Wasserstein Control of Mirror Langevin Monte Carlo

Kelvin Shuangjian Zhang SZHANG@ENS.FR
CNRS and DMA, École Normale Supérieure, Université PSL, Paris, France

Gabriel Peyré GABRIEL.PEYRE@ENS.FR
CNRS and DMA, École Normale Supérieure, Université PSL, Paris, France

Jalal Fadili JALAL.FADILI@GREYC.ENSICAEN.FR
Normandie Univ, ENSICAEN, UNICAEN, CNRS, GREYC, France

Marcelo Pereyra M.PEREYRA@HW.AC.UK
School of Mathematical and Computer Sciences, Heriot-Watt University, UK

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Discretized Langevin diffusions are efficient Monte Carlo methods for sampling from high dimensional target densities that are log-Lipschitz-smooth and (strongly) log-concave. In particular, the Euclidean Langevin Monte Carlo sampling algorithm has received much attention lately, leading to a detailed understanding of its non-asymptotic convergence properties and of the role that smoothness and log-concavity play in the convergence rate. Distributions that do not possess these regularity properties can be addressed by considering a Riemannian Langevin diffusion with a metric capturing the local geometry of the log-density. However, the Monte Carlo algorithms derived from discretizations of such Riemannian Langevin diffusions are notoriously difficult to analyze. In this paper, we consider Langevin diffusions on a Hessian-type manifold and study a discretization that is closely related to the mirror-descent scheme. We establish for the first time a non-asymptotic upper-bound on the sampling error of the resulting Hessian Riemannian Langevin Monte Carlo algorithm. This bound is measured according to a Wasserstein distance induced by a Riemannian metric ground cost capturing the squared Hessian structure and closely related to a self-concordance-like condition. The upper-bound implies, for instance, that the iterates contract toward a Wasserstein ball around the target density whose radius is made explicit. Our theory recovers existing Euclidean results and can cope with a wide variety of Hessian metrics related to highly non-flat geometries.

Keywords: Riemannian Langevin Monte Carlo, Hessian manifold, sampling, contraction, Baillon-Haddad inequality.

1. Introduction

1.1. Problem and setting

We consider the problem of sampling from a target probability distribution $d\pi = e^{-f(\mathbf{x})}d\mathbf{x}$ supported on a domain $\mathcal{X} \subset \mathbb{R}^p$, where f is differentiable on $\mathcal{X}$. We are particularly interested in sampling algorithms that scale efficiently to high dimensions. When f is Lipschitz-smooth (i.e. differentiable with Lipschitz gradient) and strongly convex on $\mathcal{X}$, then the conventional Langevin Monte Carlo (LMC) algorithm derived from an Euler-Maruyama discretization of the Langevin stochastic differential equation (SDE) is one of the most computationally efficient methods to sample from π. In this paper, we endow $\mathcal{X}$ with a carefully designed Riemannian structure and study the non-asymptotic convergence properties of a Riemannian generalization of the LMC algorithm. The motivation is that by endowing $\mathcal{X}$ with an appropriate Riemannian geometry, it is possible to obtain algorithms with better convergence properties, and which can tackle distributions that are beyond the scope of the Euclidean LMC algorithm. We consider Riemannian structures of Hessian type (Shima, 2007); the corresponding metric is induced by the Hessian $D^2\phi(\mathbf{x})$ of some $C^2(\mathcal{X})$ *Legendre-type* convex potential/entropy ϕ on $\mathcal{X}$ (see (Rockafellar, 1970, Chapter 26) for a comprehensive account on Legendre functions).

Discrete scheme. In the same vein as in Hsieh et al. (2018), we consider a sampling analogue of mirror-descent as an extension of the classical Euler-Maruyama discretization of the Langevin SDE, which reads, starting from some random vector $\mathbf{X}_0$ on $\mathcal{X}$,

$$\mathbf{X}_{k+1} \stackrel{\text{def.}}{=} \nabla\phi^*\Big(\nabla\phi(\mathbf{X}_k) - h_{k+1}\nabla f(\mathbf{X}_k) + \sqrt{2h_{k+1}[D^2\phi(\mathbf{X}_k)]}\boldsymbol{\xi}_{k+1}\Big). \tag{1}$$

Here ϕ^* is the Legendre-Fenchel conjugate of ϕ, i.e., $\phi^*(\mathbf{y}) \stackrel{\text{def.}}{=} \sup_{\mathbf{x}\in\mathcal{X}}\langle\mathbf{x},\mathbf{y}\rangle - \phi(\mathbf{x})$, $\{h_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_{++}$ is the sequence of step-sizes, and $\{\boldsymbol{\xi}_k\}_{k\in\mathbb{N}}$ is a sequence of standard normal random vectors that are mutually independent and independent of $\mathbf{X}_0$, which is either deterministic or random. Let us recall the useful fact that ϕ is of Legendre type if and only if its conjugate ϕ^* is of Legendre type. Moreover, the gradient $\nabla\phi$ of ϕ is a bijection from $\operatorname{int}\operatorname{dom}(\phi) = \mathcal{X}$ to $\operatorname{int}\operatorname{dom}(\phi^*) = \mathcal{Y}$ and its inverse obeys $(\nabla\phi)^{-1} = \nabla\phi^*$, see (Rockafellar, 1970, Theorem 26.5). Thus (1) makes perfect sense as a single-valued mapping from $\mathcal{X}$ to $\mathcal{X}$.

In the following, we call iteration (1) **Hessian Riemannian Langevin Monte Carlo (HRLMC)** algorithm. Note that Hsieh et al. (2018) does not study this method, and rather settles for a different discretization, which is simpler to analyze (being a change of variable applied to the Euclidean case) and enjoys theoretical guarantees that are markedly different from ours (we refer to Section 1.2 for a detailed comparison).

In the case where $\boldsymbol{\xi}_k = 0$ (optimization framework), one recovers the mirror descent minimization algorithm (Nemirovsky and Yudin, 1983; Bauschke et al., 2017; Lu et al., 2018). The classical Euclidean case is recovered when ϕ is the energy, i.e., $\phi(\mathbf{x}) = \|\mathbf{x}\|_2^2/2$. Other popular options to sample in $\mathcal{X} = \mathbb{R}^p_{++}$ include Shannon entropy $\phi(\mathbf{x}) = \sum_i x_i\log(x_i)$ and Burg's entropy $\phi(\mathbf{x}) = -\sum_i\log(x_i)$.

As mentioned previously, the key motivations behind switching from Euclidean LMC methods to the HRLMC scheme are that by choosing an entropy ϕ adapted to f, one can either obtain better smoothness and strong convexity properties or even recover smoothness and strong convexity relative to ϕ in cases where f is neither Lipschitz-smooth nor strongly convex in the standard Euclidean geometry. The goal of this paper is to provide the first step toward a theoretical understanding of these phenomena, by establishing a non-asymptotic upper-bound on the error in a properly designed Wasserstein distance for sampling from π using HRLMC. The terms in the bound explicitly reflect the interleaved geometries of f and ϕ.

Continuous flow. It can be shown that the HRLMC algorithm (1) can be viewed as a discretization of a Riemannian SDE. Denoting $\mathbf{Y}_t \stackrel{\text{def.}}{=} \nabla\phi(\mathbf{X}_t)$, this SDE reads

$$d\mathbf{Y}_t = -\nabla f\circ\nabla\phi^*(\mathbf{Y}_t)dt + \sqrt{2[D^2\phi^*(\mathbf{Y}_t)]^{-1}}d\mathbf{B}_t, \tag{2}$$

where $\{\mathbf{B}_t\}_{t\geq 0}$ is a standard p-dimensional Brownian motion. If moreover $\phi \in C^3(\mathcal{X})$, then Legendreness of ϕ entails that the SDE on $\mathbf{X}_t$ reads

$$d\mathbf{X}_t = \big(\theta(\mathbf{X}_t) - [D^2\phi(\mathbf{X}_t)]^{-1}\nabla f(\mathbf{X}_t)\big)dt + \sqrt{2[D^2\phi(\mathbf{X}_t)]^{-1}}d\mathbf{B}_t, \tag{3}$$

where the additional drift term $\theta(\mathbf{X}_t) \overset{\text{def.}}{=} -[D^2\phi(\mathbf{X}_t)]^{-1}\mathrm{Tr}\left(D^3\phi(\mathbf{X}_t)[D^2\phi(\mathbf{X}_t)]^{-1}\right)$. Moreover, the corresponding density can be shown to satisfy a Fokker-Planck equation that has π as its stationary solution (we omit the details for the sake of brevity). When $\phi(\mathbf{x}) = \|\mathbf{x}\|_2^2/2$, then $\mathbf{X}_t = \mathbf{Y}_t$, and (2) and (3) coincide with the standard Langevin diffusion. The SDE (3), viewed as Brownian motion on a Hessian manifold corrected by a Riemannian drift term, is then its natural generalization to a Riemannian manifold with a Hessian structure. The SDE (2) appeared in earlier preprint versions of Hsieh et al. (2018), while the SDE (3) is a particular case of the so-called Riemannian Langevin dynamics as shown in Roberts and Stramer (2002). We will show in Appendix A that both (2) and (3) are well-posed, under a self-concordance-like condition (**A1**) and a relative Lipschitz-smoothness condition (**A4**).

1.2. Previous work

The goal of this paper is to provide non-asymptotic upper-bounds on the Wasserstein distance, with an appropriate ground cost, between the distribution μ_k of $\mathbf{X}_k$ and the target distribution π.

Langevin Monte Carlo (LMC) under (strong) log-concavity. The Euclidean LMC, corresponding to $\phi(\mathbf{x}) = \|\mathbf{x}\|_2^2/2$, has been extensively studied in the literature, where non-asymptotic error bounds have been established under various sampling error metrics (Kullback-Leibler, Total-Variation, or Wasserstein). The case where f is m-strongly convex with a M-Lipschitz gradient is the one that has been most widely studied (Dalalyan, 2017a,b; Durmus and Moulines, 2017; Cheng and Bartlett, 2018; Durmus and Moulines, 2019; Dalalyan and Karagulyan, 2019; Durmus et al., 2019; Dwivedi et al., 2018). In particular, (Dalalyan and Karagulyan, 2019) have shown that, when using a constant step-size $h_k = h \in (0, \frac{2}{M})$, the distribution of LMC algorithm samples converge to the targed distribution with a contraction factor $\rho = \max(1 - mh, Mh - 1)$. More precisely,

$$\begin{aligned} W_2(\mu_k, \pi) &\leq \rho^k W_2(\mu_0, \pi) + \frac{1.65Mh^{\frac{3}{2}}p^{\frac{1}{2}}}{1-\rho} \\ &\leq (1-mh)^k W_2(\mu_0, \pi) + 1.65(M/m)(ph)^{\frac{1}{2}}, \quad \text{if } h \leq 2/(m+M), \end{aligned} \tag{4}$$

where W_2 is the 2-Wasserstein distance between two probability measures, i.e.,

$$W_2^2(\mu, \nu) \overset{\text{def.}}{=} \inf_{\mathbf{X}\sim\mu, \mathbf{X}'\sim\nu} \mathbf{E}\left[\|\mathbf{X} - \mathbf{X}'\|_2^2\right].$$

This is the best known result in Wasserstein distance.

Durmus et al. (2018) studied the case of non-Lipschitz-smooth (strongly) convex f via Moreau-Yosida regularization, and Bubeck et al. (2018); Brosse et al. (2017) the case of log-Lipschitz-smooth strongly log-concave densities supported on a convex compact set. Cheng et al. (2017); Dalalyan and Riou-Durand (2018) investigated the case of a kinetic Langevin diffusion (i.e., underdamped LMC) for the same class of densities, showing that it leads to improved dependence on the dimension and error.

Non-asymptotic sampling error bounds when f is Lipschitz-smooth and merely convex (but not strongly so) have been established in the literature in KL and TV Durmus et al. (2019), and in Wasserstein distance Dalalyan et al. (2019) for various discrete LMC schemes.

LMC beyond log-concavity. Obtaining convergence results is very difficult when f is not convex. Luu et al. (2017) considered densities that are neither necessarily smooth nor log-concave and provided asymptotic consistency guarantees. Assuming convexity at infinity, Cheng et al. (2018); Majka et al. (2018) obtained convergence results in the 1-Wasserstein distance by using results in Eberle (2016). When replacing convexity with a dissipativity condition, a non-asymptotic bound was first provided by Raginsky et al. (2017) in the 2-Wasserstein distance, and then improved by Chau et al. (2019). In Zhang et al. (2019), assumptions are further weakened by assuming only local Lipschitz continuity of ∇f and by relaxing conditions of convexity at infinity and uniform dissipativity.

Continuous Riemannian Langevin dynamics. The SDE (3) is a special case of the so-called Riemannian Langevin dynamics, which appeared in Roberts and Stramer (2002); Girolami and Calderhead (2011); Patterson and Teh (2013), when considering $\mathcal{X}$ as a Riemannian manifold with Hessian metric $D^2\phi$. For this Riemannian Langevin SDE setting, it is known since Kent (1978) that $\mathbf{X}_t$ has π as its unique invariant measure as long as $\mathbf{X}_t$ is non-explosive. For the conditions on the non-explosion of diffusions, see Stroock and Varadhan (2007). Moreover, the linear convergence theory of the corresponding Fokker-Planck equation is known since Arnold et al. (2001), relying on the positivity of Bakry-Emery tensor; see (Bakry et al., 2014) for a comprehensive account. Discretization schemes of the Riemannian Langevin SDE (3) were proposed in Roberts and Stramer (2002); Girolami and Calderhead (2011); Patterson and Teh (2013). For instance, Roberts and Stramer (2002) provided a linear convergence result of the Ozaki discretization under quite stringent conditions. In particular, for the Hessian manifold, this theory requires ϕ to be strongly convex, which in turn restricts the target distribution to be strongly log-concave.

In this paper, instead, we take the Euler-Maruyama discretization of (2) and map the process back to $\mathbf{X}_k$ by the mirror map $\mathbf{X}_k = \nabla\phi^*(\mathbf{Y}_k)$. This is a key difference between our HRLMC algorithm (1) and those proposed in Roberts and Stramer (2002); Girolami and Calderhead (2011); Patterson and Teh (2013). However, the restriction to a Hessian Riemannian geometry is crucial in our method and theory, which strongly rely on convex analysis tools and bijective duality mappings. To the best of our knowledge, there is no proof of convergence or error bounds for such Euler-Maruyama discretization of (2) or (3).

Relation to Hsieh et al. (2018). In 2018, Hsieh et al. (2018) studied a mirror-type discretization of Langevin dynamics. Though it seems that their work shares apparent similarities with ours at first glance, both their scheme and results are, however, markedly different from our HRLMC. More precisely, a key difference lies in the fact that here, we use an appropriate diffusion term entailing a Gaussian noise in the discrete scheme with iteration-dependent covariances that account for the Hessian Riemannian structure. In contrast, Hsieh et al. (2018) adopted a standard Gaussian noise instead. Moreover, they provided the existence of good mirror maps assuming f is strongly convex and gave convergence of their sampling algorithm under 1-strongly convex mirror maps. In this paper, we relax these requirements to relative versions and aim to generalize results from the literature relying on strong convexity and Lipschitz-smoothness of f.

1.3. Contributions

In this paper, by relaxing strong convexity and Lipschitz-smoothness of f to the relative versions with respect to a Legendre-type entropy ϕ, we prove that, if the step-sizes h_k are chosen sensibly, the law of discrete process (1) contracts into a Wasserstein ball centered at the desired invariant distribution, whose radius is given explicitly. This Wasserstein distance relies on a ground cost, which is a Riemannian distance that captures the squared Hessian structure of the manifold. In fact, convergence to π is not achieved in general unless ϕ is quadratic, but our bound allows us to isolate a bias term that depends on the interleaved geometries of f and ϕ. In particular, our method recovers the state-of-the-art non-asymptotic sampling error bounds in Wasserstein distance when $\phi(\mathbf{x}) = \|\mathbf{x}\|_2^2/2$ (Dalalyan and Karagulyan, 2019).

Section 2 states the main contribution of this paper, Proposition 1, whose proof relies on a more general result (Theorem 2) detailed in Section 3. In the appendices, we collect all details of the discussions and proofs. This includes discussions of our assumptions (e.g., intuition behind condition (**A1**), relation of (**A3**) and (**A4**) to relative strong convexity and relative smoothness). We also present a generalized Baillon-Haddad inequality (8) that is of independent interest, and give the detailed proofs of Proposition 1, Corollary 3, and Proposition 5. We also report some numerical experiments to illustrate and support our theoretical predictions.

Notations. Thought out the paper, $\mathcal{M}_{k\times l}$ is the ring of $k \times l$ matrices on $\mathbb{R}$. $\|\mathbf{v}\|_2$ is the Euclidean norm of a vector $\mathbf{v}$; for a matrix $\mathbf{M} \in \mathcal{M}_{k\times l}$, $\|\mathbf{M}\|_2$ stands for its spectral norm. That is, $\|\mathbf{M}\|_2 = \sqrt{\lambda_{\max}(\mathbf{M}^T\mathbf{M})}$, where $\lambda_{\max}$ represents the largest value of eigenvalues. By definition, $\|\mathbf{M}\|_2 \le \delta$ is equivalent to $\mathbf{M}^T\mathbf{M} \preceq \delta^2\mathbf{I}_p$, i.e., $\mathbf{M}^T\mathbf{M} - \delta^2\mathbf{I}_p$ is negative semi-definite. Another matrix norm we use here is the Frobenius norm $\|\mathbf{M}\|_F = \sqrt{\sum_{i,j=1}\mathbf{M}_{ij}^2} = \sqrt{\mathrm{Tr}(\mathbf{M}^T\mathbf{M})}$, where Tr is the trace operator. The commutator of two square matrices $\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}_{p\times p}$ is denoted as $[\mathbf{M}_1, \mathbf{M}_2] \overset{\text{def.}}{=} \mathbf{M}_1\mathbf{M}_2 - \mathbf{M}_2\mathbf{M}_1$.

2. Main contributions

In this section, we state our main contributions, namely that the HRLMC algorithm (1) contracts into a Wasserstein ball centered at the invariant measure.

2.1. Assumptions on ϕ and f

In the following, we assume that the domain $\mathcal{X} \subset \mathbb{R}^p$ is open, contractible and $\nabla\left(\frac{d\pi}{d\mathbf{x}}\right) = 0$ on its boundary $\partial\mathcal{X}$. To avoid technical issues, we assume that both f and ϕ are in $C^3(\mathcal{X})$ and ϕ is of Legendre type.

Self-concordance-like condition on ϕ. Our first condition imposes the existence of $\kappa \ge 0$ such that

$$\forall(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2, \quad \sqrt{2}\left\|D^2\phi(\mathbf{x})^{\frac{1}{2}} - D^2\phi(\mathbf{x}')^{\frac{1}{2}}\right\|_F \le \kappa\left\|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\right\|_2. \tag{A1}$$

In 1D, it is easy to check that this condition is equivalent to self-concordance. The general case is more intricate. (**A1**) is important to guarantee the existence and uniqueness of the strong solution of continuous dynamics (2) (see (Øksendal, 2003, Theorem 5.2.1)). In fact, if it is violated, the

Lipschitz condition of the SDE also fails, which removes the general theoretical guarantee for (2) to have an unique solution. See Appendix A for further details.

Moment condition on the Hessian of ϕ. The second constant involved in our analysis is

$$R \stackrel{\text{def.}}{=} \mathbf{E}_{\mathbf{X}\sim\pi}\left[\left\|D^2\phi(\mathbf{X})\right\|_2\right] = \int_{\mathcal{X}} \left\|D^2\phi(\mathbf{x})\right\|_2 e^{-f(\mathbf{x})}\mathrm{d}\mathbf{x} < +\infty. \tag{A2}$$

Relative strong convexity and Lipschitz-smoothness. In this paper, we relax the usual strong convexity and Lipschitz-smoothness conditions to versions relatively to ϕ: there exists $m \geq 0$, $M > 0$ such that $\forall(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2$,

$$m\left\|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\right\|_2^2 \leq \langle\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}'), \nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\rangle; \tag{A3}$$

$$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\right\|_2 \leq M\left\|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\right\|_2. \tag{A4}$$

In the Euclidean case when $\phi(\mathbf{x}) = \|\mathbf{x}\|^2/2$, one recovers the usual notion of strong convexity of f and Lipschitz continuity of its gradient. The condition (**A3**) and (**A4**) imply, respectively, the relative strong convexity and relative Lipschitz-smoothness defined in Lu et al. (2018); Bauschke et al. (2017). More precisely, they imply that $mD^2\phi(\mathbf{x}) \preceq D^2f(\mathbf{x}) \preceq MD^2\phi(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{X}$. The converse is not true in general. See details in Appendix B.

Bound on the commutator of $D^2\phi$ and D^2f. Whenever the Hessians D^2f and $D^2\phi$ do not commute, we require the following assumption to quantify the commutator:

$$\exists\delta \geq 0,\ \forall\mathbf{x} \in \mathcal{X}, \quad \left\|\left[(D^2\phi(\mathbf{x}))^{-1}, D^2f(\mathbf{x})\right]\right\|_2 \leq \delta. \tag{A5}$$

This control is crucial to prove the generalized Baillon-Haddad inequality (Proposition 4).

2.2. Wasserstein Distance

While the de-facto geodesic distance on $\mathcal{X}$ endowed with the Hessian structure is the Riemannian distance associated with $D^2\phi(\mathbf{x})$, this distance cannot be computed in closed form. We thus settle for a simpler one, which is the Riemannian distance d associated with the squared Hessian $[D^2\phi(\mathbf{x})]^2$. One can check that the diffeomorphism $\nabla\phi : (\mathcal{X}, [D^2\phi(\mathbf{x})]^2) \to (\mathcal{Y}, \mathbf{I}_p)$ is an isometry (see (do Carmo, 1992, Chapter 1) for a detailed account on the isometry of Riemannian manifolds). Therefore, $d(\mathbf{x}, \mathbf{x}') = \|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\|_2$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

With this ground distance, the natural associated geometric distance on the space of probability distributions on $\mathcal{X}$ is the Wasserstein distance

$$W_{2,\phi}^2(\mu,\nu) \stackrel{\text{def.}}{=} \inf_{\mathbf{x}\sim\mu,\mathbf{x}'\sim\nu} \mathbf{E}\left[d^2(\mathbf{x},\mathbf{x}')\right] = \inf_{\mathbf{x}\sim\mu,\mathbf{x}'\sim\nu} \mathbf{E}\left[\left\|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\right\|_2^2\right]. \tag{5}$$

When $\phi(\mathbf{x}) = \|\mathbf{x}\|^2/2$, one recovers the usual W_2 distance used in (4).

2.3. Statement of the main result

From now on, we assume that conditions (**A1**)–(**A5**) are satisfied. Denote by μ_k the distribution of the k-th iterate $\mathbf{X}_k$ in (1), i.e., $\mathbf{X}_k \sim \mu_k$, and define

$$\tilde{\kappa} \stackrel{\text{def.}}{=} \sqrt{\kappa^2 + \frac{\delta(4M+\delta)}{2(m+M)}}.$$

Our main contribution is Theorem 2, whose statement and proof will be given shortly in a forthcoming section. For the sake of clarity, we first apply it below to the case of constant step-sizes, which makes it easier to get the gist of our main result and compare it with existing works.

Proposition 1 (Constant step-size) *Assume conditions (**A1**)–(**A5**) are satisfied with* $\tilde{\kappa} < \sqrt{2m}$ *and* $h_k = h < \min\left(\frac{2m-\tilde{\kappa}^2}{m^2}, \frac{2M-\tilde{\kappa}^2}{M^2}\right)$. *Then*

$$W_{2,\phi}(\mu_k, \pi) \leq \rho^k W_{2,\phi}(\mu_0, \pi) + h^{\frac{3}{2}} p^{\frac{1}{2}} (1-\rho)^{-1} \beta_2(R, M, \kappa) + h p^{\frac{1}{2}} (1-\rho)^{-1} \beta_1(R, \kappa), \quad (6)$$

where $\rho \stackrel{\text{def.}}{=} \max\left(\sqrt{(1-mh)^2 + h\tilde{\kappa}^2}, \sqrt{(1-Mh)^2 + h\tilde{\kappa}^2}\right) < 1$, $\beta_1(R, \kappa) \stackrel{\text{def.}}{=} \kappa R^{\frac{1}{2}}$, *and* $\beta_2(R, M, \kappa) \stackrel{\text{def.}}{=} M^{\frac{1}{2}} R^{\frac{1}{2}} \left(\frac{7\sqrt{2}M}{6} + \frac{\kappa}{\sqrt{3}}\right)$ *are dimension-free constants.*

The error upper-bound is composed of three terms. The first one comes from the time finiteness that decreases exponentially, while the second corresponds to the discretization error. These two terms are standard in LMC. The last term is new and reveals the price to be paid if one trades the standard strong convexity and Lipschitz-smoothness for their relative versions in the Riemannian geometry induced by ϕ. If h is sufficiently small, one can see that $(1-\rho)^{-1} = \mathcal{O}(h^{-1})$, where the constant in the order depends on (m, M, κ, δ). In turn, the discretization error term will scale as $\mathcal{O}(\beta_2(R, M, \kappa) p^{1/2} h^{1/2})$, which vanishes as $h \to 0$, while the last term is $\mathcal{O}(\beta_1(R, \kappa) p^{1/2})$. The latter is a bias term. We conjecture that the bias is unavoidable and that our contraction analysis is tight. Indeed, this term is not an artifact of the proof since the estimates are based on sharp inequalities for which lower bounds are available. This is also confirmed by the numerics discussed in the appendix.

Moreover, our analysis recovers exactly known results for the particular case when f is m-strongly convex and has an M-Lipschitz continuous gradient, hence satisfying conditions (**A1**)–(**A5**) with $\phi(\mathbf{x}) = \|\mathbf{x}\|^2/2$, $\kappa = 0$, $R = 1$, $\delta = 0$, $\beta_1 = 0$, $\beta_2 = \frac{7\sqrt{2}M}{6}$, $\tilde{\kappa} = 0$, $\rho = \max\{1 - mh, Mh - 1\}$, and $W_{2,\phi} = W_2$. In this case, the bias term vanishes, and Proposition 1 recovers the sampling error bound of LMC from (Dalalyan and Karagulyan, 2019, Theorem 1), recalled in (4).

Besides, our proposition covers new cases not known in the literature, as shown in the forthcoming section. We want to emphasize that the condition $\tilde{\kappa} < \sqrt{2m}$ is essential as it connects the key parameters m, M, κ, δ, which summarize the interleaved geometries of f and ϕ. It requires $\kappa < \sqrt{2m}$ even if $\delta = 0$. We now illustrate this condition and assumptions (**A1**)–(**A5**) with several examples.

2.4. Examples

In this section, we provide two tables to include some examples that satisfy the assumptions (**A1**)–(**A5**) and condition $\tilde{\kappa} < \sqrt{2m}$ with explicit parameters. As κ is the only constant that depends merely on ϕ, Table 1 presents a list of entropy functions that satisfy (**A1**) or not, while Table 2 gives the constants involving interplay between ϕ and f. For instance, in the example of Gamma distribution (Table 2, middle column), one can see clearly how dimension enters the game via m and M.

1. More generally, $\phi(\mathbf{x}) = \sum_{i=1}^{p} \phi_i(x_i)$ satisfies (**A1**) with $\kappa = \sqrt{2}M'$ provided that $[(\phi_i^*)'']^{-\frac{1}{2}}$ has an M'-Lipschitz continuous gradient for each i. If $f(\mathbf{x}) = \sum_{i=1}^{p} f_i(x_i)$, then it satisfies

Table 1: Common entropy functions and the corresponding κ in (**A1**)

ϕ	κ	Domain
$\|\mathbf{x}\|^2/2$	0	$\mathbb{R}^p$
$-\sum_i \log(x_i)$	$\sqrt{2}$	$\mathbb{R}^p_{++}$
$\sum_i x_i \log(x_i)$	∞	$\mathbb{R}^p_{++}$
$-\log(x) - \log(1-x)$	$\sqrt{2}$	$(0,1)$
$\sum_i a_i x_i \log(x_i) - \sum_i (1-a_i)\log(x_i)$	$\sqrt{\frac{2}{1-\max_i a_i}}$	$\mathbb{R}^p_{++}$; $a_i \in [0,1]$
$(1-x^2)^{-1}$	1.43	$(-1,1)$
$-\log(x_2^2 - x_1^2)$	$\sqrt{2}$	$\{(x_1,x_2) : \|x_1\| < x_2\}$
$-\log(1-x^2)$	$\sqrt{2}$	$(-1,1)$

Table 2: Other parameters in the assumptions (**A2**)–(**A5**)

	$\phi = \|\mathbf{x}\|^2/2$ $f = \mathbf{x}^T\mathbf{A}\mathbf{x}/2 + C$ $(\mathbf{A}^T = \mathbf{A})$	$\phi = -\sum_{i=1}^p \log(x_i)$ $f = \sum_i (1-a_i)\log(x_i)$ $+b_i x_i + C$	$\phi = -\log(x) - \log(1-x)$ $f = (1-a_1)\log(x)$ $+(1-a_2)\log(1-x) + C$
R	1	$\sum_i (a_i-3)!/b_i^{a_i-2}$	$\frac{(a_1-3)!(a_2-1)!+(a_1-1)!(a_2-3)!}{(a_1+a_2-3)!}$
m	$\lambda_{\min}(\mathbf{A})$	$\min_i\{a_i - 1\}$	$\min\{a_1-1, a_2-1\}$
M	$\lambda_{\max}(\mathbf{A})$	$\max_i\{a_i - 1\}$	$\max\{a_1-1, a_2-1\}$
δ	0	0	0
$\tilde{\kappa} < \sqrt{2m}$	$\mathbf{A}$ is positive definite	$a_i > 2, \forall i$	$a_1, a_2 > 2$

(**A2**) and (**A5**) with $R \le \sum_i \mathbf{E}_{\mathbf{x}\sim\pi}[\phi_i''(x_i)]$ and $\delta = 0$. Besides, (**A3**) and (**A4**) are satisfied if, for each i, f_i is m-strongly convex and has an M-Lipschitz continuous gradient relatively to ϕ_i, in the sense of Lu et al. (2018).

2. Boltzmann-Shannon entropy: When $\phi(\mathbf{x}) = \sum_{i=1}^p x_i \log(x_i)$, however, condition (**A1**) is violated on $\mathbb{R}^p_{++}$.

3. Proof of the Main Result

3.1. A general non-asymptotic error bound

We are now in position to state our main theorem.

Theorem 2 (Contractibility) *Assume that* (**A1**)–(**A5**) *hold such that* $\tilde{\kappa} < \sqrt{2m}$. *Suppose* $h_{k+1} < \min\left(\frac{2m-\tilde{\kappa}^2}{m^2}, \frac{2M-\tilde{\kappa}^2}{M^2}\right)$. *Then*

$$W_{2,\phi}(\mu_{k+1}, \pi) \le \rho_{k+1} W_{2,\phi}(\mu_k, \pi) + h_{k+1} p^{\frac{1}{2}} \beta_1(R,\kappa) + h_{k+1}^{\frac{3}{2}} p^{\frac{1}{2}} \beta_2(R, M, \kappa). \tag{7}$$

Here $\rho_{k+1} \stackrel{\text{def.}}{=} \max\left(\sqrt{(1-mh_{k+1})^2 + h_{k+1}\tilde{\kappa}^2}, \sqrt{(1-Mh_{k+1})^2 + h_{k+1}\tilde{\kappa}^2}\right) < 1$, $\beta_1(R,\kappa) = \kappa R^{\frac{1}{2}}$, *and* $\beta_2(R, M, \kappa) = M^{\frac{1}{2}} R^{\frac{1}{2}} \left(\frac{7\sqrt{2M}}{6} + \frac{\kappa}{\sqrt{3}}\right)$ *are dimension-free constants.*

The main arguments to prove Theorem 2 will be given in Section 3.2 and 3.3. This result implies in particular Proposition 1 when the step-sizes are constant. Besides, the result in (7) is invariant in scalings like $\tilde{\phi} = \alpha\phi$ for any $\alpha > 0$.

Theorem 2 has the next corollary. In a nutshell, this corollary states that with vanishing step-sizes, the HRLMC algorithm contracts toward a Wasserstein ball centered at the target distribution π with radius r_0. The explicit formula of this radius is $r_0 \stackrel{\text{def.}}{=} \frac{2\kappa p^{\frac{1}{2}} R^{\frac{1}{2}}}{2m-\tilde{\kappa}^2}$, which scales as $\mathcal{O}(p^{\frac{1}{2}})$ in the dimension. This formula is derived from (7) upon applying Lemma 21 (see (44) and the proof of Corollary 3). Moreover, once entering the ball, the distribution μ_k never leaves it. When $\phi = \|\mathbf{x}\|^2/2$, it is clear that $r_0 = 0$ and therefore the algorithm converges to the stationary distribution.

In the following, we use the notation $\mathcal{B}_r(\pi) \stackrel{\text{def.}}{=} \{\mu \in \mathcal{P}(\mathcal{X}) | W_{2,\phi}(\mu, \pi) < r\}$ and $\overline{\mathcal{B}}_r(\pi) \stackrel{\text{def.}}{=} \{\mu \in \mathcal{P}(\mathcal{X}) | W_{2,\phi}(\mu, \pi) \leq r\}$, where $\mathcal{P}(\mathcal{X})$ is the space of probability distributions on $\mathcal{X}$.

Corollary 3 (Contracting to a Wasserstein ball) *Assume (A1)–(A5) hold with $\tilde{\kappa} < \sqrt{2m}$. Then the following statements hold:*

(i) For any $\mu_0 \in \mathcal{P}(\mathcal{X})$, there exist step-sizes $\{h_k\}_{k\in\mathbb{N}}$ such that $\limsup_{k\to\infty} W_{2,\phi}(\mu_k, \pi) \leq r_0$.

(ii) If $\mu_k \notin \overline{\mathcal{B}}_{r_0}(\pi)$, then there exists a step-size h_{k+1} such that $W_{2,\phi}(\mu_{k+1}, \pi) < W_{2,\phi}(\mu_k, \pi)$.

(iii) If $\mu_k \in \mathcal{B}_{r_0}(\pi)$, then there exists $h_{k+1} > 0$ such that $\mu_{k+1} \in \mathcal{B}_{r_0}(\pi)$.

(iv) If $\mu_k \in \overline{\mathcal{B}}_{r_0}(\pi) \setminus \mathcal{B}_{r_0}(\pi)$, then for any $r > r_0$, there exists $h_{k+1} > 0$, such that $\mu_{k+1} \in \mathcal{B}_r(\pi)$.

The proof can be found in Appendix D where we also construct an example of appropriate vanishing step-sizes $\{h_k\}_{k\in\mathbb{N}}$ that are in the order of $\frac{1}{k}$, and which guarantees that the claims of Corollary 3 hold.

Iteration complexity bounds. From these guarantees, for any $\varepsilon > 0$ small enough, we can now derive the smallest number of iterations K_ε (i.e., iteration complexity bound), such that the corresponding upper-bound of HRLMC with constant step-size is smaller than $r_0+\varepsilon$ after K_ε steps. More precisely, for any ε such that $0 < \varepsilon < \min\left(\frac{4\sqrt{2}p^{\frac{1}{2}}\beta_2}{m\sqrt{2m-\tilde{\kappa}^2}}, \frac{2\tilde{\kappa}^2 p^{\frac{1}{2}}\beta_1}{(2m-\tilde{\kappa}^2)^2}, \frac{32p^{\frac{1}{2}}\beta_2^2}{\tilde{\kappa}^2(4m-\tilde{\kappa}^2)^2\beta_1}\right)$, the number of iterations needed to get $W_{2,\phi}(\mu_k, \pi) < r_0 + \varepsilon$ with constant step-size is

$$K_\varepsilon \gtrsim \frac{pMR\left(\sqrt{M}+\kappa\right)^2}{(2m-\tilde{\kappa}^2)^3}\frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right).$$

When $\kappa = 0$, this becomes

$$K_\varepsilon \gtrsim \frac{p(m+M)^3M^2R}{(4m^2+4M(m-\delta)-\delta^2)^3}\frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right).$$

In the classical case when f is m-strongly convex and has an M-Lipschitz continuous gradient, the bound becomes

$$K_\varepsilon \gtrsim \frac{pM^2}{m^3\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right),$$

which coincides with the best result in the literature of Euler-Maruyama LMC (See (Durmus et al., 2019, Table 1) for an overview).

3.2. Baillon-Haddad type inequality

Baillon and Haddad showed that if the gradient of a convex and continuously differentiable function is nonexpansive, then it is firmly nonexpansive (Baillon and Haddad (1977)). This is one of the critical steps in the proof of convergence when $\phi(\mathbf{x}) = \|\mathbf{x}\|^2/2$. We extend the Baillon-Haddad theorem to the case of relative Lipschitz-smoothness (**A4**). We state a weaker version here, which is sufficient for the proof of the main theorem, and defer a stronger version with proof to the Appendix C, which is of independent interest.

Proposition 4 (Baillon-Haddad extension) *Assume f satisfies assumptions (**A3**)-(**A5**), then for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,*

$$\begin{aligned}&\langle \nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2), \nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\rangle \\ \geq & A\,\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2^2 + B\,\|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\|_2^2,\end{aligned} \tag{8}$$

where the constants are $A \stackrel{\text{def.}}{=} \frac{1}{m+M}$ and $B \stackrel{\text{def.}}{=} \frac{4mM-4M\delta-\delta^2}{4(m+M)}$.

3.3. Proof of Theorem 2

We first state a proposition that is useful in this section. Its proof is postponed to Appendix D.

Proposition 5 *Let $\mathbf{L}_0$ be any random vector drawn from π and $\mathbf{L}_t$ be a continuous dynamics satisfying (10). Then for any $s > 0$, one has*

$$\sqrt{\mathbf{E}\left[\|\nabla\phi(\mathbf{L}_0) - \nabla\phi(\mathbf{L}_s)\|_2^2\right]} \leq s\sqrt{MpR} + \sqrt{2spR}. \tag{9}$$

Proof of Theorem 2.

For notation simplicity, we use h, and ρ to represent h_{k+1}, and ρ_{k+1}, respectively. Let $\mathbf{L}_0$ be a random vector drawn from π such that $W_{2,\phi}^2(\mu_k, \pi) = \mathbf{E}\left[\|\nabla\phi(\mathbf{L}_0) - \nabla\phi(\mathbf{X}_k)\|_2^2\right]$. Let $\mathbf{B}_t = \sqrt{t}\boldsymbol{\xi}_{k+1}$, independent of $(\mathbf{X}_k, \mathbf{L}_0)$. Define a stochastic process $\mathbf{L}$ such that

$$\nabla\phi(\mathbf{L}_t) = \nabla\phi(\mathbf{L}_0) - \int_0^t \nabla f(\mathbf{L}_s)ds + \sqrt{2}\int_0^t [D^2\phi(\mathbf{L}_s)]^{\frac{1}{2}} d\mathbf{B}_s. \tag{10}$$

Then, by (**A1**), $\{\mathbf{L}_t : t \geq 0\}$ has π as its stationary distribution and $\mathbf{L}_t \sim \pi$ for all $t > 0$. On the other hand, our HRLMC algorithm reads

$$\nabla\phi(\mathbf{X}_{k+1}) = \nabla\phi(\mathbf{X}_k) - h\nabla f(\mathbf{X}_k) + \sqrt{2h[D^2\phi(\mathbf{X}_k)]}\boldsymbol{\xi}_{k+1}. \tag{11}$$

Let

$$\begin{aligned}\mathbf{A} &\stackrel{\text{def.}}{=} \nabla\phi(\mathbf{L}_0) - \nabla\phi(\mathbf{X}_k) - h(\nabla f(\mathbf{L}_0) - \nabla f(\mathbf{X}_k)),\\ \mathbf{C} &\stackrel{\text{def.}}{=} \int_0^h (\nabla f(\mathbf{L}_0) - \nabla f(\mathbf{L}_s))\,ds,\\ \mathbf{G} &\stackrel{\text{def.}}{=} \sqrt{2h}\left([D^2\phi(\mathbf{L}_0)]^{\frac{1}{2}} - [D^2\phi(\mathbf{X}_k)]^{\frac{1}{2}}\right)\boldsymbol{\xi}_{k+1},\end{aligned}$$

$$\mathbf{H} \overset{\text{def.}}{=} \sqrt{2} \int_0^h \left([D^2\phi(\mathbf{L}_s)]^{\frac{1}{2}} - [D^2\phi(\mathbf{L}_0)]^{\frac{1}{2}} \right) d\mathbf{B}_s.$$

By definition of $W_{2,\phi}^2$ and triangle inequality, one has

$$\begin{aligned} W_{2,\phi}(\mu_{k+1}, \pi) \leq & \sqrt{\mathbf{E}\left[\|\nabla\phi(\mathbf{L}_h) - \nabla\phi(\mathbf{X}_{k+1})\|_2^2 \right]} \\ = & \sqrt{\mathbf{E}\left[\|\mathbf{A} + \mathbf{C} + \mathbf{G} + \mathbf{H}\|_2^2 \right]} \\ \leq & \sqrt{\mathbf{E}\left[\|\mathbf{A} + \mathbf{G}\|_2^2 \right]} + \sqrt{\mathbf{E}\left[\|\mathbf{C}\|_2^2 \right]} + \sqrt{\mathbf{E}\left[\|\mathbf{H}\|_2^2 \right]}. \end{aligned} \tag{12}$$

Below, we estimate the three terms in the right-hand side separately.

1. Define $\rho = \sqrt{\tau^2 + h\kappa^2}$, where

$$\tau^2 = \begin{cases} (1-mh)^2 + \frac{h\delta(4M+\delta)}{2(m+M)}, \text{ for } h \in (0, \frac{2}{m+M}); \\ (1-Mh)^2 + \frac{h\delta(4M+\delta)}{2(m+M)}, \text{ for } h \in (\frac{2}{m+M}, \frac{2}{M}). \end{cases}$$

One can check that $\rho < 1$ because of $\tilde{\kappa}^2 < 2m$ and $h < \min\left(\frac{2m-\tilde{\kappa}^2}{m^2}, \frac{2M-\tilde{\kappa}^2}{M^2}\right)$. Therefore, by Proposition 4, we have

$$\begin{aligned} \mathbf{E}\left[\|\mathbf{A}\|_2^2 \right] = & \mathbf{E}\Big[\|\nabla\phi(\mathbf{L}_0) - \nabla\phi(\mathbf{X}_k)\|_2^2 + h^2 \|\nabla f(\mathbf{L}_0) - \nabla f(\mathbf{X}_k)\|_2^2 \\ & - 2h\langle \nabla f(\mathbf{L}_0) - \nabla f(\mathbf{X}_k), \nabla\phi(\mathbf{L}_0) - \nabla\phi(\mathbf{X}_k)\rangle \Big] \\ \leq & \mathbf{E}\Big[\left(1 - \frac{h(4mM - 4M\delta - \delta^2)}{2(m+M)}\right) \|\nabla\phi(\mathbf{L}_0) - \nabla\phi(\mathbf{X}_k)\|_2^2 \\ & + h\left(h - \frac{2}{m+M}\right) \|\nabla f(\mathbf{L}_0) - \nabla f(\mathbf{X}_k)\|_2^2 \Big] \\ \leq & \tau^2 W_{2,\phi}^2(\mu_k, \pi). \end{aligned} \tag{13}$$

The last inequality is derived from (**A4**) if $h \in \left(\frac{2}{m+M}, \frac{2}{M}\right)$ or (**A3**) if $h \in \left(0, \frac{2}{m+M}\right)$.

On the other hand, from Itô's isometry (Lemma 18) and assumption (**A1**), we have

$$\begin{aligned} \mathbf{E}[\|\mathbf{G}\|_2^2] = & \mathbf{E}\left[h \left\| \sqrt{2}\left([D^2\phi(\mathbf{L}_0)]^{\frac{1}{2}} - [D^2\phi(\mathbf{X}_k)]^{\frac{1}{2}} \right) \right\|_F^2 \right] \\ \leq & h\mathbf{E}\left[\kappa^2 \|\nabla\phi(\mathbf{L}_0) - \nabla\phi(\mathbf{X}_k)\|_2^2 \right] \\ = & h\kappa^2 W_{2,\phi}^2(\mu_k, \pi). \end{aligned} \tag{14}$$

Note that $\mathbf{E}[\langle \mathbf{A}, \mathbf{G}\rangle] = 0$, since $\boldsymbol{\xi}_{k+1}$ is independent of $(\mathbf{X}_k, \mathbf{L}_0)$. Therefore, combining equations (13) and (14), one has

$$\sqrt{\mathbf{E}\left[\|\mathbf{A} + \mathbf{G}\|_2^2 \right]} = \sqrt{\mathbf{E}\left[\|\mathbf{A}\|_2^2 + \|\mathbf{G}\|_2^2 \right]} \leq \sqrt{(\tau^2 + h\kappa^2)} W_{2,\phi}(\mu_k, \pi) = \rho W_{2,\phi}(\mu_k, \pi). \tag{15}$$

2. Applying Minkowski's integral inequality (Lemma 19), assumption (**A4**), and Proposition 5,

$$\begin{aligned}
\sqrt{\mathbf{E}\left[\|\mathbf{C}\|_2^2\right]} &\leq \int_0^h \sqrt{\mathbf{E}\left[\|\nabla f(\mathbf{L}_0) - \nabla f(\mathbf{L}_s)\|_2^2\right]} ds \\
&\leq M \int_0^h \sqrt{\mathbf{E}\left[\|\nabla \phi(\mathbf{L}_0) - \nabla \phi(\mathbf{L}_s)\|_2^2\right]} ds \\
&\leq M \int_0^h \left(s\sqrt{MpR} + \sqrt{2spR}\right) ds \\
&\leq \frac{7\sqrt{2}}{6} M h^{\frac{3}{2}} p^{\frac{1}{2}} R^{\frac{1}{2}}.
\end{aligned}$$

3. By Itô's isometry, assumption (**A1**), and Proposition 5,

$$\begin{aligned}
\mathbf{E}\left[\|\mathbf{H}\|_2^2\right] &= \int_0^h \mathbf{E}\left[\left\|\sqrt{2}\left([D^2\phi(\mathbf{L}_s)]^{\frac{1}{2}} - [D^2\phi(\mathbf{L}_0)]^{\frac{1}{2}}\right)\right\|_F^2\right] ds \\
&\leq \kappa^2 \int_0^h \mathbf{E}\left[\|\nabla \phi(\mathbf{L}_s) - \nabla \phi(\mathbf{L}_0)\|_2^2\right] ds \\
&\leq \kappa^2 \int_0^h \left(s\sqrt{MpR} + \sqrt{2spR}\right)^2 ds \\
&\leq \kappa^2 h^2 pR \left(1 + \sqrt{\frac{M}{3}} h^{\frac{1}{2}}\right)^2.
\end{aligned}$$

In conclusion, combining (12) and the above, we arrive at

$$\begin{aligned}
W_{2,\phi}(\mu_{k+1}, \pi) &\leq \sqrt{\mathbf{E}\left[\|\mathbf{A} + \mathbf{G}\|_2^2\right]} + \sqrt{\mathbf{E}\left[\|\mathbf{C}\|_2^2\right]} + \sqrt{\mathbf{E}\left[\|\mathbf{H}\|_2^2\right]} \\
&\leq \rho W_{2,\phi}(\mu_k, \pi) + \frac{7\sqrt{2}}{6} M h^{\frac{3}{2}} p^{\frac{1}{2}} R^{\frac{1}{2}} + \kappa h p^{\frac{1}{2}} R^{\frac{1}{2}} + \sqrt{\frac{M}{3}} \kappa h^{\frac{3}{2}} p^{\frac{1}{2}} R^{\frac{1}{2}} \\
&= \rho W_{2,\phi}(\mu_k, \pi) + h p^{\frac{1}{2}} \beta_1(R, \kappa) + h^{\frac{3}{2}} p^{\frac{1}{2}} \beta_2(R, M, \kappa).
\end{aligned}$$

■

Conclusion

In this paper, we have proposed the first theoretical guarantees for the discretized Langevin counterpart of the celebrated mirror descent algorithm to sample from distributions whose densities are not necessarily log-conncave nor log-Lipschitz-smooth. We showed that it is a stable discretization of the continuous Riemannian Langevin flow, more precisely, that it contracts toward a Wasserstein ball associated with a Hessian squared Riemannian metric. This analysis highlights the critical role played by the self-concordance of the entropy function and the relative anisotropy of the entropy and log-distribution (controlled by bounding the associated commutator).

Acknowledgments

The authors are grateful to Arnak Dalalyan, Marco Cuturi, and Paul Rolland for stimulating conversations. They also thank Institut Henri Poincaré for their hospitality throughout "The Mathematics of Imaging" semester program, during which the collaboration started. The work of the first two authors is supported by the ERC project NORIA.

References

Anton Arnold, Peter Markowich, Giuseppe Toscani, and Andreas Unterreiter. On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations. *Communications in Partial Differential Equations*, 26(1–2):43–100, 2001.

Jean-Bernard Baillon and Georges Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel J. Math.*, 26(2):137–150, 1977. ISSN 0021-2172. doi: 10.1007/BF03007664. URL https://doi.org/10.1007/BF03007664.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften book series*. Springer, 2014.

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. *Proceedings of the 2017 Conference on Learning Theory*, 65:319–342, 2017.

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

Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. arXiv1905.13142, 2019.

Xiang Cheng and Peter L Bartlett. Convergence of Langevin MCMC in KL-divergence. In *COLT*, volume 83, pages 186–211. PMLR, 2018.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Kai Lai Chung. On a stochastic approximation method. *Ann. Math. Statist.*, 25(3):463–483, 09 1954. doi: 10.1214/aoms/1177728716. URL https://doi.org/10.1214/aoms/1177728716.

Gyula Csató, Bernard Dacorogna, and Olivier Kneuss. *The pullback equation for differential forms*, volume 83. Springer Science & Business Media, 2011.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017a.

Arnak S Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. *arXiv preprint arXiv:1704.04752*, 2017b.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. arXiv preprint arXiv:1807.09382, 2018.

Arnak S Dalalyan, Lionel Riou-Durand, and Avetik Karagulyan. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. arXiv preprint arXiv:1807.09382, 2019.

Manfredo Perdigao do Carmo. *Riemannian geometry*. Birkhäuser, 1992.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

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

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov Chain Monte Carlo: when Langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *COLT*, pages 793–797, 2018.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3-4):851–886, 2016.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 2878–2887, 2018.

John Kent. Time-reversible diffusions. *Advances in Applied Probability*, 10(4):819–835, 1978.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Tung Duy Luu, Jalal Fadili, and Christophe Chesneau. Sampling from non-smooth distribution through Langevin diffusion. hal-01492056, March 2017. URL https://hal.archives-ouvertes.fr/hal-01492056.

Mateusz B Majka, Aleksandar Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without logconcavity. arXiv preprint arXiv:1808.07105, 2018.

Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley, New York, 1983.

Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in neural information processing systems*, pages 3102–3110, 2013.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *COLT*, volume 65, pages 1674–1703. PMLR, 2017. URL http://proceedings.mlr.press/v65/raginsky17a.html.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Hirohiko Shima. *The geometry of Hessian strutcures*. World Scientific Publishing, 2007.

Elias M Stein. *Singular integrals and differentiability properties of functions*, volume 2. Princeton university press, 1970.

Czeslaw Stępniak. Two orderings on a convex cone of nonnegative definite matrices. *Linear Algebra and its Applications*, 94:263–272, 1987.

Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes*. Springer, 2007.

Ying Zhang, Ömer Deniz Akyildiz, Theo Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization. arXiv1910.02008, 2019.

Appendix A. Well-posedness of (2)

Let us recall the SDE (2),

$$d\mathbf{Y}_t = -\nabla f \circ \nabla \phi^*(\mathbf{Y}_t)dt + \sqrt{2[D^2\phi^*(\mathbf{Y}_t)]^{-1}}d\mathbf{B}_t.$$

Let $\mathcal{Y} \stackrel{\text{def.}}{=} \nabla\phi(\mathcal{X})$. The following conditions are usually required for existence and uniqueness of (strong) solutions to this SDE in time interval $[0, T]$ (see (Øksendal, 2003, Theorem 5.2.1)):

- **Lipschitz condition:** there exists $K_1 > 0$, such that for all vectors $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ (and all $t \in [0, T]$),

$$\sqrt{2}\left\|D^2\phi^*(\mathbf{y}_1)^{-\frac{1}{2}} - D^2\phi^*(\mathbf{y}_2)^{-\frac{1}{2}}\right\|_F + \|\nabla f(\nabla\phi^*(\mathbf{y}_1)) - \nabla f(\nabla\phi^*(\mathbf{y}_2))\|_2 \le K_1 \|\mathbf{y}_1 - \mathbf{y}_2\|_2 . \tag{16}$$

  Let $\mathbf{x}_i = \nabla\phi^*(\mathbf{y}_i)$ for $i = 1, 2$. Then the above inequality is equivalent to, for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$\sqrt{2}\left\|D^2\phi(\mathbf{x}_1)^{\frac{1}{2}} - D^2\phi(\mathbf{x}_2)^{\frac{1}{2}}\right\|_F + \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 \le K_1 \|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\|_2 .$$

  In view of assumptions (**A1**) and (**A4**), the Lipschitz condition (16) holds with $K_1 = M + \kappa$.

- **Growth condition:** there exist $K_2 > 0$, such that for all $\mathbf{y} \in \mathcal{Y}$ (and $t \in [0, T]$),

$$2\left\|[D^2\phi^*(\mathbf{y})]^{-\frac{1}{2}}\right\|_F^2 + \|\nabla f \circ \nabla\phi^*(\mathbf{y})\|_2^2 \le K_2(1 + \|\mathbf{y}\|_2^2). \tag{17}$$

  Similarly, this is equivalent to the existence of $K_2 > 0$ such that for all $\mathbf{x} \in \mathcal{X}$,

$$2\left\|[D^2\phi(\mathbf{x})]^{\frac{1}{2}}\right\|_F^2 + \|\nabla f(\mathbf{x})\|_2^2 \le K_2(1 + \|\nabla\phi(\mathbf{x})\|_2^2).$$

  Again, owing to (**A1**) and (**A4**), one easily sees that (17) holds with K_2 depending on M and κ.

Remark 6 *Although the Lipschitz and Growth conditions are general requirements to guarantee the existence and uniqueness of solutions to SDE* (2)*, one can easily check that the Lipschitz condition implies the other one.*

Remark 7 *Examples of entropies ϕ verifying for instance* (**A1**) *are given in the text, e.g., Burg's entropy $\phi(x) = -\log(x)$ on $\mathbb{R}_{++}$. However, this does hold for the Boltzmann-Shannon $\phi(x) = x\log(x)$ on $\mathbb{R}_{++}$.*

Appendix B. Assumption (A3) v.s. relative strong convexity; and (A4) v.s. relative smoothness

Throughout, f and ϕ are assumed $C^2(\mathcal{X})$. By Cauchy-Schwarz inequality, (A3) implies

$$\exists m \geq 0, \text{ s.t. } m \left\|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{x}')\right\|_2 \leq \left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\right\|_2 \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{18}$$

Since $\mathcal{X}$ is open, for any $\mathbf{x} \in \mathcal{X}$, (18) and (A4) implies that for all $\mathbf{z} \in \mathbb{R}^p$ and t sufficiently small

$$m \left\|\nabla\phi(\mathbf{x} + t\mathbf{z}) - \nabla\phi(\mathbf{x})\right\|_2 \leq \left\|\nabla f(\mathbf{x} + t\mathbf{z}) - \nabla f(\mathbf{x})\right\|_2 \leq M \left\|\nabla\phi(\mathbf{x} + t\mathbf{z}) - \nabla\phi(\mathbf{x})\right\|_2.$$

Dividing by t and passing to the limit as $t \to 0^+$, we get

$$m \left\|D^2\phi(\mathbf{x})\mathbf{z}\right\|_2 \leq \left\|D^2 f(\mathbf{x})\mathbf{z}\right\|_2 \leq M \left\|D^2\phi(\mathbf{x})\mathbf{z}\right\|_2, \quad \forall \mathbf{x} \in \mathcal{X}, \forall \mathbf{z} \in \mathbb{R}^p.$$

Squaring, this is equivalent to

$$m^2 \left\langle (D^2\phi(\mathbf{x}))^2\mathbf{z}, \mathbf{z}\right\rangle \leq \left\langle (D^2 f(\mathbf{x}))^2\mathbf{z}, \mathbf{z}\right\rangle \leq M^2 \left\langle (D^2\phi(\mathbf{x}))^2\mathbf{z}, \mathbf{z}\right\rangle, \quad \forall \mathbf{x} \in \mathcal{X}, \forall \mathbf{z} \in \mathbb{R}^p, \tag{19}$$

or

$$(mD^2\phi(\mathbf{x}))^2 \preceq (D^2 f(\mathbf{x}))^2 \preceq (MD^2\phi(\mathbf{x}))^2, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{20}$$

where $\preceq$ is the Loewner order defined by the cone of positive semi-definite matrices. We recall the following lemma due to (Stępniak, 1987, Theorem 1).

Lemma 8 *For any positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$, if $\mathbf{A}^2 \succeq \mathbf{B}^2$, then $\mathbf{A} \succeq \mathbf{B}$.*

Applying this lemma with $\mathbf{A} = MD^2\phi(\mathbf{x})$ and $\mathbf{B} = D^2 f(\mathbf{x})$, and then with $\mathbf{A} = D^2 f(\mathbf{x})$ and $\mathbf{B} = mD^2\phi(\mathbf{x})$, we conclude that (20) implies

$$mD^2\phi(\mathbf{x}) \preceq D^2 f(\mathbf{x}) \preceq MD^2\phi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}. \tag{21}$$

According to (Bauschke et al., 2017, Proposition 1.(i, ii)), (21) is equivalent to smoothness and strong convexity of f relatively to ϕ, as defined in Lu et al. (2018).

Overall, we have proved the following claim.

Proposition 9 *Suppose that f and ϕ are $C^2(\mathcal{X})$. Then* (A3) *implies m-strong relative convexity with respect to ϕ and* (A4) *implies M-relative smoothness of f with respect to ϕ, i.e.* (21) *holds.*

Observe that the converse implication in Lemma 8 does not hold in general, see Stępniak (1987), and thus (21) $\not\Rightarrow$ (20) in general. In turn assumptions (A3) and (A4) are strictly stronger than relative smoothness and strong convexity.

Appendix C. Proof of a stronger version of Proposition 4, the Baillon-Haddad type inequality

In this section, we will prove a Baillon-Haddad type inequality, as in Proposition 4, but with weaker assumptions. This inequality serves as an essential step in the proof of Theorem 2.

In the following, we denote by $\mathcal{M}_{l\times n}$ the space of all matrices that have l rows and n columns and whose entries have real values.

Lemma 10 ((Horn and Johnson, 2012, Example 5.6.6)) *For any matrix $\mathbf{M} \in \mathcal{M}_{l\times n}$, $\|\mathbf{M}\|_2 = \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\|\mathbf{M}\mathbf{v}\|_2}{\|\mathbf{v}\|_2}$.*

Remark 11 *From the above lemma, it is clear that $\|\mathbf{M}_1\mathbf{M}_2\|_2 \le \|\mathbf{M}_1\|_2 \|\mathbf{M}_2\|_2$ for any $\mathbf{M}_1 \in \mathcal{M}_{k\times l}$ and $\mathbf{M}_2 \in \mathcal{M}_{l\times n}$.*

Definition 12 (Contractibility) *We say a domain $\mathcal{U} \subset \mathbb{R}^p$ is contractible if there exists some point $\mathbf{c} \in \mathcal{U}$ such that the constant map $\mathbf{x} \mapsto \mathbf{c}$ is homotopic to the identity map on $\mathcal{U}$.*

Definition 13 (Differential Forms) *Let $0 \le k \le p$. A differential k-form $g : \mathcal{U} \to \Lambda^k$ will be written as $g = \sum_{1\le i_1<\cdots<i_k\le p} g_{i_1\cdots i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k}$, where $g_{i_1\cdots i_k} : \mathcal{U} \to \mathbb{R}$ for every $1 \le i_1 < \cdots < i_k \le p$ and $\Lambda^k = \Lambda^k(\mathbb{R}^{p*})$ with $\mathbb{R}^{p*}$ being the dual of $\mathbb{R}^p$ as a vector space. When $g_{i_1\cdots i_k} \in C^r(\mathcal{U})$ for every $1 \le i_1 < \cdots < i_k \le p$, we will write $g \in C^r(\mathcal{U};\Lambda^k)$.*

Lemma 14 (Poincaré lemma, (Csató et al., 2011, Theorem 8.1)) *Let $r \ge 1$ and $0 \le k \le p-1$ be integers and $\mathcal{U} \subset \mathbb{R}^p$ be an open contractible set. Let $g \in C^r(\mathcal{U};\Lambda^{k+1})$ with $dg = 0$ in $\mathcal{U}$. Then there exists $G \in C^r(\mathcal{U};\Lambda^k)$ such that $dG = g$ in $\mathcal{U}$.*

Remark 15 *For relaxation on the contractibility of the domain and sharper regularity in Hölder spaces, see (Csató et al., 2011, Theorem 8.3).*

Proposition 16 (Baillon-Haddad extension) *Assume that $\mathcal{X}$ is contractible, ϕ is a Legendre function on $\mathcal{X}$, f and $\phi \in C^3(\mathcal{X})$ satisfying (**A5**), and that there exist $0 \le m \le M$ such that for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,*

$$m \|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\|_2^2 \le \langle \nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2), \nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\rangle \le M \|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\|_2^2. \tag{22}$$

Then for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have

$$\begin{aligned} &\langle \nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2), \nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\rangle \\ \ge& \frac{1}{m+M} \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2^2 + \frac{4mM - 4M\delta - \delta^2}{4(m+M)} \|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\|_2^2. \end{aligned} \tag{23}$$

Remark 17 *1. Under the same assumptions as above and assuming $D^2\phi$ and $D^2 f$ are commutable, then for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,*

$$\begin{aligned}&\langle \nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2), \nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\rangle \\ \geq & \frac{1}{m+M}\left\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\right\|_2^2 + \frac{mM}{m+M}\left\|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\right\|_2^2.\end{aligned} \tag{24}$$

2. If, in addition, $m = 0$, then the inequality becomes

$$\langle \nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2), \nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\rangle \geq \frac{1}{M}\left\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\right\|_2^2. \tag{25}$$

This is the canonical form of Baillon-Haddad inequality, which is equivalent to equation (24).
3. In general, if $m = 0$ (but δ may not), the inequality (23) implies relative Lipschitz smoothness

$$\left\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\right\|_2 \leq \left(M + \frac{\delta}{2}\right)\left\|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\right\|_2. \tag{26}$$

Proof [Proposition 16] Denote $\mathbf{A}(\mathbf{y}) := D^2 f(\nabla\phi^*(\mathbf{y}))$ and $\mathbf{B}(\mathbf{y}) := D^2\phi^*(\mathbf{y})$.

Notice that

$$\begin{aligned}& d\left(\frac{1}{2}\sum_{i,j}\left[(\mathbf{AB})_{ji} - (\mathbf{AB})_{ij}\right] dy_i \wedge dy_j\right) \\ =&\sum_{i,j,l}\frac{1}{2} d\left(\partial_{jl} f(\nabla\phi^*)\partial_{li}\phi^* - \partial_{il} f(\nabla\phi^*)\partial_{lj}\phi^*\right) \wedge dy_i \wedge dy_j \\ =&\sum_{i,j,k,l}\frac{1}{2}\Bigg(\partial_{jl} f(\nabla\phi^*)\partial_{lik}\phi^* - \partial_{il} f(\nabla\phi^*)\partial_{ljk}\phi^* + \sum_m \partial_{jlm} f(\nabla\phi^*)\partial_{mk}\phi^*\partial_{li}\phi^* - \\ & \qquad - \sum_m \partial_{ilm} f(\nabla\phi^*)\partial_{mk}\phi^*\partial_{lj}\phi^*\Bigg) dy_k \wedge dy_i \wedge dy_j \\ =&\sum_{i,j,k,l}\frac{1}{6}\cdot 0\; dy_k \wedge dy_i \wedge dy_j + \sum_{i,j,k,l,m}\frac{1}{6}\cdot 0\; dy_k \wedge dy_i \wedge dy_j \\ =&0.\end{aligned}$$

By the Poincaré lemma, there exists a 1-form ω on $\mathcal{Y}$ such that

$$d\omega = \frac{1}{2}\sum_{i,j}\left[(\mathbf{AB})_{ji} - (\mathbf{AB})_{ij}\right] dy_i \wedge dy_j \tag{27}$$

Note that ω is a 1-form on $\mathcal{Y}$, which corresponds to a vector field $\mathbf{g} : \mathcal{Y} \to \mathbb{R}^p$ such that $\omega = \mathbf{g} \cdot d\mathbf{y}$. Define $\tilde{\mathbf{g}} := \nabla f \circ \nabla\phi^* - \mathbf{g} : \mathcal{Y} \to \mathbb{R}^p$.

By Stokes-Cartan theorem, for any $\mathcal{U} \subset \mathcal{Y}$, one has

$$\begin{aligned}\int_{\partial\mathcal{U}} \nabla f \circ \nabla\phi^* \cdot d\mathbf{y} =& \int_{\mathcal{U}} d\left(\sum_{j=1} \partial_j f(\nabla\phi^*) dy_j\right) \\ =& \frac{1}{2}\int_{\mathcal{U}} \sum_{i,j=1}^{p} \left[(\mathbf{AB})_{ji} - (\mathbf{AB})_{ij}\right] dy_i \wedge dy_j \\ =& \int_{\mathcal{U}} d\omega = \int_{\partial\mathcal{U}} \omega = \int_{\partial\mathcal{U}} \mathbf{g} \cdot d\mathbf{y}.\end{aligned}$$

This implies, for any closed curve Γ on $\mathcal{Y}$, one has

$$\oint_{\Gamma} \tilde{\mathbf{g}} \cdot d\mathbf{y} = 0.$$

That is, $\tilde{\mathbf{g}}$ is path-independent. Define $\tilde{f}$ as a function on $\mathcal{Y}$ from any given point $\mathbf{y}_0 \in \mathcal{Y}$ such that $\tilde{f}(\mathbf{y}) \stackrel{\text{def.}}{=} \tilde{f}(\mathbf{y}_0) + \int_\Gamma \tilde{\mathbf{g}} \cdot d\mathbf{y}$, where Γ is any smooth curve from $\mathbf{y}_0$ to $\mathbf{y}$. Therefore,

$$\nabla \tilde{f} = \tilde{\mathbf{g}} = \nabla f \circ \nabla\phi^* - \mathbf{g}. \tag{28}$$

From (27), we know $\partial_i g_j = \frac{1}{2}\left[(\mathbf{AB})_{ji} - (\mathbf{AB})_{ij}\right]$, for all $1 \le i, j \le p$. Thus, (28) implies

$$\begin{aligned}(D^2\tilde{f})_{ji} =& \partial_i\partial_j\tilde{f} = \partial_i(\partial_j f(\nabla\phi^*) - g_j) = \sum_k \partial_{jk} f(\nabla\phi^*) \cdot \partial_{ki}\phi^* - \partial_i g_j \\ =& (\mathbf{BA})_{ij} + \frac{1}{2}[(\mathbf{AB})_{ij} - (\mathbf{BA})_{ij}] = \frac{1}{2}[(\mathbf{AB})_{ij} + (\mathbf{BA})_{ij}].\end{aligned}$$

This shows that $D^2\tilde{f}$ is symmetric and

$$D^2\tilde{f} = \frac{1}{2}(\mathbf{AB} + \mathbf{BA}) = \frac{1}{2}\left(D^2 f \circ \nabla\phi^* \cdot D^2\phi^* + D^2\phi^* D^2 f \circ \nabla\phi^*\right). \tag{29}$$

By assumption, there exist $0 \le m \le M$ such that for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$m\left\|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\right\|_2^2 \le \langle \nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2), \nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\rangle \le M\left\|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\right\|_2^2.$$

This implies for any $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$,

$$m\left\|\mathbf{y}_1 - \mathbf{y}_2\right\|_2^2 \le \langle \nabla f(\nabla\phi^*(\mathbf{y}_1)) - \nabla f(\nabla\phi^*(\mathbf{y}_2)), \mathbf{y}_1 - \mathbf{y}_2\rangle \le M\left\|\mathbf{y}_1 - \mathbf{y}_2\right\|_2^2.$$

Thus, for any $\mathbf{v} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathcal{Y}$, one has

$$m\left\|\mathbf{v}\right\|_2^2 \le \mathbf{v}^T \frac{[D(\nabla f \circ \nabla\phi^*)(\mathbf{y})]^T + [D(\nabla f \circ \nabla\phi^*)(\mathbf{y})]}{2}\mathbf{v} \le M\left\|\mathbf{v}\right\|_2^2.$$

This reads, from (29),

$$m\mathbf{I}_p \preceq D^2\tilde{f}(\mathbf{y}) \preceq M\mathbf{I}_p$$

for all $\mathbf{y} \in \mathcal{Y}$. By the classical Baillon-Haddad theorem, we know

$$\langle \nabla\tilde{f}(\mathbf{y}_1) - \nabla\tilde{f}(\mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2\rangle \ge \frac{1}{m+M}\left\|\nabla\tilde{f}(\mathbf{y}_1) - \nabla\tilde{f}(\mathbf{y}_2)\right\|_2^2 + \frac{mM}{m+M}\left\|\mathbf{y}_1 - \mathbf{y}_2\right\|_2^2. \tag{30}$$

Now let us estimate $\langle \mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2\rangle$, $\langle \nabla\tilde{f}(\mathbf{y}_1) - \nabla\tilde{f}(\mathbf{y}_2), \mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2)\rangle$, and $\|\mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2)\|_2^2$.

1. For any $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, and any $t, s \in [0,1]$, denote $\mathbf{y}_t = t\mathbf{y}_1 + (1-t)\mathbf{y}_2$ and $\mathbf{y}_s = s\mathbf{y}_1 + (1-s)\mathbf{y}_2$. Then $\mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2) = \int_0^1 d(\mathbf{g}(\mathbf{y}_t)) = \int_0^1 \nabla \mathbf{g}(\mathbf{y}_t) \cdot (\mathbf{y}_1 - \mathbf{y}_2) dt$. Since $\nabla \mathbf{g}(\mathbf{y}_t)$ is anti-symmetric,

$$\begin{aligned}\langle \mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2 \rangle &= \int_0^1 (\mathbf{y}_1 - \mathbf{y}_2)^T [\nabla \mathbf{g}(\mathbf{y}_t)]^T (\mathbf{y}_1 - \mathbf{y}_2) dt \\ &= \frac{1}{2} \int_0^1 (\mathbf{y}_1 - \mathbf{y}_2)^T [(\nabla \mathbf{g}(\mathbf{y}_t))^T + \nabla \mathbf{g}(\mathbf{y}_t)] (\mathbf{y}_1 - \mathbf{y}_2) dt = 0.\end{aligned} \tag{31}$$

2. As follows, for any $t \in [0,1]$, let $\mathbf{C}(t) := D^2 f(\nabla \phi^*(\mathbf{y}_t)) D^2 \phi^*(\mathbf{y}_t) = \mathbf{A}(\mathbf{y}_t)\mathbf{B}(\mathbf{y}_t)$. Then, by assumption, $\left\| \mathbf{C}(t)^T - \mathbf{C}(t) \right\|_2 \le \delta$ for all $t \in [0,1]$.

   Therefore,

$$\begin{aligned}&\langle \nabla \tilde{f}(\mathbf{y}_1) - \nabla \tilde{f}(\mathbf{y}_2), \mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2) \rangle \\ =& \sum_{l=1}^p (\partial_l \tilde{f}(\mathbf{y}_1) - \partial_l \tilde{f}(\mathbf{y}_2)) \cdot (g_l(\mathbf{y}_1) - g_l(\mathbf{y}_2)) \\ =& \sum_{l=1}^p \int_0^1 d(\partial_l \tilde{f}(\mathbf{y}_t)) \cdot \int_0^1 d(g_l(\mathbf{y}_s)) \\ =& \sum_{l=1}^p \int_0^1 \sum_i \partial_{il} \tilde{f}(\mathbf{y}_t) \cdot (\mathbf{y}_1 - \mathbf{y}_2)_i dt \cdot \int_0^1 \sum_j \partial_j g_l(\mathbf{y}_s) \cdot (\mathbf{y}_1 - \mathbf{y}_2)_j ds \\ =& \int_0^1 \int_0^1 \sum_{i,j,l} (\mathbf{y}_1 - \mathbf{y}_2)_i \cdot \partial_{il} \tilde{f}(\mathbf{y}_t) \cdot \partial_j g_l(\mathbf{y}_s) \cdot (\mathbf{y}_1 - \mathbf{y}_2)_j ds dt \\ =& \int_0^1 \int_0^1 \sum_{i,j,l} (\mathbf{y}_1 - \mathbf{y}_2)_i \cdot \frac{(\mathbf{A}(\mathbf{y}_t)\mathbf{B}(\mathbf{y}_t) + \mathbf{B}(\mathbf{y}_t)\mathbf{A}(\mathbf{y}_t))_{il}}{2} \\ & \qquad \cdot \frac{(\mathbf{A}(\mathbf{y}_s)\mathbf{B}(\mathbf{y}_s) - \mathbf{B}(\mathbf{y}_s)\mathbf{A}(\mathbf{y}_s))_{lj}}{2} \cdot (\mathbf{y}_1 - \mathbf{y}_2)_j ds dt \\ =& \frac{1}{4} \int_0^1 \int_0^1 (\mathbf{y}_1 - \mathbf{y}_2)^T \left[\left(\mathbf{C}(t) + \mathbf{C}(t)^T \right) \left(\mathbf{C}(s) - \mathbf{C}(s)^T \right) \right] (\mathbf{y}_1 - \mathbf{y}_2) ds dt.\end{aligned}$$

   Notice that

$$\left\| \left(\mathbf{C}(t) + \mathbf{C}(t)^T \right) \left(\mathbf{C}(s) - \mathbf{C}(s)^T \right) \right\|_2 \le \left\| \mathbf{C}(t) + \mathbf{C}(t)^T \right\|_2 \left\| \mathbf{C}(s) - \mathbf{C}(s)^T \right\|_2 \le 2M\delta.$$

   Therefore,

$$\langle \nabla \tilde{f}(\mathbf{y}_1) - \nabla \tilde{f}(\mathbf{y}_2), \mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2) \rangle \le \frac{1}{4} \int_0^1 \int_0^1 2M\delta \left\| \mathbf{y}_1 - \mathbf{y}_2 \right\|_2^2 ds dt = \frac{1}{2} M\delta \left\| \mathbf{y}_1 - \mathbf{y}_2 \right\|_2^2. \tag{32}$$

3. Similarly, one has

$$\left\| \mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2) \right\|_2^2 = \frac{1}{4} \int_0^1 \int_0^1 (\mathbf{y}_1 - \mathbf{y}_2)^T \left[(\mathbf{C}(t)^T - \mathbf{C}(t))(\mathbf{C}(s) - \mathbf{C}(s)^T) \right] (\mathbf{y}_1 - \mathbf{y}_2) ds dt,$$

and

$$\left\|(\mathbf{C}(t)^T - \mathbf{C}(t))(\mathbf{C}(s) - \mathbf{C}(s)^T)\right\|_2 \leq \left\|\mathbf{C}(t)^T - \mathbf{C}(t)\right\|_2 \left\|\mathbf{C}(s) - \mathbf{C}(s)^T\right\|_2 \leq \delta^2.$$

Thus,

$$\|\mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2)\|_2^2 \leq \frac{1}{8}\int_0^1 \int_0^1 2\delta^2 \|\mathbf{y}_1 - \mathbf{y}_2\|_2^2 \, dsdt = \frac{\delta^2}{4} \|\mathbf{y}_1 - \mathbf{y}_2\|_2^2. \tag{33}$$

Combining equations (30)-(33), one has

$$\begin{aligned}
&\langle \nabla f \circ \nabla\phi^*(\mathbf{y}_1) - \nabla f \circ \nabla\phi^*(\mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2\rangle \\
=&\langle \nabla\tilde{f}(\mathbf{y}_1) - \nabla\tilde{f}(\mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2\rangle + \langle \mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2\rangle \\
=&\langle \nabla\tilde{f}(\mathbf{y}_1) - \nabla\tilde{f}(\mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2\rangle \\
\geq&\frac{1}{m+M}\left\|\nabla\tilde{f}(\mathbf{y}_1) - \nabla\tilde{f}(\mathbf{y}_2)\right\|_2^2 + \frac{mM}{m+M}\|\mathbf{y}_1 - \mathbf{y}_2\|_2^2 \\
\geq&\frac{1}{m+M}\|\nabla f \circ \nabla\phi^*(\mathbf{y}_1) - \nabla f \circ \nabla\phi^*(\mathbf{y}_2)\|_2^2 - \frac{1}{m+M}\|\mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2)\|_2^2 \\
&- \frac{2}{m+M}\langle \nabla\tilde{f}(\mathbf{y}_1) - \nabla\tilde{f}(\mathbf{y}_2), \mathbf{g}(\mathbf{y}_1) - \mathbf{g}(\mathbf{y}_2)\rangle + \frac{mM}{m+M}\|\mathbf{y}_1 - \mathbf{y}_2\|_2^2 \\
\geq&\frac{1}{m+M}\|\nabla f \circ \nabla\phi^*(\mathbf{y}_1) - \nabla f \circ \nabla\phi^*(\mathbf{y}_2)\|_2^2 + \frac{4mM - 4M\delta - \delta^2}{4(m+M)}\|\mathbf{y}_1 - \mathbf{y}_2\|_2^2.
\end{aligned}$$

By change of variables, this implies

$$\begin{aligned}
&\langle \nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2), \nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\rangle \\
\geq&\frac{1}{m+M}\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2^2 + \frac{4mM - 4M\delta - \delta^2}{4(m+M)}\|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\|_2^2.
\end{aligned} \tag{34}$$

■

Appendix D. Proof of Proposition 1, Corollary 3, and Proposition 5

In this section, we first recall two lemmas that are used in the proof of Theorem 2, followed by the proof of Proposition 1, Corollary 3, and Proposition 5. The Itô's isometry theorem can be found, for instance, in (Øksendal, 2003, Corollary 3.1.7) for the one-dimensional case. Here we state its apparent consequence in the multidimensional case.

Lemma 18 (Itô's isometry) *Let $\mathbf{B} : [0, T] \times \Omega \to \mathbb{R}^p$ be the standard p-dimensional Brownian motion and $\mathbf{M} : [0, T] \times \Omega \to \mathbb{R}^{p\times p}$ be a matrix-valued stochastic process adapted to the natural filtration of the Brownian motion. Then*

$$\mathbf{E}\left[\left\|\int_0^T \mathbf{M}_t d\mathbf{B}_t\right\|_2^2\right] = \mathbf{E}\left[\int_0^T \|\mathbf{M}_t\|_F^2 \, dt\right], \tag{35}$$

whenever the integrals make sense.

Lemma 19 (Minkowski's integral inequality, (Stein, 1970, Appendix A)) *Suppose that $(\mathcal{S}_1, \pi_1)$ and $(\mathcal{S}_2, \pi_2)$ are two σ-finite measure spaces, $l \geq 1$ and $f : \mathcal{S}_1 \times \mathcal{S}_2 \to \mathbb{R}_+$ is measurable, then*

$$\left\{\int_{\mathcal{S}_1}\left(\int_{\mathcal{S}_2} f(\mathbf{x},\mathbf{y})d\pi_2(\mathbf{y})\right)^l d\pi_1(\mathbf{x})\right\}^{\frac{1}{l}} \leq \int_{\mathcal{S}_2}\left(\int_{\mathcal{S}_1} f^l(\mathbf{x},\mathbf{y})d\pi_1(\mathbf{x})\right)^{\frac{1}{l}} d\pi_2(\mathbf{y}). \tag{36}$$

Remark 20 *Assume the same conditions as above, and $f_i : \mathcal{S}_1 \times \mathcal{S}_2 \to \mathbb{R}_+$ are measurable for $i = 1, ...p$, then*

$$\left\{\int_{\mathcal{S}_1}\sum_{i=1}^{p}\left(\int_{\mathcal{S}_2} f_i(\mathbf{x},\mathbf{y})d\pi_2(\mathbf{y})\right)^l d\pi_1(\mathbf{x})\right\}^{\frac{1}{l}} \leq \int_{\mathcal{S}_2}\left(\int_{\mathcal{S}_1}\sum_{i=1}^{p} f_i^l(\mathbf{x},\mathbf{y})d\pi_1(\mathbf{x})\right)^{\frac{1}{l}} d\pi_2(\mathbf{y}). \tag{37}$$

It can be viewed as Minkowski's inequality applying on $(\mathcal{S}_1 \times \{1, ..., p\}, \pi_1 \times \pi_3)$ and $(\mathcal{S}_2, \pi_2)$, where π_3 is uniform measure up to a constant multiplication.

Proof [Proposition 1] From Theorem 2, one has

$$\begin{aligned}
W_{2,\phi}(\mu_k, \pi) \leq & \rho W_{2,\phi}(\mu_{k-1}, \pi) + hp^{\frac{1}{2}}\beta_1 + h^{\frac{3}{2}}p^{\frac{1}{2}}\beta_2 \\
\leq & \rho \cdot (\rho W_{2,\phi}(\mu_{k-2}, \pi) + hp^{\frac{1}{2}}\beta_1 + h^{\frac{3}{2}}p^{\frac{1}{2}}\beta_2) + hp^{\frac{1}{2}}\beta_1 + h^{\frac{3}{2}}p^{\frac{1}{2}}\beta_2 \\
\leq & \cdots \\
\leq & \rho^k W_{2,\phi}(\mu_0, \pi) + (hp^{\frac{1}{2}}\beta_1 + h^{\frac{3}{2}}p^{\frac{1}{2}}\beta_2)(1 + \rho + \cdots + \rho^{k-1}) \\
= & \rho^k W_{2,\phi}(\mu_0, \pi) + (hp^{\frac{1}{2}}\beta_1 + h^{\frac{3}{2}}p^{\frac{1}{2}}\beta_2) \cdot \frac{1-\rho^k}{1-\rho} \\
< & \rho^k W_{2,\phi}(\mu_0, \pi) + \frac{hp^{\frac{1}{2}}\beta_1 + h^{\frac{3}{2}}p^{\frac{1}{2}}\beta_2}{1-\rho}.
\end{aligned}$$

The last inequality holds because $0 < \rho < 1$. ■

Lemma 21 ((Chung, 1954, Lemma 1)) *Let $\{w_k\}_{k\in\mathbb{N}}$ be a sequence of real numbers such that, for all k,*

$$w_{k+1} \leq \left(1 - \frac{c}{k}\right) w_k + \frac{c_1}{k^{s+1}}, \tag{38}$$

where $c > s > 0$, $c_1 > 0$. Then for any k,

$$w_k \leq c_1(c-s)^{-1}k^{-s} + o(k^{-s}). \tag{39}$$

Remark 22 *The same consequence (39) holds if c_1 is replaced by $c_1 + o(1)$.*

Proof [Corollary 3] 1. For any $0 < b_1 < \frac{2m-\tilde{\kappa}^2}{2}$, there exists $a_1 > 0$ such that $h_k = \frac{a_1}{k}$ is small enough and

$$\rho_k \leq 1 - b_1 h_k$$

for all $k \in \mathbb{N}$. Thus, from Theorem 2, we get

$$\begin{aligned} W_{2,\phi}(\mu_{k+1},\pi) &\leq \rho_{k+1} W_{2,\phi}(\mu_k,\pi) + \beta_2 p^{1/2} h_{k+1}^{3/2} + \beta_1 p^{1/2} h_{k+1} \\ &\leq (1 - b_1 h_{k+1}) W_{2,\phi}(\mu_k,\pi) + p^{1/2}(\beta_1 + o(1)) h_{k+1}. \end{aligned} \tag{40}$$

For any $0 < s < a_1 b_1$, set $w_k \overset{\text{def.}}{=} h_{k+1}^s W_{2,\phi}(\mu_k,\pi)$. Multiplying both sides of (40) by h_{k+2}^s, and using the fact that $\{h_k\}_{k\in\mathbb{N}}$ is a decreasing sequence, we get

$$w_{k+1} \leq \left(1 - \frac{a_1 b_1}{k+1}\right) w_k + \frac{a_1^{s+1} p^{1/2}(\beta_1 + o(1))}{(k+1)^{s+1}}. \tag{41}$$

Applying Lemma 21 with its Remark 22, we have

$$w_k \leq a_1^{s+1} p^{1/2}(\beta_1 + o(1))(a_1 b_1 - s)^{-1}(k+1)^{-s} + o((k+1)^{-s}).$$

From the definition of w_k, we deduce that

$$W_{2,\phi}(\mu_k,\pi) \leq a_1 p^{1/2}(\beta_1 + o(1))(a_1 b_1 - s)^{-1} + o(1) = a_1 p^{1/2} \beta_1 (a_1 b_1 - s)^{-1} + o(1).$$

In turn, we conclude that

$$\limsup_{k\to\infty} W_{2,\phi}(\mu_k,\pi) \leq a_1 p^{1/2} \beta_1 (a_1 b_1 - s)^{-1},$$

for any $0 < s < a_1 b_1$. Taking the limit at both sides when $s \to 0$, one has

$$\limsup_{k\to\infty} W_{2,\phi}(\mu_k,\pi) \leq p^{1/2} \beta_1 b_1^{-1}. \tag{42}$$

This implies that $W_{2,\phi}(\mu_k,\pi) h_{k+1}^2$ has the order $o(h_{k+1})$ whenever $h_k = \frac{a}{k}$ for $a \in (0, a_1]$.

Now let $b = \frac{2m-\tilde{\kappa}^2}{2}$. There exists $a \in (0, a_1]$ such that $h_k = \frac{a}{k}$ is small enough and

$$\rho_k \leq 1 - b h_k + \frac{m^2}{2} h_k^2$$

for all $k \in \mathbb{N}$. Theorem 2 then implies

$$\begin{aligned} W_{2,\phi}(\mu_{k+1},\pi) &\leq \left(1 - b h_{k+1} + \frac{m^2}{2} h_{k+1}^2\right) W_{2,\phi}(\mu_k,\pi) + \beta_2 p^{1/2} h_{k+1}^{3/2} + \beta_1 p^{1/2} h_{k+1} \\ &\leq (1 - b h_{k+1}) W_{2,\phi}(\mu_k,\pi) + p^{1/2}(\beta_1 + o(1)) h_{k+1}. \end{aligned} \tag{43}$$

Repeating the above argument by using Remark 22 gives $\limsup_{k\to\infty} W_{2,\phi}(\mu_k,\pi) \leq p^{1/2} \beta_1 b^{-1} = r_0$ as claimed. △

2. Let $\alpha_i = p^{\frac{1}{2}}\beta_i$ for $i = 1, 2$. Define a function $r : [0, \infty) \to \mathbb{R}$ such that $r(0) = r_0$ and for all $t > 0$,

$$r(t) \overset{\text{def.}}{=} \frac{t\alpha_1 + t^{\frac{3}{2}}\alpha_2}{1 - \sqrt{(1 - mt)^2 + \tilde{\kappa}^2 t}}. \tag{44}$$

One can check that its derivative $r'(t) > 0$ for all $0 < t < \min\left(\frac{2}{m+M}, \frac{2m-\tilde{\kappa}^2}{m^2}\right)$ and $\lim_{t\to 0^+} r(t) = r_0$. If $\mu_k \notin \overline{\mathcal{B}}_{r_0}(\pi)$, i.e., $W_{2,\phi}(\mu_k, \pi) > r_0$, by the continuity of r at 0, there exists $0 < h_{k+1} < \min\left(\frac{2m-\tilde{\kappa}^2}{m^2}, \frac{2M-\tilde{\kappa}^2}{M^2}, \frac{2}{m+M}\right)$ such that $W_{2,\phi}(\mu_k, \pi) > r(h_{k+1}) = \frac{h_{k+1}\alpha_1 + h_{k+1}^{\frac{3}{2}}\alpha_2}{1-\rho_{k+1}}$. For the μ_{k+1} obtained from the algorithm (1), by Theorem 2, we know

$$\begin{aligned} W_{2,\phi}(\mu_{k+1}, \pi) \leq& \rho_{k+1} W_{2,\phi}(\mu_k, \pi) + h_{k+1}\alpha_1 + h_{k+1}^{\frac{3}{2}}\alpha_2 \\ <& \rho_{k+1} W_{2,\phi}(\mu_k, \pi) + (1 - \rho_{k+1}) W_{2,\phi}(\mu_k, \pi) \\ =& W_{2,\phi}(\mu_k, \pi). \end{aligned} \tag{45}$$

That is, the distance is strictly decreasing. △

3. If $\mu_k \in \mathcal{B}_{r_0}(\pi)$, the function

$$\sqrt{(1 - mt)^2 + \tilde{\kappa}^2 t}\,(W_{2,\phi}(\mu_k, \pi) - r_0) + t\alpha_1 + t^{\frac{3}{2}}\alpha_2$$

is continuous in t and negative at $t = 0$. Thus there exists

$$0 < h_{k+1} < \min\left(\frac{2m - \tilde{\kappa}^2}{m^2}, \frac{2M - \tilde{\kappa}^2}{M^2}, \frac{2}{m + M}\right)$$

such that $\rho_{k+1}(W_{2,\phi}(\mu_k, \pi) - r_0) + h_{k+1}\alpha_1 + h_{k+1}^{\frac{3}{2}}\alpha_2 < 0$. Therefore, by Theorem 2, we know

$$W_{2,\phi}(\mu_{k+1}, \pi) \leq \rho_{k+1} W_{2,\phi}(\mu_k, \pi) + h_{k+1}\alpha_1 + h_{k+1}^{\frac{3}{2}}\alpha_2 < \rho_{k+1} r_0 < r_0. \tag{46}$$

That is, $\mu_{k+1} \in \mathcal{B}_{r_0}(\pi)$. △

4. Suppose $W_{2,\phi}(\mu_k, \pi) = r_0$. For any $r > r_0$, there exists

$$0 < h_{k+1} < \min\left(\frac{2m - \tilde{\kappa}^2}{m^2}, \frac{2M - \tilde{\kappa}^2}{M^2}, \frac{2}{m + M}\right)$$

such that $r > r(h_{k+1}) = \frac{h_{k+1}\alpha_1 + h_{k+1}^{\frac{3}{2}}\alpha_2}{1-\rho_{k+1}}$. Therefore, by Theorem 2, we know

$$W_{2,\phi}(\mu_{k+1}, \pi) \leq \rho_{k+1} W_{2,\phi}(\mu_k, \pi) + h_{k+1}\alpha_1 + h_{k+1}^{\frac{3}{2}}\alpha_2 < \rho_{k+1} r_0 + (1 - \rho_{k+1}) r < r. \tag{47}$$

That is, $\mu_{k+1} \in \mathcal{B}_r(\pi)$. ■

The following lemma comes from (Horn and Johnson, 2012, Theorem 7.4.1.4).

Lemma 23 *For any symmetric matrix* $\mathbf{M}$ *with rank* p*, we have* $\mathrm{Tr}(\mathbf{M}) \le p \|\mathbf{M}\|_2$.

The remark below follows clearly from the definition of the spectral norm.

Remark 24 *If* $\mathbf{M}$ *is a symmetric matrix, then* $\|\mathbf{M}\|_2 = \lambda_{\max}(\mathbf{M})$.

Proof [Proposition 5] Firstly, we want to show

$$\mathbf{E}_{\mathbf{L}\sim\pi}\left[\|\nabla f(\mathbf{L})\|_2^2\right] = \mathbf{E}_{\mathbf{L}\sim\pi}\left[\mathrm{Tr}(D^2 f(\mathbf{L}))\right] \le p \cdot \mathbf{E}_{\mathbf{L}\sim\pi}\left[\left\|D^2 f(\mathbf{L})\right\|_2\right] \le MpR. \tag{48}$$

For the equality in (48), from integration by parts, we have

$$\begin{aligned}
&\mathbf{E}_{\mathbf{L}\sim\pi}\left[\|\nabla f(\mathbf{L})\|_2^2\right] \\
=&\int_{\mathcal{X}} \langle \nabla f(\mathbf{x}), \nabla f(\mathbf{x})\rangle \cdot \frac{d\pi}{d\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\
=&-\int_{\mathcal{X}} \left\langle \nabla f(\mathbf{x}), \nabla\left(\frac{d\pi}{d\mathbf{x}}\right)(\mathbf{x})\right\rangle d\mathbf{x} \\
=&-\int_{\partial\mathcal{X}} \frac{d\pi}{d\mathbf{x}}(\mathbf{x})\langle \nabla f(\mathbf{x}), \mathbf{n}\rangle d\mathcal{H}^{p-1}(\mathbf{x}) + \int_{\mathcal{X}} \frac{d\pi}{d\mathbf{x}}(\mathbf{x})\Delta f(\mathbf{x}) d\mathbf{x} \\
=&\int_{\partial\mathcal{X}} \left\langle \nabla\left(\frac{d\pi}{d\mathbf{x}}\right)(\mathbf{x}), \mathbf{n}\right\rangle d\mathcal{H}^{p-1}(\mathbf{x}) + \mathbf{E}_{\mathbf{L}\sim\pi}\left[\mathrm{Tr}(D^2 f(\mathbf{L}))\right] \\
=&\mathbf{E}_{\mathbf{L}\sim\pi}\left[\mathrm{Tr}(D^2 f(\mathbf{L}))\right].
\end{aligned}$$

The first inequality in (48) can be derived using Lemma 23 when $\mathbf{M} = D^2 f(\mathbf{x})$.

For the last inequality in (48), one only need to show $\left\|D^2 f(\mathbf{x})\right\|_2 \le M\left\|D^2\phi(\mathbf{x})\right\|_2$ for all $\mathbf{x} \in \mathcal{X}$. This can be derived from assumption (**A4**), as shown in Appendix B.

Secondly, since $\|\mathbf{M}\|_F \le \sqrt{p}\,\|\mathbf{M}\|_2$ holds for any matrix $\mathbf{M}$ with rank p, one has

$$2\left\|\left[D^2\phi(\mathbf{x})\right]^{\frac{1}{2}}\right\|_F^2 \le 2p\left\|\left[D^2\phi(\mathbf{x})\right]^{\frac{1}{2}}\right\|_2^2 = 2p \cdot \lambda_{\max}(D^2\phi(\mathbf{x})) = 2p\left\|D^2\phi(\mathbf{x})\right\|_2,$$

for every $\mathbf{x} \in \mathcal{X}$. Here the last equality comes from Remark 24. Thus, integrating at both sides against measure π gives

$$\mathbf{E}_{\mathbf{L}\sim\pi}\left[\left\|\sqrt{2}[D^2\phi(\mathbf{L})]^{\frac{1}{2}}\right\|_F^2\right] \le 2pR. \tag{49}$$

Lastly,

$$\begin{aligned}
&\sqrt{\mathbf{E}\left[\|\nabla\phi(\mathbf{L}_0) - \nabla\phi(\mathbf{L}_s)\|_2^2\right]} && (50)\\
=&\sqrt{\mathbf{E}\left[\left\|\int_0^s \nabla f(\mathbf{L}_r)dr - \sqrt{2}\int_0^s [D^2\phi(\mathbf{L}_r)]^{\frac{1}{2}} d\mathbf{B}_r\right\|_2^2\right]} && (51)\\
\le&\sqrt{\mathbf{E}\left[\left\|\int_0^s \nabla f(\mathbf{L}_r)dr\right\|_2^2\right]} + \sqrt{\mathbf{E}\left[\left\|\int_0^s \sqrt{2}[D^2\phi(\mathbf{L}_r)]^{\frac{1}{2}} d\mathbf{B}_r\right\|_2^2\right]} && (52)
\end{aligned}$$

$$=\sqrt{\mathbf{E}\left[\left\|\int_0^s \nabla f(\mathbf{L}_r)dr\right\|_2^2\right]}+\sqrt{\int_0^s \mathbf{E}\left[\left\|\sqrt{2}[D^2\phi(\mathbf{L}_r)]^{\frac{1}{2}}\right\|_F^2\right]dr} \tag{53}$$

$$\leq \int_0^s \sqrt{\mathbf{E}\left[\|\nabla f(\mathbf{L}_r)\|_2^2\right]}dr+\sqrt{\int_0^s \mathbf{E}\left[\left\|\sqrt{2}[D^2\phi(\mathbf{L}_r)]^{\frac{1}{2}}\right\|_F^2\right]dr} \tag{54}$$

$$=\int_0^s \sqrt{\mathbf{E}\left[\|\nabla f(\mathbf{L}_0)\|_2^2\right]}dr+\sqrt{\int_0^s \mathbf{E}\left[\left\|\sqrt{2}[D^2\phi(\mathbf{L}_0)]^{\frac{1}{2}}\right\|_F^2\right]dr} \tag{55}$$

$$=s\sqrt{\mathbf{E}\left[\|\nabla f(\mathbf{L}_0)\|_2^2\right]}+\sqrt{s\mathbf{E}\left[\left\|\sqrt{2}[D^2\phi(\mathbf{L}_0)]^{\frac{1}{2}}\right\|_F^2\right]} \tag{56}$$

$$\leq s\sqrt{MpR}+\sqrt{2spR}. \tag{57}$$

Here (52) comes from the triangle inequality; (53) is derived from Itô's isometry; (54) is obtained from Minkowski's inequality; (55) comes from the fact that $\mathbf{L}_r \sim \pi$ for all $r \geq 0$; and (57) is from (48) and (49). ■

Appendix E. Numerical Experiments

In this section, we support and illustrate our theoretical findings through a series of numerical simulations involving the Dirichlet distribution supported on the 1D and 2D standard simplex. Despite their simplicity, these numerical results clearly illustrate our analysis of the sampling error.

E.1. 1D Simplex

We consider sampling from π, where $d\pi \propto x^{a_1-1}(1-x)^{a_2-1}dx$ is the symmetric Dirichlet distribution in $\mathbb{R}^2$ with parameters $a_1 = a_2 = 3$. A natural choice of the entropy ϕ is that in the fourth row of Table 1. Overall, we are in the situation of the last column in Table 2 with parameters $(\kappa = \sqrt{2}, R = 2/3, m = 2, M = 2, \delta = 0)$. The choice of (a_1, a_2) complies with the condition $\tilde{\kappa} < \sqrt{2m}$ since $\tilde{\kappa} = \kappa = \sqrt{2}$. In turn, $r_0 = 2/\sqrt{3}$; recall the definition of r_0 from Section 3.1. Figure 1(a) shows the evolution of $W_{2,\phi}(\mu_k, \pi)$, where μ_k is the (empirical) distribution of the sample at iteration k of the HRLMC algorithm, with various constant step-sizes, starting from the Dirac measure at 10^{-4}. Figure 1(b) displays the empirical distribution of $\mathbf{X}_k$ with increasing time for a constant step-size $h = 0.04$ and three different initializations. One clearly sees that the stationary distribution is the same independently of initialization. From Figure 1(a), one observes that, with sufficiently small step-sizes, the Markov chain enters a Wasserstein ball of radius r_0 around π. However, even if running the HRLMC algorithm with vanishing step-sizes for a very long time, the error does not vanish, which supports our theoretical prediction that the bias term is inevitable.

E.2. 2D Simplex

We now consider sampling on a 2D simplex (represented as a triangle in $[0,1]^2$). Let $d\pi \propto e^{-f(x_1,x_2)}dx_1dx_2$ be a Dirichlet distribution on this simplex where $f(x_1, x_2) = -2\log(x_1) - 2\log(x_2) - 2\log(1-x_1-x_2) + C$, and C comes from the normalization constant in $d\pi$. We

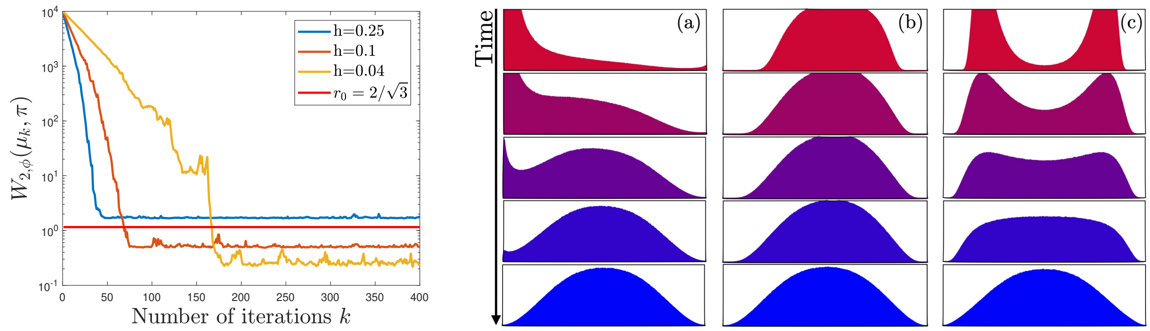

Figure 1: Results of sampling from the symmetric Dirichlet distribution in $\mathbb{R}^2$ with parameters $a_1 = a_2 = 3$ using HRLMC. Left: Evolution in time of the sampling error for various constant step-sizes. A horizontal line at $r_0 = \frac{2}{\sqrt{3}}$ materializes the size of the bias term. Right: Visual display of the evolution of the empirical distribution of $\mathbf{X}_k$ at different times, for three different initializations: (a) Dirac measure at 10^{-4}; (b) uniform measure on $[0.3, 0.8]$; (c) two Dirac measures at 0.2 and 0.8.

use $\phi(x_1, x_2) = -\log(x_1) - \log(x_2) - \log(1 - x_1 - x_2)$. Figure 2$(a)$ shows the sampling error of the HRLMC algorithm initialized with a Dirac measure at $(x_1, x_2) = (0.01, 0.99)$, and with three different constant step-sizes. We observe the same behavior as in the 1D case, where the sampling error does not vanish but rather stabilizes in a ball of radius r_0 around π. Figure 2(b) depicts the empirical distribution of $\mathbf{X}_k$ shown in contour plots with increasing time for various initializations.

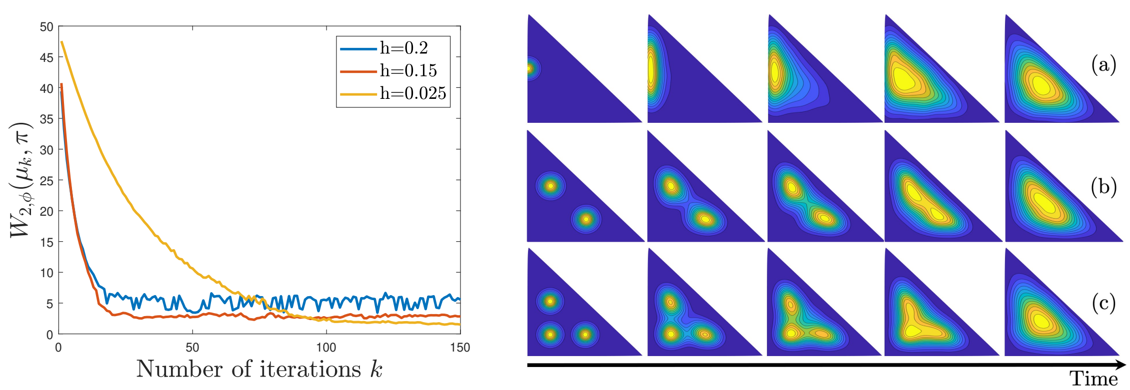

Figure 2: Results of sampling from the symmetric Dirichlet distribution on the 2D standard simplex using HRLMC. Left: evolution in time of the sampling error for various constant step-sizes. Right: visual display of the evolution of the empirical distribution of $\mathbf{X}_k$ shown as contour plots at different times, for three different initializations: (a) Dirac measure at $(0.01, 0.99)$; (b) mixture of Gaussian distributions centered at $(0.2, 0.5)$ and $(0.5, 0.2)$, respectively; (c) mixture of Gaussian distributions centered at $(0.2, 0.2)$, $(0.2, 0.5)$, and $(0.5, 0.2)$, respectively.