Mirror Langevin Monte Carlo: the Case Under Isoperimetry

Qijia Jiang*
UT Austin
qjiang@austin.utexas.edu

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 suggests that care might need to be taken in order for the discretized algorithm to achieve vanishing bias with diminishing stepsize for sampling from potentials under weaker smoothness/convexity regularity conditions.

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 shapes 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 [17] and ask the question if there's an analog of it that can adapt to geometries beyond the Euclidean case for Langevin diffusion, under isoperimetric 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 the Euclidean geometry. The continuous 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 be recast in the primal variable as $dX_t = -(\nabla^2\phi(X_t))^{-1}\nabla f(X_t)dt$, akin to natural gradient flow, which preconditions the update to adapt to local geometry. Equation (1) makes it clear that the mirror descent update

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

*Work done while at Stanford University.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

is nothing more than the 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[2] 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 $\nabla^2\phi^*(Y_t) = [\nabla^2\phi(X_t)]^{-1}$. No particular warm start is assumed for initialization. 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 is 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 that (3) has $\pi = e^{-f}$ as the 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 under mild regularity conditions. 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 the Euclidean setting under LSI [14, 23, 15] and SLC [8, 5, 7, 19, 16] 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 techniques do not apply.

For Langevin dynamics working under non-Euclidean geometry, an earlier proposal was made in [11] where an algorithm was 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 induced by ϕ. The study of dynamics (2) was initiated in [24] 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),

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

where the 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 (and slows down convergence from exponential to polynomial).

Recent work of [13] leverages the mean-square framework, which heavily exploits the contraction property of the dynamics, to show that under a modified self-concordance property, algorithm (4) in fact converges without asymptotic bias. This particular property is something we do not have here, therefore it's likely new ideas are needed if one were to improve the analysis of (4) in this setting.

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 [18].

Many natural barrier and entropy functions (e.g., log-barrier) satisfy such self-concordance property. This also guarantees solution of the continuous dynamics (2), cf. Appendix A of [24]. 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 the 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 for the distribution (subgaussian tails on Lipschitz functions) 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 [24] 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 allows us to move away from convex potentials.

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, it holds that $f : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable with*

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

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

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

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 [24, 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 [14, 23] 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 [15], 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 Time Process: Mirror LSI

One can get using LSI (and LSI alone) the exponential convergence in KL divergence for the continuous time 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 the 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(al) is replaced by $H_\pi(\rho_t)$, and the seminal work of [12] (see also later treatments by [2, 22]) 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 the 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 convergence rate to the 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* $H_\pi(\rho_t) \le e^{-2\beta t} H_\pi(\rho_0)$ *under Assumption 2.*

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

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

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 [21] 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 the 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 appealing to study compared to the two previous works [1, 24] 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 [23] 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*

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

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* $h \leq \min(1/2\zeta L, 1/16\zeta^2 d, D/\sqrt{\alpha}L, D^2/4\alpha d, M/6\beta)$*,*

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

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, this particular analysis suggests that the simplest EM discretization exhibits an irreducible bias, and similar observation was also made in [24] where the authors showed convergence to a Wasserstein ball with explicit radius, even for diminishing stepsize.

Theorem 1 (Convergence Guarantee for EM)**.** *Under Assumption 1-5, picking stepsize* $h \leq \min(1/2\zeta L, 1/16\zeta^2 d, D/\sqrt{\alpha}L, D^2/4\alpha d, M/6\beta, \delta\beta/44M^3\gamma^2 d)$*, after* $k \geq \tilde{\Omega}(M^4\gamma^2 d/\beta^2\delta)$ *iterations of Algorithm* (4)*, we have for* $x_k \sim \rho_k$ *that* $H_\pi(\rho_k) \leq \delta + R_h$*, 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,

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

where we used $1 - e^{-a} \geq 3a/4$ for $a \in (0, 1/4]$. Now using Lemma 6 for initialization, picking the assumed stepsize, after $k \geq \tilde{\Omega}(M/\beta h) \geq \tilde{\Omega}(M^4\gamma^2 d/\beta^2\delta)$ iterations, we have $H_\pi(\rho_k) \leq \delta + R_h$. As long as $\nabla^2\phi$ is not constant (therefore $\zeta \neq 0$, recall $M \geq 1$), $R_h := 15M^2\zeta^2 d^2\eta_h^2\beta^{-1} \neq 0$ 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 [24] with their guarantee in Wasserstein distance. Additionally, 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, 20].

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 the 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 [20] 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, although could improve particular parameter dependence compared to the Euclidean counterpart (as illustrated above), might 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 data points. 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 based on splitting the deterministic and stochastic part of the SDE works with a weaker notion of smoothness assumption as well. 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 might be 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 as motivation for 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 [23], 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 $x_{k+1} = x_k - \eta\nabla f(x_k)$ at each step.

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 [23], is to relate the process (18) to an SDE for which the $\hat{\mu}$ and G only involve 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 9). 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 apart (therefore local stability implied by self-concordance doesn't help). Our analysis for the algorithm works with the smoothness assumption stated below.

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

$$-\gamma\nabla^2\phi(x) \preceq \nabla^2 f(x) \preceq \gamma\nabla^2\phi(x)\,,$$

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

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. It is somewhat expected that relative-smoothness 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 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 $\mathbb{E}[\|\zeta\|_2^2] \leq d\sigma^2$. 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 maps in two experiments.

For the first 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 $[-0.01, 0.01] \times [-1, 1]$. Four 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 test 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 onto the domain. The plot below shows the samples after 500 iterations, where $\nabla \phi^*$ and the proximal operator are solved with 50 steps of gradient descent steps. Diffusion term ϕ is solved with 10 inner steps of EM. From the samples, EM seems to give qualitatively different result, suggesting the possible existence of bias.

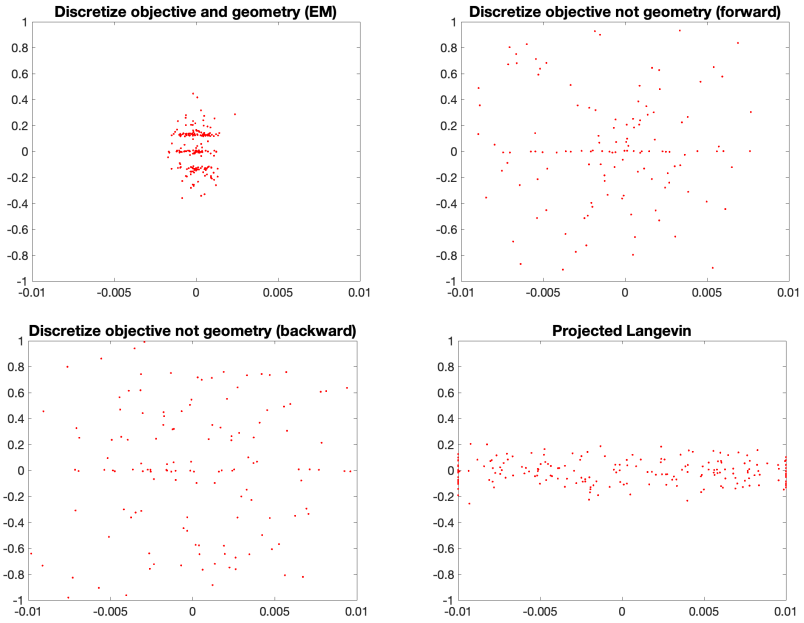

Figure 1: Uniform sampling from ill-conditioned box $[-0.01, 0.01] \times [-1, 1]$.

The second 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 $d = 50$, computing 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 $\mathcal{N}(0, I)$.

6 Discussion

Our result characterizes the interplay between ϕ and f for different discretization schemes, which can be used to guide particular choice of mirror map given the sampling problem on hand. Our newly proposed algorithm and the analysis of several previously proposed schemes in the setting of

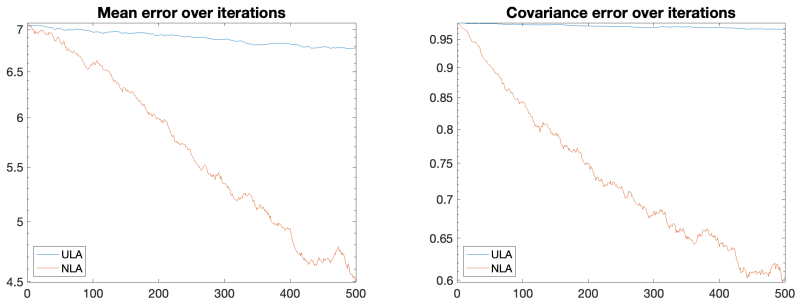

Figure 2: Newton Langevin vs. ULA for Gaussian ($\Sigma = \text{diag}(1, 2, \cdots, 50), \mu = [1, \cdots, 1]$).

sampling from nonsmooth and nonconvex potentials highlight several interesting distinctions that do not find parallel in the 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 speeding up sampling or correcting for bias? As future steps, it is an unfulfilled dream of ours to formalize and confirm a lower bound. It is conceivable that the extra bias term is unavoidable and captures 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 [19, 16] 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. On the probability theory front, it remains an intriguing open question to give a complete characterization for the Mirror-LSI condition.

Acknowledgments and Disclosure of Funding

We are grateful for the constructive comments and feedback from the anonymous reviewers. This work was partially supported under NSF-2032014.

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] Ruilin Li, Molei Tao, Santosh S. Vempala, and Andre Wibisono. The mirror langevin algorithm converges with vanishing bias, 2021.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

A Properties of Mirror Langevin SDE

A.1 Equivalence between (2) and (3)

Let $T(Y) = \nabla\phi^*(Y)$, and using that (1) $X_t = \nabla\phi^*(Y_t)$; (2) $\nabla^2\phi^*(Y_t) = [\nabla^2\phi(X_t)]^{-1}$, Itô's Lemma using equation (2) gives

$$dX_t = dT(Y_t) = -\nabla T(Y_t)^\top \nabla f(X_t)dt + \mathrm{Tr}\left(\nabla^2\phi(X_t)\nabla^2 T(Y_t)\right)dt + \sqrt{2}\nabla T(Y_t)^\top \sqrt{\nabla^2\phi(X_t)}dW_t\,.$$

Moreover, we have

$$\nabla T(Y) = [\nabla^2\phi(X)]^{-1} \tag{20}$$
$$\nabla^2 T(Y) = -[\nabla^2\phi(X)]^{-1}\frac{d\nabla^2\phi(X)}{dY}[\nabla^2\phi(X)]^{-1}$$

Therefore

$$dX_t = -[\nabla^2\phi(X_t)]^{-1}\nabla f(X_t)dt - \mathrm{Tr}\left(\frac{d\nabla^2\phi(X_t)}{dY_t}[\nabla^2\phi(X_t)]^{-1}\right)dt + \sqrt{2[\nabla^2\phi(X_t)]^{-1}}dW_t\,.$$

For the Trace operation on tensor-matrix product, we define it as $\mathrm{Tr}(\nabla^3\phi(X)G(X)) \in \mathbb{R}^d$, where the i-th element is $\mathrm{Tr}(\nabla_i\nabla^2\phi(X)G(X)) = \sum_{j,k}\frac{\partial^3\phi(X)}{\partial X_i\partial X_j\partial X_k}G(X)_{j,k}$. Looking at the i-the coordinate, the middle trace term becomes

$$\mathrm{Tr}\left([\nabla^2\phi(X)]^{-1}\sum_j \frac{\partial\nabla^2\phi(X)}{\partial X(j)}\frac{\partial X(j)}{\partial Y(i)}\right)$$
$$= \sum_j \mathrm{Tr}\left([\nabla^2\phi(X)]^{-1}\nabla_j\nabla^2\phi(X)\right)[\nabla^2\phi(X)]^{-1}_{i,j}$$

where we used that $X = T(Y)$ and equation (20). This is, of course, equal to, looking again at i-th element,

$$e_i^\top[\nabla^2\phi(X)]^{-1}\,\mathrm{Tr}\left(\nabla^3\phi(X)[\nabla^2\phi(X)]^{-1}\right)$$
$$= \sum_j [\nabla^2\phi(X)]^{-1}_{i,j}\,\mathrm{Tr}\left(\nabla_j\nabla^2\phi(X)[\nabla^2\phi(X)]^{-1}\right)\,.$$

Therefore

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

as claimed in equation (3).

A.2 Stationary distribution

Lemma 3 from [23] shows that for the SDE (under assumption $\nabla^2\phi \succ 0$)

$$dX_t = \left(\nabla\cdot[\nabla^2\phi(X_t)]^{-1} - [\nabla^2\phi(X_t)]^{-1}\nabla f(X_t)\right)dt + \sqrt{2[\nabla^2\phi(X_t)]^{-1}}dW_t\,, \tag{21}$$

the density $X_t \sim \rho_t$ satisfies the Fokker-Planck equation

$$\frac{\partial\rho_t}{\partial t} = \nabla\cdot\left(\rho_t[\nabla^2\phi]^{-1}\nabla\log\frac{\rho_t}{\pi}\right)\,, \tag{22}$$

from which it is evident that $\pi = e^{-f}$ is an invariant measure. Comparing (21) and (3), it suffices to show

$$\nabla\cdot[\nabla^2\phi(X_t)]^{-1} = -[\nabla^2\phi(X_t)]^{-1}\,\mathrm{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\right)\,. \tag{23}$$

Looking at the i-th element, we have

$$\sum_j \frac{\partial[\nabla^2\phi(X)]^{-1}_{i,j}}{\partial X(j)} = -\sum_j\sum_{s,t}[\nabla^2\phi(X)]^{-1}_{i,s}[\nabla^2\phi(X)]^{-1}_{t,j}\nabla_s[\nabla^2\phi(X)]_{t,j}$$
$$= -\sum_s [\nabla^2\phi(X)]^{-1}_{i,s}\,\mathrm{Tr}\left(\nabla_s\nabla^2\phi(X)[\nabla^2\phi(X)]^{-1}\right)\,,$$

same as RHS. This also concludes that the density of X_t from (3) follows the PDE (22).

B Properties of Mirror LSI

Lemma 2 (Stability under bounded perturbation for Mirror LSI)**.** *Suppose probability density π satisfies Mirror-LSI with parameter β, then provided $\epsilon \le \frac{d\nu}{d\pi} \le \delta$, for $\epsilon, \delta > 0$, the probability density ν satisfies Mirror-LSI with parameter $\delta/\beta\epsilon$.*

Proof. We denote the RHS of (5) as $\mathrm{Ent}_\pi[g^2]$. One could show using the variational principal of entropy that for $\nu \ll \pi$

$$\mathrm{Ent}_\nu[g^2] \le \left\| \frac{d\nu}{d\pi} \right\|_\infty \mathrm{Ent}_\pi[g^2] \le \delta \cdot \mathrm{Ent}_\pi[g^2]\,.$$

where we used the assumption $\epsilon \le \frac{d\nu}{d\pi} \le \delta$, for $\epsilon, \delta > 0$. For the LHS,

$$\begin{aligned}\frac{2}{\beta}\int \|\nabla g(x)\|^2_{[\nabla^2\phi(x)]^{-1}} d\pi &= \frac{2}{\beta}\int \|\nabla g(x)\|^2_{[\nabla^2\phi(x)]^{-1}} \frac{d\pi}{d\nu} d\nu \\ &\le \frac{2}{\beta\epsilon}\int \|\nabla g(x)\|^2_{[\nabla^2\phi(x)]^{-1}} d\nu\end{aligned}$$

Putting things together, we have

$$\frac{2\delta}{\beta\epsilon}\int \|\nabla g(x)\|^2_{[\nabla^2\phi(x)]^{-1}} d\nu \ge \mathrm{Ent}_\nu[g^2]$$

as claimed. □

There's an analogous and equivalent version of Mirror LSI (and Talagrand's inequality) in the dual y-space (as shown below) that can drive exponential convergence in the continuous dynamics (9), from which we can try to bound the discretization error. Similar analysis in the following sections, of course, can be carried out there which gives similar convergence result to $(\nabla\phi)_\#\pi$ in the dual space.

Lemma 3 (Variant of Talagrand's inequality)**.** *Mirror LSI for π as in* (6) *implies a generalized Talagrand's inequality, i.e., for all ρ,*

$$\frac{\beta}{2} W_{2,[\nabla^2\phi]^{-1}}(\nabla\phi_\#\rho, \nabla\phi_\#\pi)^2 \le H_\pi(\rho)\,,$$

where $W_{2,[\nabla^2\phi]^{-1}}(\nabla\phi_\#\rho, \nabla\phi_\#\pi)^2 := \inf_{x\sim\rho, x'\sim\pi} \mathbb{E}[\|\nabla\phi(x) - \nabla\phi(x')\|^2_{[\nabla^2\phi]^{-1}}]$.

Proof. Mirror LSI with parameter β for π in x-space implies that the density $\nabla\phi_\#\pi$ satisfies a dual Mirror-LSI with parameter β in y-space , i.e., for all ρ,

$$\int \nabla\phi_\#\rho(x) \log \frac{\nabla\phi_\#\rho(x)}{\nabla\phi_\#\pi(x)} dx \le \frac{1}{2\beta}\int \nabla\phi_\#\rho(x) \left\| \nabla \log \frac{\nabla\phi_\#\rho(x)}{\nabla\phi_\#\pi(x)} \right\|^2_{\nabla^2\phi} dx\,. \tag{24}$$

We begin by showing that the LHS is invariant under bijective mapping $\nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$. Let $\rho' = \nabla\phi_\#\rho$ and $\pi' = \nabla\phi_\#\pi$, then the change of variable formula gives

$$\frac{\rho(x)}{\pi(x)} = \frac{\rho'(\nabla\phi(x))\det(\nabla^2\phi(x))}{\pi'(\nabla\phi(x))\det(\nabla^2\phi(x))} = \frac{\rho'(\nabla\phi(x))}{\pi'(\nabla\phi(x))}\,.$$

Therefore since $\nabla\phi(x) \sim \rho'$ if $x \sim \rho$, we have

$$\mathbb{E}_{x\sim\rho'}\left[\log\frac{\rho'(x)}{\pi'(x)}\right] = \mathbb{E}_{x\sim\rho}\left[\log\frac{\rho'(\nabla\phi(x))}{\pi'(\nabla\phi(x))}\right] = \mathbb{E}_{x\sim\rho}\left[\log\frac{\rho(x)}{\pi(x)}\right]\,.$$

The RHS follows by observing that if $x \sim \rho$, $y = \nabla\phi(x) \sim \rho'$. Let $h(x) = \log\frac{\rho(x)}{\pi(x)}$ and $\tilde{h}(y) = h(\nabla\phi^*(y))$,

$$\mathbb{E}_{x\sim\rho}\left[\|\nabla_x h(x)\|^2_{[\nabla^2\phi(x)]^{-1}}\right] = \mathbb{E}_{y\sim\rho'}\left[\left\|\nabla_{\nabla\phi^*(y)}\tilde{h}(y)\right\|^2_{[\nabla^2\phi(\nabla\phi^*(y))]^{-1}}\right]$$

$$= \mathbb{E}_{y\sim\rho'}\left[\left\|\nabla^2\phi(x)\nabla_y\tilde{h}(y)\right\|^2_{[\nabla^2\phi(x)]^{-1}}\right]$$
$$= \mathbb{E}_{y\sim\rho'}\left[\left\|\nabla_y\tilde{h}(y)\right\|^2_{\nabla^2\phi(x)}\right]$$

where we used $\nabla_y\tilde{h}(y) = [\nabla_y\nabla\phi^*(y)]^\top\nabla_{\nabla\phi^*(y)}\tilde{h}(y) = [\nabla^2\phi(x)]^{-1}\nabla_{\nabla\phi^*(y)}\tilde{h}(y)$. Having established (24), applying the manifold version of LSI $\Rightarrow$ Talagrand's inequality (cf. Theorem 22.17 in [21]) on $\nabla\phi_\#\pi$, we therefore have

$$\frac{\beta}{2}W_{2,[\nabla^2\phi]^{-1}}(\nabla\phi_\#\rho, \nabla\phi_\#\pi)^2 \le H_{\nabla\phi_\#\pi}(\nabla\phi_\#\rho) = H_\pi(\rho)\,.$$

□

C Proofs for Section 4.2: EM Discretization

Proof of Lemma 1. Using (9) and let $X = T(Y) := \nabla\phi^*(Y)$, then Itô's Lemma gives

$$dX_t = dT(Y_t) = -\nabla T(Y_t)^\top\nabla f(X_0)dt + \mathrm{Tr}\left(\nabla^2\phi(X_0)\nabla^2 T(Y_t)\right)dt + \sqrt{2}\nabla T(Y_t)^\top\sqrt{\nabla^2\phi(X_0)}dW_t\,.$$

From here, a similar calculation as in Appendix A.1 shows

$$dX_t = -[\nabla^2\phi(X_t)]^{-1}\nabla f(X_0)dt - [\nabla^2\phi(X_t)]^{-1}\,\mathrm{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)[\nabla^2\phi(X_t)]^{-1}\right)dt$$
$$+ \sqrt{2}[\nabla^2\phi(X_t)]^{-1}\sqrt{\nabla^2\phi(X_0)}dW_t\,,$$

where we can easily identify G and

$$\hat{\mu} = -[\nabla^2\phi(X_t)]^{-1}\nabla f(X_0) - [\nabla^2\phi(X_t)]^{-1}\,\mathrm{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)[\nabla^2\phi(X_t)]^{-1}\right)$$
$$- \nabla\cdot G + G\nabla f(X_t)$$

as claimed. □

Below is a helper lemma stating the implication of self-concordance assumption on Hessian stability. Being an affine invariant property, it is a more natual assumption compared to those made in previous works, i.e., Lipschitz Hessian.

Lemma 4 (Self-Concordance Implication)**.** *Under Assumption 1, 3 and 5, we have for D the diameter $D := \max_{u,v}\|\nabla\phi(u) - \nabla\phi(v)\|_2$ and updates $X_t = \nabla\phi^*(Y_t)$ and $X_0 = \nabla\phi^*(Y_0)$ following* (9)*,*

$$M^{-1}\cdot[\nabla^2\phi(X_t)]^{-1} \preceq [\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)[\nabla^2\phi(X_t)]^{-1} \preceq M\cdot[\nabla^2\phi(X_t)]^{-1}$$

in expectation (w.r.t Brownian motion) for

$$M = (1-\exp(-1/16\zeta^2 t))\cdot(1-\zeta(tL+2\sqrt{td}))^{-2} + \exp(-1/16\zeta^2 t)\cdot\exp(2\zeta D/\sqrt{\alpha})\,,$$

if $t \le \min(1/2\zeta L, 1/16\zeta^2 d)$ and bounded by

$$M = \exp(2\zeta D/\sqrt{\alpha})$$

deterministically. We use the convention $M = 1$ when $\zeta = 0$ and $D = \infty$. Moreover, it implies

$$\|\,\mathrm{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\right)\|^2_{[\nabla^2\phi(X_t)]^{-1}} \le 4\zeta^2 d^2\,.$$

Proof. From the definition of self-concordance for ϕ^*, it implies from [18] that for $\|Y_t - Y_0\|_{\nabla^2\phi^*(Y_0)} \le 1/\zeta$, we have

$$\left(1-\zeta\|Y_t - Y_0\|_{\nabla^2\phi^*(Y_0)}\right)^2\nabla^2\phi^*(Y_0) \preceq \nabla^2\phi^*(Y_t) \preceq \left(1-\zeta\|Y_t - Y_0\|_{\nabla^2\phi^*(Y_0)}\right)^{-2}\nabla^2\phi^*(Y_0)\,.$$

Therefore since $\nabla\phi(X_t) - \nabla\phi(X_0) = Y_t - Y_0 = -t\cdot\nabla f(X_0) + \sqrt{2t\nabla^2\phi(X_0)}\cdot z_0$, and z_0 is independent of everything else,

$$\|Y_t - Y_0\|^2_{\nabla^2\phi^*(Y_0)} = \left\|-t\cdot\nabla f(X_0) + \sqrt{2t\nabla^2\phi(X_0)}\cdot z_0\right\|^2_{[\nabla^2\phi(X_0)]^{-1}}$$

$$= t^2\|\nabla f(X_0)\|^2_{[\nabla^2\phi(X_0)]^{-1}} + 2t\|z_0\|_2^2$$
$$\leq t^2L^2 + 2t\|z_0\|_2^2$$

Using χ^2 concentration, $\mathbb{P}(\|z_0\|_2^2 \geq (\sqrt{d}+\sqrt{\delta})^2) \leq \exp(-\delta)$, for $t \leq \min(1/2\zeta L, 1/16\zeta^2 d)$, with probability at least $1-\exp(-d) \geq 1-\exp(-1/16\zeta^2 t)$ over the draw of z_0, we have $\|Y_t - Y_0\|_{\nabla^2\phi^*(Y_0)} \leq tL + 2\sqrt{td} < 1/\zeta$, therefore

$$(1-\zeta(tL+2\sqrt{td}))^2 \cdot I \preceq \nabla^2\phi(X_0)^{1/2}[\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)^{1/2} \preceq (1-\zeta(tL+2\sqrt{td}))^{-2}\cdot I\,.$$

One the other hand, with the remaining probability $\exp(-1/16\zeta^2 t)$, consider the function $g(s) = u^\top\nabla^2\phi^*(Y_0+s(Y_t-Y_0))u =: \nabla^2\phi^*(Y_s)[u,u]$, then from self concordance we have

$$\begin{aligned}|g'(s)| = |\nabla^3\phi^*(Y_s)[u,u,Y_t-Y_0]| &\leq 2\zeta\|Y_t-Y_0\|_{\nabla^2\phi^*(Y_s)}\|u\|^2_{\nabla^2\phi^*(Y_s)}\\ &= 2\zeta\|Y_t-Y_0\|_{\nabla^2\phi^*(Y_s)}g(s)\\ &\leq \frac{2\zeta}{\sqrt{\alpha}}\|Y_t-Y_0\|_2\cdot g(s)\\ &\leq \frac{2\zeta}{\sqrt{\alpha}}D\cdot g(s)\end{aligned}$$

therefore $|\log(g(1))-\log(g(0))| \leq 2\zeta D/\sqrt{\alpha}$ implying $\exp(-2\zeta D/\sqrt{\alpha})\nabla^2\phi^*(Y_0) \preceq \nabla^2\phi^*(Y_t) \preceq \exp(2\zeta D/\sqrt{\alpha})\nabla^2\phi^*(Y_0)$ and

$$\exp(-2\zeta D/\sqrt{\alpha})\cdot I \preceq \nabla^2\phi(X_0)^{1/2}[\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)^{1/2} \preceq \exp(2\zeta D/\sqrt{\alpha})\cdot I\,.$$

Altogether this gives that in expectation w.r.t z_0, the stability parameter M is upper bounded by

$$(1-\exp(-1/16\zeta^2 t))\cdot(1-\zeta(tL+2\sqrt{td}))^{-2} + \exp(-1/16\zeta^2 t)\cdot\exp(2\zeta D/\sqrt{\alpha})$$

which goes to 1 as $t\to 0$. Self concordance also implies picking direction $[[\nabla^2\phi(X_t)]^{1/2}e_i, u, u]$

$$-2\zeta\|e_i\|_2\nabla^2\phi^*(Y_t) \preceq \nabla^3\phi^*(Y_t)[[\nabla^2\phi(X_t)]^{1/2}e_i] \preceq 2\zeta\|e_i\|_2\nabla^2\phi^*(Y_t)$$

which means using the derivation in A.1 that $\left\|\sum_j[\nabla^2\phi(X_t)]^{-1/2}_{ij}[\nabla^2\phi(X_t)]^{-1}\nabla_j\nabla^2\phi(X_t)\right\|_{op} \leq 2\zeta\ \forall i\in[d]$, therefore

$$|\sum_j[\nabla^2\phi(X_t)]^{-1/2}_{ij}\operatorname{Tr}([\nabla^2\phi(X_t)]^{-1}\nabla_j\nabla^2\phi(X_t))| \leq 2\zeta\sqrt{d}$$

and we have

$$\begin{aligned}&\|\operatorname{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\right)\|^2_{[\nabla^2\phi(X_t)]^{-1}}\\ &= \|[\nabla^2\phi(X_t)]^{-1/2}\operatorname{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\right)\|_2^2\\ &\leq d\Big|\sum_j[\nabla^2\phi(X_t)]^{-1/2}_{ij}\operatorname{Tr}\left(\nabla_j\nabla^2\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\right)\Big|^2\\ &\leq 4\zeta^2d^2\end{aligned}$$

as desired. □

We collect some useful results before giving the per-iteration progress bound.

Lemma 5 (Control on $\|\hat{\mu}\|^2_{\nabla^2\phi}$)**.** *Under Assumption 1 and 4, we have for the $\hat{\mu}$ defined in Lemma 1*

$$\|\hat{\mu}\|^2_{\nabla^2\phi} \leq 2\eta\gamma^2\|\nabla\phi(X_t)-\nabla\phi(X_0)\|^2_{[\nabla^2\phi(X_0)]^{-1}} + 8\eta^2\zeta^2d^2\,,$$

where we denote $\eta = \|[\nabla^2\phi(X_t)]^{-1}[\nabla^2\phi(X_0)]\|_{op}$.

Proof. Let $v := \nabla\cdot\left([\nabla^2\phi(X_t)]^{-1}\nabla^2\phi(X_0)[\nabla^2\phi(X_t)]^{-1}\right)$, using the fact that

$$\frac{\partial\operatorname{Tr}((X^\top CX)^{-1}A)}{\partial X} = -(CX(X^\top CX)^{-1})(A+A^\top)(X^\top CX)^{-1}\,,$$

the i-th element of v is

$$
\begin{aligned}
&-\sum_j \sum_{s,t} \Big([\nabla^2\phi(X_t)]^{-1}_{s,j} \left([\nabla^2\phi(X_t)]^{-1} \nabla^2\phi(X_0) [\nabla^2\phi(X_t)]^{-1} \right)_{i,t} \\
&\quad + [\nabla^2\phi(X_t)]^{-1}_{s,i} \left([\nabla^2\phi(X_t)]^{-1} \nabla^2\phi(X_0) [\nabla^2\phi(X_t)]^{-1} \right)_{j,t} \Big) \frac{\partial \nabla^2(X_t)_{s,t}}{\partial X_t(j)} \\
&\quad = -\sum_s \left([\nabla^2\phi(X_t)]^{-1} \right)_{i,s} \operatorname{Tr}\left(\nabla_s \nabla^2\phi(X_t) [\nabla^2\phi(X_t)]^{-1} \nabla^2\phi(X_0) \nabla^2\phi(X_t)]^{-1} \right) \\
&\quad - \sum_t \left([\nabla^2\phi(X_t)]^{-1} \nabla^2\phi(X_0) [\nabla^2\phi(X_t)]^{-1} \right)_{i,t} \operatorname{Tr}\left(\nabla_t \nabla^2\phi(X_t) [\nabla^2\phi(X_t)]^{-1} \right) .
\end{aligned}
$$

Therefore

$$
\begin{aligned}
v = &-[\nabla^2\phi(X_t)]^{-1} \operatorname{Tr}\left(\nabla^3\phi(X_t) [\nabla^2\phi(X_t)]^{-1} \nabla^2\phi(X_0) [\nabla^2\phi(X_t)]^{-1} \right) \\
&- [\nabla^2\phi(X_t)]^{-1} \nabla^2\phi(X_0) [\nabla^2\phi(X_t)]^{-1} \operatorname{Tr}\left(\nabla^3\phi(X_t) [\nabla^2\phi(X_t)]^{-1} \right) .
\end{aligned}
$$

Putting together with the expression in Lemma 1, $\|\hat{\mu}\|^2_{\nabla^2\phi}$ is

$$
\begin{aligned}
&\left\| \nabla^2\phi(X_0) [\nabla^2\phi(X_t)]^{-1} \nabla f(X_t) - \nabla f(X_0) + \nabla^2\phi(X_0) [\nabla^2\phi(X_t)]^{-1} \operatorname{Tr}\left(\nabla^3\phi(X_t) [\nabla^2\phi(X_t)]^{-1} \right) \right\|^2_{[\nabla^2\phi(X_t)]^{-1}} \\
&\le 2 \left\| [\nabla^2\phi(X_t)]^{-1/2} [\nabla^2\phi(X_0)]^{1/2} \right\|^2_{op} \left\| [\nabla^2\phi(X_0)]^{1/2} [\nabla^2\phi(X_t)]^{-1} \nabla f(X_t) - [\nabla^2\phi(X_0)]^{-1/2} \nabla f(X_0) \right\|^2_2 \\
&+ 2 \left\| [\nabla^2\phi(X_t)]^{-1/2} \nabla^2\phi(X_0) [\nabla^2\phi(X_t)]^{-1} \operatorname{Tr}\left(\nabla^3\phi(X_t) [\nabla^2\phi(X_t)]^{-1} \right) \right\|^2_2 \\
&\le 2\eta\gamma^2 \|\nabla\phi(X_t) - \nabla\phi(X_0)\|^2_{[\nabla^2\phi(X_0)]^{-1}} + 2\eta^2 \| [\nabla^2\phi(X_t)]^{-1/2} \operatorname{Tr}\left(\nabla^3\phi(X_t) [\nabla^2\phi(X_t)]^{-1} \right) \|^2_2 \\
&\le 2\eta\gamma^2 \|\nabla\phi(X_t) - \nabla\phi(X_0)\|^2_{[\nabla^2\phi(X_0)]^{-1}} + 8\eta^2\zeta^2 d^2
\end{aligned}
$$

where we used relative smoothness assumption 4 and Lemma 4. The first term will go to zero as $t \to 0$, whereas the second term will be responsible for the non-vanishing bias w.r.t the diminishing step size (as long as $\nabla^3\phi \neq 0$ so $\zeta \neq 0$). □

Now we are ready to state the main recursion, drawing doses of inspiration from [23].

Proof of Proposition 2. Denoting $G_0(x)$ and $\hat{\mu}_0(x)$ as the diffusion/drift term at time t when $x_t = x$ with x_0 at time $t = 0$, the Fokker-Planck equation for the conditional density $\rho_{t|0}(x_t|x_0)$ takes the form written below

$$
\begin{aligned}
\frac{\partial \rho_t(x)}{\partial t} &= \int \frac{\partial \rho_{t|0}(x|x_0)}{\partial t} \rho_0(x_0) dx_0 \\
&= \int \left[-\nabla \cdot (\rho_{t|0} (\nabla \cdot G_0(x) - G_0(x) \nabla f(x))) + \langle \nabla^2, \rho_{t|0} G_0(x) \rangle - \nabla \cdot (\rho_{t|0} \hat{\mu}_0(x)) \right] \rho_0(x_0) dx_0 \\
&= \nabla \cdot \left(\rho_{0|t} \int -(\rho_t (\nabla \cdot G_0(x) - G_0(x) \nabla f(x))) + \nabla \cdot (\rho_t G_0(x)) dx_0 \right) - \nabla \cdot \left(\rho_t \int \rho_{0|t} \hat{\mu}_0(x) dx_0 \right) \\
&= \nabla \cdot \left(\rho_{0|t} \int G_0(x) \nabla \rho_t + \rho_t G_0(x) \nabla f(x) dx_0 \right) - \nabla \cdot \left(\rho_t \int \rho_{0|t} \hat{\mu}_0(x) dx_0 \right) \\
&= \nabla \cdot \left(\rho_{0|t} \int \left(\rho_t G_0(x) \nabla \log \frac{\rho_t}{\pi(x)} \right) dx_0 \right) - \nabla \cdot \Big(\rho_t \underbrace{\int \rho_{0|t} \hat{\mu}_0(x) dx_0}_{\mathbb{E}_{\rho_{0|t}}[\hat{\mu}(x_0,x)|x_t=x], \text{func of } x} \Big)
\end{aligned}
$$

where for the second equality above we used Lemma 3 from [23] and (10). We will see that the first part corresponds to exponential decay to an unbiased limit (similar to what happens in Lemma 1) and the second corresponds to the biased shifted drift introduced by discretization. Let $M = \exp(2\zeta D/\sqrt{\alpha})$, since $a^\top b = 2(\sqrt{M}a)^\top (\frac{1}{2\sqrt{M}} b) \le M\|a\|_2^2 + \frac{1}{4M}\|b\|_2^2$ by Young's inequality,

$$
\frac{d}{dt} H_\pi(\rho_t) = \int \frac{d\rho_t}{dt} \log \frac{\rho_t}{\pi} dx
$$

$$
\begin{aligned}
&= \int \nabla \cdot \left(\rho_{0|t} \int \rho_t G_0 \nabla \log \frac{\rho_t}{\pi(x)} dx_0 \right) \log \frac{\rho_t}{\pi} dx - \int \nabla \cdot \left(\rho_t \int \rho_{0|t} \hat{\mu}_0 dx_0 \right) \log \frac{\rho_t}{\pi} dx \\
&= - \int \rho_{0|t} \int \rho_t \left\langle \nabla \log \frac{\rho_t}{\pi} G_0, \nabla \log \frac{\rho_t}{\pi} \right\rangle dx_0 dx + \int \rho_t \int \rho_{0|t} \langle \hat{\mu}_0, \nabla \log \frac{\rho_t}{\pi} \rangle dx_0 dx \\
&= -\mathbb{E}_{\rho_{0,t}} \left[\left\| \nabla \log \frac{\rho_t}{\pi} \right\|_G^2 \right] + \mathbb{E}_{\rho_{0,t}} \left[\langle \hat{\mu}, \nabla \log \frac{\rho_t}{\pi} \rangle \right] \\
&\leq -\frac{1}{M} \mathbb{E}_{\rho_t} \left[\left\| \nabla \log \frac{\rho_t}{\pi} \right\|_{[\nabla^2 \phi]^{-1}}^2 \right] + \mathbb{E}_{\rho_{0,t}} \left[\langle \hat{\mu}, \nabla \log \frac{\rho_t}{\pi} \rangle \right] \\
&\leq -\frac{2\beta}{M} H_\pi(\rho_t) + M \mathbb{E}_{\rho_{0,t}} [\|\hat{\mu}\|_{\nabla^2 \phi}^2] + \frac{1}{4M} \mathbb{E}_{\rho_t} \left[\left\| \nabla \log \frac{\rho_t}{\pi} \right\|_{[\nabla^2 \phi]^{-1}}^2 \right] \\
&\leq -\frac{3\beta}{2M} H_\pi(\rho_t) + M \mathbb{E}_{\rho_{0,t}} [\|\hat{\mu}\|_{\nabla^2 \phi}^2]
\end{aligned}
$$

where we used integration by parts, Mirror LSI and Lemma 4. Now using Lemma 5,

$$
\mathbb{E}_{\rho_{0,t}} [\|\hat{\mu}\|_{\nabla^2 \phi}^2] \leq 2M\gamma^2 \mathbb{E}_{\rho_{0,t}} [\|\nabla \phi(x_t) - \nabla \phi(x_0)\|_{[\nabla^2 \phi(x_0)]^{-1}}^2] + 8\zeta^2 d^2 \mathbb{E}[\|\nabla^2 \phi(x_0) [\nabla^2 \phi(x_t)]^{-1}\|_{op}^2] \,.
$$

The first term can be bounded as (since z_0 is independent from x_0)

$$
\begin{aligned}
\mathbb{E}_{\rho_{0,t}} [\|y_t - y_0\|_{\nabla^2 \phi^*(y_0)}^2] &= \mathbb{E}_{\rho_{0,t}} \left[\left\| -t \cdot \nabla f(x_0) + \sqrt{2t \nabla^2 \phi(x_0)} \cdot z_0 \right\|_{[\nabla^2 \phi(x_0)]^{-1}}^2 \right] \\
&= t^2 \mathbb{E}_{\rho_0} [\|\nabla f(x_0)\|_{[\nabla^2 \phi(x_0)]^{-1}}^2] + 2t \mathbb{E}[\|z_0\|_2^2] \\
&\leq t^2 L^2 + 2td
\end{aligned}
$$

and the second term is bounded using Lemma 4 as $8\zeta^2 d^2 \eta_t^2 := 8\zeta^2 d^2 ((1 - \exp(-1/16\zeta^2 t)) \cdot (1 - \zeta(tL + 2\sqrt{td}))^{-4} + \exp(-1/16\zeta^2 t) \cdot \exp(4\zeta D/\sqrt{\alpha}))$. Putting things together, we have for ρ_t evolving according to (10), if $0 \leq t \leq h \leq \min(1/2\zeta L, 1/16\zeta^2 d, D/\sqrt{\alpha} L, D^2/4\alpha d)$,

$$
\begin{aligned}
\frac{d}{dt} H_\pi(\rho_t) &\leq -\frac{3\beta}{2M} H_\pi(\rho_t) + 2M^2 \gamma^2 (t^2 L^2 + 2td) + 8M\zeta^2 d^2 \eta_t^2 \\
&\leq -\frac{3\beta}{2M} H_\pi(\rho_t) + 12M^2 \gamma^2 dh + 8M\zeta^2 d^2 \eta_h^2
\end{aligned}
$$

where $\eta_h^2 \to 1$ as $h \to 0$. This can be rewritten as

$$
\frac{d}{dt} \left(e^{\frac{3\beta}{2M} t} H_\pi(\rho_t) \right) \leq e^{\frac{3\beta}{2M} t} \left(12M^2 \gamma^2 dh + 8M\zeta^2 d^2 \eta_h^2 \right) \,.
$$

Integrate it for $0 \leq t \leq h$, we have for $h \leq \frac{2M}{3\beta}$,

$$
\begin{aligned}
e^{\frac{3\beta}{2M} h} H_\pi(\rho_h) - H_\pi(\rho_0) &\leq \frac{2M}{3\beta} (e^{\frac{3\beta h}{2M}} - 1)(12M^2 \gamma^2 dh + 8M\zeta^2 d^2 \eta_h^2) \\
&\leq 24M^2 \gamma^2 dh^2 + 16M\zeta^2 d^2 \eta_h^2 h \,,
\end{aligned}
$$

where we used $e^a \leq 1 + 2a$ for $a \in (0, 1]$. Therefore we end up with

$$
\begin{aligned}
H_\pi(\rho_h) &\leq e^{-\frac{3\beta}{2M} h} H_\pi(\rho_0) + e^{-\frac{3\beta}{2M} h} (24M^2 \gamma^2 dh^2 + 16M\zeta^2 d^2 \eta_h^2 h) \\
&\leq e^{-\frac{3\beta}{2M} h} H_\pi(\rho_0) + 24M^2 \gamma^2 dh^2 + 16M\zeta^2 d^2 \eta_h^2 h
\end{aligned}
$$

and identifying $x_{k+1} \sim \rho_h$ and $x_k \sim \rho_0$ finishes the proof. □

We have the following initial bound on the KL divergence with $x \sim \rho_0 = \mathcal{N}(x^*, \frac{1}{\gamma} I)$.

Lemma 6 (Initialization)**.** *Under Assumption 4 or 6 and Assumption 5, we have* $H_\pi(\rho_0) \leq f(x^*) + \frac{d}{2} \log(\frac{\gamma}{2\pi e}) + \frac{\gamma}{2\alpha} \mathbb{E}_{\rho_0} [\|\nabla \phi(x^*) - \nabla \phi(x)\|_2^2]$*, where* x^* *satisfies* $\nabla f(x^*) = 0$*. Moreover, under Assumption 7, we have* $H_\pi(\rho_0) \leq \frac{d}{2} \log(\frac{\gamma}{2\pi e}) + f(x^*) + \gamma \mathbb{E}_{\rho_0} [D_\phi(x, x^*)]$.

Proof. From relative smoothness assumption 4 or 6, let $y_t = y + t(y^* - y)$ for y, y^*, where $x^* = \nabla\phi^*(y^*)$ is the staionary point (i.e., $\nabla f(x^*) = 0$). For $\tilde{f}(y) = f(\nabla\phi^*(y))$, this also gives $\nabla\tilde{f}(y^*) = 0$ and

$$
\begin{aligned}
&|\tilde{f}(y) - [\tilde{f}(y^*) + \langle\nabla\tilde{f}(y^*), y - y^*\rangle]| \\
&= |\int_0^1 [\nabla\tilde{f}(y^*) - \nabla\tilde{f}(y_t)]^\top (y^* - y)\, dt| \\
&\le \int_0^1 \|[\nabla^2\phi(x^*)]^{-1}\nabla f(x^*) - [\nabla^2\phi(x_t)]^{-1}\nabla f(x_t)\|_{\nabla^2\phi(x_t)} \cdot \|y^* - y\|_{[\nabla^2\phi(x_t)]^{-1}} dt \\
&\le \gamma \int_0^1 \|y^* - y_t\|_{[\nabla^2\phi(x_t)]^{-1}} \cdot \|y^* - y\|_{[\nabla^2\phi(x_t)]^{-1}} dt \\
&\le \frac{\gamma}{2\alpha}\|y^* - y\|_2^2
\end{aligned}
$$

where we used Cauchy-Schwarz and ϕ being α-strongly convex implies ϕ^* is $1/\alpha$-smooth. Therefore

$$
\tilde{f}(y) \le \tilde{f}(y^*) + \langle\nabla\tilde{f}(y^*), y - y^*\rangle + \frac{\gamma}{2\alpha}\|y^* - y\|_2^2
$$

and rewriting,

$$
f(x) \le f(x^*) + \frac{\gamma}{2\alpha}\|\nabla\phi(x^*) - \nabla\phi(x)\|_2^2 . \tag{25}
$$

For $x \sim \rho_0 = \mathcal{N}(x^*, \frac{1}{\gamma}I)$ Gaussian centered at x^*, we have

$$
H_\pi(\rho_0) = -H(\rho_0) + \mathbb{E}_{\rho_0}[f] \le \frac{d}{2}\log(\frac{\gamma}{2\pi e}) + f(x^*) + \frac{\gamma}{2\alpha}\mathbb{E}_{\rho_0}[\|\nabla\phi(x^*) - \nabla\phi(x)\|_2^2] ,
$$

where we used that for normal distribution $H(\rho_0) = \frac{d}{2}\log\frac{2\pi e}{\gamma}$.

With Assumption 7, instead of (25), we have

$$
\begin{aligned}
f(x) &= f(x^*) + \int_0^1\int_0^t (x - x^*)^\top \nabla^2 f(x_s)(x - x^*) ds\, dt \\
&\le f(x^*) + \gamma\int_0^1\int_0^t (x - x^*)^\top \nabla^2\phi(x_s)(x - x^*) ds\, dt \\
&= f(x^*) + \gamma\int_0^1 (x - x^*)^\top(\nabla\phi(x_t) - \nabla\phi(x^*)) dt \\
&= f(x^*) + \gamma\left[\phi(x) - \phi(x^*) - (x - x^*)^\top\nabla\phi(x^*)\right] =: f(x^*) + \gamma D_\phi(x, x^*)
\end{aligned}
$$

which gives

$$
H_\pi(\rho_0) = -H(\rho_0) + \mathbb{E}_{\rho_0}[f] \le \frac{d}{2}\log(\frac{\gamma}{2\pi e}) + f(x^*) + \gamma\mathbb{E}_{\rho_0}[D_\phi(x, x^*)] .
$$

□

We end this section with a word about the relative smoothness assumptions.

Lemma 7 (Relative Smoothness)**.** *For the following conditions:*

1. $\nabla^2\tilde{f}(y) \preceq \gamma\nabla^2\phi^*(y)$
2. $\|[\nabla^2\phi(x)]^{-1}\nabla f(x) - [\nabla^2\phi(x')]^{-1}\nabla f(x')\|_{\nabla^2\phi(x')} \le \gamma\|\nabla\phi(x) - \nabla\phi(x')\|_{[\nabla^2\phi(x')]^{-1}}$
3. $|\tilde{f}(y) - [\tilde{f}(y') + \langle\nabla\tilde{f}(y'), y - y'\rangle]| \le \frac{\gamma}{2}\|y - y'\|^2_{\nabla^2\phi^*(y')}$
4. $\|\nabla f(x) - \nabla f(x')\|_{[\nabla^2\phi(x')]^{-1}} \le \gamma\|\nabla\phi(x) - \nabla\phi(x')\|_{[\nabla^2\phi(x')]^{-1}}$
5. $-\gamma\nabla^2\phi(x) \preceq \nabla^2 f(x) \preceq \gamma\nabla^2\phi(x)$

where $y = \nabla\phi(x)$ and $\tilde{f}(y) = f(\nabla\phi^*(y))$, we have $2 \Rightarrow 1, 3$ and $4 \Rightarrow 5$. Moreover, taking $x = x^*$ for which $\nabla f(x^*) = 0$, condition 2 becomes the same as condition 4.

Proof. We begin by showing $\nabla^2\tilde{f}(y) \succeq \gamma\nabla^2\phi^*(y)$ implies 2 does not hold. Let $g(t) = \tilde{f}(y_t)$ for $y_t = y_0 + t \cdot z$, then for any ϵ there exists some δ such that $|1/\delta \cdot (g'(\delta) - g'(0)) - g''(0)| = |1/\delta \cdot (\nabla\tilde{f}(y_\delta) - \nabla\tilde{f}(y_0))^\top z - z^\top\nabla^2\tilde{f}(y_0)z| \leq \epsilon$, therefore we have from assumption

$$(\nabla\tilde{f}(y_\delta) - \nabla\tilde{f}(y_0))^\top z \geq \delta z^\top\nabla^2\tilde{f}(y_0)z - \epsilon\delta \geq \delta\gamma\|z\|^2_{\nabla^2\phi^*(y_0)} - \epsilon\delta\,,$$

from Cauchy-Schwarz for all ϵ'

$$\|\nabla\tilde{f}(y_\delta) - \nabla\tilde{f}(y_0)\|_{[\nabla^2\phi^*(y_0)]^{-1}} \geq \delta\gamma\|z\|_{\nabla^2\phi^*(y_0)} - \epsilon' = \gamma\|y_\delta - y_0\|_{\nabla^2\phi^*(y_0)} - \epsilon'\,,$$

and chain rule $\nabla\tilde{f}(y) = [\nabla^2\phi(x)]^{-1}\nabla f(\nabla\phi^*(y))$ gives

$$\|[\nabla^2\phi(x_\delta)]^{-1}\nabla f(x_\delta) - [\nabla^2\phi(x_0)]^{-1}\nabla f(x_0)\|_{\nabla^2\phi(x_0)} \geq \gamma \cdot \|\nabla\phi(x_\delta) - \nabla\phi(x_0)\|_{[\nabla^2\phi(x_0)]^{-1}} - \epsilon'$$

for all ϵ', finishing the proof. This lets us conclude that $2 \Rightarrow 1$.

The proof of Lemma 6 shows that $2 \Rightarrow 3$.

For $4 \Rightarrow 5$, $\forall v \in \mathbb{R}^d$ we have

$$\begin{aligned}
\|\nabla^2 f(x)^\top v\|_{[\nabla^2\phi(x)]^{-1}} &= \lim_{h\to 0} \frac{\|\nabla f(x+vh) - \nabla f(x)\|_{[\nabla^2\phi(x)]^{-1}}}{h} \\
&\leq \lim_{h\to 0} \frac{\gamma \cdot \|\nabla\phi(x+vh) - \nabla\phi(x)\|_{[\nabla^2\phi(x)]^{-1}}}{h} \\
&= \gamma \cdot \|\nabla^2\phi(x)^\top v\|_{[\nabla^2\phi(x)]^{-1}}\,,
\end{aligned}$$

therefore $\nabla^2 f(x)^\top[\nabla^2\phi(x)]^{-1}\nabla^2 f(x) \preceq \gamma^2\nabla^2\phi(x)$, which in turn implies 5. □

D Proofs for Section 4.3: Alternative Forward Discretization Scheme

Proof of Proposition 3. Following the derivation for Lemma 1, we have that (15) is the same as (in the primal space)

$$dX_t = -[\nabla^2\phi(X_t)]^{-1}\nabla f(X_0)dt - [\nabla^2\phi(X_t)]^{-1}\operatorname{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\right)dt + \sqrt{2[\nabla^2\phi(X_t)]^{-1}}dW_t\,.$$

This means the process follows a weighted Langevin dynamics (10) with $G = [\nabla^2\phi(X_t)]^{-1}$ and $\hat{\mu} = [\nabla^2\phi(X_t)]^{-1}(\nabla f(X_t) - \nabla f(X_0))$ for (15) since (23) already taught us $\nabla \cdot G = -[\nabla^2\phi(X_t)]^{-1}\operatorname{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\right)$.

Now tracing the proof of Proposition 2, we have from Mirror LSI

$$\begin{aligned}
\frac{d}{dt}H_\pi(\rho_t) &= \int \frac{d\rho_t}{dt}\log\frac{\rho_t}{\pi}dx \\
&= -\mathbb{E}_{\rho_t}\left[\left\|\nabla\log\frac{\rho_t}{\pi}\right\|^2_G\right] + \mathbb{E}_{\rho_{0,t}}\left[\langle\hat{\mu}, \nabla\log\frac{\rho_t}{\pi}\rangle\right] \\
&\leq -\mathbb{E}_{\rho_t}\left[\left\|\nabla\log\frac{\rho_t}{\pi}\right\|^2_{[\nabla^2\phi]^{-1}}\right] + \mathbb{E}_{\rho_{0,t}}[\|\hat{\mu}\|^2_{\nabla^2\phi}] + \frac{1}{4}\mathbb{E}_{\rho_t}\left[\left\|\nabla\log\frac{\rho_t}{\pi}\right\|^2_{[\nabla^2\phi]^{-1}}\right] \\
&\leq -\frac{3\beta}{2}H_\pi(\rho_t) + \mathbb{E}_{\rho_{0,t}}[\|\hat{\mu}\|^2_{\nabla^2\phi}]\,.
\end{aligned}$$

Using Assumption 1, 3, 5, 6 and (14), we have for $M = \exp(2\zeta D/\sqrt{\alpha})$, $\eta_t = (1-\exp(-1/16\zeta^2 t))\cdot(1-\zeta(tL+2\sqrt{td}))^{-2} + \exp(-1/16\zeta^2 t)\cdot M$,

$$\begin{aligned}
\mathbb{E}_{\rho_{0,t}}[\|\hat{\mu}\|^2_{\nabla^2\phi}] &\leq \gamma^2\cdot\mathbb{E}_{\rho_{0,t}}\left[\|\nabla\phi(x_t) - \nabla\phi(x_0)\|^2_{[\nabla^2\phi(x_t)]^{-1}}\right] \\
&\leq \gamma^2\cdot\mathbb{E}\left[\left\|-t\nabla f(x_0) + \sqrt{2}\int_0^t[\nabla^2\phi(x_s)]^{1/2}dW_s\right\|^2_{[\nabla^2\phi(x_t)]^{-1}}\right]
\end{aligned}$$

$$\leq 2\gamma^2 t^2 \eta_t L^2 + 4t\gamma^2 Md\,,$$

where we used Itô's isometry, $(a+b)^2 \leq 2a^2 + 2b^2$ and Lemma 4.

If $0 \leq t \leq h \leq \min(1/2\zeta L, 1/16\zeta^2 d, D/\sqrt{\alpha}L, D^2/4\alpha d)$,

$$\frac{d}{dt} H_\pi(\rho_t) \leq -\frac{3\beta}{2} H_\pi(\rho_t) + 2\gamma^2 h^2 \eta_h L^2 + 4h\gamma^2 Md\,.$$

Written differently,

$$\frac{d}{dt}\left(e^{\frac{3\beta}{2}t} H_\pi(\rho_t)\right) \leq e^{\frac{3\beta}{2}t}\left(2\gamma^2 h^2 \eta_h L^2 + 4h\gamma^2 Md\right)\,.$$

Integrate it for $0 \leq t \leq h$, we have for $h \leq \frac{1}{6\beta}$,

$$\begin{aligned} e^{\frac{3\beta}{2}h} H_\pi(\rho_h) - H_\pi(\rho_0) &\leq \frac{2}{3\beta}(e^{\frac{3\beta h}{2}} - 1)(2\gamma^2 h^2 \eta_h L^2 + 4h\gamma^2 Md) \\ &\leq 24\gamma^2 h^3 \eta_h L^2 + 16 h^2 \gamma^2 Md \end{aligned}$$

where we used $e^a \leq 1 + 2a$ for $a \in (0, 1]$. Altogether this gives us

$$\begin{aligned} H_\pi(\rho_h) &\leq e^{-\frac{3\beta}{2}h} H_\pi(\rho_0) + e^{-\frac{3\beta}{2}h}(24\gamma^2 h^3 \eta_h L^2 + 16 h^2 \gamma^2 Md) \\ &\leq e^{-\frac{3\beta}{2}h} H_\pi(\rho_0) + 24\gamma^2 h^3 \eta_h L^2 + 16 h^2 \gamma^2 Md. \end{aligned}$$

Iterating the recursion,

$$\begin{aligned} H_\pi(\rho_k) &\leq e^{-\frac{3\beta}{2}hk} H_\pi(\rho_0) + \frac{24\gamma^2 h^3 \eta_h L^2}{1 - e^{-\frac{3\beta}{2}h}} + \frac{16 h^2 \gamma^2 Md}{1 - e^{-\frac{3\beta}{2}h}} \\ &\leq e^{-\frac{3\beta}{2}hk} H_\pi(\rho_0) + \frac{22\gamma^2 \eta_h L^2 h^2}{\beta} + \frac{15\gamma^2 Mhd}{\beta} \\ &\leq e^{-\frac{3\beta}{2}hk} H_\pi(\rho_0) + \frac{50 hd\gamma^2(\eta_h + M)}{\beta} \end{aligned}$$

where we used $1 - e^{-a} \geq 3a/4$ for $a \in (0, 1/4]$. Now using Lemma 6 for initialization, picking the assumed stepsize, after $k \geq \tilde{\Omega}(M\gamma^2 d/\beta^2\delta)$ iterations, we have $H_\pi(\rho_k) \leq \delta$. □

E Proof for Section 4.4: Alternative Backward Discretization Scheme

Lemma 8 (Implicit in Lemma 6 of [23])**.** *For a matrix $S = (I + t\nabla^2 f(x))^{-2}$, assuming (1) $-L \cdot I_d \preceq \nabla^2 f \preceq L \cdot I_d$; (2) $\|\nabla^2 f(x) - \nabla^2 f(y)\|_{op} \leq M\|x - y\|$ for all x, y; (3) $0 \leq t \leq \min\{1/8L, 1/M\}$, we have*

$$\|\nabla_k S\|_{op} := \left\| \frac{\partial S}{\partial x_k} \right\|_{op} \leq 4tM$$

for all $k \in [d]$. Above (2) also implies $\|\nabla_i \nabla^2 f(x)\|_{op} \leq M$ for all $i \in [d]$.

Below we give the main technical argument for this section.

Lemma 9 (SDE Derivation)**.** *If $t \leq \mathcal{O}(1/\gamma, 1/K)$, under Assumption 3,5,7, the backward discretization dynamics in* (18) *follows the SDE in* (10) *with*

$$\frac{4}{9}[\nabla^2 \phi(X_t)]^{-1} \preceq G \preceq 4[\nabla^2 \phi(X_t)]^{-1}$$

and $\|\hat{\mu}\|^2_{\nabla^2\phi} = \mathcal{O}(t^2\gamma^2 L^2 + t^2 \alpha^{-1} d^3 K^2)$. In particular, the norm decays with t therefore there is no asymptotic bias with vanishing stepsize.

Proof. For the process in (18), if we introduce $\tilde{Y}_t = Y_t + t\nabla f(\nabla \phi^*(Y_t))$, then we see that it evolves as a scaled Brownian motion:

$$d\tilde{Y}_t = \sqrt{2}[\nabla^2 \phi^*(Y_t)]^{-1/2} dW_t = \sqrt{2\nabla^2 \phi(X_t)} dW_t\,. \tag{26}$$

Therefore if we speculate that Y_t takes the form of $dY_t = \mu dt + \sqrt{2G}dW_t$, then Itô's lemma gives

$$d\tilde{Y}_t = \left(\nabla f(X_t) + \left(I + t[\nabla^2\phi(X_t)]^{-1}\nabla^2 f(X_t)\right)^\top \mu + \mathrm{Tr}(\sqrt{G}^\top T\sqrt{G})\right) dt$$
$$+ \sqrt{2}(I + t[\nabla^2\phi(X_t)]^{-1}\nabla^2 f(X_t))^\top \sqrt{G}dW_t$$

for $T = t[\nabla^2\phi(X_t)]^{-1}\frac{\partial\nabla^2 f(X_t)}{\partial Y_t} + t\frac{\partial[\nabla^2\phi(X_t)]^{-1}}{\partial Y_t}\nabla^2 f(X_t)$. Comparing this with the Brownian motion SDE for $\tilde{Y}_t$ (26), we have

$$\sqrt{G} = (I + t\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1})^{-1}[\nabla^2\phi(X_t)]^{1/2}\,,$$
$$\mu = -(I + t\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1})^{-1}[\nabla f(X_t) + \mathrm{Tr}(T\sqrt{G}\sqrt{G}^\top)]\,.$$

Now to translate to the primal X-space through mapping $\nabla\phi^*$, another application of Itô's lemma tells us

$$dX_t = d\nabla\phi^*(Y_t) = \sqrt{2}[\nabla^2\phi(X_t)]^{-1}\sqrt{G}dW_t + \left[\mathrm{Tr}\left(\sqrt{G}^\top\frac{\partial^2\nabla\phi^*(Y_t)}{\partial Y_t^2}\sqrt{G}\right) + [\nabla^2\phi(X_t)]^{-1}\mu\right]dt$$
$$= \sqrt{2}(\nabla^2\phi(X_t) + t\nabla^2 f(X_t))^{-1}[\nabla^2\phi(X_t)]^{1/2}\,dW_t + \left[\mathrm{Tr}\left(\frac{\partial^2\nabla\phi^*(Y_t)}{\partial Y_t^2}\sqrt{G}\sqrt{G}^\top\right) + [\nabla^2\phi(X_t)]^{-1}\mu\right]dt$$
$$=: \sqrt{2\tilde{G}}\,dW_t + \tilde{\mu}\,dt$$

for which $\sqrt{\tilde{G}} = (\nabla^2\phi(X_t) + t\nabla^2 f(X_t))^{-1}[\nabla^2\phi(X_t)]^{1/2} \succ 0$ by the choice of t and we can calculate the $\hat{\mu}$ in (10) as

$$\hat{\mu} = \tilde{\mu} - \nabla\cdot\tilde{G}(X_t) + \tilde{G}(X_t)\nabla f(X_t)$$
$$= -[\nabla^2\phi(X_t)]^{-1}\,\mathrm{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}G[\nabla^2\phi(X_t)]^{-1}\right) - \nabla\cdot\tilde{G} - \sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-1/2}\,\mathrm{Tr}(TG)$$
$$- t\sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-1/2}\nabla^2 f(X_t)\sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-1/2}\nabla f(X_t)$$
$$= -\nabla\cdot\tilde{G} - t\sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-1/2}\nabla^2 f(X_t)\sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-1/2}\nabla f(X_t) - [\nabla^2\phi(X_t)]^{-1}\,\mathrm{Tr}(\nabla^3\phi(X_t)\tilde{G})$$
$$- t\sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-1/2}\,\mathrm{Tr}\left(\frac{\partial\nabla^2 f(X_t)}{\partial Y_t}[\nabla^2\phi(X_t)]\tilde{G}\right)$$
$$+ t\sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-3/2}\,\mathrm{Tr}(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\nabla^2 f(X_t)[\nabla^2\phi(X_t)]\tilde{G})\,.$$

Moreover, since $0 \le t < 1/2\gamma$, under Assumption 7, we have

$$\frac{4}{9}[\nabla^2\phi(X_t)]^{-1} \preceq \tilde{G} \preceq 4[\nabla^2\phi(X_t)]^{-1}\,, \tag{27}$$

establishing the first claim. For the shifted drift $\|\hat{\mu}\|^2_{\nabla^2\phi}$, we look at it term by term: using Assumption 3 and (27),

$$\|t\sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-1/2}\nabla^2 f(X_t)\sqrt{\tilde{G}}[\nabla^2\phi(X_t)]^{-1/2}\nabla f(X_t)\|^2_{\nabla^2\phi} \le 16t^2\gamma^2\|\nabla f(X_t)\|^2_{[\nabla^2\phi(X_t)]^{-1}}$$
$$\le 16t^2\gamma^2 L^2\,.$$

For $\|-\nabla\cdot\tilde{G} - [\nabla^2\phi(X_t)]^{-1}\,\mathrm{Tr}(\nabla^3\phi(X_t)\tilde{G})\|^2_{\nabla^2\phi}$, we have using the product rule for the divergence operator, and writing $\tilde{G}$ as $[\nabla^2\phi(X_t)]^{-1}(I + t\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1})^{-2}$,

$$\Big\|[\nabla^2\phi(X_t)]^{-1}\,\mathrm{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}(I + t\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1})^{-2}\right)$$
$$- [\nabla^2\phi(X_t)]^{-1}\nabla\cdot(I + t\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1})^{-2}$$
$$- [\nabla^2\phi(X_t)]^{-1}\,\mathrm{Tr}\left(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}(I + t\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1})^{-2}\right)\Big\|^2_{\nabla^2\phi(X_t)}$$
$$= \|\nabla\cdot(I + t\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1})^{-2}\|^2_{[\nabla^2\phi(X_t)]^{-1}}$$

which can be bounded to scale with t provided $\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1}$ is Lipschitz and $-\gamma I \preceq \nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1} \preceq \gamma I$ is bounded. Using Lemma 8 with t small, for $S := \left(I + t\nabla^2 f(X_t)[\nabla^2\phi(X_t)]^{-1}\right)^{-2}$, we have $\|\nabla_k S\|_{op} \le 4tK$ for all $k \in [d]$ so with Assumption 5,

$$\|\nabla\cdot S\|^2_{[\nabla^2\phi(X_t)]^{-1}} \le \alpha^{-1}\|\nabla\cdot S\|_2^2 \le \alpha^{-1}d^2\sum_{k\in[d]}\|\nabla_k S\|^2_{op} = \mathcal{O}(\alpha^{-1}d^3t^2K^2)\,.$$

For the remaining last two terms in $\hat{\mu}$ that is up to constants equal to (by Von Neumann's trace inequality and (27) above, using a similar derivation to those in A.1)

$$\left\|t[\nabla^2\phi(X_t)]^{-1}\operatorname{Tr}(\nabla^3\phi(X_t)[\nabla^2\phi(X_t)]^{-1}\nabla^2 f(X_t)) - t[\nabla^2\phi(X_t)]^{-1}\nabla\cdot\nabla^2 f(X_t)\right\|^2_{[\nabla^2\phi(X_t)]^{-1}},$$

the Lipschitz condition required for $[\nabla^2\phi(X_t)]^{-1}\nabla^2 f(X_t)$ will give a bound on this quantity as well, as it simply being the divergence of this former expression. Hence Assumption 7, together with Lemma 8 give $\|\nabla_k([\nabla^2\phi(X_t)]^{-1}\nabla^2 f(X_t))\|_{op} \le K$ for all $k \in [d]$ and the term can be upper bounded as $t^2\alpha^{-1}\|\nabla\cdot([\nabla^2\phi(X_t)]^{-1}\nabla^2 f(X_t))\|_2^2 \le \mathcal{O}(t^2K^2d^3\alpha^{-1})$. Putting things together, we have $\|\hat{\mu}\|^2_{\nabla^2\phi} \le \mathcal{O}(t^2\gamma^2L^2 + t^2d^3K^2\alpha^{-1})$ under the assumed condition in the lemma statement. □

The important thing to note is that the diffusion and the (shifted) drift term only involves X_t and not X_0, which would introduce errors coming from the stochastic Brownian motion term and prevents a tighter control. Now we can essentially follow the template in Proposition 3 to finish the proof. The analysis is a somewhat tedious calculation.

Proof of Proposition 4. Using Mirror LSI and Lemma 9, the claim is just a stone's throw away,

$$\begin{aligned}
\frac{d}{dt}H_\pi(\rho_t) &= \int \frac{d\rho_t}{dt}\log\frac{\rho_t}{\pi}dx \\
&= -\mathbb{E}_{\rho_t}\left[\left\|\nabla\log\frac{\rho_t}{\pi}\right\|_G^2\right] + \mathbb{E}_{\rho_{0,t}}\left[\langle\hat{\mu},\nabla\log\frac{\rho_t}{\pi}\rangle\right] \\
&\le -\frac{4}{9}\mathbb{E}_{\rho_t}\left[\left\|\nabla\log\frac{\rho_t}{\pi}\right\|^2_{[\nabla^2\phi]^{-1}}\right] + \mathbb{E}_{\rho_{0,t}}[\|\hat{\mu}\|^2_{\nabla^2\phi}] + \frac{1}{4}\mathbb{E}_{\rho_t}\left[\left\|\nabla\log\frac{\rho_t}{\pi}\right\|^2_{[\nabla^2\phi]^{-1}}\right] \\
&\le -\frac{\beta}{4}H_\pi(\rho_t) + \mathbb{E}_{\rho_{0,t}}[\|\hat{\mu}\|^2_{\nabla^2\phi}] \\
&\le -\frac{\beta}{4}H_\pi(\rho_t) + C\cdot(t^2\gamma^2L^2 + t^2\alpha^{-1}d^3K^2)\,.
\end{aligned}$$

If $0 \le t \le h \le \mathcal{O}(\min(1/\gamma, 1/K))$,

$$\frac{d}{dt}H_\pi(\rho_t) \le -\frac{\beta}{4}H_\pi(\rho_t) + h^2C(\gamma^2L^2 + \alpha^{-1}d^3K^2)\,.$$

Written differently,

$$\frac{d}{dt}\left(e^{\frac{\beta}{4}t}H_\pi(\rho_t)\right) \le e^{\frac{\beta}{4}t}h^2C(\gamma^2L^2 + \alpha^{-1}d^3K^2)\,.$$

Integrate it for $0 \le t \le h$, we have for $h \le \frac{1}{\beta}$,

$$\begin{aligned}
e^{\frac{\beta}{4}h}H_\pi(\rho_h) - H_\pi(\rho_0) &\le \frac{4}{\beta}(e^{\frac{\beta h}{4}} - 1)h^2C(\gamma^2L^2 + \alpha^{-1}d^3K^2) \\
&\le 2h^3C(\gamma^2L^2 + \alpha^{-1}d^3K^2)
\end{aligned}$$

where we used $e^a \le 1 + 2a$ for $a \in (0, 1]$. Altogether this gives us

$$\begin{aligned}
H_\pi(\rho_h) &\le e^{-\frac{\beta}{4}h}H_\pi(\rho_0) + e^{-\frac{\beta}{4}h}2h^3C(\gamma^2L^2 + \alpha^{-1}d^3K^2) \\
&\le e^{-\frac{\beta}{4}h}H_\pi(\rho_0) + 2h^3C(\gamma^2L^2 + \alpha^{-1}d^3K^2).
\end{aligned}$$

Iterating the recursion,

$$\begin{aligned}
H_\pi(\rho_k) &\le e^{-\frac{\beta}{4}hk}H_\pi(\rho_0) + \frac{2h^3C(\gamma^2L^2 + \alpha^{-1}d^3K^2)}{1 - e^{-\frac{\beta}{4}h}} \\
&\le e^{-\frac{\beta}{4}hk}H_\pi(\rho_0) + \frac{12h^2C(\gamma^2L^2 + \alpha^{-1}d^3K^2)}{\beta}
\end{aligned}$$

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 $k \ge \tilde{\Omega}(\sqrt{\gamma^2L^2 + \alpha^{-1}d^3K^2}/\delta^{1/2}\beta^{3/2})$ iterations, we have $H_\pi(\rho_k) \le \delta$. □