Journal of Machine Learning Research 20 (2019) 1-46 Submitted 3/18; Revised 1/19; Published 2/19

Analysis of Langevin Monte Carlo via Convex Optimization

Alain Durmus ALAIN.DURMUS@CMLA.ENS-CACHAN.FR
CMLA - École normale supérieure Paris-Saclay,
CNRS, Université Paris-Saclay, 94235 Cachan, France

Szymon Majewski SMAJEWSKI@IMPAN.PL
Institute of Mathematics, Polish Academy of Sciences
ul. Śniadeckich 8, 00-656 Warszawa, Poland

Błażej Miasojedow B.MIASOJEDOW@MIMUW.EDU.PL
Institute of Applied Mathematics and Mechanics,
University of Warsaw, ul. Banacha 2, 02-097 Warszawa, Poland

Editor: Francois Caron

Abstract

In this paper, we provide new insights on the Unadjusted Langevin Algorithm. We show that this method can be formulated as the first order optimization algorithm for an objective functional defined on the Wasserstein space of order 2. Using this interpretation and techniques borrowed from convex optimization, we give a non-asymptotic analysis of this method to sample from log-concave smooth target distribution on $\mathbb{R}^d$. Based on this interpretation, we propose two new methods for sampling from a non-smooth target distribution. These new algorithms are natural extensions of the Stochastic Gradient Langevin Dynamics (SGLD) algorithm, which is a popular extension of the Unadjusted Langevin Algorithm for largescale Bayesian inference. Using the optimization perspective, we provide non-asymptotic convergence analysis for the newly proposed methods.

Keywords: Unadjasted Langevin Algorithm, convex optimization, Bayesian inference, gradient flow, Wasserstein metric

1. Introduction

This paper deals with the problem of sampling from a probability measure π on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which admits a density, also denoted by π, with respect to the Lebesgue measure given for all $x \in \mathbb{R}^d$ by

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

where $U : \mathbb{R}^d \to \mathbb{R}$. This problem arises in various fields such that Bayesian statistical inference (Gelman et al., 2014), machine learning (Andrieu et al., 2003), ill-posed inverse problems (Stuart, 2010), and computational physics (Krauth, 2006). Common and current methods to tackle this problem are Markov Chain Monte Carlo methods (Brooks et al., 2011), for example the Hastings-Metropolis algorithm (Metropolis et al., 1953; Hastings, 1970) or Gibbs sampling (Geman and Geman, 1984). All these methods boil down to building a Markov kernel on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ whose invariant probability distribution is π. Yet, choosing an appropriate proposal distribution for the Hastings-Metropolis algorithm is a

tricky subject. For this reason, it has been proposed to consider continuous dynamics which naturally leave the target distribution π invariant. Perhaps, one of the most famous examples is the over-damped Langevin diffusion (Rossky et al., 1978; Parisi, 1981) associated with U, which is assumed to be continuously differentiable:

$$\mathrm{d}\mathbf{Y}_t = -\nabla U(\mathbf{Y}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \;, \tag{1}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. With appropriate conditions on U, this SDE admits a unique strong solution $(\mathbf{Y}_t)_{t\geq 0}$ and defines a strong Markov semi-group $(P_t)_{t\geq 0}$ which converges to π in total variation (Roberts and Tweedie, 1996, Theorem 2.1) or Wasserstein distance (Bolley et al., 2012). However, simulating exact solutions of such stochastic differential equations is not possible in most cases, and discretizations of these equations are used instead. In addition, numerical solutions associated with these schemes define Markov kernels for which π is not invariant anymore. Therefore quantifying the error introduced by these approximations is crucial to justify their use to sample from the target π. We consider in this paper the Euler-Maruyama discretization of (1) which defines the (possibly inhomogeneous) Markov chain $(X_k)_{k\geq 0}$ given for all $k \geq 0$ by

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

where $(\gamma_k)_{k\geq 1}$ is a sequence of stepsizes which can be held constant or converges to 0, and $(G_k)_{k\geq 1}$ is a sequence of i.i.d. standard d-dimensional Gaussian random variables. The use of the Euler-Maruyama discretization (2) to approximatively sample from π is referred to as the Unadjusted Langevin Algorithm (ULA) (or the Langevin Monte Carlo algorithm (LMC)), and has already been the matter of many works. For example, weak error estimates have been obtained in Talay and Tubaro (1990); Mattingly et al. (2002) for the constant stepsize setting and in Lamberton and Pagès (2003); Lemaire (2005) when $(\gamma_k)_{k\geq 1}$ is non-increasing and goes to 0. Explicit and non-asymptotic bounds on the total variation (Dalalyan, 2016; Durmus and Moulines, 2017) or the Wasserstein distance (Durmus and Moulines, 2016) between the distribution of X_k and π have been obtained. Roughly, all these results are based on the comparison between the discretization and the diffusion process and the quantification of error accumulation throughout the algorithm. In this paper, we propose another point of view on ULA, which shares nevertheless some relations with the Langevin diffusion (1). Indeed, it has been shown in Jordan et al. (1998) that the family of distributions $(\mu_0 P_t)_{t\geq 0}$, is the solution of a gradient flow equation in the Wasserstein space of order 2 associated with the Kullback-Leibler (KL) divergence, $\mathscr{F}$, where $(P_t)_{t\geq 0}$ is the semi-group associated with (1) and μ_0 is a probability measure on $\mathcal{B}(\mathbb{R}^d)$ admitting a second moment (see Section 2). If π is invariant for $(P_t)_{t\geq 0}$, then it is a stationary solution of this equation, and is the unique minimizer of $\mathscr{F}$ if U is convex. Starting from this observation, we interpret ULA as the first order optimization algorithm on the Wasserstein space of order 2 with objective functional $\mathscr{F}$. Namely, we adapt some proofs of convergence for the gradient descent algorithm from the convex optimization literature to obtain non-asymptotic and explicit bounds between the Kullback-Leibler (KL) divergence from π to distributions of averaged distributions associated with ULA for the constant and non-increasing stepsize setting. Then, these bounds easily imply computable bounds in total variation norm and Wasserstein distance. Note that these two metrics are different in nature since convergence

in one of them does not imply convergence in the other. Convergence in one of these metrics implies a control on the bias of MCMC based estimators of the form $\hat{f}_n = n^{-1}\sum_{k=1}^n f(Y_k)$, where $(Y_k)_{k\in\mathbb{N}}$ is Markov chain ergodic with respect to the target density π, for f belonging to a certain class of functions. In the case of the total variation distance, this class is the set of measurable and bounded functions, in the case of the Wasserstein distance, it is the set of Lipschitz functions. If the potential U is strongly convex and gradient Lipschitz, we get back the results of Durmus and Moulines (2017, 2016); Cheng and Bartlett (2017), when the stepsize is held constant in (2). In the case where U is only convex and from a warm start, we get a bound on the complexity for ULA of order $d\bar{\mathcal{O}}(\varepsilon^{-2})$ and $d\bar{\mathcal{O}}(\varepsilon^{-4})$ to get one sample distributed close from π with accuracy $\varepsilon > 0$, in KL divergence and total variation distance respectively. Overview of bounds on computational complexity of ULA is presented in Table 1.

	Strongly convex U		Convex U			
	Best	Ours	Warm start		Minimizer of U	
			Best	Ours	Best	Ours
TV	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	$d\bar{\mathcal{O}}(\varepsilon^{-6})$	$d\bar{\mathcal{O}}(\varepsilon^{-4})$	$d^5\bar{\mathcal{O}}(\varepsilon^{-2})$	$d^3\bar{\mathcal{O}}(\varepsilon^{-4})$
Wasserstein	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	–	–	–	–
KL	$d\bar{\mathcal{O}}(\varepsilon^{-1})$	$d\bar{\mathcal{O}}(\varepsilon^{-1})$	$d\bar{\mathcal{O}}(\varepsilon^{-3})$	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	–	$d^3\bar{\mathcal{O}}(\varepsilon^{-2})$

Table 1: Overview of bounds on computational complexity of ULA, with constant stepsize. We present complexity to get one sample distributed close from π with accuracy $\varepsilon > 0$ in Wasserstein distance (Wasserstein), total variation distance (TV), and Kullback Leibler divergence (KL). We compare the best results (Best) from literature with ours (Ours) in the strongly convex and convex cases. In the convex case we consider two possible initial measures: a warm start, *i.e.* $W_2^2(\mu_0, \pi) \leq C$, for some absolute constant $C \geq 0$ (Warm start) and starting from Dirac delta at a minimizer of U, $x^\star$.

In addition, we propose two new algorithms to sample from a class of non-smooth log-concave distributions for which we derive computable non-asymptotic bounds as well. The first one can be applied to Lipschitz convex potential for which unbiased estimates of subgradients are available. Remarkably, the bounds we obtain for this algorithm depend on the dimension only through the initial condition and the variance of the stochastic subgradient estimates. Precisely, we get a bound on the complexity to get a sample with distribution close from π with accuracy $\varepsilon > 0$, of order $(M^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-2})$ in the case of the KL divergence and $(M^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-4})$ in the case of the total variation distance, where M is Lipschitz constant of the potential U and D^2 is a bound on the variance of the considered stochastic subgradient.

The second method we propose is a generalization of the Stochastic Gradient Langevin Dynamics algorithm (Welling and Teh, 2011), which extends ULA by replacing the gradient

with a sequence of i.i.d. unbiased estimators. In this new scheme, we assume that U can be decomposed as the sum of two functions U_1 and U_2, where U_1 is at least continuously differentiable and U_2 is only convex, and use stochastic gradient estimates for U_1 and the proximal operator associated with U_2. This new method is close to the one proposed in Durmus et al. (2018) but contrary on this work we do not need to approximate U_2 by its Moreau envelope. To get computable bounds from the target distribution π, we interpret this algorithm as a first order optimization algorithm and provide explicit bounds between the KL divergence from π to distributions associated with SGLD. In the case where U is strongly convex and gradient Lipschitz (*i.e.* $U_2 = 0$), we get back the same complexity as Dalalyan and Karagulyan (2017) which is of order $d\bar{\mathcal{O}}(\varepsilon^{-2})$ for the Wasserstein distance. We obtain the same complexity for the total variation distance and a complexity of order $d\bar{\mathcal{O}}(\varepsilon^{-1})$ for the KL divergence . In the case where U is only convex, not necessarily smooth (*i.e.* U_2 could be non-smooth), and from a warm start, we get a complexity of order $d\bar{\mathcal{O}}(\varepsilon^{-2})$ and $d\bar{\mathcal{O}}(\varepsilon^{-4})$ to get one sample distributed close from π with accuracy $\varepsilon > 0$ in KL divergence and total variation distance respectively. Overview of bounds for SGLD is presented in Table 2.

Extensive studies have also analyzed SGLD in a general setting, *i.e.* the potential U is not necessarily convex. In Vollmer et al. (2016) and Nagapetyan et al. (2017), a study of this scheme is done by weak error estimates. Finally, Raginsky et al. (2017) and Xu et al. (2017) gives some results regarding the potential use of SGLD as an optimization algorithm to minimize the potential U by targeting a target density proportional to $x \mapsto \mathrm{e}^{-\beta U(x)}$ for some $\beta > 0$.

	Strongly convex U		Convex U			
			Warm start		Minimizer of U	
	Best	Our	Best	Ours	Best	Ours
TV	–	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	–	$d\bar{\mathcal{O}}(\varepsilon^{-4})$	–	$d^3\bar{\mathcal{O}}(\varepsilon^{-4})$
Wasserstein	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	–	–	–	–
KL	–	$d\bar{\mathcal{O}}(\varepsilon^{-1})$	–	$d\bar{\mathcal{O}}(\varepsilon^{-2})$	–	$d^3\bar{\mathcal{O}}(\varepsilon^{-2})$

Table 2: Overview of bounds on computational complexity of SGLD, with constant stepsize. We present complexity to get one sample distributed close from π with accuracy $\varepsilon > 0$ in Wasserstein distance (Wasserstein), total variation distance (TV), and Kullback Leibler divergence (KL). We compare the best results (Best) from literature with ours (Ours) in the strongly convex and convex cases. In the convex case we consider two possible initial measures: warm start, *i.e.* $W_2^2(\mu_0, \pi) \leq C$, for some absolute constant $C \geq 0$ (Warm start) and starting from Dirac delta at a minimizer of U, $x^\star$.

In summary, our contributions are the following:

- We give a new interpretation of ULA and use it to get bounds on the KL divergence from π to the iterates of ULA. We recover the dependence on the dimension of Cheng and Bartlett (2017, Theorem 3) in the strongly convex case and get tighter bounds. Note that this result implies previously known bounds between π and ULA in Wasserstein distance and the total variation distance but with a completely different technique. We also give computable bounds when U is only convex which improves the results of Durmus and Moulines (2017); Dalalyan (2016); Cheng and Bartlett (2017).
- We give two new methodologies to sample from a non-smooth potential U and make a non-asymptotic analysis of them. These two new algorithms are generalizations of SGLD.

The paper is organized as follows. In Section 2, we give some intuition on the strategy we take to analyze ULA and its variants. These ideas come from gradient flow theory in Wasserstein space. In Section 3, we give the main results we obtain on ULA and their proof. In Section 4, two variants of ULA are presented and analyzed. Finally, numerical experiments on logistic regression models are presented in Section 5 to support our theoretical findings regarding our new methodologies.

1.1. Notations and Conventions

Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$, Leb the Lebesgue measure on $\mathcal{B}(\mathbb{R}^d)$, $\mathbb{F}(\mathbb{R}^d)$ the set of all Borel measurable functions on $\mathbb{R}^d$ and for $f \in \mathbb{F}(\mathbb{R}^d)$, $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$. For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $f \in \mathbb{F}(\mathbb{R}^d)$ a μ-integrable function, denote by $\mu(f)$ the integral of f w.r.t. μ. Let μ and ν be two sigma-finite measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Denote by $\mu \ll \nu$ if μ is absolutely continuous w.r.t. ν and $\mathrm{d}\mu/\mathrm{d}\nu$ the associated density. Let μ, ν be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Define the Kullback-Leibler (KL) divergence of μ from ν by

$$\mathrm{KL}\left(\mu|\nu\right) = \begin{cases} \int_{\mathbb{R}^d} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\nu(x) \;, & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise .} \end{cases}$$

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 set A of $\mathbb{R}^d$, $\zeta(\mathsf{A} \times \mathbb{R}^d) = \mu(\mathsf{A})$ and $\zeta(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})$. 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 2 as

$$W_2(\mu, \nu) = \left(\inf_{\zeta \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\zeta(x, y) \right)^{1/2} . \tag{3}$$

By (Villani, 2009, Theorem 4.1), for all μ, ν probability measures 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 $\zeta^\star$, $W_2(\mu, \nu) = \mathbb{E}[\|X - Y\|^2]^{1/2}$. This kind of transference plan (respectively coupling)

will be called an optimal transference plan (respectively optimal coupling) associated with W_2. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures with finite second moment: for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\mu(x) < +\infty$. By (Villani, 2009, Theorem 6.16), $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the Wasserstein distance W_2 of order 2 is a complete separable metric space. Denote by $\mathcal{P}^{\mathrm{a}}(\mathbb{R}^d) = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mu \ll \mathrm{Leb}\}$. For two probability measures μ and ν on $\mathbb{R}^d$, the total variation distance distance between μ and ν is defined by $\|\mu - \nu\|_{\mathrm{TV}} = \sup_{\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)} |\mu(\mathsf{A}) - \nu(\mathsf{A})|$.

Let $n \in \mathbb{N} \cup \{\infty\}$ and $\mathsf{U} \subset \mathbb{R}^d$ be an open set of $\mathbb{R}^d$. Denote by $C^n(\mathsf{U})$ the set of n-th continuously differentiable function from U to $\mathbb{R}$. Denote by $C_c^n(\mathsf{U})$ the set of n-th continuously differentiable function from U to $\mathbb{R}$ with compact support. Let $I \subset \mathbb{R}$ be an interval and $f : I \to \mathbb{R}$. f is absolutely continuous on I if for all $\varepsilon > 0$, there exists $\delta > 0$ such that for all $n \in \mathbb{N}^*$ and $t_1, \ldots, t_{2n} \in I$, $t_1 \leq \cdots \leq t_{2n}$,

$$\text{if } \sum_{k=1}^{n} \{t_{2k} - t_{2k-1}\} \leq \delta \text{ then } \sum_{k=1}^{n} |f(t_{2k}) - f(t_{2k-1})| \leq \varepsilon \, .$$

In the sequel, we take the convention that $\sum_p^n = 0$ and $\prod_p^n = 1$ for $n, p \in \mathbb{N}$, $n < p$.

2. Interpretation of ULA as an Optimization Algorithm

Throughout this paper, we assume that U satisfies the following condition for $m \geq 0$.

A 1 (m) $U : \mathbb{R}^d \to \mathbb{R}$ is m-convex, *i.e.* for all $x, y \in \mathbb{R}^d$,

$$U(tx + (1-t)y) \leq tU(x) + (1-t)U(y) - t(1-t)(m/2) \|x - y\|^2$$

Note that $\mathbf{A}1(m)$ includes the case where U is only convex when $m = 0$. We consider in this Section the following additional condition on U which will be relaxed in Section 4.

A 2 U is continuously differentiable and L-gradient Lipschitz, *i.e.* there exists $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $\|\nabla U(x) - \nabla U(y)\| \leq L \|x - y\|$

Under $\mathbf{A}\,1$ and $\mathbf{A}\,2$, the Langevin diffusion (1) has a unique strong solution $(\mathbf{Y}_t^x)_{t \geq 0}$ starting at $x \in \mathbb{R}^d$. The Markovian semi-group $(P_t)_{t \geq 0}$, given for all $t \geq 0$, $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by $P_t(x, \mathsf{A}) = \mathbb{P}(\mathbf{Y}_t^x \in \mathsf{A})$, is reversible with respect to π and π is its unique invariant probability measure, see (Ambrosio et al., 2009, Theorem 1.2 and Theorem 1.6). Using this probabilistic framework, (Roberts and Tweedie, 1996, Theorem 1.2) shows that $(P_t)_{t \geq 0}$ is irreducible with respect to the Lebesgue measure, strong Feller and $\lim_{t \to +\infty} \|P_t(x, \cdot) - \pi\|_{\mathrm{TV}} = 0$ for all $x \in \mathbb{R}^d$. But to study the properties of the semi-group $(P_t)_{t \geq 0}$, an other complementary and significant approach can be used. This dual point of view is based on the adjoint of the infinitesimal generator associated with $(P_t)_{t \geq 0}$. The *strong* generator of (1) $(\mathcal{A}, \mathrm{D}(\mathcal{A}))$ is defined for all $f \in \mathrm{D}(\mathcal{A})$ and $x \in \mathbb{R}^d$ by

$$\mathcal{A}f(x) = \lim_{t \to 0} t^{-1}(P_t f(x) - f(x)) \, ,$$

where $\mathrm{D}(\mathcal{A})$ is the subset of $C_0(\mathbb{R}^d)$ such that for all $f \in \mathrm{D}(\mathcal{A})$, there exists $g \in C_0(\mathbb{R}^d)$ such that $\lim_{t \to 0} \left\| t^{-1}(P_t f - f) - g \right\|_\infty = 0$. In particular for $f \in C_c^2(\mathbb{R}^d)$, we get by Itô's formula

$$\mathcal{A}f = \langle \nabla f, \nabla U \rangle + \Delta f \, .$$

In addition, by (Ethier and Kurtz, 1986, Proposition 1.5), for all $f \in C_c^2(\mathbb{R}^d)$, $P_t f(x) \in \mathrm{D}(\mathcal{A})$ and for $x \in \mathbb{R}^d$, $t \mapsto P_t f(x)$ is continuously differentiable,

$$\frac{\mathrm{d}P_t f(x)}{\mathrm{d}t} = \mathcal{A}P_t f(x) = P_t \mathcal{A} f(x) \; . \tag{4}$$

For all $\mu_0 \in \mathcal{P}_2^{\mathrm{a}}(\mathbb{R}^d)$ and $t > 0$, by Girsanov's Theorem (Karatzas and Shreve, 1991, Theorem 5.1, Corollary 5.16, Chapter 3), $\mu_0 P_t(\cdot)$ admits a density with respect to the Lebesgue measure denoted by ρ_t. This density is solution by (4) of the Fokker-Planck equation (in the weak sense):

$$\frac{\partial \rho_t}{\partial t} = \mathrm{div}(\nabla \rho_t + \rho_t \nabla U(x)) \; ,$$

meaning that for all $\phi \in C_c^\infty(\mathbb{R}^d)$ and $t > 0$,

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^d} \phi(y) \rho_t(\mathrm{d}y) = \int_{\mathbb{R}^d} \mathcal{A}\phi(y) \, \rho_t(\mathrm{d}y) \; . \tag{5}$$

In the landmark paper Jordan et al. (1998), the authors shows that if U is infinitely continuously differentiable, $(\rho_t)_{t>0}$ is the limit of the minimization scheme which defines a sequence of probability measures $(\tilde{\rho}_{k,\gamma})_{k\in\mathbb{N}}$ as follows. For $\gamma > 0$ set $\rho_{0,\gamma} = \mathrm{d}\mu_0/\mathrm{d}\,\mathrm{Leb}$ and

$$\tilde{\rho}_{k+1,\gamma} = \frac{\mathrm{d}\tilde{\mu}_{k+1,\gamma}}{\mathrm{d}\,\mathrm{Leb}} \; , \qquad \tilde{\mu}_{k+1,\gamma} = \operatorname*{argmin}_{\mu \in \mathcal{P}_2^{\mathrm{a}}(\mathbb{R}^d)} \; W_2^2(\tilde{\mu}_{k,\gamma}, \mu)/2 + \gamma \mathscr{F}(\mu) \; , \; k \in \mathbb{N} \; , \tag{6}$$

where $\mathscr{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is the free energy functional,

$$\mathscr{F} = \mathscr{H} + \mathscr{E} \; , \tag{7}$$

$\mathscr{H}, \mathscr{E} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ are the Boltzmann H-functional and the potential energy functional, given for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$\mathscr{H}(\mu) = \begin{cases} \int_{\mathbb{R}^d} \frac{\mathrm{d}\mu}{\mathrm{d}\,\mathrm{Leb}}(x) \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\,\mathrm{Leb}}(x)\right) \mathrm{d}x & \text{if } \mu \ll \mathrm{Leb} \\ +\infty \text{ otherwise} \; , \end{cases} \tag{8}$$

$$\mathscr{E}(\mu) = \int_{\mathbb{R}^d} U(x) \mathrm{d}\mu(x) \; . \tag{9}$$

More precisely, setting $\bar{\rho}_{0,\gamma} = \mathrm{d}\mu_0/\mathrm{d}\,\mathrm{Leb}$ and $\bar{\rho}_{t,\gamma} = \tilde{\rho}_{k,\gamma}$ for $t \in [k\gamma, (k+1)\gamma)$, (Jordan et al., 1998, Theorem 5.1) shows that for all $t > 0$, $\bar{\rho}_{t,\gamma}$ converges to ρ_t weakly in $\mathrm{L}^1(\mathbb{R}^d)$ as γ goes to 0. Note that the minimization scheme in (6) can be seen as a proximal type algorithm (see Martinet (1970) and Rockafeller (1976)) on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ used to minimize the functional $\mathscr{F}$. On $\mathbb{R}^d$, for continuous convex function f the proximal update with step size γ corresponds to one step of backward Euler discretization of the gradient flow ordinary differential equation (ODE) $\mathrm{d}x(t)/\mathrm{d}t = -\nabla f(x(t))$ with parameter γ. Therefore piecewise constant functions $(\bar{\rho}_{t,\gamma})_{\gamma>0}$ can be interpreted as backward Euler discretizations of an informal ODE $\partial \mu_t/\partial t = -\nabla \mathscr{F}(\mu_t)$ and their limit as $\gamma \to 0$, can be interpreted as a solution to this equation. This idea has been formalized and extended

to construct the framework of gradient flows in the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, see Ambrosio et al. (2008). We provide a short introduction to this topic in Section A and present useful concepts and results for our proofs.

The following lemma shows that π is the unique minimizer of $\mathscr{F}$. As a result, the distribution of the Langevin diffusion is the steepest descent flow of $\mathscr{F}$ and we get back intuitively that this process converges to the target distribution π.

Lemma 1 *Assume* $\boldsymbol{A}\,1(0)$. *The following holds:*

a) $\pi \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathscr{E}(\pi) < +\infty$ *and* $\mathscr{H}(\pi) < +\infty$.

b) For all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *satisfying* $\mathscr{E}(\mu) < \infty$

$$\mathscr{F}(\mu) - \mathscr{F}(\pi) = \mathrm{KL}\left(\mu|\pi\right) \ . \tag{10}$$

Proof The proof is postponed to Section 7.1. ■

Based on this interpretation, we could think about minimizing $\mathscr{F}$ on the Wasserstein space to get close to π using the minimization scheme (6). However, while this scheme is shown in Jordan et al. (1998) to be well-defined, finding explicit recursions $(\tilde{\rho}_{k,\gamma})_{k\in\mathbb{N}}$ is as difficult as minimizing $\mathscr{F}$ and can not be used in practice. In addition, to the authors knowledge, there is no efficient and practical schemes to optimize this functional. On the other hand, discretization schemes have been used to approximate the Langevin diffusion $(\mathbf{Y}_t)_{t\geq 0}$ (1) and its long-time behavior. One of the most popular method is the Euler-Maruyama discretization $(X_k)_{k\in\mathbb{N}}$ given in (2). While most work study the theoretical properties of this discretization to ensure to get samples close to the target distribution π, by comparing the distributions of $(X_k)_{k\in\mathbb{N}}$ and $(\mathbf{Y}_t)_{t\geq 0}$ through couplings or weak error expansions, we interpret this scheme as the first order optimization algorithm for the objective functional $\mathscr{F}$.A similar approach has been recently applied in Wibisono (2018) and in Bernton (2018). We postpone the comparison of our contributions and the results of this two papers in Section 4.2 where it is the most relevant.

3. Main Results for the Unadjusted Langevin Algorithm

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex continuously differentiable objective function with $x_f \in \arg\min_{\mathbb{R}^d} f \neq \emptyset$. The *inexact* or *stochastic* gradient descent algorithm used to estimate $f(x_f)$ defines the sequence $(x_k)_{k\in\mathbb{N}}$ starting from $x_0 \in \mathbb{R}^d$ by the following recursion for $k \in \mathbb{N}$:

$$x_{k+1} = x_k - \gamma_{k+1}\nabla f(x_k) + \gamma_{k+1}\Xi(x_k) \ ,$$

where $(\gamma_k)_{k\in\mathbb{N}^*}$ is a non-increasing sequence of stepsizes and $\Xi : \mathbb{R}^d \to \mathbb{R}^d$ is a deterministic or/and stochastic perturbation of ∇f. To get explicit bound on the convergence (in expectation) of the sequence $(f(x_k))_{k\in\mathbb{N}}$ to $f(x_f)$, one possibility (see e.g. Beck and Teboulle (2009)) is to show that the following inequality holds: for all $k \in \mathbb{N}$,

$$2\gamma_{k+1}(f(x_{k+1}) - f(x_f)) \leq \|x_k - x_f\|^2 - \|x_{k+1} - x_f\|_2^2 + C\gamma_{k+1}^2 \ , \tag{11}$$

for some constant $C \geq 0$. In a similar manner as for inexact gradient algorithms, in this section we will establish that ULA satisfies an inequality of the form (11) with the objective function $\mathscr{F}$ defined by (7) on $\mathcal{P}_2(\mathbb{R}^d)$, but instead of the Euclidean norm, the Wasserstein distance of order 2 will be used.

Consider the family of Markov kernels $(R_{\gamma_k})_{k\in\mathbb{N}^*}$ associated with the Euler-Maruyama discretization $(X_k)_{k\in\mathbb{N}}$, (2), for a sequence of stepsizes $(\gamma_k)_{k\in\mathbb{N}^*}$, given for all $\gamma > 0, x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$R_\gamma(x, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(-\left\|y - x - \gamma\nabla U(x)\right\|^2 / (4\gamma)\right) \mathrm{d}y \ . \tag{12}$$

Proposition 2 *Assume* ***A*** *1(m) for $m \geq 0$ and* ***A*** *2. For all $\gamma \in (0, L^{-1}]$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we have*

$$2\gamma \left\{\mathscr{F}(\mu R_\gamma) - \mathscr{F}(\pi)\right\} \leq (1 - m\gamma) W_2^2(\mu, \pi) - W_2^2(\mu R_\gamma, \pi) + 2\gamma^2 L d \ , \tag{13}$$

where $\mathscr{F}$ is defined in (7)*.*

The main difficulty in establishing Proposition 2 is to deal with the entropy function $\mathscr{H}$ defined by (8) in $\mathscr{F}$. To obtain the desired result, we decompose R_γ for all $\gamma > 0$ in the product of two elementary kernels S_γ and T_γ given for all $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$S_\gamma(x, \mathsf{A}) = \delta_{x - \gamma\nabla U(x)}(\mathsf{A}) \ , \ T_\gamma(x, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(-\left\|y - x\right\|^2 / (4\gamma)\right) \mathrm{d}y \ . \tag{14}$$

We take the convention that $S_0 = T_0 = \mathrm{Id}$ is the identity kernel given for all $x \in \mathbb{R}^d$ by $\mathrm{Id}(x, \{x\}) = 1$. S_γ is the deterministic part of the Euler-Maruyama discretization, which corresponds to gradient descent step relative to U for the $\mathscr{E}$ functional, whereas T_γ is the random part, that corresponds to going along the gradient flow of $\mathscr{H}$. Note then $R_\gamma = S_\gamma T_\gamma$ and consider the following decomposition

$$\mathscr{F}(\mu R_\gamma) - \mathscr{F}(\pi) = \mathscr{E}(\mu R_\gamma) - \mathscr{E}(\mu S_\gamma) + \mathscr{E}(\mu S_\gamma) - \mathscr{E}(\pi) + \mathscr{H}(\mu R_\gamma) - \mathscr{H}(\pi) \ . \tag{15}$$

The proof of Proposition 2 then consists in bounding each difference in the decomposition above. This is the matter of the following Lemmas. While the proofs of the bounds for the first two terms are quite elementary, the one for the final term uses results from gradient flow theory which are summarized in Section A. It is worthwhile to observe that we do not apply this theory to the Langevin semi-group but only to the heat semi-group.

Lemma 3 *Assume* ***A*** *2. For all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma > 0$,*

$$\mathscr{E}(\mu T_\gamma) - \mathscr{E}(\mu) \leq L d \gamma \ .$$

Proof First note that by (Nesterov, 2004, Lemma 1.2.3), for all $x, \tilde{x} \in \mathbb{R}^d$, we have

$$\left|U(\tilde{x}) - U(x) - \langle \nabla U(x), \tilde{x} - x\rangle\right| \leq (L/2) \left\|\tilde{x} - x\right\|^2 \ . \tag{16}$$

Therefore, for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma > 0$, we get

$$\begin{aligned}\mathscr{E}(\mu T_\gamma) - \mathscr{E}(\mu) &= (4\pi\gamma)^{-d/2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \{U(x+y) - U(x)\} \, \mathrm{e}^{-\|y\|^2/(4\gamma)} \mathrm{d}y \mathrm{d}\mu(x) \\ &\leq (4\pi\gamma)^{-d/2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left\{ \langle \nabla U(x), y \rangle + (L/2) \|y\|^2 \right\} \mathrm{e}^{-\|y\|^2/(4\gamma)} \mathrm{d}y \mathrm{d}\mu(x) \, ,\end{aligned}$$

which concludes the proof. ■

Lemma 4 *Assume* $\boldsymbol{A}\,1(m)$ *for* $m \geq 0$ *and* $\boldsymbol{A}\,2$*. For all* $\gamma \in (0, L^{-1}]$ *and* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$*,*

$$2\gamma \left\{ \mathscr{E}(\mu S_\gamma) - \mathscr{E}(\nu) \right\} \leq (1 - m\gamma) W_2^2(\mu, \nu) - W_2^2(\mu S_\gamma, \nu) - \gamma^2 (1 - \gamma L) \int_{\mathbb{R}^d} \|\nabla U(x)\|^2 \, \mathrm{d}\mu(x) \, ,$$

where $\mathscr{E}$ *and* T_γ *are defined in* (9) *and* (14) *respectively.*

Proof Using (16) and $\mathbf{A}1(m)$, for all $x, y \in \mathbb{R}^d$, we get

$$\begin{aligned}U(x - \gamma \nabla U(x)) - U(y) &= U(x - \gamma \nabla U(x)) - U(x) + U(x) - U(y) \\ &\leq -\gamma(1 - \gamma L/2) \|\nabla U(x)\|^2 + \langle \nabla U(x), x - y \rangle - (m/2) \|y - x\|^2 \, .\end{aligned}$$

Multiplying both sides by 2γ we obtain:

$$\begin{aligned}2\gamma \{U(x - \gamma \nabla U(x)) - U(y)\} \leq (1 - m\gamma) \|x - y\|^2 - \|x - \gamma \nabla U(x) - y\|^2 \\ - \gamma^2 (1 - \gamma L) \|\nabla U(x)\|^2 \, . \quad (17)\end{aligned}$$

Let now (X, Y) be an optimal coupling between μ and ν. Then by definition and (17), we get

$$\begin{aligned}2\gamma \left\{ \mathscr{E}(\mu S_\gamma) - \mathscr{E}(\nu) \right\} \leq (1 - m\gamma) W_2^2(\mu, \nu) - \mathbb{E}\left[\|X - \gamma \nabla U(X) - Y\|^2 \right] \\ - \gamma^2 (1 - \gamma L) \mathbb{E}\left[\|\nabla U(X)\|^2 \right] \, .\end{aligned}$$

Using that $W_2^2(\mu S_\gamma, \nu) \leq \mathbb{E}[\|X - \gamma \nabla U(X) - Y\|^2]$ concludes the proof. ■

Lemma 5 *Let* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$*,* $\mathscr{H}(\nu) < \infty$*. Then for all* $\gamma > 0$*,*

$$2\gamma \left\{ \mathscr{H}(\mu T_\gamma) - \mathscr{H}(\nu) \right\} \leq W_2^2(\mu, \nu) - W_2^2(\mu T_\gamma, \nu) \, ,$$

where T_γ *is given in* (14)*.*

Proof Denote for all $t \geq 0$ by $\mu_t = \mu T_t$. Since $(T_t)_{t \geq 0}$ is the Markov semi-group associated with the Brownian motion, then $(\mu_t)_{t \geq 0}$ is the solution (in the sense of distribution) of the Fokker-Plank equation:

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t \, ,$$

and μ_t goes to μ as t goes to 0 in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma > 0$. Then by Theorem 31, for all $\epsilon \in (0, \gamma)$, there exists $(\delta_t) \in \mathrm{L}^1((\epsilon, \gamma))$ such that

$$W_2^2(\mu_\gamma, \nu) - W_2^2(\mu_\epsilon, \nu) = \int_\epsilon^\gamma \delta_s \mathrm{d}s \tag{18}$$

$$\delta_s/2 \leq \mathscr{H}(\nu) - \mathscr{H}(\mu_s) \ , \text{ for almost all } s \in (\epsilon, \gamma) \ . \tag{19}$$

In addition by (Villani, 2009, Particular case 24.3), $s \mapsto \mathscr{H}(\mu_s)$ is non-increasing on $\mathbb{R}_+^*$ and therefore (19) becomes

$$\delta_s/2 \leq \mathscr{H}(\nu) - \mathscr{H}(\mu_\gamma) \ , \text{ for almost all } s \in (\epsilon, \gamma) \ .$$

Plugging this bound in (18) yields that for all $\epsilon \in \mathbb{R}_+^*$,

$$W_2^2(\mu_t, \nu) - W_2^2(\mu_\epsilon, \nu) \leq 2(\gamma - \epsilon) \left\{ \mathscr{H}(\nu) - \mathscr{H}(\mu_\gamma) \right\} \ .$$

Taking $\epsilon \to 0$ concludes the proof. ■

We now have all the tools to prove Proposition 2.

Proof [Proof of Proposition 2] Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \mathbb{R}_+^*$. By Lemma 3, we get

$$\mathscr{E}(\mu R_\gamma) - \mathscr{E}(\mu S_\gamma) = \mathscr{E}(\mu S_\gamma T_\gamma) - \mathscr{E}(\mu S_\gamma) \leq L d \gamma \ .$$

By Lemma 4 since $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ (see Lemma 1-a)),

$$2\gamma \left\{ \mathscr{E}(\mu S_\gamma) - \mathscr{E}(\pi) \right\} \leq (1 - m\gamma) W_2^2(\mu, \nu) - W_2^2(\mu S_\gamma, \nu) \ .$$

By Lemma 5 and Lemma 1-a),

$$\begin{aligned} 2\gamma \left\{ \mathscr{H}(\mu R_\gamma) - \mathscr{H}(\pi) \right\} &= 2\gamma \left\{ \mathscr{H}((\mu S_\gamma) T_\gamma) - \mathscr{H}(\pi) \right\} \\ &\leq W_2^2(\mu S_\gamma, \pi) - W_2^2(\mu R_\gamma, \pi) \ . \end{aligned}$$

Plugging these bounds in (15) concludes the proof. ■

Based on inequalities of the form (11) and using the convexity of f, for all $n \in \mathbb{N}$, non-asymptotic bounds (in expectation) between $f(\bar{x}_n)$ and $f(x_f)$ can be derived, where $(\bar{x}_k)_{k \in \mathbb{N}}$ is the sequence of averages of $(x_k)_{k \in \mathbb{N}}$ given for all $n \in \mathbb{N}$ by $\bar{x}_n = n^{-1} \sum_{k=1}^n x_k$. Besides, if f is assumed to be strongly convex, a bound on $\mathbb{E}[\|x_n - x_f\|^2]$ can be established. We will adapt this methodology to get some bounds on the convergence of sequences of averaged measures defined as follows. Let $(\gamma_k)_{k \in \mathbb{N}^*}$ and $(\lambda_k)_{k \in \mathbb{N}^*}$ be two non-increasing sequences of reals numbers referred to as the sequence of stepsizes and weights respectively. Define for all $n, N \in \mathbb{N}$, $n \geq 1$,

$$\Gamma_{N,N+n} = \sum_{k=N+1}^{N+n} \gamma_k \ , \qquad \Lambda_{N,N+n} = \sum_{k=N+1}^{N+n} \lambda_k \ . \tag{20}$$

Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ be an initial distribution. The sequence of probability measures $(\nu_n^N)_{n\in\mathbb{N}^*}$ is defined for all $n, N \in \mathbb{N}$, $n \geq 1$, by

$$\nu_n^N = \Lambda_{N,N+n}^{-1} \sum_{k=N+1}^{N+n} \lambda_k \, \mu_0 Q_\gamma^k \, , \qquad Q_\gamma^k = R_{\gamma_1} \cdots R_{\gamma_k} \, , \text{ for } k \in \mathbb{N}^* \, , \tag{21}$$

where R_γ is defined by (12) and N is a burn-in time. We take in the following, the convention that Q_γ^0 is the identity operator.

Contrary to most works on ULA, we state our next results in terms of average measures ν_n^N, defined by (21) for $N \in \mathbb{N}$ and $n \in \mathbb{N}^*$ instead of the final iterates $\mu_0 Q_\gamma^n$. Indeed, in the case where $m = 0$, Proposition 2 does not imply very informative bounds for Q_γ^n. However using that $\mathrm{KL}(\cdot|\pi)$ is convex and applying Proposition 2 allow to use averaging trick from optimization to obtain useful bounds on $\mathrm{KL}(\nu_n^N|\pi)$.

Theorem 6 *Assume* ***A*** *1(m) for $m \geq 0$ and* ***A*** *2. Let $(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ be two non-increasing sequences of positive real numbers satisfying $\gamma_1 \leq L^{-1}$, and for all $k \in \mathbb{N}^*$, $\lambda_{k+1}(1 - m\gamma_{k+1})/\gamma_{k+1} \leq \lambda_k/\gamma_k$. Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $N \in \mathbb{N}$. Then for all $n \in \mathbb{N}^*$, it holds:*

$$\begin{aligned}
&\mathrm{KL}\left(\nu_n^N|\pi\right) + \lambda_{N+n} W_2^2\left(\mu_0 Q_\gamma^{N+n}, \pi\right)/(2\gamma_{N+n}\Lambda_{N,N+n}) \\
&\quad \leq \lambda_{N+1}(1 - m\gamma_{N+1}) W_2^2\left(\mu_0 Q_\gamma^N, \pi\right)/(2\gamma_{N+1}\Lambda_{N,N+n}) + (Ld/\Lambda_{N,N+n}) \sum_{k=N+1}^{N+n} \gamma_k \lambda_k \, ,
\end{aligned}$$

where ν_n^N and Q_γ^N are defined in (21)*.*

Proof Using the convexity of KL divergence (see (Cover and Thomas, 2006, Theorem 2.7.2) or (van Erven and Harremos, 2014, Theorem 11)) and Proposition 2, we obtain

$$\begin{aligned}
\mathrm{KL}\left(\nu_n^N|\pi\right) &\leq \Lambda_{N,N+n}^{-1} \sum_{k=N+1}^{N+n} \lambda_k \, \mathrm{KL}\left(\mu_0 Q_\gamma^k \middle| \pi\right) \\
&\leq (2\Lambda_{N,N+n})^{-1} \left[\frac{(1 - m\gamma_{N+1})\lambda_{N+1}}{\gamma_{N+1}} W_2^2\left(\mu_0 Q_\gamma^N, \pi\right) - \frac{\lambda_{N+n}}{\gamma_{N+n}} W_2^2\left(\mu_0 Q_\gamma^{N+n}, \pi\right) \right. \\
&\qquad \left. + \sum_{k=N+1}^{N+n-1} \left\{ \frac{(1 - m\gamma_{k+1})\lambda_{k+1}}{\gamma_{k+1}} - \frac{\lambda_k}{\gamma_k} \right\} W_2^2\left(\mu_0 Q_\gamma^k, \pi\right) + \sum_{k=N+1}^{N+n} Ld\lambda_k\gamma_k \right] .
\end{aligned}$$

We get the thesis using that $\lambda_{k+1}(1 - m\gamma_{k+1})/\gamma_{k+1} \leq \lambda_k/\gamma_k$ for all $k \in \mathbb{N}^*$. ■

Corollary 7 *Assume* ***A*** *1(0) and* ***A*** *2. Let $\varepsilon > 0$ and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Let*

$$\gamma_\varepsilon \leq \min\left\{\varepsilon/(2Ld), L^{-1}\right\} \, , \qquad n_\varepsilon \geq \lceil W_2^2(\mu_0, \pi)\gamma_\varepsilon^{-1}\varepsilon^{-1} \rceil \, .$$

Then it holds $\mathrm{KL}\left(\nu_{n_\varepsilon}|\pi\right) \leq \varepsilon$ where $\nu_{n_\varepsilon} = n_\varepsilon^{-1} \sum_{k=1}^{n_\varepsilon} \mu_0 R_{\gamma_\varepsilon}^k$.

Proof We apply Theorem 6 with $\gamma_k = \gamma_\varepsilon$ and $\lambda_k = 1$ for all $k \geq 1$. We obtain

$$\mathrm{KL}\left(\nu_{n_\varepsilon}|\pi\right) + W_2^2\left(\mu_0 Q_\gamma^{n_\varepsilon}, \pi\right)/(2\gamma_\varepsilon n_\varepsilon) \leq W_2^2\left(\mu_0, \pi\right)/(2\gamma_\varepsilon n_\varepsilon) + (Ld/n_\varepsilon)\sum_{k=1}^{n_\varepsilon} \gamma_\varepsilon \;,$$

and the proof is concluded by a straightforward calculation using the definition of γ_ε and n_ε. ■

Corollary 8 *Assume* ***A*** *1(m) for $m \geq 0$ and* ***A*** *2. Let $\alpha \in (0,1)$. Define $(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ for all $k \in \mathbb{N}^*$ by $\gamma_k = \gamma_1/k^\alpha$, $\lambda_k = \gamma_1/(k+1)^\alpha$, $\gamma_1 \in \left(0, L^{-1}\right)$. Then, there exists $C \geq 0$ such that for all $n \in \mathbb{N}^*$ we have $\mathrm{KL}\left(\nu_n^0|\pi\right) \leq C\max(n^{\alpha-1}, n^{-\alpha})$, if $\alpha \neq 1/2$, and for $\alpha = 1/2$, we have $\mathrm{KL}\left(\nu_n^0|\pi\right) \leq C(\ln(n)+1)n^{-1/2}$, where ν_n^0 is defined by* (21).

Proof The proof is postponed to Section 7.2. ■

In the case where a warm start is available for the Wasserstein distance, *i.e.* $W_2^2(\mu_0, \pi) \leq C$, for some absolute constant $C \geq 0$, then Corollary 7 implies that the complexity of ULA to obtain a sample close from π in KL with a precision $\varepsilon > 0$ is of order $d\bar{\mathcal{O}}(\varepsilon^{-2})$. In addition, by Pinsker inequality, we have for all probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $\|\mu - \pi\|_{\mathrm{TV}} \leq \{2\,\mathrm{KL}(\mu|\pi)\}^{1/2}$, which implies that the complexity of ULA for the total variation distance is of order $d\bar{\mathcal{O}}(\varepsilon^{-4})$.

In addition if we have access to $\eta > 0$ and $M_\eta \geq 0$, independent of the dimension, such that for all $x \in \mathbb{R}^d$, $x \notin \mathrm{B}(x^\star, M_\eta)$, $U(x) - U(x^\star) \geq \eta\,\|x - x^\star\|$, $x^\star \in \arg\min_{\mathbb{R}^d} U$, Proposition 32 in Appendix B shows that for all $\int_{\mathbb{R}^d} \|x - x^\star\|^2 \,\mathrm{d}\pi(x) \leq 2\eta^{-2}d(1+d) + M_\eta^2$. Therefore, starting at $\delta_{x^\star}$, the overall complexity for the KL is in this case $d^3\bar{\mathcal{O}}(\varepsilon^{-2})$ and $d^3\bar{\mathcal{O}}(\varepsilon^{-4})$ for the total variation distance. This discussion justifies the bound we state in Table 1.

In the case where $m > 0$, based on Proposition 2, we can directly get a bound on the Wasserstein distance between the final iterate of ULA and π.

Theorem 9 *Assume* ***A*** *1(m) for $m > 0$ and* ***A*** *2. Let $(\gamma_k)_{k\in\mathbb{N}^*}$ be a non-increasing sequence of positive real numbers, $\gamma_1 \in \left(0, L^{-1}\right]$, and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then for all $n \in \mathbb{N}^*$, it holds*

$$W_2^2\left(\mu_0 Q_\gamma^n, \pi\right) \leq \left\{\prod_{k=1}^n (1 - m\gamma_k)\right\} W_2^2\left(\mu_0, \pi\right) + 2Ld\sum_{k=1}^n \gamma_k^2 \prod_{i=k+1}^n (1 - m\gamma_i) \;,$$

where Q_γ^n is defined in (21).

Proof Using Proposition 2 and since the KL divergence is non-negative, we get for all $k \in \{1, \ldots, n\}$,

$$W_2^2\left(\mu_0 Q_\gamma^k, \pi\right) \leq (1 - m\gamma_k) W_2^2\left(\mu_0 Q_\gamma^{k-1}, \pi\right) + 2Ld\gamma_k^2 \;.$$

The proof then follows from a direct induction. ■

Corollary 10 *Assume* ***A*** *1(m) for $m > 0$ and* ***A*** *2. Let $\varepsilon > 0$ and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Define:*

$$\gamma_\varepsilon \leq \min\left\{m\varepsilon/(4Ld), L^{-1}\right\} , \qquad n_\varepsilon \geq \lceil \ln(2W_2^2(\mu_0,\pi)/\varepsilon)\gamma_\varepsilon^{-1}m^{-1}\rceil .$$

Then we have $W_2^2\left(\mu_0 R_{\gamma_\varepsilon}^{n_\varepsilon}, \pi\right) \leq \varepsilon$, where R_{γ_ε} is defined by (12).

Proof By Theorem 9, we have

$$W_2^2\left(\mu_0 Q_\gamma^{n_\varepsilon}, \pi\right) \leq (1-m\gamma_\varepsilon)^{n_\varepsilon} W_2^2\left(\mu_0,\pi\right) + 2Ld\sum_{k=1}^{n_\varepsilon}\gamma_\varepsilon^2(1-m\gamma_\varepsilon)^{n_\varepsilon-k} .$$

On one hand, by definition of γ_ε, we get $2Ld\sum_{k=1}^{n_\varepsilon}\gamma_\varepsilon^2(1-m\gamma_\varepsilon)^{n_\varepsilon-k} \leq 2Ld\gamma_\varepsilon/m \leq \varepsilon/2$. On the other hand, using that for all $t \in \mathbb{R}_+$, $1-t \leq \exp(-t)$ and the definition of n_ε, we obtain $(1-m\gamma_\varepsilon)^{n_\varepsilon} W_2^2\left(\mu_0,\pi\right) \leq \exp(-m\gamma_\varepsilon n_\varepsilon) W_2^2\left(\mu_0,\pi\right) \leq \varepsilon/2$. Then the thesis of the corollary follows directly from the above inequalities. ■

Note that the bound in the right hand side of Theorem 9 is tighter than the previous bound given in Dalalyan and Karagulyan (2017, Theorem 1) (for constant stepsize) and in Durmus and Moulines (2016, Theorem 5) (for both constant and non-increasing stepsizes). Indeed Dalalyan and Karagulyan (2017, Theorem 1) shows that, in the constant stepsize setting $\gamma_k = \gamma$, for all $k \in \mathbb{N}$,

$$W_2(\mu_0 Q_\gamma^k, \pi) \leq (1-m\gamma)^k W_2(\mu_0,\pi) + 1.65(L/m)(\gamma d)^{1/2} .$$

On the other hand, the inequality $(t+s)^{1/2} \leq t^{1/2} + s^{1/2}$ for $t, s \geq 0$ and Theorem 9 imply that for all $k \in \mathbb{N}$,

$$W_2(\mu_0 Q_\gamma^k, \pi) \leq (1-m\gamma)^{k/2} W_2(\mu_0,\pi) + \{2\gamma dL/m\}^{1/2} . \tag{22}$$

Thus, the dependency on the condition number L/m is improved. This bound is in agreement for the case where π is the zero-mean d-dimensional Gaussian distribution with covariance matrix Σ. In that case, all the iterates $(X_k)_{k\in\mathbb{N}^*}$ defined by (2) for $\gamma > 0$, starting from $x \in \mathbb{R}^d$, follows the Gaussian distribution with mean $(\mathrm{Id} - \gamma\Sigma)^k x$ and covariance matrix $2\gamma\sum_{i=0}^{k-1}(1-\gamma\Sigma)^{2i}$. Since the Wasserstein distance between d-dimensional Gaussian distributions can be explicitly computed, see Givens and Shortt (1984), denoting by L and m the largest and smallest eigenvalues of Σ respectively, we have by an explicit calculation for $\gamma \in \left(0, L^{-1}\right]$,

$$W_2(\mu_0 Q_\gamma^k, \pi) \leq (1-m\gamma)^k W_2(\mu_0,\pi) + (d/m)^{1/2}\left\{(1-\gamma L/2)^{-1/2} - 1\right\} .$$

Since for $t \in [0, 1/2]$, $(1-t)^{-1/2} - 1 - t \leq 0$, we get

$$W_2(\mu_0 Q_\gamma^k, \pi) \leq (1-m\gamma)^k W_2(\mu_0,\pi) + 2^{-1}\gamma(d/m)^{1/2}\left\{(1-\gamma L)^{-1/2} - 1\right\} .$$

Using that $\gamma \leq L^{-1}$, we get that the second term in the right hand side is bounded by $(dL\gamma/m)^{1/2}$, which is precisely the order we get from (22).

Finally, if $(\gamma_k)_{k\in\mathbb{N}^*}$ is given for all $k \in \mathbb{N}^*$, by $\gamma_k = \gamma_1/k^\alpha$, for $\alpha \in (0,1)$, then using (Durmus and Moulines, 2016, Lemma 7) and the same calculation of (Durmus and Moulines, 2015, Section 6.1), we get that there exists $C \geq 0$ such that for all $n \in \mathbb{N}^*$, $W_2(\mu_0 Q_\gamma^n, \pi) \leq Cn^{-\alpha/2}$.

Based on Theorem 9, we can improve Corollary 7 in the case where U is strongly convex using an appropriate burn-in time.

Corollary 11 *Assume* ***A*** *1(m) for $m > 0$ and* ***A*** *2. Let $\varepsilon > 0$, $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and*

$$\gamma_\varepsilon \leq \min\left\{m\varepsilon/(4Ld), L^{-1}\right\}\ , \qquad \tilde{\gamma}_\varepsilon \leq \min\left\{\varepsilon/2Ld, L^{-1}\right\}\ ,$$
$$N_\varepsilon \geq \lceil \ln(2W_2^2(\mu_0,\pi)/\varepsilon)(\gamma_\varepsilon m)^{-1}\rceil\ , \qquad n_\varepsilon \geq \lceil \tilde{\gamma}_\varepsilon^{-1}\rceil\ .$$

Let $(\gamma_k)_{k\in\mathbb{N}}$ defined by $\gamma_k = \gamma_\varepsilon$ for $k \in \{1,\dots,N_\varepsilon\}$ and $\gamma_k = \tilde{\gamma}_\varepsilon$ for $k > N_\varepsilon$. Then we have $\mathrm{KL}\left(\nu_{n_\varepsilon}^{N_\varepsilon}|\pi\right) \leq \varepsilon$ where $\nu_{n_\varepsilon}^{N_\varepsilon} = n_\varepsilon^{-1}\sum_{k=1}^{n_\varepsilon} \mu_0 R_{\gamma_\varepsilon}^{N_\varepsilon} R_{\tilde{\gamma}_\varepsilon}^k$.

Proof Using Corollary 10, we have $W_2^2\left(\mu_0 Q_\gamma^{N_\varepsilon}, \pi\right) \leq \varepsilon$. Now applying Theorem 6 we get:

$$\mathrm{KL}\left(\nu_{n_\varepsilon}^{N_\varepsilon}|\pi\right) \leq W_2^2\left(\mu_{N_\varepsilon}, \pi\right)/(2\tilde{\gamma}_\varepsilon n_\varepsilon) + (Ld/n_\varepsilon\tilde{\gamma}_\varepsilon)\sum_{k=N_\varepsilon+1}^{N_\varepsilon+n_\varepsilon}(\tilde{\gamma}_\varepsilon)^2 \leq \varepsilon/(2\tilde{\gamma}_\varepsilon n_\varepsilon) + Ld\tilde{\gamma}_\varepsilon \leq \varepsilon$$

■

By (Durmus and Moulines, 2016, Proposition 1), we have $\int_{\mathbb{R}^d}\|x - x^\star\|^2 \mathrm{d}\pi(x) \leq d/m$, where $x^\star = \arg\min_{\mathbb{R}^d} U$. Therefore we have that in the constant stepsize setting, $\gamma_k = \gamma \in (0, L^{-1}]$ for all $k \in \mathbb{N}^*$, Corollary 10 implies that the sufficient number of iterations to have $W_2(\delta_{x^\star} Q_\gamma^n, \pi) \leq \varepsilon$ is of order $d\bar{\mathcal{O}}(\varepsilon^{-2})$. Then Corollary 11 implies that the sufficient number of iterations to get $\mathrm{KL}\left(\nu_n^N|\pi\right) \leq \varepsilon$, for $\varepsilon > 0$, is of order $d\bar{\mathcal{O}}(\varepsilon^{-1})$. By Pinsker inequality, we obtain that a sufficient number of iterations to get $\|\nu_n^N - \pi\|_{\mathrm{TV}} \leq \varepsilon$, for $\varepsilon > 0$, is of order $d\bar{\mathcal{O}}(\varepsilon^{-2})$.

For a sufficiently small constant stepsize γ, ULA produces a Markov Chain with a stationary measure π_γ. In general this measure is different from the measure of interest π. Based on our previous results, we establish computable bounds on the distance between π and π_γ.

Theorem 12 *Assume* ***A*** *1(m) for $m \geq 0$ and* ***A*** *2. Let $\gamma \in \left(0, L^{-1}\right]$. Then there exists a measure π_γ, such that $\pi_\gamma R_\gamma = \pi_\gamma$ where R_γ is defined by (12). In addition, we have*

$$\mathrm{KL}\left(\pi_\gamma|\pi\right) \leq Ld\gamma, \qquad \|\pi_\gamma - \pi\|_{\mathrm{TV}} \leq \sqrt{2Ld\gamma}$$

Furthermore, if $m > 0$ we also have $W_2^2\left(\pi_\gamma, \pi\right) \leq 2Ld\gamma/m$.

Proof Under **A** 1 and **A** 2, (Durmus and Moulines, 2017, Proposition 13) shows that R_γ satisfies a geometric Foster-Lyapunov drift condition for $\gamma \leq L^{-1}$. In addition, it is easy to see that R_γ is Leb-irreducible and weak Feller and therefore by (Meyn and Tweedie, 2009, Theorem 6.0.1 together with Theorem 5.5.7), all compact sets are small. Then, by (Meyn and Tweedie, 2009, Theorem 16.0.1), R_γ has a unique invariant distribution π_γ.

Second, taking $\mu = \pi_\gamma$ in Proposition 2 we obtain:

$$2\gamma \,\mathrm{KL}\,(\pi_\gamma R_\gamma | \pi) \leq (1 - m\gamma) W_2^2(\pi_\gamma, \pi) - W_2^2(\pi_\gamma R_\gamma, \pi) + 2\gamma^2 L d \,, \tag{23}$$

and because $\pi_\gamma R_\gamma = \pi_\gamma$, the above implies $2\,\mathrm{KL}\,(\pi_\gamma|\pi) + m W_2^2(\pi_\gamma, \pi) \leq 2Ld\gamma$. Since both the KL divergence and Wasserstein distance are positive, the desired bounds in KL and W_2^2 follow. The bound in total variation follows from the bound in KL-divergence and Pinsker inequality. ■

4. Extensions of ULA

In this section, two extensions of ULA are presented and analyzed. These two algorithms can be applied to non-continuously differentiable convex potential $U : \mathbb{R}^d \to \mathbb{R}$ and therefore **A**2 is not assumed anymore. In addition, for the two new algorithms we present, only i.i.d. unbiased estimates of (sub)gradients of U are necessary as in Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011). The main difference in these two approaches is that one relies on the subgradient of U while the other is based on proximal operators which are tools commonly used in non-smooth optimization. However, the theoretical results that we can show for these two algorithms, hold for different sets of conditions.

4.1. Stochastic SubGradient Langevin Dynamics

Note that if U is convex and l.s.c then for any point $x \in \mathbb{R}^d$, its subdifferential $\partial U(x)$ defined by

$$\partial U(x) = \left\{ v \in \mathbb{R}^d \,:\, U(y) \geq U(x) + \langle v, y - x \rangle \text{ for all } y \in \mathbb{R}^d \right\} \,, \tag{24}$$

is non empty, see (Rockafellar and Wets, 1998, Proposition 8.12, Theorem 8.13). For all $x \in \mathbb{R}^d$, any elements of $\partial U(x)$ is referred to as a subgradient of U at x. Consider the following condition on U which assumes that we have access to unbiased estimates of subgradients of U at any point $x \in \mathbb{R}^d$.

A 3 (i) The potential U is M-Lipschitz, *i.e.* for all $x, y \in \mathbb{R}^d$, $|U(x) - U(y)| \leq M \|x - y\|$.

(ii) There exists a measurable space $(\mathsf{Z}, \mathcal{Z})$, a probability measure η on $(\mathsf{Z}, \mathcal{Z})$ and a measurable function $\Theta : \mathbb{R}^d \times \mathsf{Z} \to \mathbb{R}^d$ for all $x \in \mathbb{R}^d$,

$$\int_{\mathsf{Z}} \Theta(x, z) \mathrm{d}\eta(z) \in \partial U(x) \,.$$

Note that under **A**3-(i), for all $x \in \mathbb{R}^d$ and $v \in \partial U(x)$,

$$\|v\| \leq M \,. \tag{25}$$

Assumption **A**3 is satisfied for example in the case where $U = U_1 + U_2$, U_1 is L-gradient Lipschitz and Lipschitz and U_2 is non-smooth but Lipschitz, if there exists a measurable $\tilde{\Theta} : \mathbb{R}^d \times \mathsf{Z} \to \mathbb{R}^d$ such that $\int_{\mathsf{Z}} \tilde{\Theta}(x, z) \mathrm{d}\eta(z) = \nabla U_1(x)$ for any $x \in \mathbb{R}^d$. Then by (Bauschke

and Combettes, 2011, Corollary 16.38), unbiased and i.i.d. estimates of ∇f can be computed setting $\Theta = \partial g + \tilde{\Theta}$.

Let $(Z_k)_{k\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables distributed according to η, $(\gamma_k)_{k\in\mathbb{N}^*}$ be a sequence of non-increasing stepsizes and $\bar{X}_0$ distributed according to $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Stochastic SubGradient Langevin Dynamics (SSGLD) defines the sequence of random variables $(\bar{X}_k)_{k\in\mathbb{N}}$ starting at $\bar{X}_0$ for $n \geq 0$ by

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}\Theta(\bar{X}_n, Z_{n+1}) + \sqrt{2\gamma_{n+2}}G_{n+1} \;, \tag{26}$$

where $(G_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables, independent of $(Z_k)_{k\in\mathbb{N}^*}$, see Algorithm 1. Consequently this method defines a new sequence of Markov kernels $(\bar{R}_{\gamma_k,\gamma_{k+1}})_{k\in\mathbb{N}^*}$ given for all $\gamma, \tilde{\gamma} > 0$, $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\bar{R}_{\gamma,\tilde{\gamma}}(x, \mathsf{A}) = (4\pi\tilde{\gamma})^{-d/2} \int_{\mathsf{A}\times\mathsf{Z}} \exp\left(-\left\|y - x + \gamma\Theta(x,z)\right\|^2 / (4\tilde{\gamma})\right) \mathrm{d}\eta(z)\mathrm{d}y \;. \tag{27}$$

Algorithm 1: SSGLD

Data: initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, non-increasing sequence $(\gamma_k)_{k\geq 1}$, U, Θ, η satisfying **A**3
Result: $(\bar{X}_k)_{k\in\mathbb{N}}$
begin
  Draw $\bar{X}_0 \sim \mu_0$;
  for $k \geq 0$ **do**
    Draw $G_{k+1} \sim \mathcal{N}(0, \mathrm{Id})$ and $Z_{k+1} \sim \eta$;
    Set $\bar{X}_{k+1} = \bar{X}_k - \gamma_{k+1}\Theta(\bar{X}_k, Z_{k+1}) + \sqrt{2\gamma_{k+2}}G_{k+1}$

Let $(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ be two non-increasing sequences of reals numbers and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ be an initial distribution. The weighted averaged distribution associated with (26) $(\bar{\nu}_n^N)_{n\in\mathbb{N}}$ is defined for all $N, n \in \mathbb{N}$, $n \geq 1$ by

$$\bar{\nu}_n^N = \Lambda_{N,N+n}^{-1} \sum_{k=N+1}^{N+n} \lambda_k \, \mu_0 \bar{Q}_\gamma^k \;, \qquad \bar{Q}_\gamma^k = \bar{R}_{\gamma_1,\gamma_2} \cdots \bar{R}_{\gamma_k,\gamma_{k+1}} \;, \text{ for } k \in \mathbb{N}^* \;, \tag{28}$$

where N is a burn-in time and $\Lambda_{N,N+n}$ is defined in (20). We take in the following the convention that $\bar{Q}_\gamma^0$ is the identity operator.

Under **A**3, define for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$v_\Theta(\mu) = \int_{\mathbb{R}^d\times\mathsf{Z}} \left\| \Theta(x,z) - \int_{\mathsf{Z}} \Theta(x,\tilde{z})\mathrm{d}\eta(\tilde{z}) \right\|^2 \mathrm{d}\eta(z)\mathrm{d}\mu(x) = \mathbb{E}\left[\left\| \Theta(\bar{X}_0, Z_1) - v \right\|^2 \right] \;, \tag{29}$$

where $\bar{X}_0, Z_1$ are independent random variables with distribution μ and η respectively and $v \in \partial U(X_0)$ almost surely. In addition, consider $\bar{S}_\gamma$, the Markov kernel on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ defined for all $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\bar{S}_\gamma(x, \mathsf{A}) = \int_{\mathsf{Z}} \mathbb{1}_{\mathsf{A}}\left(x - \gamma\Theta(x,z)\right) \mathrm{d}\eta(z) \;. \tag{30}$$

Theorem 13 *Assume* ***A*** *1(0) and* ***A*** *3. Let* $(\gamma_k)_{k\in\mathbb{N}^*}$ *and* $(\lambda_k)_{k\in\mathbb{N}^*}$ *be two non-increasing sequences of positive real numbers satisfying for all* $k \in \mathbb{N}^*$, $\lambda_{k+1}/\gamma_{k+2} \leq \lambda_k/\gamma_{k+1}$. *Let* $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ *and* $N \in \mathbb{N}$. *Then for all* $n \in \mathbb{N}^*$, *it holds*

$$\mathrm{KL}\left(\bar{\nu}_n^N \middle| \pi\right) \leq \lambda_{N+1} W_2^2\left(\mu_0 \bar{Q}_\gamma^N \bar{S}_{\gamma_{N+1}}, \pi\right)/(2\gamma_{N+2}\Lambda_{N,N+n}) + (2\Lambda_{N,N+n})^{-1} \sum_{k=N+1}^{N+n} \left\{\gamma_{k+1}\lambda_k \left(M^2 + \upsilon_\Theta(\mu_0 \bar{Q}_\gamma^k)\right)\right\} ,$$

where $\bar{\nu}_n^N$ *and* $\bar{Q}_\gamma^N$ *are defined in* (28).

Proof The proof is postponed to Section 7.3.1.. ■

Note that in the bound given by Theorem 13, we need to control the ergodic average of the variance of the stochastic gradient estimates. When **A** 3 is satisfied, a possible assumption is that $x \mapsto \upsilon(\delta_x)$ is uniformly bounded. This assumption will be satisfied for example when the potential U is a sum of Lipschitz continuous functions.

Corollary 14 *Assume* ***A*** *1(0) and* ***A*** *3. Assume that* $\sup_{x\in\mathbb{R}^d} \upsilon_\Theta(\delta_x) \leq D^2 < \infty$. *Let* $(\gamma_k)_{k\in\mathbb{N}^*}$ *and* $(\lambda_k)_{k\in\mathbb{N}^*}$ *given for all* $k \in \mathbb{N}^*$ *by* $\lambda_k = \gamma_k = \gamma > 0$. *Let* $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. *Then for any* $N \in \mathbb{N}, n \in \mathbb{N}^*$ *we have*

$$\mathrm{KL}\left(\bar{\nu}_n^N \middle| \pi\right) \leq W_2^2\left(\mu_0 \bar{Q}_\gamma^N \bar{S}_\gamma, \pi\right)/(2n\gamma) + (\gamma/2)\left(M^2 + D^2\right) .$$

Furthermore, let $\varepsilon > 0$ *and*

$$\gamma_\varepsilon \leq \varepsilon/(M^2 + D^2) , \qquad n_\varepsilon \geq \lceil W_2^2(\mu_0 \bar{S}_\gamma, \pi)(\gamma_\varepsilon \varepsilon)^{-1} \rceil .$$

Then for $\gamma = \gamma_\varepsilon$ *we have* $\mathrm{KL}\left(\bar{\nu}_{n_\varepsilon}^0 \middle| \pi\right) \leq \varepsilon$.

Proof The first inequality is a direct consequence of Theorem 13. The bound for $\mathrm{KL}\left(\bar{\nu}_{n_\varepsilon}^0 \middle| \pi\right)$ follows directly from this inequality and definitions of γ_ε and n_ε. ■

In the case where a warm start is available for the Wasserstein distance, *i.e.* $W_2^2(\mu_0, \pi) \leq C$, for some absolute constant $C \geq 0$, then Corollary 14 implies that the complexity of SSGLD to obtain a sample close from π in KL with a precision target $\varepsilon > 0$ is of order $(M^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-2})$. Therefore, this complexity bound depends on the dimension only trough M and D^2 contrary to ULA. In addition, Pinsker inequality implies that the complexity of SSGLD for the total variation distance is of order $(M^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-4})$.

In addition if we have access to $\eta > 0$ and $M_\eta \geq 0$, independent of the dimension, such that for all $x \in \mathbb{R}^d$, $x \notin \mathrm{B}(x^\star, M_\eta)$, $U(x) - U(x^\star) \geq \eta \|x - x^\star\|$, where $x^\star \in \arg\min_{\mathbb{R}^d} U$, Proposition 32 and **A**3-(i) imply that starting at $\delta_{x^\star}$, the overall complexity of SSGLD for the KL is in this case $(\eta^{-2}d^2 + M_\eta^2)(M^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-2})$ and $(\eta^{-2}d^2 + M_\eta^2)(M^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-4})$ for the total variation distance.

If $(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ are given for all $k \in \mathbb{N}^*$ by $\gamma_k = \lambda_k = \gamma_1/k^{-\alpha}$, with $\alpha \in (0, 1)$, then by the same reasoning as in the proof of Corollary 8, we obtain that there exists $C \geq 0$

such that for all $n \in \mathbb{N}^*$, we have $\mathrm{KL}\left(\bar{\nu}_n^0 | \pi\right) \leq C \max(n^{\alpha-1}, n^{-\alpha})$, if $\alpha \neq 1/2$, and for $\alpha = 1/2$, we have $\mathrm{KL}\left(\bar{\nu}_n^0 | \pi\right) \leq C(\ln(n)+1)n^{-1/2}$.

We can have a better control on the variance terms using the following conditions on Θ.

A 4 There exists $\tilde{L} \geq 0$ such that for η-almost every $z \in \mathsf{Z}$, $x \mapsto \Theta(x,z)$ is $1/\tilde{L}$-cocoercive, *i.e.* for all $x \in \mathbb{R}^d$,

$$\langle \Theta(x,z) - \Theta(y,z), x-y \rangle \geq (1/\tilde{L}) \left\| \Theta(x,z) - \Theta(y,z) \right\|^2 .$$

This assumption is for example satisfied if η-almost every z, $x \mapsto \Theta(x,z)$ is the gradient of a continuously differentiable convex function with Lipschitz gradient, see Nesterov (2004, Thereom 2.1.5) and Zhu and Marcotte (1995). Note that in general **A** 4 is not implied by **A** 1(0) and **A** 2. Indeed, **A** 4 is a regularity condition on the stochastic (sub)gradient of U, Θ, while **A** 1 and **A** 2 depend only on U. However, if U is continuously differentiable, Jensen inequality and **A** 4 imply that **A** 2 is satisfied with L equals $\tilde{L}$.

Proposition 15 *Assume **A** 3 and **A** 4. Then we have for all $x \in \mathbb{R}^d$ and $\gamma, \tilde{\gamma} > 0$, $\gamma \leq \tilde{L}^{-1}$*

$$2\gamma(\tilde{L}^{-1} - \gamma) \upsilon_\Theta(\delta_x) \leq \|x - x^\star\|^2 - \int_{\mathbb{R}^d} \|y - x^\star\|^2 \bar{R}_{\gamma,\tilde{\gamma}}(x, \mathrm{d}y) + 2\gamma^2 \upsilon_\Theta(\delta_{x^\star}) + 2\tilde{\gamma} d ,$$

where υ_Θ is defined by (29).

Proof Consider $\bar{X}_1 = x - \gamma\Theta(x, Z_1) + \sqrt{2\tilde{\gamma}} G_1$, where Z_1 and G_1 are two independent random variables, Z_1 has distribution η and G_1 is the standard Gaussian random variables. Then using **A** 4, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{X}_1 - x^\star\right\|^2\right] &= \mathbb{E}\left[\|x - \gamma\Theta(x, Z_1) - x^\star\|\right] + 2\tilde{\gamma} d \\
&= \|x - x^\star\|^2 + \mathbb{E}\left[\gamma^2 \|\Theta(x, Z_1)\|^2 - 2\gamma \langle \Theta(x, Z_1), x - x^\star \rangle\right] + 2\tilde{\gamma} d \\
&\leq \|x - x^\star\|^2 - 2\gamma(\tilde{L}^{-1} - \gamma)\mathbb{E}\left[\|\Theta(x, Z_1) - \Theta(x^\star, Z_1)\|^2\right] \\
&\qquad + 2\gamma^2 \mathbb{E}\left[\|\Theta(x^\star, Z_1)\|^2\right] + 2\tilde{\gamma} d .
\end{aligned}$$

The proof is completed upon noting that $\upsilon_\Theta(\delta_x) \leq \mathbb{E}[\|\Theta(x, Z_1) - \Theta(x^\star, Z_1)\|^2]$ and $\upsilon_\Theta(\delta_{x^\star}) = \mathbb{E}\left[\|\Theta(x^\star, Z_1)\|^2\right]$ ■

Combining Theorem 13 and Proposition 15, we get the following result.

Corollary 16 *Assume **A** 1(0)-**A** 3 and **A** 4. Let $(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ defined for all $k \in \mathbb{N}^*$ by $\gamma_k = \lambda_k = \gamma \in (0, \tilde{L}^{-1})$. Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then for all $N \in \mathbb{N}$ and $n \in \mathbb{N}^*$, we have*

$$\begin{aligned}
&\mathrm{KL}\left(\bar{\nu}_n^N | \pi\right) \leq W_2^2\left(\mu_0 \bar{R}_{\gamma,\gamma}^N \bar{S}_\gamma, \pi\right) \Big/ (2\gamma n) \\
&\quad + \gamma M^2/2 + (2(\tilde{L}^{-1} - \gamma))^{-1} \left\{ (2n)^{-1} \int_{\mathbb{R}^d} \|x - x^\star\|^2 \, \mathrm{d}\mu_0 \bar{R}_{\gamma,\gamma}^{N+1}(x) + \gamma^2 \upsilon_\Theta(\delta_{x^\star}) + \gamma d \right\} .
\end{aligned}$$

Furthermore, let $\varepsilon > 0$ *and*

$$\gamma_\varepsilon \leq \min\left[\varepsilon \Big/ \left\{2M^2 + 4\tilde{L}d\right\}, \sqrt{\varepsilon\left(4\tilde{L}\upsilon_\Theta(\delta_{x^\star})\right)^{-1}}, (2\tilde{L})^{-1}\right] ,$$

$$n_\varepsilon \geq 2\max\left\{\left\lceil W_2^2(\mu_0\bar{S}_{\gamma_\varepsilon}, \pi)(\gamma_\varepsilon\varepsilon)^{-1}\right\rceil, \left\lceil \tilde{L}\varepsilon^{-1}\int_{\mathbb{R}^d}\|x - x^\star\|^2 \,\mathrm{d}\mu_0\bar{R}_{\gamma_\varepsilon,\gamma_\varepsilon}(x)\right\rceil\right\} .$$

Then for $\gamma = \gamma_\varepsilon$, *then we have* $\mathrm{KL}\left(\bar{\nu}^0_{n_\varepsilon}|\pi\right) \leq \varepsilon$.

Proof The proof is postponed to Section 7.3.2. ■

Note that compared to Corollary 14, the dependence on the variance of the stochastic subgradients in the bound on n_ε, given in Corollary 16, is less significant since n_ε scales as $(\upsilon_\Theta(\delta_{x^\star}))^{1/2}$ and not as $\sup_{x\in\mathbb{R}^d}\upsilon_\Theta(\delta_x)$. However, the dependency on the dimension deteriorates a little.

4.2. Stochastic Proximal Gradient Langevin Dynamics

In this section, we propose and analyze an other algorithm to handle non-smooth target distribution using stochastic gradient estimates and proximal operators. For $m \geq 0$, consider the following assumptions on the gradient.

A 5 (m) There exists $U_1 : \mathbb{R}^d \to \mathbb{R}$ and $U_2 : \mathbb{R}^d \to \mathbb{R}$ such that $U = U_1 + U_2$ and satisfying the following assumptions:

(i) U_1 satisfies **A** 1(m) and **A** 2. In addition, there exists a measurable space $(\tilde{\mathsf{Z}}, \tilde{\mathcal{Z}})$, a probability measure $\tilde{\eta}_1$ on $(\tilde{\mathsf{Z}}, \tilde{\mathcal{Z}})$ and a measurable function $\tilde{\Theta}_1 : \mathbb{R}^d \times \tilde{\mathsf{Z}} \to \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$,

$$\int_{\tilde{\mathsf{Z}}} \tilde{\Theta}_1(x, \tilde{z})\mathrm{d}\tilde{\eta}_1(\tilde{z}) = \nabla U_1(x) .$$

(ii) U_2 satisfies **A**1(0) and is M_2-Lipschitz.

Under **A** 5, consider the proximal operator associated with U_2 with parameter $\gamma > 0$ (see Rockafellar and Wets (1998, Chapter 1 Section G)), defined for all $x \in \mathbb{R}^d$ by

$$\mathrm{prox}^\gamma_{U_2}(x) = \operatorname*{arg\,min}_{y\in\mathbb{R}^d}\left\{U_2(y) + (2\gamma)^{-1}\|x - y\|^2\right\} .$$

Note that taking the derivative in the right hand side of this equation, we get that for any $x \in \mathbb{R}^d$ and $\gamma > 0$,

$$\mathrm{prox}^\gamma_{U_2}(x) \in x - \gamma\partial U_2(\mathrm{prox}^\gamma_{U_2}(x)) . \tag{31}$$

Let $(\tilde{Z}_k)_{k\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables distributed according to $\tilde{\eta}_1$, $(\gamma_k)_{k\in\mathbb{N}^*}$ be a sequence of non-increasing stepsizes and $\tilde{X}_0$ distributed according to $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Stochastic Proximal Gradient Langevin Dynamics (SPGLD) defines the sequence of random variables $(\tilde{X}_n)_{n\in\mathbb{N}}$ starting at $\tilde{X}_0$ for $n \geq 0$ by

$$\tilde{X}_{n+1} = \mathrm{prox}^{\gamma_{n+1}}_{U_2}(\tilde{X}_n) - \gamma_{n+2}\tilde{\Theta}_1\{\mathrm{prox}^{\gamma_{n+1}}_{U_2}(\tilde{X}_n), \tilde{Z}_{n+1}\} + \sqrt{2\gamma_{n+2}}G_{n+1} , \tag{32}$$

where $(G_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables, independent of $(\tilde{Z}_k)_{k\in\mathbb{N}^*}$. The recursion (32) is associated with the family of Markov kernels $(\tilde{R}_{\gamma_k,\gamma_{k+1}})_{k\in\mathbb{N}^*}$ given for all $\gamma,\tilde{\gamma}>0$, $x\in\mathbb{R}^d$ and $\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)$ by

$$\begin{aligned}&\tilde{R}_{\gamma,\tilde{\gamma}}(x,\mathsf{A})\\&=(4\pi\tilde{\gamma})^{-d/2}\int_{\mathsf{A}\times\tilde{\mathsf{Z}}}\exp\left(-\left\|y-\mathrm{prox}_{U_2}^{\gamma}(x)+\tilde{\gamma}\tilde{\Theta}_1\{\mathrm{prox}_{U_2}^{\gamma}(x),z\}\right\|^2\Big/(4\tilde{\gamma})\right)\mathrm{d}\tilde{\eta}_1(z)\mathrm{d}y\ .\end{aligned}\tag{33}$$

Note that for all $\gamma,\tilde{\gamma}>0$, $\tilde{R}_{\gamma,\tilde{\gamma}}$ can be decomposed as the product $\tilde{S}^2_{\gamma}\tilde{S}^1_{\tilde{\gamma}}T_{\tilde{\gamma}}$ where $T_{\tilde{\gamma}}$ is defined by (14) and for all $x\in\mathbb{R}^d$ and $\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)$

$$\tilde{S}^1_{\tilde{\gamma}}(x,\mathsf{A})=\int_{\tilde{\mathsf{Z}}}\mathbb{1}_{\mathsf{A}}(x-\tilde{\gamma}\tilde{\Theta}_1(x,z))\mathrm{d}\tilde{\eta}_1(z)\ ,\qquad \tilde{S}^2_{\gamma}(x,\mathsf{A})=\delta_{\mathrm{prox}_{U_2}^{\gamma}(x)}(\mathsf{A})\ .\tag{34}$$

Algorithm 2: SPGLD

Data: initial distribution $\mu_0\in\mathcal{P}_2(\mathbb{R}^d)$, non-increasing sequence $(\gamma_k)_{k\geq 1}$, $U=U_1+U_2,\tilde{\Theta}_1,\tilde{\eta}_1$ satisfying **A**5

Result: $(\tilde{X}_k)_{k\in\mathbb{N}}$

begin

Draw $\tilde{X}_0\sim\mu_0$;

for $k\geq 1$ **do**

Draw $G_{k+1}\sim\mathcal{N}(0,\mathrm{Id})$ and $\tilde{Z}_{k+1}\sim\tilde{\eta}_1$;

Set $\tilde{X}_{k+1}=\mathrm{prox}_{U_2}^{\gamma_{k+1}}(\tilde{X}_k)-\gamma_{k+2}\tilde{\Theta}_1(\mathrm{prox}_{U_2}^{\gamma_{k+1}}(\tilde{X}_k),\tilde{Z}_{k+1})+\sqrt{2\gamma_{k+2}}G_k$

Let$(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ be two non-increasing sequences of reals numbers and $\mu_0\in\mathcal{P}_2(\mathbb{R}^d)$ be an initial distribution. The weighted averaged distribution associated with (32) $(\tilde{\nu}_n^N)_{n\in\mathbb{N}}$ is defined for all $N,n\in\mathbb{N}$, $n\geq 1$ by

$$\tilde{\nu}_n^N=\Lambda_{N,N+n}^{-1}\sum_{k=N+1}^{N+n}\lambda_k\,\mu_0\tilde{Q}_{\gamma}^k\ ,\qquad \tilde{Q}_{\gamma}^k=\tilde{R}_{\gamma_1,\gamma_2}\cdots\tilde{R}_{\gamma_k,\gamma_{k+1}}\ ,\ \text{for } k\in\mathbb{N}^*\ ,\tag{35}$$

where N is a burn-in time and $\Lambda_{N,N+n}$ is defined in (20). We take in the following the convention that $\tilde{Q}_{\gamma}^0$ is the identity operator.

Under **A**3, define for all $\mu\in\mathcal{P}_2(\mathbb{R}^d)$,

$$\begin{aligned}\upsilon_1(\mu)&=\int_{\mathbb{R}^d\times\tilde{\mathsf{Z}}}\left\|\tilde{\Theta}_1(x,z)-\int_{\tilde{\mathsf{Z}}}\tilde{\Theta}_1(x,\tilde{z})\mathrm{d}\tilde{\eta}_1(\tilde{z})\right\|^2\mathrm{d}\eta(z)\mathrm{d}\mu(x)\\&=\mathbb{E}\left[\left\|\tilde{\Theta}_1(\tilde{X}_0,\tilde{Z}_1)-\nabla U_1(\tilde{X}_0)\right\|^2\right]\ ,\end{aligned}\tag{36}$$

where $\tilde{X}_0,\tilde{Z}_1$ are independent random variables with distribution μ and $\tilde{\eta}_1$ respectively.

Theorem 17 *Assume* ***A***$5(m)$*, for $m\geq 0$. Let $(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ be two non-increasing sequences of positive real numbers satisfying $\gamma_1\in\left(0,L^{-1}\right]$, and for all $k\in\mathbb{N}^*$, $\lambda_{k+1}/\gamma_{k+2}\leq$*

λ_k/γ_{k+1}. *Let* $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ *and* $N \in \mathbb{N}$. *Then for all* $n \in \mathbb{N}^*$, *we have*

$$\mathrm{KL}\left(\tilde{\nu}_n^N \middle| \pi\right) \leq \lambda_{N+1} W_2^2\left(\mu_0 \tilde{Q}_\gamma^N \tilde{S}^2_{\gamma_{N+1}}, \pi\right) \Big/ (2\gamma_{N+2}\Lambda_{N,N+n}) + (2\Lambda_{N,N+n})^{-1} \sum_{k=N+1}^{N+n} \lambda_k \gamma_{k+1}\{2Ld + (1+\gamma_{k+1}L)\upsilon_1(\mu_0 Q_\gamma^{k-1}\tilde{S}^2_{\gamma_k}) + 2M_2^2\} .$$

Proof The proof is postponed to Section 7.4.1. ■

Corollary 18 *Assume* **A** 5(m), *for* $m \geq 0$. *Assume that* $\sup_{x\in\mathbb{R}^d} \upsilon_1(\delta_x) \leq D^2 < \infty$. *Let* $(\gamma_k)_{k\in\mathbb{N}^*}$ *and* $(\lambda_k)_{k\in\mathbb{N}^*}$ *given for all* $k \in \mathbb{N}^*$ *by* $\lambda_k = \gamma_k = \gamma \in \left(0, L^{-1}\right]$. *Let* $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. *Then for any* $N \in \mathbb{N}, n \in \mathbb{N}^*$ *we have*

$$\mathrm{KL}\left(\tilde{\nu}_n^N \middle| \pi\right) \leq W_2^2\left(\mu_0 \bar{Q}_\gamma^N \bar{S}_\gamma, \pi\right)/(2n\gamma) + \gamma\left(Ld + M_2^2 + D^2\right) ,$$

Furthermore, let $\varepsilon > 0$ *and*

$$\gamma_\varepsilon \leq \min\left\{\varepsilon/(2(Ld + M_2^2 + D^2)), L^{-1}\right\} , \qquad n_\varepsilon \geq \lceil W_2^2(\mu_0 \tilde{S}^2_{\gamma_1}, \pi)(\gamma_\varepsilon \varepsilon)^{-1}\rceil .$$

Then we have $\mathrm{KL}\left(\tilde{\nu}^0_{n_\varepsilon} \middle| \pi\right) \leq \varepsilon$.

In the case where a warm start is available for the Wasserstein distance, *i.e.* $W_2^2(\mu_0, \pi) \leq C$, for some absolute constant $C \geq 0$, then Corollary 18 implies that the complexity of SPGLD to obtain a sample close from π in KL with a precision target $\varepsilon > 0$ is of order $(d + M_2^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-2})$. In addition, Pinsker inequality implies that the complexity of SPGLD for the total variation distance is of order $(d + M_2^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-4})$.

In addition if we have access to $\eta > 0$ and $M_\eta \geq 0$, independent of the dimension, such that for all $x \in \mathbb{R}^d$, $x \notin \mathrm{B}(x^\star, M_\eta)$, $U(x) - U(x^\star) \geq \eta \left\|x - x^\star\right\|$, Proposition 32, **A**5-(ii), (31) and (25) imply that starting at $\delta_{x^\star}$, the overall complexity of SPGLD for the KL is in this case $(\eta^{-2}d^2 + M_\eta^2)(d + M_2^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-2})$ and $(\eta^{-2}d^2 + M_\eta^2)(d + M_2^2 + D^2)\bar{\mathcal{O}}(\varepsilon^{-4})$ for the total variation distance If $(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ are given for all $k \in \mathbb{N}^*$ by $\gamma_k = \lambda_k = \gamma_1/k^{-\alpha}$, $\gamma_1 \in \left(0, L^{-1}\right]$. Then by the same reasoning as in the proof of Corollary 8, we obtain that there exists $C \geq 0$ such that for all $n \in \mathbb{N}^*$, we have $\mathrm{KL}\left(\bar{\nu}_n^0 \middle| \pi\right) \leq C \max(n^{\alpha-1}, n^{-\alpha})$, if $\alpha \neq 1/2$, and for $\alpha = 1/2$, we have $\mathrm{KL}\left(\bar{\nu}_n^0 \middle| \pi\right) \leq C(\ln(n) + 1)n^{-1/2}$.

If $\sup_{x\in\mathbb{R}^d} \upsilon_1(\delta_x) < +\infty$ does not hold, we can control the variance of stochastic gradient estimates using **A**4 again based on this following result.

Proposition 19 *Assume* **A** 5 *and* $\tilde{\Theta}_1$ *satisfies* **A** 4. *Then we have for all* $x \in \mathbb{R}^d$ *and* $\gamma \in (0, \tilde{L}^{-1}]$

$$2\gamma(\tilde{L}^{-1} - \gamma)\upsilon_1(\delta_x) \leq \left\|x - x^\star\right\|^2 - \int_{\mathbb{R}^d} \left\|y - x^\star\right\|^2 (\tilde{S}^1_\gamma T_\gamma \tilde{S}^2_\gamma)(x, \mathrm{d}y) + 2\gamma^2 \upsilon_1(\delta_{x^\star}) + 2\gamma d ,$$

where $\tilde{S}^1_\gamma, \tilde{S}^2_\gamma$ *and* υ_1 *are defined by* (34)-(36) *respectively.*

Proof Let $\gamma > 0$, $x \in \mathbb{R}^d$ and consider $\tilde{X}_1 = \mathrm{prox}_{U_2}^{\gamma}\left\{x - \gamma\tilde{\Theta}_1(x, \tilde{Z}_1) + \sqrt{2\gamma}G_1\right\}$, where $\tilde{Z}_1$ and G_1 are two independent random variables, $\tilde{Z}_1$ has distribution $\tilde{\eta}_1$ and G_1 is the standard Gaussian random variable, so that $\tilde{X}_1$ has distribution $\tilde{S}_\gamma^1 T_\gamma \tilde{S}_\gamma^2(x, \cdot)$. First by (Bauschke and Combettes, 2011, Theorem 26.2(vii)), we have that $x^\star = \mathrm{prox}_{U_2}^{\gamma}(x^\star - \gamma\nabla U_1(x^\star))$ and by (Bauschke and Combettes, 2011, Proposition 12.27), the proximal is non-expansive, for all $x, y \in \mathbb{R}^d$, $\|\mathrm{prox}_{U_2}^{\gamma}(x) - \mathrm{prox}_{U_2}^{\gamma}(y)\| \leq \|x - y\|$. Using these two results and the fact that $\tilde{\Theta}_1$ satisfies **A** 4, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\tilde{X}_1 - x^\star\right\|^2\right] &= \mathbb{E}\left[\left\|\mathrm{prox}_{U_2}^{\gamma}\left\{x - \gamma\tilde{\Theta}_1(x, \tilde{Z}_1) + \sqrt{2\gamma}G_1\right\} - \mathrm{prox}_{U_2}^{\gamma}\{x^\star - \gamma\nabla U_1(x^\star)\}\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\left(x - \gamma\tilde{\Theta}_1(x, \tilde{Z}_1) + \sqrt{2\gamma}G_1\right) - (x^\star - \gamma\nabla U_1(x^\star))\right\|^2\right] \\
&\leq \|x - x^\star\|^2 \\
&\quad + \mathbb{E}\left[2\gamma\left\langle x - x^\star, \nabla U_1(x^\star) - \tilde{\Theta}_1(x, \tilde{Z}_1)\right\rangle + \gamma^2\left\|\nabla U_1(x^\star) - \tilde{\Theta}_1(x, \tilde{Z}_1)\right\|^2\right] + 2\gamma d \\
&\leq \|x - x^\star\|^2 - 2\gamma(\tilde{L}^{-1} - \gamma)\mathbb{E}\left[\left\|\tilde{\Theta}_1(x, \tilde{Z}_1) - \tilde{\Theta}_1(x^\star, \tilde{Z}_1)\right\|^2\right] \\
&\qquad + 2\gamma^2\mathbb{E}\left[\left\|\tilde{\Theta}_1(x^\star, \tilde{Z}_1) - \nabla U_1(x^\star)\right\|^2\right] + 2\gamma d\,.
\end{aligned}
$$

The proof is completed upon noting that $\upsilon_1(\delta_x) \leq \mathbb{E}[\left\|\tilde{\Theta}_1(x, \tilde{Z}_1) - \tilde{\Theta}_1(x^\star, \tilde{Z}_1)\right\|^2]$. ■

Combining Theorem 17 and Proposition 19, we get the following result.

Corollary 20 *Assume* ***A*** *5(m) for $m \geq 0$ and that $\tilde{\Theta}_1$ satisfies* ***A*** *4. Let $(\gamma_k)_{k\in\mathbb{N}^*}$ and $(\lambda_k)_{k\in\mathbb{N}^*}$ be two non-increasing sequences of positive real numbers given for all $k \in \mathbb{N}^*$ by $\gamma_k = \lambda_k = \gamma \in (0, L^{-1}]$, $\gamma < \tilde{L}^{-1}$. Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $N \in \mathbb{N}$. Then for all $n \in \mathbb{N}^*$, it holds*

$$
\begin{aligned}
\mathrm{KL}\left(\tilde{\nu}_n^N | \pi\right) &\leq W_2^2\left(\mu_0\tilde{Q}_\gamma^N\tilde{S}_{\gamma_{N+1}}^2, \pi\right)\Big/(2\gamma n) + \gamma(Ld + M_2^2) \\
&\quad + (1 + \gamma L)(2(\tilde{L}^{-1} - \gamma))^{-1}\left\{(2n)^{-1}\int_{\mathbb{R}^d}\|x - x^\star\|^2\,\mathrm{d}\mu_0\tilde{Q}_\gamma^N\tilde{S}_\gamma^2(y) + \gamma^2\upsilon_1(\delta_{x^\star}) + \gamma d\right\}\,.
\end{aligned}
$$

Furthermore, for $\varepsilon > 0$, consider stepsize and a number of iterations satisfying:

$$
\begin{aligned}
\gamma_\varepsilon &\leq \min\left[\varepsilon\Big/\left\{4M_2^2 + 4Ld + 8\tilde{L}d\right\}, \sqrt{\varepsilon/\left(8\tilde{L}\upsilon_1(\delta_{x^\star})\right)}, L^{-1}, (2\tilde{L})^{-1}\right]\,, \\
n_\varepsilon &\geq 2\max\left\{\left\lceil W_2^2(\mu_0\tilde{S}_{\gamma_\varepsilon}^2, \pi)(\gamma_\varepsilon\varepsilon)^{-1}\right\rceil, \left\lceil 2\tilde{L}\varepsilon^{-1}\int_{\mathbb{R}^d}\|x - x^\star\|^2\,\mathrm{d}\mu_0\tilde{S}_\gamma^2(y)\right\rceil\right\}\,.
\end{aligned}
$$

Then, we have $\mathrm{KL}\left(\tilde{\nu}_{n_\varepsilon}^0 | \pi\right) \leq \varepsilon$.

Proof The proof of the corollary is a direct consequence of Theorem 17 and Proposition 19, and is postponed to Section 7.4.2. ■

Note that the dependency on the variance of the stochastic gradients is improved compared to the bound given by Corollary 18. We specify once again the result of Theorem 17 for strongly convex potential.

Theorem 21 *Assume* ***A*** *5(m), for $m > 0$. Let $(\gamma_k)_{k\in\mathbb{N}^*}$ be a non-increasing sequences of positive real numbers satisfying for all $k \in \mathbb{N}^*$, $\gamma_k \in \left(0, L^{-1}\right]$. Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then for all $n \in \mathbb{N}^*$, it holds*

$$W_2^2(\mu_0 \tilde{Q}_\gamma^n \tilde{S}^2_{\gamma_{n+1}}, \pi) \leq \left\{ \prod_{k=1}^{n} (1 - m\gamma_{k+1}) \right\} W_2^2(\mu_0 \tilde{S}^2_{\gamma_1}, \pi) + \sum_{k=1}^{n} \gamma_{k+1}^2 \left\{ \prod_{i=k+2}^{n+1} (1 - m\gamma_i) \right\} \{2Ld + (1 + \gamma_{k+1}L)\upsilon_1(\mu_0 \tilde{Q}_\gamma^{k-1} \tilde{S}^2_{\gamma_k}) + 2M_2^2\} .$$

Proof The proof is postponed to Section 7.4.3. ■

Corollary 22 *Assume* ***A*** *5(m), for $m > 0$. Assume that $\sup_{x\in\mathbb{R}^d} \upsilon_1(\delta_x) \leq D^2 < \infty$. Let $\varepsilon > 0$, $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, and*

$$\gamma_\varepsilon \leq \min\left\{ m\varepsilon/(4(Ld + D^2 + M_2^2)), L^{-1} \right\} , \qquad n_\varepsilon \geq \lceil \ln(2W_2^2(\mu_0 \tilde{S}^2_{\gamma_\varepsilon}, \pi)/(\varepsilon\gamma_\varepsilon m)^{-1} \rceil .$$

Then $W_2^2(\mu_0 \tilde{R}^{n_\varepsilon}_{\gamma_\varepsilon,\gamma_\varepsilon} \tilde{S}^2_{\gamma_\varepsilon}, \pi) \leq \varepsilon$, where $\tilde{R}_{\gamma,\gamma}$ and $\tilde{S}^2_\gamma$ are defined by (33) and (34) respectively.

Proof Since $\gamma_\varepsilon \leq L^{-1}$, we have $(1 + \gamma_\varepsilon L)\upsilon_1(\mu_0 \tilde{R}^k_{\gamma_\varepsilon} \tilde{S}^2_\gamma) \leq 2D^2$ for all $k \geq 1$. Using Theorem 21 then concludes the proof. ■

Note that the bounds given by Theorem 21 are tighter the one given by Dalalyan and Karagulyan (2017, Theorem 3) which shows under **A**5 with $U_2 = 0$ and $\sup_{x\in\mathbb{R}^d} \upsilon_1(\delta_x) \leq D^2$ that

$$W_2(\mu_0 \tilde{R}_{\gamma,\gamma}, \pi) \leq (1 - mh)W_2(\mu_0, \pi) + 1.65(L/m)(\gamma d)^{1/2} + D^2(\gamma d)^{1/2}/(1.65L + Dm) .$$

Indeed, for constant stepsize $\gamma_k = \gamma \in (0, L^{-1}]$ for all $k \in \mathbb{N}^*$, Theorem 21 implies with the same assumptions that

$$W_2(\mu_0 \tilde{R}_{\gamma,\gamma}, \pi) \leq (1 - mh)^{1/2} W_2(\mu_0, \pi) + (2Ld\gamma/m)^{1/2} + ((1 + \gamma)\gamma/m)^{1/2} D .$$

As for ULA, the dependency on the condition number L/m is improved.

In the strongly convex case, we can improve the dependency on the variance of the stochastic gradient under the following condition.

A 6 There exist $\tilde{L}_1, \tilde{m}_1 > 0$ such that for all for $\tilde{\eta}_1$-almost every $z \in \tilde{\mathsf{Z}}$, for all $x, y \in \mathbb{R}^d$, we have

$$\left\langle \tilde{\Theta}_1(x, z) - \tilde{\Theta}_1(y, z), x - y \right\rangle \geq \tilde{m}_1 \|x - y\|^2 + (1/\tilde{L}_1) \left\| \tilde{\Theta}_1(x, z) - \tilde{\Theta}_1(y, z) \right\|^2 .$$

The condition **A**6 is for example satisfied if η-almost surely, $x \mapsto \tilde{\Theta}_1(x,z)$ is strongly convex, see (Nesterov, 2004, Theorem 2.1.12).

Proposition 23 *Assume* ***A*** *5(m) for $m > 0$ and* ***A*** *6. Then for all $\gamma > 0$ we have*

$$2\gamma(\tilde{L}_1^{-1} - \gamma)\upsilon_1(\delta_x) \leq (1 - \tilde{m}_1\gamma)\left\|x - x^\star\right\|^2 - \int_{\mathbb{R}^d} \left\|y - x^\star\right\|^2 (\tilde{S}_\gamma^1 T_\gamma \tilde{S}_\gamma^2)(x, \mathrm{d}y) + 2\gamma^2 \upsilon_1(\delta_{x^\star}) + 2\gamma d \ ,$$

where $\tilde{S}_\gamma^1, \tilde{S}_\gamma^2$ and υ_1 are defined by (34)-(36) *respectively.*

Proof The proof is similar to the proof of Proposition 19 and is postponed to Section 7.4.4. ■

Corollary 24 *Assume* ***A*** *5(m), for $m > 0$ and that $\tilde{\Theta}_1$ satisfies* ***A*** *6. Let $(\gamma_k)_{k\in\mathbb{N}^*}$ defined for all $k \in \mathbb{N}^*$ by $\gamma_k = \gamma \in (0, L^{-1} \wedge (2\tilde{L}_1)^{-1}]$. Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Define $\tilde{m} = \min(m, \tilde{m}_1)$ and*

$$\begin{aligned} \Delta_1 &= 2(Ld + M_2)/m + \{2\tilde{L}_1(1+\gamma L)/\tilde{m}\}d \\ \Delta_2 &= \{2\tilde{L}_1(1+\gamma L)/\tilde{m}\}\upsilon_1(\delta_{x^\star}) \\ \Delta_3 &= \gamma\tilde{L}_1(1+\gamma L)\left\{\int_{\mathbb{R}^d} \left\|x - x^\star\right\|^2 \mathrm{d}\mu_0 \tilde{S}_{\gamma_\varepsilon}^2(x)\right\} \ . \end{aligned} \tag{37}$$

Then for all $n \in \mathbb{N}^*$, it holds

$$W_2^2(\mu_0 \tilde{R}_{\gamma,\gamma}^n \tilde{S}_\gamma^2, \pi) \ \leq \ (1 - m\gamma)^n W_2^2(\mu_0 \tilde{S}_\gamma^2, \pi) + (1 - \tilde{m}\gamma)^n \Delta_3 + \gamma\Delta_1 + \gamma^2\Delta_2 \ , \tag{38}$$

where $\tilde{R}_{\gamma,\gamma}$ and $\tilde{S}_\gamma^2$ are defined by (33) *and* (34).

Therefore, for $\varepsilon > 0$ and

$$\begin{aligned} \gamma_\varepsilon &\leq \min\left\{\varepsilon/(4\Delta_1), [\varepsilon/(4\Delta_2)]^{1/2}, L^{-1}, (2\tilde{L}_1)^{-1}\right\} \\ n_\varepsilon &\geq \max\left\{\lceil \ln(4W_2^2(\mu_0\tilde{S}_{\gamma_\varepsilon}^2, \pi)/\varepsilon)(\gamma_\varepsilon m)^{-1}\rceil, \lceil \ln(4\Delta_3/\varepsilon)(\gamma_\varepsilon\tilde{m})^{-1}\rceil\right\} \ , \end{aligned}$$

it holds $W_2^2(\mu_0 \tilde{R}_{\gamma_\varepsilon,\gamma_\varepsilon}^{n_\varepsilon} \tilde{S}_{\gamma_\varepsilon}^2, \pi) \leq \varepsilon$.

Proof The proof of the corollary is postponed to Section 7.4.5. ■

Corollary 25 *Assume* ***A*** *5(m), for $m > 0$ and that $\tilde{\Theta}_1$ satisfies* ***A*** *6. Define $\tilde{m} = \min(m, \tilde{m}_1)$. Let $\varepsilon > 0$, $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and*

$$\begin{aligned} \gamma_\varepsilon &\leq \min\left\{\varepsilon/(4\Delta_1), [\varepsilon/(4\Delta_2)]^{1/2}, L^{-1}, (2\tilde{L}_1)^{-1}\right\} \ , \\ N_\varepsilon &\geq \max\left\{\lceil \ln(4W_2^2(\mu_0\tilde{S}_{\gamma_\varepsilon}^2, \pi)/\varepsilon)(\gamma_\varepsilon m)^{-1}\rceil, \lceil \ln(4\Delta_3/\varepsilon)(\gamma_\varepsilon\tilde{m})^{-1}\rceil\right\} \\ \tilde{\gamma}_\varepsilon &\leq \min\left[\varepsilon \Big/ \left\{4M_2^2 + 4Ld + 8\tilde{L}d\right\}, \sqrt{\varepsilon / \left(8\tilde{L}\upsilon_1(\delta_{x^\star})\right)}, L^{-1}, (2\tilde{L})^{-1}\right] \ , \\ n_\varepsilon &\geq 2\max\left\{\left\lceil \gamma_\varepsilon^{-1}\right\rceil, \left\lceil 2\tilde{L}\varepsilon^{-1}\int_{\mathbb{R}^d} \left\|x - x^\star\right\|^2 \mathrm{d}\mu_0 \tilde{R}_{\gamma_\varepsilon,\gamma_\varepsilon}^{N_\varepsilon} \tilde{S}_\gamma^2(y)\right\rceil\right\} \ , \end{aligned}$$

where $\Delta_1, \Delta_2, \Delta_3$ *are defined in* (37) *and* $\tilde{R}_{\gamma,\gamma}$ *and* $\tilde{S}^2_\gamma$ *are defined by* (33) *and* (34). *Let* $(\gamma_k)_{k\in\mathbb{N}}$ *defined by* $\gamma_k = \gamma_\varepsilon$ *for* $k \in \{1, \dots, N_\varepsilon\}$ *and* $\gamma_k = \tilde{\gamma}_\varepsilon$ *for* $k > N_\varepsilon$. *Then we have* $\mathrm{KL}\left(\tilde{\nu}^{N_\varepsilon}_{n_\varepsilon} | \pi\right) \leq \varepsilon$ *where* $\tilde{\nu}^{N_\varepsilon}_{n_\varepsilon} = n_\varepsilon^{-1} \sum_{k=1}^{n_\varepsilon} \mu_0 \tilde{R}^{N_\varepsilon}_{\gamma_\varepsilon,\gamma_\varepsilon} \tilde{R}^k_{\tilde{\gamma}_\varepsilon,\tilde{\gamma}_\varepsilon}$.

Proof Corollary 24 implies that after the burn in phase of N_ε steps with stepsize γ_ε, we have $W_2^2(\mu_0 \tilde{Q}^{N_\varepsilon}_\gamma \tilde{S}^2_{\gamma_\varepsilon}, \pi) \leq \varepsilon$. Then, since we can treat $\mu_0 \tilde{Q}^{N_\varepsilon}_\gamma$ as a new starting measure, Corollary 20 concludes the proof. ■

4.2.1. Discussion on Related Works

Note that the SPGLD is different from the algorithm MYULA proposed by Durmus et al. (2018) which approximates U_2 by its Moreau envelope, and under **A**5 defines the Markov chain $(X^{\mathrm{M}}_k)_{k\in\mathbb{N}}$ by the recursion:

$$X^{\mathrm{M}}_{k+1} = (1 - \gamma/\lambda^{\mathrm{M}}) + (\gamma/\lambda^{\mathrm{M}}) \operatorname{prox}^{\lambda^{\mathrm{M}}}_{U_2}(X^{\mathrm{M}}_k) - \gamma \nabla U_1(X^{\mathrm{M}}_k) + \sqrt{2\gamma} G_{k+1} \;,$$

for a stepsize $\gamma > 0$, $(G_k)_{k\in\mathbb{N}^*}$ a sequence of i.i.d. d-dimensional standard Gaussian random variables and $\lambda^{\mathrm{M}} > 0$ a regularization parameter. So taking $\lambda^{\mathrm{M}} = \gamma$, we get that the recursion boils down to

$$X^{\mathrm{M}}_{k+1} = \operatorname{prox}^{\gamma}_{U_2}(X^{\mathrm{M}}_k) - \gamma \nabla U_1(X^{\mathrm{M}}_k) + \sqrt{2\gamma} G_{k+1} \;.$$

Then the main difference with (32) setting $\gamma_k = \gamma$ for all $k \in \mathbb{N}^*$ and $\tilde{\Theta}_1 = \nabla U_1$ is that $\nabla U_1(X^{\mathrm{M}}_k)$ is replaced in (32) by $\nabla U_1(\operatorname{prox}^{\gamma}_{U_2}(X^{\mathrm{M}}_k))$.

Recently, two papers have independently interpreted ULA as an algorithm that optimizes $\mathscr{F}$ on the Wasserstein space. They both relies on this interpretation to derive and analyze different algorithms for sampling log-concave target measures, that use the proximal operator associated with U.

First, Wibisono (2018) proposed and analyzed the symmetrized Langevin algorithm (SLA) in order to reduce the discretization bias. SLA combines backward and forward steps, *i.e.* it defines the Markov chain $(X^{\mathrm{SLA}}_k)_{k\in\mathbb{N}}$ by the recursion

$$X^{\mathrm{SLA}}_{k+1} = \operatorname{prox}^{\gamma}_{U}\left\{X^{\mathrm{SLA}}_k - \gamma \nabla U(X^{\mathrm{SLA}}_k) + \sqrt{4\gamma} G_{k+1}\right\} \;, \tag{39}$$

for a stepsize $\gamma > 0$ and $(G_k)_{k\in\mathbb{N}^*}$ a sequence of i.i.d. d-dimensional standard Gaussian random variables. Wibisono (2018) shows that in the Gaussian case, *i.e.* for any $x \in \mathbb{R}^d$, $U(x) = \langle (x - \bar{x})\Sigma^{-1}, (x - \bar{x})\rangle/2$ for some mean $\bar{x} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, SLA is unbiased and converges with an exponential rate to π. For general strongly convex potentials U, exponential convergence in Wasserstein distance of the sequence of distributions associated with $(X^{\mathrm{SLA}}_k)_{k\in\mathbb{N}}$ to a biased limit $\pi^{\mathrm{SLA}}_\gamma$ is established as $k \to +\infty$. However, no explicit bound on $W_2(\pi_\gamma, \pi^{\mathrm{SLA}})$ or nonasymptotic convergence rates are given.

In a parallel work, Bernton (2018) considers the proximal Langevin algorithm (PLA). This algorithm defines the Markov chain $(X^{\mathrm{PLA}}_k)_{k\in\mathbb{N}}$ by the recursion

$$X^{\mathrm{PLA}}_{k+1} = \operatorname{prox}^{\gamma_{k+1}}_{U}(X^{\mathrm{PLA}}_k) - \gamma\sqrt{2\gamma_{k+1}} G_{k+1} \;.$$

Applying techniques from Ambrosio et al. (2008), quantitative results on the Wasserstein distance between the above discretization and gradient flow of KL divergence are obtained. From those results, bounds on the Wasserstein distance between iterates and target distributions are given, in the case where U is strongly convex. The complexity bounds for PLA obtained in Bernton (2018) are of order $d\bar{\mathcal{O}}(\varepsilon^{-2})$ when U is smooth and strongly convex, and are equivalent to our results up to dependence on the starting measure. In the case of $U = U_1 + U_2$ where U_1 is strongly convex and U_2 is Lipschitz, bounds in Bernton (2018) are of order $d^2\bar{\mathcal{O}}(\varepsilon^{-4})$ while our bounds are still of the same order as in the smooth case.

5. Numerical Experiments

In this section, we experiment SPGLD and SSGLD on a Bayesian logistic regression problem, see e.g. (Holmes and Held, 2006; Gramacy and Polson, 2012; Park and Hastie, 2007). Consider i.i.d. observations $(X_i, Y_i)_{i\in\{1,\dots,N\}}$, where $(Y_i)_{i\in\{1,\dots,N\}}$ are binary response variables and $(X_i)_{i\in\{1,\dots,N\}}$ are d-dimensional covariance variables. For all $i \in \{1,\dots,N\}$, Y_i is assumed to be a Bernoulli random variable with parameter $\Phi(\beta^{\mathrm{T}} X_i)$ where β is the parameter of interest and for all $u \in \mathbb{R}$, $\Phi(u) = \mathrm{e}^u/(1+\mathrm{e}^u)$. We choose as prior distributions (see Genkin et al. (2007) and Li and Lin (2010)) a d-dimensional Laplace distribution and a combination of the Laplace distribution and the Gaussian distribution, with density with respect to the Lebesgue measure given respectively for all $\beta \in \mathbb{R}^d$ by

$$\mathrm{p}_1(\beta) \propto \exp\left(-a_1 \sum_{i=1}^{d} |\beta_i|\right) , \qquad \mathrm{p}_{1,2}(\beta) \propto \exp\left(-a_1 \sum_{i=1}^{d} |\beta_i| - a_2 \sum_{i=1}^{d} \beta_i^2\right) ,$$

where a_1 is set to 1 in the case of p_1 and $a_1 = 0.9,\ a_2 = 0.1$ in the case of $\mathrm{p}_{1,2}$. The chosen priors on the one hand reduce impact of the irrelevant features by shrinking them close to zero. On the other hand this choice of priors leads to the log-concave posteriors. Both these property are highly desirable in the high dimensional setting. We obtain then the two different a posteriori distributions $\mathrm{p}_1(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$ and $\mathrm{p}_{1,2}(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$ with potentials given, respectively, by

$$\beta \mapsto \sum_{n=1}^{N} \ell_n(\beta) + a_1 \sum_{i=1}^{d} |\beta_i| , \qquad \beta \mapsto \sum_{n=1}^{N} \ell_n(\beta) + a_2 \sum_{i=1}^{d} \beta_i^2 + a_1 \sum_{i=1}^{d} |\beta_i| .$$

where

$$\ell_n(\beta) = -Y_n \beta^{\mathrm{T}} X_n + \log[1 + \exp(\beta^{\mathrm{T}} X_n)] .$$

We consider the three data sets from UCI repository (Dua and Efi, 2017) Heart disease dataset ($N = 270$, $d = 14$), Australian Credit Approval dataset ($N = 690$, $d = 34$) and Musk dataset ($N = 476$, $d = 166$). We approximate $\mathrm{p}_1(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$ using SPGLD and SSGLD, since the associated potential is Lipschitz, whereas regarding $\mathrm{p}_{1,2}(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$ we only apply SPGLD.

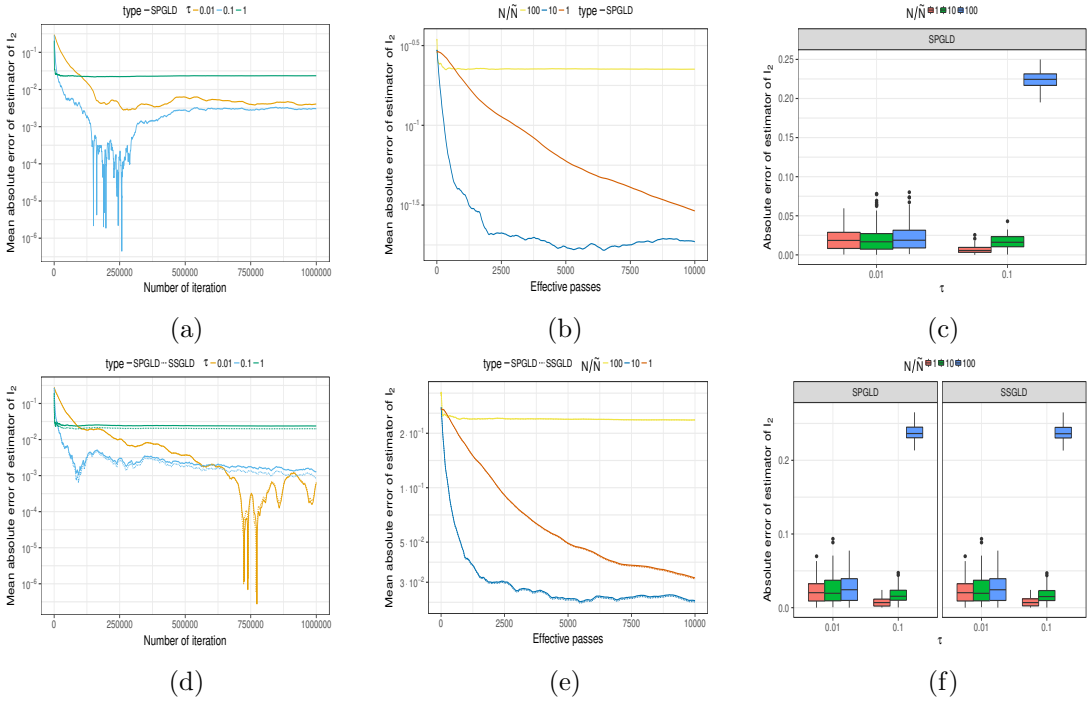

Figure 1: Mean absolute error of estimator of I_2 for Australian Credit Approval dataset: (top row) results for $\mathrm{p}_{1,2}(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$; (a) convergence of SPGLD for $\tilde{N} = N$, (b) convergence of SPGLD in terms of effective passes for $\tau = 0.1$, (c) boxplot of SPGLD for full runs; (bottom row) results for $\mathrm{p}_1(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$; (d) convergence of SPGLD and SSGLD for $\tilde{N} = N$, (e) convergence of SPGLD and SSGLD in terms of effective passes for $\tau = 0.1$, (f) boxplot of SPGLD and SSGLD for full run.

SPGLD is performed using the following stochastic gradient

$$\tilde{\Theta}_1(\beta, Z) = (N/\tilde{N}) \sum_{n\in Z} \nabla \ell_n(\beta) + a_2\beta \,,$$

where a_2 is set to 0 in the case of $\mathrm{p}_1(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$ and Z is a uniformly distributed random subset of $\{1,\dots,N\}$ with cardinal $\tilde{N} \in \{1,\dots,N\}$. In addition, the proximal operator associated with $\beta \mapsto a_1 \sum_{i=1}^d |\beta_i|$ is given for all $\beta \in \mathbb{R}^d$ and $\gamma > 0$ by (see e.g. Parikh and Boyd (2013))

$$(\mathrm{prox}^{\gamma}_{a_1,\ell_1}(\beta))_i = \mathrm{sign}(\beta_i)\max(|\beta_i| - a_1\gamma, 0) \,, \quad \text{for } i \in \{1,\dots,d\} \,.$$

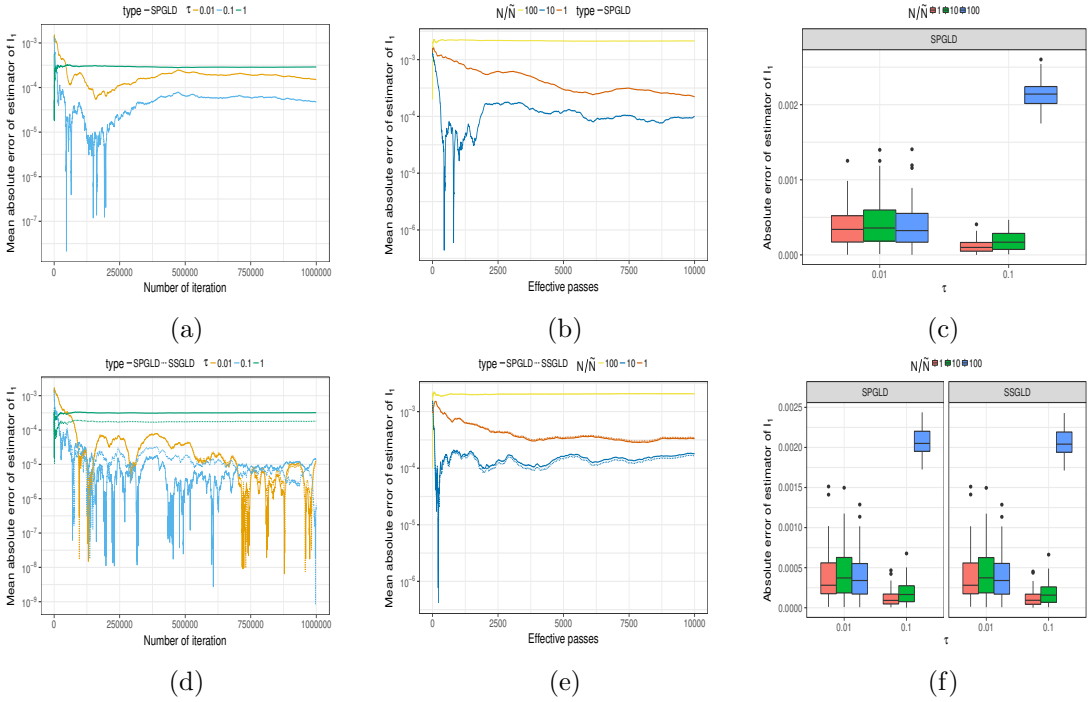

Figure 2: Mean absolute error of estimator of I_1 for Australian Credit Approval dataset: (top row) results for $\mathrm{p}_{1,2}(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$; (a) convergence of SPGLD for $\tilde{N} = N$, (b) convergence of SPGLD in terms of effective passes for $\tau = 0.1$, (c) boxplot of SPGLD for full runs; (bottom row) results for $\mathrm{p}_1(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$; (d) convergence of SPGLD and SSGLD for $\tilde{N} = N$, (e) convergence of SPGLD and SSGLD in terms of effective passes for $\tau = 0.1$, (f) boxplot of SPGLD and SSGLD for full run.

SSGLD is performed using the following stochastic subgradient

$$\Theta(\beta, Z) = (N/\tilde{N}) \sum_{n\in Z} \nabla \ell_n(\beta) + a_1 \sum_{i=1}^{d} \mathrm{sign}(\beta_i)\mathbf{e}_i \ ,$$

where $(\mathbf{e}_i)_{i\in\{1,\dots,d\}}$ denotes the canonical basis and Z is a uniformly distributed random subset of $\{1,\dots,N\}$ with cardinal $\tilde{N} \in \{1,\dots,N\}$.

Based on the results of SPGLD and SSGLD, we estimate the posterior mean I_1 and I_2 of the test functions $\beta \mapsto \beta_1$ and $\beta \mapsto (1/d)\sum_{i=1}^{d} \beta_i^2$. For our experiments, we use constant stepsizes γ of the form $\tau(L+m)^{-1}$ with $\tau = 0.01, 0.1, 1$ and for stochastic (sub) gradient we use $\tilde{N} = N, \lfloor N/10 \rfloor, \lfloor N/100 \rfloor$. For all datasets and all settings of τ, $\tilde{N}$ we run 100 independent runs of SPGLD (SSGLD), where each run was of length 10^6. For each set of parameters we estimate I_1, I_2 and we compute the absolute errors, where the true value were obtained by prox-MALA (see Pereyra (2015)) with 10^7 iterations and stepsize corresponding to optimal acceptance ratio ≈ 0.5, see (Roberts and Rosenthal, 1998). The results for I_2 are presented on Figure 1, Figure 3 and Figure 5 for Australian Credit Approval dataset,

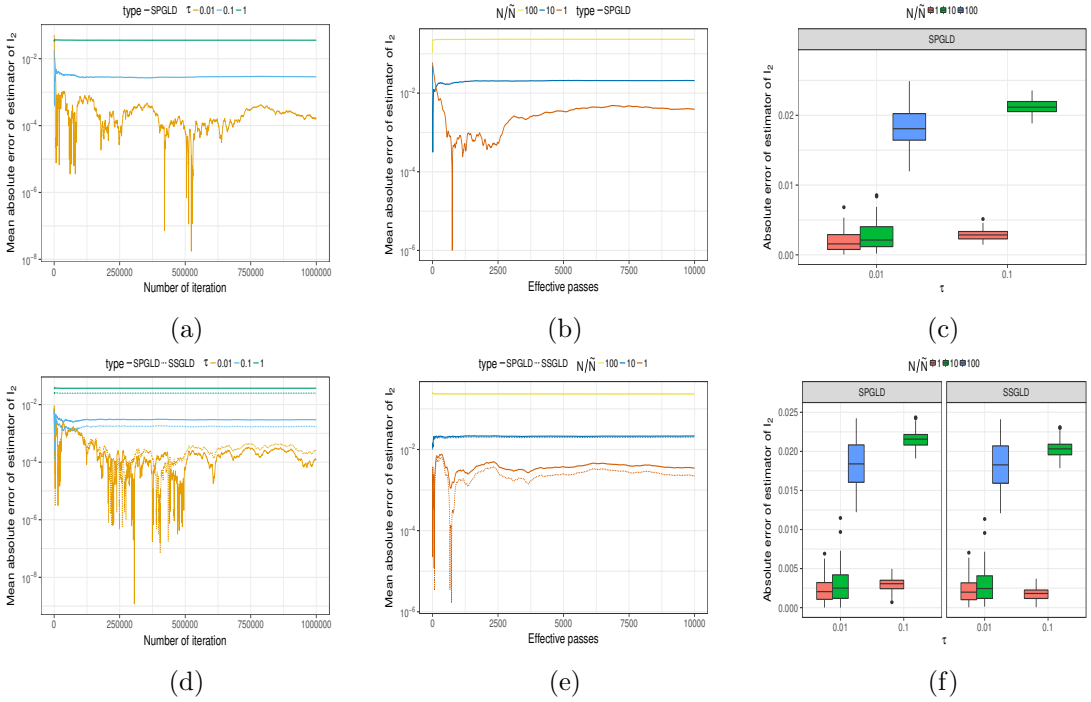

Figure 3: Mean absolute error of estimator of I_2 for Heart disease dataset: (top row) results for $\mathrm{p}_{1,2}(\cdot|(X,Y)_{i\in\{1,\ldots,N\}})$; (a) convergence of SPGLD for $\tilde{N} = N$, (b) convergence of SPGLD in terms of effective passes for $\tau = 0.1$, (c) boxplot of SPGLD for full run; (bottom row) results for $\mathrm{p}_1(\cdot|(X,Y)_{i\in\{1,\ldots,N\}})$; (d) convergence of SPGLD and SSGLD for $\tilde{N} = N$, (e) convergence of SPGLD and SSGLD in terms of effective passes for $\tau = 0.1$, (f) boxplot of SPGLD and SSGLD for full run.

Heart disease dataset and Musk data respectively. The results for I_1 are presented on Figure 2, Figure 4 and Figure 6 for Australian Credit Approval dataset, Heart disease dataset and Musk data respectively. We note that in the all cases, bias decreases but convergence becomes slower with decreasing γ. When we look for stochastic (sub)gradient then the bias of estimators and also their variance increase when we decrease $\tilde{N}$. However if we look for the effective passes, i.e. the number of iteration is scaled with the cost of computing gradients, we observe that convergence is faster with reasonably small $\tilde{N}$. If we compare SSGLD with SPGLD we see that in the almost all cases, except Musk dataset, SSGLD leads to slightly smaller bias. For the Musk dataset differences between SSGLD and SPGLD are negligible and we do not present the results for SPGLD. In the presented experiments, all results agrees with our theoretical findings and suggest that SPGLD or SSGLD could be an alternative for other MCMC methods.

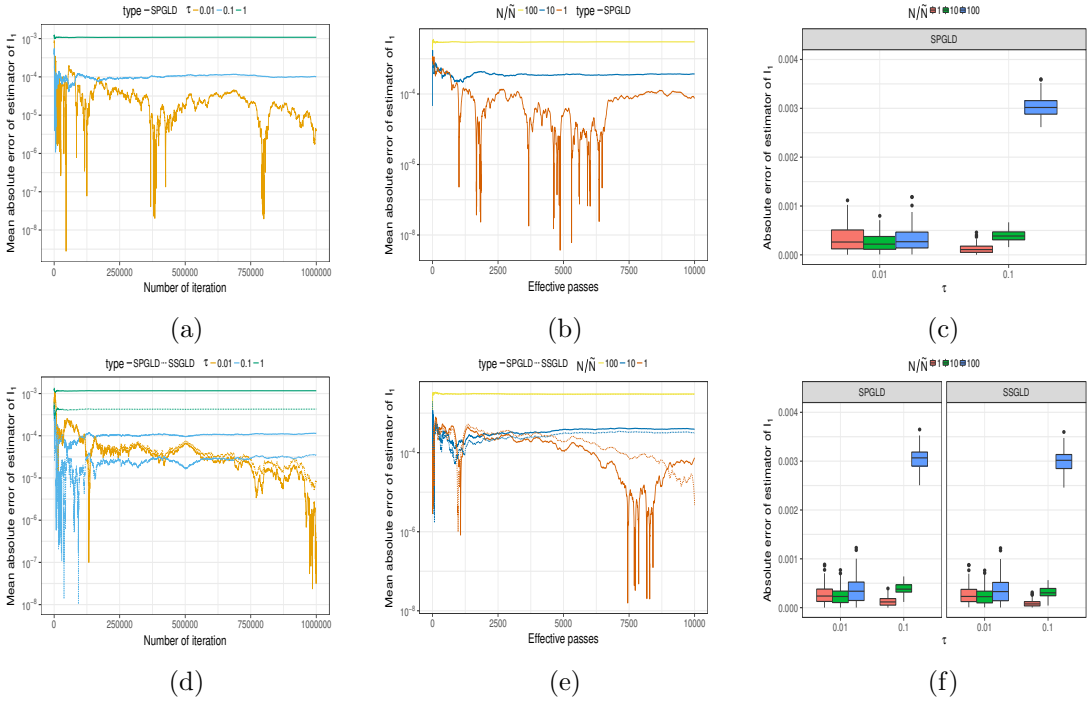

Figure 4: Mean absolute error of estimator of I_1 for Heart disease dataset: (top row) results for $\mathrm{p}_{1,2}(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$; (a) convergence of SPGLD for $\tilde{N} = N$, (b) convergence of SPGLD in terms of effective passes for $\tau = 0.1$, (c) boxplot of SPGLD for full run; (bottom row) results for $\mathrm{p}_1(\cdot|(X,Y)_{i\in\{1,\dots,N\}})$; (d) convergence of SPGLD and SSGLD for $\tilde{N} = N$, (e) convergence of SPGLD and SSGLD in terms of effective passes for $\tau = 0.1$, (f) boxplot of SPGLD and SSGLD for full run.

6. Discussion

In this paper, we presented a novel interpretation of the Unadjusted Langevin Algorithm as the first order optimization algorithm, and a new technique of proving non-asymptotic bounds for ULA, based on the proof techniques known from convex optimization. Our proof technique gives simpler proofs of some of the previously known non-asymptotic results for ULA. It can be also used to prove non-asymptotic bound that were previously unknown. Specifically, to the best of the authors knowledge, we provide the first non-asymptotic results for Stochastic Gradient ULA in the non-strongly convex case, as well as the first non-asymptotic results in the non-smooth non-strongly convex case. Furthermore, our technique extends effortlessly to the stochastic non-smooth case, and to the best of the authors knowledge we provide the first non-asymptotic analysis of that case.

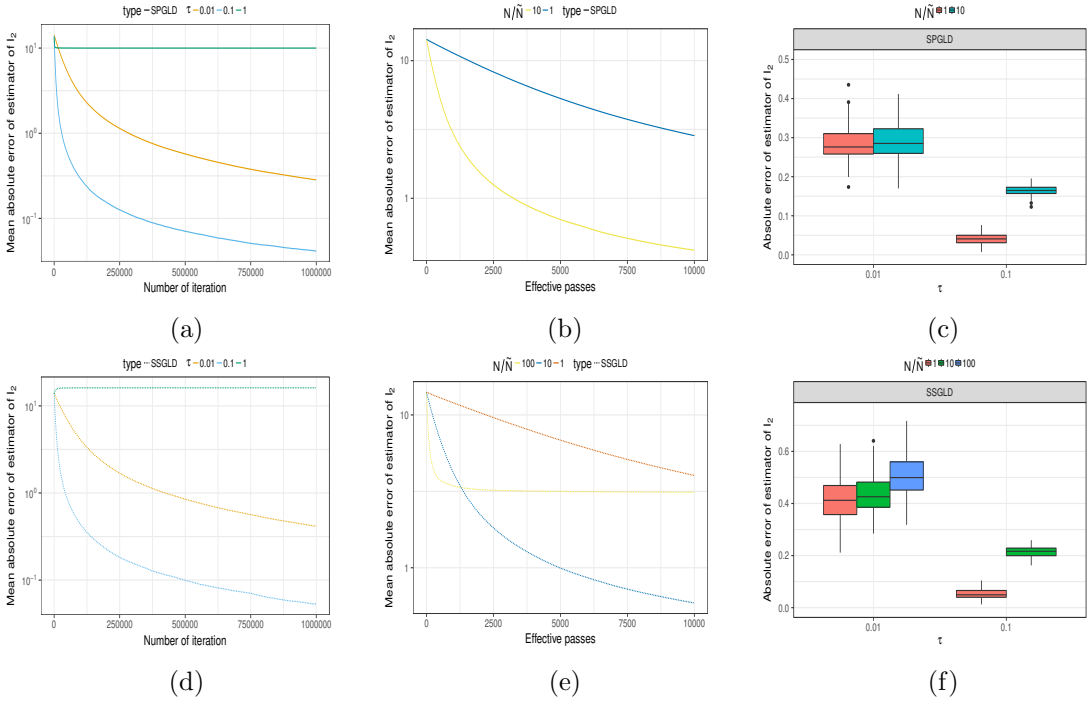

Figure 5: Mean absolute error of estimator of I_2 for Musk dataset: (top row) results for p_{1-2} prior; (a) convergence of SPGLD for $\tilde{N} = N$, (b) convergence of SPGLD in terms of effective passes for $\tau = 0.1$, (c) boxplot of SPGLD for full run; (bottom row) results for p_1 prior; (d) convergence of SSGLD for $\tilde{N} = N$, (e) convergence of SSGLD in terms of effective passes for $\tau = 0.1$, (f) boxplot of SSGLD for full run.

Furthermore our new perspective on the Unadjusted Langevin Algorithm, provides a starting point for the further research into connections between Langevin Monte Carlo and Optimization. In particular, we believe that a very promising direction for future research is to try to modify well-known effective optimization algorithms to minimize the KL divergence with respect to some target density π in Wasserstein space.

7. Postponed Proofs

7.1. Proof of Lemma 1

a) Since e^{-U} is integrable with respect to the Lebesgue measure, under $\mathbf{A}1(m)$ for $m \geq 0$, by (Brazitikos et al., 2014, Lemma 2.2.1), there exists $C_1, C_2 > 0$ such that for all $x \in \mathbb{R}^d$, $U(x) \geq C_1 \|x\| - C_2$. This inequality and $\mathbf{A}2$ implies that $\pi \in \mathcal{P}_2(\mathbb{R}^d)$. In addition, since the function $x \mapsto U(x)\mathrm{e}^{-U(x)/2}$ is bounded on $[-C_2, +\infty)$, we have for all $x \in \mathbb{R}^d$,

$$\left| \left(U(x)\mathrm{e}^{-U(x)/2} \right) \mathrm{e}^{-U(x)/2} \right| \leq C_3 \mathrm{e}^{-U(x)/2}$$

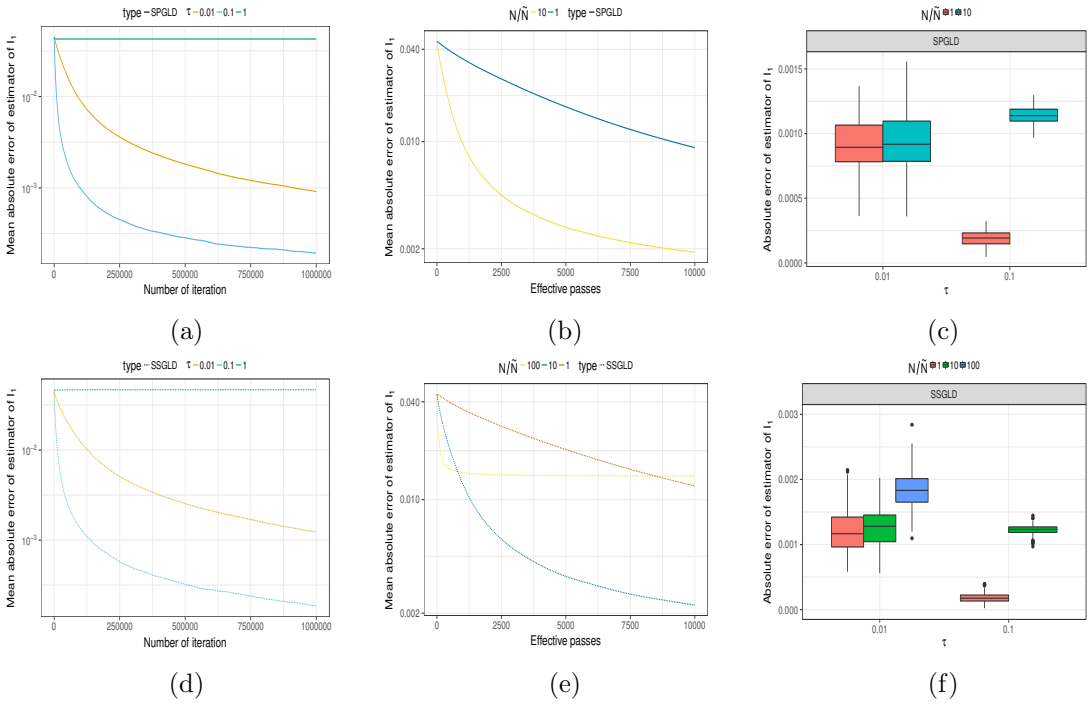

Figure 6: Mean absolute error of estimator of I_1 for Musk dataset: (top row) results for p_{1-2} prior; (a) convergence of SPGLD for $\tilde{N} = N$, (b) convergence of SPGLD in terms of effective passes for $\tau = 0.1$, (c) boxplot of SPGLD for full run; (bottom row) results for p_1 prior; (d) convergence of SSGLD for $\tilde{N} = N$, (e) convergence of SSGLD in terms of effective passes for $\tau = 0.1$, (f) boxplot of SSGLD for full run.

for some constant C_3. From this, and $U(x) \geq C_1 \|x\| - C_2$ we conclude that $\mathscr{E}(\pi) < +\infty$. Using the same reasoning, we have $\mathscr{H}(\pi) < +\infty$ which finishes the proof of the first part. b) First, if μ does not admit a density with respect to Lebesgue measure, then both sides of (10) are $+\infty$. Second, if μ admits a density still denoted by μ with respect to the Lebesgue measure, we have by (7):

$$\mathscr{F}(\mu) - \mathscr{F}(\pi) = \mathrm{KL}\left(\mu|\pi\right) + \int_{\mathbb{R}^d} \{\mu(x) - \pi(x)\} \{U(x) + \log(\pi(x))\} \, \mathrm{d}x = \mathrm{KL}\left(\mu|\pi\right) \ .$$

7.2. Proof of Corollary 8

Using Theorem 6 we first get

$$\mathrm{KL}\left(\nu_n|\pi\right) \leq W_2^2\left(\mu_0, \pi\right)/(2\Gamma_{0,n}) + (Ld/\Gamma_{0,n}) \sum_{k=1}^{n} \gamma_k^2 \ . \tag{40}$$

Note that using a simple integral test, we have $\Gamma_{0,n} \geq C_1 n^{1-\alpha}$ for some constant $C_1 \geq 0$. On the other hand, for some constant $C_2 \geq 0$ we have $\sum_{k=1}^{n} \gamma_k^2 \leq C_2(1 + n^{1-2\alpha})$ if $\alpha \neq 1/2$,

and $\sum_{k=1}^{n} \gamma_k^2 \leq C_2(1+\log(n))$ if $\alpha = 1/2$. Combining all these inequalities in (40) concludes the proof.

7.3. Proofs of Section 4.1

Note that for all $\gamma, \tilde{\gamma} > 0$, $\bar{R}_{\gamma,\tilde{\gamma}}$ can be decomposed as $\bar{S}_\gamma T_{\tilde{\gamma}}$ where $T_{\tilde{\gamma}}$ is defined in (14) and $\bar{S}_\gamma$ is given by (30). Then similarly to the proof of Theorem 6, we first give a preliminary bound on $\mathscr{F}(\mu\bar{R}_{\gamma,\tilde{\gamma}}) - \mathscr{F}(\pi)$ for $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma, \tilde{\gamma} > 0$ as in Proposition 2.

Lemma 26 *Assume* ***A*** *1(0) and* ***A*** *3. For all $\gamma > 0$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$2\gamma \left\{\mathscr{E}(\mu) - \mathscr{E}(\pi)\right\} \leq W_2^2(\mu, \pi) - W_2^2(\mu\bar{S}_\gamma, \pi) + \gamma^2 \left\{M^2 + \upsilon_\Theta(\mu)\right\} ,$$

where $\mathscr{E}$ and T_γ are defined in (9) *and* (14) *respectively, $\upsilon_\Theta(\mu)$ in* (29) *and $\bar{S}_\gamma$ in* (30)*.*

Proof Let Z be a random variable with distribution η, $\gamma > 0$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. For all $x, y \in \mathbb{R}^d$, we have using the definition of $\partial U(x)$ (24) and **A** 3-(ii)

$$\begin{aligned}\|y - x + \gamma\Theta(x, Z)\|^2 &= \|y - x\|^2 + 2\gamma \langle \Theta(x, Z), y - x\rangle + \gamma^2 \|\Theta(x, Z)\|^2 \\ &\leq \|y - x\|^2 - 2\gamma \left\{U(x) - U(y)\right\} + 2\gamma \langle \Theta(x, Z) - \mathbb{E}\left[\Theta(x, Z)\right], y - x\rangle + \gamma^2 \|\Theta(x, Z)\|^2 .\end{aligned}$$

Let (X, Y) be an optimal coupling between μ and π independent of Z. Then by **A** 3-(ii) and rearranging the terms in the previous inequality, we obtain

$$2\gamma \left\{\mathscr{E}(\mu) - \mathscr{E}(\nu)\right\} \leq W_2^2(\mu, \pi) - \mathbb{E}\left[\|Y - X + \gamma\Theta(X, Z)\|^2\right] + \gamma^2 \mathbb{E}\left[\|\Theta(X, Z)\|^2\right] .$$

The proof is concluded upon noting that $W_2^2(\mu\bar{S}_\gamma, \pi) \leq \mathbb{E}[\|Y - X + \gamma\Theta(X, Z)\|^2]$ and $\mathbb{E}[\|\Theta(X, Z)\|^2] \leq M^2 + \upsilon_\Theta(\mu)$. ■

Proposition 27 *Assume* ***A*** *1(0) and* ***A*** *3. For all $\gamma, \tilde{\gamma} > 0$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$2\tilde{\gamma} \left\{\mathscr{F}(\mu\bar{R}_{\gamma,\tilde{\gamma}}) - \mathscr{F}(\pi)\right\} \leq \left\{W_2^2\left(\mu\bar{S}_\gamma, \pi\right) - W_2^2\left(\mu\bar{R}_{\gamma,\tilde{\gamma}}\bar{S}_{\tilde{\gamma}}, \pi\right)\right\} + \tilde{\gamma}^2 \left\{M^2 + \upsilon_\Theta(\mu\bar{R}_{\gamma,\tilde{\gamma}})\right\} .$$

where $\mathscr{F}$ is defined in (9)*, $\upsilon_\Theta(\mu)$ in* (29)*, $\bar{R}_{\gamma,\tilde{\gamma}}$ and $\bar{S}_\gamma$ in* (27) *in* (30) *respectively.*

Proof Note that by Lemma 26, we have

$$2\tilde{\gamma} \left\{\mathscr{E}(\mu\bar{R}_{\gamma,\tilde{\gamma}}) - \mathscr{E}(\pi)\right\} \leq W_2^2\left(\mu\bar{R}_{\gamma,\tilde{\gamma}}, \pi\right) - W_2^2\left(\mu\bar{R}_{\gamma,\tilde{\gamma}}\bar{S}_{\tilde{\gamma}}, \pi\right) + \tilde{\gamma}^2 \left\{M^2 + \upsilon_\Theta(\mu\bar{R}_{\gamma,\tilde{\gamma}})\right\} . \quad (41)$$

In addition by Lemma 5, it holds

$$2\tilde{\gamma} \left\{\mathscr{H}(\mu\bar{R}_{\gamma,\tilde{\gamma}}) - \mathscr{H}(\pi)\right\} \leq W_2^2\left(\mu\bar{S}_\gamma, \pi\right) - W_2^2\left(\mu\bar{R}_{\gamma,\tilde{\gamma}}, \pi\right) .$$

The proof then follows from combining this inequality with (41). ■

7.3.1. Proof of Theorem 13

By Proposition 27, for all $k \in \mathbb{N}^*$, we have

$$\mathscr{F}(\mu\bar{Q}_\gamma^k) - \mathscr{F}(\pi) \leq (2\gamma_{k+1})^{-1} \left\{ W_2^2 \left(\mu\bar{Q}_\gamma^{k-1} \bar{S}_{\gamma_k}, \pi \right) - W_2^2 \left(\mu\bar{Q}_\gamma^k \bar{S}_{\gamma_{k+1}}, \pi \right) \right\} + (\gamma_{k+1}/2) \left\{ M^2 + \upsilon_\Theta(\mu\bar{Q}_\gamma^k) \right\} .$$

Similarly to the proof of Theorem 6 using the convexity of KL divergence and the condition that $(\lambda_k/\gamma_{k+1})_{k\in\mathbb{N}^*}$ is non-increasing concludes the proof.

7.3.2. Proof of Corollary 16

On the one hand, using Theorem 13, we get:

$$\mathrm{KL}\left(\tilde{\nu}_n^N \middle| \pi\right) \leq (2\gamma n)^{-1} W_2^2 \left(\mu_0 \bar{Q}_\gamma^N \bar{S}_\gamma, \pi\right) + \gamma M^2/2 + (\gamma/(2n)) \sum_{k=N+1}^{N+n} \upsilon_\Theta(\mu_0 \bar{Q}_\gamma^k) \ .$$

On the other hand, using Proposition 15, we obtain:

$$2\gamma(\tilde{L}^{-1} - \gamma) \left(\sum_{k=N+1}^{N+n} \upsilon_\Theta(\mu_0 \bar{Q}_\gamma^k) \right) \leq \int_{\mathbb{R}^d} \|x - x^\star\|^2 \, \mathrm{d}\mu_0 \bar{Q}_\gamma^{N+1}(x) - \int_{\mathbb{R}^d} \|x - x^\star\|^2 \, \mathrm{d}\mu_0 \bar{Q}_\gamma^{N+n+1} + 2n\gamma^2 \upsilon_\Theta(\delta_{x^\star}) + 2n\gamma d \ .$$

Combining the two inequalities above finishes the proof of the first part of Corollary 16. For the second part, first observe that since $\gamma_\varepsilon \leq (2\tilde{L})^{-1}$ we have $(2(\tilde{L}^{-1} - \gamma))^{-1} \leq \tilde{L}$. Furthermore, from the definition of γ_ε we have $\gamma_\varepsilon(\frac{M^2}{2} + \tilde{L}d) \leq \varepsilon/4$, as well as $\gamma_\varepsilon^2 \tilde{L} \upsilon_\Theta(\delta_{x^\star}) \leq \varepsilon/4$. On the other hand, from the definition of n_ε we have $W_2^2(\mu_0 \bar{S}_{\gamma_\varepsilon}, \pi)/(2\gamma\varepsilon n_\varepsilon) \leq \varepsilon/4$ as well as $\tilde{L}(2n_\varepsilon)^{-1} \int_{\mathbb{R}^d} \|x - x^\star\|^2 \, \mathrm{d}\mu_0 \bar{R}_{\gamma_\varepsilon, \gamma_\varepsilon}(x) \leq \varepsilon/4$. Combining those four bounds together finishes the proof.

7.4. Proof of Section 4.2

We proceed for the proof of Theorem 17 similarly to the one of Theorem 6, by decomposing $\mathscr{F}(\mu\tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{F}(\pi) = \mathscr{E}(\mu\tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{E}(\pi) + \mathscr{H}(\mu\tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{H}(\pi)$, for $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma, \tilde{\gamma} > 0$. The main difference is that we now need to handle carefully the proximal step in the first term of the decomposition. To this end, we decompose the potential energy functional according to the decomposition of U, $\mathscr{E} = \mathscr{E}_1 + \mathscr{E}_2$ where for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathscr{E}_1(\mu) = \int_{\mathbb{R}^d} U_1 \mathrm{d}\mu(x) \ , \qquad \mathscr{E}_2(\mu) = \int_{\mathbb{R}^d} U_2 \mathrm{d}\mu(x) \ , \tag{42}$$

and consider

$$\mathscr{F}(\mu\tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{F}(\pi) = \mathscr{E}_1(\mu\tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{E}_1(\mu\tilde{S}_\gamma^2\tilde{S}_{\tilde{\gamma}}^1) + \mathscr{E}_1(\mu\tilde{S}_\gamma^2\tilde{S}_{\tilde{\gamma}}^1) - \mathscr{E}_1(\pi) + \mathscr{E}_2(\mu\tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{E}_2(\pi) + \mathscr{H}(\mu\tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{H}(\pi) \ . \tag{43}$$

The first and last terms in the right hand side will be controlled using Lemma 3 and Lemma 5. In the next lemmas, we bound the other terms separately.

Lemma 28 *Assume* **A** 5(m), *for* $m \geq 0$. *For all* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ *and* $\gamma \in (0, L^{-1}]$,

$$2\gamma\{\mathscr{E}_1(\mu\tilde{S}_\gamma^1) - \mathscr{E}_1(\nu)\} \leq (1 - m\gamma)W_2^2(\mu,\nu) - W_2^2(\mu\tilde{S}_\gamma^1, \nu) - \gamma^2(1-\gamma L)\int_{\mathbb{R}^d} \|\nabla U_1(x)\|^2 \,\mathrm{d}\mu(x) + \gamma^2(1+\gamma L)\upsilon_1(\mu) \,,$$

where $\mathscr{E}_1, \tilde{S}_\gamma^1$ *is defined by* (42)-(34) *and* $\upsilon_1(\mu)$ *by* (36).

Proof Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma > 0$. Since U_1 satisfies **A** 2 by (Nesterov, 2004, Lemma 1.2.3), for all $x, \tilde{x} \in \mathbb{R}^d$, we have $|U_1(\tilde{x}) - U_1(x) - \langle \nabla U_1(x), \tilde{x} - x\rangle| \leq (L/2)\|\tilde{x} - x\|^2$. Using that U_1 is m-strongly convex by **A** 5(m), for all $x, y \in \mathbb{R}^d$, $z \in \mathsf{Z}$, we get

$$\begin{aligned} U_1(x - \gamma\tilde{\Theta}_1(x,z)) - U_1(y) &= U_1(x - \gamma\tilde{\Theta}_1(x,z)) - U_1(x) + U_1(x) - U_1(y) \\ &\leq -\gamma\left\langle \nabla U_1(x), \tilde{\Theta}_1(x,z)\right\rangle + (L\gamma^2/2)\left\|\tilde{\Theta}_1(x,z)\right\|^2 + \langle \nabla U_1(x), x - y\rangle - (m/2)\|y - x\|^2 \,. \end{aligned}$$

Then multiplying both sides by γ, we obtain

$$\begin{aligned} 2\gamma\left\{U_1(x - \gamma\tilde{\Theta}_1(x,z)) - U_1(y)\right\} &\leq (1 - m\gamma)\|x - y\|^2 - \left\|x - \gamma\tilde{\Theta}_1(x,z) - y\right\|^2 \\ &- 2\gamma^2\left\langle \nabla U_1(x), \tilde{\Theta}_1(x,z)\right\rangle + \gamma^2(1+\gamma L)\left\|\tilde{\Theta}_1(x,z)\right\|^2 + 2\gamma\left\langle \nabla U_1(x) - \tilde{\Theta}_1(x,z), x - y\right\rangle \,. \end{aligned} \tag{44}$$

Let now (X, Y) be an optimal coupling between μ and ν and Z with distribution η independent of (X, Y). Note that **A** 5 implies that $\mathbb{E}[\tilde{\Theta}_1(X, Z)|(X, Y)] = \nabla U_1(X)$. Then by definition and (44), we get

$$\begin{aligned} 2\gamma\left\{\mathscr{E}(\mu\tilde{S}_\gamma^1) - \mathscr{E}(\nu)\right\} &\leq (1 - m\gamma)W_2^2(\mu,\nu) - \mathbb{E}\left[\left\|X - \gamma\tilde{\Theta}_1(X) - Y\right\|^2\right] \\ &\quad - 2\gamma^2\mathbb{E}\left[\|\nabla U_1(X)\|^2\right] + \gamma^2(1+\gamma L)\mathbb{E}\left[\left\|\tilde{\Theta}_1(X)\right\|^2\right] \\ &\leq (1 - m\gamma)W_2^2(\mu,\nu) - \mathbb{E}\left[\left\|X - \gamma\tilde{\Theta}_1(X) - Y\right\|^2\right] \\ &\quad - \gamma^2(1-\gamma L)\mathbb{E}\left[\|\nabla U_1(X)\|^2\right] + \gamma^2(1+\gamma L)\upsilon_1(\mu) \,. \end{aligned}$$

Using that $W_2^2(\mu\tilde{S}_\gamma^1, \nu) \leq \mathbb{E}[\|X - \gamma\tilde{\Theta}_1(X) - Y\|^2]$ concludes the proof. ∎

Lemma 29 *Assume* **A** 5(m) *for* $m \geq 0$. *For all* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ *and* $\gamma > 0$, *we have*

$$2\gamma\left\{\mathscr{E}_2(\mu) - \mathscr{E}_2(\nu)\right\} \leq W_2^2(\mu,\nu) - W_2^2(\mu\tilde{S}_\gamma^2, \nu) + 2\gamma^2 M_2^2 \,,$$

where $\mathscr{E}_2, \tilde{S}_\gamma^2$ *are defined by* (42) *and* (34) *respectively.*

Proof Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma > 0$. First we bound for any $x, y \in \mathbb{R}^d$, $U_2(x) - U_2(y)$ using the decomposition $U_2(x) - U_2(\mathrm{prox}^{\gamma}_{U_2}(x)) + U_2(\mathrm{prox}^{\gamma}_{U_2}(x)) - U_2(y)$. For any $x, y \in \mathbb{R}^d$, we have using that $\gamma^{-1}(x - \mathrm{prox}^{\gamma}_{U_2}(x)) \in \partial U_2(\mathrm{prox}^{\gamma}_{U_2}(x))$ (see Rockafellar and Wets (1998, Chapter 1 Section G)), where ∂U_2 is the subdifferential of U_2 defined by (24),

$$U_2(\mathrm{prox}^{\gamma}_{U_2}(x)) - U_2(y) \leq \gamma^{-1} \left\langle x - \mathrm{prox}^{\gamma}_{U_2}(x), \mathrm{prox}^{\gamma}_{U_2}(x) - y \right\rangle .$$

Since $\|x - y\|^2 = \|x - \mathrm{prox}^{\gamma}_{U_2}(x)\|^2 + \|\mathrm{prox}^{\gamma}_{U_2}(x) - y\|^2 + 2\langle x - \mathrm{prox}^{\gamma}_{U_2}(x), \mathrm{prox}^{\gamma}_{U_2}(x) - y\rangle$, we get for all $x, y \in \mathbb{R}^d$,

$$U_2(\mathrm{prox}^{\gamma}_{U_2}(x)) - U_2(y) \leq (2\gamma)^{-1}(\|x - y\|^2 - \|\mathrm{prox}^{\gamma}_{U_2}(x) - y\|^2) . \tag{45}$$

Second, since U_2 is M_2-Lipschitz, we get for any $x \in \mathbb{R}^d$, $|U_2(x) - U_2(\mathrm{prox}^{\gamma}_{U_2}(x))| \leq M_2 \|x - \mathrm{prox}^{\gamma}_{U_2}(x)\|$. Then using that $\gamma^{-1}(x - \mathrm{prox}^{\gamma}_{U_2}(x)) \in \partial U_2(\mathrm{prox}^{\gamma}_{U_2}(x))$, and for any $v \in \partial U_2(\mathrm{prox}^{\gamma}_{U_2}(x))$, since U_2 is M_2-Lipschitz, $\|v\| \leq M_2$, we obtain $|U_2(x) - U_2(\mathrm{prox}^{\gamma}_{U_2}(x))| \leq \gamma M_2^2$. Combining this result and (45) yields for any $x, y \in \mathbb{R}^d$

$$2\gamma \{U_2(x) - U_2(y)\} \leq \|x - y\|^2 - \|\mathrm{prox}^{\gamma}_{U_2}(x) - y\|^2 + 2\gamma^2 M_2^2 .$$

Let (X, Y) be an optimal coupling for μ and ν. The proof then follows from using the inequality above for (X, Y), taking the expectation and because $W_2^2(\mu \tilde{S}^1_{\gamma}, \nu) \leq \|\mathrm{prox}^{\gamma}_{U_2}(X) - Y\|^2$. ∎

Lemma 30 *Assume* ***A*** *5(m), for $m \geq 0$. For all $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma, \tilde{\gamma} \in \left(0, L^{-1}\right]$,*

$$2\tilde{\gamma}\{\mathscr{F}(\mu_0 \tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{F}(\pi)\} \leq (1 - m\tilde{\gamma}) W_2^2(\mu_0 \tilde{S}^2_{\gamma}, \pi) - W_2^2(\mu_0 \tilde{R}_{\gamma,\tilde{\gamma}} \tilde{S}^2_{\tilde{\gamma}}, \pi) + \tilde{\gamma}^2 \{2Ld + (1 + \tilde{\gamma}L) \upsilon_1(\mu_0 \tilde{S}^2_{\gamma}) + 2M_2^2\} ,$$

where $\mathscr{F}$, $\tilde{R}_{\gamma,\tilde{\gamma}}$ and $\tilde{S}^2_{\gamma}$ are defined by (7)-(33)-(34) *respectively.*

Proof Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma, \tilde{\gamma} \in \left(0, L^{-1}\right]$. By Lemma 3 and since $\tilde{R}_{\gamma,\tilde{\gamma}} = \tilde{S}^2_{\gamma} \tilde{S}^1_{\tilde{\gamma}} T_{\tilde{\gamma}}$, we have

$$\mathscr{E}_1(\mu_0 \tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{E}_1(\mu_0 \tilde{S}^2_{\gamma} \tilde{S}^1_{\tilde{\gamma}}) \leq 2Ld\tilde{\gamma} . \tag{46}$$

By Lemma 28 since $\tilde{\gamma} \leq 1/L$,

$$2\tilde{\gamma}\{\mathscr{E}_1(\mu_0 \tilde{S}^2_{\gamma} \tilde{S}^1_{\tilde{\gamma}}) - \mathscr{E}_1(\pi)\} \leq (1 - \tilde{\gamma}m) W_2^2(\mu_0 \tilde{S}^2_{\gamma}, \pi) - W_2^2(\mu_0 \tilde{S}^2_{\gamma} \tilde{S}^1_{\tilde{\gamma}}, \pi) + \tilde{\gamma}^2 (1 + \tilde{\gamma}L) \upsilon_1(\mu_0 \tilde{S}^2_{\gamma}) . \tag{47}$$

By Lemma 29, we have

$$2\tilde{\gamma}\{\mathscr{E}_2(\mu_0 \tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{E}_2(\pi)\} \leq W_2^2(\mu_0 \tilde{R}_{\gamma}, \pi) - W_2^2(\mu_0 \tilde{R}_{\gamma,\tilde{\gamma}} \tilde{S}^2_{\tilde{\gamma}}, \pi) + 2\tilde{\gamma}^2 M_2^2 . \tag{48}$$

Finally by Lemma 5, we have

$$2\tilde{\gamma}\{\mathscr{H}(\mu_0 \tilde{R}_{\gamma,\tilde{\gamma}}) - \mathscr{H}(\pi)\} \leq W_2^2(\mu_0 \tilde{S}^2_{\gamma} \tilde{S}^1_{\tilde{\gamma}}, \pi) - W_2^2(\mu_0 \tilde{R}_{\gamma,\tilde{\gamma}}, \pi) . \tag{49}$$

Combining (46)-(47)-(48)-(49) in (43) concludes the proof. ∎

7.4.1. Proof of Theorem 17

Using the convexity of KL divergence and Lemma 30, we obtain

$$\begin{aligned}
\mathrm{KL}\left(\tilde{\nu}_n^N|\pi\right) &\leq \Lambda_{N,N+n}^{-1} \sum_{k=N+1}^{N+n} \lambda_k \,\mathrm{KL}\left(\mu_0 \tilde{Q}_\gamma^k \middle| \pi\right) \\
&\leq (2\Lambda_{N,N+n})^{-1} \left[\frac{(1-m\gamma_{N+2})\lambda_{N+1}}{\gamma_{N+2}} W_2^2\left(\mu_0 \tilde{Q}_\gamma^N \tilde{S}^2_{\gamma_{N+1}}, \pi\right) \right. \\
&\quad - \frac{\lambda_{N+n}}{\gamma_{N+n+1}} W_2^2\left(\mu_0 \tilde{Q}_\gamma^{N+n} \tilde{S}^2_{\gamma_{N+n+1}}, \pi\right) \\
&\quad + \sum_{k=N+1}^{N+n-1} \left\{ \frac{(1-m\gamma_{k+2})\lambda_{k+1}}{\gamma_{k+2}} - \frac{\lambda_k}{\gamma_{k+1}} \right\} W_2^2\left(\mu_0 Q_\gamma^{k-1} \tilde{S}^2_{\gamma_{k+1}}, \pi\right) \\
&\quad \left. + \sum_{k=N+1}^{N+n} \lambda_k \gamma_{k+1} \{2Ld + (1+\gamma_{k+1}L)\upsilon_1(\mu_0 Q_\gamma^k \tilde{S}^2_{\gamma_k}) + 2M_2^2\} \right] .
\end{aligned}$$

We get the thesis using that $\lambda_{k+1}(1-m\gamma_{k+2})/\gamma_{k+2} \leq \lambda_k/\gamma_{k+1}$ for all $k \in \mathbb{N}$.

7.4.2. Proof of Corollary 20

Using Theorem 17 we get:

$$\mathrm{KL}\left(\tilde{\nu}_n^N|\pi\right) \leq W_2^2\left(\mu_0 \tilde{Q}_\gamma^N \tilde{S}^2_\gamma, \pi\right) \Big/ (2\gamma n) + \gamma(Ld + M_2^2) + \frac{\gamma}{2n} \sum_{k=N+1}^{N+n} (1+\gamma L)\upsilon_1(\mu_0 Q_\gamma^k \tilde{S}^2_\gamma)$$

and using Proposition 19 we obtain:

$$\begin{aligned}
2\gamma(\tilde{L}^{-1} - \gamma) \left(\sum_{k=N+1}^{N+n} \upsilon_1(\mu_0 Q_\gamma^k \tilde{S}^2_\gamma) \right) &\leq \int_{\mathbb{R}^d} \|y - x^\star\|^2 \,\mathrm{d}\mu_0 Q_\gamma^{N+1} \tilde{S}^2_\gamma(y) \\
&\quad - \int_{\mathbb{R}^d} \|y - x^\star\|^2 \,\mathrm{d}\mu_0 Q_\gamma^{N+n+1} \tilde{S}^2_\gamma(y) + 2n\gamma^2 \upsilon_1(\delta_{x^\star}) + 2n\gamma d \,,
\end{aligned}$$

Combining the two inequalities above finishes the proof of the first part of Corollary 20. For the second part, observe that since $\gamma_\varepsilon \leq L^{-1}$ and $\gamma_\varepsilon \leq (2\tilde{L})^{-1}$ we have $(1+\gamma L)(2(\tilde{L}^{-1} - \gamma))^{-1} \leq 2\tilde{L}$. Therefore from definition of γ_ε we have $\gamma_\varepsilon(Ld + M_2^2 + 2\tilde{L}d) \leq \varepsilon/4$, as well as $\gamma_\varepsilon^2 2\tilde{L}\upsilon_1(\delta_{x^\star}) \leq \varepsilon/4$. On the other hand, from definition of n_ε we have $W_2^2(\mu_0 \tilde{S}^2_{\gamma_\varepsilon}, \pi)/(2n_\varepsilon\gamma_\varepsilon) \leq \varepsilon/4$ as well as $2\tilde{L}(2n_\varepsilon)^{-1} \int_{\mathbb{R}^d} \|x - x^\star\|^2 \,\mathrm{d}\mu_0 \tilde{S}^2_\gamma(y) \leq \varepsilon/4$. Combining this four bounds we get the thesis.

7.4.3. Proof of Theorem 21

Using Lemma 30 and since the KL divergence is non-negative, we get for all $k \in \{1, \ldots, n\}$,

$$\begin{aligned}
W_2^2\left(\mu_0 \tilde{Q}_\gamma^k \tilde{S}^2_{\gamma_{k+1}}, \pi\right) &\leq (1 - m\gamma_{k+1}) W_2^2\left(\mu_0 Q_\gamma^{k-1} \tilde{S}^2_{\gamma_k}, \pi\right) \\
&\quad + \gamma_{k+1}^2 \{2Ld + (1+\gamma_{k+1}L)\upsilon_1(\mu_0 Q_\gamma^{k-1} \tilde{S}^2_{\gamma_k}) + 2M_2^2\} \,.
\end{aligned}$$

The proof then follows from a direct induction.

7.4.4. Proof of Proposition 23

Let $\gamma > 0$, $x \in \mathbb{R}^d$ and consider $\tilde{X}_1 = \mathrm{prox}_{U_2}^{\gamma}\left\{x - \gamma\tilde{\Theta}_1(x, Z_1) + \sqrt{2\gamma}G_1\right\}$, where Z_1 and G_1 are two independent random variables, Z_1 has distribution $\tilde{\eta}_1$ and G_1 is a standard Gaussian random variable, so that $\tilde{X}_1$ has distribution $\tilde{S}_\gamma^1 T_\gamma \tilde{S}_\gamma^2(x, \cdot)$. First by (Bauschke and Combettes, 2011, Theorem 26.2(vii)), we have that $x^\star = \mathrm{prox}_{U_2}^{\gamma}(x^\star - \gamma\nabla U_1(x^\star))$ and by (Bauschke and Combettes, 2011, Proposition 12.27), the proximal is non-expansive, for all $x, y \in \mathbb{R}^d$, $\|\mathrm{prox}_{U_2}^{\gamma}(x) - \mathrm{prox}_{U_2}^{\gamma}(y)\| \leq \|x - y\|$. Using these two results and the fact that $\tilde{\Theta}_1$ satisfies **A**4, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\tilde{X}_1 - x^\star\right\|^2\right] &= \mathbb{E}\left[\left\|\mathrm{prox}_{U_2}^{\gamma}\left\{x - \gamma\tilde{\Theta}_1(x, Z_1) + \sqrt{2\gamma}G_1\right\} - \mathrm{prox}_{U_2}^{\gamma}\{x^\star - \gamma\nabla U_1(x^\star)\}\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\left(x - \gamma\tilde{\Theta}_1(x, Z_1) + \sqrt{2\gamma}G_1\right) - (x^\star - \gamma\nabla U_1(x^\star))\right\|^2\right] \\
&\leq \|x - x^\star\|^2 \\
&\quad + \mathbb{E}\left[2\gamma\left\langle x - x^\star, \nabla U_1(x^\star) - \tilde{\Theta}_1(x, Z_1)\right\rangle + \gamma^2\left\|\nabla U_1(x^\star) - \tilde{\Theta}_1(x, Z_1)\right\|^2\right] + 2\gamma d \\
&\leq (1 - \tilde{m}_1\gamma)\|x - x^\star\|^2 - 2\gamma(\tilde{L}_1^{-1} - \gamma)\mathbb{E}\left[\left\|\tilde{\Theta}_1(x, Z_1) - \tilde{\Theta}_1(x^\star, Z_1)\right\|^2\right] \\
&\qquad + 2\gamma^2\mathbb{E}\left[\left\|\tilde{\Theta}_1(x^\star, Z_1) - \nabla U_1(x^\star)\right\|^2\right] + 2\gamma d\,.
\end{aligned}
$$

The proof is completed upon noting that $\upsilon_1(\delta_x) \leq \mathbb{E}[\|\Theta(x, Z_1) - \Theta(x^\star, Z_1)\|^2]$.

7.4.5. Proof of Corollary 24

Using Theorem 21 we get:

$$
\begin{aligned}
W_2^2(\mu_0\tilde{R}_{\gamma,\gamma}^n\tilde{S}_\gamma^2, \pi) &\leq (1 - m\gamma)^n W_2^2(\mu_0\tilde{S}_\gamma^2, \pi) \\
&\quad + \gamma^2\sum_{k=1}^{n}(1 - m\gamma)^{n-k}\left(2Ld + (1 + \gamma L)\upsilon_1(\mu_0\tilde{R}_{\gamma,\gamma}^k\tilde{S}_\gamma^2) + 2M_2^2\right) \\
&\leq (1 - m\gamma)^n W_2^2(\mu_0\tilde{S}_\gamma^2, \pi) + 2(Ld + M_2)\gamma/m \\
&\quad + \gamma^2\sum_{k=1}^{n}(1 - \tilde{m}\gamma)^{n-k}(1 + \gamma L)\upsilon_1(\mu_0\tilde{R}_{\gamma,\gamma}^k\tilde{S}_\gamma^2)\,. \qquad (50)
\end{aligned}
$$

In addition, using Proposition 23 and $\gamma \leq (2\tilde{L}_1)^{-1}$, we have

$$
\begin{aligned}
\gamma\tilde{L}_1^{-1}\sum_{k=1}^{n}(1 - \tilde{m}\gamma)^{n-k}\upsilon_1(\mu_0\tilde{R}_{\gamma,\gamma}^k\tilde{S}_\gamma^2) &\leq 2\gamma\sum_{k=1}^{n}(1 - \tilde{m}\gamma)^{n-k}(\gamma\upsilon_1(\delta_{x^\star}) + d) \\
&\quad + \sum_{k=1}^{n}(1 - \tilde{m}\gamma)^{n-k+1}\int_{\mathbb{R}^d}\|x - x^\star\|^2\,\mathrm{d}\mu_0 R_{\gamma,\gamma}^k\tilde{S}_\gamma^2(x) \\
&\quad - \sum_{k=1}^{n}(1 - \tilde{m}\gamma)^{n-k}\int_{\mathbb{R}^d}\|x - x^\star\|^2\,\mathrm{d}\mu_0 R_{\gamma,\gamma}^{k+1}\tilde{S}_\gamma^2(x)\,.
\end{aligned}
$$

Combining this result and (50) concludes the proof of (38).

Now, for $\gamma_\varepsilon, n_\varepsilon$ as defined in the thesis of the corollary we have $\gamma_\varepsilon \Delta_1 \leq \varepsilon/4$ and $\Delta_2 \gamma_\varepsilon^2 \leq \varepsilon/4$. Furthermore, $(1 - m\gamma_\varepsilon)^{n_\varepsilon} W_2^2(\mu_0 \tilde{S}_{\gamma_\varepsilon}^2, \pi) \leq \exp(-n_\varepsilon m \gamma_\varepsilon) W_2^2(\mu_0 \tilde{S}_{\gamma_\varepsilon}^2, \pi) \leq \varepsilon/4$, and $(1 - \gamma_\varepsilon \tilde{m})\Delta_3 \leq \varepsilon/4$ similarly. Together, the above inequalities conclude the proof.

Acknowledgments

A. D. acknowledges support from Chaire BayeScale "P. Laffitte". B. M. is supported by Polish National Science Center grant no. 2015/17/D/ST1/01198.

Appendix A. Definitions and Useful Results from Theory of Gradient Flows

For a continously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, the gradient flow associated with f starting at x_0 is the solution $(x(t))_{t \in \mathbb{R}_+}$ of the ordinary differential equation: $\mathrm{d}x(t)/\mathrm{d}t = -\nabla f(x(t))$, with $x(0) = x_0$. Classical theory of gradient flows was developed for functions defined on $\mathbb{R}^d$, and later extended to functionals on Banach spaces. Main motivation for this development were connections between gradient flows in Banach spaces and some partial differential equation - for example the heat equation can be formulated as the gradient flow of $u \mapsto \int_{\mathbb{R}^d} \|\nabla u\|^2 \, \mathrm{d}x$ (which is called the Dirichlet energy) on $\mathrm{L}^2(\mathbb{R}^d) = \{u : \mathbb{R}^d \to \mathbb{R} : u \text{ is measurable and } \int_{\mathbb{R}^d} \|u\|^2 \, \mathrm{d}x < +\infty\}$. Similarly, widespread interest in the theory of gradient flows in metric spaces started with the work of Jordan et al. (1998), which showed that gradient flow of the free energy functional defined in (7) in the space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a measure valued solution of the Fokker-Planck equation.

For a brief overview of the theory of gradient flows in Euclidean and metric spaces, with a focus on Wasserstein spaces, we refer to (Santambrogio, 2017). For a detailed introduction to the theory of gradient flows in metric spaces we refer the reader to (Ambrosio et al., 2008). Below, we only introduce definitions and results from the theory of gradient flows in the space of probability measures, which are relevant to our work.

Let $I \subset \mathbb{R}$ be an open interval of $\mathbb{R}$ and $(\mu_t)_{t \in I}$ be a curve on $\mathcal{P}_2(\mathbb{R}^d)$, *i.e.* a family of probability measures belonging to $\mathcal{P}_2(\mathbb{R}^d)$. $(\mu_t)_{t \in I}$ is said to be absolutely continuous if there exists $\ell \in \mathrm{L}^1(I)$ such that for all $s, t \in I$, $s \leq t$, $W_2(\mu_s, \mu_t) \leq \int_s^t |\ell|(u) \mathrm{d}u$. Denote by $\mathrm{AC}(I)$ the set of absolutely continuous curves on I and

$$\mathrm{AC}_{\mathrm{loc}}(\mathbb{R}_+^*) = \left\{ (\mu_t)_{t \geq 0} \ : (\mu_t)_{t \in I} \in \mathrm{AC}(I) \text{ for any open interval } I \subset \mathbb{R}_+^* \right\} .$$

Note that if $(\mu_t)_{t \in I} \in \mathrm{AC}(I)$, then for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, $t \mapsto W_2(\nu, \mu_t)$ is absolutely continuous on I (as a curve from I to $\mathbb{R}_+$). Therefore by (Nielsen, 1997, Theorem 20.8) and (Mitrovic and Zubrinic, 1997, Exercise 4, p.45), $t \mapsto W_2(\nu, \mu_t)$ has derivative for the almost all $t \in I$ and there exists $\delta : I \to \mathbb{R}$ satisfying

$$\int_I |\delta|(u) \mathrm{d}u < +\infty \text{ and } W_2^2(\nu, \mu_t) - W_2^2(\nu, \mu_s) = \int_s^t \delta(u) \mathrm{d}u \ , \text{ for all } s, t \in I \tag{51}$$

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. A constant speed geodesic $(\lambda_t)_{t \in [0,1]}$ between μ and ν is a curve in $\mathcal{P}_2(\mathbb{R}^d)$ such that $\lambda_0 = \mu$, $\lambda_1 = \nu$ and for all for all $s, t \in [0, 1]$, $W_2(\lambda_s, \lambda_t) = |t - s| \, W_2(\mu, \nu)$.

Note that by the triangle inequality, this definition is equivalent to the following: for all $s, t \in [0,1]$, $W_2(\lambda_s, \lambda_t) \leq |t-s|\, W_2(\mu,\nu)$. Indeed by the triangle inequality and the assumption $W_2(\lambda_s,\lambda_t) \leq |t-s|\, W_2(\mu,\nu)$, we have for all $s,t \in [0,1]$, $s<t$,

$$W_2(\mu,\nu) \leq W_2(\mu,\lambda_t) + W_2(\lambda_t,\lambda_s) + W_2(\lambda_s,\nu) \leq W_2(\mu,\nu) \; .$$

Therefore the first inequality is in fact an equality, and therefore using again the assumption for $W_2(\mu,\lambda_t)$ and $W_2(\lambda_s,\nu)$ concludes the proof. By definition of the Wasserstein distance of order 2, a constant speed geodesic $(\lambda_t)_{t\in[0,1]}$ between μ and ν is given for all $t \in [0,1]$ by $\lambda_t = (t\,\mathrm{proj}_1 + (1-t)\,\mathrm{proj}_2)_\sharp \zeta$ where ζ is an optimal transport plan between μ and ν and $\mathrm{proj}_1, \mathrm{proj}_2 : \mathbb{R}^{2d} \to \mathbb{R}^d$ are the projections on the first and the last d components respectively.

Let $\mathscr{S} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$. The functional $\mathscr{S}$ is said to be lower semi-continuous if for all $M \in \mathbb{R}$, $\{\mathscr{S} \leq M\}$ is a closed set of $\mathcal{P}_2(\mathbb{R}^d)$ and m-geodesically convex for $m \geq 0$ if for any $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d)$ there exists a constant speed geodesic $(\lambda_t)_{t\in[0,1]}$ between μ and ν such that for all $t \in [0,1]$

$$\mathscr{S}(\lambda_t) \leq t\mathscr{S}(\mu) + (1-t)\mathscr{S}(\nu) - t(1-t)(m/2)W_2^2(\mu,\nu) \; .$$

If $m = 0$, $\mathscr{S}$ will be simply said geodesically convex.

A curve $(\mu_t)_{t>0} \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R}_+^*)$ is said to be a gradient flow for the lower semi-continuous and m-geodesically convex function $\mathscr{S} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ if for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathscr{S}(\nu) < +\infty$, and for almost all $t \in \mathbb{R}_+^*$,

$$(1/2)\delta_t + (m/2)W_2^2(\mu_t,\nu) \leq \mathscr{S}(\nu) - \mathscr{S}(\mu_t) \; ,$$

where $\delta : \mathbb{R}_+^* \to \mathbb{R}$ satisfies (51) for all open interval of $\mathbb{R}_+^*$. We say that $(\mu_t)_{t\in\mathbb{R}_+^*}$ starts at μ if $\lim_{t\to 0} W_2(\mu_t,\mu) = 0$ and then set $\mu_0 = \mu$. By (Ambrosio et al., 2008, Theorem 11.1.4), there exists at most one gradient flow associated with $\mathscr{S}$.

Consider the functional $\tilde{\mathscr{F}} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ given by $\tilde{\mathscr{F}} = \mathscr{H} + \tilde{\mathscr{E}}$ where $\mathscr{H}$ is defined by (8) and $\tilde{\mathscr{E}}$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$\tilde{\mathscr{E}}(\mu) = \int_{\mathbb{R}^d} V(x) \mathrm{d}\mu(x) \; ,$$

where $V : \mathbb{R}^d \to (-\infty, +\infty]$ is a convex lower-semicontinuous function (for all $M \geq 0$, $\{V \leq M\}$ is closed subset of $\mathbb{R}^d$) with $\{V < +\infty\} \neq \emptyset$ and the interior of this set is non empty as well. By (Ambrosio et al., 2008, Proposition 9.3.2, Theorem 9.4.12), $\tilde{\mathscr{F}}$ is geodesically convex and (Ambrosio et al., 2008, Theorem 11.2.8,Theorem 11.1.4) shows that there exists the unique gradient flow $(\mu_t)_{t\geq 0}$ starting at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and this curve is the unique solution of the Fokker-Plank equation (in the sense of distributions) :

$$\frac{\partial \mu_t}{\partial t} = \mathrm{div}(\nabla \mu_t^x + \mu_t^x \nabla V(x)) \; ,$$

i.e. for all $\phi \in C_c^\infty(\mathbb{R}^d)$ and $t > 0$,

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^d} \phi(y) \mu_t(\mathrm{d}y) = \int_{\mathbb{R}^d} \mathcal{A}\phi(y)\, \mu_t(\mathrm{d}y) \; .$$

In addition for all $t > 0$, μ_t is absolutely continuous with respect to the Lebesgue measure. In particular for $V = 0$, we get the following result.

Theorem 31 *For all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a unique solution of the Fokker-Plank equation (in the sense of distributions) :*

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t \ .$$

In addition $(\mu_t)_{t\geq 0} \in \mathrm{AC}(\mathbb{R}_+^*)$ and satisfies for almost all $t \in \mathbb{R}_+^*$,

$$\delta_t/2 \leq \mathscr{H}(\nu) - \mathscr{H}(\mu_t) \ ,$$

where δ_t is given in (51).

Appendix B. On the Second Order Moment of Log-Concave Measures

A 7 *There exist $\eta > 0$, $M_\eta \geq 0$ such that for all $x \in \mathbb{R}^d$, $x \notin \mathrm{B}(0, M_\eta)$,*

$$U(x) - U(x^\star) \geq \eta \left\| x - x^\star \right\| \ .$$

In this section, we give some bounds on to deal with the distance of the initial condition of the algorithms from π in W_2.

Proposition 32 *Assume **A** 1(0) and **A** 7. Then, we have*

$$\int_{\mathbb{R}^d} \left\| x - x^\star \right\|^2 \mathrm{d}\pi(x) \leq 2\eta^{-2} d(1+d) + M_\eta^2 \ .$$

Proof Note that under **A**7, we have

$$\begin{aligned} &\int_{\mathbb{R}^d} \left\| x - x^\star \right\|^2 \mathrm{d}\pi(x) \leq \eta^{-2} \int_{\mathbb{R}^d} \left| U(x) - U(x^\star) \right|^2 \mathrm{d}\pi(x) + M_\eta^2 \\ &\leq 2\eta^{-2} \int_{\mathbb{R}^d} \left| U(x) + \log(\mathrm{Z}) + \mathscr{H}(\pi) \right|^2 \mathrm{d}\pi(x) + 2\eta^{-2} \left| -\mathscr{H}(\pi) - \log(\mathrm{Z}) - U(x^\star) \right|^2 + M_\eta^2 \ . \end{aligned} \tag{52}$$

where $\mathscr{H}$ is defined by (8) and $\mathrm{Z} = \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y$. Then, by (Bobkov and Madiman, 2011, Proposition I.2), $\left| -\mathscr{H}(\pi) - \log(\mathrm{Z}) - U(x^\star) \right| \leq d$ and by (Fradelizi et al., 2016, Theorem 2.3), (see also Nguyen (2013) and Wang (2014)), $\int_{\mathbb{R}^d} \left| U(x) + \log(\mathrm{Z}) + \mathscr{H}(\pi) \right|^2 \mathrm{d}\pi(x) \leq d$. Combining these two results in (52) concludes the proof.

■

References

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures (Lectures in Mathematics. ETH Zürich)*. Birkhäuser, 2008. ISBN 978-3-7643-8721-1.

L. Ambrosio, G. Savaré, and L. Zambotti. Existence and stability for Fokker–Planck equations with log-concave reference measure. *Probability Theory and Related Fields*, 145(3): 517–564, 2009. doi: 10.1007/s00440-008-0177-3.

C. Andrieu, N. De Freitas, A. Doucet, and M. I Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50(1/2):5–43, 2003. doi: 10.1023/a:1020281327116.

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 1st edition, 2011. ISBN 1441994661, 9781441994660.

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

E. Bernton. Langevin Monte Carlo and JKO splitting. *arXiv preprint arXiv:1802.08671*, 2018.

S. Bobkov and M. Madiman. The Entropy Per Coordinate of a Random Vector is Highly Constrained Under Convexity Conditions. *IEEE Transactions on Information Theory*, 57(8):4940–4954, 2011.

F. Bolley, I. Gentil, and A. Guillin. Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. *Journal of Functional Analysis*, 263(8):2430–2457, 2012. doi: 10.1016/j.jfa.2012.07.007.

S. Brazitikos, A. Giannopoulos, P. Valettas, and B.-H. Vritsiou. *Geometry of isotropic convex bodies*, volume 196. American Mathematical Society Providence, 2014.

S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. Handbook of Markov chain Monte Carlo, 2011.

X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. ISBN 978-0-471-24195-9; 0-471-24195-4.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. doi: 10.1111/rssb.12183.

A. S Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

D. Dua and K.T. Efi. UCI Machine Learning Repository, 2017. URL `http://archive.ics.uci.edu/ml`.

A. Durmus and É. Moulines. Supplement to "High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm", 2015. https://hal.inria.fr/hal-01176084/.

A. Durmus and É. Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. 2016.

A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 06 2017. doi: 10.1214/16-AAP1238.

A. Durmus, É. Moulines, and M. Pereyra. Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018. doi: 10.1137/16m1108340.

S.N. Ethier and T.G. Kurtz. *Markov processes: characterization and convergence*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 1986. URL `http://books.google.fr/books?id=BAWnAAAAIAAJ`.

M. Fradelizi, M. Madiman, and L. Wang. Optimal concentration of information content for log-concave densities. In *High dimensional probability VII*, pages 45–60. Springer, 2016.

A. Gelman, J. B Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

A. Genkin, D. D Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

R. B. Gramacy and N. G. Polson. Simulation-based Regularized Logistic Regression. *Bayesian Analysis*, 7(3):567–590, 09 2012. doi: 10.1214/12-BA719.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi: 10.1093/biomet/57.1.97.

C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 03 2006. doi: 10.1214/06-BA105.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

I. Karatzas and S.E. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer New York, 1991. ISBN 9780387976556.

W. Krauth. *Statistical mechanics: algorithms and computations*, volume 13. OUP Oxford, 2006.

D. Lamberton and G. Pagès. Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift. *Stochastics and Dynamics*, 3(4): 435–451, 2003. doi: 10.1142/S0219493703000838.

V. Lemaire. *Estimation de la mesure invariante d'un processus de diffusion*. PhD thesis, Université Paris-Est, 2005.

Q. Li and N. Lin. The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010. doi: 10.1214/10-BA506.

B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Recherche Opérationnelle*, 4:154–158, 1970.

J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002. doi: 10.1016/S0304-4149(02)00150-3.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 23(2): 1087–1092, 1953.

S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 0521731828, 9780521731829.

D. Mitrovic and D. Zubrinic. *Fundamentals of applied functional analysis*, volume 91. CRC Press, 1997.

T. Nagapetyan, A. B Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. C. Zygalakis. The true cost of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004. ISBN 9781402075537.

V.H. Nguyen. *Inégalités Fonctionnelles Et Convexité*. 2013.

O.A. Nielsen. *An Introduction to Integration and Measure Theory*. Wiley-Interscience and Canadian Mathematics Series of Monographs and Texts. Wiley, 1997. ISBN 9780471595182.

N. Parikh and S. Boyd. *Proximal Algorithms*. Foundations and Trends(r) in Optimization. Now Publishers, 2013. ISBN 9781601987167.

G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180:378–384, 1981.

M. Y. Park and T. Hastie. L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 69(4):659–677, 2007. doi: 10.1111/j.1467-9868.2007.00607.x.

M. Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, pages 1–16, 2015. doi: 10.1007/s11222-015-9567-4.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60 (1):255–268, feb 1998. doi: 10.1111/1467-9868.00123.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. doi: 10.2307/3318418.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998. ISBN 3-540-62772-3. doi: 10.1007/978-3-642-02431-3.

T. Rockafeller. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization*, 14:877–898, 1976.

P. J. Rossky, J. D. Doll, and H. L. Friedman. Brownian dynamics as smart monte carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, nov 1978. doi: 10.1063/1.436415.

F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, mar 2017. doi: 10.1007/s13373-017-0101-1.

A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. ISSN 0962-4929. doi: 10.1017/S0962492910000061.

D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509 (1991), 1990. doi: 10.1080/07362999008809220.

T. van Erven and P. Harremos. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3.

S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):5504–5548, 2016.

L. Wang. *Heat capacity bound, energy fluctuations and convexity*. Yale University, 2014.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, pages 681–688, 2011.

A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. *arXiv preprint arXiv:1802.08089*, 2018.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *arXiv preprint arXiv:1707.06618*, 2017.

D. Zhu and P. Marcotte. New classes of generalized monotonicity. *Journal of Optimization Theory and Applications*, 87(2):457–471, nov 1995. doi: 10.1007/bf02192574.