Efficient constrained sampling via the mirror-Langevin algorithm

Kwangjun Ahn
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA 02139
kjahn@mit.edu

Sinho Chewi
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139
schewi@mit.edu

Abstract

We propose a new discretization of the mirror-Langevin diffusion and give a crisp proof of its convergence. Our analysis uses relative convexity/smoothness and self-concordance, ideas which originated in convex optimization, together with a new result in optimal transport that generalizes the displacement convexity of the entropy. Unlike prior works, our result both (1) requires much weaker assumptions on the mirror map and the target distribution, and (2) has vanishing bias as the step size tends to zero. In particular, for the task of sampling from a log-concave distribution supported on a compact set, our theoretical results are significantly better than the existing guarantees.

1 Introduction

We consider the following canonical sampling problem. Let $V : \mathbb{R}^d \to \mathbb{R} \cup \{ \infty \}$ be a convex function and let $\pi$ be the density on $\mathbb{R}^d$ which is proportional to $\exp(-V)$. The task is to output a sample which is (approximately) distributed according to $\pi$, given query access to the gradients of $V$.

The sampling problem has attracted considerable attention recently within the machine learning and statistics communities. This renewed interest in sampling is spurred, on one hand, by a wide breadth of applications ranging from Bayesian inference [RC04, DM+19] and its use in inverse problems [DS17], to neural networks [GPAM+14, JR20]. On the other hand, there is a deep and fruitful connection between sampling and the field of optimization, introduced in the seminal work [JKO98], which has resulted in the rapid development of sampling algorithms inspired by optimization methods such as: proximal/splitting methods [Ber18, Wib18, Wib19, SKL20], coordinate descent [DLLW21a, DLLW21b], mirror descent [HKRC18, CLGL+20, ZPP20], Nesterov’s accelerated gradient descent [CCBJ18, MCC+21, DRD20], and Newton methods [MWBG12, SBCR16, CLGL+20, WL20].

To describe this connection, we recall the Langevin diffusion, which is the solution to the following stochastic differential equation (SDE):

$$dX_t = -\nabla V(X_t) \, dt + \sqrt{2} \, dB_t.$$  \hspace{1cm} (LD)

Under standard assumptions on the potential $V$, the SDE is well-posed and it converges in distribution, as $t \to \infty$, to its unique stationary distribution $\pi$. Thus, once suitably discretized, it yields a popular algorithm for the sampling problem. The Langevin diffusion is classically studied using techniques from Markov semigroup theory [see, e.g., BGL14, Pav14], but there is a more insightful perspective which views the diffusion (LD) through the lens of optimization [JKO98]. Specifically, if $\mu_t$ denotes the law of the process (LD) at time $t$, then the curve $(\mu_t)_{t \geq 0}$ is the gradient flow of the KL divergence $\mathcal{D}_{KL}(\cdot \| \pi)$ in the Wasserstein space of probability measures. This perspective has not only inspired

new analyses of Langevin [CB18 Wib18 DMM19 NW19], but has also emboldened the possibility of bringing to bear the extensive toolkit of optimization onto the problem of sampling; see the references listed above.

However, the vanilla Langevin diffusion notably fails when the support of the target distribution $\pi$ is not all of $\mathbb{R}^d$. This task of constrained sampling, named in analogy to constrained optimization, arises in applications such as latent Dirichlet allocation [BNJ03], ordinal data models [JA06], survival time analysis [KM06], regularized regression [CEAM+12], and Bayesian matrix factorization [PBJ14]. Despite such a broad range of applications, the constrained sampling problem has proven to be challenging. In particular, most prior works have focused on domain-specific algorithms [GSL92 PP14 LS16], and the first general-purpose algorithms for this task are recent [BDMP17 BEL18].

In this work, we tackle the constrained sampling problem via the mirror-Langevin algorithm (MLA). MLA is a discretization of the mirror-Langevin diffusion [HKRC18 ZPFP20], which is the sampling analogue of mirror descent. Namely, if $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a mirror map, then the mirror-Langevin diffusion is the solution to the SDE

$$X_t = \nabla \phi^* (Y_t), \quad dY_t = -\nabla V (X_t) \, dt + \sqrt{2} \left[ \nabla^2 \phi (X_t) \right]^{1/2} dB_t.$$  

(MLD)

**Technical motivation.** Recently, Zhang et al. [ZPFP20] analyze an Euler-Maruyama discretization of [MLD] see (2.2) for details. The most curious aspect of their result is that their convergence guarantee has a bias term that does not vanish even when step size tends to zero and the number of iterations tends to infinity. Moreover, they conjecture that this bias term is unavoidable. This is in contrast to known results for standard Langevin, which raises the main question of this paper:

*Can a different discretization of [MLD] lead to a vanishing bias?*

**Our contributions.** We propose a new discretization of the mirror-Langevin diffusion, given in (MLA) and illustrated in Figure 1. Our proposed discretization has the same cost as the standard Euler-Maruyama discretization of [MLD] in terms of the number of queries to the gradient oracle for $V$. We remark that our scheme for the case $\phi = \|\cdot\|^2/2$ recovers the unadjusted Langevin algorithm. The most important aspect of our result is that the bias of our algorithm vanishes as the step size tends to zero unlike the result by Zhang et al. [ZPFP20].

By adapting the analysis of Durmus et al. [DMM19], we provide a clean convergence analysis of our algorithm which theoretically validates our discretization scheme. Notably, our analysis only requires standard assumptions/definitions which are well-studied in optimization. In particular, we establish a stronger link between sampling and optimization without relying on technical assumptions of Zhang et al. [ZPFP20] (e.g. commutation conditions for Hessians; see (A5) therein).

Moreover, our analysis combines ideas from optimization with the calculus of optimal transport. In particular, we establish a new generalization of a celebrated fact, namely that the entropy functional is
displacement convex along Wasserstein geodesics, to the setting of Bregman divergences (Theorem 4). This inequality has interesting consequences in its own right; as we discuss in Corollary 1 our result already implies the transport inequality of Cordero-Erausquin [CE17].

We provide convergence guarantees for the following classes of potentials: (1) convex and relatively smooth (Theorem 1); (2) strongly relatively convex and relatively smooth (Theorem 2); and (3) convex and Lipschitz (Theorem 3). Our results largely match state-of-the-art results for the discretization of the Langevin algorithm for unconstrained sampling. Our work paves the way for the practical deployment of mirror-Langevin methods for sampling applications, paralleling the successes of mirror descent in optimization [NY83, JN11, Bub15].

In Section 5 we demonstrate the strength of our convergence guarantees compared with the previous works [BDMP17, BEL18] in an application to Bayesian logistic regression; further applications are given in Appendix E. We also corroborate our theoretical findings with numerical experiments.

Other related works. Recently, a few works have proposed modifications of the Langevin algorithm for the task of constrained sampling. Bubeck et al. [BEL13] studied the projected Langevin algorithm (PLA), which simply projects each step of the Langevin algorithm onto dom(V). A different approach was taken in Brosse et al. [BDMP17], which applies the Langevin algorithm to a smooth approximation of V given by the Moreau-Yosida envelope. The latter approach was later interpreted and further analyzed by Salim and Richtarik [SR20] using the primal-dual optimality framework from convex optimization.

A different line of work, more closely related to ours, uses a mirror map to change the geometry of the sampling problem [HKRC18, CLGL+20, ZPEP20]. In particular, the mirror-Langevin diffusion (MLD) was first introduced in an earlier draft of [HKRC18], as well as in [ZPEP20]. The diffusion was further studied in [CLGL+20], which provided a simple convergence analysis in continuous time using the sampling analog of Polyak-Łojasiewicz inequalities [KNS16]. We also remark that the idea of changing the geometry via a mirror map also played a crucial role for the problem of sampling from the uniform distribution over a polytope [KN12, LV17, CDWY18, LV18, GN20, LLV20].

Lastly, our work follows the trend of applying ideas from optimization to the task of sampling. Specifically, our analysis adopts the framework of relative convexity and smoothness, which was advocated as a more flexible framework for optimization in [BBT17, LFN18].

2 The mirror-Langevin algorithm

2.1 Background

In this section, we list basic definitions and assumptions that we employ in this work.

Convex functions of Legendre type. Throughout, we assume familiarity with the basic notions of convex analysis [see e.g. [Roc70, BL06].

Definition 1 (Convex functions of Legendre type [Roc70 §26]). A proper convex lower semicontinuous function \( \phi : \mathbb{R}^d \to \mathbb{R} \cup \{ \infty \} \) is of Legendre type if

(i) \( \mathcal{Q} := \text{int} (\text{dom}(\phi)) \neq \emptyset \),

(ii) \( \phi \) is strictly convex and differentiable on \( \mathcal{Q} \), and

(iii) \( \lim_{k \to \infty} \| \nabla \phi(x_k) \| = \infty \) whenever \( \{ x_k \}_{k \in \mathbb{N}} \) is a sequence in \( \mathcal{Q} \) converging to \( \partial \mathcal{Q} \).

The key properties of convex functions of Legendre type are listed below:

- The subdifferential \( \partial \phi \) is single-valued and hence \( \partial \phi = \{ \nabla \phi \} \) [Roc70 Theorem 26.1].
- \( \phi \) is a convex function of Legendre type if and only if its Fenchel conjugate \( \phi^* \) is a convex function of Legendre type [Roc70 Theorem 26.5].
- The gradient \( \nabla \phi \) forms a bijection between \( \text{int} (\text{dom}(\phi)) \) and \( \text{int} (\text{dom}(\phi^*)) \) with \( \nabla \phi^* = (\nabla \phi)^{-1} \) [Roc70 Theorem 26.5].

We refer readers to [Roc70 §26] for more details. We henceforth assume that our mirror map \( \phi \) is a convex function of Legendre type.
The natural notion of “distance” associated with the mirror map \( \phi \) is given by the Bregman divergence [see, e.g. \[Bre67\] §4]:

**Definition 2 (Bregman divergence \([Bre67]\)).** For a convex function \( \phi \) of Legendre type, the Bregman divergence \( D_\phi(x,y) \) associated to \( \phi \) is defined as

\[
D_\phi(x,y) := \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle, \quad \forall x, y \in Q := \text{int}(\text{dom}(\phi)).
\]

The Bregman divergence behaves like a squared distance; indeed, as \( x \to y \) a Taylor expansion shows that \( D_\phi(x,y) \sim \frac{1}{2} \langle x - y, \nabla^2 \phi(y)(x - y) \rangle \). We refer to \[Bre67\] §4 for other basic properties of the Bregman divergence. Hereinafter, when we use a Bregman divergence, we implicitly assume that its associated function \( \phi \) is a mirror map.

An important case to keep in mind is the mirror map \( \phi = \frac{1}{2} \| \cdot \|^2 \), where \( \| \cdot \| \) denotes the Euclidean norm, in which case the Bregman divergence simply becomes \( D_\phi(x,y) = \frac{1}{2} \| x - y \|^2 \).

**Self-concordance.** We recall the definition of self-concordance, which has been extensively used in applications such as interior-point methods \[NN94\]. Given a \( C^2 \) strictly convex function \( \phi \), the local norm at \( x \in \text{int}(\text{dom}(\phi)) \) with respect to \( \phi \) is defined as

\[
\| u \|_{\nabla^2 \phi(x)} := \sqrt{\langle \nabla^2 \phi(x) u, u \rangle} \quad \text{for all } u \in \mathbb{R}^d.
\]

The dual local norm at \( x \in \text{int}(\text{dom}(\phi)) \) with respect to \( \phi \) is

\[
\| u \|_{\nabla^2 \phi(x)^{-1}} = \sqrt{\langle \langle \nabla^2 \phi(x)^{-1} u, u \rangle \rangle} \quad \text{for all } u \in \mathbb{R}^d.
\]

**Definition 3 (Self-concordant function \([Nes18]\) §5.1.3)).** We say that a \( C^3 \) convex function \( \phi \) is self-concordant with a constant \( M_\phi \geq 0 \) if for any \( x \in \text{int}(\text{dom}(\phi)) \)

\[
|\nabla^3 \phi(x)|_{u,u,u} \leq 2M_\phi \| u \|_{\nabla^2 \phi(x)}^3 \quad \text{for all } u \in \mathbb{R}^d.
\]

**Relative convexity/smoothness.** We recall the following definitions:

**Definition 4 (Relative convexity \([BBT17, LFN18]\)).** \( V \) is \( \alpha \)-convex relative to \( \phi \) if

\[
V(y) \geq V(x) + \langle \nabla V(x), y - x \rangle + \alpha D_\phi(y,x) \quad \forall x, y \in Q.
\]

**Definition 5 (Relative smoothness \([BBT17, LFN18]\)).** \( V \) is \( \beta \)-smooth relative to \( \phi \) if

\[
V(y) \leq V(x) + \langle \nabla V(x), y - x \rangle + \beta D_\phi(y,x) \quad \forall x, y \in Q.
\]

We give some basic facts about these definitions in Appendix B.

In the rest of this work, we assume that \( V \in C^2(\mathcal{X}) \) where \( \mathcal{X} := \text{int}(\text{dom}(V)) \), that \( \mathcal{X} \subseteq \overline{Q} \), and \( \mathcal{X} \cap Q \neq \emptyset \). Also, we assume that \( \exp(-V) \) is integrable so that \( \pi \) is well-defined; this holds if and only if \( V(x) \geq a \| x \| - b \) for some \( a, b > 0 \) \[BGVV14\] Lemma 2.2.1.

**Optimal transport.** Given a lower semicontinuous cost function \( c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty) \), we can define the **optimal transport** cost between two probability measures \( \mu \) and \( \nu \) on \( \mathbb{R}^d \) to be

\[
\inf \{ \mathbb{E} c(X, Y) \mid X \sim \mu, \ Y \sim \nu \}.
\]

(2.1)

Here, the infimum is taken over pairs of random variables \( (X, Y) \) defined on the same probability space, with marginal laws \( \mu \) and \( \nu \) respectively. It is known that the infimum in (2.1) is always attained; we refer to the standard introductory texts \[Vil03, Vil09, San15\] for this and other basic facts in optimal transport.

In this work, we are most concerned with the case when the cost function \( c \) is the Bregman divergence associated with a mirror map:

**Definition 6 (Bregman transport cost).** The **Bregman transport cost** is defined as

\[
D_\phi(\mu, \nu) := \inf \{ \mathbb{E} D_\phi(X, Y) \mid X \sim \mu, \ Y \sim \nu \}.
\]

The Bregman transport cost was also studied in \[CE17\].

In particular, when \( \phi = \frac{1}{2} \| \cdot \|^2 \), we obtain an important special case:
Definition 7 (2-Wasserstein distance). The 2-Wasserstein distance $W_2$ is defined as
\[
W_2^2(\mu, \nu) := \inf \{ \mathbb{E}[\|X - Y\|^2] \mid X \sim \mu, \ Y \sim \nu \}.
\]

The $W_2$ optimal transport cost indeed defines a metric over the space of probability measures on $\mathbb{R}^d$ with finite second moment [Vil03, Theorem 7.3]; we refer to this metric space as the Wasserstein space. The $W_2$ metric is particularly important because it arises from a formal Riemannian structure on the Wasserstein space. This perspective was introduced in [Ott01] and applied to the Langevin diffusion in [JKO98, OV00]; in particular, these latter two works justify the perspective of the Langevin diffusion as a gradient flow of the Kullback-Leibler divergence in Wasserstein space. A rigorous exposition to Wasserstein calculus can be found in [AGS08, Vil03].

Here, we give a brief and informal introduction to the calculation rules of optimal transport. For any regular curve of measures $(\mu_t)_{t \geq 0}$, there is a corresponding family of tangent vectors $(v_t)_{t \geq 0}$ [see AGS08, Theorem 8.3.1]; here, $v_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a vector field on $\mathbb{R}^d$. Also, if $\mathcal{F}$ is any well-behaved functional defined over Wasserstein space, then at each regular measure $\mu$ one can define the Wasserstein gradient of $\mathcal{F}$ at $\mu$, which we denote $\nabla_{W_2} \mathcal{F}(\mu)$; it is also a mapping $\mathbb{R}^d \rightarrow \mathbb{R}^d$. Then, we have the calculation rule
\[
\partial_t \mathcal{F}(\mu_t) = \mathbb{E}\langle \nabla_{W_2} \mathcal{F}(\mu_t)(X_t), v_t(X_t) \rangle
\]
for any regular curve of measures $(\mu_t)_{t \geq 0}$ with corresponding tangent vectors $(v_t)_{t \geq 0}$, where $X_t \sim \mu_t$.

We will use this calculation rule in Appendix C.

2.2 Discretization of the mirror-Langevin diffusion

In order to turn a continuous-time diffusion such as [MLD] into an implementable algorithm, it is necessary to first discretize the stochastic process. The discretization considered in the prior works [ZPPF20, CLGL+20] is a simple Euler-Maruyama discretization: fixing $\eta > 0$, we define a sequence of iterates $(X_k)_{k \in \mathbb{N}}$ via
\[
\nabla \phi(X_{k+1}) = \nabla \phi(X_k) - \eta \nabla V(X_k) + \sqrt{2\eta} \, \nabla^2 \phi(X_k)^{1/2} \xi_k , \tag{2.2}
\]
where $(\xi_k)_{k \in \mathbb{N}}$ is a sequence of i.i.d. standard Gaussians in $\mathbb{R}^d$.

However, many other discretizations are possible. Indeed, in many machine learning applications, the most costly step is the evaluation of $\nabla V$, which may require a sum over a large training set, whereas the mirror map $\phi$ may be chosen to have a simple form. For the purpose of obtaining a more efficient sampling algorithm, it may therefore be a favorable trade-off to use a high-precision implementation of the diffusion step at the cost of additional computation time (which nonetheless does not require additional query access to the gradients of $V$). Motivated by these considerations, we propose a new discretization (see Figure 1 for an illustration):

The mirror-Langevin algorithm (MLA):

\[
X_{k+1} := \arg \min_{x \in \Omega} \left[ \langle \eta \nabla V(X_k), x \rangle + D_\phi(x, X_k) \right], \tag{MLA:1}
\]

\[
X_{k+1} := \nabla^\star(\eta V(\eta)), \quad \text{where} \quad \begin{cases}
  dW_t = \sqrt{2} \, \nabla^2 \phi^\star(W_t)^{-1/2} \, dB_t , \\
  W_0 = \nabla \phi(X_{k+1/2}) .
\end{cases} \tag{MLA:2}
\]

In [MLA:2], the stochastic processes $(W_t)_{t \geq 0}$ are assumed to be driven by independent Brownian motions at each iteration. When $\phi = \|\cdot\|^2/2$, MLA recovers the unadjusted Langevin algorithm.

Practicality. Although [MLA:2] is defined using the exact solution of an SDE (and we analyze the exact step [MLA:2] for simplicity), it should be understood as capturing the idea of discretizing the diffusion step more finely (e.g., through multiple inner iterations of an Euler-Maruyama discretization) than the gradient step. This is indeed amenable to practical implementation since, as previously discussed, the gradient step is typically much more costly than the diffusion step. Moreover, this is justified by our theoretical results in the next section, which together with the conjecture of [ZPPF20] suggest that fine discretization of the diffusion step is potentially crucial for attaining vanishing bias. Nevertheless, it is indeed the case that a single iteration of [MLA] is more costly than a single step of the Euler-Maruyama discretization of [MLD] and this represents a limitation of our work.
Remark 1. Our proposed discretization can be understood as a more faithful discretization of the mirror-Langevin diffusion (MLD), à la [GWS21]. It can also be understood as the forward-flow discretization of MLD in the interpretation of [Wib18].

Remark 2. In order for MLD to be well-defined, we require assumptions on $\phi$ such that the diffusion $(W_t)_{t \geq 0}$ is non-explosive, i.e., it does not exit $\text{int}(\text{dom}(\phi^*))$ in finite time. This holds under very mild assumptions on $\phi$; see [GK96]. For situations of interest, the assumptions of [GK96] can be checked directly.

3 Convergence rates of mirror-Langevin algorithm

First, we state the main assumptions which are used for our main results.

Assumption 1 (Self-concordance of $\phi$). We assume that $\phi$ is $M_{\phi}$-self-concordant (Definition 3).

Assumption 2 (Relative Lipschitzness). We assume that $V$ is $L$-relatively Lipschitz with respect to $\phi$, in the sense that $\|\nabla V(x)\|_{\nabla^2 \phi(x)^{-1}} \leq L$ for all $x \in \mathcal{X}$ (see Section 2.7 for the definition of the local norm used here).

Assumption 3 (Relative convexity and smoothness). We assume that $V$ is $\alpha$-convex relative to $\phi$ and $\beta$-smooth relative to $\phi$, where $0 \leq \alpha \leq \beta \leq \infty$ (Definitions 4 and 5).

Remark 3. Our analysis works under weaker assumptions than those of [ZPFP20]. In particular, our analysis does not assume the moment condition on the Hessian ((A2) therein) and the bound on the commutator between $\nabla^2 \phi$ and $\nabla^2 V$ ((A5) therein). Moreover, our analysis uses weaker (and more standard) definitions of self-concordance.

Throughout this section, we assume the conditions listed above, and we present convergence results for MLD under various sets of assumptions. Our first two results pertain to the smooth case, i.e. $\beta < \infty$. Define the parameter

$$\beta' := \beta + 2M_{\phi}L.$$ 

One might wonder how large $\beta'$ is for typical applications. First, we only need $\phi$ to be self-concordant, not a self-concordant barrier. Hence the appearance of $M_{\phi}$ is typically not problematic; for instance, a log barrier with $m$ constraints is $O(1)$-self-concordant. The smoothness parameter could be large and dimension-dependent in general. However, such situations are actually where our approach could be potentially advantageous. In Appendix E.2 we demonstrate an example where the smoothness parameter becomes much smaller by choosing $\phi$ carefully.

Theorem 1 (Weakly convex case). Suppose that Assumptions 1, 2, 3 hold with $\alpha = 0$ and $\beta' > 0$. For a target accuracy $\epsilon > 0$, let $X_k \sim \mu_k$ denote the iterates of MLD with step size $\eta = \min\{\frac{\alpha\epsilon}{2\beta' d}, \frac{1}{\beta' d}\}$. Then, the following convergence rate holds for the distribution $\tilde{\mu}_N := \frac{1}{N} \sum_{k=1}^N \mu_k$:

$$\mathcal{D}_{KL}(\tilde{\mu}_N \| \pi) \leq \epsilon, \quad \text{provided that } N \geq \frac{4\beta'd}{\epsilon^2} \frac{\mathcal{D}_\phi(\pi, \mu_0)}{\epsilon^2} \max\{1, \epsilon/2d\}. \quad (3.1)$$

Proof. See Appendix C.2.

Theorem 2 (Strongly relatively convex case). Suppose that Assumptions 1, 2, 3 hold with $\alpha, \beta' > 0$.

1. (Convergence in Bregman transport cost) For a target accuracy $\epsilon > 0$, let $X_k \sim \mu_k$ denote the iterates of MLD with step size $\eta = \min\{\frac{\alpha\epsilon}{2\beta' d}, \frac{1}{\beta' d}\}$. Then,

$$\mathcal{D}_\phi(\pi, \mu_N) \leq \epsilon, \quad \text{provided that } N \geq \frac{2\beta'd}{\alpha^2 \epsilon} \ln\left(\frac{2\mathcal{D}_\phi(\pi, \mu_0)}{\epsilon}\right) \max\{1, \epsilon/2d\}. \quad (3.1)$$

2. (Convergence in KL divergence) For a target accuracy $\epsilon > 0$, suppose that $X_0 \sim \mu_0$ satisfies $\mathcal{D}_\phi(\pi, \mu_0) \leq \epsilon/\alpha$. Let $X_k \sim \mu_k$ denote the iterates of MLD with step size $\eta = \min\{\frac{\alpha\epsilon}{2\beta' d}, \frac{1}{\beta' d}\}$. Then, the following convergence rate holds for the distribution $\tilde{\mu}_N := \frac{1}{N} \sum_{k=1}^N \mu_k$:

$$\mathcal{D}_{KL}(\tilde{\mu}_N \| \pi) \leq \epsilon, \quad \text{provided that } N \geq \frac{4\beta'd}{\alpha \epsilon} \max\{1, \epsilon/2d\}. \quad (3.1)$$
Proof. See Appendix C.3.

Note that the initialization assumption $\mathcal{D}_\phi(\pi, \mu_0) \leq \epsilon/\alpha$ in the second assertion of Theorem 2 can be obtained from the first guarantee of Theorem 2. Chaining together the two parts of the theorem, we therefore obtain the following guarantee: suppose that we initialize MLA at a distribution $\mu_0$. Then, with step size $\eta = \min\{\epsilon/\alpha^2, 1\}$, we obtain

$$
\mathcal{D}_{KL}\left(\frac{1}{N} \sum_{k=N_0+1}^{N_0+N_1} \mu_k \parallel \pi\right) \leq \epsilon,
$$

provided that

$$
\begin{aligned}
N_0 &\geq \frac{2\beta d}{\alpha^2} \ln\left(\frac{2N\mathcal{D}_\phi(\pi, \mu_0)}{\alpha}\right) \max\{1, \frac{\epsilon}{2\alpha}\}, \\
N_1 &\geq \frac{4\beta d}{\alpha^2} \max\{1, \frac{\epsilon}{2\alpha}\}.
\end{aligned}
$$

Observe also that for the case $\phi = \frac{||\cdot||^2}{2}$, Theorems 1 and 2 recover the corresponding convergence guarantees for the unadjusted Langevin algorithm [Corollary 7, and Corollaries 10 and 11 respectively in DMM19].

Next, we present our guarantee for the non-smooth case $\beta = \infty$. For this result, we assume that $\phi$ is strongly convex w.r.t. a norm $||\cdot||$ on $\mathbb{R}^d$, and that $V$ is $L$-Lipschitz in this norm. Since the norm of the gradient should be measured in the dual norm $||\cdot||^\ast$, this means precisely that

$$
||\nabla V(x)||^\ast \leq L, \quad \text{for all } x \in \mathcal{X}.
$$

We also note that the next result does not require self-concordance of $\phi$.

**Theorem 3** (Non-smooth case). Assume $\phi$ is $1$-strongly convex w.r.t. a norm $||\cdot||$ on $\mathbb{R}^d$, that $V$ is $L$-Lipschitz in this norm (in the sense of (3.2)), and that $\alpha = 0$ (i.e., $V$ is convex). For a target accuracy $\epsilon > 0$, let $X_k \sim \mu_k$ denote the iterates of MLA with step size $\eta = \epsilon/L^2$. Then, the following convergence rate holds for the mixture distribution $\overline{\mu}_N := \frac{1}{N} \sum_{k=1}^N \mu_k$:

$$
\mathcal{D}_{KL}(\overline{\mu}_N \parallel \pi) \leq \epsilon, \quad \text{provided that } N \geq \frac{2L^2\mathcal{D}_\phi(\pi, \mu_1/2)}{\epsilon^2}.
$$

Proof. See Appendix C.4.

The assumption (3.2) is stronger than relative Lipschitzness: if $V$ satisfies (3.2) and $\phi$ is $1$-strongly convex w.r.t. $||\cdot||$, then $V$ is $L$-relatively Lipschitz with respect to $\phi$ with $L \leq \overline{L}$. When $||\cdot||$ is the Euclidean norm and $\phi = \frac{||\cdot||^2}{2}$, then we recover a special case of DMM19 Corollary 14.

We now make a number of remarks about our result.

**Remark 4** (Implementing $\overline{\mu}_N$). One can output a sample from $\overline{\mu}_N$ by simply outputting one of the iterates $\{X_k\}_{k=1}^N$ chosen uniformly at random.

**Remark 5** (Convergence in other metrics). Using standard inequalities, our results for convergence in KL divergence imply convergence in a number of other information divergences such as the total variation distance, see [Tsy09 §2.4].

When $V$ is $\alpha$-strongly convex (w.r.t. $\frac{||\cdot||^2}{2}$), then the $T_2$ transportation-cost inequality [Tsy09] Theorem 22.14 $\alpha W_2^2(\mu, \pi) \leq \mathcal{D}_{KL}(\mu \parallel \pi)$ implies convergence in the $W_2$ distance as well. In general, we do not have convergence in $W_2$, but we can always obtain convergence with respect to a different optimal transport cost, namely, the Bregman transport cost $\mathcal{D}_V$ associated with $V$. This is a consequence of Corollary 1 [also see CE17 Proposition 1], which asserts that $\mathcal{D}_V(\mu, \pi) \leq \mathcal{D}_{KL}(\mu \parallel \pi)$.

**Remark 6** (Dimension dependence). Ignoring for now the dependence on $\beta'$ (which may also have a dimension dependence depending on the application), the Bregman divergence $\mathcal{D}_\phi(\pi, \mu_0)$ term is typically of size $O(d)$ (see Section 5 for a particular instance of this). Thus, our overall dimension dependence is $O(d^5)$ for the weakly convex case and $O(d)$ for the strongly convex case. Overall, this is a significantly better dependence on the dimension as compared to the previous works [BDMPT17, BEL18], we perform a detailed comparison in Section 5 for a specific setting.

We also remark that mirror descent has classically been used for dimension reduction by changing the geometry of the algorithm from $\ell_2$ to $\ell_1$. We investigate the possibility of doing the same for sampling in Appendix C.2.

1 In this case, $M_\phi = 0$, so the Lipschitz constant $L$ does not enter the final result. In particular, it is not contradictory to assume strong convexity ($\alpha > 0$).
Remark 7 (Comparison with [HKRC18]). [HKRC18] show that for strictly log-concave targets, there exists a good mirror map \( \phi \) for which the pushforward of the target distribution via \( \nabla \phi \) enjoys the same guarantees as ordinary Langevin. However, this result is only existential and gives no guidance on how to construct the mirror map. In contrast, our theorems hold for any choice of mirror map which satisfies our assumptions, and provide guidance on how to choose the mirror map. Also, our relative smoothness condition allows for potentials which blow up at the boundary of their domain (i.e. the target distribution vanishes near the boundary of its support), whereas this is forbidden by the assumptions of [HKRC18]. Lastly, our algorithm does not require computing the third derivative of the mirror map, whereas this is required for [HKRC18].

Remark 8 (Comparison with [ZPFP20]). [ZPFP20] performs an analysis of the Euler-Maruyama discretization of [MLD] which we temporarily refer to as MLA'. Our result guarantees that for any desired accuracy \( \epsilon \), it is possible to choose the step size sufficiently small so that MLA achieves the target accuracy; in contrast, the result of [ZPFP20] only guarantees that MLA' contracts to within a ball around \( \pi \) of radius \( O(\sqrt{d}) \) (measured w.r.t. a modified Wasserstein distance). Moreover, [ZPFP20] conjecture that their bias term is unavoidable.

In light of our result, we believe that it is an interesting open question to resolve their conjecture. Their conjecture, if true, suggests that replacing MLA', which we temporarily refer to as MLA, with a single step of the Euler-Maruyama discretization has disastrous effects on the convergence of the algorithm, and therefore provides further support for considering MLA instead of MLA'.

Recently, after the first draft of our paper was published online, Li et al. [LTVW21] gave an analysis of MLA' under a subset of the assumptions of [ZPFP20] which indeed exhibits vanishing bias, provided that the relative strong convexity parameter of the potential is sufficiently large compared to the modified self-concordance parameter of the mirror map. It remains an open question to remove this latter restriction from their work, and moreover to obtain similar results under the more usual definitions of relative convexity/smoothness and self-concordance that we adopt in this work.

In order to generalize the discretization analysis from the vanilla Langevin algorithm to the mirror-Langevin algorithm, in the next section we prove a new displacement convexity result for the entropy with respect to the Bregman transport cost which may be of independent interest.

4 Convexity of the entropy with respect to the Bregman divergence

It is well-known that the entropy functional \( \mathcal{H} \) is displacement convex along \( W_2 \) geodesics [AGS08, Theorem 9.4.11]. In fact, this displacement convexity is crucial in showing that \( D_{KL}(\cdot \mid \pi) = \mathcal{E} + \mathcal{H} \) is displacement convex (when \( \pi \) is log-concave), which in turn is used to analyze the convergence of [LD] to the target measure. Therefore, in order to understand the convergence of [MLD], it is crucial to see if such a result is true when \( W_2 \) is replaced by \( D_\phi \). We prove that indeed the displacement convexity-like property holds for \( \mathcal{H} \) under \( D_\phi \)-optimal couplings.

**Theorem 4** ("Convexity" of the entropy with respect to the Bregman divergence). Let \( \mu, \nu \) be probability measures on \( \mathbb{R}^d \) and let \( X \sim \mu \), \( Y \sim \nu \) be coupled according to the Bregman transport cost \( D_\phi(\mu, \nu) \). Then, it holds that

\[
\mathcal{H}(\nu) \geq \mathcal{H}(\mu) + \mathbb{E}\langle \nabla W_2 \mathcal{H}(\mu) \rangle(X, Y - X).
\]

As a corollary, we can use the calculus of optimal transport in order to recover the transportation-cost inequality of [CE17]. In the following, we do not carry out the approximation arguments necessary to make the proof fully correct because a rigorous proof of the statement is already given in [CE17]. Rather, our main purpose in giving this argument is simply to point out the convexity principle which underlies the transport inequality.

**Corollary 1** ([CE17 Proposition 1]). For any probability measure \( \mu \) on \( \mathbb{R}^d \),

\[
D_{KL}(\mu \mid \pi) \geq D_V(\mu, \pi).
\]

**Proof Sketch.** Let \((X, Y)\) be optimally coupled according to the Bregman transport cost \( D_V(\cdot \mid \cdot) \) between \( \mu \) and \( \pi \). We decompose \( D_{KL}(\mu \mid \pi) = \mathbb{E} V(X) + \mathcal{H}(\mu) \). On one hand, the first term is

\[
\mathbb{E} V(X) = \mathbb{E} V(Y) + \mathbb{E} \langle \nabla V(Y) \rangle, X - Y + \mathbb{E} D_V(X, Y).
\]
On the other hand, the convexity result (Theorem 4) shows that
\[ \mathcal{H}(\mu) \geq \mathcal{H}(\pi) + \mathbb{E}(\nabla W_2 \mathcal{H}(\pi)(Y), X - Y). \]

Putting these together, we obtain
\[
\mathcal{D}_{\text{KL}}(\mu \parallel \pi)
\geq \mathbb{E} V(Y) + \mathbb{E}(\nabla V(Y), X - Y) + \mathbb{E} D_V(X, Y) + \mathcal{H}(\pi) + \mathbb{E}(\nabla W_2 \mathcal{H}(\pi)(Y), X - Y)
\]
\[= \mathcal{D}_{\text{KL}}(\pi \parallel \pi) + \mathbb{E}(\nabla V + \nabla W_2 \mathcal{H}(\pi))(Y), X - Y) + \mathbb{E} D_V(X, Y) = \mathbb{E} D_V(X, Y), \]

since \(\nabla V + \nabla W_2 \mathcal{H}(\pi)\), the \(W_2\) gradient of \(\mathcal{D}_{\text{KL}}(\cdot \parallel \pi)\) at \(\pi\), is zero. This proves the result. \qed

5 Application to Bayesian logistic regression

In this section, we apply our main result to Bayesian logistic regression. For more applications of our result, such as the possibility of dimension reduction and sampling from non-smooth distributions, we refer the readers to Appendix E.

We recall the setting of Bayesian logistic regression: we observe pairs \((X_i, Y_i), i = 1, \ldots, n\), where \(X_i \in \mathbb{R}^d\) and \(Y_i \in \{0, 1\}\). The data are assumed to follow the model

\[ Y_i \sim \text{Bernoulli}\left(\frac{\exp \langle \theta, X_i \rangle}{1 + \exp \langle \theta, X_i \rangle}\right), \quad \text{independently for } i = 1, \ldots, n. \tag{5.1} \]

Here, the parameter \(\theta\) itself is assumed to be a random variable taking values in \(\mathbb{R}^d\). If we assume that \(\theta\) has a prior density \(\lambda\) with respect to Lebesgue measure, then the posterior distribution is

\[ \pi(\theta) \propto \lambda(\theta) \exp \left[ \sum_{i=1}^{n} (Y_i \langle \theta, X_i \rangle - \ln(1 + \exp \langle \theta, X_i \rangle)) \right]. \]

Since it may be computationally infeasible to explicitly compute the normalizing constant for the posterior distribution, we turn towards sampling algorithms.

When we take a prior \(\lambda\) which has full support on \(\mathbb{R}^d\), e.g. a Gaussian prior, then we may apply off-the-shelf methods such as the Langevin diffusion \([4]\). However, if we choose a prior which has compact support, then the unadjusted Langevin algorithm is no longer an acceptable option because it outputs samples outside the support of the posterior. In this case, we must turn to other methods, such as the projected Langevin algorithm \([BEL18]\). Here, we explore the use of the mirror-Langevin algorithm \([MLA]\) for constrained sampling.

For the rest of this section, we will focus on a particular problem for concreteness and interpretability: we consider the uniform prior \(\lambda\) on the \(\ell_1\)-ball \([-1, 1]^d\). By duality, this is an attractive model when the data \((X_i)_{i=1}^n\) have small \(\ell_1\)-norm, i.e., are approximately sparse. A natural choice of mirror map for this problem is the logarithmic barrier

\[ \phi(\theta) = \sum_{i=1}^{d} \left( \ln \frac{1}{1 - \theta[i]} + \ln \frac{1}{1 + \theta[i]} \right), \]

where we use \(\theta[i]\) to denote the coordinates of \(\theta \in \mathbb{R}^d\). Then, \(\phi\) is 1-self-concordant \(([Nes13] \S 5.1.3)\).

We remark that the separability of the mirror map in this example implies that the diffusion step \([MLA, 2]\) can be simulated in \(O(d)\) steps, rather than \(O(d^2)\).

For this setting, we compare the guarantees of \([MLA]\) with the Projected Langevin Algorithm (PLA) \([BEL18]\) and the Moreau-Yosida unadjusted Langevin algorithm (MYULA) \([BDMP17]\); see Table 1. The details of the comparison are given in Appendix E.1.

We also perform a numerical experiment to compare the practical performance of \([MLA]\) with PLA. We take \(\theta^* := (0.9, \ldots, 0.9) \in \mathbb{R}^{10}\) as the ground truth, and we generate 1000 i.i.d. pairs \((X_i, Y_i)\) where \(X_i\) is sampled uniformly from the \(\ell_1\) ball and \(Y_i\) is generated from \(X_i\) according to (5.1) with \(\theta = \theta^*\). We generate 30 samples using both \([MLA]\) and PLA (both with step size \(\eta = 0.005\)). At each iteration, we average the samples to obtain an estimate \(\theta_k\) for the posterior mean, and we plot the error \(\|\theta_k - \theta^*\|_2\) in Figure 2, averaged over 10 trials. We implement \([MLA, 2]\) by performing 10 inner iterations of an Euler-Maruyama discretization.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Guarantee</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLA</td>
<td>$O(d/\epsilon^2)$</td>
</tr>
<tr>
<td>PLA</td>
<td>$O(d^{15}/\epsilon^6)$</td>
</tr>
<tr>
<td>MYULA</td>
<td>$O(d^9/\epsilon^3)$</td>
</tr>
</tbody>
</table>

Table 1: Comparison of MLA with other constrained sampling algorithms for the number of iterations required to output a sample whose squared total variation distance to $\pi$ is at most $\epsilon$. For simplicity, we focus on the dependence with respect to dimension and the accuracy $\epsilon$, and we defer details of the comparison to Appendix E.1.

Figure 2: We plot the error for estimators of the posterior mean computed using both MLA and PLA, each taken with step size $\eta = 0.005$.

Acknowledgments. Kwangjun Ahn was supported by graduate assistantship from the NSF Grant (CAREER: 1846088) and by Kwanjeong Educational Foundation. Sinho Chewi was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. We thank Philippe Rigollet for helpful suggestions which greatly improved the presentation of this paper.

References


A Open questions

We conclude by discussing some questions for future research.

1. As we discuss in Remark 8, it is an open question to determine if the analyses of [ZPFP20, LTVW21] can be improved to obtain vanishing bias for the Euler-Maruyama discretization of MLD under weaker assumptions.

2. In our work, we analyze the sampling analogue of mirror descent under the assumption that the mirror map is self-concordant. This notably bears resemblance to the development of interior-point methodology in optimization [NN94], and it is an interesting problem to develop further sampling analogues of interior-point algorithms.

3. In Appendix E.2, we conduct a preliminary investigation into the possibility that MLA can alleviate the dependence on dimension for some sampling problems. However, Metropolis-adjusted variants of the Langevin algorithm enjoy significantly better dependence on the dimension as compared to their unadjusted counterparts; see [RR98, PST12, CLA+21]. Thus, the Metropolis-adjusted version of MLA may be a more appropriate setting in which to investigate this dimension reduction question, which we leave to future work.

B Details on relative convexity and smoothness

For the reader’s convenience, we list basic facts regarding relative convexity and smoothness.

Proposition 1 ([LFN18, Proposition 1.1]). The following conditions are equivalent:

- $f$ is $\beta$-smooth relative to $h$.
- $\beta h - f$ is convex on $Q$.
- Under twice differentiability, $\nabla^2 f(x) \leq \beta \nabla^2 h(x)$ for any $x \in \text{int}(Q)$.
- $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \beta \langle \nabla h(x) - \nabla h(y), x - y \rangle$ for all $x, y \in \text{int}(Q)$.

Furthermore, the following conditions are equivalent:

- $f$ is $\alpha$-convex relative to $h$.
- $f - \alpha h$ is convex on $Q$.
- Under twice differentiability, $\nabla^2 f(x) \geq \alpha \nabla^2 h(x)$ for any $x \in \text{int}(Q)$.
- $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \langle \nabla h(x) - \nabla h(y), x - y \rangle$ for all $x, y \in \text{int}(Q)$.

C Proof of the convergence rates

C.1 Per-iteration progress bound

For the convergence rates of MLA, we first prove the following per-iterate progress bound, from which Theorems 1 and 2 will be easily deduced.

Lemma 1 (Per-iteration progress bound). Assume $\beta > 0$. For $0 \leq \eta \leq \frac{1}{\beta}$, let $X_k \sim \mu_k$ be the iterates of MLA with step size $\eta$. Then, under Assumptions [L] the following holds:

$$\eta \mathcal{D}_{KL}(\mu_{k+1} \parallel \pi) \leq (1 - \alpha \eta) \mathcal{D}_{\phi}(\pi, \mu_k) - \mathcal{D}_{\phi}(\pi, \mu_{k+1}) + (\beta + 2M_P L)d\eta^2. \tag{C.1}$$

Proof. We decompose the KL divergence into two parts:

$$\mathcal{D}_{KL}(\mu \parallel \pi) = \int_Q V(x) \, d\mu(x) + \int_Q \mu(x) \ln \mu(x) \, dx =: \mathcal{E}(\mu) + \mathcal{H}(\mu).$$

Here and throughout the paper, we abuse notation by identifying a measure $\mu$ with its density.
The first term above has the interpretation of energy, while the second term has the interpretation of (negative) entropy. The basic scheme of the proof follows the method in \cite{DMM19}, which views the two steps of the update rule as alternately dissipating the energy and the entropy. More specifically, we will show that \textit{MLA:1} dissipates \( \mathcal{E} \) and \textit{MLA:2} dissipates \( \mathcal{H} \), while the two steps do not badly interfere with each other.

Our analysis proceeds by controlling each term in the following decomposition:

\[
\mathcal{D}_{\text{KL}}(\mu_{k+1} \parallel \pi) = \mathcal{E}(\mu_{k+1}) + \mathcal{H}(\mu_{k+1}) - \mathcal{E}(\pi) - \mathcal{H}(\pi) \\
= \mathcal{E}(\mu_{k+1/2}) - \mathcal{E}(\pi) + \mathcal{E}(\mu_{k+1}) - \mathcal{E}(\mu_{k+1/2}) + \mathcal{H}(\mu_{k+1}) - \mathcal{H}(\pi) .
\]

Before we go into the analysis of each term, we outline our proof strategy. Term \( \text{(1)} \) corresponds to a deterministic step of the mirror descent algorithm, and we adapt the analysis of mirror descent based on the Bregman proximal inequality \cite[Lemma 3.2]{CT93}. For terms \( \text{(2)} \) and \( \text{(3)} \), it will be important to understand the stochastic process \(( Z_t )_{t \in [0, \eta]} \) in \textit{MLA} where \( Z_t := \nabla \phi^*(W_t) \), along with the corresponding marginal laws \(( \nu_t )_{t \in [0, \eta]} \). There are two important and distinct perspectives we can adopt. On one hand, the stochastic process \(( Z_t )_{t \in [0, \eta]} \) is a diffusion, and can be studied via stochastic calculus. On the other hand, the laws \(( \nu_t )_{t \in [0, \eta]} \) follow a Wasserstein “mirror flow” of the entropy functional \( \mathcal{H} \), in the sense that it evolves continuously in Wasserstein space with tangent vector \(-\nabla^2 \phi \) \( \nabla_w \mathcal{H}(\nu_t) \) (see Section 2.1) for a brief introduction to Wasserstein calculus, and \cite{CLGL+20} for a discussion of \textit{MLD} from this perspective. In turn, these two perspectives offer different calculation rules: stochastic calculus provides Itô’s formula (see \cite[Theorem 5.10]{LG16} or \cite[§3.3]{Str18}), while Wasserstein calculus provides the rule

\[
\frac{d}{dt} \mathcal{F}(\nu_t) = -E(\nabla W_2 \mathcal{F}(\nu_t)(Z_t), [\nabla^2 \phi(Z_t)]^{-1} \nabla W_2 \mathcal{H}(\nu_t)(Z_t)),
\]

for any sufficiently well-behaved functional \( \mathcal{F} \) on Wasserstein space. Both of these perspectives are insightful, and we will employ both.

For term \( \text{(2)} \), we show that \textit{MLA:2} does not greatly increase the energy, and we accomplish this via calculations using Itô’s formula together with the relative smoothness and self-concordance assumptions. Finally, we control term \( \text{(3)} \) by developing a new displacement convexity result (Theorem \[ \text{[4]} \]) for the entropy functional \( \mathcal{H} \), which is crucial for applying Wasserstein calculus.

1: Let \( Y \) be a random variable (defined on the same probability space) which is distributed according to \( \pi \). Then,

\[
\mathcal{E}(\mu_{k+1/2}) - \mathcal{E}(\pi) = \mathbb{E}[V(X_{k+1/2})] - \mathbb{E}[V(Y)] \\
= \mathbb{E}[V(X_{k+1/2})] - \mathbb{E}[V(X_k)] + \mathbb{E}[V(X_k)] - \mathbb{E}[V(Y)] \\
\leq \mathbb{E}[(\nabla V(X_k), X_{k+1/2} - X_k) + \beta D_\phi(X_{k+1/2}, X_k)] \\
+ \mathbb{E}[(\nabla V(X_k), X_k - Y) - \alpha D_\phi(Y, X_k)] \\
= \mathbb{E}[(\nabla V(X_k), X_{k+1/2} - Y) + \beta D_\phi(X_{k+1/2}, X_k) - \alpha D_\phi(Y, X_k)],
\]

where the inequality follows due to the \( \alpha \)-relative strong convexity and \( \beta \)-relative smoothness of \( V \). Now to control \( \text{(2)} \), we invoke a standard tool from optimization:

**Lemma 2** (Bregman proximal inequality \cite[Lemma 3.2]{CT93}). For a convex function \( f \) and a convex function \( \phi \) of Legendre type, suppose that

\[
x_+ := \arg \min_{z \in \mathcal{Q}} [f(z) + D_\phi(z, x)].
\]

Then,

\[
f(x_+) - f(y) \leq D_\phi(y, x) - D_\phi(y, x_+) - D_\phi(x_+, x) \quad \forall y \in \mathcal{Q}.
\]
Applying the Bregman proximal inequality (Lemma 2) with \( f(x) = \eta (\nabla V(X_k), x) \),

\[ (C.2) \leq \mathbb{E}\left[ \frac{1}{\eta} - \alpha \right] \mathbb{D}_\phi(Y, X_k) - \frac{1}{\eta} \mathbb{D}_\phi(Y, X_{k+1/2}) + (\beta - \frac{1}{\eta}) \mathbb{D}_\phi(X_{k+1/2}, X_k) \]

\[ \leq \mathbb{E}\left[ \frac{1}{\eta} - \alpha \right] \mathbb{D}_\phi(Y, X_k) - \frac{1}{\eta} \mathbb{D}_\phi(Y, X_{k+1/2}) \]

provided that \( \frac{1}{\eta} \geq \beta \Leftrightarrow \eta \leq \beta^{-1} \). Choosing \( Y \) so that the coupling \((Y, X_k)\) minimizes \( \mathbb{E}[\mathbb{D}_\phi(Y, X_k)] \), we obtain

\[ \eta \{ \mathbb{E}[\mu_{k+1/2}] - \mathbb{E}(\pi) \} \leq (1 - \alpha\eta) \mathbb{D}_\phi(\pi, \mu_k) - \mathbb{E}[\mathbb{D}_\phi(Y, X_{k+1/2})] \]

\[ \leq (1 - \alpha\eta) \mathbb{D}_\phi(\pi, \mu_k) - \mathbb{D}_\phi(\pi, \mu_{k+1/2}) . \]

2: First, note from [MLA] that

\[ \mathbb{E}[\mu_{k+1}] - \mathbb{E}[\mu_{k+1/2}] = \mathbb{E}[V(\nabla \phi^*(W_\eta)) - V(\nabla \phi^*(W_0))] . \]

To compute the above term, we define \( f(x) := V(\nabla \phi^*(x)) \) and apply Itô’s formula to the random variable \( f(W_\eta) - f(W_0) \). To that end, we first compute the Hessian of \( f \):

\[
\nabla f(x) = \nabla V(\nabla \phi^*(x))^T \nabla^2 \phi^*(x) = \nabla V(\nabla \phi^*(x))^T [\nabla^2 \phi(\nabla \phi^*(x))]^{-1},
\]

\[
\nabla^2 f(x) = \nabla^2 V(\nabla \phi^*(x)) [\nabla^2 \phi(\nabla \phi^*(x))]^{-1} [\nabla^2 \phi^*(x)]
\]

\[
+ \nabla V(\nabla \phi^*(x))^T [\nabla^2 \phi(\nabla \phi^*(x))]^{-1} [\nabla^3 \phi(\nabla \phi^*(x))] [\nabla^2 \phi(\nabla \phi^*(x))]^{-2} .
\]

Itô’s formula now decomposes \( f(W_\eta) - f(W_0) \) into the sum of an integral and a stochastic integral. Intuitively, the stochastic integral has mean zero (since it is a local martingale), and this can be rigorously argued using the standard technique of localization; we give the argument at the end of this step. Thus, we concentrate on the expectation of the first term. Writing \( Z_t := \nabla \phi^*(W_t) \), the above Hessian calculation gives

\[ \mathbb{E}[f(W_\eta) - f(W_0)] \]

\[
= \mathbb{E}\left[ \int_0^\eta \langle \nabla^2 V(Z_t) [\nabla^2 \phi(\nabla \phi^*(x))]^{-2}, \nabla^2 \phi(\nabla \phi^*(x)) \rangle dt \right]
\]

\[
+ \mathbb{E}\left[ \int_0^\eta \langle \nabla V(Z_t)^T [\nabla^2 \phi(\nabla \phi^*(x))]^{-1} [\nabla^3 \phi(\nabla \phi^*(x))] [\nabla^2 \phi(\nabla \phi^*(x))]^{-2}, \nabla \phi(\nabla \phi^*(x)) \rangle dt \right]
\]

\[ \leq \mathbb{E}\left[ \int_0^\eta \langle \nabla^2 V(Z_t), [\nabla^2 \phi(\nabla \phi^*(x))]^{-1} \rangle dt \right]
\]

\[ + \mathbb{E}\left[ \int_0^\eta \text{tr}(\nabla V(Z_t)^T [\nabla^2 \phi(\nabla \phi^*(x))]^{-1} [\nabla^3 \phi(\nabla \phi^*(x))] [\nabla^2 \phi(\nabla \phi^*(x))]^{-1} ) dt \right] . \]

(C.5)

We can control (C.4) easily based on the relative smoothness of \( V \): indeed, since \( \nabla^2 V \leq \beta \nabla^2 \phi \) (see Appendix B),

\[ (C.4) \leq \beta \eta . \]

To control (C.5), we use the self-concordance of \( \phi \). We recall here the following result:

Proposition 2 ([Nes18 Corollary 5.1.1]). A function \( \phi \) is self-concordant with a constant \( M_\phi \geq 0 \) if and only if for any \( x \in \text{dom}(\phi) \) and any direction \( u \in \mathbb{R}^n \) we have

\[ \nabla^3 \phi(x) u \leq 2M_\phi \| u \| \nabla^2 \phi(x) \nabla^2 \phi(x) \].

Using Proposition 2, it follows that

\[ \text{(C.5)} \leq 2M_\phi \int_0^\eta \mathbb{E}\left[ \|\nabla^2 \phi(Z_t)\|^{-1} \nabla V(Z_t) \|\nabla^2 \phi(Z_t)\|^{-1} \right] dt \]

\[ \leq 2M_\phi \int_0^\eta \mathbb{E}\left[ \|\nabla V(Z_t)\| \|\nabla^2 \phi(Z_t)\|^{-1} \right] dt \leq 2M_\phi \mathbb{L} \eta . \]
Thus, our calculation shows that
\[
E(\mu_{k+1}) - E(\mu_{k+1/2}) \leq (\beta + 2M_\phi L)d\eta. \tag{C.6}
\]
We now sketch the localization argument. Let \((\tau_\ell)_{\ell \in \mathbb{N}}\) be a localizing sequence for \((W_\ell)_{\ell \in [0,\eta]}\). The argument above may be applied rigorously for the stopped process \((W_\ell \wedge \tau_\ell)_{\ell \in [0,\eta]}\) to obtain \(E V(Z_\ell \wedge \tau_\ell) - E V(Z_0) \leq (\beta + 2M_\phi L)d\eta\). Since \(V\) is bounded below, we use Fatou’s lemma to pass \(\ell \to \infty\) and deduce (C.6).

3: Let \(\nu_t\) denote the law of \(Z_t := \nabla \phi^*(W_t)\). For this step, we start with a calculation of the derivative of \(t \mapsto D_\phi(\pi, \nu_t)\). Noting that \(\nabla_2 D_\phi(y, x) = -\nabla^2 \phi(x)(y - x)\) and that \(\nu_t\) follows the Wasserstein tangent vector \(-\nabla^2 \phi^{-1} \nabla W_2 \mathcal{H}(\nu_t)\), we expect that
\[
\frac{d}{dt} D_\phi(\pi, \nu_t) = E\langle [\nabla^2 \phi(Z_t)]^{-1} \nabla W_2 \mathcal{H}(\nu_t)(Z_t), \nabla^2 \phi(Z_t)(Y - Z_t) \rangle
\]
where \((Y, Z_t)\) are optimally coupled for \(\pi\) and \(\nu_t\) for the Bregman transport cost. In general, the differentiability properties of optimal transport costs can be quite subtle, but thankfully it is much easier to establish the superdifferentiability
\[
\frac{d}{dt} D_\phi(\pi, \nu_t) \leq E\langle \nabla W_2 \mathcal{H}(\nu_t)(Z_t), Y - Z_t \rangle
\]
as shown below, we use Fatou’s lemma to pass \(t \to \infty\) and deduce (C.6).

Next, we apply a result which can be interpreted as convexity of the entropy functional with respect to the Bregman divergence; it will be given as Theorem 4 in the next section. It implies that for \(t \in [0, \eta]\),
\[
\frac{d}{dt} D_\phi(\pi, \nu_t) \leq E\langle \nabla W_2 \mathcal{H}(\nu_t)(Z_t), Y - Z_t \rangle \leq \mathcal{H}(\pi) - \mathcal{H}(\nu_t) \leq \mathcal{H}(\pi) - \mathcal{H}(\nu_\eta),
\]
where the last inequality follows since
\[
\frac{d}{dt} \mathcal{H}(\nu_t) = -E\langle (\nabla W_2 \mathcal{H}(\nu_t)(Z_t), [\nabla^2 \phi(Z_t)]^{-1} \nabla W_2 \mathcal{H}(\nu_t)(Z_t)) \rangle \leq 0,
\]
which implies \(\mathcal{H}(\nu_t) \leq \mathcal{H}(\nu_t)\) for any \(t \in [0, \eta]\). Integrating from 0 to \(\eta\), we obtain
\[
D_\phi(\pi, \nu_\eta) - D_\phi(\pi, \nu_0) \leq \eta \{\mathcal{H}(\pi) - \mathcal{H}(\nu_\eta)\},
\]
which is the same as
\[
\eta \{\mathcal{H}(\mu_{k+1}) - \mathcal{H}(\pi)\} \leq D_\phi(\pi, \mu_{k+1/2}) - D_\phi(\pi, \mu_{k+1}).
\]
Combining the upper bounds from (1), (2), and (3), the proof of Lemma 1 is complete. \(\Box\)

C.2 Proof of Theorem 1

From the per-iteration progress bound (Lemma 1), we have for any \(k \in \mathbb{N}\)
\[
\eta D_{KL}(\mu_{k+1} \| \pi) \leq D_\phi(\pi, \mu_k) - D_\phi(\pi, \mu_{k+1}) + \beta' d\eta^2. \tag{C.7}
\]
Summing (C.7) over \(k = 0, 1, \ldots, N - 1\),
\[
\eta \sum_{k=1}^N D_{KL}(\mu_k \| \pi) \leq D_\phi(\pi, \mu_0) - D_\phi(\pi, \mu_N) + \beta' d\eta^2 N.
\]
Using the convexity of the KL divergence [which follows from the Gibbs variational principle; see [RAS15] §5.1],
\[
D_{KL}(\mathcal{M}_N \| \pi) \leq \frac{D_\phi(\pi, \mu_0)}{\eta} \leq \frac{\beta' d\eta^2 N}{2} + \frac{\epsilon}{2},
\]
where the last inequality follows from the choice \(N \geq \frac{2D_\phi(\pi, \mu_0)}{\beta' d\eta^2} \) and \(\eta \leq \frac{\epsilon}{2\beta' d}\).
C.3 Proof of Theorem 2

Let us first prove the convergence in Bregman transport cost. For any \( k \in \mathbb{N} \), the per-iteration progress bound (Lemma 1) together with the fact \( D_{KL}(\mu_{k+1} \parallel \pi) \geq 0 \) imply

\[
D_{\phi}(\pi, \mu_{k+1}) \leq (1 - \alpha \eta) D_{\phi}(\pi, \mu_{k}) + \beta' d\eta^2.
\]  
(C.8)

Recursively applying (C.8) for \( k = 0, 1, \ldots, N - 1 \), we obtain

\[
D_{\phi}(\pi, \mu_{N}) \leq (1 - \alpha \eta)^N D_{\phi}(\pi, \mu_{0}) + \beta' d\eta^2 \sum_{k=0}^{N-1} (1 - \alpha \eta)^k
\]

\[
\leq (1 - \alpha \eta)^N D_{\phi}(\pi, \mu_{0}) + \beta' d\eta^2 \sum_{k=0}^{\infty} (1 - \alpha \eta)^k
\]

\[
\leq \exp(-\alpha \eta N) D_{\phi}(\pi, \mu_{0}) + \frac{\beta' d\eta^2}{\alpha} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2},
\]

where the last inequality follows since \( N \geq \frac{1}{\alpha \eta} \ln \frac{2D_{KL}(\mu_{0} \parallel \pi)}{\epsilon} \) and \( \eta \leq \frac{\alpha \epsilon}{2\beta' d} \). Having proved the convergence in terms of the Bregman transport cost, the convergence in terms of the KL divergence follows by applying Theorem 1.

C.4 Analysis for the non-smooth case (Theorem 3)

The analysis for the non-smooth case proceeds in a similar manner to the smooth case. We first prove the following per-iterate progress bound.

**Lemma 3** (Per-iteration progress bound; non-smooth case). Let \( X_k \sim \mu_k \) be the iterates of [MLA] with step size \( \eta > 0 \). Assume that \( \phi \) is 1-strongly convex w.r.t \( \| \cdot \| \), and that \( V \) is convex and \( L \)-Lipschitz w.r.t \( \| \cdot \| \). Then, the following holds:

\[
\eta D_{KL}(\mu_{k+1} \parallel \pi) \leq D_{\phi}(\pi, \mu_{k+1/2}) - D_{\phi}(\pi, \mu_{k+3/2}) + \frac{\eta^2 L^2}{2}.
\]  
(C.9)

**Proof.** Our analysis proceeds by controlling each term in the following decomposition:

\[
D_{KL}(\mu_{k+1} \parallel \pi) = \underbrace{E(\mu_{k+1}) - E(\pi)}_{A} + \underbrace{\mathcal{H}(\mu_{k+1}) - \mathcal{H}(\pi)}_{B}.
\]

For term \( A \), we invoke the following upper bound (from the analysis of term \( \beta \) in the proof of Lemma 1):

\[
\eta \{ \mathcal{H}(\mu_{k+1}) - \mathcal{H}(\pi) \} \leq D_{\phi}(\pi, \mu_{k+1/2}) - D_{\phi}(\pi, \mu_{k+1}).
\]  
(C.10)

Let us turn to \( \beta \), and let \( Y \) be a random variable (defined on the same probability space) which is distributed according to \( \pi \). Since we have

\[
X_{k+3/2} = \arg\min_{x \in \mathcal{Q}} [\eta \nabla V(X_{k+1}), x] + D_{\phi}(x, X_{k+1}),
\]

applying the Bregman proximal inequality (Lemma 2) with \( f(x) = \eta \langle \nabla V(X_{k+1}), x \rangle \) gives

\[
\eta \langle \nabla V(X_{k+1}), X_{k+3/2} - Y \rangle \leq D_{\phi}(Y, X_{k+1}) - D_{\phi}(Y, X_{k+3/2}) - D_{\phi}(X_{k+3/2}, X_{k+1}),
\]

which after rearranging becomes

\[
D_{\phi}(Y, X_{k+3/2}) - D_{\phi}(Y, X_{k+1}) \leq \eta \langle \nabla V(X_{k+1}), Y - X_{k+3/2} \rangle - D_{\phi}(X_{k+3/2}, X_{k+1}).
\]  
(C.11)

On the other hand, the right hand side of (C.11) can be controlled using the convexity, the Lipschitz-ness of \( V \), and strong convexity of \( \phi \):

**RHS of (C.11)**

\[
\eta \langle \nabla V(X_{k+1}), Y - X_{k+1} \rangle + \eta \langle \nabla V(X_{k+1}), X_{k+1} - X_{k+3/2} \rangle - D_{\phi}(X_{k+3/2}, X_{k+1})
\]

\[
\leq \eta [V(Y) - V(X_{k+1})] + \eta \| \nabla V(X_{k+1}) \|_* \| X_{k+1} - X_{k+3/2} \| - \frac{1}{2} \| X_{k+1} - X_{k+3/2} \|^2
\]

\[
\leq \eta [V(Y) - V(X_{k+1})] + \frac{\eta^2 L^2}{2} \| \nabla V(X_{k+1}) \|_*
\]

\[
\leq \eta [V(Y) - V(X_{k+1})] + \frac{\eta^2 L^2}{2}.
\]
For the LHS of (C.11), choose $Y$ so that the coupling $(Y, X_{k+1})$ minimizes $\mathbb{E}[D_\phi(Y, X_{k+1})]$ to obtain
\[
\mathbb{E}[\text{LHS of (C.11)}] = \mathbb{E}[D_\phi(Y, X_{k+3/2})] - D_\phi(\pi, \mu_{k+1}) \geq D_\phi(\pi, \mu_{k+3/2}) - D_\phi(\pi, \mu_{k+1}).
\]
Combining these upper and lower bounds, (C.11) becomes:
\[
\eta [\mathbb{E}(\mu_{k+1}) - \mathbb{E}(\pi)] \leq D_\phi(\pi, \mu_{k+1}) - D_\phi(\pi, \mu_{k+3/2}) + \frac{\eta^2 \tilde{L}^2}{2}.
\]
Together with (C.10), the proof is complete.

Now using Lemma 3 we prove Theorem 3.

**Proof of Theorem 3.** From Lemma 3 we have for any $k \in \mathbb{N}$
\[
\eta D_{KL}(\mu_{k+1} || \pi) \leq D_\phi(\pi, \mu_{k+1/2}) - D_\phi(\pi, \mu_{k+3/2}) + \frac{\eta^2 \tilde{L}^2}{2}.
\]
(C.12)

Summing (C.12) over $k = 0, 1, \ldots, N - 1$,
\[
\eta \sum_{k=1}^{N} D_{KL}(\mu_k || \pi) \leq D_\phi(\pi, \mu_{1/2}) - D_\phi(\pi, \mu_{N+1/2}) + \frac{\eta^2 \tilde{L}^2}{2} N.
\]
Again using the convexity of the KL divergence, we obtain
\[
D_{KL}(\mu_N || \pi) \leq \frac{D_\phi(\pi, \mu_{1/2})}{N \eta} + \frac{\eta \tilde{L}^2}{2} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2},
\]
where the last inequality follows from the choice $N \geq \frac{2D_\phi(\pi, \mu_{1/2})}{\eta \epsilon}$ and $\eta \leq \frac{\epsilon}{\tilde{L}^2}$.

### D Proofs for the convexity of entropy

To prove Theorem 4, we will use the known result about the convexity of $\mathcal{H}$ along generalized geodesics [AGS08, Theorem 9.4.11]. To that end, the first step is to obtain a characterization of the optimal Bregman transport coupling which is analogous to Brenier’s theorem. The following theorem is of independent interest:

**Theorem 5** (Brenier’s theorem for the Bregman transport cost). Let $\mu, \nu$ be probability measures on $\mathbb{R}^d$. The optimal Bregman transport coupling $(X, Y)$ for $\mu$ and $\nu$ is of the form
\[
\nabla \phi(X) - \nabla \phi(Y) = \nabla h(X),
\]
where $h : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ is such that $\phi - h$ is convex.

**Proof.** From the general theory of optimal transport duality, it holds that
\[
\nabla_1 D_\phi(X, Y) = \nabla h(X),
\]
where $h$ is a $D_\phi$-concave function [see [Vil09, Theorem 10.28]]. The left-hand side of this equation evaluates to $\nabla \phi(X) - \nabla \phi(Y)$, so we simply have to check that $D_\phi$-concavity of $h$ implies that $\phi - h$ is convex (which is in fact equivalent to saying that $h$ is 1-relatively smooth with respect to $\phi$, see Appendix B).

Recall that the $D_\phi$-concavity of $h$ means there exists a function $\tilde{h} : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ such that
\[
h(x) = \inf_{y \in \mathbb{R}^d} \{D_\phi(x, y) - \tilde{h}(y)\},
\]
In fact, there are three assumptions for [Vil09, Theorem 10.28]. Here, we explicitly check them one by one for clarity. (i) **Super-differentiability:** $D_\phi$ is clearly differentiable on $\mathbb{R}$ as $\phi$ is of class $C^3$. (ii) **Injectivity of gradient:** $\nabla_1 D_\phi(x, \cdot) = \nabla \phi(x)$ is injective as $\phi$ is of Legendre type. (iii) **$\mu$-almost-sure differentiability of $D_\phi$-concave functions:** In (D.1), we actually show that for any $D_\phi$-concave function $h$, $\phi - h$ is convex and thus differentiable Lebesgue a.e. [Koc70, Theorem 25.5]. Since $\mu$ is absolutely continuous w.r.t. Lebesgue measure and $\phi$ is differentiable, $h$ must be differentiable $\mu$-almost surely.
We now prove Theorem 4 using Theorem 5:

We may compute

\[ \phi(x) - h(x) = \sup_{y \in \mathbb{R}^d} \{ \langle \nabla \phi(y), x - y \rangle + \tilde{h}(y) + \phi(y) \}. \tag{D.1} \]

As a supremum of affine functions, we see that \( \phi - h \) is convex, which completes the proof. \( \square \)

We now prove Theorem 4 using Theorem 5.

**Proof of Theorem 4.** By Theorem 5, the optimal Bregman transport coupling is of the form \( \nabla \phi(Y) = \nabla (\phi - h)(X) = \nabla \phi(X) \), where we have defined the convex function \( \zeta := \phi - h \). Hence, letting \( \tau \) denote the law of \( Y := \nabla \phi(Y) \), it follows that \( (X, \nabla \zeta(X)) \) is a \( W_2 \) optimal coupling between \( \mu \) and \( \tau \). Furthermore, since \( \phi \) is a convex function of Legendre type, \( (\nabla \zeta(X), \nabla \phi^* \circ \nabla \zeta(X)) \) is also a \( W_2 \) optimal coupling between \( \tau \) and \( \nu \). Noting that

\[ \nabla \phi^* \circ \nabla \zeta(X) = \nabla \phi^* \circ \nabla \phi(Y) = Y, \]

it follows that \( (X, Y) \) is a generalized geodesic according to \( W_2 \). Therefore, the convexity of \( \mathcal{H} \) along generalized geodesics \( [AGS08 \text{ Theorem 9.4.11}] \) concludes the proof (see \[SKL20 \text{ Lemma 4}] \). \( \square \)

**Remark 9.** For reader’s convenience, we provide a direct calculation that (formally) shows the convexity result. Since we have shown \( Y = \nabla \phi^* \circ \nabla \zeta(X) \), the change of variable formula gives

\[ \mathcal{H}(\nu) = \int \nu(y) \ln \nu(y) \, dy = \int \mu(x) \ln \mu(x) \, dx \]

\[ = \int \mu(x) \ln \det \nabla(\nabla \phi^* \circ \nabla \zeta(x)) \, dx. \]

(Here the change of variables is valid since \( [\nabla(\nabla \phi^* \circ \nabla \zeta)](x) = [\nabla^2 \phi^*(\nabla \zeta(x))] [\nabla^2 \zeta(x)] \geq 0 \).)

Thus, using the convexity of \( -\ln \det \) and integrating by parts, we obtain

\[ \mathcal{H}(\mu) - \mathcal{H}(\nu) = -\int \mu(x) \ln \det \nabla(\nabla \phi^* \circ \nabla \zeta)(x) \, dx \]

\[ \geq -\int \mu(x) \text{tr}[\nabla(\nabla \phi^* \circ \nabla \zeta)(x) - I_d] \, dx = \int (\nabla \mu(x), (\nabla \phi^* \circ \nabla \zeta)(x) - x) \, dx \]

\[ = \int (\nabla \ln \mu(x), (\nabla \phi^* \circ \nabla \zeta)(x) - x) \, d\mu(x). \]

Recalling that \( \nabla_{W_2} \mathcal{H}(\mu) = \nabla \ln \mu \) and \( Y = \nabla \phi^*(\nabla \zeta(X)) \), the convexity result follows.

**E Further applications**

**E.1 Details for the Bayesian logistic regression application**

We may compute

\[ V(\theta) = \sum_{i=1}^{n} \left( -Y_i \langle \theta, X_i \rangle + \ln(1 + \exp \langle \theta, X_i \rangle) \right), \quad \phi(\theta) = \sum_{i=1}^{d} \left( \ln \frac{1}{1 - \theta[i]} + \ln \frac{1}{1 + \theta[i]} \right) \]

\[ \nabla V(\theta) = -\sum_{i=1}^{n} \left( Y_i - \frac{\exp \langle \theta, X_i \rangle}{1 + \exp \langle \theta, X_i \rangle} \right) X_i, \quad \nabla \phi(\theta) = \sum_{i=1}^{d} \left( \frac{1}{1 - \theta[i]} - \frac{1}{1 + \theta[i]} \right) e_i, \]

\[ \nabla^2 V(\theta) = \sum_{i=1}^{n} \left( \frac{\exp \langle \theta, X_i \rangle}{(1 + \exp \langle \theta, X_i \rangle)^2} \right) X_i X_i^\top, \quad \nabla^2 \phi(\theta) = \text{diag}\left[ \frac{1}{(1 - \theta)^2} + \frac{1}{(1 + \theta)^2} \right]. \]

\[ ^3 \text{In Villani’s book, he works with the definition of } c\text{-convexity rather than } c\text{-concavity, but this is merely a matter of convention; c.f. } [Vil03 \text{ §2.4}] \text{ for the conventions regarding } c\text{-concavity.} \]
From these expressions, we see that
\[ 0 \leq \nabla^2 V \preceq \sum_{i=1}^{n} X_i X_i^\top, \quad 2I_d \preceq \nabla^2 \phi. \]

Let \( L := \sup_{[-1,1]^d} \| \nabla V \| \) denote the (ordinary) Lipschitz constant of \( V \), and let \( \beta \) denote the (ordinary) smoothness parameter of \( V \) (from above we see that \( \beta \) can be taken to be the largest eigenvalue of \( \sum_{i=1}^{n} X_i X_i^\top \)). Note that the 2-strong convexity of \( \phi \) implies that \( V \) is \( L/\sqrt{2} \)-relatively Lipschitz and \( \beta/2 \)-relatively smooth with respect to \( \phi \), so Theorem 1 holds with
\[ \beta' = \frac{\beta}{2} + \sqrt{2L}. \]

In order to fully understand the quantitative convergence rate provided by Theorem 1 we must also bound the Bregman divergence \( D_{\phi}(\pi, \mu_0) \). We have:

**Lemma 4.** Let \( \mu_0 = \delta_0 \) be the point mass at 0. Then, for the logarithmic barrier mirror map \( \phi \) defined above, we have \( D_{\phi}(\pi, \mu_0) \leq 4.1(1 + \beta + L) d \).

**Proof.** See Appendix E.1

From Theorem 1 we can deduce that using \( N \) iterations of MLA we can obtain a distribution \( \mu_N^{MLA} \) such that
\[ 2\| \mu_N^{MLA} - \pi \|_{TV}^2 \leq \mathcal{D}_{KL}(\mu_N^{MLA} \parallel \pi) \leq \epsilon, \quad \text{provided} \quad N \geq \frac{23(1 + \beta + L)^2 d^2}{\epsilon^2} \max\{1, \frac{\epsilon}{2d}\}, \]
where \( \beta \) is the largest eigenvalue of \( \sum_{i=1}^{n} X_i X_i^\top \) and \( L := \sup_{[-1,1]^d} \| \nabla V \| \) is the usual Lipschitz constant of \( V \); for details, see Appendix E.1. In fact, if we use the non-smooth guarantee in Theorem 3 then we can improve this to \( O(L^2 d/\epsilon^2) \) iterations. For comparison purposes, the guarantee for the projected Langevin algorithm (PLA) \[BEL18, \text{Theorem 1} \] with \( R = \sqrt{d} \) implies
\[ \| \mu_N^{PLA} - \pi \|_{TV}^2 \leq \epsilon, \quad \text{provided} \quad N \geq \tilde{\Omega}\left(\frac{(\sqrt{d} + \beta + L)^{12} d^3}{\epsilon^6}\right). \]

On the other hand, the Moreau-Yosida unadjusted Langevin algorithm (MYULA) \[BDMPT7, \text{Theorem 2} \] with \( R = \sqrt{d} \) provides the guaranteed\(^6\)
\[ \| \mu_N^{MYULA} - \pi \|_{TV}^2 \leq \epsilon, \quad \text{provided} \quad N \geq \tilde{\Omega}\left(\frac{\beta^4 d^9}{\epsilon^3}\right). \]

### E.2 Better dimension dependency via mirror Langevin

As described in \[Bub15, \S 4.3\], a classical application of mirror descent is to obtain better dependence on the dimension by changing the geometry of the optimization algorithm from \( \ell_2 \) to \( \ell_1 \). We investigate the possibility of analogous improvements in the setting of constrained sampling.

We consider a simple toy problem in which the constraint set is the interior of the filled-in simplex \( \mathcal{Q} := \{ x \in \mathbb{R}^d \mid x > 0, \sum_{i=1}^{d} x[i] < 1 \} \), and we take the potential to be a quadratic
\[ V(x) := \frac{1}{2} \langle x, Ax \rangle, \]
where \( A \in \mathbb{R}^{d \times d} \) is a symmetric positive semidefinite matrix with all entries bounded in magnitude by 1. We choose as our mirror map the barrier:
\[ \phi(x) := \sum_{i=1}^{d} \ln \frac{1}{x[i]} + \ln \frac{1}{1 - \sum_{i=1}^{d} x[i]}. \]

\(^6\)To be precise, their bound on the number of iterates required reads \( N \geq \tilde{\Omega}(\Delta_2 d^7/\epsilon^3) \), where \( \Delta_2 \) is a parameter measuring how close the domain \( \text{dom}(V) \) is to an isotropic convex body. For concreteness, we bound this parameter by \( \beta R \) following \[BDMPT7, \text{pg. 7}\].
This map is self-concordant with parameter 1.

We can compute
\[
\nabla \phi(x) = \sum_{i=1}^{d} \left( -\frac{1}{x[i]} + \frac{x[i]}{1 - \sum_{j=1}^{d} x[j]} \right) e_i,
\]
\[
\nabla^2 \phi(x) = \text{diag} \frac{1}{x^2} + \frac{I_d}{1 - \sum_{i=1}^{d} x[i]} + \frac{xx^T}{(1 - \sum_{i=1}^{d} x[i])^2}.
\]

Since \(x[i] < 1\) for all \(i = 1, \ldots, d\), it follows that \(\langle v, \text{diag}(1/x^2)v \rangle \geq \langle v, \text{diag}(1/x)v \rangle \geq \|v\|_1^2\), where the second inequality follows from the strong convexity of the entropy with respect to the \(\ell_1\)-norm. Hence, \(\phi\) is 1-strongly convex with respect to the \(\ell_1\)-norm. From our assumption on \(A\),
\[
\|\nabla V(x)\|_{\nabla^2 \phi(x)^{-1}} \leq \|\nabla V(x)\|_\infty \leq 1,
\]
\[
\langle v, \nabla^2 V(x)v \rangle \leq \left| \sum_{i,j=1}^{d} A_{i,j} v_i v_j \right| \leq \sum_{i,j=1}^{d} |v_i| |v_j| \leq \|v\|_1^2,
\]

which implies that \(V\) is 1-relatively Lipschitz and 1-relatively smooth with respect to \(\phi\), and the assumptions of Theorem 1 hold with \(\beta = 3\). In contrast, if we had instead considered the \(\ell_2\)-norm, then the Lipschitz constant of \(V\) could be as large as \(\sqrt{d}\), and the smoothness parameter of \(V\) could be as large as \(d\). Together with a warm start, this suggests that \(\text{MLA}\) could attain a better dimension dependence for this example.

**Remark 10.** Alternatively, we can apply Theorem 3 with the entropic mirror map
\[
\phi(x) = \sum_{i=1}^{d} x[i] \ln x[i] + \left( 1 - \sum_{i=1}^{d} x[i] \right) \ln \left( 1 - \sum_{i=1}^{d} x[i] \right)
\]
to the above setting; note that Theorem 3 only requires standard assumptions for mirror descent guarantees (e.g., Theorem 4.2), and does not require the mirror map to be self-concordant. In particular, \(V\) is 1-Lipschitz w.r.t. \(\|\cdot\|_1\) and \(\phi\) is strongly convex w.r.t. \(\|\cdot\|_1\), so Theorem 3 implies that \(\mathcal{D}_{\text{KL}}(\pi_N \| \pi) \leq \epsilon\) after \(N = O\left( \frac{D_{\phi}(\pi, \mu)}{\epsilon^2} \right)\) iterations. For comparison, note that the approach of Hsieh et al. [HKR18] does not apply to this example, because the pushforward of the distribution via the entropic mirror map is not log-concave.

In Figure 3 we report the results of a preliminary numerical study which indicates that \(\text{MLA}\) may have an advantage in high dimension. In this example, we generate a \(100 \times 100\) matrix \(A\) with i.i.d. entries drawn uniformly from \([-1, 1]\), and we take \(A\) to be the matrix \(\hat{A} \hat{A}^T\), rescaled so the largest magnitude of the entries is 1. We compare the performance of \(\text{MLA}\) (with step size \(\eta = 5 \times 10^{-3}\) and 10 inner iterations of the Euler-Maruyama discretization for \(\text{MLA:2}\)) with PLA (taken with step size \(\eta = 10^{-6}\)). The step sizes were tuned to be as large as possible while avoiding instabilities in the algorithms. We plot the convergence in \(W_2^2\), averaged over 10 trials.

### E.3 Sampling from non-smooth distributions

Thus far, we have focused on distributions whose potential \(V\) is bounded within its domain \(\text{dom}(V)\). However, in many applications, one is required to sample from a distribution whose potential \(V\) blows up near the boundary of its domain. Such distributions violate the standard assumptions of Lipschitzness and smoothness and hence are beyond the scope of the existing guarantees. In this subsection, we demonstrate that one can still sample from such distributions via \(\text{MLA}\), together with the relative Lipschitzness and relative smoothness.

Consider the Dirichlet distribution \(\pi\) which is defined on the interior of the filled-in simplex \(Q := \{ x \in \mathbb{R}^d \mid x > 0, \sum_{i=1}^{d} x[i] < 1 \}\) by the potential
\[
V(x) = a_0 \ln \frac{1}{1 - \sum_{i=1}^{d} x[i]} + \sum_{i=1}^{d} a_i \ln \frac{1}{x[i]},
\]
Figure 3: 30 iterations of MLA vs. PLA for a 100-dimensional example with step size $5 \times 10^{-3}$ for MLA and $10^{-6}$ for PLA. The step sizes were tuned to be as large as possible while avoiding instability. The results are averaged over 10 runs to smooth out the curves.

for some constants $a_0, a_1, \ldots, a_d > 0$, and we take $V = \phi$. Then, it is well-known that $V$ is $(\max_{i=0,1,\ldots,d} a_i^{-1/2})$-self-concordant [Nes18 Theorem 5.1.1]. Also, from $(\sum_{i=0}^d a_i)$-exp-concavity of $V$ [Nes18 Theorem 5.3.2], it holds that $\|\nabla V(x)\|_{(\nabla^2 V(x))^{-1}} \leq (\sum_{i=0}^d a_i)^{1/2}$. Therefore, it follows that $V$ is $(\sum_{i=0}^d a_i)^{1/2}$-Lipschitz, 1-convex, and 1-smooth relative to $V$, which implies that the assumptions of Theorem 2 holds with

$$\beta' = 1 + 2 \left( \max_{i=0,1,\ldots,d} a_i^{-1/2} \right) \left( \sum_{i=0}^d a_i \right)^{1/2} \leq 3\sqrt{d} \sqrt{\frac{a_{\max}}{a_{\min}}},$$

where $a_{\max} := \max_{i=0,1,\ldots,d} a_i$ and $a_{\min} := \min_{i=0,1,\ldots,d} a_i$. Therefore, one can obtain a mixture distribution $\pi_N$ after $N$ iterations of MLA such that

$$D_{KL}(\pi_N \| \pi) \leq \epsilon, \quad \text{provided that } N \geq \tilde{\Omega} \left( \sqrt{\frac{a_{\max}}{a_{\min}}} \frac{d^{3/2}}{\epsilon} \right).$$

Using this example, we perform a numerical experiment to investigate the effect of simulating the diffusion step (MLA2) more faithfully. In Figure 4, we run MLA for 30 iterations with step size 0.005. The potential $V$ and the mirror map $\phi$ are taken as above, with $a_0 = a_1 = \cdots = a_{10} = 2$. When implementing MLA2, we use $k$ inner iterations of an Euler-Maruyama discretization, where $k$ ranges in $\{1, 5, 10, 20\}$, and the results are averaged over 10 trials. We do indeed observe that a more precise implementation of MLA2 yields better results, although the difference is subtle.

F Auxiliary results

**Lemma 5.** Let $\pi$ be a probability distribution supported on $[-1, 1]^d$ which has density proportional to $\exp(-V)$. Assume that $V : [-1, 1]^d \to \mathbb{R}^d$ is $L$-Lipschitz and $\beta$-smooth. Then, we have the following bound on the marginal density $\pi_1$ of $\pi$ on the first coordinate:

$$\sup_{[-1,1]} \pi_1 \leq 3(1 + \sqrt{\beta} + L).$$
This yields the lower bound
\[
\pi_1(\theta_1) = \frac{1}{Z} \int_{[-1,1]^{d-1}} \exp(-V(\theta_1, \theta_{-1})) \, d\theta_{-1} \\
\geq \frac{\exp[-(\beta + L^2) (\theta_1 - \theta^*)^2/2 - 1/2]}{Z} \int_{[-1,1]^{d-1}} \exp(-V(\theta_1^*, \theta_{-1})) \, d\theta_{-1} \\
= \exp\left[-\frac{1}{2} (\beta + L^2) (\theta_1 - \theta^*)^2 - \frac{1}{2}\right] \sup_{[-1,1]} \pi_1.
\]

Next,
\[
1 = \int_{[-1,1]} \pi_1(\theta_1) \, d\theta_1 \geq \frac{\sup_{[-1,1]} \pi_1}{\sqrt{e}} \int_{[-1,1]} \exp\left[-\frac{1}{2} (\beta + L^2) (\theta_1 - \theta^*)^2\right] \, d\theta_1 \\
\geq \frac{\sup_{[-1,1]} \pi_1}{\sqrt{e}} \int_0^1 \exp\left[-\frac{1}{2} (\beta + L^2) x^2\right] \, dx.
\]

Let \( c := \int_0^1 \exp(-x^2) \, dx \). By splitting into the two cases \( \beta + L^2 \leq 1 \) and \( \beta + L^2 \geq 1 \), we can deduce the inequality
\[
1 \geq \frac{c \sup_{[-1,1]} \pi_1}{\sqrt{e}} \left( \frac{1}{\sqrt{\beta + L^2}} \right) (1).
\]

It yields
\[
\sup_{[-1,1]} \pi_1 \leq \frac{\sqrt{e}}{e} (\sqrt{\beta + L^2} \vee 1) \leq \frac{\sqrt{e}}{c} ((\sqrt{\beta + L}) \vee 1) \leq \frac{\sqrt{e}}{c} (1 + \sqrt{\beta + L}),
\]
which is the result. \( \square \)

\footnote{In Section 5 we used the notation \( \theta[i] \) for the \( i \)th coordinate of \( \theta \), but for the sake of simplicity we switch to the notation \( \theta_i \) for this proof.}
Proof of Lemma 4. Let $\pi_i$ denote the $i$th marginal of $\pi$. Then, since $\phi(0) = \nabla \phi(0) = 0$, we must estimate

$$D_\phi(\pi, \mu_0) = \int_{[-1,1]^d} \sum_{i=1}^d \ln \frac{1}{1 - \theta[i]} \pi(\theta) d\theta = \sum_{i=1}^d \int_{[-1,1]} \ln \frac{1}{1 - \theta[i]} \pi_i(\theta[i]) d\theta[i]$$

$$\leq C (1 + \sqrt{\beta + L}) d \int_{[-1,1]} \ln \frac{1}{1 - x^2} dx \leq \frac{3}{2} C (1 + \beta + L) d \int_{[-1,1]} \ln \frac{1}{1 - x^2} dx,$$

where $C$ is the constant from the proof of Lemma 5. It yields the result. \qed