CONVERGENCE ANALYSIS OF THE WASSERSTEIN PROXIMAL ALGORITHM BEYOND CONVEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The proximal algorithm is a powerful tool to minimize nonlinear and nonsmooth functionals in a general metric space. Motivated by the recent progress in studying the training dynamics of the noisy gradient descent algorithm on two-layer neural networks in the mean-field regime, we provide in this paper a simple and self-contained analysis for the convergence of the general-purpose Wasserstein proximal algorithm without assuming geodesic convexity on the objective functional. Under a natural Wasserstein analog of the Euclidean Polyak-Łojasiewicz inequality, we establish that the proximal algorithm achieves an unbiased and linear convergence rate. Our convergence rate improves upon existing rates of the proximal algorithm for solving Wasserstein gradient flows under strong geodesic convexity. We also extend our analysis to the inexact proximal algorithm for geodesically semiconvex objectives. In our numerical experiments, proximal training demonstrates a faster convergence rate than the noisy gradient descent algorithm on mean-field neural networks.

1 INTRODUCTION

Minimizing a cost functional over the space of probability distributions has recently drawn widespread statistical and machine learning applications such as variational inference (Lambert et al., 2022; Ghosh et al., 2022; Yao & Yang, 2023), sampling (Wibisono, 2018; Vempala & Wibisono, 2022; Chewi et al., 2024) and generative modeling (Xu et al., 2024; Cheng et al., 2024), among many others. In this work, we consider the following general optimization problem:

$$\min_{\rho \in \mathcal{P}_2(\Theta)} F(\rho), \tag{1}$$

where F is a real-valued target functional defined on the space of probability distributions $\mathcal{P}_2(\Theta)$ with finite second moments on $\Theta \subset \mathbb{R}^d$. Our motivation for studying this problem stems from analyzing training dynamics of the Gaussian noisy gradient descent algorithm on infinitely wide neural networks, which can be viewed as a forward time-discretization of the mean-field Langevin dynamics (MFLD) (Mei et al., 2019; Hu et al., 2021). Given the connection between sampling and optimization, the continuous-time MFLD is an important example of the Wasserstein gradient flow corresponding to minimizing an entropy-regularized total objective function of large interacting particle systems (cf. Section 2.1).

On the other hand, the Wasserstein gradient flow is conventionally constructed by the following proximal algorithm,

$$\rho_{n+1} = prox_{F,\xi}(\rho_n) := \operatorname*{arg\,min}_{\tilde{\rho} \in \mathcal{P}_2(\Theta)} \left\{ F(\tilde{\rho}) + \frac{1}{2\xi} \mathcal{W}_2^2(\tilde{\rho}, \rho_n) \right\}, \tag{2}$$

where $\xi > 0$ is the time-discretization step size. The Wasserstein proximal algorithm (2) is an iterative *backward time-discretization* procedure for solving (1) and it is also known as the JKO scheme introduced in a seminal work (Jordan et al., 1998). In contrast to various forward-discretization methods such as gradient descent over $\mathcal{P}_2(\Theta)$ and the Langevin sampling algorithms (Durmus et al., 2019; Vempala & Wibisono, 2022; Chewi et al., 2024), proximal algorithms are often *unbiased* in the sense that their convergence guarantees do not depend on the dimension-dependent discretization error with positive step size and they are often more stable than the forward gradient descent

algorithms without strong smoothness condition (Yao & Yang, 2023; Yao et al., 2024; Salim et al., 2021; Cheng et al., 2024). While the proximal algorithms for geodesically convex functionals are well studied in the literature, it remains an open question **whether they can maintain similar unbiased and linear convergence guarantees in discrete time beyond the geodesic convexity**. Current work fills this important gap by establishing linear convergence results without assuming geodesic convexity on the objective functional F.

Our general quantitative convergence rate for the Wasserstein proximal algorithm offers an alternative training scheme to the noisy gradient descent for two-layer neural networks in the mean-field regime. Specifically, a two-layer neural network is parameterized as

$$f(x;\boldsymbol{\theta}) := \frac{1}{m}\sum_{j=1}^{m}\varphi(\theta_j;x) = \int_{\Theta}\varphi(\theta;x)d\rho^m(\theta), \tag{3}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m) \in \mathbb{R}^{d\times m}$ and $\rho^m = \frac{1}{m}\sum_{j=1}^{m}\delta_{\theta_j}$ is the empirical distribution of the hidden neuron parameters. The perceptron $\varphi(\theta_j;x)$ in (3) can take the form $\varphi(\theta_j;x) = \sigma(\theta_j^\top x)$ where σ is some nonlinear activation function. Given a training dataset $(x_i, y_i)_{i=1}^N \sim p(x,y)$ and a convex loss function $l(\cdot)$ (such as the squared loss and logistic loss), the L^2-regularized training risk is defined as

$$\begin{aligned} R(\rho^m) &= \frac{1}{N}\sum_{i=1}^{N} l(f(x_i;\boldsymbol{\theta}), y_i) + \frac{\lambda}{m}\sum_{j=1}^{m}\|\theta_j\|^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} l\left(\int_{\Theta}\varphi(\theta;x_i)d\rho^m(\theta), y_i\right) + \lambda\int_{\Theta}\|\theta\|^2 d\rho^m(\theta). \end{aligned} \tag{4}$$

where $\lambda > 0$ is the coefficient of L_2-regularization. The Gaussian noisy gradient descent algorithm on the L^2-regularized training risk can be written as the following stochastic recursion

$$\theta_j^{n+1} = \theta_j^n - \xi\nabla\frac{\delta R}{\delta\rho}\left(\frac{1}{m}\sum_{\ell=1}^{m}\delta_{\theta_\ell^n}\right)(\theta_j^n) + \sqrt{2\xi\tau}z_{nj}, \tag{5}$$

where z_{nj} are i.i.d. $\mathcal{N}(0, I)$ and $\tau > 0$ represents the Gaussian noise variance. In (5), $\frac{\delta R}{\delta\rho}(\rho)$ is the first variation of R at ρ (cf. Definition A.2). The limiting dynamics of (5) under $m \to \infty$ and $\xi \to 0$ is called the continuous-time MFLD (Hu et al., 2021). Under a uniform log-Sobolev inequality (LSI) assumption (cf. Definition C.1), linear convergence of MFLD to the optimal value of the total objective (L^2-training risk plus an entropy term) is established in Nitanda et al. (2022); Chizat (2022), and the noisy gradient descent algorithm is subject to a dimension-dependent time-discretization error (Nitanda et al., 2022), which may slow down the convergence.

To remove the time-discretization error, we may instead train the neural network with the Wasserstein proximal algorithm (2). Since such neural network architecture satisfies the uniform LSI which in turn implies a Wasserstein Polyak-Łojasiewicz (PL) inequality (cf. Definition 3.2), our algorithm can achieve an unbiased linear rate of convergence to a global minimum of total objective.

1.1 Contributions

In this work, we give a simple and self-contained convergence rate analysis of the Wasserstein proximal algorithm (2) for minimizing the objective function satisfying a PL-type inequality without resorting to any geodesic convexity assumption. Below we summarize our main contributions.

- To the best of our knowledge, current work is among the first work to obtain an unbiased and linear convergence rate of the general-purpose Wasserstein proximal algorithm for optimizing a functional under **merely a PL-type inequality**. When restricted to μ-convex ($\mu > 0$) objective functional, our result yields a sharper linear convergence rate (in function value and minimizer under $\mathcal{W}_2$ distance) than the existing literature (Yao & Yang, 2023; Cheng et al., 2024).
- The linear convergence guarantee provides a new training scheme for two-layer wide neural networks in the mean-field regime. Our numeric experiments show a faster training phase (up to particle discretization error) than the (forward) noisy gradient descent method.
- We also analyze the inexact proximal algorithm for geodesically semiconvex objectives under PL-type inequality.

1.2 Literature Review

Recently, various time-discretization methods have been proposed for minimizing a functional over a single distribution. Different from the proximal algorithm, some explicit forward schemes that can be seen as gradient descent in Wasserstein space are proposed (Chewi et al., 2020; Liu & Wang, 2016). For example, Chewi et al. (2020) studies a gradient descent algorithm for solving the barycenter problem on Bures-Wasserstein manifold of Gaussian distributions. The Langevin algorithm, as another forward discretization of Wasserstein gradient flows via its stochastic differential equation (SDE) recursion, is widely used in the sampling literature. Numerous works (Durmus et al., 2019; Vempala & Wibisono, 2022; Chewi et al., 2024) have been devoted to the analysis of the Langevin algorithm under different settings and its variants (Zhang et al., 2023; Wu et al., 2022). However, Langevin algorithms are naturally biased for a positive step size. Salim et al. (2021) introduced a hybridized forward-backward discretization, namely the Wasserstein proximal gradient descent, and proved convergence guarantees for geodesically convex objective, akin to the proximal gradient descent algorithm in Euclidean spaces.

Existing rate analysis for proximal algorithm. Though convergence rate analysis for Langevin algorithms under strong convexity is well-developed, it is until recently that convergence rate of proximal algorithm on geodesically convex objectives is obtained. One advantage of proximal algorithm is that it ensures a dimension-independent convergence guarantee directly for any starting distribution. Yao & Yang (2023); Cheng et al. (2024) proved an unbiased linear convergence result for the μ-strongly convex objective. The condition is relaxed to geodesic convexity and quadratic growth of functional in (Yao et al., 2024). However, convergence analysis for non-geodesically convex objective functionals is missing.

Convergence rate of different time-discretizations under PL-type inequality. Vempala & Wibisono (2022) obtained a biased linear convergence result for Langevin dynamics under the log-Sobolev inequality (LSI) and smoothness condition. Nitanda et al. (2022) extended this result to MFLD with similar techniques. Proximal Langevin algorithm proposed by Wibisono (2019), attains a biased linear convergence rate under the LSI, while an extra smoothness condition of the second derivative of the sampling function is required. Proximal sampling algorithm (Chen et al., 2022), assuming access to samples from an oracle distribution, achieves an unbiased linear convergence for sampling from Langevin dynamics under the LSI, while the analysis requires geodesic semiconvexity (cf. Definition A.3). Fan et al. (2023); Liang & Chen (2024) improved the results, however, they still concentrate on sampling on a fixed function and cannot be applied to MFLD.

To highlight the distinction between our contributions and existing results from the literature, we make the following comparison between explicit convergence guarantees of the Wasserstein proximal algorithm (our result) and Langevin algorithms for optimizing the KL divergence in Table 1. Similar comparison on the convergence rates can be made between our result and the forward time-discretization of MFLD under further assumptions (Nitanda et al., 2022; Chizat, 2022).

Table 1: Comparison between Langevin and Wasserstein proximal algorithms for KL divergence.

Algorithm	Assumptions	Step size	Convergence guarantee at n-th iteration
Langevin	μ-LSI L-smooth (on Θ)	$0 < \xi < \frac{\mu}{4L^2}$	$e^{-n\mu\xi} D_{\mathrm{KL}}(\rho_0 \| \nu) + \frac{8\xi d L^2}{\mu}$ Vempala & Wibisono (2022)
Proximal	μ-LSI semiconvex (on Θ)	$\xi > 0$	$\frac{1}{(1+\mu\xi)^{2n}} D_{\mathrm{KL}}(\rho_0 \| \nu)$ Ours
Proximal	μ-strongly convex (on Θ)	$\xi > 0$	$\frac{1}{(1+\mu\xi)^{n}} D_{\mathrm{KL}}(\rho_0 \| \nu) \to \frac{1}{(1+\mu\xi)^{2n}} D_{\mathrm{KL}}(\rho_0 \| \nu)$ Yao & Yang (2023) Ours

The remainder of this paper is organized as follows. In Section 2, we provide some background knowledge for Wasserstein gradient flows and a strong lemma that can be applied to Wasserstein space. In Section 3, we present our main convergence results. In Section 4, we discuss how to apply the proximal algorithm for MFLD of a two-layer neural network and provide numerical experiments exploring the behavior of the proximal algorithm.

Notations. We assume $\Theta = \mathbb{R}^d$ (by default) unless explicitly indicating that it is a compact subset of $\mathbb{R}^d$. Let $\mathcal{P}_2(\Theta)$ be the collection of all probability measures with finite second moment and $\mathcal{P}_2^a(\Theta) \subset \mathcal{P}_2(\Theta)$ be the absolutely continuous measures. For a measurable map $T : \Theta \to \Theta$, let $T_\# : \mathcal{P}_2(\Theta) \to \mathcal{P}_2(\Theta)$ be the corresponding pushforward operator. For probability measures μ and ν, we shall use T_μ^ν to denote the optimal transport (OT) map from μ to ν and $\mathbf{id}$ to denote the identity map. We use $\mathcal{W}_2(\cdot,\cdot)$ to denote the Wasserstein-2 distance. We denote $\partial F(\rho)$ to be the strong subdifferential at $\rho \in \mathcal{P}_2(\Theta)$, $\mathcal{D}(F) := \{\rho \in \mathcal{P}_2(\Theta) \mid F(\rho) < \infty\}$ to be the domain of F that has finite functional value, and $D(|\partial F|)$ to be the domain of F that has finite metric slope, see [Lemma 10.1.5, Ambrosio et al. (2005)]. We refer to Appendix A for more notions and definitions.

2 Preliminaries

In this section, we review the connection between Wasserstein gradient flows and the associated Langevin dynamics.

2.1 Wasserstein gradient flows and continuous-time Langevin dynamics

Gradient flows in the Wasserstein space of probability distributions provide a powerful means to understand and develop practical algorithms for solving diffusion-type equations (Ambrosio et al., 2005). For a smooth Wasserstein gradient flow, noisy gradient descent algorithms over relative entropy functionals are often used for space-time discretization via the stochastic differential equation (SDE). Below we illustrate two main Wasserstein gradient flow examples involving the linear and nonlinear Fokker-Planck equations, which model the diffusion behavior of probability distributions.

Langevin dynamics via the Fokker-Planck equation. The Langevin dynamics for the target distribution $\nu = e^{-f}$ is defined as an SDE,

$$d\theta_t = -\nabla f(\theta_t)dt + \sqrt{2}dW_t \tag{6}$$

where W_t is the standard Brownian motion in Θ with zero initialization. It is well-known that, see e.g., Chapter 8 of (Santambrogio, 2015), if the process (θ_t) evolves according to the Langevin dynamics in (6), then their marginal probability density distributions $\rho_t(\theta)$ satisfy the Fokker-Planck equation

$$\partial_t \rho_t - \Delta \rho_t - \nabla \cdot (\rho_t \nabla f) = 0,$$

which is the Wasserstein gradient flow for minimizing the KL divergence

$$D_{\mathrm{KL}}(\rho\|\nu) := \int_\Theta \rho(\theta) \log \frac{\rho(\theta)}{\nu(\theta)} d\theta = \int_\Theta f(\theta)\rho(\theta)d\theta + \int_\Theta \rho(\theta)\log\rho(\theta)d\theta.$$

If ν satisfies a log-Sobolev inequality (LSI) with constant $\mu > 0$, i.e., if for all $\rho \in \mathcal{P}_2^a(\Theta)$, we have

$$D_{\mathrm{KL}}(\rho\|\nu) \le \frac{1}{2\mu} J(\rho\|\nu), \tag{7}$$

where $J(\rho\|\nu) = \int \|\nabla \log \frac{\rho}{\nu}\|^2 d\rho$ is the relative Fisher information, then the continuous-time Langevin ρ_t convergences to ν exponentially fast.

Mean-field Langevin dynamics (MFLD) via the McKean-Vlasov equation. In an interacting m-particle system, the potential energy contains a nonlinear interaction term in addition to $\int f d\rho$. More generally, in the mean-field limit as $m \to \infty$, the nonlinear Langevin dynamics can be described as

$$d\theta_t = -\nabla \frac{\delta R}{\delta \theta}(\theta_t)dt + \sqrt{2\tau}dW_t, \tag{8}$$

where $R : \mathcal{P}_2(\Theta) \to \mathbb{R}$ is a cost functional such as the L^2-regularized training risk of mean-field neural networks in (4) and $\tau > 0$ is a temperature parameter. For a convex loss l, the risk R has the linear convexity in (4). Process evolving according to (8) solves the following McKean-Vlasov equation (Yao et al., 2022),

$$\partial_t \rho_t - \tau \Delta \rho_t - \nabla \cdot (\rho_t \nabla \frac{\delta R}{\delta \rho}(\rho_t)) = 0,$$

which is the Wasserstein gradient flow of the entropy-regularized total objective,

$$F_\tau(\rho) = R(\rho) + \tau \int_\Theta \rho \log \rho. \tag{9}$$

Similarly, as in the linear Langevin case if the proximal Gibbs distribution of ρ (cf. Definition C.1) satisfies a uniform LSI (cf. Definition C.1), then MFLD converges to the optimal value exponentially fast in continuous time (Chizat, 2022; Nitanda et al., 2022) and, in the case of infinite-width neural networks in mean-field regime, it subjects to a dimension-dependent time-discretization error (Nitanda et al., 2022).

3 Convergence Rate Analysis

In this section, we first introduce a natural PL inequality in the Wasserstein space and then provide the convergence rate analysis for the Wasserstein proximal algorithm under such a weak assumption. Then, we shall extend our analysis to the inexact proximal algorithm setting. In the whole section, we make the following regularity assumption,

Assumption 1 (Regularity assumption)**.** *The functional $F : \mathcal{P}_2(\Theta) \to (-\infty, +\infty]$ is proper, weakly lower semicontinuous with $\mathcal{D}(F) \subset \mathcal{P}_2^a(\Theta)$.*

Assumption 1 ensures that the proximal operator (2) admits a minimizer (cf. Lemma B.2) and all minimizers belong to $\mathcal{P}_2^a(\Theta)$. We refer the reader to Lemma B.3 for conditions that guarantee weakly lower semicontinuity.

Definition 3.1 (Hopf-Lax formula)**.** *Let $\xi > 0$. The Hopf-Lax formula $u(\rho, \xi)$ of a functional $F : \mathcal{P}_2(\Theta) \to \mathbb{R}$ is defined as*

$$u(\rho, \xi) := F(\rho_\xi) + \frac{1}{2\xi}\mathcal{W}_2^2(\rho_\xi, \rho) = \min_{\tilde{\rho} \in \mathcal{P}_2(\Theta)} \left\{ F(\tilde{\rho}) + \frac{1}{2\xi}\mathcal{W}_2^2(\tilde{\rho}, \rho) \right\}, \tag{10}$$

where $\rho_\xi = prox_{F,\xi}(\rho)$.

The Hopf-Lax formulation in (10) is also known as the Moreau-Yoshida approximation (Ambrosio et al., 2005). Below, we present a key connection between the time-derivative of the Hopf-Lax semigroup and the squared Wasserstein distance between the proximal and the initial point.

Lemma 3.1. *Let $\rho \in \mathcal{D}(F)$. Under Assumption 1, we have*

$$\partial_\xi u(\rho, \xi) = -\frac{1}{2\xi^2}\mathcal{W}_2^2(\rho_\xi, \rho) \tag{11}$$

holds for $\xi \in (0, +\infty)$ with at most countable exceptions.

Proof of Lemma 3.1 is provided in Appendix C. The proof essentially follows from Proposition 3.1 and Proposition 3.3 in (Ambrosio et al., 2013), which are summarized together in Lemma B.1 for completeness.

Remark 3.2 (Computation of the proximal operator)**.** *To compute ρ_{n+1} in (2), we can reformulate the proximal algorithm into an optimization problem in functional space. Finding ρ_{n+1} is equivalent to finding an optimal transport (OT) map T such that $T_\# \rho_n$ minimizes (2),*

$$T_{\rho_n}^{\rho_{n+1}} = \arg\min_{T:\Theta\to\Theta} F(T_\#\rho_n) + \frac{1}{2\xi}\int_\Theta \|T(\theta) - \theta\|^2 d\rho_n. \tag{12}$$

3.1 Convergence Rates of Exact Proximal Algorithm

In this subsection, we establish the convergence rate for the Wasserstein proximal algorithm (2). First, we define the PL inequality in Wasserstein space as in (Boufadène & Vialard, 2023).

Definition 3.2 (Polyak-Łojasiewicz inequality)**.** *For any $\rho \in \mathcal{D}(F)$, the objective functional F satisfies the following inequality with $\mu > 0$,*

$$\int_\Theta \|\nabla \frac{\delta F}{\delta \rho}(\rho)\|^2 d\rho \geq 2\mu(F(\rho) - F(\rho^*)), \tag{13}$$

where ρ^* is any global minimizer of F. Denote $F^* = F(\rho^*)$. The Wasserstein PL inequality generalizes the classical PL inequality in Euclidean space $\|\nabla f(\theta)\|^2 \geq 2\mu(f(\theta) - f(\theta^*))$ where $f : \Theta \to \mathbb{R}$ (Karimi et al., 2016).

For KL divergence, the Wasserstein analog of the Euclidean PL inequality is the LSI. Proving an LSI (7) is often difficult since it is almost exclusively used to study the linear convergence for KL divergence-type objective functionals. We remark that with a convex function f, the quadratic growth of f implies the PL inequality in Euclidean space (Karimi et al., 2016) and the quadratic growth of its KL objective implies LSI (Yao et al., 2024). Previous works (Boufadène & Vialard, 2023; Kondratyev et al., 2016; Chizat, 2022) show that under certain regularity conditions, the continuous-time dynamics have linear convergence under the Wasserstein PL inequality. Our paper considers the problem of minimizing a general functional in (1), where the convergence analysis of the proximal algorithm is directly based on Wasserstein PL inequality (13).

Assumption 2 (Proximal trajectory)**.** *For every $\rho \in \mathcal{D}(F)$ and every $\rho_\xi = prox_{F,\xi}(\rho)$,*

(a) $\nabla \frac{\delta F}{\delta \rho}(\rho_\xi) = \frac{T^\rho_{\rho_\xi} - \mathbf{id}}{\xi}$, *$\rho_\xi$-a.e., or*

(b) $\|\nabla \frac{\delta F}{\delta \rho}(\rho_\xi)\|_{L_2(\rho_\xi)} \leq \|\frac{T^\rho_{\rho_\xi} - \mathbf{id}}{\xi}\|_{L_2(\rho_\xi)}$.

Remark 3.3. *By Lemma 10.1.2 in (Ambrosio et al., 2005), $\rho_\xi \in D(|\partial F|)$ and $(T^\rho_{\rho_\xi} - \mathbf{id})/\xi$ is a strong subdifferential of F at ρ_ξ. Then Assumption 2(b) requires that the Wasserstein gradient $\nabla \frac{\delta F}{\delta \rho}(\rho_\xi)$ is the **minimal norm selection** of the strong subdifferential of F along the proximal algorithm trajectory. If Θ is a compact set in $\mathbb{R}^d$, Assumption 2(a) automatically holds due to the existence of the first variation of $\mathcal{W}_2(\cdot, \rho)$ distance for any fixed $\rho \in \mathcal{P}_2(\Theta)$ (cf. Lemma B.5). Moreover, MFLD under the assumptions of Corollary 3.5 and Langevin dynamics under the assumptions of Corollary 3.6 satisfy Assumption 2(b).*

Now we are ready to state the main theorem of this paper.

Theorem 3.4 (Convergence rate of the exact proximal algorithm under PL inequality)**.** *Under Assumptions 1 and 2, if the objective functional F in (1) satisfies the PL inequality (13), then for any $\xi > 0$, the Wasserstein proximal algorithm (2) satisfies*

$$F(\rho_n) - F^* \leq \frac{1}{(1+\xi\mu)^{2n}}(F(\rho_0) - F^*).$$

Proof of Theorem 3.4. It suffices to prove one-step contraction. We begin with

$$\begin{aligned}
\partial_\xi(u(\rho,\xi) - F^*) &= -\frac{\mu}{2\xi(1+\mu\xi)}\mathcal{W}_2^2(\rho_\xi,\rho) - \frac{1}{2(1+\mu\xi)\xi^2}\mathcal{W}_2^2(\rho_\xi,\rho) && \text{(by Lemma 3.1)}\\
&= -\frac{\mu}{2\xi(1+\mu\xi)}\mathcal{W}_2^2(\rho_\xi,\rho) - \frac{1}{2(1+\mu\xi)\xi^2}\int_\Theta \|T^\rho_{\rho_\xi} - \mathbf{id}\|^2 d\rho_\xi && \text{(by Brenier's theorem)}\\
&\leq -\frac{\mu}{2\xi(1+\mu\xi)}\mathcal{W}_2^2(\rho_\xi,\rho) - \frac{1}{2(1+\mu\xi)}\int_\Theta \|\nabla\frac{\delta F}{\delta\rho}(\rho_\xi)\|^2 d\rho_\xi && \text{(by Assumption 2)}\\
&\leq -\frac{\mu}{2\xi(1+\mu\xi)}\mathcal{W}_2^2(\rho_\xi,\rho) - \frac{\mu}{(1+\mu\xi)}(F(\rho_\xi) - F^*) && \text{(by PL (13))}\\
&= -\frac{\mu}{1+\mu\xi}(u(\rho,\xi) - F^*). && \text{(by Def 3.1)}
\end{aligned}$$

Using Lemma B.6 to deal with the technique issue of almost everywhere differentiability, we have

$$u(\rho,\xi) - F^* \leq (u(\rho,0) - F^*)\exp\left(\int_0^\xi -\frac{\mu}{1+\mu t}dt\right) = (F(\rho) - F^*)\frac{1}{1+\mu\xi}.$$

Invoking the PL inequality (13) once again, we obtain that

$$\begin{aligned}
(F(\rho) - F^*)\frac{1}{1+\mu\xi} &\geq u(\rho,\xi) - F^* = F(\rho_\xi) - F^* + \frac{1}{2\xi}\mathcal{W}_2^2(\rho_\xi,\rho)\\
&\geq F(\rho_\xi) - F^* + \frac{\xi}{2}\int_\Theta \|\nabla\frac{\delta F}{\delta\rho}(\rho_\xi)\|^2 d\rho_\xi \geq (1+\mu\xi)(F(\rho_\xi) - F^*).
\end{aligned}$$

□

Next, we specialize our general-purpose convergence guarantee for the Wasserstein proximal algorithm to MFLD induced from training a two-layer neural network (3) in the mean-field regime, and Langevin dynamics, which is a special case of MFLD.

Corollary 3.5 (Wasserstein proximal algorithm on MFLD)**.** *Let F_τ be the total objective functional defined in (9). Suppose that the loss function $l(\cdot)$ is either squared loss or logistic loss. If the perceptron $\varphi(\theta; x)$ is bounded by K and $sup_{\theta,x}\|\nabla_\theta \varphi(\theta; x)\|$ is finite, then F_τ satisfies Assumptions 1, 2(b) and the PL inequality (13). Consequently, the Wasserstein proximal algorithm (2) satisfies for any $\xi > 0$,*

$$F_\tau(\rho_n) - F_\tau^* \leq \frac{1}{(1+\xi\tau\mu_\tau)^{2n}}(F_\tau(\rho_0) - F_\tau^*), \tag{14}$$

$$\mathcal{W}_2(\rho^*, \rho_n) \leq \sqrt{\frac{2}{\mu_\tau \tau}(F_\tau(\rho_0) - F_\tau^*)}\left(\frac{1}{1+\xi\mu_\tau\tau}\right)^n. \tag{15}$$

Our Theorem 3.4 can also be applied to derive the convergence rate for backward time-discretized KL divergence (i.e., linear Langevin dynamics).

Corollary 3.6 (Wasserstein proximal algorithm on Langevin dynamics)**.** *If $f : \Theta \to \mathbb{R}$ is semiconvex, and $\nu = \exp(-f)$ satisfies the μ-LSI condition (7), Assumptions 1, 2(b) and the PL inequality (13) are satisfied. Consequently, the Wasserstein proximal algorithm (2) satisfies for any $\xi > 0$,*

$$D_{\mathrm{KL}}(\rho_n\|\nu) \leq \frac{1}{(1+\mu\xi)^{2n}} D_{\mathrm{KL}}(\rho_0\|\nu) \quad \textit{and} \quad \mathcal{W}_2(\rho_n, \nu) \leq \sqrt{\frac{2}{\mu} D_{\mathrm{KL}}(\rho_0\|\nu)}\frac{1}{(1+\mu\xi)^n}.$$

Remark 3.7. *Moreover, if Θ is a compact set, f is lower semicontinuous and bounded from below, and $\nu = exp(-f)$ satisfies μ-LSI condition (7), then Assumptions 1, 2(a) and the PL inequality are satisfied. Therefore, for any $\xi > 0$, the exact Wasserstein proximal algorithm (2) satisfies*

$$D_{\mathrm{KL}}(\rho_n\|\nu) \leq \frac{1}{(1+\mu\xi)^{2n}} D_{\mathrm{KL}}(\rho_0\|\nu) \quad \textit{and} \quad \mathcal{W}_2(\rho_n, \nu) \leq \sqrt{\frac{2}{\mu} D_{\mathrm{KL}}(\rho_0\|\nu)}\frac{1}{(1+\mu\xi)^n}.$$

Applying similar proof techniques to the μ-strongly convex objective functional F (1), we obtain a faster convergence rate than those in existing literature, see Remark 3.9 below. Specially, we are able to avoid evoking Assumption 2 with careful modifications to our proof.

Theorem 3.8 (Sharper convergence rates of the exact proximal algorithm: strongly geodesically convex objective)**.** *Under Assumption 1, if the objective functional F in (1) is μ-strongly convex along geodesics, then for any $\xi > 0$, the Wasserstein proximal algorithm (2) satisfies*

$$F(\rho_n) - F^* \leq \frac{1}{(1+\mu\xi)^{2n}}(F(\rho_0) - F^*) \ \textit{and} \ \mathcal{W}_2(\rho^*, \rho_n) \leq \sqrt{\frac{2}{\mu}(F(\rho_0) - F^*)}\frac{1}{(1+\mu\xi)^{2n}}.$$

Remark 3.9. *The bound obtained in Corollary 3.8 is faster than those in (Yao & Yang, 2023) and (Cheng et al., 2024) for μ-strongly convex objective, where the convergence rate $(1+\mu\xi)^{-n}(F(\rho_0) - F^*)$ is proved. Corollary 3.8 applies to functionals corresponding to interacting particle systems $F(\rho) = \int f(\theta)d\rho(\theta) + \int w(\theta_1, \theta_2)d\rho(\theta_1)d\rho(\theta_2) + \int \log \rho(\theta)d\rho(\theta)$, where both the external force $f(\cdot)$ and interaction potential $w(\cdot, \cdot)$ are both convex and at least one of them is strongly convex (Yao et al., 2024).*

3.2 Convergence Rates of Inexact Proximal Algorithm

This subsection investigates the inexact proximal algorithm where numerical errors are allowed in each iteration. Let $\overline{\rho}_n$ be the inexact solution of the proximal algorithm at iteration n. We need an additional smoothness assumption on the proximal flow to provide a quantitative analysis for the proximal algorithm when the OT map at each iteration is allowed to be estimated with errors.

Assumption 3 (Smoothness of inexact proximal flow)**.** *If $\overline{\rho}_n \in \mathcal{C}^1(\Theta)$, then $\overline{\rho}_{n+1} \in \mathcal{C}^1(\Theta)$, for all $n \in \mathbb{N}$.*

When the proximal algorithm (2) is initialized with $\overline{\rho}_1 \in \mathcal{C}^1(\Theta)$, Assumption 3 ensures that the inexact proximal flow $\overline{\rho}_n$ remains $\mathcal{C}^1(\Theta)$. In practice, to optimize over (12) via $T_{\overline{\rho}_n}^{\overline{\rho}_{n+1}}$, the learned

OT map is typically restricted to a specific class of functionals, such as a normalizing flow (Xu et al., 2024) or a neural network (Yao & Yang, 2023) with a certain structure. Even though this assumption is mainly for technical purpose, we provide Lemma B.8 in Appendix, showing that restricting the learned OT map to some classes can ensure Assumption 3. Next we define the error $\beta_{n+1} : \Theta \to \mathbb{R}$ in estimating the Wasserstein gradient in the $(n+1)$-th iterate as,

$$\beta_{n+1} = T_{\overline{\rho}_{n+1}}^{\overline{\rho}_n} - \mathbf{id} - \xi \nabla \frac{\delta F}{\delta \rho}(\overline{\rho}_{n+1}). \tag{16}$$

Under Assumption 3, if $\overline{\rho}_{n+1}$ is the exact solution of the Wasserstein proximal algorithm (2), then $\beta_{n+1} = 0$ (cf. Lemma B.10). Therefore, we utilize $\int \|\beta_{n+1}\|^2 d\overline{\rho}_{n+1}$ to depict the error induced by solving (2) inexactly,

$$\int_{\Theta} \|\beta_{n+1}\|^2 d\overline{\rho}_{n+1} \leq \epsilon_{n+1},$$

which can be viewed as measuring the norm of strong subdifferential as in Euclidean case. Furthermore, we need the additional geodesic semiconvexity assumption apart from PL inequality to control the inexact error, which differs from Section 3.1 which only relies on PL inequality. Examples that satisfy both semiconvexity and PL inequality include the objective of MFLD under the assumptions of Corollary 3.5 and KL divergence objective under the assumptions of Corollary 3.6. Now we are in the position to quantify the impact of numerical errors on the proximal algorithm.

Theorem 3.10 (Convergence rates of the inexact proximal algorithm under PL inequality)**.** *Suppose $\overline{\rho}_n \in D(|\partial F|)$ for every $n \in \mathbb{N}$. Under Assumptions 1 and 3, if F is $(-L)$- geodesically semiconvex, F satisfies the PL inequality (13), and $0 < \xi \leq \frac{1}{L}$, then we have the following cases.*
(a) If $\epsilon_n \leq C_{exp}\gamma^n$ with $\gamma \in (0,1)$, then there exists $C_1 = C_1(\mu, L, \gamma, C_{exp})$ such that

$$F(\overline{\rho}_n) - F^* \leq \left(\frac{1}{1+\mu\xi}\right)^n (F(\rho_0) - F^*) + C_1 \max\left\{\frac{1}{1+\mu\xi}, \gamma\right\}^{n+1}.$$

(b) If $\epsilon_n \leq C_{poly} n^{-\zeta}$ with $\zeta > 0$, then there exists $C_2 = C_2(\mu, L, \zeta, C_{poly})$ such that,

$$F(\overline{\rho}_n) - F^* \leq \left(\frac{1}{1+\mu\xi}\right)^n (F(\rho_0) - F^*) + \frac{C_2}{n^{\zeta}}.$$

Remark 3.11. *Theorem 3.10 demonstrates how the decay of inexact error impacts the convergence behavior under PL inequality. If the numerical error decays at an exponential rate, the linear convergence rate still holds. However, if the numerical error decays at a polynomial rate, the linear convergence will degrade to a sublinear rate. Under Assumption 1, the inexact proximal algorithm has been well-studied under strong geodesic convexity in [Section 4.3, Yao et al. (2024)].*

4 Numerical Experiments

In this section, we first present the application of the exact proximal algorithm on the KL divergence functional, for which the particle and the distribution updates can be computed explicitly. Then we show how to apply the proximal algorithm for the regularized training objective of two-layer neural networks in the mean-field regime.

4.1 Linear Langevin dynamics

In this subsection, we apply the proximal algorithm on KL divergence with the target distribution $\nu = \exp(-\frac{1}{2}\theta^2)$ where $\theta \in \mathbb{R}$. Note that $D_{KL}(\cdot\|\nu)$ is 1-strongly convex. We provide numerical experiments to explore the dynamical behavior of the proximal algorithm.

When both initialization and the target distributions are Gaussian, problem (12) can be explicitly solved and ρ_n remains Gaussian for every n. In particular, closed forms of the particle and distribution updates are available in (Wibisono, 2018). Additionally, the $\mathcal{W}_2$ distance between two Gaussian distributions, known as the Bures-Wasserstein distance, can be computed explicitly. In the experiment, we set the initialization Gaussian distribution to be $\mathcal{N}(0, 100)$, step size $\xi = 0.1$, and iterations equal to 60.

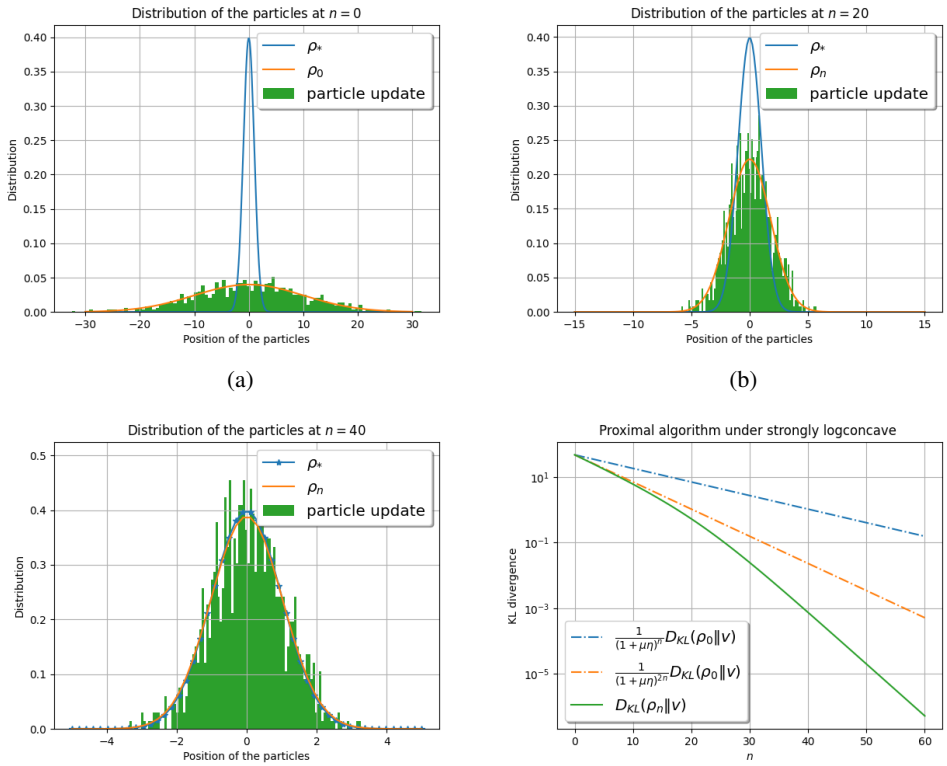


Figure 1: Wasserstein proximal algorithm on the KL divergence with $\nu = \exp(-\|\theta\|^2/2)$.

In Figure 1, the distribution of particles, represented by the histogram, approximates ρ_n and converges to ρ^* after several iterations (approximately 40 iterations in this experiment). The linear converge result of $D_{KL}(\rho_n \| \nu)$ in Figure 1d demonstrates a sharper bound holds for μ-convex objective with respect to (Yao & Yang, 2023; Cheng et al., 2024), as Corollary 3.8 suggests.

4.2 Mean-field neural network training with entropy regularization

In practice, the optimization problem (12) typically lacks an explicit solution. Therefore, we can use particle methods to approximate the time-evolving probability distributions, and we can solve an approximate T^{n+1} using functional approximation methods. When applied to our entropy-regularized total objective of mean-field neural network (9), the functional approximation method can be expressed as,

$$T^{n+1} = \arg\min_{T} \frac{1}{N}\sum_{i=1}^{N} l\Big(\frac{1}{m}\sum_{j=1}^{m}\varphi(T(\theta_j^n), x_i), y_i\Big) + \frac{\lambda}{m}\sum_{j=1}^{m}\|T(\theta_j^n)\|^2 - \frac{\tau}{m}\sum_{j=1}^{m}\log|\det\nabla T(\theta_j^n)| + \frac{1}{2m\xi}\sum_{j=1}^{m}\|T(\theta_j^n) - \theta_j^n\|^2, \tag{17}$$

where the change of variable for entropy (Mokrov et al., 2021) is utilized. In our work, we specifically employ a shallow neural network to learn the optimal transport map T^{n+1} at each iteration, using the right-hand side of (17) as the loss function.

Experiments. In our experiments, we aim to optimize the MFLD entropy-regularized total objective (9) with $\varphi(\theta, x) = \tanh(\theta^T x)$ for $\theta, x \in \Theta$. The parameters are set as $d = 2$,$\lambda = 0.1$, $\tau = \{0.04, 0.1\}$, with the number of particles $m = 100$ and a discretized step size of $\xi = 0.1$. We generate $N = 1000$ training data samples using a teacher model $y = sin(\alpha^T x)$, where $x \sim \mathcal{N}(0, I)$. Our goal is to compare the proximal algorithm with the neural network-based functional approximation (17) with the noisy gradient descent algorithm (5).

We first randomly generated a dataset and then conducted 5 repeated experiments for both $\tau = 0.04$ and $\tau = 0.1$ on the same generated data. For each experiment, a new weight empirical distribution is generated from standard Gaussian distribution, and both algorithms will use the same weight as the initial value. For the learning of the optimal transport map, we train a one-hidden layer fully-connected neural network of the form $g(\theta) = W_2\sigma(W_1\theta)$ for each iteration using Adam optimizer with learning rate 0.004, where $W_1 \in \mathbb{R}^{q\times d}$, $W_2 \in \mathbb{R}^{d\times q}$ where $\sigma(\cdot) = ReLU(\cdot)$ and $q = 1000$.

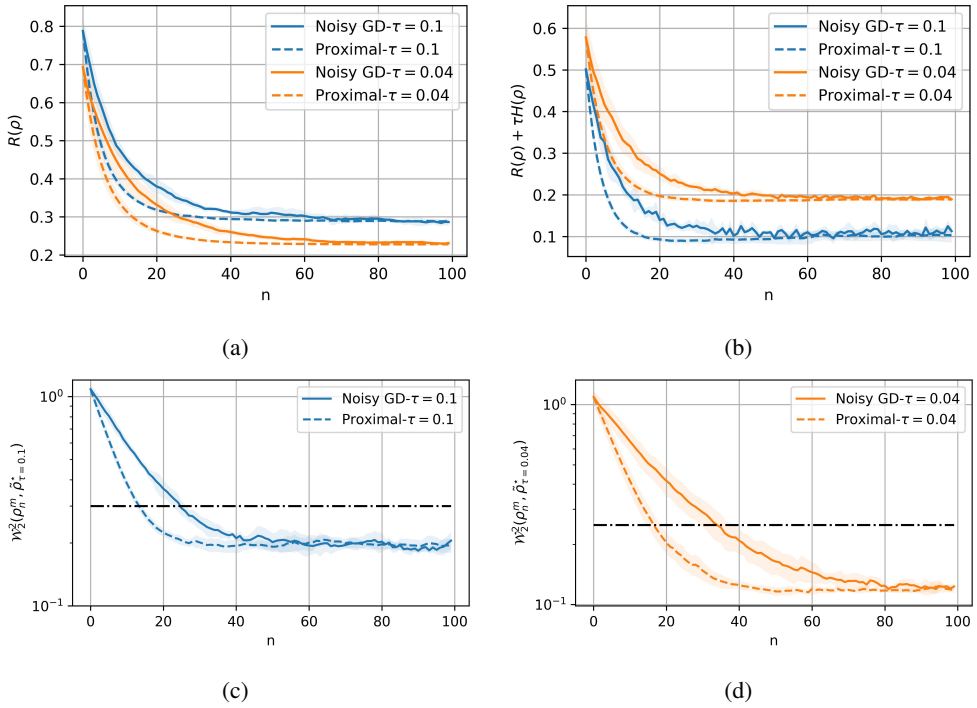


Figure 2: Wasserstein proximal algorithm on MFLD objective. Note that for Figure 2c and Figure 2d, the y-axis is on log-scale.

In Figure 2a and Figure 2b, we observe that both the L^2-regularized loss R and the total objective $F = R + \tau H$ converge under two algorithms, where the nearest neighbor estimator (Kozachenko & Leonenko, 1987) is utilized for estimating $H(\rho) = \int \rho \log \rho$. To better depict the convergence rate of the Wasserstein proximal algorithm and the Langevin algorithm (forward time-discretization of MFLD), we obtain a reference $\tilde{\rho}_\tau^*$ of ρ^*, by running the noisy gradient descent algorithm with very small step size 10^{-3} and $m = 1000$ particles. In the early training phase of both algorithms, $\mathcal{W}_2^2(\rho_n^m, \tilde{\rho}_\tau^*)$ is dominated by $\mathcal{W}_2^2(\rho_n^m, \rho^*)$ and exhibits a linear convergence above the black dash-dot line as shown in Figure 2c and Figure 2d. Within this phase, the Wasserstein proximal algorithm demonstrates a faster linear rate thanks to the unbiased linear convergence nature of $\mathcal{W}_2^2(\rho_n^m, \rho^*)$ (cf. Corollary 3.5). However, both algorithm has similar bias at convergence, for which we conjecture that the particle discretization error of ρ_n^m dominates $\mathcal{W}_2^2(\rho_n^m, \tilde{\rho}_\tau^*)$ while close to convergence. We validate our conjecture through further experiments and discussions in Appendix D.1.

5 Conclusion

In this work, we provided a convergence analysis of the Wasserstein proximal algorithm without assuming any geodesic convexity, which improves upon the existing rates when strong geodesic convexity indeed holds. We also analyzed the inexact gradient variant under an extra geodesic semiconvexity condition. Applying to the proximal training of mean-field neural networks, linear convergence of the entropy-regularized total objective is guaranteed, which is faster than the noisy gradient descent algorithm as observed in our empirical experiments. One future work would be the study of particle discretization effect of the Wasserstein proximal algorithm in the setting of MFLD (Kook et al., 2024; Fu & Wilson, 2024).

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below. *Inventiones mathematicae*, 195(2):289–391, February 2013. ISSN 1432-1297. doi: 10.1007/s00222-013-0456-1. URL http://dx.doi.org/10.1007/s00222-013-0456-1.

Siwan Boufadène and François-Xavier Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. 2023.

Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field Langevin dynamics, 2023. URL https://arxiv.org/abs/2212.03050.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2984–3014. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/chen22c.html.

Xiuyuan Cheng, Jianfeng Lu, Yixin Tan, and Yao Xie. Convergence of flow-based generative models via proximal gradient descent in Wasserstein space. *IEEE Transactions on Information Theory*, pp. 1–1, 2024. doi: 10.1109/TIT.2024.3422412.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pp. 1276–1304. PMLR, 2020.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Matthew S Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *Foundations of Computational Mathematics*, pp. 1–51, 2024.

Lénaïc Chizat. Mean-field Langevin dynamics: exponential convergence and annealing, 2022. URL https://arxiv.org/abs/2202.01009.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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

Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1473–1521. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/fan23a.html.

Qiang Fu and Ashia Camage Wilson. Mean-field underdamped Langevin dynamics and its space-time discretization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14175–14206. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fu24g.html.

Soumyadip Ghosh, Yingdong Lu, Tomasz Nowicki, and Edith Zhang. On representations of mean-field variational inference, 2022. URL https://arxiv.org/abs/2210.11385.

Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 57, pp. 2043–2065. Institut Henri Poincaré, 2021.

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 795–811, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46128-1.

Vilmos Komornik. *Lectures on Functional Analysis and the Lebesgue Integral*, volume 2. Springer, 2016.

Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite radon measures. 2016.

Yunbum Kook, Matthew S. Zhang, Sinho Chewi, Murat A. Erdogdu, and Mufan (Bill) Li. Sampling from the mean-field stationary distribution. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3099–3136. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/kook24a.html.

Lyudmyla F Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022.

Jiaming Liang and Yongxin Chen. Proximal oracles for optimization and sampling, 2024. URL https://arxiv.org/abs/2404.02239.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit, 2019.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-scale Wasserstein gradient flows. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 9741–9757. PMLR, 2022.

Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein proximal gradient algorithm, 2021. URL https://arxiv.org/abs/2002.03035.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer, 2015.

Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices, 2022. URL https://arxiv.org/abs/1903.08568.

Andre Wibisono. Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2093–3027. PMLR, 06–09 Jul 2018.

Andre Wibisono. Proximal Langevin algorithm: rapid convergence under isoperimetry, 2019. URL https://arxiv.org/abs/1911.01469.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270): 1–63, 2022. URL http://jmlr.org/papers/v23/21-1184.html.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by JKO scheme. *Advances in Neural Information Processing Systems*, 36, 2024.

Rentian Yao and Yun Yang. Mean-field variational inference via Wasserstein gradient flow, 2023. URL https://arxiv.org/abs/2207.08074.

Rentian Yao, Xiaohui Chen, and Yun Yang. Mean-field nonparametric estimation of interacting particle systems. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2242–2275. PMLR, 02–05 Jul 2022.

Rentian Yao, Xiaohui Chen, and Yun Yang. Wasserstein proximal coordinate gradient algorithms, 2024. URL https://arxiv.org/abs/2405.04628.

Shunshi Zhang, Sinho Chewi, Mufan Li, Krishna Balasubramanian, and Murat A Erdogdu. Improved discretization analysis for underdamped Langevin Monte Carlo. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 36–71. PMLR, 2023.

A Backgound on Optimal Transport and Wasserstein Space

A.1 Wasserstein distance and optimal transport

The squared 2-Wasserstein distance is defined as the solution to the Kantorovich problem

$$\mathcal{W}_2^2(\rho, \tilde{\rho}) := \min_{\pi \in \Pi(\rho, \tilde{\rho})} \int_{\Theta \times \Theta} \|\theta - \tilde{\theta}\|^2 d\pi(\theta, \tilde{\theta}),$$

where $\Pi(\rho, \tilde{\rho}) \subset \mathcal{P}_2(\Theta \times \Theta)$ is the set of all coupling distributions with marginals ρ and $\tilde{\rho}$. The optimal solution π^* is called the optimal transport plan. When $\rho \in \mathcal{P}_2^a(\Theta)$, it is known from Brenier's theorem that the solution $T_\rho^{\tilde{\rho}}$ of Monge's problem exists, and the optimal transport plan is $\pi^* = (\mathbf{id}, T_\rho^{\tilde{\rho}})_\# \rho$.

A.2 Wasserstein Subdifferential and μ-convex functionals

Definition A.1 (Frechet subdifferential, [Definition 10.1.1 Ambrosio et al. (2005)])**.** *Let $F : \mathcal{P}_2(\Theta) \to (-\infty, +\infty]$ be proper, lower semicontinuous, and let $\rho \in D(|\partial F|)$ where $|\partial F|(\rho)$ denotes the metric slope (Ambrosio et al., 2005). We say that $\mathbf{v} \in L^2(\rho; \Theta)$ belongs to the Frechet subdifferential $\partial F(\rho)$ of F, written as $\mathbf{v} \in \partial F(\rho)$, if*

$$F(\tilde{\rho}) \geq F(\rho) + \int_\Theta \langle \mathbf{v}(\theta), T_\rho^{\tilde{\rho}}(\theta) - \theta \rangle d\rho(\theta) + o(\mathcal{W}_2(\rho, \tilde{\rho})).$$

And $\mathbf{v}$ is called strong subdifferential. Also, $\mathbf{v} \in \partial F(\rho)$ with minimal $\|\cdot\|_{L_2(\rho)}$ norm is denoted as $\partial^\circ F(\rho)$, and it is unique, see [Lemma 10.1.5, Ambrosio et al. (2005)].

Definition A.2 (First variation)**.** *Let $F : \mathcal{P}_2(\Theta) \to (-\infty, +\infty]$ be proper, lower semicontinuous. Let $\rho \in \mathcal{P}_2(\Theta)$, the first variation $\frac{\delta F}{\delta \rho}(\rho) : \Theta \to \mathbb{R}$ exists if,*

$$\frac{d}{d\varepsilon} F(\mu + \varepsilon \chi)\Big|_{\varepsilon=0} = \int \frac{\delta F}{\delta \rho}(\rho) d\chi$$

for any perturbation $\chi = \tilde{\rho} - \rho$ with $\tilde{\rho} \in \mathcal{P}_2(\Theta)$.

Definition A.3 (μ-convexity along geodesic, [Section 10.1.1, Ambrosio et al. (2005)])**.** *A proper, lower semicontinuous functional F is said to be μ-convex on geodesic ($\mu \in \mathbb{R}$) at $\rho \in \mathcal{D}(F) \subset \mathcal{P}_2^a(\Theta)$, if for all $\tilde{\rho} \in \mathcal{D}(F)$,*

$$F(\tilde{\rho}) \geq F(\rho) + \int_\Theta \langle \mathbf{v}(\theta), T_\rho^{\tilde{\rho}}(\theta) - \theta \rangle d\rho(\theta) + \frac{\mu}{2} \mathcal{W}_2^2(\rho, \tilde{\rho}), \tag{18}$$

*where $\mathbf{v} \in \partial F(\rho)$. In particular, if $\mu \geq 0$, we call F is **geodesically convex**; If $\mu < 0$, we call F is **geodesically semiconvex**. We refer to [Section 9, Ambrosio et al. (2005)] for definitions of μ-convexity (and semiconvexity) in Euclidean space, which is similar to the definition above.*

Definition A.4 (Weak Convergence)**.** *Let $\mathcal{C}_b(\Theta)$ be the set of all continuous bounded functions on Θ and $\mathcal{M}(\Theta)$ be the set of all finite signed measures on Θ. We say that $\rho_k \in \mathcal{M}(\Theta)$ converges to $\rho \in \mathcal{M}(\Theta)$ weakly if for every $\phi \in \mathcal{C}_b(\Theta)$,*

$$\lim_{k \to \infty} \int_\Theta \phi d\rho_k = \int_\Theta \phi d\rho.$$

The weak convergence is also called narrow convergence in the literature Santambrogio (2015).

B Technical Lemmas

Lemma B.1 (Proposition 3.1 and 3.3, Ambrosio et al. (2013))**.** *Let (Z, d) be a general metric space, $z \in Z$ such that $f(z) < +\infty$, the Hopf-Lax semigroup is defined as*

$$u(z, \xi) = \inf_{z' \in Z} f(z') + \frac{1}{2\xi} d(z', z)^2.$$

We define

$$D_+(z,\xi) := \sup \limsup_{n\to\infty} d(z, z_n'), \qquad D_-(z,\xi) := \inf \liminf_{n\to\infty} d(z, z_n'),$$

where the supremum and the infimum run among all minimizing sequences (z_n')*. For* $z \in Z$ *such that* $f(z) < +\infty$*, we define* $t^*(z) = \sup\{t > 0 : u(t,\xi) > -\infty\}$*. Then if* $f(z) < +\infty$*, we have*

(a) $D_+(z,\xi) = D_-(z,\xi)$ *holds for* $\xi \in (0, t^*(z))$ *except for at most countable exceptions;*

(b) If and only if $D_+(z,\xi) = D_-(z,\xi)$*, the map* $\xi \to u(z,\xi)$ *is differentiable in* $(0, t^*(z))$ *and*

$$\frac{d}{d\xi} u(z,\xi) = -\frac{D_+(z,\xi)}{2\xi^2} = -\frac{D_-(z,\xi)}{2\xi^2}.$$

Lemma B.2 (Existence of minimizer)**.** *If* $F : \mathcal{P}_2(\Theta) \to (-\infty, \infty]$ *is weakly lower semicontinuous, then proximal algorithm (2) admits a minimizer.*

Proof. The proof is essentially contained in Section 10.1 of Ambrosio et al. (2005). For the sake of completeness, we provide a proof here. We only need to show $\mathcal{B}_r(\rho) = \{v | \mathcal{W}_2^2(\rho, v) < r\}$ is weakly-precompact for any fixed $r > 0$. By Prokhorov's theorem, it suffices to prove the tightness of $\mathcal{B}_r(\rho)$, i.e., there is a sequence of compact sets $(K_i)_{i\in\mathbb{N}}$ such that

$$v(\Theta \backslash K_i) \le 1/i, \forall v \in \mathcal{B}_r(\rho).$$

We prove for any $\varepsilon > 0$, we can find compact set K such that

$$v(\Theta \backslash K) \le \varepsilon, \forall v \in \mathcal{B}_r(\rho).$$

by contradiction. Assume there exists ε, for any compact K, there exists $v \in \mathcal{B}_r(\rho)$ such that $v(\Theta \backslash K) > \varepsilon$. As a singleton $\{\rho\}$ constitutes a tight family, we can find a compact set $K_{\rho,\varepsilon}$ such that

$$\rho(K_{\rho,\varepsilon}) < \varepsilon.$$

We define a compact set $U_R = \{\theta' | \|\theta - \theta'\|^2 \le R, \forall \theta \in K_{\rho,\varepsilon}\}$. Let $R = 3r/\varepsilon$, then there exists v such that $v(\Theta \backslash U_{3r/\varepsilon}) > \varepsilon$. However, $\mathcal{W}_2^2(\rho, v) \ge 3r$, contradiction. Furthermore, note that since Θ is not compact, $\mathcal{B}_r(\rho)$ under the Wasserstein metric is not compact. □

Lemma B.3 (Conditions on weakly lower semicontinuity)**.** *(a) If* $F : \mathcal{P}_2(\Theta) :\to (-\infty, +\infty]$ *is proper, lower semicontinuous (with respect to* $\mathcal{W}_2$ *topology), and bounded from below, then* F *is weakly lower semicontinuous.*

(b) If $F : \mathcal{P}_2(\Theta) :\to (-\infty, +\infty]$ *is proper, lower semicontinuous (with respect to* $\mathcal{W}_2$ *topology), and* μ*-convex (*$\mu \in \mathbb{R}$*) along generalized geodesics , then* F *is weakly lower semicontinuous.*

Proof of Lemma B.3. See Section 10.3 in Ambrosio et al. (2005). □

Remark B.4. *We give several examples that the functional* F *satisfies either (a) or (b) in Lemma B.3.*

- *If* $F = \int f d\rho$ *where* $f : \mathbb{R}^d \to (-\infty, +\infty]$ *is lower semicontinuous and the negative part of* f *has a 2-growth condition (or bounded from below), then Part (a) of Lemma B.3 implies that* F *is weakly lower semicontinuous, [Section 9.3, Ambrosio et al. (2005)].*
- *If* $F = \int f d\rho$ *where* $f : \mathbb{R}^d \to (-\infty, +\infty]$ *is lower semicontinuous and* μ*-convex (*$\mu \in \mathbb{R}$*) in Euclidean sense, then Part (b) of Lemma B.3 implies that* F *is weakly lower semicontinuous.*

For conditions that ensure the weakly lower semicontinuity of internal energy, we refer to [Section 9.3, Ambrosio et al. (2005)] for details. Specifically, $\int \rho log \rho$ *is weakly lower semicontinuous.*

Lemma B.5 (Satisfication of Assumption 2(a) on $\mathcal{P}_2(\Theta)$ with compact set Θ)**.** *Under Assumption 1, if* Θ *is compact, then Assumption 2(a) holds.*

Proof. By [Proposition 7.17, Santambrogio (2015)], when Θ is compact, for fixed $\rho \in \mathcal{P}_2^a(\Theta)$, the first variation of $\mathcal{W}_2(\cdot, \rho)$ is well-defined for all $\tilde{\rho} \in \mathcal{P}_2^a(\Theta)$. Similar to [Proposition 8.7, Santambrogio (2015)], by standard calculus of variation followed by gradient operation, we have

$$\nabla \frac{\delta F}{\delta \rho}(\rho_\xi) = \frac{T_{\rho_\xi}^{\rho} - \mathbf{id}}{\xi}.$$

See [Lemma B.1, Yao et al. (2024)] for similar arguments. If Θ is not compact, the first variation of $\mathcal{W}_2$ distance is not well defined. □

Lemma B.6. *Assume $u : [0, \xi] \to (0, +\infty)$ is a decreasing function, $\partial_\xi u(\xi) \leq -C(\xi)u(\xi)$ almost everywhere for $\xi \in [0, \xi]$, $C(\xi) > 0$ for every $\xi \in [0, \xi]$, then*

$$u(\xi) \leq u(0) \exp\left(\int_0^\xi -C(\xi)d\xi\right).$$

Remark B.7. *Lemma B.6 extends the classical Gronwall lemma, which requires everywhere differentiability, to the case with almost everywhere differentiability and monotonicity.*

Proof. We construct a function $g(\xi) = \ln(u(\xi))$. It is a decreasing function and $\frac{dg(\xi)}{d\xi} = \frac{\partial_\xi u(\xi)}{u(\xi)} \leq -C(\xi)$ almost everywhere on $[0, \xi]$. By properties of Lebesgue integral, we have

$$\int_0^\xi \frac{dg(\xi)}{d\xi} d\xi \leq \int_0^\xi -C(\xi)d\xi.$$

Since g is decreasing, by [Proposition 6.6, Komornik (2016)],

$$g(\xi) - g(0) \leq \int_0^\xi \frac{dg(\xi)}{d\xi} d\xi.$$

Thus,

$$u(\xi) \leq u(0) \exp\left(\int_0^\xi -C(\xi)d\xi\right).$$

□

Lemma B.8. *If $T_{\overline{\rho}_n}^{\overline{\rho}_{n+1}}$ is C^2 diffeomorphism, then if $\overline{\rho}_n \in \mathcal{C}^1(\Theta)$, then $\overline{\rho}_{n+1} \in \mathcal{C}^1(\Theta)$.*

Proof. The change variable formula of probability density is,

$$T_\# \rho(\theta) = \rho(T^{-1}(\theta)) \cdot |detD(T^{-1})(\theta)|$$

where $D(T^{-1})$ is the jacobian matrix of T^{-1}. Since T is a $\mathcal{C}^2$ diffeomorphism, then T^{-1} is $\mathcal{C}^2$ mapping. Thus, $\rho \circ (T^{-1})$ is $\mathcal{C}^1(\Theta)$. Since T^{-1} is diffeomorphism, then $D(T^{-1})$ is not singular and $|detD(T^{-1})|$ is $C^1(\Theta)$. And thus $T_\# \rho$ is $\mathcal{C}^1(\Theta)$. □

Remark B.9. *We have an example from normalizing flow that can be $\mathcal{C}^2$ diffeomorphism. Real NVP (Dinh et al., 2016) has the following structure,*

$$T(\theta_{1:\tilde{d}}) = \theta_{1:\tilde{d}}$$
$$T(\theta_{\tilde{d}+1:d}) = \theta_{\tilde{d}+1:d} * exp(s(\theta_{1:\tilde{d}}) + t(\theta_{1:\tilde{d}}))$$

where $*$ refers to the pointwise product. T is naturally reversible (The Jacobian matrix is always non-singular) and the reverse is,

$$T^{-1}(\theta_{1:\tilde{d}}) = \theta_{1:\tilde{d}}$$
$$T^{-1}(\theta_{\tilde{d}+1:d}) = (\theta_{\tilde{d}+1:d} - t(\theta_{1:\tilde{d}})) * exp(-s(\theta_{1:\tilde{d}}))$$

It is not hard to see that if $s(\cdot) : \mathbb{R}^{\tilde{d}} \to \mathbb{R}^{d-\tilde{d}}$ and $t(\cdot) : \mathbb{R}^{\tilde{d}} \to \mathbb{R}^{d-\tilde{d}}$ are $\mathcal{C}^2$ maps (i.e. represented by fully neural network with smooth activation function), then T is restricted to be $\mathcal{C}^2$ diffeomorphism.

Lemma B.10. *Under Assumption 3, if $\overline{\rho}_n \in \mathcal{C}^1(\Theta)$ and $\overline{\rho}_{n+1}$ is the exact solution of the Wasserstein proximal algorithm (2), then $\beta_{n+1} = 0$.*

Proof. Since $\overline{\rho}_{n+1}$ is the exact solution of (2), then $\overline{\rho}_{n+1} \in D(|\partial F|)$ and $\frac{T_{\overline{\rho}_{n+1}}^{\overline{\rho}_n} - \mathbf{id}}{\xi} \in \partial F(\overline{\rho}_{n+1})$ by [Lemma 10.1.2, Ambrosio et al. (2005)]. As $\overline{\rho}_{n+1} \in \mathcal{C}^1(\Theta)$, then $\nabla \frac{\delta F}{\delta \rho}(\overline{\rho}_{n+1})$ is the unique strong subdifferential by [Lemma 10.4.1, Ambrosio et al. (2005)]. □

C PROOFS

Proof of Lemma 3.1. Lemma B.2 guarantees the existence of a minimizer of the JKO scheme or proximal algorithm. Thus we can define

$$z_\xi = \arg\min_{z\in\mathcal{P}_2(\mathbb{R}^d)} u(z,\xi).$$

Then Lemma B.1 implies that $D_+(z,\xi) = D_-(z,\xi) = d(z,z_\xi)$ in $(0,t^*(z))$ except for at most countable exceptions, from which we can conclude the desired Lemma 3.1. □

Definition C.1 (Uniform log-Sobolev inequality)**.** *There is a constant $\mu > 0$ such that for any $\rho \in \mathcal{P}_2(\Theta)$, its Gibbs proximal distribution π_ρ defined as,*

$$\pi_\rho(\theta) \propto \exp\left(-\frac{1}{\tau}\frac{\delta R(\rho)}{\delta\rho}(\theta)\right)$$

satisfies the log-Sobolev inequality (7) with the constant μ.

Proof of Corollary 3.5. We divide our proof into four parts,
(1) Firstly, we prove the satisfaction of Assumption 1 and the geodesic semiconvexity of F_τ;
(2) Secondly, we prove the PL inequality;
(3) Thirdly, we prove the satisfaction of Assumption 2(b);
(4) With the previous three parts, we can get the linear convergence of function value. The last part is devoted to obtain a bound on $\mathcal{W}_2$ distance using some structure of MFLD.

Our proof is as follows,
(1) The weakly lower semicontinuity of F_τ is verified in [Section 5.1, Chizat (2022)]. The geodesic semiconvexity follows [Lemma A.2, Chizat (2022)] and the proof actually relies on (19) below.

(2) Now we prove PL inequality. Since the training risk $R : \mathcal{P}_2(\Theta) \to (-\infty,+\infty]$ has linear convexity if the loss function l is convex (in the Euclidean sense). By [Proposition 5.1, (Chizat, 2022)], the L^2-regularized training risk R in (4) satisfies the uniform LSI assumption [Assumption 3, Chizat (2022)]. Next, we shall show that these assumptions imply the relaxed PL-inequality defined in (13). By the entropy sandwich bound [Lemma 3.4, (Chizat, 2022)],

$$\tau D_{\mathrm{KL}}(\rho\|q_\rho) \geq F_\tau(\rho) - F_\tau(\rho^*).$$

Therefore,

$$\begin{aligned}\int_\Theta \|\nabla\frac{\delta F_\tau}{\delta\rho}(\rho)\|^2 d\rho &= \int_\Theta \|\nabla\frac{\delta R}{\delta\rho}(\rho) + \tau\nabla\log(\rho)\|^2 d\rho \\ &= \tau^2 J_{q_\rho}(\rho) \geq 2\mu_\tau\tau^2 D_{\mathrm{KL}}(\rho\|q_\rho) \geq 2\mu_\tau\tau(F_\tau(\rho) - F_\tau(\rho^*)),\end{aligned}$$

Thus, the functional F_τ satisfies the Wasserstein PL-inequality with parameter $\tau\mu_\tau$.

(3) Under Assumption 1, $(T^\rho_{\rho_\xi} - \mathbf{id})/\xi$ is a strong subdifferential at $\rho_\xi \in \mathcal{P}_2^a(\Theta)$, and $\rho_\xi \in D(|\partial F_\tau|)$ by [Lemma 10.1.2, Ambrosio et al. (2005)]. To prove that Assumption 2(b) holds, we only need to prove that,

$$\textit{If } \rho \in D(|\partial F_\tau|) \cap \mathcal{P}_2^a(\Theta), \textit{ then } \nabla\tfrac{\delta F_\tau}{\delta\rho}(\rho) = \partial^\circ F_\tau(\rho).$$

Our proof for (3) below highly relies on the proof of [Theorem 10.4.13, Ambrosio et al. (2005)].
Step 1. We first need to derive some conditions similar to [(10.4.58), (10.4.59), Ambrosio et al. (2005)]. Under the assumptions of Corollary 3.5, R satisfies the following smoothness condition with $\tilde{L} > 0$ by [Proposition 5.1, Chizat (2022)].

$$\forall\theta,\tilde{\theta}\in\Theta, \forall\rho,\tilde{\rho}\in\mathcal{P}_2(\Theta), \quad \|\nabla\frac{\delta R}{\delta\rho}(\rho)(\theta) - \nabla\frac{\delta R}{\delta\rho}(\tilde{\rho})(\tilde{\theta})\| \leq \tilde{L}(\|\theta-\tilde{\theta}\|_2 + \mathcal{W}_2(\rho,\tilde{\rho})). \tag{19}$$

By [Lemma A.2, Chizat (2022)], relying on (19), we have

$$\lim_{t\to 0}\frac{R((\mathbf{id}+t\mathbf{t})_\#\rho) - R(\rho)}{t} = \int\langle\nabla\frac{\delta R}{\delta\rho}(\rho), \mathbf{r}-\mathbf{id}\rangle d\rho. \tag{20}$$

(20) is the key condition similar to [(10.4.58), (10.4.59), Ambrosio et al. (2005)] that we want to obtain.

Furthermore, a by-product is that (20) suggests that $\nabla \frac{\delta R}{\delta \rho}(\rho)$ is a (unique) strong subdifferential and thus $\|\nabla \frac{\delta R}{\delta \rho}(\rho)\|_{L_2(\rho)} = |\partial R(\rho)|$ for all $\rho \in \mathcal{P}_2^a(\Theta)$, see [Lemma A.2, Chizat (2022)]. It is not hard to verify that $\|\nabla \frac{\delta R}{\delta \rho}(\rho)\|_{L_2(\rho)}$ is finite at any $\rho \in \mathcal{P}_2(\Theta)$ because (19) ensures 2-growth of $\|\nabla \frac{\delta R}{\delta \rho}(\rho)(\cdot)\|^2$.

In **Step 2**, we conduct a proof similar to [Theorem 10.4.13, Ambrosio et al. (2005)].

Step 2. By [Lemma 10.4.4, Ambrosio et al. (2005)], (20), and the fact that $\rho \in \mathcal{P}_2^a(\Theta) \cap D(|\partial F|)$, for $\mathbf{t} \in C_c^\infty(\Theta)$

$$-\int \tau \rho \nabla \cdot \mathbf{t} d\theta + \int \langle \nabla \frac{\delta R}{\delta \rho}(\rho), \mathbf{t} \rangle d\rho \geq -|\partial F_\tau(\rho)| \|\mathbf{t}\|_{L_2(\rho)}.$$

By (21), $\forall \theta, \tilde{\theta} \in \Theta, \forall \rho \in \mathcal{P}_2(\Theta)$,

$$\|\nabla \frac{\delta R}{\delta \rho}(\rho)(\theta) - \nabla \frac{\delta R}{\delta \rho}(\rho)(\tilde{\theta})\| \leq \tilde{L}(\|\theta - \tilde{\theta}\|_2). \tag{21}$$

Therefore, $\nabla \frac{\delta R}{\delta \rho}(\rho)(\cdot)$ is Lipschitz and locally bounded. Therefore, following the same argument of [Theorem 10.4.13, Ambrosio et al. (2005)], we obtain that

$$\nabla \frac{\delta F_\tau}{\delta \rho}(\rho) = \tau \frac{\nabla \rho}{\rho} + \nabla \frac{\delta R}{\delta \rho}(\rho) \in L_2(\rho) \text{ and } \|\nabla \frac{\delta F_\tau}{\delta \rho}(\rho)\|_{L_2(\rho)} \leq |\partial F_\tau(\rho)|.$$

where we define $\frac{\nabla \rho}{\rho} = 0$ if $\rho(\theta) = 0$. Furthermore, it is straightforward to prove $\nabla \frac{\delta F_\tau}{\delta \rho}(\rho) \in \partial F_\tau(\rho)$.[1] Since $\rho \in D(|\partial F_\tau|)$ and $\rho \in D(|\partial R|)$, we have $\rho \in D(|\partial H_\tau|)$ where $H_\tau(\rho) = \int \tau \rho log \rho$. Therefore, $\tau \frac{\nabla \rho}{\rho} \in \partial H_\tau(\rho)$ [Theorem 10.4.6, Ambrosio et al. (2005)]. With $\nabla \frac{\delta R}{\delta \rho}(\rho) \in \partial R(\rho)$, we have proved that $\nabla \frac{\delta F_\tau}{\delta \rho}(\rho) \in \partial F_\tau(\rho)$. Thus, $\nabla \frac{\delta F_\tau}{\delta \rho}(\rho) = \partial^\circ F_\tau(\rho)$.

(4) With the previous three parts, we can get (14) with Theorem 3.4. Next, we prove (15). Using [Lemma 3.4, (Chizat, 2022)] once again, we get

$$\tau D_{\mathrm{KL}}(\rho \| \rho^*) \leq F_\tau(\rho) - F_\tau(\rho^*) \leq \tau D_{\mathrm{KL}}(\rho \| q_\rho).$$

Since ρ^* also satisfies μ_τ-LSI, by Talagrand's inequality, we have

$$\mathcal{W}_2^2(\rho^*, \rho_n) \leq \frac{2}{\mu_\tau} D_{\mathrm{KL}}(\rho_n \| \rho^*) \leq \frac{2}{\mu_\tau \tau}(F_\tau(\rho_n) - F_\tau(\rho^*)) \leq \frac{2}{\mu_\tau \tau}(F_\tau(\rho_0) - F_\tau(\rho^*))(\frac{1}{1 + \mu_\tau \tau \xi})^{2n}.$$

□

Proof of Corollary 3.6. Similar to proof of Corollary 3.5, we divide the proof into four parts.

(1) If f is semiconvex and lower-semicontinuous, then $\int f d\rho$ is weakly lower semicontinuous by Lemma B.3. In addition, $\int \log \rho d\rho$ is also weakly lower semicontinuous by Lemma B.3. Thus, F is weakly lower continuous, and Assumption 1 satisfies.

Furthermore, $\int f d\rho$ is semiconvex along geodesics and $\int \log \rho d\rho$ is convex along geodesics, see [Section 9.3, Ambrosio et al. (2005)]. Thus, F is semiconvex along geodesics.

(2) The PL inequality directly follows from LSI condition.

(3) We prove that Assumption 2(b) is satisfied. [Theorem 10.4.13, Ambrosio et al. (2005)] already incorporates KL divergence as a special case: Assume f is semiconvex and lower semicontinuous, and $\{\theta | V(\theta) < \infty\}$ is not empty. Then for $F(\rho) = \int f d\rho + \int \rho log \rho$ (here we assume the interaction

[1] Here the proof is slightly different from [Theorem 10.4.13, Ambrosio et al. (2005)] for simplicity, as we already have the finite slope property of R.

energy to be 0), if $\rho \in D(|\partial F|) \cap \mathcal{P}_2^a(\Theta)$, then $\nabla \frac{\delta F}{\delta \rho}(\rho) = \nabla V + \nabla \log(\rho) = \partial^\circ F(\rho)$, where we assume $\nabla log(\rho)(\theta) = 0$ if $\rho(\theta) = 0$. With Assumption 1 to guarantee $\rho_\xi \in \mathcal{P}_2^a(\Theta)$ and [Lemma 10.1.2, Ambrosio et al. (2005)] to guarantee that $\rho_\xi \in D(|\partial F|)$, Assumption 2(b) holds.

(4) The convergence rate on $\mathcal{W}_2$ distance follows Talagrand inequality. □

Proof of Theorem 3.8. In this proof, we will not invoke Assumption 2.
Step 1. We want to prove that μ-geodesic convex implies, for any fixed ρ and for any ξ,

$$F(\rho^*) \geq F(\rho_\xi) - \frac{1}{2\mu\xi^2}\mathcal{W}_2^2(\rho_\xi, \rho). \tag{22}$$

Note $\dfrac{T_{\rho_\xi}^\rho - I}{\xi}$ is a strong subdifferential at ρ_ξ,

$$\begin{aligned} F(\tilde{\rho}) &\geq F(\rho_\xi) + \int_\Theta \langle \frac{T_{\rho_\xi}^\rho(\theta) - \theta}{\xi}, T_{\rho_\xi}^{\tilde{\rho}}(\theta) - \theta \rangle d\rho_\xi(\theta) + \frac{\mu}{2}\mathcal{W}_2^2(\rho_\xi, \tilde{\rho}) \\ &= F(\rho_\xi) + \int_\Theta \left(\langle \frac{T_{\rho_\xi}^\rho(\theta) - (\theta)}{\xi}, T_{\rho_\xi}^{\tilde{\rho}}(\theta) - \theta \rangle + \frac{\mu}{2}\|T_{\rho_\xi}^{\tilde{\rho}}(\theta) - \theta\|^2 \right) d\rho_\xi(\theta). \end{aligned}$$

Then, we minimize both sides of Eqn (18) with respect to $\tilde{\rho} \in \mathcal{P}_2^a(\Theta)$. Clearly, ρ^* minimizes the left side. For the integral term on the right-hand side, we define $T_{\rho_\xi}^{\tilde{\rho}}$ in the following,

$$T_{\rho_\xi}^{\tilde{\rho}}(\theta) - \theta = -\frac{1}{\mu\xi}\left(T_{\rho_\xi}^\rho(\theta) - \theta\right), \quad \rho_\xi\text{-}a.e.$$

then it minimizes the integral as it minimizes the term inside the integral almost everywhere. Therefore,

$$F(\rho^*) \geq F(\rho_\xi) - \frac{1}{2\mu\xi^2}\int_\Theta \|T_{\rho_\xi}^\rho(\theta) - \theta\|^2 d\rho_\xi(\theta) = F(\rho_\xi) - \frac{1}{2\mu\xi^2}W_2^2(\rho_\xi, \rho).$$

Step 2.

$$\begin{aligned} \partial_\xi(u(\rho,\xi) - F^*) &= -\frac{\mu}{2\xi(1+\mu\xi)}\mathcal{W}_2^2(\rho_\xi,\rho) - \frac{1}{2(1+\mu\xi)\xi^2}\mathcal{W}_2^2(\rho_\xi,\rho) && \text{(by (11))} \\ &\leq -\frac{\mu}{2\xi(1+\mu\xi)}\mathcal{W}_2^2(\rho_\xi,\rho) - \frac{\mu}{(1+\mu\xi)}(F(\rho_\xi) - F^*) && \text{(by (22) in Step 1)} \\ &= -\frac{\mu}{1+\mu\xi}(u(\rho,\xi) - F^*). && \text{(by Def 3.1)} \end{aligned}$$

Using Lemma B.6 to deal with the technique issue of almost everywhere differentiability, we have

$$u(\rho,\xi) - F^* \leq (u(\rho,0) - F^*)\exp\left(\int_0^\xi -\frac{\mu}{1+\mu t}dt\right) = (F(\rho) - F^*)\frac{1}{1+\mu\xi}.$$

Invoking (22) once again, we obtain that

$$(F(\rho) - F^*)\frac{1}{1+\mu\xi} \geq u(\rho,\xi) - F^* = F(\rho_\xi) - F^* + \frac{1}{2\xi}\mathcal{W}_2^2(\rho_\xi,\rho) \geq (1+\mu\xi)(F(\rho_\xi) - F^*).$$

Step 3. We derive a bound for $\mathcal{W}_2(\rho^*, \rho_n)$ here. Since for any $\rho \in \mathcal{D}(F)$,

$$F(\rho) - F(\rho^*) \geq \int \langle \mathbf{0}, T_{\rho^*}^\rho \rangle d\rho^*(\theta).$$

Therefore, $\mathbf{0}$ is a strong subdifferential by definition of geodesic convexity. Thus, we have

$$F(\rho) - F(\rho^*) \geq \int \langle \mathbf{0}, T_{\rho^*}^\rho \rangle d\rho^*(\theta) + \frac{\mu}{2}\mathcal{W}_2^2(\rho^*,\rho) = \frac{\mu}{2}\mathcal{W}_2^2(\rho^*,\rho).$$

by definition of μ-convexity along geodesics. Therefore,

$$\mathcal{W}_2(\rho^*, \rho_n) \leq \sqrt{\frac{2}{\mu}F(\rho_0) - F(\rho^*)}\frac{1}{(1+\mu\xi)^n}.$$

□

Proof of Theorem 3.10. It suffices to prove the case $\xi = 1/L$ because for $\xi < 1/L$, a functional that is $(-L)$ geodesically convex is also $-(1/\xi)$ geodesically convex. Under Assumption 3, since we assume $\overline{\rho}_{n+1} \in D(\partial F)$, then $\nabla \frac{\delta F}{\delta \rho}(\overline{\rho}_{n+1})$ is the unique strong subdifferential. Thus,

$$
\begin{aligned}
F(\overline{\rho}_{n+1}) - F(\overline{\rho}_n) &\leq -\xi \int_\Theta \langle \nabla \frac{\delta F}{\delta \rho}(\overline{\rho}_{n+1}), \frac{T_{\overline{\rho}_{n+1}}^{\overline{\rho}_n} - \mathbf{id}}{\xi} \rangle d\overline{\rho}_{n+1} + \frac{L}{2} \mathcal{W}_2^2(\overline{\rho}_{n+1}, \overline{\rho}_n) \\
&= -\xi \int_\Theta \langle \nabla \frac{\delta F}{\delta \rho}(\overline{\rho}_{n+1}), \nabla \frac{\delta F}{\delta \rho}(\overline{\rho}_{n+1}) + \frac{1}{\xi} \beta_{n+1} \rangle d\overline{\rho}_{n+1} \\
&\quad + \frac{L}{2} \int_\Theta \| \xi \nabla \frac{\delta F}{\delta \rho}(\overline{\rho}_{n+1}) + \beta_{n+1} \|^2 d\overline{\rho}_{n+1} && \text{(by Eqn 16)} \\
&= (-\frac{1}{2L}) \int_\Theta \| \nabla \frac{\delta F}{\delta \rho}(\overline{\rho}_{n+1}) \|^2 d\overline{\rho}_{n+1} + \frac{L}{2} \int_\Theta \| \beta_{n+1} \|^2 d\overline{\rho}_{n+1} && \text{(let } \xi = \frac{1}{L}\text{)} \\
&\leq -\frac{\mu}{L}(F(\overline{\rho}_{n+1}) - F(\rho^*)) + \frac{L}{2}(\epsilon_{n+1}). && \text{(by PL (13))}
\end{aligned}
$$

Thus,

$$
(1 + \frac{\mu}{L})(F(\overline{\rho}_{n+1}) - F^*) \leq (F(\overline{\rho}_n) - F^*) + \frac{L}{2}\epsilon_{n+1}.
$$

By [Lemma 14, (Yao et al., 2024)], we set $A = \frac{1}{1+\mu/L}$, $B = \frac{L/2}{1+\mu/(L)}$.

(a) If $\epsilon_n \leq C_{exp}\gamma^n$ with $\gamma \in (0, 1)$,

$$
F(\overline{\rho}_n) - F(\rho^*) \leq A^n(F(\rho_0) - F(\rho^*)) + \frac{BC_{exp}}{|A - \gamma|} max\{A, \gamma\}^{n+1}.
$$

(b) If $\epsilon_n \leq C_{poly} n^{-\zeta}$ with $\zeta > 0$, there is a constant $C(\zeta, A)$,

$$
F(\overline{\rho}_n) - F(\rho^*) \leq A^n(F(\rho_0) - F(\rho^*)) + \frac{BC(\zeta, A)\zeta}{n^\zeta}.
$$

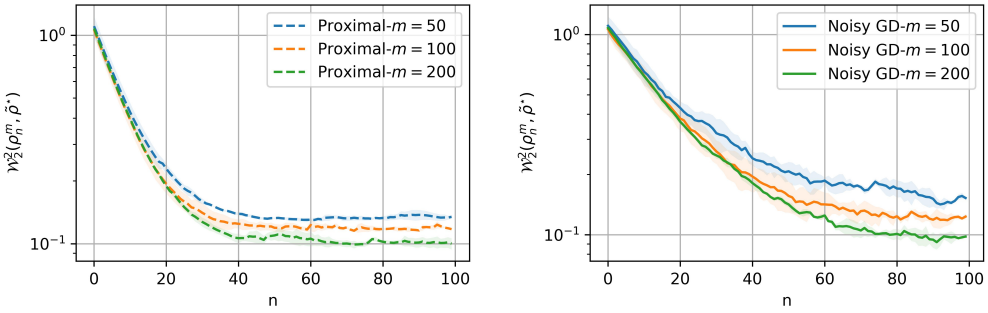
□

D Further Numerical Experiments and Discussions

D.1 Further discussions on MFLD Experiments

As shown in Figure 3, while m increases, $\mathcal{W}_2(\rho_n^m, \tilde{\rho}_\tau^*)$ at convergence decreases for both algorithms. This experiment supports our conjecture that partial discretization error dominates the bias, if we assume $\tilde{\rho}_\tau^*$ is a good approximate of ρ^*.

The particle discretization error is well studied for noisy gradient descent and a quantitative uniform-in-time propagation of chaos result for the MFLD has been established in (Chen et al., 2023). Specifically, the "distance" between the finite particle dynamics and the mean-field dynamics converges at rate $O(1/m)$ for all $t > 0$ under the uniform LSI condition. Note that even though the "distance" in theoretical analysis of uniform propagation chaos is not simply defined to be the $\mathcal{W}_2$ distance between empirical measure of finite particle dynamics and absolutely continuous measure of mean-field dynamics, it is empirically observed that the $\mathcal{W}_2$ distance also demonstrates similar propagation-of-chaos property in Figure 3b. The remaining challenge is how to theoretically characterize the particle discretization error of the proximal algorithm.



(a) (b)

Figure 3: Particle discretization error with different number of particles. We follow all the experiment settings in Section 4.2.