Autour de l'algorithme du Langevin en grande dimension: extensions et applications.

Mots Clefs : Méthodes de Monte Carlo par Chaînes de Markov, algorithme du Langevin, simulation, statistiques bayésiennes

Résumé : Cette thèse porte sur le problème de l'échantillonnage en grande dimension et est basée sur l'algorithme de Langevin non ajusté (ULA). Dans une première partie, nous proposons deux extensions d'ULA et fournissons des garanties de convergence précises pour ces algorithmes. Nous nous intéressons en particulier aux cas où la distribution cible est à support compact et lorsque les queues de la distribution cible sont trop fines. Dans une deuxième partie, nous donnons deux applications d'ULA. Nous fournissons un algorithme pour estimer les constantes de normalisation de densités log concaves. En comparant ULA avec la diffusion de Langevin, nous développons une nouvelle méthode de variables de contrôle basée sur la variance asymptotique de la diffusion de Langevin. Dans une troisième partie, nous analysons Stochastic Gradient Langevin Dynamics (SGLD), qui diffère de ULA seulement dans l'estimation stochastique du gradient.

Title : Around the Langevin algorithm in high dimension: extensions and applications

Keys words : Markov Chain Monte Carlo, Langevin algorithm, simulation, Bayesian statistics

Abstract : This thesis addresses the problem of sampling in high dimension and is based on the unadjusted Langevin algorithm (ULA). In the first part, we propose two extensions of ULA and provide precise convergence guarantees for these algorithms. We are particularly interested in cases where the target distribution is compactly supported and the tails of the target distribution are too thin. In a second part, we give two applications of ULA. We provide an algorithm to estimate the normalisation constants of log-concave densities. By comparing ULA with the Langevin diffusion, we develop a new method of control variates based on the asymptotic variance of the Langevin diffusion. In a third part, we analyze Stochastic Gradient Langevin Dynamics (SGLD), which differs from ULA only in the stochastic estimation of the gradient.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France