Mirror Langevin Monte Carlo: the Case Under Isoperimetry

Anonymous Author(s)
Affiliation
Address
email

Abstract

Motivated by the connection between sampling and optimization, we study a mirror descent analogue of Langevin dynamics and analyze three different discretization schemes, giving nonasymptotic convergence rate under functional inequalities such as Log-Sobolev in the corresponding metric. Compared to the Euclidean setting, the result reveals intricate relationship between the underlying geometry and the target distribution and show that care needs to be taken in order for the discretized algorithm to achieve vanishing bias with diminishing stepsize for sampling from non-smooth and non-convex potentials.

1 Introduction

It has been widely recognized that optimization and sampling are deeply connected. One one hand, optimization can be viewed as performing sampling in the limit, and on the other, since the influential work of Jordan-Kinderlehrer-Otto [12], Langevin dynamics takes on the interpretation as performing deterministic optimization (gradient flow) in the space of probability measures. This profoundly shape the way we view and understand traditional MCMC sampling algorithms, deviating from the Markov semigroup path. While huge amount of progress has been made on the optimization front in the past few decades, its sampling counterpart, finding far-reaching applications in Bayesian statistical inference and inverse problems, hasn't been fully explored to leverage the advancements offered by the optimization toolbox. In this paper, we draw inspiration from mirror descent [16] and ask the question if there's an analog of it that can adapt to geometries beyond the Euclidean case for Langevin diffusion, under isoperemetric inequalities such as Log-Sobolev for the target distribution, rather than Strong-Log-Concavity, where we recall the celebrated result of Bakry and Émery [3] prescribes that the latter implies the former.

1.1 Mirror Flow and Mirror Descent

In optimization, the extension to arbitrary geometry through the choice of a mirror map ϕ can often give better smoothness/strong convexity parameter dependence, or even handle cases where strong-convexity and Lipschitz gradient of f do not hold in Euclidean geometry. The continous limit of mirror descent can be written as

$$dY_t = -\nabla f(X_t)dt, \quad X_t = \nabla\phi^*(Y_t) \tag{1}$$

which is equivalent to $dY_t/dt = d[\nabla\phi(X_t)]/dt = \nabla^2\phi(X_t)dX_t/dt = -\nabla f(X_t)$, therefore the mirror flow can also be written as $dX_t = -(\nabla^2\phi(X_t))^{-1}\nabla f(X_t)dt$, akin to natural gradient flow, which precondition the update to adapt to local geometry. Equation (1) makes it clear that mirror descent update

$$x_{k+1} = \nabla\phi^*(\nabla\phi(x_k) - h_{k+1}\nabla f(x_k)) = \arg\min_x \langle x, \nabla f(x_k)\rangle + h_{k+1}^{-1}D_\phi(x, x_k)$$

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

is nothing more than (forward-discretized) gradient descent in the dual y-space through mirror mapping. Here $D_\phi(x, x_k) = \phi(x) - \phi(x_k) - \nabla\phi(x_k)^\top(x - x_k) \geq 0$ is the Bregman divergence and ϕ^* is the Fenchel conjugate of ϕ. We assume that the mirror map $\phi\colon \mathbb{R}^d \to \mathbb{R}$ is of Legendre type[1] and strictly convex throughout. Common choices include $\phi(x) = \|x\|_2^2/2$, which reduces to classical gradient descent as $x_{k+1} = x_k - h_{k+1}\nabla f(x_k)$, and $\phi(x) = -\sum_i x_i \log(x_i)$, which gives multiplicative weight update. In the special case when $\phi = f$, one readily recovers Newton's method.

1.2 Mirror Langevin Dynamics and Mirror Langevin Monte Carlo

For sampling, we consider the Mirror Langevin stochastic differential equation (SDE) where for $Y_t = \nabla\phi(X_t)$, and target distribution $\pi = e^{-f}$,

$$dY_t = -\nabla f(\nabla\phi^*(Y_t))dt + \sqrt{2[\nabla^2\phi^*(Y_t)]^{-1}}dW_t \tag{2}$$

for W_t the standard Brownian motion in $\mathbb{R}^d$. If ϕ is three-times-differentiable, it is equivalent to

$$dX_t = \left(-[\nabla^2\phi(X_t)]^{-1}\left[\mathrm{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\right) - \nabla f(X_t)\right]\right)dt + \sqrt{2[\nabla^2\phi(X_t)]^{-1}}dW_t \tag{3}$$

and the corresponding Euler-Maruyama (EM) discretized version (in dual y-space) becomes

$$x_{k+1} = \nabla\phi^*\left(\nabla\phi(x_k) - h_{k+1}\nabla f(x_k) + \sqrt{2h_{k+1}[\nabla^2\phi(x_k)]}\cdot z_{k+1}\right) \tag{4}$$

for h_{k+1} the stepsize and $z_{k+1} \in \mathbb{R}^d$ an independent standard Gaussian random vector, where we used $\nabla^2\phi^*(Y_t) = [\nabla^2\phi(X_t)]^{-1}$. Initialization can be done with a dirac mass at $\mathbb{1}_{x=x_0}$ without any particular warm start. It is worth noting that while the continuous dynamics in the primal X-space involves the 3rd-order derivative tensor, the implementation of (4) only requires access to a gradient oracle for f. For $\phi(x) = \|x\|_2^2/2$, one recovers the classical (overdamped) Langevin dynamics, whose Euler-Maruyama discretization is ULA: $x_{k+1} = x_k - h_{k+1}\nabla f(x_k) + \sqrt{2h_{k+1}}\cdot z_{k+1}$. In the case when there's no closed-form expression, the inversion $\nabla\phi^*$ can be solved numerically (and therefore approximately) using $\nabla\phi^*(z) = \arg\max_x\{z^\top x - \phi(x)\}$, which is a convex optimization problem. A derivation for the equivalence between (2) and (3), along with the fact (3) has $\pi = e^{-f}$ as stationary distribution are given in Appendix A.

It is evident (and reminiscent of the classical Langevin algorithm) that the discretized algorithm will converge to a biased limit π^h. A Metropolis-Hastings step can be applied on top to correct for the bias but we focus on the unadjusted case in this paper. The main question we aim to address in this paper is – what is the non-asymptotic rate of convergence for (2) using different discretization schemes, under functional inequalities such as Log-Sobolev Inequality (LSI), which encompasses broader classes of distributions compared to the more restrictive and well-studied Strong-Log-Concavity (SLC) setting. In particular, it is known that LSI is preserved under bounded perturbation [10] and Lipschitz mapping, which capture cases when f is far from convex, e.g., multi-modal such as Gaussian mixtures.

2 Related Work

Various discretizations of underdamped, overdamped Langevin dynamics in Euclidean setting under LSI [13, 22, 14] and SLC [8, 5, 7, 18, 15] are the main focus of a series of developments, for which non-asymptotic error bounds are established for several metrics including KL, Wasserstein and TV distance. Convergence of discretized algorithm under LSI, in general, introduces considerably more challenges, as commonly used synchronous/reflection coupling technique do not apply.

For Langevin dynamics working under non-Euclidean geometry, an earlier proposal was made in [11] where an algorithm is designed to converge to $(\nabla\phi)_\#\pi$, being essentially a change of measure from the classical Langevin dynamics. Crucially, their dynamics is different in that the diffusion term isn't scaled as the one we consider to take into account the Riemannian metric structure. The study of dynamics (2) was initiated in [23] under a relaxed-SLC assumption, where the authors show convergence of (4) to a Wasserstein ball with non-vanishing bias. [1] studied under a similar relaxed-SLC assumption, but with a different discretization scheme as opposed to (4), where the

[1] so that $\nabla\phi$ is invertible and $\nabla\phi^* = (\nabla\phi)^{-1}$ is a single-valued mapping that makes sense for (1)

bias decreases to zero with diminishing stepsize. Closer to our work is [6], in which the authors investigated convergence of the continuous process (2) under functional inequalities resembling the ones we consider and focused on $\phi = f$, leaving the analysis of discretized algorithm for future work. In this case one does not need to resort to smoothness assumption etc. for deriving contraction/handling bias issues – it is only when studying stable discretization schemes that they become necessary.

3 Notation and Assumptions

Notation We use $\nabla \cdot v(x) = \sum_i \frac{\partial v_i(x)}{\partial x_i} \in \mathbb{R}$ to denote the divergence operator for a vector field $v : \mathbb{R}^d \to \mathbb{R}^d$, and for a matrix-valued function $G : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, we let $\nabla \cdot G(x) \in \mathbb{R}^d$ with the i-th element as $\sum_j \frac{\partial G(x)_{ij}}{\partial x_j}$. We denote the Laplacian operator as Δ where $\Delta\phi(x) = \mathrm{Tr}(\nabla^2\phi(x))$. Moreover, let $\langle \nabla^2, G(x)\rangle = \nabla \cdot (\nabla \cdot G(x)) = \sum_{i,j} \frac{\partial^2 G_{ij}(x)}{\partial x_i \partial x_j} \in \mathbb{R}$. The norm induced by a positive definite G is defined as $\|z\|_G^2 := z^\top G z$. Wasserstein-p distance is defined as $W_p(\rho,\pi) = \inf_{x\sim\rho, x'\sim\pi} \mathbb{E}[\|x - x'\|^p]^{1/p}$ for $p \geq 1$. From Monge-Kantorvich's duality $W_1(\rho,\pi) = \sup_{\|f\|_{lip}\leq 1} \mathbb{E}_\rho[f] - \mathbb{E}_\pi[f]$ and monotonicity property $W_1(\rho,\pi) \leq W_2(\rho,\pi)$, bound on Wasserstein-2 metric captures many functions of potential interest. We define the pushforward measure $\bar{\rho} = \nabla\phi_\# \rho$ as $\bar{\rho}(\mathcal{B}) = \rho(\nabla\phi^{-1}(\mathcal{B}))$ for every borel set $\mathcal{B}$.

We state the assumptions that we will make below, some of which we will relax later.

Assumption 1 (ζ-Self-Concordance)**.** *There exists a constant $\zeta \geq 0$ such that the conjugate mirror map ϕ^* satisfies that $\forall y, u, s, v$,*

$$|\nabla^3\phi^*(y)[u,s,v]| \leq 2\zeta \cdot (u^\top\nabla^2\phi^*(y)u)^{1/2}(s^\top\nabla^2\phi^*(y)s)^{1/2}(v^\top\nabla^2\phi^*(y)v)^{1/2}\,.$$

Moreover, this property is preserved under Fenchel conjugation (with the same parameter), affine transformation and summation [17].

Many natural barrier and entropy functions (e.g., log-barrier) satisfy such self-concordance property. This also guarantees uniqueness of the (strong) solution of the continuous dynamics (2), cf. Appendix A of [23]. One can show that this affine-invariant condition implies a form of Hessian stability as $M^{-1}\nabla^2\phi(x) \preceq \nabla^2\phi(x') \preceq M\nabla^2\phi(x)$ for some $M \geq 1$ (cf. Lemma 4), entailing that the underlying geometry isn't rapidly changing. This form of self-concordance also appears in interior point method in optimization and previous work on Mirror Langevin [1] and suggests that locally the function behaves like a quadratic.

Assumption 2 (β-Mirror-Log-Sobolev)**.** *The target distribution π satisfies Mirror LSI with constant β w.r.t a given mirror map ϕ, i.e., for every locally lipschitz function g, it holds that π satisfies*

$$\frac{2}{\beta}\int \|\nabla g(x)\|^2_{[\nabla^2\phi(x)]^{-1}} d\pi \geq \int g(x)^2 \log g(x)^2 d\pi - \left(\int g(x)^2 d\pi\right)\log\left(\int g(x)^2 d\pi\right) \tag{5}$$

taking $g(x) = \sqrt{d\rho(x)/d\pi(x)}$ one gets that for all ρ,

$$H_\pi(\rho) := \int \rho(x)\log\frac{\rho(x)}{\pi(x)}dx \leq \frac{1}{2\beta}\int \rho(x)\left\|\nabla\log\frac{\rho(x)}{\pi(x)}\right\|^2_{[\nabla^2\phi(x)]^{-1}} dx =: \frac{1}{2\beta}J^\phi_\pi(\rho)\,. \tag{6}$$

This is the gradient-domination condition for KL-divergence in Wasserstein metric, where the (weighted) Fisher information on the RHS is the squared norm of the gradient for KL divergence.

LSI is an isoperimetric inequality that implies concentration (subgaussian tails) for the distribution and plays an important role in many results in probability theory. A discussion of its implication in our context is further expanded in Section 4.1. We note that both [1] and [23] require relative μ-strong convexity of f w.r.t ϕ, which means $\nabla^2 f \succeq \mu\nabla^2\phi \succ 0$. Therefore this assumption based on LSI is strictly weaker, allowing us to move away from convex potential.

The following two recover the familiar smoothness and Lipschitz condition when $\phi(x) = \|x\|_2^2/2$.

Assumption 3 (L-Relative Lipschitz)**.** *For all $x, x' \in dom(\phi)$, it holds that*

$$\|\nabla f(x)\|_{[\nabla^2\phi(x)]^{-1}} \leq L\,. \tag{7}$$

Assumption 4 (γ-Relative Smooth)**.** *For all $x, x' \in dom(\phi)$,*

$$\|[\nabla^2\phi(x)]^{-1}\nabla f(x) - [\nabla^2\phi(x')]^{-1}\nabla f(x')\|_{\nabla^2\phi(x')} \leq \gamma \cdot \|\nabla\phi(x) - \nabla\phi(x')\|_{[\nabla^2\phi(x')]^{-1}} . \quad (8)$$

One could show that (8) is slightly stronger than assuming $\nabla^2\tilde{f}(y) \preceq \gamma\nabla^2\phi^*(y)$ for $\tilde{f}(y) = f(\nabla\phi^*(y))$ or that $\tilde{f}(y)$ is smooth, i.e., $\nabla^2\tilde{f}(y) \preceq c \cdot I$ (cf. Lemma 7). When $\phi(x) = \|Ax\|_2^2/2$, condition (8) reduces to $\|\nabla f(x) - \nabla f(x')\|_{(A^\top A)^{-1}} \leq \gamma \cdot \|x - x'\|_{A^\top A}$, for which the dual norm on the two sides makes sense as a generalization, whereas one could check that the relative smoothness as defined in [23, 1] do not admit such a natural interpretation. We will comment more on this assumption in later sections on discretization. Condition (7) is also common in previous works [1].

Assumption 5 (α-Strongly Convex)**.** *The mirror map ϕ is three-times differentiable, and let $\alpha := \lambda_{\min}(\nabla^2\phi) > 0$.*

All best known results under LSI [13, 22] in the Euclidean setting assume third order smoothness on f (i.e., Lipschitz Hessian) and Lipschitz gradient (i.e., $-L \cdot I \preceq \nabla^2 f \preceq L \cdot I$), sometimes with an additional disspativity assumption [14], whereas we only require a weak notion of smoothness on $f \circ \nabla\phi^*$. The study of discretized sampling algorithms for *non-smooth* and *non-convex* potentials, to the best of our knowledge, is a road less traversed.

4 Convergence Analysis

4.1 Continuous Process: Mirror LSI

One can get using LSI (and LSI alone) the exponential convergence in KL divergence for the continuous process, which is a manifestation of convergence under Polyak-Łojasiewicz (PL) inequality from optimization. In optimization, steepest descent gradient flow curve $y_t^* \in \mathbb{R}^d$ is defined as

$$\min_{y_t} \langle \nabla f(y_t), \dot{y}_t \rangle + \frac{1}{2}\|\dot{y}_t\|^2$$

which implies $\dot{y}_t^* = -\nabla f(y_t^*)$ and $df(y_t^*)/dt = -\|\nabla f(y_t^*)\|^2$, from which m-gradient dominance condition on f (weaker than strong convexity)

$$f(y_t^*) - \min_y f(y) \leq \frac{1}{m}\|\nabla f(y_t^*)\|^2$$

gives exponential convergence for objective function f as $d(f(y_t^*) - \min_y f(y))/dt \leq -m(f(y_t^*) - \min_y f(y))$. In sampling within the space of probability measures, the objective function is replaced by $H_\pi(\rho_t)$, and the seminal work of [12] (see also later treatments by [2, 21]) show that the density $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ followed by the (overdamped) Langevin dynamics satisfies

$$\dot{\rho}_t = \arg\min_{v_t \text{ tangent to } \rho_t} \mathbb{E}_{x\sim\rho_t}\left[\langle \nabla_{W_2} H_\pi(\rho_t)(x), v_t(x)\rangle\right] + \frac{1}{2}\mathbb{E}_{x\sim\rho_t}[\|v_t(x)\|_2^2]$$

where $\nabla_{W_2} H_\pi(\rho_t)(x) = \nabla_x \frac{\partial H_\pi(\rho_t)}{\partial \rho_t}(x)$ is Wasserstein-2 gradient of KL divergence at ρ_t. Therefore

$$\frac{dH_\pi(\rho_t^*)}{dt} = -\mathbb{E}_{x\sim\rho_t^*}\left[\left\|\nabla_x \frac{\partial H_\pi(\rho_t^*)}{\partial \rho_t}(x)\right\|_2^2\right] = -\mathbb{E}_{x\sim\rho_t^*}\left[\left\|\nabla x\left(1 + \log\left(\frac{\rho_t^*(x)}{\pi(x)}\right)\right)\right\|_2^2\right],$$

where for the second step we used $\frac{\partial H_\pi(\rho_t)}{\partial \rho_t} = 1 + \log(\frac{\rho_t}{\pi})$ [2]. The RHS becomes the negative of the Fisher information $J_\pi^\phi(\rho)$ therefore LSI (6) reduces to gradient dominance condition for KL divergence (in this case with $\nabla^2\phi = I$ for classical Langevin dynamics), from which one can derive exponential contraction rate to target distribution π as carried out in the optimization land sans bias. This casts Langevin diffusion as precisely gradient flow w.r.t KL divergence in Wasserstein metric in the space of probability measures, from which LSI (weaker than SLC) suffices for convergence. We show in the lemma below that there is an extension of this result for Mirror-LSI that gives exponential convergence for the continuous process (3).

Proposition 1 (Convergence under Mirror-LSI)**.** *Along the dynamics of (3), we have that $H_\pi(\rho_t) \leq e^{-2\beta t} H_\pi(\rho_0)$ under Assumption 2.*

Proof. Using (22), which is the density followed by (3) and the integration by parts formula $\int \langle \nabla\phi(x), v(x)\rangle dx = -\int \phi(x)\nabla\cdot v(x)dx$,

$$\begin{aligned}\frac{d}{dt}H_\pi(\rho_t) &= \int \frac{d\rho_t}{dt}\log\frac{\rho_t}{\pi}dx + \int \pi\frac{1}{\pi}\frac{d\rho_t}{dt}dx \\ &= \int \nabla\cdot\left(\rho_t[\nabla^2\phi]^{-1}\nabla\log\frac{\rho_t}{\pi}\right)\log\frac{\rho_t}{\pi}dx + \frac{d}{dt}\int\rho_t dx \\ &= -\int \rho_t\left\|\nabla\log\frac{\rho_t}{\pi}\right\|^2_{[\nabla^2\phi]^{-1}}dx + 0 \\ &\le -2\beta\cdot H_\pi(\rho_t)\end{aligned}$$

where we used Mirror LSI in the last step. This implies from Grönwall's inequality that $H_\pi(\rho_t) \le e^{-2\beta t}H_\pi(\rho_0)$. □

For π that satisfies Mirror LSI with constant β, thanks to the strong convexity of $\nabla^2\phi \succeq \alpha I$, it also satisfies Talagrand's inequality [20] with parameter $\beta\cdot\alpha$, which means the Wasserstein-2 distance is upper bounded by KL divergence as $\frac{\alpha\beta}{2}W_2(\rho,\pi)^2 \le H_\pi(\rho)$ for any ρ. Such transportation-cost inequality has the interpretation of quadratic growth in the iterate space. Therefore Mirror LSI has the additional nice property of allowing us to translate guarantee in objective value (i.e., KL divergence) to iterate space (that involves optimal coupling between iterates $x\sim\rho, x'\sim\pi$).

Stability One can show that Mirror-LSI, similar to its Euclidean counterpart, is stable under bounded perturbation. Therefore if the potential function f of interest is not exactly relative-smooth w.r.t to a "nice" mirror map, it suffices for it to be close to one, in the sense made precise in Appendix B. This is not something one can hope for with strong-convexity assumption, where perturbation from convex function usually breaks the assumption. This could be especially useful when the potential takes a composite form of $f+g$, where one part is smooth and the other part isn't. It is also known that operations such as convolution with gaussian (or other density that satisfy LSI) preserves LSI as well [4], which offers the option of smoothing to perform approximate sampling from a nicer proxy potential $\tilde{f}$. All these reasons make target measure satisfying Mirror-LSI attractive to study compared to the two previous works [1, 23] studying mirror Langevin for $\nabla^2 f \succ 0$.

4.2 EM-Discretized Process: Interpolation with Weighted Dynamics

For analyzing the EM discretization, we build upon the idea initiated in [22] and view the discretized mirror Langevin Monte Carlo (4) as following a *weighted* Langevin dynamics. It is clear that $y_{k+1} = \nabla\phi(x_{k+1})$ is the value at time $t = h_{k+1}$ of the stochastic process

$$Y_t = Y_0 - t\cdot\nabla f(\nabla\phi^*(Y_0)) + \sqrt{2[\nabla^2\phi^*(Y_0)]^{-1}}W_t$$

starting from $Y_0 = y_k$, or written in differential equation form,

$$dY_t = -\nabla f(\nabla\phi^*(Y_0))dt + \sqrt{2[\nabla^2\phi^*(Y_0)]^{-1}}dW_t\,. \tag{9}$$

Through the mapping $X_t = \nabla\phi^*(Y_t)$, one can study the corresponding dynamics in X-space, which evolves following a *weighted* Langevin dynamics with shifted drift $\hat{\mu}$ as shown in Lemma 1 (that is responsible for convergence to a biased limit $\pi^h \ne \pi$), i.e.,

$$dX_t = (\nabla\cdot G(X_t) - G(X_t)\nabla f(X_t) + \hat{\mu})dt + \sqrt{2G(X_t)}dW_t\,. \tag{10}$$

Lemma 1 (Shifted Drift and Covariance)**.** *For the dynamics written in* (10) *following* (9)*, we have* $G(X_t) = [\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)[\nabla^2\phi(X_t)]^{-1} \succ 0$ *and*

$$\begin{aligned}\hat{\mu} = &-[\nabla^2\phi(X_t)]^{-1}\nabla f(X_0) - [\nabla^2\phi(X_t)]^{-1}\operatorname{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)[\nabla^2\phi(X_t)]^{-1}\right) \\ &- \nabla\cdot\left([\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)[\nabla^2\phi(X_t)]^{-1}\right) + [\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)[\nabla^2\phi(X_t)]^{-1}\nabla f(X_t)\,.\end{aligned}$$

Having established this, a combination of Mirror LSI and careful bounding of the discretization error using self-concordance and smoothness properties can be brought together to derive the per-iteration progress. Deatiled proofs for this section can be found in Appendix C.

Proposition 2 (Progress in One Step of EM Discretization)**.** *In one iteration of Algorithm (4) with $x_k \sim \rho_0, x_{k+1} \sim \rho_h$, under Assumption 1-5, define $D := \max_{u,v} \|\nabla\phi(u) - \nabla\phi(v)\|_2$, we have for stepsize $h \le \min(1/2\zeta L, 1/16\zeta^2 d, D/\sqrt{\alpha}L, D^2/4\alpha d, M/6\beta)$,*

$$H_\pi(\rho_h) \le e^{-\frac{3\beta}{2M}h} H_\pi(\rho_0) + 24M^2\gamma^2 dh^2 + 16M\zeta^2 d^2 \eta_h^2 h\,,$$

where we denote $M = \exp(2\zeta D/\sqrt{\alpha})$ and $\eta_h^2 = (1 - \exp(-1/16\zeta^2 h)) \cdot (1 - \zeta(hL + 2\sqrt{hd}))^{-4} + \exp(-1/16\zeta^2 h) \cdot M^2$. We use the convention $M = 1$ when $\zeta = 0$ and $D = \infty$.

Picking appropriate stepsize gives the following result – perhaps surprisingly, the simplest EM discretization exhibit an irreducible bias, and similar observation was also made in [23] where the authors show convergence to a Wasserstein ball with explicit radius, even for diminishing stepsize.

Theorem 1 (Convergence Guarantee for EM)**.** *Under Assumption 1-5, picking stepsize $h \le \min(1/2\zeta L, 1/16\zeta^2 d, D/\sqrt{\alpha}L, D^2/4\alpha d, M/6\beta, \delta\beta/44M^3\gamma^2 d)$, after $k \ge \tilde{\Omega}(M^4\gamma^2 d/\beta^2\delta)$ iterations of Algorithm (4), we have for $x_k \sim \rho_k$ that $H_\pi(\rho_k) \le \delta + R_h$, where the nonvanishing bias $R_h = \mathcal{O}(M^2\zeta^2 d^2/\beta)$. In the above, $D = \max_{u,v} \|\nabla\phi(u) - \nabla\phi(v)\|_2$ and $M = \exp(2\zeta D/\sqrt{\alpha})$.*

Proof. Iterating the inequality in Proposition 2 for k iterations,

$$\begin{aligned} H_\pi(\rho_k) &\le e^{-\frac{3\beta}{2M}hk} H_\pi(\rho_0) + \frac{24M^2\gamma^2 dh^2}{1 - e^{-\frac{3\beta}{2M}h}} + \frac{16M\zeta^2 d^2\eta_h^2 h}{1 - e^{-\frac{3\beta}{2M}h}} \\ &\le e^{-\frac{3\beta}{2M}hk} H_\pi(\rho_0) + \frac{22M^3\gamma^2 dh}{\beta} + \frac{15M^2\zeta^2 d^2\eta_h^2}{\beta} \end{aligned}$$

where we used $1 - e^{-a} \ge 3a/4$ for $a \in (0, 1/4]$. Now using Lemma 6 for initialization, picking the assumed stepsize, after $k \ge \tilde{\Omega}(M/\beta h) \ge \tilde{\Omega}(M^4\gamma^2 d/\beta^2\delta)$ iterations, we have $H_\pi(\rho_k) \le \delta + R_h$. As long as $\nabla^2\phi$ is not constant (therefore $\zeta \ne 0$, recall $M \ge 1$), $R_h := 15M^2\zeta^2 d^2\eta_h^2\beta^{-1} \ne 0$ as $h \to 0$ and the asymptotic bias R_h scale as $\mathcal{O}(M^2\zeta^2 d^2/\beta)$ since $\eta_h^2 \to 1$ as $h \to 0$. □

Remark. This convergence rate for KL divergence is stronger than [23] with their guarantee in Wasserstein distance. By Pinsker's inequality, TV distance is upper bounded by KL divergence therefore one could also get an analogous guarantee in that metric. Variations on the argument will likely generalize to other metrics such as χ^2 and Rényi divergence, with other appropriate (mirror-version of) functional inequalities such as Poincaré [6, 19].

In the case when $\phi(x) = \|Ax\|^2/2$, Theorem 1 gives no asymptotic bias, as we have $M = 1$ and $\zeta = 0$. Moreover, the assumption in this case says f is smooth in $\|\cdot\|_{A^\top A}$ although not necessarily convex, and our algorithm gives an update of the form $x_{k+1} = x_k - h_{k+1}(A^\top A)^{-1}\nabla f(x_k) + \sqrt{2h_{k+1}(A^\top A)^{-1}}z_{k+1}$, which coincides with the classical Langevin performed on function $g(\tilde{x}) = f(A^{-1}\tilde{x})$ for $x = A^{-1}\tilde{x}$. One could check that g is smooth $-\gamma \cdot I \preceq \nabla^2 g \preceq \gamma \cdot I$ and satisfies the classical LSI, which means that f is globally nice w.r.t a fixed geometry, after a change of basis. If applying the Euclidean result of [19] on $g(\cdot)$, Theorem 1 recover the same $\tilde{\mathcal{O}}(d/\delta)$ complexity, which is the best-known rate without third-order smoothness assumption on f.

A closer look at the theorem points out that M plays a prominent role in the rate, where M is effectively the Hessian stability parameter (cf. Lemma 4). However, this dependence on M necessarily means that ϕ needs to be smooth on its domain (and M will be roughly the condition number of ϕ). Although this extends beyond the Euclidean case when ϕ needs to be constant – hence allowing for slowly-changing geometry, it is still far from satisfying. An important motivation for relaxing smoothness assumption is in constrained sampling (e.g., uniform sampling from a convex body), where one typically would pick ϕ to be a self-concordant barrier function that blows up on the boundary as a proxy for the nonsmooth constraint for approximate sampling. Such functions do not have a bounded M. This, in additional to the non-vanishing bias, all suggest that EM discretization, natural as it may seem, while could improve particular parameter dependence compared to its Euclidean counterpart (as illustrated above), does not fully benefit from the use of a mirror map.

4.3 Alternative Forward Discretization Scheme

In practical applications, it is often the case that the cost of evaluating $\nabla\phi$ is considerably cheaper than the cost of computing ∇f, which could involve a finite sum over a large number of samples. Taking

hints from this observation, in [1], a slightly modified mirror Langevin algorithm was considered where at iteration k, with step size η

$$x_{k+1/2} = \arg\min_v \ \eta \nabla f(x_k)^\top v + D_\phi(v, x_k) = \nabla\phi^*(\nabla\phi(x_k) - \eta\nabla f(x_k)) \tag{11}$$

$$\text{solve } dy_t = \sqrt{2[\nabla^2\phi^*(y_t)]^{-1}}dW_t \ \text{ for } \ y_0 = \nabla\phi(x_{k+1/2}) \tag{12}$$

$$x_{k+1} = \nabla\phi^*(y_\eta) \tag{13}$$

The oracle complexity (i.e., number of queries for ∇f) of the above algorithm is the same as the one in (4), but aiming at a higher accuracy implementation for the diffusion part involving ϕ. The inner step (12) can be implemented approximately using e.g., Euler-Maruyama. Integrating both sides of (12), it is not hard to see that $y_{k+1} = \nabla\phi(x_{k+1})$ is the value at time $t = \eta$ of the continuous process

$$Y_t = Y_0 - t\nabla f(X_0) + \sqrt{2}\int_0^t [\nabla^2\phi^*(Y_s)]^{-1/2}dW_s \tag{14}$$

given $X_0 = \nabla\phi^*(Y_0) = x_k$ from the previous iteration. Written in differential form,

$$dY_t = -\nabla f(\nabla\phi^*(Y_0))dt + \sqrt{2[\nabla^2\phi^*(Y_t)]^{-1}}dW_t\,. \tag{15}$$

Compared to (9), the difference is in the second term where we traded Y_0 for Y_t, therefore this formulation amounts to discretizing the objective but not the geometry and will turn out to be crucial for removing the asymptotic bias. This is also in line with the observation from optimization [9], from which the authors argue that $\dot{x}_t = -\nabla^2\phi(x_t)^{-1}\nabla f(x_{\lfloor t\rfloor})$ gives a more "faithful" discretization compared to $\dot{x}_t = -\nabla^2\phi(x_{\lfloor t\rfloor})^{-1}\nabla f(x_{\lfloor t\rfloor})$. Indeed for the process (15), one gets in the X-space another weighted Langevin dynamics (10) with $G = [\nabla^2\phi(X_t)]^{-1}$ and $\hat{\mu} = [\nabla^2\phi(X_t)]^{-1}(\nabla f(X_t) - \nabla f(X_0))$, the discretization error of the gradient in the local ϕ metric.

It turns out that this algorithm works with a weaker notion of smoothness assumption as well (see also Lemma 7). In particular, this definition of relative smoothness only involves the local metric $\nabla^2\phi$ at a single point, whereas the previous Assumption 4 requires Lipschitz gradient across different metrics $\nabla^2\phi$, which is essentially unavoidable if one is discretizing the geometry as well.

Assumption 6 (Weaker γ-Relative Smooth)**.** *For all $x, x' \in dom(\phi)$, it holds that*

$$\|\nabla f(x) - \nabla f(x')\|_{[\nabla^2\phi(x')]^{-1}} \le \gamma \cdot \|\nabla\phi(x) - \nabla\phi(x')\|_{[\nabla^2\phi(x')]^{-1}}\,.$$

When $\phi(x) = \|Ax\|_2^2/2$, Assumption 6 reduces to $\|\nabla f(x) - \nabla f(x')\|_{(A^\top A)^{-1}} \le \gamma \cdot \|x - x'\|_{A^\top A}$; and when $\phi = f$ we always have $\gamma = 1$ (not the case for Assumption 4). We have the following result for the forward-discretized Mirror Langevin algorithm. Proofs for this section can be found in Appendix D.

Proposition 3 (Convergence Guarantee for Forward Discretization)**.** *For the Algorithm in (11)-(13), under Assumption 1-3,5,6, let $M = \exp(2\zeta D/\sqrt{\alpha})$ and $D = \max_{u,v}\|\nabla\phi(u) - \nabla\phi(v)\|_2$, picking stepsize $h \le \min(1/2\zeta L, 1/16\zeta^2 d, D/\sqrt{\alpha}L, D^2/4\alpha d, 1/6\beta, \delta\beta/100M\gamma^2 d)$, after $k \ge \tilde{\Omega}(M\gamma^2 d/\beta^2\delta)$ iterations, we have $H_\pi(\rho_k) \le \delta$.*

While it is reassuring that the algorithm has vanishing bias with diminishing stepsize for any self-concordant mirror map ϕ (at a higher cost of computation for each step), we still see the appearance of M in the rate, which as we discussed earlier, substantially limit the use case for handling weakly smooth potentials through the introduction of a mirror map. To give a concrete example, for logistic regression, one might be interested in sampling from a posterior $\pi(\theta) \propto \lambda(\theta) \cdot \exp[\sum_i y_i\theta^\top x_i - \log(1 + \exp(\theta^\top x_i))]$, where $\lambda(\theta)$ could be a constrained prior such as uniform on ℓ_∞ ball $[-1, 1]^d$. In such a setting, one natural choice is to pick a mirror map which is a self-concordant barrier for the constraint set to enforce the constraint on the drawn samples, e.g., $\phi(\theta) = \sum_i \log((1 - \theta_i)^{-1}) + \log((1 + \theta_i)^{-1})$ with $dom(\phi) = (-1, 1)^d$. One could check that with the potential of interest as $f(\theta) = \sum_i -y_i\theta^\top x_i + \log(1 + \exp(\theta^\top x_i))$, it does not have a bounded M.

4.4 Alternative Backward Discretization Scheme

Backward discretization is known to be more stable compared to forward discretization in optimization and it is also known to give the best rate under LSI for Langevin diffusion in the Euclidean setting

with weaker assumptions [22], albeit at a higher cost of solving a proximal step $x_{k+1} = x_k - \eta\nabla f(x_{k+1}) = \arg\min_x f(x) + (2\eta)^{-1}\|x - x_k\|_2^2$ compared to $x_{k+1} = x_k - \eta\nabla f(x_k)$.

It is relatively straightforward to see that a backward discretization for the dynamics using the same philosophy as Section 4.3 can be implemented with step size η as (assuming $\eta f + \phi$ is convex - guaranteed if η small enough)

$$\text{solve } dy_t = \sqrt{2[\nabla^2\phi^*(y_t)]^{-1}}dW_t \;\text{ for }\; y_0 = \nabla\phi(x_k) \tag{16}$$

$$x_{k+1} = \arg\min_v \; \eta f(v) + \phi(v) - y_\eta^\top v \Leftrightarrow \eta\nabla f(x_{k+1}) + \nabla\phi(x_{k+1}) - y_\eta = 0 \tag{17}$$

and $y_{k+1} = \nabla\phi(x_{k+1})$ is the value at time $t = \eta$ of the continuous process

$$Y_t = Y_0 - t\nabla f(\nabla\phi^*(Y_t)) + \sqrt{2}\int_0^t [\nabla^2\phi^*(Y_s)]^{-1/2}dW_s \tag{18}$$

given $x_k = \nabla\phi^*(Y_0)$ from the previous iteration. Here step (16) can again be solved iteratively using Euler-Maruyama and (17) is another convex optimization for which we can implement approximately.

We include the argument of this scheme in Appendix E, from which we see that one gets a better rate compared to the previous forward-discretized method ($\Omega(\delta^{-1/2})$ vs. $\Omega(\delta^{-1})$) while maintaining no bias. The crucial step, which was also used in the work of [22], is to relate the process (18) to an SDE for which the $\hat{\mu}$ and G only involves samples at X_t and no other time points (e.g., X_0), contrary to what happens for EM and forward discretization. This allows us to get a tighter control on the relevant quantities for bounding the discretization error and hence a better final rate (cf. Lemma 8). The difficulty for the other two schemes lies in the fact that due to the stochastic Brownian motion term, if we were to bound the discretization error between two time points X_0 and X_t, one would need a more global stability control (hence dependence on M) to account for the small probability that they are far (therefore local stability implied by self-concordance doesn't help).

Even more importantly, it works with the weakest smoothness assumption stated below.

Assumption 7 (Even Weaker γ-Relative Smooth)**.** *For all $x, x' \in dom(\phi)$, it holds that*

$$-\gamma\nabla^2\phi(x) \preceq \nabla^2 f(x) \preceq \gamma\nabla^2\phi(x)\,,$$
$$\|[\nabla^2\phi(x)]^{-1}\nabla^2 f(x) - [\nabla^2\phi(x')]^{-1}\nabla^2 f(x')\|_{op} \le K\|x - x'\|\,.$$

Examples abound for such an assumption. We give an example here where f is not smooth, yet satisfy this relative smoothness condition above. For $f(x) = x\log(x)$ which does not have Lipschitz gradient, picking the strongly convex $\phi(x) = x\log(x) + (1-x)\log(1-x)$ where $dom(\phi) = [0, 1]$, it's easy to see $f''/\phi'' = x^{-1}/(x^{-1} + (1-x)^{-1})$ satisfy both requirements.

Below we give the result for our algorithm. As alluded to earlier, this new proposal removes the M dependence altogether and uses the weakest set of assumptions. It is somewhat expected that relative-smooth type assumption would show up in the result, the lack of which necessarily implies that the potential f is "misaligned" w.r.t the underlying changing geometry, for which sampling is unequivocally expected to be hard. But what's surprising is that this is in fact all that's required from ϕ for our method, and ϕ in itself doesn't have to be smooth or even self-concrodant in this case.

Proposition 4 (Convergence Guarantee for Backward Discretization)**.** *For the Algorithm (16)-(17), under Assumption 2,3,5,7 and stepsize $h = \mathcal{O}(\min\{1/\gamma, 1/K, 1/\beta, \sqrt{\delta\beta/(\gamma^2L^2 + \alpha^{-1}d^3K^2)}\})$, after $k \ge \tilde{\Omega}(\sqrt{\gamma^2L^2 + \alpha^{-1}d^3K^2}/\delta^{1/2}\beta^{3/2})$ iterations, we have $H_\pi(\rho_k) \le \delta$.*

Let us mention in passing some possible extensions to the framework presented above. In the case when we are dealing with potential with finite sum structure, i.e., $f(x) = \sum_i f_i(x)$, as is often the case in machine learning problems, one could execute instead of (17) the update

$$x_{k+1} = \arg\min_v \; \eta\sum_{i\in B} f_i(v) + \phi(v) - y_\eta^\top v, \tag{19}$$

where B is a random batch of data points. Such algorithm effectively assumes that we have stochastic (and therefore) noisy access to ∇f, where $\hat{\nabla} f(x) = \nabla f(x) + \zeta$ for ζ an independent random noise vector with $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}[\|\zeta\|_2^2] \le d\sigma^2$. Essentially this amounts to adding another independent Brownian motion on top of the interpolated process (18) with σdB_t at each step. Basic considerations suggest that this stochastic variant of the algorithm will converge to a noise ball with radius that scales with the variance of the noise σ, but is nevertheless more efficient from a computational perspective.

5 Numerical Experiments

In this section, we test out the induced bias and the benefit of using mirror map in two experiments.

For the 1st experiment, we take the similar setup as in [6], where we pick $\phi = f$ (i.e., Newton) and consider uniform sampling from a 2D box $[-0.01, 0.01] \times [-1, 1]$. 4 methods are compared. For Newton Langevin, we aim to target $\pi_\beta \propto \exp(-\beta \cdot \phi)$, taking $\phi(x) = -\log(1-x_1^2) - \log(0.01^2 - x_2^2)$ as the barrier. We tested out the 3 different discretization schemes with $\beta = 10^{-4}$ so that $\pi_\beta \approx \pi$. Stepsize is chosen to be $h = 10^{-5}$. Projected Langevin is taken to be another option for dealing with constraints, which targets the uniform distribution π directly and simply performs ULA followed by projection. The plot below shows the samples after 500 iterations, where $\nabla\phi^*$ and proximal operator is solved with 50 steps of gradient descent steps. Diffusion of ϕ is solved with 10 inner steps of EM. From the samples, EM seems to give qualitatively different result, suggesting the existence of bias.

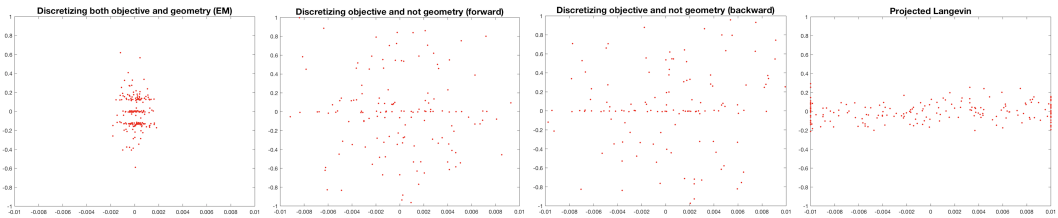

Figure 1: Uniform sampling from ill-conditioned box $[-0.01, 0.01] \times [-1, 1]$.

The 2nd experiment concerns ill-conditioned Gaussian potential (no bias in this case), for which we compare the speed of convergence for Mirror-Langevin (with EM discretization) vs. ULA. We take $\phi = f = (x-\mu)^\top \Sigma^{-1}(x-\mu)/2$, and repeat the process 200 times with $d = 50$, compute the error in empirical mean $\|\hat{\mu} - \mu\|_2$ and covariance $\|\hat{\Sigma} - \Sigma\|_F / \|\Sigma\|_F$ and plot them across iterations below. Stepsize h is picked to be 10^{-3} in both cases and initialization as $\mathcal{N}(0, I)$.

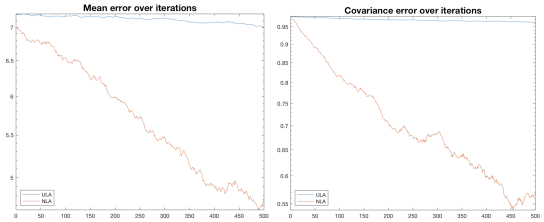

Figure 2: Newton Langevin vs. ULA for Gaussian ($\Sigma = \mathrm{diag}(1, 2, \cdots, 50), \mu = [1, \cdots, 1]$).

6 Discussion

Our result characterizes the precise interplay between ϕ and f for different discretization schemes, which can be used to guide particular choice of mirror map given sampling problems on hand. Not aware of any previous work looking into sampling from *nonsmooth* and *nonconvex* potentials, we think our newly proposed algorithm and the analysis of several previously proposed scheme in this setting highlight several interesting distinctions that do not find parallel in traditional Euclidean setup.

In optimization, one typically requires quite strong assumption for global stable convergence of Newton's method (for which $\phi = f$). Are various tricks such as Trust-Region, Cubic-Regularized Newton justified here as well for either speed up sampling or correcting for bias? As future steps, it would be good to formalize and confirm if there's a lower bound. It is likely that the extra bias term is unavoidable and capture the price we have to pay for discretizing ϕ while asking for weaker smoothness. It's also interesting to ask whether higher-order discretizers such as Runge-Kutta would yield better rate. In a different vein, one could explore the possibility of higher-order dynamics (à la underdampled Langevin) for which more sophisticated integrator [18, 15] could potentially be leveraged. Conventional wisdom suggests that the introduction of auxiliary variable that handles the non-smoothness of Brownian motion can often lead to design of better discrete sampling algorithms.

References

[1] Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-langevin algorithm, 2020.

[2] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2005.

[3] Dominique Bakry and Michel Émery. Diffusions hypercontractives. *Séminaire de probabilités de Strasbourg*, 19:177–206, 1985.

[4] Djalil Chafaï. Entropies, convexity, and functional inequalities, on ϕ-entropies and ϕ-sobolev inequalities. *Journal of Mathematics of Kyoto University - J MATH KYOTO UNIV*, 44, 01 2004.

[5] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018.

[6] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-langevin diffusions. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19573–19585. Curran Associates, Inc., 2020.

[7] Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient, 2018.

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

[9] Suriya Gunasekar, Blake Woodworth, and Nathan Srebro. Mirrorless mirror descent: A natural derivation of mirror descent. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2305–2313. PMLR, 13–15 Apr 2021.

[10] R. Holley and D. Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.

[11] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored langevin dynamics. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[12] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[13] Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942 – 1992, 2021.

[14] Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. Improved bounds for discretization of langevin diffusions: Near-optimal rates without convexity, 2019.

[15] Wenlong Mou, Yi-An Ma, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. High-order langevin diffusion yields an accelerated mcmc algorithm. *Journal of Machine Learning Research*, 22(42):1–41, 2021.

[16] A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization: Wiley, new york, 1983. *Mathematics and Computers in Simulation*, 26(1):75, 1984.

[17] Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

[18] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[19] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[20] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2016.

[21] Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 06–09 Jul 2018.

[22] Andre Wibisono. Proximal Langevin Algorithm: Rapid Convergence Under Isoperimetry. *arXiv e-prints*, page arXiv:1911.01469, November 2019.

[23] Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal M. Fadili, and Marcelo Pereyra. Wasserstein Control of Mirror Langevin Monte Carlo. In *Proceedings of Machine Learning Research, Conference on Learning Theory (COLT)*, volume 125, pages 1 – 28, Graz, Austria, July 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]