Mirrored Langevin Dynamics

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, Volkan Cevher
Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
{ya-ping.hsieh, ali.kavis, paul.rolland, volkan.cevher}@epfl.ch

Abstract

We consider the problem of sampling from *constrained* distributions, which has posed significant challenges to both non-asymptotic analysis and algorithmic design. We propose a unified framework, which is inspired by the classical mirror descent, to derive novel first-order sampling schemes. We prove that, for a general target distribution with strongly convex potential, our framework implies the existence of a first-order algorithm achieving $\tilde{O}(\epsilon^{-2}d)$ convergence, suggesting that the state-of-the-art $\tilde{O}(\epsilon^{-6}d^5)$ can be vastly improved. With the important Latent Dirichlet Allocation (LDA) application in mind, we specialize our algorithm to sample from Dirichlet posteriors, and derive the first non-asymptotic $\tilde{O}(\epsilon^{-2}d^2)$ rate for first-order sampling. We further extend our framework to the mini-batch setting and prove convergence rates when only stochastic gradients are available. Finally, we report promising experimental results for LDA on real datasets.

1 Introduction

Many modern learning tasks involve sampling from a high-dimensional and large-scale distribution, which calls for algorithms that are scalable with respect to both the dimension and the data size. One approach [41] that has found wide success is to discretize the **Langevin Dynamics**:

$$\mathrm{d}\mathbf{X}_t = -\nabla V(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \tag{1.1}$$

where $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ presents a target distribution and $\mathbf{B}_t$ is a d-dimensional Brownian motion. Such a framework has inspired numerous first-order sampling algorithms [1, 12, 19, 21, 26, 27, 35, 38], and the convergence rates are by now well-understood for unconstrained and log-concave distributions [13, 17, 20].

However, applying (1.1) to sampling from *constrained* distributions (i.e., when V has a bounded convex domain) remains a difficult challenge. From the theoretical perspective, there are only two existing algorithms [6, 7] that possess non-asymptotic guarantees, and theif rates are significantly worse than the unconstrained scenaria under the same assumtions; *cf.*, Table 1. Furthermore, many important constrained distributions are inherently non-log-concave. A prominent instance is the **Dirichlet posterior**, which, in spite of the presence of several tailor-made first-order algorithms [26, 35], is still lacking a non-asymptotic guarantee.

In this paper, we aim to bridge these two gaps at the same time. For general constrained distributions with a strongly convex potential V, we prove the existence of a first-order algorithm that achieves the same convergence rates as if there is no constraint at all, suggesting the state-of-the-art $\tilde{O}(\epsilon^{-6}d^5)$ can be brought down to $\tilde{O}(\epsilon^{-2}d)$. When specialized to the important case of simplex constraint, we provide the first non-asymptotic guarantee for Dirichlet posteriors, $\tilde{O}(\epsilon^{-2}d^2R_0)$ for deterministic and $\tilde{O}\left(\epsilon^{-2}(Nd+\sigma^2)R_0\right)$ for the stochastic version of our algorithms; *cf.*, **Example 1** and **2** for the involved parameters.

Our framework combines ideas from the **Mirror Descent** [2, 33] algorithm for optimization and the theory of Optimal Transport [40]. Concretely, for constrained sampling problems, we propose to use the *mirror map* to transform the target into an unconstrained distribution, whereby many existing methods apply. Optimal Transport theory then comes in handy to relate the convergence rates between the original and transformed problems. For simplex constraints, we use the entropic mirror map to design practical first-order algorithms that possess rigorous guarantees, and are amenable to mini-batch extensions.

The rest of the paper is organized as follows. We briefly review the notion of push-forward measures in Section 2. In Section 3, we propose the **Mirrored Langevin Dynamics** and prove its convergence rates for constrained sampling problems. Mini-batch extensions are derived in Section 4. Finally, in Section 5, we provide synthetic and real-world experiments to demonstrate the empirical efficiency of our algorithms.

1.1 Related Work

First-Order Sampling Schemes with Langevin Dynamics: There exists a bulk of literature on (stochastic) first-order sampling schemes derived from Langevin Dynamics or its variants [1, 6, 7, 11, 13, 14, 17, 20, 22, 28, 35, 41]. However, to our knowledge, this work is the first to consider mirror descent extensions of the Langevin Dynamics.

The authors in [29] proposed a formalism that can, in principle, incorporate any variant of Langevin Dynamics for a given distribution $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$. The Mirrored Langevin Dynamics, however, is targeting the push-forward measure $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$ (see Section 3.1), and hence our framework is not covered in [29].

For Dirichlet posteriors, there is a similar variable transformation as our entropic mirror map in [35] (see the "reduced-natural parametrization" therein). The dynamics in [35] is nonetheless drastically different from ours, as there is a position-dependent matrix multiplying the Brownian motion, whereas our dynamics has no such feature; see (3.2).

Mirror Descent-Type Dynamics for Stochastic Optimization: Although there are some existing work on mirror descent-type dynamics for *stochastic optimization* [25, 32, 36, 42], we are unaware of any prior result on sampling.

Convergence Rates for Sampling from Dirichlet Posteriors: The work [15] proposed a zero$^{\text{th}}$ order method that achieves $\tilde{O}\left(T^{-1/2}\right)$ convergence in relative entropy for Dirichlet posteriors, which requires $O(dT^2)$ computation per iteration. Our method achieves the same rate with $O(d)$-complexity per iteration.

2 Preliminaries

2.1 Notation

In this paper, all Lipschitzness and strong convexity are with respect to the Euclidean norm $\|\cdot\|$. We use $\mathcal{C}^k$ to denote k-times differentiable functions with continuous k^{th} derivative. The Fenchel dual

[37] of a function h is denoted by $h^\star$. Given two mappings T, F of proper dimensions, we denote their composite map by $T \circ F$. For a probability measure μ, we write $\mathbf{X} \sim \mu$ to mean that "$\mathbf{X}$ is a random variable whose probability law is μ".

2.2 Push-Forward and Optimal Transport

Let $\mathrm{d}\mu = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ be a probability measure with support $\mathcal{X} \coloneqq \mathrm{dom}(V) = \{\mathbf{x} \in \mathbb{R}^d \mid V(\mathbf{x}) < +\infty\}$, and h be a convex function on $\mathcal{X}$. Throughout the paper we assume:

Assumption 1. *h is closed, proper, $h \in \mathcal{C}^2$, and $\nabla^2 h \succ 0$ on $\mathcal{X} \subset \mathbb{R}^d$.*

Assumption 2. *All measures have finite second moments.*

Assumption 3. *All measures vanish on sets with Hausdorff dimension [30] at most $d-1$.*

The gradient map ∇h induces a new probability measure $\mathrm{d}\nu \coloneqq e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$ through $\nu(E) = \mu\left(\nabla h^{-1}(E)\right)$ for every Borel set E on $\mathbb{R}^d$. We say that ν is the **push-forward measure** of μ under ∇h, and we denote it by $\nabla h \# \mu = \nu$. If $\mathbf{X} \sim \mu$ and $\mathbf{Y} \sim \nu$, we will sometimes abuse the notation by writing $\nabla h \# \mathbf{X} = \mathbf{Y}$ to mean $\nabla h \# \mu = \nu$.

If $\nabla h \# \mu = \nu$, the triplet (μ, ν, h) must satisfy the Monge-Ampère equation:

$$e^{-V} = e^{-W \circ \nabla h} \det \nabla^2 h. \tag{2.1}$$

Using $(\nabla h)^{-1} = \nabla h^\star$ and $\nabla^2 h \circ \nabla h^\star = \nabla^2 h^{\star -1}$, we see that (2.1) is equivalent to

$$e^{-W} = e^{-V \circ \nabla h^\star} \det \nabla^2 h^\star \tag{2.2}$$

which implies $\nabla h^\star \# \nu = \mu$.

The 2-Wasserstein distance between μ_1 and μ_2 is defined by[1]

$$\mathcal{W}_2^2(\mu_1, \mu_2) \coloneqq \inf_{T : T\#\mu_1 = \mu_2} \int \|\mathbf{x} - T(\mathbf{x})\|^2 \mathrm{d}\mu_1(\mathbf{x}). \tag{2.3}$$

3 Mirrored Langevin Dynamics

This section demonstrates a framework for transforming constrained sampling problems into unconstrained ones. We then focus on applications to sampling from strongly log-concave distributions and simplex-constrained distributions, even though the framework is more general and future-proof.

3.1 Motivation and Algorithm

We begin by briefly recalling the mirror descent (MD) algorithm for optimization. In order to minimize a function over a bounded domain, say $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$, MD uses a mirror map h to transform the primal variable $\mathbf{x}$ into the dual space $\mathbf{y} \coloneqq \nabla h(\mathbf{x})$, and then performs gradient updates in the dual: $\mathbf{y}^+ = \mathbf{y} - \beta \nabla f(\mathbf{x})$ for some step-size β. The mirror map h is chosen to adapt to the geometry of the constraint $\mathcal{X}$, which can often lead to faster convergence [33] or, more pivotal to this work, an **unconstrained** optimization problem [2].

Inspired by the MD framework, we would like to use the mirror map idea to remove the constraint for sampling problems. Toward this end, we first establish a simple fact [39]:

[1] In general, (2.3) is ill-defined; see [39]. The validity of (2.3) is guaranteed by McCann's theorem [31] under **Assumption 2** and **3**.

Theorem 1. *Let h satisfy **Assumption 1**. Suppose that $\mathbf{X} \sim \mu$ and $\mathbf{Y} = \nabla h(\mathbf{X})$. Then $\mathbf{Y} \sim \nu := \nabla h \# \mu$ and $\nabla h^\star(\mathbf{Y}) \sim \mu$.*

Proof. For any Borel set E, we have $\nu(E) = \mathbb{P}(\mathbf{Y} \in E) = \mathbb{P}\left(\mathbf{X} \in \nabla h^{-1}(E)\right) = \mu\left(\nabla h^{-1}(E)\right)$. Since ∇h is one-to-one, $\mathbf{Y} = \nabla h(\mathbf{X})$ if and only if $\mathbf{X} = \nabla h^{-1}(\mathbf{Y}) = \nabla h^\star(\mathbf{Y})$. □

In the context of sampling, **Theorem 1** suggests the following simple procedure: For any target distribution $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ with support $\mathcal{X}$, we choose a mirror map h on $\mathcal{X}$ satisfying **Assumption 1**, and we consider the **dual distribution** associated with $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ and h:

$$e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y} := \nabla h \# e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}. \tag{3.1}$$

Theorem 1 dictates that if we are able to draw a sample $\mathbf{Y}$ from $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$, then $\nabla h^\star(\mathbf{Y})$ immediately gives a sample for the desired distribution $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$. Furthermore, suppose for the moment that $\mathrm{dom}(h^\star) = \mathbb{R}^d$, so that $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$ is unconstrained. Then we can simply exploit the classical Langevin Dynamics (1.1) to efficiently take samples from $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$.

The above reasoning leads us to set up the **Mirrored Langevin Dynamics** (MLD):

$$\mathbf{MLD} \equiv \left\{ \begin{array}{l} \mathrm{d}\mathbf{Y}_t = -(\nabla W \circ \nabla h)(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \\ \mathbf{X}_t = \nabla h^\star(\mathbf{Y}_t) \end{array} \right. . \tag{3.2}$$

Notice that the stationary distribution of $\mathbf{Y}_t$ in MLD is $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$, since $\mathrm{d}\mathbf{Y}_t$ is nothing but the Langevin Dynamics (1.1) with $\nabla V \leftarrow \nabla W$. As a result, we have $\mathbf{X}_t \to \mathbf{X}_\infty \sim e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$.

Using (2.1), we can equivalently write the $\mathrm{d}\mathbf{Y}_t$ term in (3.2) as

$$\mathrm{d}\mathbf{Y}_t = -\nabla^2 h(\mathbf{X}_t)^{-1}\Big(\nabla V(\mathbf{X}_t) + \nabla \log\det \nabla^2 h(\mathbf{X}_t)\Big)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t.$$

In order to arrive at a practical algorithm, we then discretize the MLD, giving rise to the following equivalent iterations:

$$\mathbf{y}^{t+1} - \mathbf{y}^t = \left\{ \begin{array}{l} -\beta^t \nabla W(\mathbf{y}^t) + \sqrt{2\beta^t}\boldsymbol{\xi}^t \\ -\beta^t \nabla^2 h(\mathbf{x}^t)^{-1}\Big(\nabla V(\mathbf{x}^t) + \nabla \log\det \nabla^2 h(\mathbf{x}^t)\Big) + \sqrt{2\beta^t}\boldsymbol{\xi}^t \end{array} \right. \tag{3.3}$$

where in both cases $\mathbf{x}^{t+1} = \nabla h^\star(\mathbf{y}^{t+1})$, $\boldsymbol{\xi}^t$'s are i.i.d. standard Gaussian, and β^t's are step-sizes. The first formulation in (3.3) is useful when ∇W has a tractable form, while the second one can be computed using solely the information of V and h.

Next, we turn to the convergence of discretized MLD. Since $\mathrm{d}\mathbf{Y}_t$ in (3.2) is the classical Langevin Dynamics, and since we have assumed that W is unconstrained, it is typically not difficult to prove the convergence of $\mathbf{y}^t$ to $\mathbf{Y}_\infty \sim e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$. However, what we ultimately care about is the guarantee on the primal distribution $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$. The purpose of the next theorem is to fill the gap between primal and dual convergence.

We consider three most common metrics in evaluating approximate sampling schemes, namely the 2-Wasserstein distance $\mathcal{W}_2$, the total variation d_{TV}, and the relative entropy $D(\cdot\|\cdot)$.

Theorem 2 (Convergence in $\mathbf{y}^t$ implies convergence in $\mathbf{x}^t$)**.** *For any h satisfying **Assumption 1**, we have $d_{\mathrm{TV}}(\nabla h \# \mu_1, \nabla h \# \mu_2) = d_{\mathrm{TV}}(\mu_1, \mu_2)$ and $D(\nabla h \# \mu_1 \| \nabla h \# \mu_2) = D(\mu_1 \| \mu_2)$. In particular, we have $d_{\mathrm{TV}}(\mathbf{y}^t, \mathbf{Y}_\infty) = d_{\mathrm{TV}}(\mathbf{x}^t, \mathbf{X}_\infty)$ and $D(\mathbf{y}^t \| \mathbf{Y}_\infty) = D(\mathbf{x}^t \| \mathbf{X}_\infty)$ in (3.3).*

If, furthermore, h is ρ-strongly convex: $\nabla^2 h \succeq \rho I$. Then $\mathcal{W}_2(\mathbf{x}^t, \mathbf{X}_\infty) \leq \frac{1}{\rho}\mathcal{W}_2(\mathbf{y}^t, \mathbf{Y}_\infty)$.

Proof. See **Appendix A**. □

Assumption	$D(\cdot\|\cdot)$	$\mathcal{W}_2$	d_{TV}	Algorithm
$LI \succeq \nabla^2 V \succeq mI$	unknown	unknown	$\tilde{O}\left(\epsilon^{-6}d^5\right)$	MYULA [6]
$LI \succeq \nabla^2 V \succeq 0$	unknown	unknown	$\tilde{O}\left(\epsilon^{-12}d^{12}\right)$	PLMC [7]
$\nabla^2 V \succeq mI$	$\tilde{O}\left(\epsilon^{-1}d\right)$	$\tilde{O}\left(\epsilon^{-2}d\right)$	$\tilde{O}\left(\epsilon^{-2}d\right)$	MLD; this work
$LI \succeq \nabla^2 V \succeq mI$, V unconstrained	$\tilde{O}\left(\epsilon^{-1}d\right)$	$\tilde{O}\left(\epsilon^{-2}d\right)$	$\tilde{O}\left(\epsilon^{-2}d\right)$	Langevin Dynamics [13, 16, 20]

Table 1: Convergence rates for sampling from $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ with $\mathrm{dom}(V)$ bounded

3.2 Applications to Sampling from Constrained Distributions

We now consider applications of MLD. For strongly log-concave distributions with general constraint, we prove matching rates to that of unconstrained ones; see Section 3.2.1. In Section 3.2.2, we consider the important case where the constraint is a probability simplex.

3.2.1 Sampling from a strongly log-concave distribution with constraint

As alluded to in the introduction, the existing convergence rates for constrained distributions are significantly worse than their unconstrained counterparts; see Table 1 for a comparison. The main result of this subsection is the existence of a "good" mirror map for *arbitrary* constraint, with which the dual distribution $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$ becomes unconstrained:

Theorem 3 (Existence of a good mirror map for MLD)**.** *Let $\mathrm{d}\mu(\mathbf{x}) = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ be a probability measure with bounded convex support such that $V \in \mathcal{C}^2$, $\nabla^2 V \succeq mI \succ 0$, and V is bounded away from $+\infty$ in the interior of the support. Then there exists a mirror map $h \in \mathcal{C}^2$ such that the discretized MLD* (3.3) *yields*

$$D\left(\mathbf{x}^T\|\mathbf{X}_\infty\right) = \tilde{O}\left(\frac{d}{T}\right), \quad \mathcal{W}_2\left(\mathbf{x}^T, \mathbf{X}_\infty\right) = \tilde{O}\left(\sqrt{\frac{d}{T}}\right), \quad d_{\mathrm{TV}}\left(\mathbf{x}^T, \mathbf{X}_\infty\right) = \tilde{O}\left(\sqrt{\frac{d}{T}}\right).$$

Proof. See **Appendix B**. □

Remark 1. *We remark that **Theorem 3** is only an existential result, not an actual algorithm. Practical algorithms are considered in the next subsection.*

3.2.2 Sampling Algorithms on Simplex

We apply the discretized MLD (3.3) to the task of sampling from distributions on the probability simplex $\Delta_d \coloneqq \{\mathbf{x} \in \mathbb{R}^d \mid \sum_{i=1}^d x_i \leq 1, x_i \geq 0\}$, which is instrumental in many fields of machine learning and statistics.

On a simplex, the most natural choice of h is the entropic mirror map [2], which is well-known to be 1-strongly convex:

$$h(\mathbf{x}) = \sum_{\ell=1}^d x_i \log x_\ell + \left(1 - \sum_{\ell=1}^d x_\ell\right) \log \left(1 - \sum_{\ell=1}^d x_\ell\right), \text{ where } \ 0 \log 0 \coloneqq 0. \tag{3.4}$$

In this case, the associated dual distribution can be computed explicitly.

Lemma 1 (Sampling on a simplex with entropic mirror map)**.** *Let $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ be the target distribution on Δ_d, h be the entropic mirror map (3.4), and $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y} \coloneqq \nabla h \# e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$. Then the potential W of the push-forward measure admits the expression*

$$W(\mathbf{y}) = V \circ \nabla h^\star(\mathbf{y}) - \sum_{\ell=1}^{d} y_\ell + (d+1)h^\star(\mathbf{y}) \tag{3.5}$$

where $h^\star(\mathbf{y}) = \log\left(1 + \sum_{\ell=1}^{d} e^{y_\ell}\right)$ is the Fenchel dual of h, which is strictly convex and 1-Lipschitz gradient.

Proof. See **Appendix C**. □

Crucially, we have $\mathrm{dom}(h^\star) = \mathbb{R}^d$, so that the Langevin Dynamics for $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$ is **unconstrained**.

Based on **Lemma 1**, we now present the surprising case of the *non-log-concave* Dirichlet posteriors, a distribution of central importance in topic modeling [3], for which the dual distribution $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}$ becomes strictly *log-concave.*

Example 1 (Dirichlet Posteriors)**.** Given parameters $\alpha_1, \alpha_2, ..., \alpha_{d+1} > 0$ and observations $n_1, n_2, ..., n_{d+1}$ where n_ℓ is the number of appearance of category ℓ, the probability density function of the Dirichlet posterior is

$$p(\mathbf{x}) = \frac{1}{C}\prod_{\ell=1}^{d+1} x_\ell^{n_\ell+\alpha_\ell-1}, \quad \mathbf{x} \in \mathrm{int}\left(\Delta_d\right) \tag{3.6}$$

where C is a normalizing constant and $x_{d+1} \coloneqq 1 - \sum_{\ell=1}^{d} x_\ell$. The corresponding V is

$$V(\mathbf{x}) = -\log p(\mathbf{x}) = \log C - \sum_{\ell=1}^{d+1}(n_\ell + \alpha_\ell - 1)\log x_\ell, \quad \mathbf{x} \in \mathrm{int}\left(\Delta_d\right).$$

The interesting regime of the Dirichlet posterior is when it is **sparse**, meaning the majority of the n_ℓ's are zero and a few n_k's are large, say of order $O(d)$. It is also common to set $\alpha_\ell < 1$ for all ℓ in practice. Evidently, V is neither convex nor concave in this case, and no existing non-asymptotic rate can be applied. However, plugging V into (3.5) gives

$$W(\mathbf{y}) = \log C - \sum_{\ell=1}^{d}(n_\ell + \alpha_\ell)y_\ell + \left(\sum_{\ell=1}^{d+1}(n_\ell + \alpha_\ell)\right)h^\star(\mathbf{y}) \tag{3.7}$$

which, magically, becomes strictly convex and $O(d)$-Lipschitz gradient **no matter what the observations and parameters are!** In view of **Theorem 2** and **Corollary 7** of [20], one can then apply (3.3) to obtain an $\tilde{O}\left(\epsilon^{-2}d^2R_0\right)$ convergence in relative entropy, where $R_0 \coloneqq \mathcal{W}_2^2(\mathbf{y}^0, e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y})$ is the initial Wasserstein distance to the target. □

4 Stochastic Mirrored Langevin Dynamics

We have thus far only considered **deterministic** methods based on exact gradients. In practice, however, evaluating gradients typically involves one pass over the full data, which can be time-consuming in large-scale applications. In this section, we turn attention to the **mini-batch** setting, where one can use a small subset of data to form stochastic gradients.

Algorithm 1 Stochastic Mirrored Langevin Dynamics (SMLD)

```
Require: Target distribution $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ where $V = \sum_{i=1}^N V_i$, step-sizes $\beta^t$, batch-size $b$
Find $W_i$ such that $e^{-NW_i} \propto \nabla h \# e^{-NV_i}$ for all $i$.
for $t \leftarrow 0, 1, \cdots, T-1$ do
    Pick a mini-batch $B$ of size $b$ uniformly at random.
    Update $\mathbf{y}^{t+1} = \mathbf{y}^t - \frac{\beta^t N}{b} \sum_{i \in B} \nabla W_i(\mathbf{y}^t) + \sqrt{2\beta^t}\boldsymbol{\xi}^t$
    $\mathbf{x}^{t+1} = \nabla h^\star(\mathbf{y}^{t+1})$        ▷ Update only when necessary.
end for
return $\mathbf{x}^T$
```

Toward this end, we assume:

Assumption 4 (Primal Decomposibility)**.** *The target distribution $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ admits a decomposable structure $V = \sum_{i=1}^N V_i$ for some functions V_i.*

The above assumption is often met in machine learning applications, where each V_i represents one data. If there is an additional prior term (that is, $V = \sum_{i=1}^N V_i + U$ for some U), then one can redefine $V_i' := V_i + \frac{1}{N}U$ so that **Assumption 4** still holds.

Consider the following common scheme in obtaining stochastic gradients. Given a batch-size b, we randomly pick a mini-batch B from $\{1, 2, \ldots, N\}$ with $|B| = b$, and form an unbiased estimate of ∇V by computing

$$\tilde{\nabla} V := \frac{N}{b} \sum_{i \in B} \nabla V_i. \tag{4.1}$$

The following lemma asserts that exactly the same procedure can be carried out in the dual.

Lemma 2. *Assume that h is 1-strongly convex. For $i = 1, 2, ..., N$, let W_i be such that*

$$e^{-NW_i} = \nabla h \# \frac{e^{-NV_i}}{\int e^{-NV_i}}. \tag{4.2}$$

Define $W := \sum_{i=1}^N W_i$ and $\tilde{\nabla} W := \frac{N}{b} \sum_{i \in B} \nabla W_i$, where B is chosen as in (4.1). Then:

1. *Primal decomposibility implies dual decomposability: There is a constant C such that $e^{-(W+C)} = \nabla h \# e^{-V}$.*
2. *For each i, the gradient ∇W_i depends only on ∇V_i and the mirror map h.*
3. *The gradient estimate is unbiased: $\mathbb{E}\tilde{\nabla} W = \nabla W$.*
4. *The dual stochastic gradient is more accurate: $\mathbb{E}\|\tilde{\nabla} W - \nabla W\|^2 \leq \mathbb{E}\|\tilde{\nabla} V - \nabla V\|^2$.*

Proof. See **Appendix D**. □

Lemma 2 furnishes a template for the mini-batch extension of MLD. The pseudocode is detailed in **Algorithm 1**, whose convergence rate is given by the next theorem.

Theorem 4. *Let $e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ be a distribution satisfying **Assumption 4**, and h a 1-strongly convex mirror map. Let $\sigma^2 := \mathbb{E}\|\tilde{\nabla} V - \nabla V\|^2$ be the variance of the stochastic gradient of V in (4.1).*

Suppose that the corresponding dual distribution $e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y} = \nabla h \# e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ *satisfies* $LI \succeq \nabla^2 W \succeq 0$*. Then, applying SMLD with constant step-size* $\beta^t = \beta$ *yields*[2]*:*

$$D\left(\mathbf{x}^T \| e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}\right) \leq \sqrt{\frac{2\mathcal{W}_2^2\left(\mathbf{y}^0, e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}\right)\left(Ld+\sigma^2\right)}{T}} = O\left(\sqrt{\frac{Ld+\sigma^2}{T}}\right), \tag{4.3}$$

provided that $\beta \leq \min\left\{\left[2T\mathcal{W}_2^2\left(\mathbf{y}^0, e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}\right)\left(Ld+\sigma^2\right)\right]^{-\frac{1}{2}}, \frac{1}{L}\right\}$*.*

Proof. See **Appendix E**. □

Example 2 (SMLD for Dirichlet Posteriors)**.** For the case of Dirichlet posteriors, we have seen in (3.7) that the corresponding dual distribution satisfies $(N+\Gamma)I \succeq \nabla^2 W \succ 0$, where $N \coloneqq \sum_{\ell=1}^{d+1} n_\ell$ and $\Gamma \coloneqq \sum_{\ell=1}^{d+1} \alpha_\ell$. Furthermore, it is easy to see that the stochastic gradient $\tilde{\nabla} W$ can be efficiently computed (see **Appendix F**):

$$\tilde{\nabla} W(\mathbf{y})_\ell \coloneqq \frac{N}{b}\sum_{i\in B} \nabla W_i(\mathbf{y})_\ell = -\left(\frac{Nm_\ell}{b} + \alpha_\ell\right) + (N+\Gamma)\frac{e^{y_\ell}}{1+\sum_{k=1}^d e^{y_k}}, \tag{4.4}$$

where m_ℓ is the number of observations of category ℓ in the mini-batch B. As a result, **Theorem 4** states that SMLD achieves

$$D\left(\mathbf{x}^T \| e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}\right) \leq \sqrt{\frac{2\mathcal{W}_2^2\left(\mathbf{y}^0, e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}\right)\left((N+\Gamma)(d+1)+\sigma^2\right)}{T}} = O\left(\sqrt{\frac{(N+\Gamma)d+\sigma^2}{T}}\right)$$

with a constant step-size. □

5 Experiments

We conduct experiments with a two-fold purpose. First, we use a low-dimensional synthetic data, where we can evaluate the total variation error by comparing histograms, to verify the convergence rates in our theory. Second, We demonstrate that the SMLD, modulo a necessary modification for resolving numerical issues, outperforms state-of-the-art first-order methods on the Latent Dirichlet Allocation (LDA) application with Wikipedia corpus.

5.1 Synthetic Experiment for Dirichlet Posterior

We implement the deterministic MLD for sampling from an 11-dimensional Dirichlet posterior (3.6) with $n_1 = 10{,}000$, $n_2 = n_3 = 10$, and $n_4 = n_5 = \cdots = n_{11} = 0$, which aims to capture the sparse nature of real observations in topic modeling. We set $\alpha_\ell = 0.1$ for all ℓ.

As a baseline comparison, we include the Stochastic Gradient Riemannian Langevin Dynamics (SGRLD) [35] with the expanded-mean parametrization. SGRLD is a tailor-made first-order scheme

[2] Our guarantee is given on a randomly chosen iterate from $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^T\}$, instead of the final iterate $\mathbf{x}^T$. In practice, we observe that the final iterate always gives the best performance, and we will ignore this minor difference in the theorem statement.

for simplex constraints, and it remains one of the state-of-the-art algorithms for LDA. For fair comparison, we use deterministic gradients for SGRLD.

We perform a grid search over the constant step-size for both algorithms, and we keep the best three for MLD and the best five for SGRLD. For each iteration, we build an empirical distribution by running 2,000,000 independent trials, and we compute its total variation with respect to the histogram generated by the true distribution.

Figure 1(a) reports the total variation error along the first dimension, where we can see that MLD outperforms SGRLD by a substantial margin. As dictated by our theory, all the MLD curves decay at the $O(T^{-1/2})$ rate until they saturate at the dicretization error level. In contrast, SGRLD lacks non-asymptotic guarantees, and there is no clear convergence rate we can infer from Figure 1(a).

5.2 Latent Dirichlet Allocation with Wikipedia Corpus

An influential framework for topic modeling is the Latent Dirichlet Allocation (LDA) [3], which, given a text collection, requires to infer the posterior word distributions without knowing the exact topic for each word. The full model description is standard but somewhat convoluted; we refer to the classic [3] for details.

Each topic k in LDA determines a word distribution $\boldsymbol{\pi}_k$, and suppose there are in total K topics and $W+1$ words. The variable of interest is therefore $\boldsymbol{\pi} := (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, ..., \boldsymbol{\pi}_K) \in \Delta_W \times \Delta_W \times \cdots \Delta_W$. Since this domain is a Cartesian product of simplices, we propose to use $\tilde{h}(\boldsymbol{\pi}) := \sum_{k=1}^{K} h(\boldsymbol{\pi}_k)$, where h is the entropic mirror map (3.4), for SMLD. It is easy to see that all of our computations for Dirichlet posteriors generalize to this setting.

5.2.1 Experimental Setup

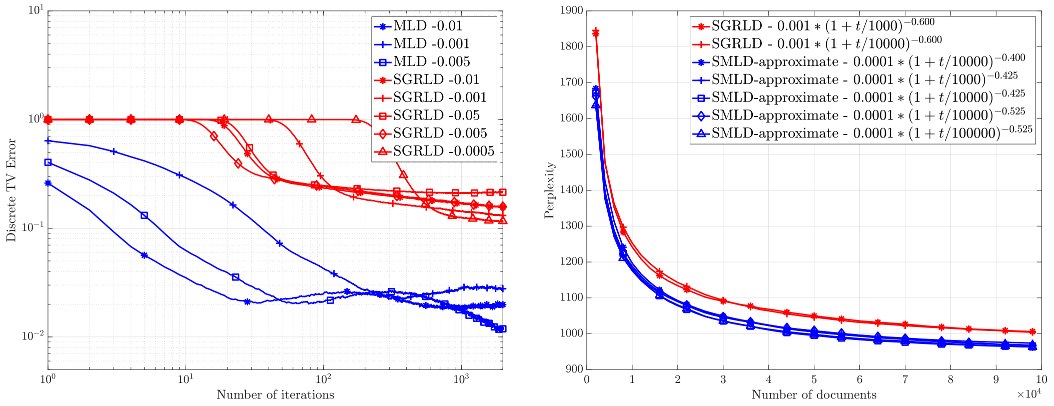
We implement the SMLD for LDA on the Wikipedia corpus with 100,000 documents, and we compare the performance against the SGRLD [35]. In order to keep the comparison fair, we adopt exactly the same setting as in [35], including the model parameters, the batch-size, the Gibbs sampler steps, etc. See Section 4 and 5 in [35] for omitted details.

Another state-of-the-art first-order algorithm for LDA is the SGRHMC in [29], for which we skip the implementation, due to not knowing how the $\hat{B}_t$ was chosen in [29]. Instead, we will repeat the same experimental setting as [29] and directly compare our results versus the ones reported in [29]. See **Appendix G** for comparison against SGRHMC.

5.2.2 A Numerical Trick and the SMLD-approximate Algorithm

A major drawback of the SMLD in practice is that the stochastic gradients (4.4) involve exponential functions, which are unstable for large-scale problems. For instance, in python, `np.exp(800) = inf`, whereas the relevant variable regime in this experiment extends to 1600. To resolve such numerical issues, we appeal to the linear approximation[3] $\exp(\mathbf{y}) \simeq \max\{0, 1+\mathbf{y}\}$. Admittedly, our theory no longer holds under such numerical tricks, and we shall not claim that our algorithm is provably convergent for LDA. Instead, the contribution of MLD here is to identify the dual dynamics associated with (3.7), which would have been otherwise difficult to perceive. We name the resulting algorithm "SMLD-approximate" to indicate its heuristic nature.

[3] One can also use a higher-order Taylor approximation for $\exp(\mathbf{y})$, or add a small threshold $\exp(\mathbf{y}) \simeq \max\{\epsilon, 1+\mathbf{y}\}$ to prevent the iterates from going to the boundary. In practice, we observe that these variants do not make a huge impact on the performance.



(a) Synthetic data.

(b) LDA on Wikipedia corpus.

5.2.3 Results

Figure 1(b) reports the perplexity on the test data up to 100,000 documents, with the five best step-sizes we found via grid search for SMLD-approximate. For SGRLD, we use the best step-sizes reported in [35].

From the figure, we can see a clear improvement, both in terms of convergence speed and the saturation level, of the SMLD-approximate over SGRLD. One plausible explanation for such phenomenon is that our MLD, as a simple unconstrained Langevin Dynamics, is less sensitive to discretization. On the other hand, the underlying dynamics for SGRLD is a more sophisticated Riemannian diffusion, which requires finer discretization than MLD to achieve the same level of approximation to the original continuous-time dynamics, and this is true even in the presence of noisy gradients and our numerical heuristics

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^o 725594 - time-data).

References

[1] Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1771–1778, 2012.

[2] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math*, 305(19):805–808, 1987.

[5] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[6] Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 319–342. PMLR, 07–10 Jul 2017.

[7] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *arXiv preprint arXiv:1507.02564*, 2015.

[8] Luis A Caffarelli. A localization property of viscosity solutions to the monge-ampere equation and their strict convexity. *Annals of Mathematics*, 131(1):129–134, 1990.

[9] Luis A Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[10] Luis A Caffarelli. Monotonicity properties of optimal transportation and the fkg and related inequalities. *Communications in Mathematical Physics*, 214(3):547–563, 2000.

[11] Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pages 2278–2286, 2015.

[12] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

[13] Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR, 07–09 Apr 2018.

[14] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

[15] Bo Dai, Niao He, Hanjun Dai, and Le Song. Provable bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994, 2016.

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

[17] Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

[18] Guido De Philippis and Alessio Figalli. The monge–ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

[19] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in neural information processing systems*, pages 3203–3211, 2014.

[20] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *arXiv preprint arXiv:1802.09188*, 2018.

[21] Alain Durmus, Umut Simsekli, Eric Moulines, Roland Badeau, and Gaël Richard. Stochastic gradient richardson-romberg markov chain monte carlo. In *Advances in Neural Information Processing Systems*, pages 2047–2055, 2016.

[22] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! *arXiv preprint arXiv:1801.02309*, 2018.

[23] Bela A Frigyik, Amol Kapila, and Maya R Gupta. Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*, 2010.

[24] Alexander V Kolesnikov. Mass transportation and contractions. *arXiv preprint arXiv:1103.1479*, 2011.

[25] Walid Krichene and Peter L Bartlett. Acceleration and averaging in stochastic descent dynamics. In *Advances in Neural Information Processing Systems*, pages 6799–6809, 2017.

[26] Shiwei Lan and Babak Shahbaba. Sampling constrained probability distributions using spherical augmentation. In *Algorithmic Advances in Riemannian Geometry and Applications*, pages 25–71. Springer, 2016.

[27] Chang Liu, Jun Zhu, and Yang Song. Stochastic gradient geodesic mcmc methods. In *Advances in Neural Information Processing Systems*, pages 3009–3017, 2016.

[28] Tung Luu, Jalal Fadili, and Christophe Chesneau. Sampling from non-smooth distribution through langevin diffusion. 2017.

[29] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.

[30] Benoit B Mandelbrot. *The fractal geometry of nature*, volume 173. WH freeman New York, 1983.

[31] Robert J McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–324, 1995.

[32] Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, 2018.

[33] AS Nemirovsky and DB Yudin. Problem complexity and method efficiency in optimization. 1983.

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

[35] Sam Patterson and Yee Whye Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.

[36] Maxim Raginsky and Jake Bouvrie. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 6793–6800. IEEE, 2012.

[37] Ralph Tyrell Rockafellar. *Convex analysis.* Princeton university press, 1970.

[38] Umut Simsekli, Roland Badeau, Taylan Cemgil, and Gaël Richard. Stochastic quasi-newton langevin monte carlo. In *International Conference on Machine Learning*, pages 642–651, 2016.

[39] Cédric Villani. *Topics in optimal transportation.* Number 58. American Mathematical Soc., 2003.

[40] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[41] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

[42] Pan Xu, Tianhao Wang, and Quanquan Gu. Accelerated stochastic mirror descent: From continuous-time dynamics to discrete-time algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1087–1096, 2018.

A Proof of Theorem 2

We first focus on the convergence for total variation and relative entropy, since they are in fact quite trivial. The proof for the 2-Wasserstein distance requires a bit more work.

A.1 Total Variation and Relative Entropy

Since h is strictly convex, ∇h is one-to-one, and hence

$$\begin{aligned} d_{\mathrm{TV}}(\nabla h\#\mu_1, \nabla h\#\mu_2) &= \frac{1}{2}\sup_E |\nabla h\#\mu_1(E) - \nabla h\#\mu_2(E)| \\ &= \frac{1}{2}\sup_E \left|\mu_1\big(\nabla h^{-1}(E)\big) - \mu_2\big(\nabla h^{-1}(E)\big)\right| \\ &= d_{\mathrm{TV}}(\mu_1, \mu_2). \end{aligned}$$

On the other hand, it is well-known that applying a one-to-one mapping to distributions leaves the relative entropy intact. Alternatively, we may also simply write (letting $\nu_i = \nabla h\#\mu_i$):

$$\begin{aligned} D(\nu_1\|\nu_2) &= \int \log\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_2}\mathrm{d}\nu_1 \\ &= \int \log\left(\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_2}\circ\nabla h\right)\mathrm{d}\mu_1 && \text{by (A.5) below} \\ &= \int \log\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\mathrm{d}\mu_1 && \text{by (2.1)} \\ &= D(\mu_1\|\mu_2) \end{aligned}$$

The “in particular” part follows from noticing that $\mathbf{y}^t \sim \nabla h\#\mathbf{x}^t$ and $\mathbf{Y}_\infty \sim \nabla h\#\mathbf{X}_\infty$.

A.2 2-Wasserstein Distance

Now, let h be ρ-strongly convex. The most important ingredient of the proof is **Lemma 3** below, which is conceptually clean. Unfortunately, for the sake of rigor, we must deal with certain intricate regularity issues in the Optimal Transport theory. If the reader wishes, she/he can simply assume that the quantities (A.1) and (A.2) below are well-defined, which is always satisfied by any practical mirror map, and skip all the technical part about the well-definedness proof.

For the moment, assume $h \in \mathcal{C}^5$; the general case is given at the end. Every convex h generates a Bregman divergence via $B_h(\mathbf{x}, \mathbf{x}') \coloneqq h(\mathbf{x}) - h(\mathbf{x}') - \langle \nabla h(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle$. The following key lemma allows us to relate guarantees in $\mathcal{W}_2$ between $\mathbf{x}^t$'s and $\mathbf{y}^t$'s. It can be seen as a generalization of the classical duality relation (A.4) in the space of probability measures.

Lemma 3 (Duality of Wasserstein Distances)**.** *Let μ_1, μ_2 be probability measures satisfying **Assumptions 2** and **3**. If h is ρ-strongly convex and $\mathcal{C}^5$, then the* (A.1) *and* (A.2) *below are well-defined:*

$$\mathcal{W}_{B_h}(\mu_1, \mu_2) \coloneqq \inf_{T : T\#\mu_1 = \mu_2} \int B_h\left(\mathbf{x}, T(\mathbf{x})\right) \mathrm{d}\mu_1(\mathbf{x}) \tag{A.1}$$

and (notice the exchange of inputs on the right-hand side)

$$\mathcal{W}_{B_{h^\star}}(\nu_1, \nu_2) \coloneqq \inf_{T : T\#\nu_1 = \nu_2} \int B_{h^\star}\left(T(\mathbf{y}), \mathbf{y}\right) \mathrm{d}\nu_1(\mathbf{y}). \tag{A.2}$$

Furthermore, we have

$$\mathcal{W}_{B_h}(\mu_1, \mu_2) = \mathcal{W}_{B_{h^\star}}(\nabla h \# \mu_1, \nabla h \# \mu_2). \tag{A.3}$$

Before proving the lemma, let us see that the relation in $\mathcal{W}_2$ is a simple corollary of **Lemma 3**. Since h is ρ-strongly convex, it is classical that, for any $\mathbf{x}$ and $\mathbf{x}'$,

$$\frac{\rho}{2}\|\mathbf{x} - \mathbf{x}'\|^2 \leq B_h(\mathbf{x}, \mathbf{x}') = B_{h^\star}(\nabla h(\mathbf{x}'), \nabla h(\mathbf{x})) \leq \frac{1}{2\rho}\|\nabla h(\mathbf{x}) - \nabla h(\mathbf{x}')\|^2. \tag{A.4}$$

Using **Lemma 3** and the fact that $\mathbf{y}^t \sim \nabla h \# \mathbf{x}^t$ and $\mathbf{Y}_\infty \sim \nabla h \# \mathbf{X}_\infty$, we conclude $\mathcal{W}_2(\mathbf{x}^t, \mathbf{X}_\infty) \leq \frac{1}{\rho}\mathcal{W}_2(\mathbf{y}^t, \mathbf{X}_\infty)$. It hence remains to prove **Lemma 3** when $h \in \mathcal{C}^5$.

A.2.1 Proof of Lemma 3 When $h \in \mathcal{C}^5$

We first prove that (A.2) is well-defined by verifying the sufficient conditions in **Theorem 3.6** of [18]. Specifically, we will verify **(C0)-(C2)** in p.554 of [18] when the transport cost is $B_{h^\star}$.

Since h is ρ-strongly convex, ∇h is injective, and hence $\nabla h^\star = (\nabla h)^{-1}$ is also injective, which implies that $h^\star$ is strictly convex. On the other hand, the strong convexity of h implies $\nabla^2 h^\star \preceq \frac{1}{\rho} I$, and hence $B_{h^\star}$ is globally upper bounded by a quadratic function.

We now show that the conditions **(C0)-(C2)** are satisfied. Since we have assumed $h \in \mathcal{C}^5$, we have $B_{h^\star} \in \mathcal{C}^4$. Since $B_{h^\star}$ is upper bounded by a quadratic function, the condition **(C0)** is trivially satisfied. On the other hand, since $h^\star$ is strictly convex, simple calculation reveals that, for any $\mathbf{y}'$, the mapping $\mathbf{y} \to \nabla_{\mathbf{y}'} B_{h^\star}(\mathbf{y}, \mathbf{y}')$ is injective, which is **(C1)**. Similarly, for any $\mathbf{y}$, the mapping $\mathbf{y}' \to \nabla_{\mathbf{y}} B_{h^\star}(\mathbf{y}, \mathbf{y}')$ is also injective, which is **(C2)**. By **Theorem 3.6** in [18], (A.2) is well-defined.

We now turn to (A.3), which will automatically establish the well-definedness of (A.1). We first need the following equivalent characterization of $\nabla h\#\mu = \nu$ [40]:

$$\int f \mathrm{d}\nu = \int f \circ \nabla h \mathrm{d}\mu \tag{A.5}$$

for all measurable f. Using (A.5) in the definition of $\mathcal{W}_{B_{h^\star}}$, we get

$$\begin{aligned} \mathcal{W}_{B_{h^\star}}(\nabla h\#\mu_1, \nabla h\#\mu_2) &= \inf_T \int B_{h^\star}\left(T(\mathbf{y}), \mathbf{y}\right) \mathrm{d}\nabla h\#\mu_1(\mathbf{y}) \\ &= \inf_T \int B_{h^\star}\Big((T \circ \nabla h)(\mathbf{x}), \nabla h(\mathbf{x})\Big) \mathrm{d}\mu_1(\mathbf{x}), \end{aligned}$$

where the infimum is over all T such that $T\#(\nabla h\#\mu_1) = \nabla h\#\mu_2$. Using the classical duality $B_h(\mathbf{x}, \mathbf{x}') = B_{h^\star}(\nabla h(\mathbf{x}'), \nabla h(\mathbf{x}))$ and $\nabla h \circ \nabla h^\star(\mathbf{x}) = \mathbf{x}$, we may further write

$$\mathcal{W}_{B_{h^\star}}(\nabla h\#\mu_1, \nabla h\#\mu_2) = \inf_T \int B_h\Big(\mathbf{x}, (\nabla h^\star \circ T \circ \nabla h)(\mathbf{x})\Big) \mathrm{d}\mu_1(\mathbf{x}) \tag{A.6}$$

where the infimum is again over all T such that $T\#(\nabla h\#\mu_1) = \nabla h\#\mu_2$. In view of (A.6), the proof would be complete if we can show that $T\#(\nabla h\#\mu_1) = \nabla h\#\mu_2$ if and only if $(\nabla h^\star \circ T \circ \nabla h)\#\mu_1 = \mu_2$.

For any two maps T_1 and T_2, we claim that

$$(T_1 \circ T_2)\#\mu = T_1\#\left(T_2\#\mu\right). \tag{A.7}$$

Indeed, for any Borel set E, we have, by definition of the push-forward,

$$\begin{aligned} (T_1 \circ T_2)\#\mu(E) &= \mu\big((T_1 \circ T_2)^{-1}(E)\big) \\ &= \mu\big((T_2^{-1} \circ T_1^{-1})(E)\big). \end{aligned}$$

On the other hand, recursively applying the definition of push-forward to $T_1\#\left(T_2\#\mu\right)$ gives

$$\begin{aligned} T_1\#\left(T_2\#\mu\right)(E) &= T_2\#\mu\big(T^{-1}(E)\big) \\ &= \mu\big((T_2^{-1} \circ T_1^{-1})(E)\big) \end{aligned}$$

which establishes (A.7).

Assume that $T\#(\nabla h\#\mu_1) = \nabla h\#\mu_2$. Then we have

$$\begin{aligned} (\nabla h^\star \circ T \circ \nabla h)\#\mu_1 &= \nabla h^\star\#(T\#(\nabla h\#\mu_1)) && \text{by (A.7)} \\ &= \nabla h^\star\#(\nabla h\#\mu_2) && \text{since } T\#(\nabla h\#\mu_1) = \nabla h\#\mu_2 \\ &= (\nabla h^\star \circ \nabla h)\#\mu_2 && \text{by (A.7) again} \\ &= \mu_2. \end{aligned}$$

On the other hand, if $(\nabla h^\star \circ T \circ \nabla h)\#\mu_1 = \mu_2$, then composing both sides by ∇h and using (A.7) yields $T\#(\nabla h\#\mu_1) = \nabla h\#\mu_2$, which finishes the proof.

A.2.2 When h is only $\mathcal{C}^2$

When h is only $\mathcal{C}^2$, we will directly resort to (A.4). Let T be any map such that $T\#(\nabla h\#\mu_1) = \nabla h\#\mu_2$, and consider the optimal transportation problem $\inf_T \int \|\mathbf{y} - T(\mathbf{y})\|^2 \mathrm{d}\nabla h\#\mu_1(\mathbf{y})$. By (A.4) and (A.5), we have

$$\begin{aligned}\inf_T \int \|\mathbf{y} - T(\mathbf{y})\|^2 \mathrm{d}\nabla h\#\mu_1(\mathbf{y}) &= \inf_T \int \|\nabla h(\mathbf{x}) - (T \circ \nabla h)(\mathbf{x}))\|^2 \mathrm{d}\mu_1(\mathbf{x}) \\ &\geq \rho^2 \inf_T \int \|\mathbf{x} - (\nabla h^\star \circ T \circ \nabla h)(\mathbf{x}))\|^2 \mathrm{d}\mu_1(\mathbf{x})\end{aligned}$$

where the infimum is over all T such that $T\#(\nabla h\#\mu_1) = \nabla h\#\mu_2$. But as proven in **Appendix A.2.1**, this is equivalent to $(\nabla h^\star \circ T \circ \nabla h)\#\mu_1 = \mu_2$. The proof is finished by noting $\mathbf{y}^t \sim \nabla h\#\mathbf{x}^t$ and $\mathbf{Y}_\infty \sim \nabla h\#\mathbf{X}_\infty$.

B Proof of Thereom 3

In previous sections, we are given a target distribution e^{-V} and a mirror map h, and we derive the induced distribution e^{-W} through the Monge-Ampère equation (2.1). The high-level idea of this proof is to reverse the direction: We start with two good distributions e^{-V} and e^{-W}, and we invoke deep results in Optimal Transport to deduce the existence of a good mirror map h.

First, notice that if V has bounded domain, then the strong convexity of V implies $V > -\infty$. Along with the assumption that V is bounded away from $+\infty$ in the interior, we see that e^{-V} is bounded away from 0 and $+\infty$ in the interior of support.

Let $\mathrm{d}\nu(\mathbf{x}) \propto e^{-\frac{\|\mathbf{x}\|^2}{2}}\mathrm{d}\mathbf{x}$ be the standard d-dimensional Gaussian measure. By Brenier's polarization theorem [4, 5] and **Assumption 2**, **3**, there exists a convex function $h^\star$ whose gradient solves the $\mathcal{W}_2\left(e^{-\frac{\|\mathbf{x}\|^2}{2}}\mathrm{d}\mathbf{x}, \mu\right)$ optimal transportation problem. Caffarelli's regularity theorem [8, 9, 10] then implies that the Brenier's map $h^\star$ is in $\mathcal{C}^2$. Finally, a slightly stronger form of Caffarelli's contraction theorem [24] asserts:

$$\nabla^2 h^\star \preceq \frac{1}{m} I, \tag{B.1}$$

which implies $h = (h^\star)^\star$ is m-strongly convex.

Let us consider the discretized MLD (3.3) corresponding to the mirror map h. The SDE governing $\mathbf{Y}_t$ is simply the Ornstein-Uhlenbeck Process [34]:

$$\mathrm{d}\mathbf{Y}_t = -\mathbf{Y}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t. \tag{B.2}$$

Invoking **Theorem 3** of [13], for each iteration $\mathbf{y}^T$ from (3.3) applied to (B.2), we have $D(\mathbf{y}^T \| \mathbf{Y}_\infty) = \tilde{O}\left(\frac{d}{T}\right)$, which in turn implies $\mathcal{W}_2(\mathbf{y}^T, \mathbf{Y}_\infty) = \tilde{O}\left(\sqrt{\frac{d}{T}}\right)$ and $d_{\mathrm{TV}}(\mathbf{y}^T, \mathbf{Y}_\infty) = \tilde{O}\left(\sqrt{\frac{d}{T}}\right)$. **Theorem 2** then completes the proof.

C Proof of Lemma 1

Straightforward calculations in convex analysis shows

$$\frac{\partial h}{\partial x_i} = \log \frac{x_i}{x_{d+1}}, \qquad \frac{\partial^2 h}{\partial x_i \partial x_j} = \delta_{ij} x_i^{-1} + x_{d+1}^{-1},$$
$$h^\star(\mathbf{y}) = \log \left(1 + \sum_{i=1}^d e^{y_i}\right), \quad \frac{\partial h^\star}{\partial y_i} = \frac{e^{y_i}}{1 + \sum_{i=1}^d e^{y_i}}, \tag{C.1}$$

which proves that h is 1-strongly convex.

Let $\mu = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ be the target distribution and define $\nu = e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y} := \nabla h \# \mu$. By (2.1), we have

$$W \circ \nabla h = V + \log\det \nabla^2 h. \tag{C.2}$$

Since $\nabla^2 h(\mathbf{x}) = \mathrm{diag}[x_i^{-1}] + x_{d+1}^{-1} \mathbb{1}\mathbb{1}^\top$ where $\mathbb{1}$ is the all 1 vector, the well-known matrix determinant lemma "$\det(A + \mathbf{u}\mathbf{v}^\top) = (1 + \mathbf{v}^\top A^{-1}\mathbf{u}) \det A$" gives

$$\begin{aligned}\log\det \nabla^2 h(\mathbf{x}) &= \log \left(1 + x_{d+1}^{-1} \sum_{i=1}^d x_i\right) \cdot \prod_{i=1}^d x_i^{-1} \\ &= -\sum_{i=1}^{d+1} \log x_i = -\sum_{i=1}^d \log x_i - \log\left(1 - \sum_{i=1}^d x_i\right).\end{aligned} \tag{C.3}$$

Composing both sides of (C.2) with $\nabla h^\star$ and using (C.1), (C.3), we then finish the proof by computing

$$\begin{aligned}W(\mathbf{y}) &= V \circ \nabla h^\star(\mathbf{y}) - \sum_{i=1}^d y_i + (d+1)\log\left(1 + \sum_{i=1}^d e^{y_i}\right) \\ &= V \circ \nabla h^\star(\mathbf{y}) - \sum_{i=1}^d y_i + (d+1) h^\star(\mathbf{y}).\end{aligned}$$

D Proof of Lemma 2

The proof relies on rather straightforward computations.

1. In order to show $e^{-(W+C)} = \nabla h \# e^{-V}$ for some constant C, we will verify the Monge-Ampère equation:
$$e^{-V} = e^{-(W\circ\nabla h + C)} \det \nabla^2 h \tag{D.1}$$
for $V = \sum_{i=1}^N V_i$ and $W = \sum_{i=1}^N W_i$, where W_i is defined via (4.2). By (4.2), it holds that
$$\frac{1}{C_i} e^{-NV_i} = e^{-NW_i \circ \nabla h} \det \nabla^2 h, \quad C_i := \frac{1}{\int e^{-NV_i}}. \tag{D.2}$$

Multiplying (D.2) for $i = 1, 2, ..., N$, we get

$$\prod_{i=1}^{N} \frac{1}{C_i} e^{-NV} = e^{-NW \circ \nabla h} \left(\det \nabla^2 h\right)^N . \tag{D.3}$$

The first claim follows by taking the N^{th} root of (D.3).

2. The second claim directly follows by (D.2).
3. Trivial.
4. By (D.1) and (D.2) and using $\nabla h^\star \circ \nabla h(\mathbf{x}) = \mathbf{x}$, we get

$$W_i = V_i \circ \nabla h^\star + \frac{1}{N} \log\det \nabla^2 h(\nabla h^\star) - \log C_i, \tag{D.4}$$

$$W = V \circ \nabla h^\star + \log\det \nabla^2 h(\nabla h^\star) - C, \tag{D.5}$$

which implies $N\nabla W_i - \nabla W = \nabla^2 h^\star \left(N\nabla V_i \circ \nabla h^\star - \nabla V \circ \nabla h^\star\right)$. Since h is 1-strongly convex, $h^\star$ is 1-Lipschitz gradient, and therefore the spectral norm of $\nabla^2 h^\star$ is upper bounded by 1. In the case of $b = 1$, the final claim follows by noticing

$$\mathbb{E}\|\tilde{\nabla} W - \nabla W\|^2 = \frac{1}{N} \sum_{i=1}^{N} \|N\nabla W_i - \nabla W\|^2 \tag{D.6}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \|\nabla^2 h^\star \left(N\nabla V_i \circ \nabla h^\star - \nabla V \circ \nabla h^\star\right)\|^2 \tag{D.7}$$

$$\leq \frac{\|\nabla^2 h^\star\|^2_{\text{spec}}}{N} \sum_{i=1}^{N} \|N\nabla V_i \circ \nabla h^\star - \nabla V \circ \nabla h^\star\|^2 \tag{D.8}$$

$$\leq \mathbb{E}\|\tilde{\nabla} V - \nabla V\|^2. \tag{D.9}$$

The proof for general batch-size b is exactly the same, albeit with more cumbersome notation.

E Proof of Theorem 4

The proof is a simple combination of the existing result in [20] and our theory in Section 3.

By **Theorem 2**, we only need to prove that the inequality (4.3) holds for $D(\tilde{\mathbf{y}}^T \| e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y})$, where $\tilde{\mathbf{y}}^T$ is to be defined below. By assumption, W is unconstrained and satisfies $LI \succeq \nabla^2 W \succeq 0$. By **Lemma 2**, the stochastic gradient $\tilde{\nabla} W$ is unbiased and satisfies

$$\mathbb{E}\|\tilde{\nabla} W - \nabla W\|^2 \leq \mathbb{E}\|\tilde{\nabla} V - \nabla V\|^2 = \sigma^2.$$

Pick a random index[4] $t \in \{1, 2, ..., T\}$ and set $\tilde{\mathbf{y}}^T := \mathbf{y}^t$. Then **Corollary 18** of [20] with $D^2 = \sigma^2$ and $M_2 = 0$ implies $D(\tilde{\mathbf{y}}^T \| e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y}) \leq \epsilon$, provided

$$\beta \leq \min\left\{ \frac{\epsilon}{2\left(Ld + \sigma^2\right)}, \frac{1}{L} \right\}, \quad T \geq \frac{\mathcal{W}_2^2(\mathbf{y}^0, e^{-W(\mathbf{y})}\mathrm{d}\mathbf{y})}{\beta\epsilon}. \tag{E.1}$$

Solving for T in terms of ϵ establishes the theorem.

[4] The analysis in [20] provides guarantees on the probability measure $\nu_T := \frac{1}{N}\sum_{t=1}^{T} \nu_t$ where $\mathbf{y}^t \sim \nu_t$. The $\tilde{\mathbf{y}}^T$ defined here has law ν_T.

F Stochastic Gradients for Dirichlet Posteriors

In order to apply SMLD, one must have, for each term V_i, the corresponding dual W_i defined via (4.2). In this appendix, we derive a closed-form expression in the case of the Dirichlet posterior (3.6).

Recall that the Dirichlet posterior (3.6) consists of a Dirichlet prior and categorical data observations [23]. Let $N := \sum_{\ell=1}^{d+1} n_\ell$, where n_ℓ is the number of observations for category ℓ, and suppose that the parameters α_ℓ's are given. If the i^{th} data is in category $c_i \in \{1, 2, ..., d+1\}$, then we can define $V_i(\mathbf{x}) := -\sum_{\ell=1}^{d+1} \mathbb{I}_{\{\ell=c_i\}} \log x_\ell - \frac{1}{N}\sum_{\ell=1}^{d+1}(\alpha_\ell - 1)\log x_\ell$ so that **Assumption 4** holds. In view of **Lemma 1**, The corresponding dual W_i is, up to a constant, given by

$$W_i(\mathbf{y}) = -\sum_{\ell=1}^{d} \mathbb{I}_{\{\ell=c_i\}} y_\ell - \sum_{\ell=1}^{d} \frac{\alpha_\ell}{N} y_\ell + h^\star + \left(\sum_{\ell=1}^{d+1} \frac{\alpha_\ell}{N}\right) h^\star(\mathbf{y}). \tag{F.1}$$

Similarly, if we take a mini-batch B of the data with $|B| = b$, then

$$\frac{N}{b}\tilde{W}(\mathbf{y}) := \frac{N}{b}\sum_{i\in B} W_i(\mathbf{y}) = -\sum_{\ell=1}^{d}\left(\frac{Nm_\ell}{b} + \alpha_\ell\right) y_\ell + \left(N + \sum_{\ell=1}^{d+1}\alpha_\ell\right) h^\star(\mathbf{y}), \tag{F.2}$$

where m_ℓ is the number of observations of category ℓ in the set B. Apparently, the gradient of (F.2) is (4.4).

G Comparison against SGRHMC for Latent Dirichlet Allocation

The only difference between the experimental setting of [29] and the main text is the number of topics (50 vs. 100). In this appendix, we run SMLD-approximate under the setting of [29] and directly compare against the results reported in [29]. We have also included the SGRLD as a baseline.

Figure 1 reports the perplexity on the test data. According to [29], the best perplexity achieved by SGRHMC up to 10,000 documents is approximately 1400, which is worse than the 1323 by SMLD-approximate. Moreover, from Figure 3 of [29], we see that the SGRHMC yields comparable performance as SGRLD for 2 out 3 independent runs, especially in the beginning phase, whereas the SMLD-approximate has sizeable lead over SGRLD at any stage of the experiment. The potential reason for this improvement is, similar to SGRLD, that the SGRHMC exploits the Riemannian Hamiltonian dynamics, which is more complicated than MLD and hence more sensitive to the discretization error.

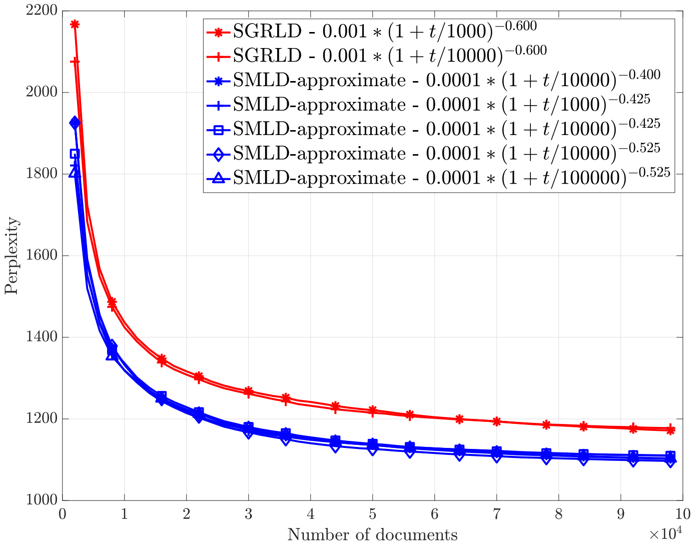


Figure 1: LDA for Wikipedia, 50 topics.