Journal of Machine Learning Research 23 (2022) 1-38 Submitted 6/20; Revised 11/21; Published 9/22

Bounding the Error of Discretized Langevin Algorithms for Non-Strongly Log-Concave Targets

Arnak S. Dalalyan ARNAK.DALALYAN@ENSAE.FR
ENSAE, CREST, IP Paris
5 av. Le Chatelier, 91120 Palaiseau, France

Avetik Karagulyan AVETIK.KARAGULYAN@KAUST.EDU.SA
King Abdullah University of Science and Technology
Thuwal 23955, Saudi Arabia

Lionel Riou-Durand LIONEL.RIOU-DURAND@WARWICK.AC.UK
University of Warwick, Coventry CV4 7AL, UK

Editor: David Wipf

Abstract

In this paper, we provide non-asymptotic upper bounds on the error of sampling from a target density over $\mathbb{R}^p$ using three schemes of discretized Langevin diffusions. The first scheme is the Langevin Monte Carlo (LMC) algorithm, the Euler discretization of the Langevin diffusion. The second and the third schemes are, respectively, the kinetic Langevin Monte Carlo (KLMC) for differentiable potentials and the kinetic Langevin Monte Carlo for twice-differentiable potentials (KLMC2). The main focus is on the target densities that are smooth and log-concave on $\mathbb{R}^p$, but not necessarily strongly log-concave. Bounds on the computational complexity are obtained under two types of smoothness assumption: the potential has a Lipschitz-continuous gradient and the potential has a Lipschitz-continuous Hessian matrix. The error of sampling is measured by Wasserstein-q distances. We advocate for the use of a new dimension-adapted scaling in the definition of the computational complexity, when Wasserstein-q distances are considered. The obtained results show that the number of iterations to achieve a scaled-error smaller than a prescribed value depends only polynomially in the dimension.

Keywords: Approximate sampling, log-concave distributions, Langevin Monte Carlo

Contents

1. Introduction

The two most popular techniques for defining estimators or predictors in statistics and machine learning are the M estimation, also known as empirical risk minimization, and the Bayesian method (leading to posterior mean, median, etc.). In practice, it is necessary to devise a numerical method for computing an approximation of these estimators. Optimization algorithms are used for approximating M-estimators, while Monte Carlo algorithms are employed for approximating Bayesian estimators. In statistical learning theory, over past decades, a concentrated effort was made for getting non asymptotic guarantees on the error of an optimization algorithm. For smooth optimization, sharp results were obtained in the strongly convex and convex cases (see Bubeck, 2015), the case of non-convex smooth optimization being much more delicate (see Jain and Kar, 2017). As for Monte Carlo algorithms, past three years or so witnessed considerable progress on theory of sampling from strongly log-concave densities. Some results for (non strongly) log-concave densities were obtained as well. However, to the best of our knowledge, there is no paper providing a systematic account on the error bounds for sampling from (non strongly) log-concave densities over unbounded domains. The main goal of this paper is to fill this gap.

A good starting point for accomplishing the aforementioned task is perhaps a result from (Durmus et al., 2019) for the sampling error measured by the Kullback-Leibler divergence. The result is established for the Langevin Monte Carlo (LMC) algorithm, which is the "sampling analogue" of the gradient descent. Let $\pi : \mathbb{R}^p \to [0, +\infty)$ be a probability density function (with respect to Lebesgue's measure) given for a potential function f, by

$$\pi(\boldsymbol{\theta}) = \frac{e^{-f(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f(\boldsymbol{v})} d\boldsymbol{v}}.$$

The goal of sampling is to generate a random vector in $\mathbb{R}^p$ having a distribution close to the target distribution defined by π. In the sequel, we make repeated use of the moments $\mu_k(\pi) = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta}\|_2^k]^{1/k}$, where $\|\boldsymbol{v}\|_q = (\sum_j |v_j|^q)^{1/q}$ is the usual ℓ_q-norm for any $q \geq 1$. When there is no risk of confusion, we write μ_k instead of $\mu_k(\pi)$.

To define the LMC algorithm, we need a sequence of positive parameters $\boldsymbol{h} = \{h_k\}_{k\in\mathbb{N}}$, referred to as the step-sizes and an initial point $\boldsymbol{\vartheta}_{0,\boldsymbol{h}} \in \mathbb{R}^p$ that may be deterministic or random. The successive iterations of the LMC algorithm are given by the update rule

$$\boldsymbol{\vartheta}_{k+1,\boldsymbol{h}} = \boldsymbol{\vartheta}_{k,\boldsymbol{h}} - h_{k+1}\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}}) + \sqrt{2h_{k+1}}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{1}$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots$ is a sequence of independent, and independent of $\boldsymbol{\vartheta}_{0,\boldsymbol{h}}$, centered Gaussian vectors with identity covariance matrices. Let ν_K denote the distribution of the K-th iterate of the LMC algorithm, assuming that all the step-sizes are equal ($h_k = h$ for every $k \in \mathbb{N}$) and the initial point is $\boldsymbol{\vartheta}_{0,h} = \mathbf{0}_p$. We will also define the distribution $\bar\nu_K = (1/K)\sum_{k=1}^K \nu_k$, obtained by choosing uniformly at random one of the elements of the sequence $\{\boldsymbol{\vartheta}_{1,\boldsymbol{h}}, \ldots, \boldsymbol{\vartheta}_{K,\boldsymbol{h}}\}$. It is proved in (Durmus et al., 2019, Cor. 7) that if the gradient ∇f is Lipschitz continuous with the Lipschitz constant M, then for every $K \in \mathbb{N}$, the Kullback-Leibler divergence between $\bar\nu_K$ and π satisfies

$$D_{\rm KL}(\bar\nu_K\|\pi) \leq \frac{\mu_2^2(\pi)}{2Kh} + Mph, \qquad D_{\rm KL}(\bar\nu_K^{\rm opt}\|\pi) \leq \sqrt{\frac{2Mp}{K}}\,\mu_2(\pi).$$

Note that the second inequality above is derived from the first one by using the step-size $h_{\rm opt} = \mu_2(\pi)/\sqrt{2KMp}$ obtained by minimizing the right hand side of the first inequality. Therefore, if we assume that the second-order moment μ_2 of π satisfies the condition $M\mu_2^2 \leq \kappa p^\beta$, for some dimension-free positive constants β and κ, we get

$$D_{\rm KL}(\bar\nu_K^{\rm opt}\|\pi) \leq \sqrt{\frac{2\kappa p^{1+\beta}}{K}}.$$

A natural measure of complexity of the LMC with averaging is, for every $\varepsilon > 0$, the number of gradient evaluations that is sufficient for getting a sampling error bounded from above by ε. From the last display, taking into account the Pinsker inequality, $d_{\rm TV}(\bar\nu_K, \pi) \leq \sqrt{D_{\rm KL}(\bar\nu_K, \pi)/2}$, and the fact that each iterate of the LMC requires one evaluation of the gradient of f, we obtain the following result. The number of gradient evaluations $K_{\rm LMCa,TV}(p, \varepsilon)$ sufficient for the total-variation-error of the LMC with averaging (hereafter, LMCa) to be smaller than ε is

$$K_{\rm LMCa,TV}(p, \varepsilon) = \frac{\kappa p^{1+\beta}}{2\varepsilon^4}.$$

The main goal of the present work is to provide this type of bounds on the complexity of various versions of the Langevin algorithm under different measures of the quality of sampling. The most important feature that we wish to uncover is the explicit dependence of the complexity $K(\varepsilon)$ on the dimension p, the inverse-target-precision $1/\varepsilon$ and the condition number κ. We will focus only on those measures of quality of sampling that can be directly used for evaluating the quality of approximating expectations.

The three main contributions of this paper are:

- Bounds on the sampling error measured in the Wasserstein W_q-distance (for $q \in [1, 2]$) in two settings: (a) convex and gradient-Lipschitz potential and (b) convex, gradient-Lipschitz and Hessian-Lipschitz potential. Tight bounds are established for the strong-convexified versions of three algorithms: Langevin Monte Carlo, kinetic Langevin Monte Carlo and kinetic Langevin Monte Carlo for twice-differentiable potentials.
- Bounds on the moments of log-concave distributions, especially in the settings where the potential function f is strongly convex inside a ball and (weakly) convex everywhere else, or strongly convex outside a ball and (weakly) convex everywhere else. These bounds are low-degree polynomials in dimension and in the other relevant parameters.
- Upper bounds on the mixing time of the aforementioned three versions of the Langevin algorithm, obtained by combining the W_q-error bounds and moment evaluations.

The remainder of the paper is structured as follows. We begin in Section 2 by introducing the sampling methods based on the Langevin diffusion and by formulating the main assumptions that are used throughout this work. Section 3 and Section 4 contain, respectively, a discussion on the choice of the complexity measure and an overview of our main contributions. Relation to prior work is discussed in Section 5. Section 6 and Section 7 contain formal statements and proofs of the main error- and complexity-bounds for the vanilla and the kinetic Langevin, respectively. Upper bounds for the moments of two classes of log-concave distributions are presented in Section 8. We conclude with a discussion in Section 9.

2. Further Precisions on the Analyzed Methods

Since our main motivation for considering the sampling problem comes from applications in statistics and machine learning, we will focus on the Monge-Kantorovich-Wasserstein distances W_q defined by

$$W_q(\nu,\nu') = \inf\Big\{\mathbf{E}[\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}'\|_2^q]^{1/q} : \boldsymbol{\vartheta}\sim\nu \text{ and } \boldsymbol{\vartheta}'\sim\nu'\Big\}, \qquad q\geq 1.$$

The infimum above is over all the couplings between ν and ν'. In view of the Hölder inequality, the mapping $q\mapsto W_q(\nu,\nu')$ is increasing for every pair (ν,ν').

Our main contributions are upper bounds on quantities of the form $W_q(\nu_K,\pi)$where π is a log-concave target distribution and ν_K is the distribution of the Kth iterate of various discretization schemes of Langevin diffusions. More precisely, we consider two types of Langevin processes: the kinetic Langevin diffusion and the vanilla Langevin diffusion. The latter is the highly overdamped version of the former, see (Nelson, 1967). The Langevin diffusion, having π as invariant distribution, is defined as a solution[1] to the stochastic differential equation

$$d\boldsymbol{L}_t^{\mathsf{LD}} = -\nabla f(\boldsymbol{L}_t^{\mathsf{LD}})\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \qquad t\geq 0, \tag{2}$$

where $\boldsymbol{W}$ is a p-dimensional standard Brownian motion independent of the initial value $\boldsymbol{L}_0$. An illustration of this process is given in Figure 1. The LMC algorithm presented in (1) is merely the Euler-Maruyama discretization of the process $\boldsymbol{L}$. The kinetic Langevin diffusion $\{\boldsymbol{L}_t^{\mathsf{KLD}} : t\geq 0\}$, also known as the second-order Langevin process, is defined by

$$d\begin{bmatrix}\boldsymbol{V}_t\\ \boldsymbol{L}_t^{\mathsf{KLD}}\end{bmatrix} = \begin{bmatrix}-(\gamma\boldsymbol{V}_t+\nabla f(\boldsymbol{L}_t^{\mathsf{KLD}}))\\ \boldsymbol{V}_t\end{bmatrix}dt + \sqrt{2\gamma}\begin{bmatrix}\mathbf{I}_p\\ \mathbf{0}_{p\times p}\end{bmatrix}d\boldsymbol{W}_t, \qquad t\geq 0, \tag{3}$$

where $\gamma>0$ is the friction coefficient. The process $\boldsymbol{V}_t$ is often called the velocity process since the second row in (3) implies that $\boldsymbol{V}_t$ is the time derivative of $\boldsymbol{L}_t^{\mathsf{KLD}}$. The continuous-time Markov process $(\boldsymbol{L}_t^{\mathsf{KLD}},\boldsymbol{V}_t)$ is positive recurrent and has a unique invariant distribution, which has the following density with respect to the Lebesgue measure on $\mathbb{R}^{2p}$:

$$p_*(\boldsymbol{\theta},\boldsymbol{v}) \propto \exp\Big\{-f(\boldsymbol{\theta})-\frac{1}{2}\|\boldsymbol{v}\|_2^2\Big\}, \qquad \boldsymbol{\theta}\in\mathbb{R}^p,\ \boldsymbol{v}\in\mathbb{R}^p.$$

If $(\boldsymbol{L},\boldsymbol{V})$ is a pair of random vectors drawn from the joint density p_*, then $\boldsymbol{L}$ and $\boldsymbol{V}$ are independent, $\boldsymbol{L}$ is distributed according to the target π, whereas $\boldsymbol{V}$ is a standard Gaussian vector. Therefore, at equilibrium, the random variable $\boldsymbol{L}_t^{\mathsf{KLD}}$ has the target distribution π.

Time-discretized versions of Langevin diffusion processes (2) and (3) are used for (approximately) sampling from π. In order to guarantee that the discretization error is not too large, as well as that the process $\{\boldsymbol{L}_t\}$ converges fast enough to its invariant distribution, we need to impose some assumptions on f. In the present work, we will assume that either Conditions 1, 2 or Conditions 1, 2, 3 presented below are satisfied.

[1] Under the conditions imposed on f, namely the convexity and the Lipschitzness of the gradient, all the considered SDEs have unique strong solutions. Furthermore, all conditions (see e.g. (Pavliotis, 2014)) ensuring that π and p^* are invariant densities of, respectively, processes (2) and (3) are fulfilled.

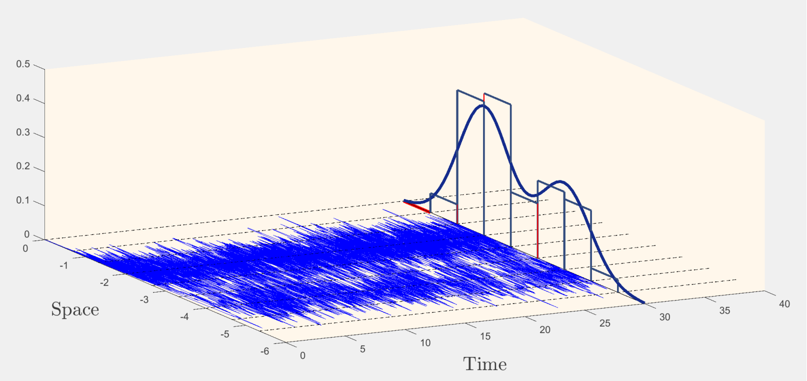

Figure 1: Illustration of Langevin dynamics. The blue lines represent different paths of a Langevin process. We see that the histogram of the state at time $t = 30$ is close to the target density (the dark blue line).

Condition 1 *The function f is continuously differentiable on $\mathbb{R}^p$ and its gradient ∇f is M-Lipschitz for some $M > 0$: $\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \le M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$.*

From now on, we will always assume that the Langevin (vanilla or kinetic) diffusion under consideration has the initial point $\boldsymbol{L}_0 = \mathbf{0}$. Some of the conditions presented below implicitly require that this initialization is not too far away from the "center" of the target distribution π. In many statistical problems where π is the Bayesian posterior, one can come close to these assumptions by shifting the distribution using a simple initial estimator.

Condition 2 *The function f is convex on $\mathbb{R}^p$. Furthermore, for some positive constants D and β, we have $\mu_2^2(\pi) = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta}\|_2^2] \le Dp^\beta$.*

Under Condition 2, the centered second moment of π scales polynomially with the dimension with power $\beta > 0$, while the flatness of the distribution is controlled by the parameter $D > 0$. Remarkably, Condition 2 implies that all the moments $\{\mu_q(\pi)\}_{q\ge 1}$ scale polynomially with p, provided that $\mathbf{E}_{\vartheta\sim\pi}[\|\boldsymbol{\vartheta}\|_2^2]$ also does. This fact is a consequence of Borell's lemma (Giannopoulos et al., 2014, Theorem 2.4.6), stating that for any $q \ge 1$, there is a numerical constant B_q that depends only on q such that $\mu_q(\pi) \le B_q\mu_2(\pi)$. An attempt to provide optimized constants in this inequality is stated in Lemma 5.

In the sequel, we show that the smoothness and the flatness of π have a combined impact on the sampling error considered. It turns out that the important parameter with respect to the hardness of the sampling problem is the product

$$\kappa := MD.$$

For m-strongly convex functions f, Condition 2 is satisfied with $D = 1/m$ and $\beta = 1$, according to Brascamp-Lieb inequality (Brascamp and Lieb, 1976). In this case, the parameter $\kappa = M/m$ is referred to as the condition number. We will show that Condition 2 is also satisfied for functions f that are convex everywhere and strongly convex inside a ball, as well as for functions f that are convex everywhere and strongly convex only outside a ball.

In the next assumption, we use notation $\|\mathbf{M}\|$ for the spectral norm (the largest singular value) of a matrix $\mathbf{M}$.

Condition 3 *The function f is twice differentiable in $\mathbb{R}^p$ with a M_2-Lipschitz Hessian $\nabla^2 f$ for some $M_2 > 0$: $\|\nabla^2 f(\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta}')\| \le M_2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$.*

Condition 3 ensures further smoothness of the potential f. When it holds, the Lipschitz continuity of the Hessian and the flatness of π also have a combined impact on the sampling error. A second important parameter with respect to the hardness of the sampling problem in such a case is the product

$$\kappa_2 := M_2^{2/3} D.$$

The case of an m-strongly convex function f has been studied in several recent papers. As a matter of fact, global strong convexity implies exponentially fast mixing of processes (2) and (3), with dimension-free rates e^{-mt} and $e^{-mt/(M+m)^{1/2}}$, respectively. When only (weak) convexity is assumed, such results do not hold in general.

The strategy we adopt here relies on a convexification trick, which consists in sampling from a distribution that is provably close to the target, but has the advantage of being strongly log-concave. More precisely, for some small positive α, the surrogate potential is defined by $f_\alpha(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + \alpha\|\boldsymbol{\theta}\|_2^2/2$. Therefore, the corresponding surrogate distribution has the density

$$\pi_\alpha(\boldsymbol{\theta}) := \frac{e^{-f_\alpha(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f_\alpha(\boldsymbol{v})} d\boldsymbol{v}}.$$

We stress the fact that the quadratic penalty $\alpha\|\boldsymbol{\theta}\|_2^2/2$ added to the potential f is centered at the origin. This is closely related to the fact that the diffusion is assumed to have the origin as initial point, and also to the fact that the origin is assumed here to be a good guess of the "center" of π. The parameter α, together with the step-size $\boldsymbol{h}$, is considered as a tuning parameter of the algorithms to be calibrated. Large values of α will result in fast convergence to π_α but a poor approximation of π by π_α. On the other hand, smaller values of α will lead to a small approximation error but also slow convergence. The next result quantifies the approximation of π by π_α, for different distances.

Proposition 1 *For any $\alpha \ge 0$ and $q \in [1, +\infty)$, there is a numerical constant C_q depending only on q such that*

$$d_{\rm TV}(\pi, \pi_\alpha) \le \alpha\mu_2^2(\pi), \qquad W_q^q(\pi, \pi_\alpha) \le C_q\alpha\mu_2(\pi)^{q+2}.$$

Here the constant C_q can be bounded for every q. In particular, $C_1 \le 11$ and $C_2 \le 111$.

This result allows us to control the bias induced by replacing the target distribution by the surrogate one and paves the way for choosing the "optimal" α by minimizing an upper bound on the sampling error. We draw the attention of the reader to the fact that, for W_q distance, the dependence on α of the upper bound is $\alpha^{1/q}$, which slows down when q

increases (recall that α is a small parameter). This explains the deterioration with increasing q of the complexity bounds presented in forthcoming sections. Let us define the constant

$$C_q = \inf\{C : W_q^q(\pi, \pi_\alpha) \leq C\alpha\mu_2(\pi)^{q+2},\ \forall\, \pi \text{ log-concave}\}, \tag{4}$$

which will repeatedly appear in the statements of the theorems.

3. How to Measure the Complexity of a Sampling Scheme?

We have already introduced the notation $K_{\mathsf{Alg},\mathsf{Crit}}(p, \varepsilon)$, the number of iterations that guarantee that algorithm Alg has an error—measured by criterion Crit—smaller than ε. If we choose a criterion, this quantity can be used to compare two methods, the iterates of which have comparable computational complexity. For example, LMC and KLMC being discretized versions of the Langevin process (2) and the kinetic Langevin process (3), respectively, are such that one iteration requires one evaluation of ∇f and generation of one realization of a Gaussian vector of dimension p or $2p$. Thus, the iterations are of comparable computational complexity and, therefore, it is natural to prefer LMC if $2K_{\text{LMC},\mathsf{Crit}}(p, \varepsilon) \leq K_{\text{KLMC},\mathsf{Crit}}(p, \varepsilon)$ and to prefer KLMC if the opposite inequality is true.

A delicate question that has not really been discussed in literature is a notion of complexity that allows to compare the quality of a given sampling method for two different criteria. To be more precise, assume that we are interested in the LMC algorithm and wish to figure out whether it is "more difficult" to perform approximate sampling for the TV-distance or for the Wasserstein distance. It is a well-known fact that the TV-distance induces the uniform strong convergence of measures whereas the Wasserstein distances induce the weak convergence. Therefore, at least intuitively, approximate sampling for the TV-distance should be harder than approximate sampling for the Wasserstein distance[2]. However, under Condition 1 and m-strong convexity of f, the available results for the LMC provide the same order of magnitude, p/ε^2, both for $K_{\text{LMC},\text{TV}}$ (Dalalyan, 2017b; Durmus and Moulines, 2019) and K_{LMC,W_2} (Durmus and Moulines, 2019; Dalalyan and Tsybakov, 2009). The point we want to put forward is that the origin of this discrepancy between the intuitions and mathematical results is the inappropriate scaling of the target accuracy in the definition of K_{LMC,W_2}.

To further justify the importance of choosing the right scaling of the target accuracy, let us make the following observation. The total-variation distance, on the one hand, serves to approximate probabilities, which are adimensional and scale-free quantities belonging to the interval $[0, 1]$. The Wasserstein distances, on the other hand, are useful for approximating moments[3], the latter depending both on the dimension and on the scale. For this reason, we suggest the following definition of the analogue of K in the case of Wasserstein distances:

$$K_{\mathsf{Alg},W_q}(p, \varepsilon) = \min\{k \in \mathbb{N} : W_q(\nu_k^{\mathsf{Alg}}, \pi) \leq \varepsilon\mu_2(\pi),\ \forall\, \pi \in \mathscr{P}\}, \tag{5}$$

where Alg is a Markov Chain Monte Carlo or another method of sampling, k is generally the number of calls to the oracle and $\mathscr{P}$ is a class of target distributions. Examples of oracle

[2] We underline here that the aforementioned hardness argument is based only on the topological consideration, since it is not possible, in general, to upper bound the Wasserstein distance W_q, for $q \geq 1$ by the TV-distance or a function of it.

[3] Recall that by the triangle inequality, one has $|\mu_q(\nu) - \mu_q(\pi)| \leq W_q(\nu, \pi)$.

call are the evaluation of the gradient of the potential at a given point or the computation of the product of the Hessian of f at a given point and a given vector. Note also that $\mu_2(\pi)$ is the W_2 distance between the Dirac mass at the origin and the target distribution.

Definition (5), as opposed to those used in prior work, has the advantage of being scale invariant and reflecting the fact that we deal with objects that might be large if the dimension is large. Note that the idea of scaling the error in order to make the complexity measure scale-invariant has been recently used in (Tat Lee et al., 2018; Baker et al., 2019) as well. Indeed, in the context of m-strongly log-concave distributions, Tat Lee et al. (2018) propose to find the smallest k such that $W_2(\nu_k^{\mathsf{Alg}}, \pi) \leq \varepsilon/\sqrt{m}$. This is close to our proposal, since in the case of m-strongly log-concave distributions, it follows from the Brascamp-Lieb inequality that $\sup_\pi \mu_2(\pi) = \sqrt{p/m}$ (the sup is attained for Gaussian distributions).

4. Overview of Main Contributions

In this work, we analyze three methods, LMC, KLMC (Cheng et al., 2018b) and KLMC2 (Dalalyan and Riou-Durand, 2020), applied to the strong-convexified potential $f_\alpha(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + (\alpha/2)\|\boldsymbol{\theta}\|_2^2$ in order to cope with the lack of strong convexity. We briefly recall these algorithms and present a summary of the main contributions of this work.

4.1 Considered Markov Chain Monte-Carlo Methods

We first recall the definition of the Langevin Monte Carlo algorithms. For the LMC algorithm introduced in (1), we will only use the constant step-size form, the update rule of which is given by

$$\boldsymbol{\vartheta}_{k+1} = (1-\alpha h)\boldsymbol{\vartheta}_k - h\nabla f(\boldsymbol{\vartheta}_k) + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{α-LMC}$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots$ is a sequence of mutually independent, independent of $\boldsymbol{\vartheta}_0$, centered Gaussian vectors with covariance matrices equal to identity. We will refer to this version of the LMC algorithm as α-LMC.

We now recall the definition of the first and second-order Kinetic Langevin Monte Carlo algorithms. We suppose that, for some initial distribution ν_0 chosen by the user, both KLMC and KLMC2 algorithms start from $(\boldsymbol{v}_0, \boldsymbol{\vartheta}_0) \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) \otimes \nu_0$. Before stating the update rules, we specify the structure of the random perturbation generated at each step. In what follows, $\{(\boldsymbol{\xi}_k^{(1)}, \boldsymbol{\xi}_k^{(2)}, \boldsymbol{\xi}_k^{(3)}, \boldsymbol{\xi}_k^{(4)}) : k \in \mathbb{N}\}$ will stand for a sequence of iid $4p$-dimensional centered Gaussian vectors, independent of the initial condition $(\boldsymbol{v}_0, \boldsymbol{\vartheta}_0)$.

To specify the covariance structure of these Gaussian variables, we define two sequences of functions (ψ_k) and (φ_k) as follows. For every $t > 0$, let $\psi_0(t) = e^{-\gamma t}$, then for every $k \in \mathbb{N}$, define $\psi_{k+1}(t) = \int_0^t \psi_k(s)\,ds$ and $\varphi_{k+1}(t) = \int_0^t e^{-\gamma(t-s)}\psi_k(s)\,ds$. Now, let us denote by $\xi_{k,j}$ for the j-th component of the vector $\boldsymbol{\xi}_k$ (a scalar), and assume that for any fixed k, the 4-dimensional random vectors $\{(\xi_{k,j}^{(1)}, \xi_{k,j}^{(2)}, \xi_{k,j}^{(3)}, \xi_{k,j}^{(4)}) : 1 \leq j \leq p\}$ are iid with the covariance matrix

$$\mathbf{C}_{h,\gamma} = \int_0^h [\psi_0(t);\, \psi_1(t);\, \varphi_2(t);\, \varphi_3(t)]^\top [\psi_0(t);\, \psi_1(t);\, \varphi_2(t);\, \varphi_3(t)]\,dt.$$

The KLMC algorithm, introduced by Cheng et al. (2018b), is a sampler derived from a suitable time-discretization of the kinetic diffusion. When applied to the strong-convexified potential f_α, for a step-size $h > 0$, its update rule reads as follows

$$\begin{bmatrix}\boldsymbol{v}_{k+1}\\ \boldsymbol{\vartheta}_{k+1}\end{bmatrix} = \begin{bmatrix}\psi_0(h)\boldsymbol{v}_k - \psi_1(h)(\nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k)\\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)(\nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k)\end{bmatrix} + \sqrt{2\gamma}\begin{bmatrix}\boldsymbol{\xi}^{(1)}_{k+1}\\ \boldsymbol{\xi}^{(2)}_{k+1}\end{bmatrix}. \qquad (\alpha\text{-KLMC})$$

Roughly speaking, this formula is obtained from (3) by replacing the function $t \mapsto \nabla f(\boldsymbol{L}_t)$ by a piecewise constant approximation. Such an approximation is made possible by the fact that f is gradient-Lipschitz.

It is natural to expect that further smoothness of f may allow one to improve upon the aforementioned piecewise constant approximation. This is done by the KLMC2 algorithm, introduced by Dalalyan and Riou-Durand (2020), which takes advantage of the existence and smoothness of the Hessian of f in order to use a local-linear approximation. At any iteration $k \in \mathbb{N}$ with a current value $\boldsymbol{\vartheta}_k$, define the gradient $\boldsymbol{g}_{k,\alpha} = \nabla f(\boldsymbol{\vartheta}_k) + \alpha\boldsymbol{\vartheta}_k$ and the Hessian $\mathbf{H}_{k,\alpha} = \nabla^2 f(\boldsymbol{\vartheta}_k) + \alpha\mathbf{I}_p$. When applied to the modified strongly convex potential f_α, for $h > 0$, the update rule of the KLMC2 algorithm is

$$\begin{bmatrix}\boldsymbol{v}_{k+1}\\ \boldsymbol{\vartheta}_{k+1}\end{bmatrix} = \begin{bmatrix}\psi_0(h)\boldsymbol{v}_k - \psi_1(h)\boldsymbol{g}_{k,\alpha} - \varphi_2(h)\mathbf{H}_{k,\alpha}\boldsymbol{v}_k\\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\boldsymbol{g}_{k,\alpha} - \varphi_3(h)\mathbf{H}_{k,\alpha}\boldsymbol{v}_k\end{bmatrix} + \sqrt{2\gamma}\begin{bmatrix}\boldsymbol{\xi}^{(1)}_{k+1} - \mathbf{H}_{k,\alpha}\boldsymbol{\xi}^{(3)}_{k+1}\\ \boldsymbol{\xi}^{(2)}_{k+1} - \mathbf{H}_{k,\alpha}\boldsymbol{\xi}^{(4)}_{k+1}\end{bmatrix}. \qquad (\alpha\text{-KLMC2})$$

Notice that if we apply KLMC2 with $\mathbf{H}_{k,\alpha} = 0$, we recover the KLMC algorithm. These two algorithms, derived from the kinetic Langevin diffusion, will be referred to as α-KLMC and α-KLMC2.

4.2 Summary of the Obtained Complexity Bounds

Without going into details here, we mention in the tables below the order of magnitude of the number of iterations required by different algorithms for getting an error bounded by ε for various metrics. For improved legibility, we do not include logarithmic factors and report the order of magnitude of $K_{\square,\square}(p,\varepsilon)$ in the case when the parameter β in Condition 2 is fixed to a particular value. We present hereafter the case where $\beta = 1$, which is of particular interest as discussed in Section 8.

$\beta = 1$	LMCa	α-LMC		α-KLMC	α-KLMC2
Cond.	1-2	1-2	1-3	1-2	1-3
W_2	–	$\kappa p^2/\varepsilon^6$	$(\kappa_2^{1.5}p^{0.5} + \kappa^{1.5})p^2/\varepsilon^5$	$\kappa^{1.5}p^2/\varepsilon^5$	$\kappa^{0.5}\kappa_2{}^{1.5}p^2/\varepsilon^4$
W_1	–	$\kappa p^2/\varepsilon^4$	$(\kappa_2^{1.5}p^{0.5} + \kappa^{1.5})p^2/\varepsilon^3$	$\kappa^{1.5}p^2/\varepsilon^3$	$\kappa^{0.5}\kappa_2{}^{1.5}p^2/\varepsilon^2$
d_{TV}	$\kappa p^2/\varepsilon^4$ △	$\kappa^2p^3/\varepsilon^4$ □	–	–	–

The results indicated by △ describe the behavior of the Langevin Monte Carlo with averaging established in (Durmus et al., 2019, Cor. 7). To date, these results have the best known dependence (under conditions 1 and 2 only) on p. The results indicated by □ summarize the behavior of the Langevin Monte Carlo established in (Dalalyan, 2017b). All the remaining cells of the table are filled in by the results obtained in the present work. One can observe that the results for W_1 are strictly better than those for W_2. Similar hierarchy was already reported in (Majka et al., 2020, Remark 1.9). It is also worth mentioning here, that using Metropolis-Hastings adjustment of the LMC (termed MALA), Dwivedi et al. (2018); Chen et al. (2020) obtained[4] the complexity

$$K_{\text{MALA,TV}}(p,\varepsilon) = O\bigg(\frac{p^2\kappa^{3/2}}{\varepsilon^{3/2}}\,\log\big(p\kappa/\varepsilon\big)\bigg).$$

It is still an open question whether this type of result can be proved for Wasserstein distances.

We would also like to comment on the relation between the third and the fourth columns of the table, corresponding to the α-LMC algorithm under different sets of assumptions. The result for the more constrained Hessian-Lipschitz case is not always better than the result when only gradient Lipschitzness is assumed. For instance, for W_2, the latter is better than the former when $\kappa \lesssim (\kappa_2^{1.5}p^{0.5}+\kappa^{1.5})\varepsilon$, which is equivalent to $M \lesssim (M_2p^{1/2}+M^{3/2})D^{1/2}\varepsilon$. At a very high level, this reflects the fact that when the condition number is large, the Hessian-Lipschitzness does not help to get an improved result. Note that the same phenomenon occurs in the strongly log-concave case.

4.3 The General Approach Based on a Log-Strongly-Concave Surrogate

We have already mentioned that the strategy we adopt here is the one described in (Dalalyan, 2017b), consisting of replacing the potential of the target density by a strongly convex surrogate. Prior to instantiating this approach to various sampling algorithms under various conditions and error measuring distances, we provide here a more formal description of it. In the remaining dist is a general distance on the set of all probability measures.

We will denote by $\nu_{k,\alpha}^{\mathsf{Alg}}$ the distribution of the random vector obtained after performing k iterations of the algorithm Alg with the surrogate potential $f_\alpha(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + \alpha\|\boldsymbol{\theta}\|_2^2/2$. Our first goal is to establish an upper bound on the distance between the sampling distribution $\nu_{k,\alpha}^{\mathsf{Alg}}$ and the target π. The methods we analyze here depend on the step-size h, as they are discretizations of continuous-time diffusion processes. Thus, the obtained bound will depend on h. This bound should be so that one can make it arbitrarily small by choosing small α and h and a large value of k. In a second stage, the goal is to exploit the obtained error-bounds in order to assess the order of magnitude of the computational complexity K, defined in Section 3, as a function of p, ε and the condition number κ.

To achieve this goal, we first use the triangle inequality

$$\mathsf{dist}(\nu_{k,\alpha}^{\mathsf{Alg}},\pi) \le \mathsf{dist}(\nu_{k,\alpha}^{\mathsf{Alg}},\pi_\alpha) + \mathsf{dist}(\pi_\alpha,\pi).$$

Then, the second term of the right hand side of the last displayed equation is bounded using Proposition 1. Finally, the distance between the sampling density $\nu_{k,\alpha}^{\mathsf{Alg}}$ and the surrogate π_α is bounded using the prior work on sampling for log-strongly-concave distributions. Optimizing over α leads to the best bounds on precision and complexity.

[4] This rate is not explicitly present in the cited papers, but it can be derived from (Chen et al., 2020, Theorem 5) using the approach outlined in (Dwivedi et al., 2018, Section 3.3).

5. Prior Work

Mathematical analysis of MCMC methods defined as discretizations of diffusion processes is an active area of research since several decades. Important early references are (Roberts and Tweedie, 1996; Roberts and Rosenthal, 1998; Roberts and Stramer, 2002; Douc et al., 2004) and the references therein. Although those papers do cover the multidimensional case, the guarantees they provide do not make explicit the dependence on the dimension. In a series of work analyzing ball walk and hit-and-run MCMCs, Lovasz and Vempala (2006); Lovász and Vempala (2006) put forward the importance of characterizing the dependence of the number of iterations on the dimension of the state space.

More recently, Dalalyan (2017b) advocated for analyzing MCMCs obtained from continuous time diffusion processes by decomposing the error into two terms: a non-stationarity error of the continuous-time process and a discretization error. A large number of works applied this kind of approach in various settings. (Bubeck et al., 2018; Durmus and Moulines, 2017, 2019; Durmus et al., 2018) improved the results obtained by Dalalyan and extended them in many directions including non-smooth potentials and variable step-sizes. While previous work studied the sampling error measured by the total variation and Waserstein distances, (Cheng and Bartlett, 2018) proved that similar results hold for the Kulback-Leibler divergence. (Cheng et al., 2018b,a; Dalalyan and Riou-Durand, 2020) investigated the case of a kinetic Langevin diffusion, showing that it leads to improved dependence on the dimension. A promising line of related research, initiated by (Wibisono, 2018; Bernton, 2018), is to consider the sampling distributions as a gradient flow in a space of measures. The benefits of this approach were demonstrated in (Durmus et al., 2019; Ma et al., 2019).

Motivated by applications in Statistics and Machine Learning, many recent papers developed theoretical guarantees for stochastic versions of algorithms, based on noisy gradients, see (Baker et al., 2019; Chatterji et al., 2018; Dalalyan and Karagulyan, 2019; Dalalyan, 2017a; Zou et al., 2019; Raginsky et al., 2017) and the references therein. A related topic is non-asymptotic guarantees for the Hamiltonian Monte Carlo (HMC). There is a growing literature on this in recent years, see (Mangoubi and Smith, 2017; Tat Lee et al., 2018; Chen and Vempala, 2019; Mangoubi and Smith, 2019) and the references therein.

In all these results, the dependence of the number of iterations on the inverse precision is polynomial. (Dwivedi et al., 2018; Chen et al., 2020; Mangoubi and Vishnoi, 2019) proved that one can reduce this dependence to logarithmic by using Metropolis adjusted versions of the algorithms.

6. Precision and Computational Complexity of the LMC

In this section, we present non-asymptotic upper bounds in the (non-strongly) convex case for the suitably adapted LMC algorithm for Wasserstein and bounded-Lipschitz error measures under two sets of assumptions: Conditions 1-2 and Conditions 1-3. To refer to these settings, we will call them “Gradient-Lipschitz” and “Hessian-Lipschitz”, respectively. The main goal is to provide a formal justification of the rates included in columns 2 and 3 of the table presented in Section 4.2. To ease notation, and since there is no risk of confusion, we write μ_2 instead $\mu_2(\pi)$.

6.1 The Gradient-Lipschitz Setting

First we consider the Gradient-Lipschitz setting and give explicit conditions on the parameters α, h and K to have a theoretical guarantee on the sampling error, measured in the Wasserstein distance, of the LMC algorithm.

Theorem 1 *Suppose that the potential function f is convex and satisfies Condition 1. Let $q \in [1, 2]$. Then, for every $\alpha \leq M/20$ and $h \leq 1/(M + \alpha)$, we have*

$$W_q(\nu_K^{\alpha\text{-}\mathsf{LMC}}, \pi) \leq \underbrace{\mu_2(1-\alpha h)^{K/2}}_{\substack{\text{error due to the}\\ \text{time finiteness}}} + \underbrace{(2.1hMp/\alpha)^{1/2}}_{\text{discretization error}} + \underbrace{\Big(C_q\alpha\mu_2^{q+2}\Big)^{1/q}}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}},$$

where C_q is a dimension free constant given by (4).

The proof of this result is postponed to the end of this section. Let us consider its consequences in the cases $q = 1$ and $q = 2$ presented in the table of Section 4.2. The general strategy is to choose the value of α by minimizing the sum of the discretization error and the error caused by the lack of strong convexity. Then, the parameter h is chosen so that the sum of the two aforementioned errors is smaller than 99% of the target precision $\varepsilon\mu_2$. Finally, the number of iterations K is selected in such a way that the error due to the time finiteness is also smaller than 1% of the target precision.

Implementing this strategy for $q = 1$ and $q = 2$, we get the optimized value of α and the corresponding value of h,

$q = 1$		$q = 2$	
$\alpha = \dfrac{(2.1hMp)^{1/3}}{44^{2/3}\mu_2^2}$	$h = \dfrac{\varepsilon^3}{322Mp}$	$\alpha = \dfrac{(2.1hMp)^{1/2}}{111^{1/2}\mu_2^2}$	$h = \dfrac{\varepsilon^4}{3900Mp}$

These values of α and h satisfy the conditions imposed in Theorem 1. They imply that the computational complexity of the method, for $\nu_0 = \delta_0$ (the Dirac mass at the origin), is given by

$$K_{\alpha\text{-LMC},W_1}(p,\varepsilon) \leq \frac{2}{\alpha h}\log\Big(\frac{100W_2(\nu_0,\pi_\alpha)}{\varepsilon\mu_2}\Big) \leq 4.3\times 10^4 M\frac{\mu_2^2 p}{\varepsilon^4}\log(100/\varepsilon)$$

$$K_{\alpha\text{-LMC},W_2}(p,\varepsilon) \leq \frac{2}{\alpha h}\log\Big(\frac{100W_2(\nu_0,\pi_\alpha)}{\varepsilon\mu_2}\Big) \leq 3.6\times 10^6 M\frac{\mu_2^2 p}{\varepsilon^6}\log(100/\varepsilon).$$

In both inequalities, the second passage is due to the monotone behaviour of the function $\alpha \mapsto \mu_2(\pi_\alpha)$. This property, formulated in Lemma 1, implies that

$$W_2(\delta_0, \pi_\alpha) = \mu_2(\pi_\alpha) \leq \mu_2(\pi). \tag{6}$$

Combining Condition 2, $M\mu_2^2(\pi) \leq \kappa p^\beta$, with the last display, we check that $K_{\alpha\text{-LMC},W_1}(p,\varepsilon) \leq C\kappa(p^{1+\beta}/\varepsilon^4)\log(100/\varepsilon)$ and $K_{\alpha\text{-LMC},W_2}(p,\varepsilon) \leq C\kappa(p^{1+\beta}/\varepsilon^6)\log(100/\varepsilon)$. For $\beta = 1$, this matches well with the rates reported in the table of Section 4.2. Unfortunately, the numerical constant C, just like the factors 4.3×10^4 and 3.6×10^6 in the last display, is too large to be useful for practical purposes. Getting similar bounds with better numerical constants is an open question. The same remark applies to all the results presented in the subsequent sections.

Proof of Theorem 1 To ease notation, we write ν_K instead of $\nu_K^{\alpha\text{-LMC}}$. The triangle inequality and the monotonicity of W_q with respect to q imply that

$$W_q(\nu_K,\pi) \le W_2(\nu_K,\pi_\alpha) + W_q(\pi_\alpha,\pi).$$

Recall that π_α is α-strongly log-concave and has f_α as its potential function. By definition, f_α has also a Lipschitz continuous gradient with the Lipschitz constant at most $M+\alpha$. As we assume that $h \le 1/(M+\alpha)$, we can apply (Durmus et al., 2019, Theorem 9). It implies that

$$\begin{aligned}W_2(\nu_K,\pi_\alpha) &\le (1-\alpha h)^{K/2} W_2(\nu_0,\pi_\alpha) + \big(2h(M+\alpha)p/\alpha\big)^{1/2}\\ &\le (1-\alpha h)^{K/2} W_2(\nu_0,\pi_\alpha) + (2.1hMp/\alpha)^{1/2}.\end{aligned}$$

The last inequality is true due to the fact that $\alpha \le M/20$. Combining this inequality with Proposition 1, we obtain

$$\begin{aligned}W_q(\nu_K,\pi) &\le (1-\alpha h)^{K/2} W_2(\nu_0,\pi_\alpha) + (2.1hMp/\alpha)^{1/2} + W_q(\pi_\alpha,\pi)\\ &\le \mu_2(\pi_\alpha)(1-\alpha h)^{K/2} + (2.1hMp/\alpha)^{1/2} + \big(C_q\alpha\mu_2^{q+2}\big)^{1/q}.\end{aligned}$$

Thus, applying (6), we get the claim of the theorem. ■

6.2 The Hessian-Lipschitz Setting

It has been noticed by Durmus and Moulines (2019), see also (Dalalyan and Karagulyan, 2019, Theorem 5), that if the potential f has a Lipschitz-continuous Hessian matrix, then the LMC algorithm, without any modification, is more accurate than in the Gradient-Lipschitz setting. These improvements were obtained under the condition of strong convexity of the potential, showing that the computational complexity drops down from p/ε^2 to p/ε. The goal of this section is to understand how this additional smoothness assumption impacts the computational complexity of the α-LMC algorithm.

Theorem 2 *Suppose that the potential function f satisfies conditions 1 and 3. Let $q \in [1,2]$. For every $\alpha \le M/20$ and $h \le 1/(M+\alpha)$, we have*

$$W_q(\nu_K^{\alpha\text{-LMC}},\pi) \le \underbrace{\mu_2(1-\alpha h)^K}_{\substack{\text{error due to the}\\ \text{time finiteness}}} + \underbrace{\frac{M_2hp}{2\alpha} + \frac{2.8M^{3/2}hp^{1/2}}{\alpha}}_{\text{discretization error}} + \underbrace{\Big(C_q\alpha\mu_2^{q+2}\Big)}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}}{}^{1/q},$$

where C_q is a dimension free constant given by (4).

In order to provide more insight on the complexity bounds implied by the latter result, let us instantiate it for $q=1$ and $q=2$. Optimizing the sum of the two last error terms with respect to α, then choosing this sum to be equal to $0.99\varepsilon\mu_2$, we arrive at the following values

$q=1$		$q=2$	
$\alpha = \left(\dfrac{hQp}{44\mu_2^3}\right)^{1/2}$	$h = \dfrac{\varepsilon^2}{45\mu_2 Qp}$	$\alpha = \dfrac{(hQp)^{2/3}}{(111\mu_2^4)^{1/3}}$	$h = \dfrac{\varepsilon^3}{387\mu_2 Qp}$

Here Q is defined as $(M_2 + 5.6M^{3/2}p^{-1/2})$. These values of α and h satisfy the conditions imposed in Theorem 2. They imply that the computational complexity of the method, for $\nu_0 = \delta_0$ (the Dirac mass at the origin), is given by

$$K_{\alpha\text{-LMC},W_1}(p,\varepsilon) \leq \frac{2}{\alpha h}\log\Big(\frac{100W_2(\nu_0,\pi_\alpha)}{\varepsilon\mu_2}\Big) \leq 2\times 10^3\mu_2^3 Q(p/\varepsilon^3)\log(100/\varepsilon)$$

$$K_{\alpha\text{-LMC},W_2}(p,\varepsilon) \leq \frac{2}{\alpha h}\log\Big(\frac{100W_2(\nu_0,\pi_\alpha)}{\varepsilon\mu_2}\Big) \leq 9.9\times 10^4\mu_2^3 Q(p/\varepsilon^5)\log(100/\varepsilon).$$

Combining Condition 2 and the last display, we check that

$$K_{\alpha\text{-LMC},W_1}(p,\varepsilon) \leq C\varepsilon^{-3}(\kappa_2^{3/2}p^{(2+3\beta)/2} + \kappa^{3/2}p^{(1+3\beta)/2})\log(100/\varepsilon),$$

$$K_{\alpha\text{-LMC},W_2}(p,\varepsilon) \leq C\varepsilon^{-5}(\kappa_2^{3/2}p^{(2+3\beta)/2} + \kappa^{3/2}p^{(1+3\beta)/2})\log(100/\varepsilon).$$

The latter is true, since by definition κ_2 is equal to $M_2^{2/3}D$. For $\beta = 1$, this matches well with the rates reported in the table of Section 4.2.

Proof of Theorem 2 We repeat the same steps as in the proof of Theorem 1, except that instead of (Durmus et al., 2019, Theorem 9) we use (Dalalyan and Karagulyan, 2019, Theorem 5). To ease notation, we write ν_K instead of $\nu_K^{\alpha\text{-}\mathsf{LMC}}$. One easily checks that π_α is α-strongly log-concave with potential function f_α. Furthermore, the latter is $(M+\alpha)$-gradient-Lipschitz and M_2-Hessian-Lipschitz. Therefore, for $h \leq 2/(M+\alpha)$, Theorem 5 from (Dalalyan and Karagulyan, 2019) implies that

$$\begin{aligned} W_2(\nu_K,\pi_\alpha) &\leq (1-\alpha h)^K W_2(\nu_0,\pi_\alpha) + \frac{M_2hp}{2\alpha} + \frac{13(M+\alpha)^{3/2}hp^{1/2}}{5\alpha} \\ &\leq (1-\alpha h)^K W_2(\nu_0,\pi_\alpha) + \frac{M_2hp}{2\alpha} + \frac{2.8M^{3/2}hp^{1/2}}{\alpha}, \end{aligned}$$

where the second inequality follows from the fact that $\alpha \leq M/20$. The triangle inequality and the monotonicity of W_q with respect to q yield $W_q(\nu_K,\pi) \leq W_2(\nu_K,\pi_\alpha) + W_q(\pi_\alpha,\pi)$, which leads to

$$W_q(\nu_K,\pi) \leq \mu_2(\pi_\alpha)(1-\alpha h)^K + \frac{M_2hp}{2\alpha} + \frac{2.8M^{3/2}hp^{1/2}}{\alpha} + W_q(\pi,\pi_\alpha).$$

Replacing the last term above by its upper bound provided by Proposition 1 and applying (6), we get the claimed result. ■

7. Precision and Computational Complexity of KLMC and KLMC2

Several recent studies showed that for some classes of targets, including the strongly log-concave densities, the sampling error of discretizations of the kinetic Langevin diffusion scales better with the large dimension than discretizations of the Langevin diffusion. However, the dependence of the available bounds on the condition number is better for the Langevin diffusion. In this section we show a similar behavior in the case of (non-strongly) log-concave densities. This is done by providing quantitative upper bounds on the error of sampling using the kinetic Langevin process.

Theorem 3 *Suppose that the potential function f satisfies Condition 1. Let $q \in [1, 2]$. Then for every $\alpha \le M/20$, $\gamma \ge \sqrt{M + 2\alpha}$ and $h \le \alpha/(4\gamma(M + \alpha))$, we have*

$$W_q(\nu_K^{\alpha\text{-}\mathsf{KLMC}}, \pi) \le \underbrace{\sqrt{2}\,\mu_2\Big(1 - \frac{3\alpha h}{4\gamma}\Big)^K}_{\text{error due to the time finiteness}} + \underbrace{1.5Mp^{1/2}(h/\alpha)}_{\text{discretization error}} + \underbrace{\Big(C_q\alpha\mu_2^{q+2}\Big)^{1/q}}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}}.$$

where C_q is a dimension free constant given by (4)*.*

The proof of this result is postponed to the end of this section. Since the contraction rate is an increasing function of γ, we choose its lowest possible value achieved for $\gamma = \sqrt{M + 2\alpha}$. Then the strategy is the same as for the previous section, that is to choose the value of α by minimizing the sum of the discretization error and the error caused by the lack of strong convexity. Then, the parameter h is chosen so that the sum of the two aforementioned errors is smaller than 99% of the target precision $\varepsilon\mu_2$. The number of iterations K is selected in such a way that the error due to the time finiteness is also smaller than 1% of the target precision. Implementing this strategy for $q = 1$ and $q = 2$, we get the optimized value of α and the corresponding value of h,

$q = 1$		$q = 2$	
$\alpha = \dfrac{(1.5hMp^{1/2})^{1/2}}{(21\mu_2^3)^{1/2}}$	$h = \dfrac{\varepsilon^2}{143M\mu_2p^{1/2}}$	$\alpha = \dfrac{(3hMp^{1/2})^{2/3}}{(111\mu_2^4)^{1/3}}$	$h = \dfrac{\varepsilon^4}{1200M\mu_2p^{1/2}}$

These values of α and h satisfy the conditions imposed in Theorem 3. They imply that the computational complexity of the method, for $\nu_0 = \delta_0$ (the Dirac mass), is given by

$$K_{\alpha\text{-KLMC},W_1}(p, \varepsilon) \le \frac{4\gamma}{3\alpha h}\log\Big(\frac{150}{\varepsilon}\Big) \le 9.2 \times 10^3(M\mu_2^2)^{3/2}(p^{1/2}/\varepsilon^3)\log(150/\varepsilon)$$
$$K_{\alpha\text{-KLMC},W_2}(p, \varepsilon) \le \frac{4\gamma}{3\alpha h}\log\Big(\frac{150}{\varepsilon}\Big) \le 4.4 \times 10^5(M\mu_2^2)^{3/2}(p^{1/2}/\varepsilon^5)\log(150/\varepsilon).$$

Recall that Condition 2 implies $M\mu_2^2 \le \kappa p^\beta$. Combining this inequality with the last display, we check that

$$K_{\alpha\text{-KLMC},W_q}(p, \varepsilon) \le C\kappa^{3/2}(p^{(1+3\beta)/2}/\varepsilon^{2q+1})\log(150/\varepsilon), \qquad q = 1, 2.$$

For $\beta = 1$, this matches well with the rates reported in the table of Section 4.2.

Proof of Theorem 3 To ease notation, we write ν_K instead of $\nu_K^{\alpha\text{-}\mathsf{KLMC}}$. The triangle inequality and the monotonicity of W_q with respect to q imply that $W_q(\nu_K, \pi) \le W_2(\nu_K, \pi_\alpha) + W_q(\pi_\alpha, \pi)$. Recall that π_α is a α-strongly log-concave distribution with potential function f_α. By definition, f_α has also a Lipschitz continuous gradient with the Lipschitz constant at most $M + \alpha$. As we assumed that $\alpha \le M/20$, $\gamma \ge \sqrt{M + 2\alpha}$ and $h \le \alpha/(4\gamma(M + \alpha))$, we can apply (Dalalyan and Riou-Durand, 2020, Theorem 2). The latter implies that

$$\begin{aligned} W_2(\nu_K, \pi_\alpha) &\le \sqrt{2}(1 - 3\alpha h/(4\gamma))^K W_2(\nu_0, \pi_\alpha) + \sqrt{2}(M + \alpha)p^{1/2}(h/\alpha) \\ &\le \sqrt{2}\,\mu_2(\pi_\alpha)(1 - 3\alpha h/(4\gamma))^K + 1.5Mp^{1/2}(h/\alpha). \end{aligned}$$

The last inequality is true thanks to the fact that $\alpha \leq M/20$. Thus, (6) yields

$$W_q(\nu_K, \pi) \leq \sqrt{2}\, \mu_2(\pi)(1 - 3\alpha h/(4\gamma))^K + 1.5 M p^{1/2}(h/\alpha) + W_q(\pi_\alpha, \pi).$$

Owing to Proposition 1, the last term of the last display can be bounded by $(C_q \alpha \mu_2^{q+2})^{1/q}$. This completes the proof. ∎

The rest of this section is devoted to the sampling guarantees for the KLMC2 algorithm. Recall that this algorithm requires accurate evaluations of the Hessian of the potential function f to be available at each given point.

Theorem 4 *Suppose that the potential function f satisfies conditions 1 and 3. Let $q \in [1,2]$ and $Q = M_2 + M^{3/2} p^{-1/2}$. Then, for every $\alpha, h, \gamma > 0$ such that*

$$\alpha \leq \frac{M}{20}, \qquad \gamma \geq \sqrt{M + 2\alpha}, \qquad h \leq \frac{\alpha}{5\gamma(M+\alpha)} \bigvee \frac{\alpha}{4M_2\sqrt{5p}},$$

we have

$$W_q(\nu_K^{\alpha\text{-}\mathsf{KLMC2}}, \pi) \leq \underbrace{\sqrt{2}\, \mu_2 \left(1 - \frac{\alpha h}{4\gamma}\right)^K}_{\substack{\text{error due to the}\\ \text{time finiteness}}} + \underbrace{\frac{2h^2 Q p}{\alpha} + \frac{1.6}{\sqrt{M}} \exp\left\{-\frac{(\alpha/h)^2}{160 M_2^2}\right\}}_{\text{discretization error}} + \underbrace{\left(C_q \alpha \mu_2^{q+2}\right)^{1/q}}_{\substack{\text{error due to the lack}\\ \text{of strong-convexity}}},$$

where C_q is a dimension free constant given by (4).

The proof of this result is postponed to the end of this section. The contraction rate is an increasing function of γ, therefore we choose its lowest possible value achieved for $\gamma = \sqrt{M + 2\alpha}$. In this case the strategy for finding h and α is slightly different from the previous ones. Here, we first choose the parameter h so that the two terms of the discretization error are respectively bounded by 1% and 2% of the target precision $\varepsilon\mu_2$. This yields the following choice for the time step h:

$$h = \alpha \left(160 M_2^2 \log\left(\frac{160}{\varepsilon \mu_2 \sqrt{M}}\right) \bigvee \frac{100\alpha Q p}{\varepsilon \mu_2}\right)^{-1/2}.$$

The parameter α is then chosen so that the error due to the lack of strong convexity is lower than 96% of the target precision. Implementing this strategy for $q = 1$ and $q = 2$, we get the following value for α

$q = 1$	$q = 2$
$\alpha = \dfrac{\varepsilon}{23\mu_2^2}$	$\alpha = \dfrac{\varepsilon^2}{116\mu_2^2}$

Finally, the number of iterations K is selected in such a way that the error due to the time finiteness is also smaller than 1% of the target precision. This yields, that

$$K = \frac{4\gamma}{\alpha h} \log\left(\frac{142}{\varepsilon}\right)$$

is sufficient to reach the target precision. The values of γ, α and h imply that the computational complexity of the method is given by

$$K_{\alpha\text{-KLMC2},W_1}(p,\varepsilon) = 2.2\times 10^4\,\frac{M^{1/2}M_2\mu_2^4}{\varepsilon^2}\bigg\{1.6\log\Big(\frac{160}{\varepsilon\mu_2\sqrt{M}}\Big)\bigvee\frac{Qp}{23M_2^2\mu_2^3}\bigg\}^{1/2}\log\Big(\frac{142}{\varepsilon}\Big)$$

$$K_{\alpha\text{-KLMC2},W_2}(p,\varepsilon) = 5.4\times 10^6\,\frac{M^{1/2}M_2\mu_2^4}{\varepsilon^4}\bigg\{1.6\log\Big(\frac{160}{\varepsilon\mu_2\sqrt{M}}\Big)\bigvee\frac{\varepsilon Qp}{116M_2^2\mu_2^3}\bigg\}^{1/2}\log\Big(\frac{142}{\varepsilon}\Big).$$

Since according to Condition 2, $\mu_2 \le Dp^\beta$, the last display implies that up to logarithmic factors $K_{\alpha\text{-KLMC2},W_1}(p,\varepsilon)$ scales as $\kappa^{1/2}\kappa_2^{3/2}p^{2\beta}/\varepsilon^2$ and $K_{\alpha\text{-KLMC2},W_2}(p,\varepsilon)$ scales as $\kappa^{1/2}\kappa_2^{3/2}p^{2\beta}/\varepsilon^4$. For $\beta = 1$, this matches well with the rates reported in the table of Section 4.2.

Proof of Theorem 4 To ease notation, we write ν_K instead of $\nu_K^{\alpha\text{-}\mathsf{KLMC2}}$. As already checked in the proof of Theorem 2, the distribution π_α is α-strongly log-concave with potential function f_α. Furthermore, the latter is $(M+\alpha)$-gradient-Lipschitz and M_2-Hessian-Lipschitz. In view of (Dalalyan and Riou-Durand, 2020, Theorem 3), since the parameters $\alpha,\gamma,h>0$ are such that

$$\alpha \le \frac{M}{20},\qquad \gamma\ge\sqrt{M+2\alpha},\qquad h\le\frac{\alpha}{5\gamma(M+\alpha)}\bigwedge\frac{\alpha}{4M_2\sqrt{5p}},$$

the distribution of the KLMC2 sampler after k iterates satisfies

$$\begin{aligned}W_2(\nu_k,\pi_\alpha) &\le \sqrt{2}\,\mu_2(\pi_\alpha)\Big(1-\frac{\alpha h}{4\gamma}\Big)^k + \frac{2h^2M_2p}{\alpha} + \frac{h^2(M+\alpha)^{3/2}\sqrt{2p}}{\alpha}\\ &\qquad\qquad\qquad\qquad + \frac{8h(M+\alpha)}{\alpha}\exp\Big\{-\frac{\alpha^2}{160M_2^2h^2}\Big\}\\ &\le \sqrt{2}\,\mu_2(\pi_\alpha)\Big(1-\frac{\alpha h}{4\gamma}\Big)^K + \frac{2h^2(M_2p+M^{3/2}p^{1/2})}{\alpha} + \frac{1.6}{\sqrt{M}}\exp\Big\{-\frac{(\alpha/h)^2}{160M_2^2}\Big\},\end{aligned}$$

where the second inequality follows from the fact that $\alpha\le M/20$ and $h\le\alpha/(5\gamma(M+\alpha))$. The triangle inequality and the monotonicity of W_q with respect to q yields $W_q(\nu_K,\pi)\le W_2(\nu_K,\pi_\alpha)+W_q(\pi_\alpha,\pi)$, which leads to

$$W_q(\nu_K,\pi)\le\sqrt{2}\,\mu_2(\pi_\alpha)\Big(1-\frac{\alpha h}{4\gamma}\Big)^K + \frac{2h^2Qp}{\alpha} + \frac{1.6}{\sqrt{M}}\exp\Big\{-\frac{(\alpha/h)^2}{160M_2^2}\Big\} + W_q(\pi,\pi_\alpha).$$

Replacing the last term above by its upper bound provided by Proposition 1 and applying (6), we obtain the claim of the theorem. ■

8. Bounding the Moments

From the user's perspective, the choice of α and h requires the computation of the second moment of the distribution π. In most cases, this moment is an intractable integral. However, when some additional information on π is available, this moment can be replaced by a tractable upper bound. In this section, we provide upper bounds on the moments

$$\mu_a^* := \big(\mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta}-\boldsymbol{\theta}^*\|_2^a]\big)^{1/a},\qquad a\ge1,$$

centered at the minimizer of the potential $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{\theta})$. The knowledge of the second moment is enough to compute the mixing times presented in Section 6 and Section 7. However, providing bounds on general moments is of interest in order to highlight the dependence on the dimension, but also to obtain sharp numerical constants when the second moment is unknown. For instance, the proof of Proposition 1 shows that results for the W_1 and W_2 metrics essentially rely on some bounds over the third and fourth moments of π, which could be better understood in some specific contexts.

In this section, we investigate two particular classes of convex functions: (a) those which are m-strongly convex inside a ball of radius R around the mode $\boldsymbol{\theta}^*$, and (b) those which are m-strongly convex outside a ball of radius R around the mode $\boldsymbol{\theta}^*$. We provide user-friendly bounds on μ_a^* with fairly small constants. In the aforementioned two cases, if m and R are dimension free, we show that μ_a^* scales respectively as $p\log p$ and $(p\log p)^{1/2}$. This scaling with the dimension is sharp, up to logarithmic factors, and matches Condition 2 with $\beta = 2$ for the class (a) and Condition 2 with $\beta = 1$ for the class (b).

Proposition 2 *Assume that for some positive numbers m and R, we have $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \le R$. Then, for every $a \ge 2$, we have*

$$\mu_a^* \le A \vee B + \frac{3}{mR\,\Gamma(p/2)^{1/a}}$$

where[5]

$$A = \frac{3p}{mR}\bigg((1 + a/p)\log(1 + a/p) + \log_+\Big(\frac{2M}{m^2R^2}\Big)\bigg) \quad \textit{and} \quad B = \Big(\frac{p}{m}\Big)^{1/2}\Big\{2 + \frac{a}{2p}\Big\}^{\mathbb{1}_{a>2}}.$$

In the case of fixed a and p tending to infinity, Proposition 2 implies that

$$\mu_a^* = \widetilde{O}\bigg(\frac{p}{mR} \bigvee \Big(\frac{p}{m}\Big)^{1/2}\bigg),$$

where $\widetilde{O}$ hides constants and polylogarithmic factors. In the bound of Proposition 2, the term A is the dominating one when p/m is large as compared to R^2, while B is the dominating term when R^2 is of a higher order of magnitude than p/m. The residual term $3/(mR\Gamma(p/2)^{1/a})$ goes to zero whenever p or R tend to infinity. If m and R are assumed to be dimension free constants, then μ_a^* scales as $p\log p$. This rate is optimal within a poly-log factor, which is proven in Lemma 4. Note also that when R goes to infinity we exactly recover the bound of the strongly convex case proven in Lemma 2.

We now switch to bounding the moments of π under the condition that f is convex everywhere and strongly convex outside the ball of radius R around $\boldsymbol{\theta}^*$.

Proposition 3 *Assume that for some positive m and R, we have $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 > R$. If $p \ge 3$, then, for every $a > 0$, we have*

$$\mu_a^* \le \Big(1 + \frac{2}{\Gamma(p/2)}\Big)^{1/a}\bigg\{(4R) \bigvee \Big(\frac{4(p+a)}{m}\log\Big(\frac{pM}{m}\Big)\Big)^{1/2}\bigg\}.$$

[5] We denote by $\log_+ x$ the positive part of $\log x$, $\log_+ x = max(0, \log x)$.

Under the assumptions of Proposition 3, we obtain $\mu_a^* = \widetilde{O}\big(R \vee (p/m)^{1/2}\big)$. In the bound of Proposition 3, if m and R are assumed to be dimension free constants, then μ_a^* scales as $(p \log p)^{1/2}$. When R is not large, this rate is improved in Proposition 4 below to $p^{1/2}$, which is optimal. However, the bound of Proposition 3 is sharper when R is large.

Proposition 4 *Assume that for some positive m and R, we have $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 > R$. Then for every $a > 0$ we have*

$$\mu_a^* \leq e^{mR^2/2a} \left(\frac{p}{m}\right)^{1/2} \left\{2 + \frac{a}{2p}\right\}^{\mathbb{1}_{a>2}}.$$

Note that when R approaches zero, this bound matches the one of the strongly convex case; see, for instance, Lemma 2. To close this section, let us note that in the setting considered in Proposition 3 and 4, one can also apply the bounds obtained by reflection coupling (Majka et al., 2020; Cheng et al., 2018a). Quite surprisingly, they do not lead to better bounds than those obtained in the present work by the simple convexification trick.

9. Discussion

In this section we highlight some consequences of the bounds on the sampling error measured in Wasserstein distance and provide a discussion on how our results compare to those obtained by other relevant approaches.

9.1 Moment Approximation Bounds Berived from Bounds on W_q-Distances

A glance at the table of Section 4.2 is enough to notice that the reported sampling guarantees for the W_1-distance are much better than those for W_2. Are there situations where using sampling guarantees in W_1 distance is more suitable than W_2 so that we can take advantage of improved rates? The answer to this question is positive; it is formalized in the next proposition. In a nutshell, one can rely on guarantees in W_1-distance in the problem of approximating the expectation of a Lipschitz function of $\boldsymbol{\vartheta} \sim \pi$, while guarantees in W_2-distance are used for approximating the standard-deviation of Lipschitz functions.

Proposition 5 *Let π and ν be two probability distribution on $\mathbb{R}^p$; π is the target distribution while ν is the sampling distribution. Let $\varphi : \mathbb{R}^p \mapsto \mathbb{R}$ be a 1-Lipschitz function. For $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_N$ independently drawn from ν, define*

$$\widehat{m}_N(\varphi^q) = \frac{1}{N}\sum_{i=1}^N \varphi^q(\boldsymbol{\vartheta}_i), \qquad m_\pi(\varphi^q) = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\varphi^q(\boldsymbol{\vartheta})], \qquad m_\nu(\varphi^q) = \mathbf{E}_{\boldsymbol{\vartheta}\sim\nu}[\varphi^q(\boldsymbol{\vartheta})]$$

for any $q \in \mathbb{N}$. It holds that

$$\mathbf{E}\big[\big(\widehat{m}_N(\varphi) - m_\pi(\varphi)\big)^2\big]^{1/2} \leq W_1(\nu, \pi) + \sqrt{\frac{m_\nu(\varphi^2)}{N}},$$

$$\mathbf{E}\big[\big(\widehat{m}_N^{1/2}(\varphi^2) - m_\pi^{1/2}(\varphi^2)\big)^2\big]^{1/2} \leq W_2(\nu, \pi) + \sqrt{\frac{m_\nu(\varphi^4)}{N m_\nu(\varphi^2)}}.$$

Proof Using the bias-variance decomposition and the fact that $\mathbf{E}\big[\widehat{m}_N(\varphi)\big] = m_\nu(\varphi)$, one can check that

$$\begin{aligned}\mathbf{E}\big[\big(\widehat{m}_N(\varphi) - m_\pi(\varphi)\big)^2\big] &= \big(m_\nu(\varphi) - m_\pi(\varphi)\big)^2 + \mathbf{Var}\big[\widehat{m}_N(\varphi)\big]\\ &\le W_1^2(\nu,\pi) + \frac{\mathbf{Var}[\varphi(\boldsymbol{\vartheta}_1)]}{N},\end{aligned}$$

where the last inequality follows from the dual formulation of the Wasserstein distance. To complete the proof of the first claim, it suffices to note that $\mathbf{Var}[\varphi(\boldsymbol{\vartheta}_1)] \le m_\nu(\varphi^2)$.

For the second claim, we start by applying the triangle inequality

$$\mathbf{E}\big[\big(\widehat{m}_N^{1/2}(\varphi^2) - m_\pi^{1/2}(\varphi^2)\big)^2\big]^{1/2} \le \mathbf{E}\big[\big(\widehat{m}_N^{1/2}(\varphi^2) - m_\nu^{1/2}(\varphi^2)\big)^2\big]^{1/2} + \big|m_\nu^{1/2}(\varphi^2) - m_\pi^{1/2}(\varphi^2)\big|.$$

On the one hand, for any $\boldsymbol{\vartheta} \sim \nu$ and $\boldsymbol{\vartheta}' \sim \pi$, we have

$$\begin{aligned}\big|m_\nu^{1/2}(\varphi^2) - m_\pi^{1/2}(\varphi^2)\big| &= \big|\mathbf{E}^{1/2}[\varphi^2(\boldsymbol{\vartheta})] - \mathbf{E}^{1/2}[\varphi^2(\boldsymbol{\vartheta}')]\big|\\ &\le \mathbf{E}^{1/2}\big[\big(\varphi(\boldsymbol{\vartheta}) - \varphi(\boldsymbol{\vartheta}')\big)^2\big]\\ &\le \mathbf{E}^{1/2}\big[\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2^2\big].\end{aligned}$$

Since this is true for any coupling $(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}')$ of ν and π, and the infimum of the right hand side of the last display over all the couplings is the W_2 distance, we get

$$\big|m_\nu^{1/2}(\varphi^2) - m_\pi^{1/2}(\varphi^2)\big| \le W_2(\nu,\pi).$$

On the other hand,

$$\mathbf{E}\big[\big(\widehat{m}_N^{1/2}(\varphi^2) - m_\nu^{1/2}(\varphi^2)\big)^2\big] \le \frac{\mathbf{E}\big[\big(\widehat{m}_N(\varphi^2) - m_\nu(\varphi^2)\big)^2\big]}{m_\nu(\varphi^2)} = \frac{\mathbf{Var}\big[\varphi^2(\boldsymbol{\vartheta}_1)\big]}{N m_\nu(\varphi^2)}.$$

To complete the proof of the second claim, it suffices to note that $\mathbf{Var}[\varphi^2(\boldsymbol{\vartheta}_1)] \le m_\nu(\varphi^4)$. ■

The simplest application of the last result is when $\varphi(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{v}$ with a unit vector $\boldsymbol{v}$. From the last proposition, we infer that the error of approximating the mean of the target distribution by the sample mean of an N-sample drawn from the sampling distribution is controlled by the W_1 distance plus a small term of order $N^{-1/2}$. If, in addition to estimating the mean, we also want to estimate the second-order moment, then the approximation error is bounded by the W_2 distance plus a small term of order $N^{-1/2}$. Thus, the fact that the guarantees for W_1 are better than those for W_2 illustrates that it is computationally less expensive to approximate the first-order moment rather than the second-order moment. Furthermore, the rates reported in the table of Section 4.2 provide a quantitiave assessment of the computational gain.

9.2 Suboptimality of Wasserstein Bounds Derived from Total-Variation Bounds

In Section 4.2, we presented a summary of rates for the LMC with a quadratic penalty in Wasserstein distance obtained in this manuscript, and compared them to existing TV mixing times Dalalyan (2017b); Durmus et al. (2019). Since the topology induced by the total-variation distance is stronger than the topology of weak convergence induced by the Wasserstein distance, one can wonder whether existing results for TV-distance may directly lead to mixing times for Wasserstein distances. If the answer to this question is positive, the next question is how do they compare to the results obtained in this manuscript. The following proposition and the subsequent discussion aim to clarify this point.

Proposition 6 *Let ν and ν' be arbitrary probability distributions on $\mathbb{R}^p$. For every $q, r, s \geq 1$ such that $1/r + 1/s = 1$, we have*

$$W_q(\nu, \nu') \leq \big(\mu_{qr}(\nu) + \mu_{qr}(\nu')\big)\, d_{\rm TV}(\nu, \nu')^{1/(qs)}.$$

Proof The proof follows from the definition of the Wasserstein and the total variation distances in terms of optimal couplings. Let $\Gamma(\nu, \nu')$ be the set of joint distributions on $\mathbb{R}^p \times \mathbb{R}^p$ with marginals ν and ν' and let $q, r, s \geq 1$ such that $1/r + 1/s = 1$. Choose an arbitrary coupling $\gamma \in \Gamma(\nu, \nu')$. Applying successively Hölder's and Minkowski's inequalities, we arrive at

$$\begin{aligned}
W_q^q(\nu, \nu') &\leq \mathbf{E}_{(X,Y)\sim\gamma}\left[\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2^q \mathbb{1}_{\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}'}\right]\\
&\leq \mathbf{E}_\gamma[\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2^{qr}]^{1/r}\mathbf{P}_\gamma(\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}')^{1/s}\\
&\leq \big(\mu_{qr}(\nu) + \mu_{qr}(\nu')\big)^q\, \mathbf{P}_\gamma(\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}')^{1/s}.
\end{aligned}$$

Choosing as γ the coupling that minimizes $\mathbf{P}_\gamma\left(\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}'\right)$, we get the claim of the proposition. ■

Proposition 6, combined with the available bounds on the total variation distance, can be used to derive bounds on the Wasserstein error of LMCa or α-LMC. In particular, for $\nu = \pi$ and $\nu' = \nu_k^{\rm LMCa}$, one can infer from Proposition 6 that

$$d_{\rm TV}(\pi, \nu_k^{\rm LMCa}) \leq \left(\frac{\varepsilon\mu_2(\pi)}{\mu_{qr}(\pi) + \mu_{qr}(\nu_k^{\rm LMCa})}\right)^{qs} \qquad \Longrightarrow \qquad W_q(\pi, \nu_k^{\rm LMCa}) \leq \varepsilon\mu_2(\pi),$$

for any $q, r, s \geq 1$ such that $1/r + 1/s = 1$. As mentioned in Section 2, just after Condition 2, there is a constant B_{qr} such that $\mu_{qr}(\pi) \leq B_{qr}\mu_2(\pi)$ for every log-concave distribution π. Even though the constant B_{qr} does not depend on the target distribution, it blows up whenever $r \to \infty$ (known upper bounds on this constant are at least linear in r; see, for instance, Lemma 5). Therefore, assuming that there is a constant $C > 0$ such that $\mu_{qr}(\nu_k^{\rm LMCa}) \leq C\mu_{qr}(\pi)$, we get

$$d_{\rm TV}(\pi, \nu_k^{\rm LMCa}) \leq \sup_{r>1}\left(\frac{\varepsilon}{(1+C)B_{qr}}\right)^{qr/(r-1)} \qquad \Longrightarrow \qquad W_q(\pi, \nu_k^{\rm LMCa}) \leq \varepsilon\mu_2(\pi)$$

for any $q \geq 1$. If we neglect that the constant B_{qr} blows up, the method sketched above leads to the best scalings with respect to ε when $r \to +\infty$. In concrete examples, however, deriving from the last display the sharpest bound on the W_q-mixing-time for a fixed precision level $\varepsilon > 0$ would require a non-trivial optimization with respect to r.

Even if we disregard the fact that B_{qr} is not bounded and if we admit that $\mu_{qr}(\nu_k^{\rm LMCa}) \leq C\mu_{qr}(\pi)$, it turns out that the outlined method will lead to looser W_q-mixing rate than the one obtained by the direct approach developed in this paper. Indeed, choosing $r = (1+\eta)/\eta$ for some $\eta > 0$ and taking into account that the TV-mixing-time for the LMCa obtained in (Durmus et al., 2019) is of the order $O(p^2/\varepsilon^4)$, we get a W_q-mixing-time of order $O(p^2/\varepsilon^{4q(1+\eta)})$. For $q = 1, 2$, these rates are worse than the rates p^2/ε^{2q+2} reported in the table of Section 4.2 for the α-LMC algorithm.

The take away message is that the direct method of proof used in Theorem 1-4 allows for a better control of W_q distances than the combination of existing results for the TV-distance with Proposition 6.

9.3 Convexification Versus Reflection Coupling

When the potential f is strongly convex outside a ball of radius R, as in Proposition 3, an alternative to the approach developed in the present paper is to apply the results obtained by reflection coupling (Majka et al., 2020; Cheng et al., 2018a) under more general dissipativity assumption. We can thus compare our results with those of these papers. It turns out that in the natural setting of large R our results provide much tighter bounds than those by reflection coupling.

Indeed, let us compare Theorem 1 for $q = 1, 2$ with equations (2.28) and (2.29), Theorem 2.9, from (Majka et al., 2020). In our results the geometric contraction takes place at the rate $e^{-\alpha hK/2}$, this rate is e^{-chK} with a parameter c exponentially small in R, $c = O(e^{-aMR^2})$ for some $a > 0$. The choice of the parameter h also involves a term which is exponentially small in MR^2. The same factor e^{aMR^2} appears in (Cheng et al., 2018a). This implies that the number of gradient evaluations to achieve ε-accuracy, derived from (Majka et al., 2020; Cheng et al., 2018a), is exponential in MR^2. In our results, derived from Theorem 1 and Proposition 3, this dependence is polynomial. In a high-dimensional setting, the parameter MR^2 will most likely be polynomial in p, leading thus to a substantial improvement obtained by our results as compared to those inferred from the reflection coupling.

9.4 Centering the Target Distribution

Although the theorems stated in the previous section apply to general loc-concave distributions π having nonzero density on the whole $\mathbb{R}^p$, they are meaningful when the "center of the distribution" is close to the origin. This has been quickly mentioned in Section 2, just before Condition 2; we provide below more details on the meaning and the potential cost of centering.

Clearly, if the density $\pi(\cdot)$ is log-concave with gradient-Lipschitz potential f, then the same is true for $\pi_{\boldsymbol{\theta}_0}(\cdot) = \pi(\cdot + \boldsymbol{\theta}_0)$, whatever the value $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is. The only quantity appearing in our upper bounds that is impacted by such a transformation of π is the second-order moment $\mu_2^2(\pi_{\theta_0}) = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\|\boldsymbol{\vartheta} - \boldsymbol{\theta}_0\|_2^2]$. Ideally, we would like to choose $\boldsymbol{\theta}_0$ as the minimizer of $\mu_2(\pi_{\theta_0})$, which corresponds to $\boldsymbol{\theta}_0 = \mathbf{E}_{\boldsymbol{\vartheta}\sim\pi}[\boldsymbol{\vartheta}]$. This value, unfortunately, is rarely available. Instead, one can choose as $\boldsymbol{\theta}_0$ any minimizer of f. In view of Proposition 2 and Proposition 3, the second-order moment of the resulting centered distribution has suitably bounded moments. It should be noted, however, that computing a minimizer of the convex function f requires $O(1/\varepsilon^2)$ gradient calls. Another approach that can be adopted in a statistical setting—where π is a posterior distribution—consists in choosing $\boldsymbol{\theta}_0$ as an initial estimator of the true parameter. It can be, for instance, based on the method of moments.

Acknowledgments

This work was partially supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047). LRD was partially supported by the EPSRC grant EP/R034710/1.

Appendix A. Postponed Proofs

This section contains proofs of the propositions stated in previous sections as well as those of some technical lemmas used in the proofs of the propositions.

A.1 Proof of Proposition 1

Without loss of generality we may assume that $\int_{\mathbb{R}^p} \exp(-f(\boldsymbol{\theta}))\,d\boldsymbol{\theta} = 1$. We first derive upper and lower bounds for the normalizing constant of π_α, that is

$$c_\alpha := \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta})\, e^{-\alpha\|\boldsymbol{\theta}\|_2^2/2}\, d\boldsymbol{\theta}.$$

To do so, we introduce the notation

$$r_\alpha := \frac{2}{\alpha}\log\frac{1}{c_\alpha}$$

so that $\log(\pi_\alpha/\pi)(\boldsymbol{\theta}) = (\alpha/2)(r_\alpha - \|\boldsymbol{\theta}\|_2^2)$. One can check that $c_\alpha \le 1$. To get a lower bound, we note that c_α is an expectation with respect to the density π, hence it can be lower bounded using Jensen's inequality, applied to the convex map $x \mapsto e^{-x}$. These two facts yield $\exp\{-\alpha\mu_2^2/2\} \le c_\alpha \le 1$. Therefore, by definition of r_α, we have

$$0 \le r_\alpha \le \mu_2^2. \tag{7}$$

For any fixed $\boldsymbol{\theta} \in \mathbb{R}^p$, we now split the Euclidean distance between $\pi(\boldsymbol{\theta})$ and $\pi_\alpha(\boldsymbol{\theta})$ between its positive and negative parts:

$$|\pi(\boldsymbol{\theta}) - \pi_\alpha(\boldsymbol{\theta})| = \underbrace{\pi(\boldsymbol{\theta})\left[1 - e^{-(\alpha/2)(\|\boldsymbol{\theta}\|_2^2 - r_\alpha)}\right]\mathbb{1}_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha}}_{:=(\pi-\pi_\alpha)_+(\boldsymbol{\theta})} + \underbrace{\pi(\boldsymbol{\theta})\left[e^{-(\alpha/2)(r_\alpha - \|\boldsymbol{\theta}\|_2^2)} - 1\right]\mathbb{1}_{\|\boldsymbol{\theta}\|_2^2 < r_\alpha}}_{:=(\pi-\pi_\alpha)_-(\boldsymbol{\theta})}.$$

In order to bound the positive part, we make use of the inequality $1 - e^{-x} \le x$ for $x > 0$. Therefore:

$$(\pi - \pi_\alpha)_+(\boldsymbol{\theta}) \le \frac{\alpha}{2}\pi(\boldsymbol{\theta})(\|\boldsymbol{\theta}\|_2^2 - r_\alpha)\mathbb{1}_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha}. \tag{8}$$

The total variation distance between densities π and π_α is twice the integral of the positive part, $d_{\rm TV}(\pi_\alpha, \pi) = 2\int_{\mathbb{R}^p}(\pi - \pi_\alpha)_+(\boldsymbol{\theta})\,d\boldsymbol{\theta}$. Therefore,

$$d_{\rm TV}(\pi_\alpha, \pi) \le \alpha \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta})(\|\boldsymbol{\theta}\|_2^2 - r_\alpha)\mathbb{1}_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha} d\boldsymbol{\theta} \le \alpha \int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^2 \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

This yields the first claim of the proposition.

The proof of the bound for Wasserstein distances is inspired by the arguments from (Villani, 2008, Theorem 6.15, page 115). We consider a suitable coupling between π and π_α, defined by keeping fixed the mass shared by π and π_α while distributing the rest of the mass with a product measure. Letting $C := (\pi - \pi_\alpha)_+(\mathbb{R}^p) = (\pi-\pi_\alpha)_-(\mathbb{R}^p)$, we define the joint distribution

$$\gamma(d\boldsymbol{\theta}, d\boldsymbol{\theta}') := (\pi \wedge \pi_\alpha)(d\boldsymbol{\theta})\delta_{\boldsymbol{\theta}'=\boldsymbol{\theta}} + \frac{1}{C}(\pi-\pi_\alpha)_+(d\boldsymbol{\theta})(\pi-\pi_\alpha)_-(d\boldsymbol{\theta}').$$

The joint distribution γ defines a coupling of π and π_α. Therefore for any $q \geq 1$, by definition of the Wasserstein distance we get

$$\begin{aligned}
W_q^q(\mu,\nu) &\leq \int_{\mathbb{R}^p\times\mathbb{R}^p} \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^q \gamma(d\boldsymbol{\theta}, d\boldsymbol{\theta}')\\
&= \frac{1}{C}\int_{\mathbb{R}^p\times\mathbb{R}^p} \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^q (\pi-\pi_\alpha)_+(d\boldsymbol{\theta})(\pi-\pi_\alpha)_-(d\boldsymbol{\theta}')\\
&\leq \frac{1}{C}\int_{\mathbb{R}^p\times\mathbb{R}^p} (\|\boldsymbol{\theta}\|_2 + \sqrt{r_\alpha})^q\,(\pi-\pi_\alpha)_+(d\boldsymbol{\theta})(\pi-\pi_\alpha)_-(d\boldsymbol{\theta}')\\
&= \int_{\mathbb{R}^p} (\|\boldsymbol{\theta}\|_2 + \sqrt{r_\alpha})^q\,(\pi-\pi_\alpha)_+(d\boldsymbol{\theta})
\end{aligned}$$

where the third line follows from the fact that $(\pi-\pi_\alpha)_-(d\boldsymbol{\theta}')$ has positive mass only inside the ball $\{\|\boldsymbol{\theta}'\|_2 \leq \sqrt{r_\alpha}\}$. We now define the quantity

$$J_{q,\alpha}(\pi) := \frac{1}{2}\int_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha} (\|\boldsymbol{\theta}\|_2 + \sqrt{r_\alpha})^q \left(\|\boldsymbol{\theta}\|_2^2 - r_\alpha\right)\pi(d\boldsymbol{\theta})$$

and remark that inequality (8) yields

$$W_q^q(\mu,\nu) \leq \alpha J_{q,\alpha}(\pi).$$

The claim of the proposition follows from the fact that there is a numerical constant C_q that only depends on q such that

$$J_{q,\alpha}(\pi) \leq C_q \mu_2^{q+2}(\pi).$$

This is a combined consequence of (7) and Lemma 5. Indeed, we have

$$J_{q,\alpha}(\pi) \leq \frac{1}{2}\int_{\|\boldsymbol{\theta}\|_2^2 > r_\alpha} (2\|\boldsymbol{\theta}\|_2)^q \|\boldsymbol{\theta}\|_2^2 \pi(d\boldsymbol{\theta}) \leq 2^{q-1}\mu_{q+2}^{q+2}(\pi) \leq 2^{q-1}(B_{q+2}\mu_2(\pi))^{q+2}.$$

As shown below, we can get better values for C_q when $q=1$ and $q=2$. We have

$$J_{1,\alpha}(\pi) \leq \frac{1}{2}\int_{\mathbb{R}^p} (\|\boldsymbol{\theta}\|_2 + \mu_2)\|\boldsymbol{\theta}\|_2^2\,\pi(d\boldsymbol{\theta}) = (\mu_3^3 + \mu_2^3)/2 \leq 21\mu_2^3$$

and

$$\begin{aligned}
J_{2,\alpha}(\pi) &= \frac{1}{2}\int_{\|\boldsymbol{\theta}\|_2^2>r_\alpha} (\|\boldsymbol{\theta}\|_2 + \sqrt{r_\alpha})^2(\|\boldsymbol{\theta}\|_2^2 - r_\alpha)\pi(d\boldsymbol{\theta}) \\
&\le \frac{1}{2}\int_{\|\boldsymbol{\theta}\|_2^2>r_\alpha} (\|\boldsymbol{\theta}\|_2^4 + 2\|\boldsymbol{\theta}\|_2^3\sqrt{r_\alpha})\pi(d\boldsymbol{\theta}) \\
&\le \frac{1}{2}\int_{\mathbb{R}^p} (\|\boldsymbol{\theta}\|_2^4 + 2\|\boldsymbol{\theta}\|_2^3\mu_2)\pi(d\boldsymbol{\theta}) = (\mu_4^4 + 2\mu_3^3\mu_2)/2 \le 262\mu_2^4.
\end{aligned}$$

In both calculations, inequality (7) is used to bound r_α, while the last inequality follows from Lemma 5. It turns out that in the particular cases $q = 1$ and $q = 2$, the constant C_q can be further improved using, respectively, (Bolley and Villani, 2005, Corollary 4) and the transportation cost inequality; see for instance (Gozlan and Léonard, 2010, Corollary 7.2). Since π_α is α-strongly log-concave, we have

$$W_1^2(\pi, \pi_\alpha) \le 2\mu_2^2(\pi) D_{\mathrm{KL}}(\pi||\pi_\alpha), \qquad W_2^2(\pi, \pi_\alpha) \le (2/\alpha) D_{\mathrm{KL}}(\pi||\pi_\alpha).$$

The computation of the Kullback-Leibler divergence yields

$$D_{\mathrm{KL}}(\pi||\pi_\alpha) = \int_{\mathbb{R}^p} \pi(\boldsymbol{\theta})(\alpha/2)(\|\boldsymbol{\theta}\|_2^2 - r_\alpha)\, d\boldsymbol{\theta} = \alpha\mu_2^2/2 + \log c_\alpha.$$

Using the inequality $e^{-x} \le 1 - x + x^2/2$ for $x > 0$ yields

$$c_\alpha = \int_{\mathbb{R}^{\shortmid}} \pi(\boldsymbol{\theta}) e^{-(\alpha/2)\|\boldsymbol{\theta}\|_2^2} d\boldsymbol{\theta} \le 1 - \alpha\mu_2^2/2 + \alpha^2\mu_4^4/8.$$

Since $\log(1+x) \le x$ for $x > -1$ we get $D_{\mathrm{KL}}(\pi||\pi_\alpha) \le \alpha^2\mu_4^4/8$. Combining this inequality with the bound on μ_4 from Lemma 5, we get

$$W_1(\pi, \pi_\alpha) \le \alpha\mu_2\mu_4^2/2 \le 11\alpha\mu_2^3, \qquad W_2^2(\pi, \pi_\alpha) \le \alpha\mu_4^4/4 \le 111\alpha\mu_2^4.$$

This shows that for $q = 1$ and $q = 2$ the constants can be improved to $C_1 = 11$ and $C_2 = 111$. Therefore we get the claim of the proposition.

A.2 Proof of Proposition 2

We assume without loss of generality that $\boldsymbol{\theta}^* = \mathbf{0}_p$. Let $A \ge R$ and $a > 0$. Define $B_A = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\|_2 \le A\}$. We split the integral into two parts

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int_{B_A} \|\boldsymbol{\theta}\|_2^a\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} + \int_{B_A^c} \|\boldsymbol{\theta}\|_2^a\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Let us bound the integral over B_A^c. It follows from the assumptions of the proposition that for any $\boldsymbol{\theta} \in \mathbb{R}^p$, $\nabla^2 f(\boldsymbol{\theta}) \succeq m(\|\boldsymbol{\theta}\|_2)\mathbf{I}_p$, for the map $m(r) := m\mathbb{1}_{(0,R)}(r)$. One can show that (see Lemma 3 for the precise statement and the proof)

$$\int_{B_A^c} \|\boldsymbol{\theta}\|_2^a\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \le \frac{2(M/2)^{p/2}}{\Gamma(p/2)} \int_A^{+\infty} r^{p+a-1} e^{-\widetilde{m}(r) r^2/2} dr, \tag{9}$$

where $\widetilde{m}(r)\ := 2\int_0^1 (1-t)\, m(rt)\, dt$. Using the fact that $m(r) := m\mathbb{1}_{(0,R)}(r)$, for all $r > R$, we get

$$\widetilde{m}(r) = 2m\int_0^{1\wedge R/r}(1-t)dt = m\left(\frac{2R}{r} - \frac{R^2}{r^2}\right).$$

Since in (9) the integration is with respect to $r \geq A \geq R$, we have

$$\begin{aligned}\int_{B_A^c} \|\boldsymbol{\theta}\|_2^a\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} &\leq \frac{2(M/2)^{p/2}}{\Gamma(p/2)} e^{mR^2/2}\int_A^{+\infty} r^{p+a-1}e^{-mRr}dr\\ &= \frac{2(M/2)^{p/2}}{\Gamma(p/2)}\,\frac{e^{mR^2/2}}{(mR)^{a+p}}\int_{mRA}^{+\infty} y^{p+a-1}\, e^{-y}\, dy.\end{aligned}$$

We now use the following inequality on the incomplete Gamma function from (Natalini and Palumbo, 2000), see also Borwein et al. (2009): For all $q \geq 1$ and all $x \geq 2(q-1)$,

$$\int_x^{+\infty} y^{q-1}e^{-y}dy \leq 2x^{q-1}e^{-x}. \tag{10}$$

We apply this inequality for $q = p + a$. For $A \geq 2(p+a-1)/(mR)$, we have

$$\int_{mRA}^{+\infty} y^{p+a-1}\, e^{-y}\, dy \leq 2(mRA)^{p+a-1}e^{-mRA}.$$

Now, we make use of the fact that $mRA \geq mRA/2$. The latter yields

$$\int_{B_A^c} \|\boldsymbol{\theta}\|_2^a\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \leq \frac{2^{a+1}}{(mR)^a\Gamma(p/2)}\left(\frac{2M}{m^2R^2}\right)^{p/2}\left(\frac{mRA}{2}\right)^{p+a-1} e^{-mRA/2}.$$

The last bound ensures that the inequality

$$\int_{B_A^c} \|\boldsymbol{\theta}\|_2^a\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \leq \frac{2^{a+1}}{(mR)^a\Gamma(p/2)} \tag{11}$$

is fulfilled whenever $\varphi(x) := x - c\log(x) - b \geq 0$, where

$$x = \frac{mRA}{2}, \qquad c = p + a - 1, \qquad b = \frac{p}{2}\log\left(\frac{2M}{m^2R^2}\right).$$

We now establish for which values of x (or equivalently, A) we have $\varphi(x) \geq 0$. Taylor's expansion around $y_c := 1.5(c+1)\log(c+1)$ yields

$$\varphi\big(y_c + 3b_+\big) = \varphi(y_c) + \varphi'(y) \times 3b_+$$

for some $y \geq y_c$. The latter implies that

$$\varphi'(y) = 1 - \frac{c}{y} \geq 1 - \frac{c}{y_c} \geq 1/3.$$

Hence, $\varphi\big(y_c+3b_+\big)\geq\varphi(y_c)+b_+\geq y_c-c\log y_c+b_+-b\geq 0$. Since the map φ is increasing on $[c,+\infty)$ and $y_c+3b_+\geq c$, we conclude that (11) is fulfilled for any

$$A\geq A_0:=\frac{3}{mR}\left((p+a)\log(p+a)+p\log_+\left(\frac{2M}{m^2R^2}\right)\right).$$

We choose $A=A_0\vee R$. If $R<A_0$, we have $A=A_0$ and we use the obvious inequality

$$\int_{B_{A_0}}\|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}\leq A_0^a.$$

The second case to consider is $R\geq A_0$. The map $f(\boldsymbol{\theta})=-\log\pi(\boldsymbol{\theta})$ being m-strongly convex on the ball $B_A=B_R$, Lemma 2 yields

$$\bigg(\int_{B_A}\|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}\bigg)^{1/a}\leq\Big(\frac{p}{m}\Big)^{1/2}\bigg\{2+\frac{a}{2p}\bigg\}^{\mathbb{1}_{a>2}}.$$

Since inequality (11) is valid in both cases, the claim of Proposition 2 follows.

A.3 Proof of Proposition 3

Note that for any $\boldsymbol{\theta}\in\mathbb{R}^p$, $\nabla^2 f(\boldsymbol{\theta})\succeq m(\|\boldsymbol{\theta}\|_2)\mathbf{I}_p$, where $m(\cdot)$ is defined as below:

$$m(r)=m\mathbb{1}_{(R,+\infty)}(r).$$

We begin by computing the map $\widetilde{m}(r):=2\int_0^1 m(ry)(1-y)dy$. Using the definition of $\widetilde{m}$, we have:

$$\begin{aligned}\widetilde{m}(r)&=2\int_0^1 m\mathbb{1}_{(R,+\infty)}(ry)(1-y)dy\\&=2m\mathbb{1}_{r>R}\int_{R/r}^1(1-y)dy\\&=m\,(1-R/r)^2\,\mathbb{1}_{r>R}.\end{aligned}$$

Let $A\geq 4R$ and $a>0$. We assume without loss of generality that $\boldsymbol{\theta}^*=\mathbf{0}_p$. Define $B_A=\{\boldsymbol{\theta}\in\mathbb{R}^p:\|\boldsymbol{\theta}\|_2\leq A\}$. We will use the following bound:

$$\int_{\mathbb{R}^p}\|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}\leq A^a+\int_{B_A^c}\|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

For the second term, Lemma 3 yields

$$\int_{B_A^c}\|\boldsymbol{\theta}\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}\leq\frac{2(M/2)^{p/2}}{\Gamma(p/2)}\int_A^{+\infty}r^{p+a-1}e^{-mr^2/8}dr.$$

This is true due to the fact that, for every $r \geq A \geq 4R$, we have $\widetilde{m}(r) \geq m/2$. We now use inequality (10) with $B = 2$, $q = (p+a)/2$ and $mA^2/4 \geq (p+a) - 1/2$:

$$\begin{aligned}\int_A^{+\infty} r^{p+a-1} e^{-mr^2/8} dr &= 2^{-1}\left(\frac{4}{m}\right)^{(p+a)/2} \int_{mA^2/4}^{+\infty} y^{(p+a)/2-1} e^{-y} dy \\ &\leq \left(\frac{4}{m}\right)^{(p+a)/2} \left(\frac{mA^2}{4}\right)^{(p+a)/2-1} e^{-mA^2/4} \\ &= A^a \left(\frac{4}{m}\right)^{p/2} \left(\frac{mA^2}{4}\right)^{p/2-1} e^{-mA^2/4}.\end{aligned}$$

This yields

$$\int_{B_A^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \leq \frac{2A^a}{\Gamma(p/2)} \left(\frac{2M}{m}\right)^{p/2} \left(\frac{mA^2}{4}\right)^{p/2-1} e^{-mA^2/4}.$$

The last bound ensures that the inequality

$$\int_{B_A^c} \|\boldsymbol{\theta}\|_2^a \, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \leq \frac{2A^a}{\Gamma(p/2)} \tag{12}$$

is fulfilled whenever $\varphi(x) := x - c\log(x) - b \geq 0$, where

$$x = \frac{mA^2}{4}, \qquad c = \frac{p}{2} - 1, \qquad b = \frac{p}{2}\log\left(\frac{2M}{m}\right) > 0.$$

Taylor's expansion around $y_c := 2(c+1)\log(c+1)$ yields

$$\varphi(y_c + 2b) = \varphi(y_c) + \varphi^{'}(y) \times 2b$$

for some $y \geq y_c$. The latter implies that

$$\varphi^{'}(y) = 1 - \frac{c}{y} \geq 1 - \frac{c}{y_c} \geq 1/3.$$

We get $\varphi(y_c + 2b) \geq y_c - c\log(y_c) + b - b \geq 0$. Since the map φ is increasing on $[c, +\infty)$ and $y_c + 2b \geq c$, we conclude that (12) is fulfilled for any

$$A^2 \geq \frac{4}{m}\left(p\log(p/2) + p\log(2M/m)\right) = \frac{4p}{m}\log\left(\frac{pM}{m}\right).$$

Finally, we choose A such that this inequality and the two additional assumptions: $A \geq 2R$ and $mA^2/4 \geq (p+a) - 1/2$ hold. If $p \geq 3$ we can choose

$$A = (4R) \bigvee \left(\frac{4(p+a)}{m}\log\left(\frac{pM}{m}\right)\right)^{1/2}.$$

This yields the claim of Proposition 3.

A.4 Proof of Proposition 4

Define $f = -\log \pi$ and for any $\boldsymbol{\theta} \in \mathbb{R}^p$:

$$\bar{f}(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + \frac{m}{2} \left(\|\boldsymbol{\theta}\|_2 - R\right)^2 \mathbb{1}_{\|\boldsymbol{\theta}\|_2 \leq R}.$$

For any $\boldsymbol{\theta} \in \mathbb{R}^p$, we have $\bar{f}(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}) + mR^2/2$, this yields

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq e^{mR^2/2} \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a e^{-\bar{f}(\boldsymbol{\theta})} d\boldsymbol{\theta}.$$

Now we define the normalising constant

$$\bar{C} := \int_{\mathbb{R}^p} e^{-\bar{f}(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

and the corresponding probability density $\bar{\pi}(\boldsymbol{\theta}) := e^{-\bar{f}(\boldsymbol{\theta})}/\bar{C}$. The constant $\bar{C} \leq 1$ since $f(\boldsymbol{\theta}) \leq \bar{f}(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$. Therefore we have

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq e^{mR^2/2} \int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a \, \bar{\pi}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}.$$

By construction the density $\bar{\pi}$ is m-strongly log-concave. We apply Lemma 2 on this last term and get the claim of Proposition 4.

A.5 Technical Lemmas

Lemma 1 *Suppose that π has a finite fourth-order moment. Then $\alpha \mapsto \mu_2(\pi_\alpha)$ is continuously differentiable and non-increasing, when $\alpha \in [0, +\infty)$.*

Proof For $k \in \mathbb{N} \cup \{0\}$, define

$$h_k(\alpha) = \int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta}) - \alpha \|\boldsymbol{\theta}\|_2^2/2\right) d\boldsymbol{\theta}.$$

If $\pi \in \mathcal{P}_k(\mathbb{R}^p)$ then the function h_k is continuous on $[0; +\infty)$. Indeed, if the sequence $\{\alpha_n\}_n$ converges α_0, when $n \to +\infty$, then the function $\|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta}) - ({}^1\!/\!_2)\alpha_n \|\boldsymbol{\theta}\|_2^2\right)$ is upper-bounded by $\|\boldsymbol{\theta}\|_2^k \exp\left(-f(\boldsymbol{\theta})\right)$. Thus in view of the dominated convergence theorem, we can interchange the limit and the integral. Since, by definition,

$$\mu_k^k(\pi_\alpha) = \frac{h_k(\alpha)}{h_0(\alpha)},$$

we get the continuity of $\mu_2(\pi_\alpha)$ and $\mu_4(\pi_\alpha)$. Let us now prove that $h_k(t)$ is continuously differentiable, when $\pi \in \mathcal{P}_{k+2}(\mathbb{R}^p)$. The integrand function in the definition of h_k is a continuously differentiable function with respect to t. In addition, its derivative is continuous and is as well integrable on $\mathbb{R}^p$, as we supposed that π has the $(k+2)$-th moment. Therefore, the Leibniz integral rule yields the following

$$h_k'(\alpha) = -\frac{1}{2} \int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^{k+2} \exp\left(-f(\boldsymbol{\theta}) - \alpha \|\boldsymbol{\theta}\|_2^2/2\right) d\boldsymbol{\theta} = -\frac{1}{2} h_{k+2}(t).$$

The latter yields the smoothness of h_k. Finally, in order to prove the monotonicity of $\mu_2^2(\pi_\alpha)$, we will simply compute its derivative

$$\begin{aligned}\big(\mu_2^2(\pi_\alpha)\big)' &= -\frac{1}{2h_0(\alpha)}h_4(\alpha) - \frac{h_0'(\alpha)}{h_0(\alpha)^2}h_2(\alpha)\\ &= -\frac{1}{2}\mu_4^4(\pi_\alpha) + \frac{h_2^2(\alpha)}{2h_0(\alpha)^2}\\ &= \frac{1}{2}\left(\mu_2^4(\pi_\alpha) - \mu_4^4(\pi_\alpha)\right).\end{aligned}$$

Since the latter is always negative, this completes the proof of the lemma. ■

Lemma 2 *Let $a > 0$ and $m > 0$. Assume $f = -\log\pi$ is m-strongly convex. Then*

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} \le \Big(\frac{p}{m}\Big)^{1/2}(2 + a/2p)^{\mathbb{1}_{a>2}}.$$

Proof In view of Durmus and Moulines (2019),

$$\int_{\mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} \le \frac{p}{m}.$$

The monotonicity of the $\mathbb{L}_a$-norm directly yields the claim of the lemma for $a \le 2$.

In the case $a > 2$, we use Theorem 1 from Hargé (2004). The result is formulated as follows. Assume that $X \sim \mathcal{N}_p(\mu, \Sigma)$ with density φ and Y with density $\varphi\cdot\psi$ where ψ is a log-concave function. Then for any convex map $g : \mathbb{R}^p \mapsto \mathbb{R}$ we have

$$\mathbf{E}[g(Y - \mathbf{E}[Y])] \le \mathbf{E}[g(X - \mathbf{E}[X])].$$

Since $f = -\log\pi$ is m-strongly convex, the particular choice $\mu = \mathbf{0}_p$ and $\Sigma = m\mathbf{I}_p$ yields the log-concavity of π/φ. Applied to the convex map $g : \boldsymbol{\theta} \mapsto \|\boldsymbol{\theta}\|_2^a$, the inequality of Hargé (2004) yields

$$\mathbf{E}_\pi[\|\boldsymbol{\vartheta} - \mathbf{E}_\pi[\boldsymbol{\vartheta}]\|_2^a] \le \mathbf{E}[\|X\|_2^a] = \Big(\frac{p}{m}\Big)^{a/2}\frac{\Gamma((p+a)/2)}{\Gamma(p/2)(p/2)^{a/2}}$$

using known moments of the chi-square distribution.

For any $y > 0$ the map $x \mapsto x^{-y}\Gamma(x+y)/\Gamma(x)$ goes to 1 when x goes to infinity. For convenience, we use an explicit bound from (Qi et al., 2012, Theorem 4.3), that is

$$\forall y \ge 1, \qquad x^{-y}\Gamma(x+y)/\Gamma(x) \le (1 + y/x)^{y-1}.$$

When applied to $x = p/2$ and $y = a/2 > 1$, this yields

$$\mathbf{E}_\pi[\|\boldsymbol{\vartheta} - \mathbf{E}_\pi[\boldsymbol{\vartheta}]\|_2^a] \le \Big(\frac{p}{m}\Big)^{a/2}(1 + a/p)^{a/2-1}. \tag{13}$$

We now bound the distance between the mean and the mode

$$\|\mathbf{E}_\pi[\boldsymbol{\vartheta}] - \boldsymbol{\theta}^*\|_2 \le \mathbf{E}_\pi[\|\boldsymbol{\vartheta} - \boldsymbol{\theta}^*\|_2] \le (p/m)^{1/2}. \tag{14}$$

Using the triangle inequality, followed by (13) and (14), this yields

$$\begin{aligned}\bigg(\int_{\mathbb{R}^p}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}\bigg)^{1/a} &\le \big(\mathbf{E}[\|\boldsymbol{\theta}-\mathbf{E}_\pi[\boldsymbol{\theta}]\|_2^a]\big)^{1/a}+\|\mathbf{E}_\pi[\boldsymbol{\theta}]-\boldsymbol{\theta}^*\|_2\\ &\le (p/m)^{1/2}\,(1+a/p)^{1/2}+(p/m)^{1/2}.\end{aligned}$$

Using the inequality $(1+a/p)^{1/2}\le 1+a/(2p)$, we get the claim of the lemma for $a>2$. ∎

Lemma 3 *Let $f:\mathbb{R}^p\to\mathbb{R}$ be a twice differentiable convex function such that $\nabla^2 f(\boldsymbol{\theta})\preceq M\mathbf{I}_p$ for every $\boldsymbol{\theta}\in\mathbb{R}$, and let $\boldsymbol{\theta}^*\in\mathbb{R}^p$ be a minimizer of f. Assume that there exist a measurable map $m:[0,+\infty)\mapsto[0,M]$ such that $\nabla^2 f(\boldsymbol{\theta})\succeq m(\|\boldsymbol{\theta}\|_2)\mathbf{I}_p$ for any $\boldsymbol{\theta}\in\mathbb{R}^p$. Let $a>0$ and $A>0$. Define the ball $B_A=\{\boldsymbol{\theta}\in\mathbb{R}^p:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le A\}$. We have*

$$\int_{B_A^c}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^a\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}\le\frac{2(M/2)^{p/2}}{\Gamma(p/2)}\int_A^{+\infty} r^{p+a-1}e^{-\widetilde{m}(r)\,r^2/2}dr,$$

where

$$\widetilde{m}(r)=2\int_0^1(1-t)\,m(tr)\,dt.$$

Proof Without loss of generality, we assume that $\boldsymbol{\theta}^*=\mathbf{0}_p$ and $f(\mathbf{0}_p)=0$. Therefore, the density π is such that $\pi(\boldsymbol{\theta})=e^{-f(\boldsymbol{\theta})}/C$ where

$$C=\int_{\mathbb{R}^p}e^{-f(\boldsymbol{\theta})}d\boldsymbol{\theta}\ge\int_{\mathbb{R}^p}\exp\big\{-M\|\boldsymbol{\theta}\|_2^2/2\big\}d\boldsymbol{\theta}$$

by the fact that $\nabla^2 f(\boldsymbol{\theta})\preceq M\mathbf{I}_p$ for every $\boldsymbol{\theta}\in\mathbb{R}^p$.

Now, for any $r>0$ and any $\boldsymbol{\theta}\in\mathbb{R}^p$ such that $\|\boldsymbol{\theta}\|_2=r$, Taylor's expansion around the minimum $\mathbf{0}_p$ yields

$$\begin{aligned}f(\boldsymbol{\theta})-f(\mathbf{0}_p)&=\boldsymbol{\theta}^\top\bigg(\int_0^1\int_0^s\nabla^2 f(t\boldsymbol{\theta})\,dt\,ds\bigg)\boldsymbol{\theta}\\ &\ge\|\boldsymbol{\theta}\|_2^2\int_0^1\int_0^s m(t\|\boldsymbol{\theta}\|_2)\,dt\,ds\\ &=r^2\int_0^1\int_0^s m(tr)\,dt\,ds\\ &=\frac{r^2}{2}\times\underbrace{2\int_0^1(1-t)\,m(tr)\,dt}_{=\widetilde{m}(r)}.\end{aligned}$$

We combine this fact with the lower bound on C to get

$$\begin{aligned}
\int_{B_A^c} \|\boldsymbol{\theta}\|_2^a \,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} &\le C^{-1}\int_{\|\boldsymbol{\theta}\|_2\ge A} \|\boldsymbol{\theta}\|_2^a e^{-f(\boldsymbol{\theta})} d\boldsymbol{\theta}\\
&\le \bigg(\int_{\mathbb{R}^p} e^{-M\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta}\bigg)^{-1} \int_{\|\boldsymbol{\theta}\|_2\ge A} \|\boldsymbol{\theta}\|_2^a e^{-\widetilde m(\|\boldsymbol{\theta}\|_2)\|\boldsymbol{\theta}\|_2^2/2} d\boldsymbol{\theta}\\
&= \bigg(\int_0^{+\infty} r^{p-1} e^{-Mr^2/2} dr\bigg)^{-1} \int_A^{+\infty} r^{a+p-1} e^{-\widetilde m(r)r^2/2} dr\\
&= \frac{2(M/2)^{p/2}}{\Gamma(p/2)} \int_A^{+\infty} r^{a+p-1} e^{-\widetilde m(r)r^2/2} dr
\end{aligned}$$

where the first equality follows from a change of variables in polar coordinates, where the volume of the sphere cancels out in the ratio. ■

Lemma 4 *Assume that $\pi(\boldsymbol{\theta}) \propto e^{-f(\boldsymbol{\theta})}$, where*

$$f(\boldsymbol{\theta}) = 0.5\|\boldsymbol{\theta}\|_2^2 \mathbb{1}_{\|\boldsymbol{\theta}\|_2\le 1} + \|\boldsymbol{\theta}\|_2 \mathbb{1}_{\|\boldsymbol{\theta}\|_2>1}.$$

For any $a > 0$ and for any $p \ge 2 \vee (a-1)$,

$$\mu_a^a(\pi) \ge 0.1\Gamma(p+a)/\Gamma(p) \ge 0.1(p-1)^a.$$

Consequently, under assumptions of Proposition 2 (here with $m = R = 1$), the upper bound $O(p)$ on μ_a is not improvable.

Proof Remark first that $f(\boldsymbol{\theta}) = \varphi(\|\boldsymbol{\theta}\|_2)$ where $\varphi(r) := 0.5r^2\mathbb{1}_{r\le 1} + r\mathbb{1}_{r>1}$. We compute explicitly the moment by a change of variable in polar coordinates

$$\begin{aligned}
\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a \,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} &= \bigg(\int_0^{+\infty} r^{p-1} e^{-\varphi(r)} dr\bigg)^{-1} \int_0^{+\infty} r^{p+a-1} e^{-\varphi(r)} dr\\
&= \frac{\Gamma(p+a) + \int_0^1 r^{p+a-1}(e^{-r^2/2} - e^{-r})dr}{\Gamma(p) + \int_0^1 r^{p-1}(e^{-r^2/2} - e^{-r})dr}.
\end{aligned}$$

Using the fact that $(0.2)r \le e^{-r^2/2} - e^{-r} \le r$ for $0 < r < 1$ yields

$$\begin{aligned}
\int_{\mathbb{R}^p} \|\boldsymbol{\theta}\|_2^a \,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} &\ge \frac{\Gamma(p+a) + 0.2/(p+a+1)}{\Gamma(p) + 1/(p+1)}\\
&\ge \frac{\Gamma(p+a) + 0.1/(p+1)}{\Gamma(p) + 1/(p+1)}\\
&\ge (0.1)\Gamma(p+a)/\Gamma(p)
\end{aligned}$$

where the second inequality follows from the fact that $a \le p+1$ by assumption, while the last inequality follows from the fact that $\Gamma(\cdot)$ is an increasing function on $[2, +\infty)$. This proves the claim of the lemma. ■

Lemma 5 *Let $\Gamma(k,x)$ be the upper incomplete Gamma function. Let $k>2$ be a real number, then $\mu_k \le A_k^{1/k}\mu_2$ where $A_k = \min_{\lambda>2,\gamma>1} A_k(\lambda,\gamma)$ with*

$$A_k(\lambda,\gamma) = \frac{\sqrt{\lambda-1}}{\lambda}\bigg[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\bigg]^k k\Gamma\Big(k,\frac{\gamma^{1/2}\log(\lambda-1)}{2}\Big) + \frac{k(\gamma\lambda)^{k/2-1}-2}{k-2}. \tag{15}$$

Proof Let $\lambda>1$ be fixed throughout the proof and define $\mathcal{A} = \{\boldsymbol{\theta}\in\mathbb{R}^p : \|\boldsymbol{\theta}\|_2^2 \le \lambda\mu_2^2\}$. From Markov's inequality we have

$$\pi(\mathcal{A}) \ge 1 - \frac{\mathbf{E}_\pi[\|\boldsymbol{\vartheta}\|_2^2]}{\lambda\mu_2^2} = 1-\frac{1}{\lambda}.$$

The set $\mathcal{A}$ being symmetric, Proposition 2.14 from (Ledoux, 2001) implies that

$$1-\pi(s\mathcal{A}) \le \pi(\mathcal{A})\left(\frac{1-\pi(\mathcal{A})}{\pi(\mathcal{A})}\right)^{(s+1)/2},$$

for every real number s larger than 1. Since the right-hand side is a decreasing function of $\pi(\mathcal{A})$, we obtain the following bound on $\pi(s\mathcal{A}^\complement)$:

$$\pi(s\mathcal{A}^\complement) \le \frac{1}{\lambda(\lambda-1)^{(s-1)/2}}, \qquad \forall s\ge 1. \tag{16}$$

Let us introduce the random variable η as $\|\boldsymbol{\vartheta}\|_2/\mu_2$, where $\boldsymbol{\vartheta}\sim\pi$. It is clear that (15) is equivalent to

$$\mathbf{E}[\eta^k] \le \frac{\sqrt{\lambda-1}}{\lambda}\bigg[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\bigg]^k k\Gamma\Big(k,\frac{\gamma^{1/2}\log(\lambda-1)}{2}\Big) + \frac{k(\gamma\lambda)^{k/2-1}-2}{k-2}.$$

Since $\eta>0$ almost surely,

$$\mathbf{E}[\eta^k] = \int_0^\infty \mathbf{P}(\eta^k>u)\,du = k\int_0^\infty t^{k-1}\mathbf{P}(\eta>t)\,dt.$$

Thus, the proof of the lemma reduces to bound the tail of η. The definition of η and inequality (16) yield

$$\mathbf{P}(\eta>t) = \mathbf{P}(\|\boldsymbol{\vartheta}\|_2 > t\mu_2) = \pi\left(\frac{t}{\sqrt{\lambda}}\cdot\mathcal{A}^\complement\right) \le \frac{1}{\lambda(\lambda-1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}},$$

for every $t>\sqrt{\lambda}$. We choose $\gamma>1$ and apply this inequality to $t>\sqrt{\gamma\lambda}$. For the other values of t, that is when $t\le\sqrt{\gamma\lambda}$, we apply Markov's inequality to get $\mathbf{P}(\eta>t)\le 1\wedge t^{-2}$. Combining these two bounds, we arrive at

$$\begin{aligned}\mathbf{E}[\eta^k] &\le k\int_{\sqrt{\gamma\lambda}}^\infty \frac{t^{k-1}}{\lambda\cdot(\lambda-1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}dt + \int_0^{\sqrt{\gamma\lambda}} kt^{k-1}(1\wedge t^{-2})dt\\ &= k\int_{\sqrt{\gamma\lambda}}^\infty \frac{t^{k-1}}{\lambda\cdot(\lambda-1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}dt + \frac{k(\gamma\lambda)^{k/2-1}-2}{k-2}.\end{aligned}$$

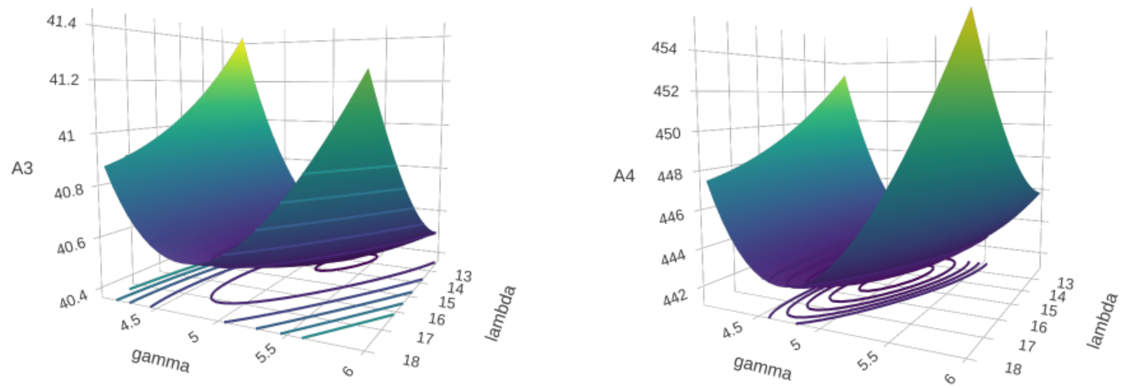

Figure 2: Shapes of the surfaces defined by the functions $A_3(\cdot,\cdot)$ and $A_4(\cdot,\cdot)$, see Lemma 5.

The first integral of the last sum can be calculated using the upper incomplete gamma function $\Gamma(k,z)$. Indeed, the change of variable $z = t\log(\lambda-1)/(2\sqrt{\lambda})$ yields

$$\begin{aligned}\int_{\sqrt{\gamma\lambda}}^{\infty}\frac{t^{k-1}}{\lambda\cdot(\lambda-1)^{(t-\sqrt{\lambda})/2\sqrt{\lambda}}}\,dt &= \frac{\sqrt{\lambda-1}}{\lambda}\int_{\sqrt{\gamma\lambda}}^{\infty} t^{k-1}\exp\Big(-\log(\lambda-1)\frac{t}{2\sqrt{\lambda}}\Big)\,dt\\ &= \frac{\sqrt{\lambda-1}}{\lambda}\bigg[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\bigg]^k\int_{\frac{\gamma^{1/2}\log(\lambda-1)}{2}}^{\infty} z^{k-1}e^{-z}\,dz\\ &= \frac{\sqrt{\lambda-1}}{\lambda}\bigg[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\bigg]^k\Gamma\Big(k,\frac{\gamma^{1/2}\log(\lambda-1)}{2}\Big).\end{aligned}$$

Finally, we obtain

$$\mathbf{E}[\eta^k] \le k\cdot\frac{\sqrt{\lambda-1}}{\lambda}\bigg[\frac{2\sqrt{\lambda}}{\log(\lambda-1)}\bigg]^k\Gamma\Big(k,\frac{\gamma^{1/2}\log(\lambda-1)}{2}\Big)+\frac{k(\gamma\lambda)^{k/2-1}-2}{k-2}.$$

This concludes the proof. ■

Remark 1 *We displayed*[6] *in Figure 2 the plots of the function $A_k(\cdot,\cdot)$ for $k=3$ and $k=4$. Numerically, we find that the optimal choice for (λ,γ) is approximately $\lambda = 15.89$ and $\gamma = 4.4$ for $k=3$ and $\lambda = 14.97$ and $\gamma = 4.8$ for $k=4$. This leads to the numerical bounds*

$$A_k \le \begin{cases}40.40, & k=3,\\ 441.43, & k=4.\end{cases}$$

These constants are by no means optimal, but we are not aware of any better bound available in the literature. Inequalities of type $\mathbf{E}[\|\boldsymbol{\vartheta}\|_2^k] \le A_k\mathbf{E}[\|\boldsymbol{\vartheta}\|_2^2]^{k/2}$ are often referred to as

[6] The R notebook for generating this figure can be found here `https://rpubs.com/adalalyan/Khintchine_constant`

the Kintchine inequality (Khintchine, 1923). According to (Cattiaux and Guillin, 2018), Corollary 4.3 from (Bobkov, 1999) implies that $A_4 \leq 49$ for one-dimensional $\boldsymbol{\vartheta}$ with log-concave density. Getting such a small constant in the multidimensional case would be of interest for applications to MCMC sampling.

References

J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. Control variates for stochastic gradient mcmc. *Statistics and Computing*, 29(3):599–615, May 2019.

E. Bernton. Langevin Monte-Carlo and JKO splitting. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1777–1798. PMLR, 2018.

S. G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27(4):1903–1921, 1999. ISSN 0091-1798.

F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 14(3):331–352, 2005.

J. M. Borwein, O. Chan, et al. Uniform bounds for the complementary incomplete gamma function. *Mathematical Inequalities and Applications*, 12:115–121, 2009.

H. J. Brascamp and E. H. Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366 – 389, 1976.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with projected Langevin Monte-Carlo. *Discrete & Computational Geometry*, 59(4):757–783, Jun 2018.

P. Cattiaux and A. Guillin. On the Poincaré constant of log-concave measures. *arXiv preprint arXiv:1810.08369*, 2018.

N. Chatterji, N. Flammarion, Y. Ma, P. Bartlett, and M. Jordan. On the theory of variance reduction for stochastic gradient Monte-Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018.

Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast mixing of Metropolized Hamiltonian Monte-Carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21:92:1–92:72, 2020.

Z. Chen and S. S. Vempala. Optimal convergence rate of Hamiltonian Monte-Carlo for strongly logconcave distributions. *CoRR*, abs/1905.02313, 2019.

X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of ALT2018*, 2018. URL `http://proceedings.mlr.press/v83/cheng18a.html`.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *CoRR*, abs/1805.01648, 2018a.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR, 2018b.

A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte-Carlo and gradient descent. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689, 07–10 Jul 2017a.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. *J. R. Stat. Soc. B*, 79:651 – 676, 2017b.

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte-Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, June 18-21, 2009*, pages 1–10, 2009.

R. Douc, E. Moulines, and J. S. Rosenthal. Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann. Appl. Probab.*, 14(4):1643–1665, 2004.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 06 2017. doi: 10.1214/16-AAP1238. URL `http://dx.doi.org/10.1214/16-AAP1238`.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 11 2019.

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), 2018. URL `https://hal.archives-ouvertes.fr/hal-01267115`.

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte-Carlo via convex optimization. *J. Mach. Learn. Res.*, 20:73–1, 2019.

R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on Learning Theory*, pages 793–797. PMLR, 2018.

A. Giannopoulos, S. Brazitikos, P. Valettas, and B.-H. Vritsiou. *Geometry of Isotropic Convex Bodies*. 05 2014.

N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Process. Related Fields*, 16(4):635–736, 2010.

G. Hargé. A convex/log-concave correlation inequality for gaussian measure and an application to abstract Wiener spaces. *Probability theory and related fields*, 130(3):415–440, 2004.

P. Jain and P. Kar. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, 2017.

A. Khintchine. Über dyadische Brüche. *Math. Z.*, 18(1):109–116, 1923.

M. Ledoux. *The concentration of measure phenomenon.* Number 89. American Mathematical Soc., 2001.

L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM J. Comput.*, 35(4):985–1005 (electronic), 2006.

L. Lovasz and S. Vempala. Fast algorithms for logconcave functions: Sampling, Rounding, Integration and Optimization. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 57–68, 2006.

Y.-A. Ma, N. Chatterji, X. Cheng, N. Flammarion, P. Bartlett, and M. I. Jordan. Is there an analog of Nesterov Acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

M. B. Majka, A. Mijatović, and Łukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, 30(4):1534 – 1581, 2020.

O. Mangoubi and A. Smith. Rapid mixing of Hamiltonian Monte-Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

O. Mangoubi and A. Smith. Mixing of Hamiltonian Monte-Carlo on strongly log-concave distributions 2: Numerical integrators. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 586–595, 2019.

O. Mangoubi and N. K. Vishnoi. Nonconvex sampling with the metropolis-adjusted Langevin algorithm. pages 2259–2293, 2019.

P. Natalini and B. Palumbo. Inequalities for the incomplete gamma function. *Math. Inequal. Appl*, 3(1):69–77, 2000.

E. Nelson. *Dynamical Theories of Brownian Motion.* Department of Mathematics. Princeton University, 1967.

G. A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics.* Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

F. Qi, Q.-M. Luo, et al. Bounds for the ratio of two gamma functions—from Wendel's and related inequalities to logarithmically completely monotonic functions. *Banach Journal of Mathematical Analysis*, 6(2):132–158, 2012.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, 07–10 Jul 2017.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268, 1998.

G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4(4):337–357 (2003), 2002.

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

Y. Tat Lee, Z. Song, and S. S. Vempala. Algorithmic Theory of ODEs and Sampling from Well-conditioned Logconcave Densities. *arXiv e-prints*, art. arXiv:1812.06243, Dec 2018.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 06–09 Jul 2018.

D. Zou, P. Xu, and Q. Gu. Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 2936–2945, 2019.