Mean-field Underdamped Langevin Dynamics and its Spacetime Discretization

Qiang Fu [1] **Ashia Wilson** [2]

Abstract

We propose a new method called the N-particle underdamped Langevin algorithm for optimizing a special class of non-linear functionals defined over the space of probability measures. Examples of problems with this formulation include training mean-field neural networks, maximum mean discrepancy minimization and kernel Stein discrepancy minimization. Our algorithm is based on a novel spacetime discretization of the mean-field underdamped Langevin dynamics, for which we provide a new, fast mixing guarantee. In addition, we demonstrate that our algorithm converges globally in total variation distance, bridging the theoretical gap between the dynamics and its practical implementation.

1. Introduction

The mean-field Langevin dynamics (MLD) has recently received renewed interest due to its connection to gradient-based techniques used in supervised learning problems such as training neural networks in a limiting regime (Mei et al., 2018). Several recent works have focused on the theoretical characterizations of the convergence properties of MLD in particular (Hu et al., 2019; Chizat, 2022; Nitanda et al., 2022; Chen et al., 2022; Claisse et al., 2023; Suzuki et al., 2023). More generally, MLD can be used to solve problems that can be posed as an entropy regularized mean-field optimization (EMO) problem. Other examples of such problems include density estimation via maximum mean discrepancy (MMD) minimization (Gretton et al., 2006; Arbel et al., 2019; Chizat, 2022; Suzuki et al., 2023) and sampling via kernel Stein discrepancy (KSD) minimization (Liu et al., 2016; Chwialkowski et al., 2016; Suzuki et al., 2023).

The MLD has also been the subject of a number of recent theoretical developments. Hu et al. (2019), for example, show that MLD finds EMO solutions asymptotically when problems can be expressed as optimizing a convex functional. If in addition, the EMO satisfies a uniform logarithmic Sobolev inequality, several studies have established that this convergence occurs exponentially quickly (Chizat, 2022; Nitanda et al., 2022; Chen et al., 2022). However, implementing MLD is not a straightforward task; to arrive at a practical algorithm requires both spatial and temporal discretizations of the dynamics. Nitanda et al. (2022) study a time-discretization of MLD by extending an interpolation argument introduced by Vempala & Wibisono (2019) to a non-linear Fokker-Planck equation. They establish a non-asymptotic rate of convergence for the discrete-time process. Chen et al. (2022) study a space-discretization consisting of a finite-particle approximation to the density of MLD (referred to as a finite-particle system) and show the finite-particle system finds the solution to the EMO problem exponentially fast, with a bias related to the number of particles. More practically, Suzuki et al. (2023) analyze a spacetime discretization of the MLD and establish the non-asymptotic convergence of the resulting algorithm to a biased limit related to both the number of particles used and stepsize. Their analysis applies to several important learning problems and improves the results of the standard gradient Langevin dynamics.

A natural candidate method for finding solutions to EMO problems faster is the mean-field *underdamped* Langevin dynamics (MULD). MULD resemble several techniques for adding momentum to gradient descent in optimization, many of which are known to result in provably faster convergence in a variety of settings (Nesterov, 1983; Wilson et al., 2016; Laborde & Oberman, 2020; Hinder et al., 2020; Fu et al., 2023; Srinivasan & Wilson, 2022). Moreover, training neural networks using momentum-based gradient descent is considered effective in several applications (Sutskever et al., 2013; Kingma & Ba, 2014; Ruder, 2016). Kazeykina et al. (2020) and Chen et al. (2023) confirm that a naive spacetime discretization of MULD has impressive empirical performance when compared to a naive discretization of the MLD on applications such as training mean-field neural networks.

Chen et al. (2023) introduce a space-discretization of MULD consisting of a finite particle approximation to the density

[1] Department of Computer Science, Yale University, New Haven, CT, USA [2] Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA. Correspondence to: Qiang Fu <qiang.fu@yale.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

and show it finds the EMO solution exponentially fast, albeit with several additional assumptions that are easy to verify for the problem of training mean-field neural networks. In addition, Chen et al. (2023) implement an Euler-Maruyama discretization of the finite-particle system and show that it performs empirically faster when compared with the space-time discretization of the mean-field Langevin dynamics in training a toy neural network model. However, space-time discretizations of MULD are not yet theoretically well understood. Furthermore, the rate obtained by Chen et al. (2023) for the dynamics does not resemble an "accelerated rate" when compared with recent results for MLD.

A remaining question is whether we can theoretically characterize the behavior of an implementable algorithm based on discretizing MULD. If so, and there is a limiting bias, how does it scale with the number of particles and other problem parameters? Ideally, these characterizations would give a sharper rate of convergence than Suzuki et al. (2023)'s space-time discretization of the mean-field Langevin dynamics, suggesting there might be an advantage to adding momentum in the mean-field setting (at least in the worst case).

Contributions We contribute the following:

1. We introduce the *N-particle underdamped Langevin algorithm* (N-ULA), a fast implementable algorithm based on the MULD for solving EMO problems. We prove that the N-ULA converges to a small limiting bias under a set of assumptions that subsumes many problems of interest.
2. We sharpen the convergence bound for MULD and its space-discretization established by Chen et al. (2023) under the same set of assumptions utilized by Chen et al. (2023) (Theorems 3.1 and 3.2 and Table 1).
3. We show the global convergence of N-ULA in total variation (TV) distance (Theorem 3.4). Importantly, our results improve on Suzuki et al. (2023)'s analysis of the spacetime discretization of the MLD. While we require additional assumptions 2.7-2.9, our results hold in several real-world applications including training neural networks, density estimation via MMD minimization and sampling via KSD minimization.

Organization Section 2 presents the formal definitions, assumptions and related work. Section 3 proposes our main methods and theoretical results. Section 4 discusses the application of our methods to some classical problems. Section 5 presents our numerical results.

2. Preliminaries

We begin by introducing general notation that will be used throughout this work.

2.1. Notation

The Euclidean and operator norms are denoted by $\|\cdot\|$ and $\|\cdot\|_{\mathsf{op}}$. The space of probability measures on $\mathbb{R}^d$ with finite second moment is denoted by $\mathcal{P}_2(\mathbb{R}^d)$. Throughout, let ρ and μ denote general distributions in $\mathcal{P}_2(\mathbb{R}^d)$ and $\mathcal{P}_2(\mathbb{R}^{2d})$ respectively. The TV distance between ρ and $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ is denoted by $\|\rho - \pi\|_{\mathsf{TV}} := \sup |\rho(A) - \pi(A)|$ where the sup is over all Borel measurable sets $A \subset \mathbb{R}^d$. The p-Wasserstein distance and Kullback-Leibler divergence between ρ and π is denoted by $W_p(\rho,\pi) := \inf_\Pi \mathbb{E}_\Pi[\|x-y\|^p]^{1/p}$ where the infimum is over joint distributions Π of (x,y) with the marginals $x \sim \rho, y \sim \pi$ and $\mathsf{KL}(\rho\|\pi) := \int \rho \log \frac{\rho}{\pi}$. The relative Fisher information is denoted by $\mathsf{FI}(\rho\|\pi) := \mathbb{E}_\rho \|\nabla \log \frac{\rho}{\pi}\|^2$, and more generally we use the notation $\mathsf{FI}_S(\rho\|\pi) := \mathbb{E}_\rho\|S^{1/2}\nabla \log \frac{\rho}{\pi}\|^2$ for a positive definite symmetric matrix S. $\mathrm{Ent}(\rho) := \int \rho \log \rho$ denotes the negative entropy of ρ. The functional and intrinsic derivatives of F are denoted by $\frac{\delta F}{\delta \rho} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ and $D_\rho F := \nabla \frac{\delta F}{\delta \rho} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$, respectively. A d-dimensional Brownian motion is denoted by B_t. We use notation $a \lesssim b$, $a_n = \Theta(b_n)$ and $a_n = \widetilde{\Theta}(b_n)$ to denote that there exist $c, C > 0$ such that $a \le Cb$, $cb_n \le a_n \le Cb_n$ for $n \ge N'$ and $a_n = \Theta(b_n)$ up to logarithmic factors, respectively.

2.2. Background

We consider the following problem described by minimizing the entropy regularized mean-field objective (EMO),

$$\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} F(\rho) + \lambda \mathrm{Ent}(\rho), \tag{1}$$

where $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is a potentially non-linear functional and $\lambda > 0$ is a regularization constant. As done in several prior works (Chen et al., 2022; 2023), we will take $\lambda = 1$ throughout and focus on other important problem parameters. Hu et al. (2019) study the gradient flow dynamics of the EMO in 2-Wasserstein metric called the *mean-field Langevin dynamics* (MLD):

$$\mathrm{d}x_t = -D_\rho F(\rho_t, x_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathrm{B}_t, \tag{MLD}$$

where $\rho_t := \mathrm{Law}(x_t) \in \mathcal{P}_2(\mathbb{R}^d)$. Under mild conditions, the MLD finds the solution to the EMO, given by $\rho_*(x) \propto \exp\left(-\frac{\delta F}{\delta \rho}(\rho_*, x)\right)$ (Hu et al., 2019).

This paper introduces a new sharp mixing-time bound for the *mean-field underdamped Langevin dynamics* (MULD):

$$\begin{aligned} \mathrm{d}x_t &= v_t \mathrm{d}t, \\ \mathrm{d}v_t &= -\gamma v_t \mathrm{d}t - D_\rho F(\mu_t^X, x_t)\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}\mathrm{B}_t. \end{aligned} \tag{MULD}$$

Here, $\mu_t := \mathrm{Law}(x_t, v_t) \in \mathcal{P}_2(\mathbb{R}^{2d})$, $\gamma > 0$ is the damping coefficient, and $\mu_t^X := \mathrm{Law}(x_t) = \int \mu_t(x,v)\mathrm{d}v$ is the X-marginal of μ_t. The limiting distribution of MULD is the solution to the augmented EMO problem,

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^{2d})} F(\mu^X)+\mathrm{Ent}(\mu)+\int\frac{1}{2}\|v\|^2\mu(\mathrm{d}x\mathrm{d}v), \quad (2)$$

where a momentum term is added to the EMO. The minimizer of the augmented EMO is given by $\mu_*(x,v)\propto\exp\left(-\frac{\delta F}{\delta\rho}(\mu_*^X,x)-\frac{1}{2}\|v\|^2\right)$. We provide details of the derivation of the limiting distributions of MLD and MULD in Appendices A.1 and A.3 respectively. To obtain the solution of the EMO problem, the minimizer $\mu_*(x,v)$ can be X-marginalized. This work also sharpens the analysis of the space-discretization of MULD introduced by Chen et al. (2023), which we refer to as the *N-particle underdamped Langevin dynamics* (N-ULD) for $i=1,...,N$:

$$\begin{aligned}\mathrm{d}x_t^i&=v_t^i\mathrm{d}t,\\ \mathrm{d}v_t^i&=-\gamma v_t^i\mathrm{d}t-D_\rho F(\mu_{\mathbf{x}_t},x_t^i)\mathrm{d}t+\sqrt{2\gamma}\mathrm{dB}_t^i,\end{aligned} \quad \text{(N-ULD)}$$

where $\mu_{\mathbf{x}_t}:=\frac{1}{N}\sum_{i=1}^N\delta_{x_t^i}$, $\mu_t^i:=\mathrm{Law}(x_t^i,v_t^i)$ and $(\mathrm{B}_t^i)_{i=1}^N$ are d-dimensional Brownian motions. Notably, a time-discretization of N-ULD is necessary to run the method on a machine. We explore two time-discretization techniques in this paper.

To motivate our algorithm as a time-discretization of N-ULD, we review discretizations of the *underdamped Langevin dynamics* (ULD), a special case of MULD where $F(\rho)=\int V(x)\rho(\mathrm{d}x)$ is a linear functional of μ:

$$\begin{aligned}\mathrm{d}x_t&=v_t\mathrm{d}t,\\ \mathrm{d}v_t&=-\gamma v_t\mathrm{d}t-\nabla V(x_t)\mathrm{d}t+\sqrt{2\gamma}\mathrm{dB}_t.\end{aligned} \quad \text{(ULD)}$$

The ULD was first studied in Kolmogoroff (1934) and Hörmander (1967). Under functional inequalities such as Poincaré's inequality on the target distribution $\rho_*\propto\exp(-V)$, the convergence guarantee of the ULD was studied by Villani using a hypocoercivity approach (Villani, 2001; 2009), but without capturing the acceleration phenomenon when compared to the overdamped Langevin dynamics. Cao et al. (2023) are the first to show ULD converges in χ^2-divergence at an accelerated rate when V is convex and the target distribution ρ_* satisfies LSI defined in (5) with $\mathscr{C}_{\mathsf{LSI}}>0$. They prove that when $\mathscr{C}_{\mathsf{LSI}}\ll 1$, the decaying rate of ULD is $O(\sqrt{\mathscr{C}_{\mathsf{LSI}}})$ whereas the decaying rate of the overdamped Langevin dynamics is $O(\mathscr{C}_{\mathsf{LSI}})$. We will refer to the time-discretization of ULD as an *underdamped Langevin Monte Carlo* (ULMC) algorithm. The *Euler-Maruyama* (EM) discretization (Kloeden et al., 1995; Platen & Bruti-Liberati, 2010) is perhaps the well-studied ULMC algorithm and it incurs the largest discretization error in several metrics including KL divergence and Wasserstein distance. Recently, however, several works have studied the ULMC obtained from a more precise discretization scheme called the the *exponential integrator* (EI) (Cheng et al., 2018):

$$\begin{aligned}\mathrm{d}x_t&=v_t\mathrm{d}t,\\ \mathrm{d}v_t&=-\gamma v_t\mathrm{d}t-\nabla V(x_{kh})\mathrm{d}t+\sqrt{2\gamma}\mathrm{dB}_t,\end{aligned} \quad \text{(EI-ULMC)}$$

for $t\in[kh,(k+1)h]$. Unlike the EM integrator, EI only fixes the drift term in each small interval, creating a group of linear stochastic differential equations (SDE) that can be exactly integrated. Leimkuhler et al. (2023) show that the EI incurs weaker stepsize restriction when compared with EM scheme. Other works have derived its convergence in Wasserstein distance (Cheng et al., 2018), KL divergence (Ma et al., 2021) and Rényi divergence (Zhang et al., 2023). Other discretization schemes are proposed in Shen & Lee (2019); Li et al. (2019); He et al. (2020); Foster et al. (2021); Monmarché (2021); Foster et al. (2022); Johnston et al. (2023), whose convergence guarantee are obtained in Wasserstein distance without achieving better dependence on terms such as the smoothness and LSI constants. In this work, we show that EI can be applied to discretize both MULD and N-ULD to achieve fast convergence.

2.3. Definitions and Assumptions

For each method considered, we study their behavior in settings where the minimizing distribution satisfies a Log-Sobolev inequality.

Definition 2.1 (LSI)**.** A measure $\pi\in\mathcal{P}_2(\mathbb{R}^d)$ satisfies Log-Sobolev Inequality (LSI) with parameter $\mathscr{C}_{\mathsf{LSI}}>0$, if for any $\rho\in\mathcal{P}_2(\mathbb{R}^d)$

$$\mathsf{KL}(\rho\|\pi)\le\frac{1}{2\mathscr{C}_{\mathsf{LSI}}}\mathsf{FI}(\rho\|\pi). \quad (5)$$

We also work with the following distribution $\hat{\mu}\in\mathcal{P}_2(\mathbb{R}^{2d})$ that appears in the Fokker-Planck equation (26) of MULD (see Appendix A.3). Note that the limiting distribution $\mu_*\in\mathcal{P}_2(\mathbb{R}^{2d})$ of MULD satisfies $\mu_*=\hat{\mu}_*$.

Definition 2.2. Throughout, we define the distribution $\hat{\mu}$ associated with the X-marginal of distribution μ and a functional F to be

$$\hat{\mu}(x,v)\propto\exp\left(-\frac{\delta F}{\delta\rho}(\mu^X,x)-\frac{1}{2}\|v\|^2\right). \quad (6)$$

We also introduce the same three assumptions on F as Chen et al. (2023) for establishing the non-asymptotic convergence of the MULD and N-ULD.

Assumption 2.3 (Convexity)**.** F is convex in the linear sense, which means for any $\rho_1,\rho_2\in\mathcal{P}_2(\mathbb{R}^d)$ and $t\in[0,1]$ the functional satisfies

$$F(t\rho_1+(1-t)\rho_2)\le tF(\rho_1)+(1-t)F(\rho_2). \quad (7)$$

Assumption 2.4 ($\mathscr{L}$-smoothness). F is smooth, which means the intrinsic derivative exists and for any $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R}^d)$, $x_1, x_2 \in \mathbb{R}^d$ and some $1 \leq \mathscr{L} < \infty$ satisfies

$$\begin{aligned} &\|D_\rho F(\rho_1, x_1) - D_\rho F(\rho_2, x_2)\| \\ &\qquad \leq \mathscr{L}(W_1(\rho_1, \rho_2) + \|x_1 - x_2\|) \end{aligned} \tag{8}$$

Assumption 2.5 (LSI). The distribution (6) satisfies LSI with constant $0 < \mathscr{C}_{\mathsf{LSI}} \leq 1$ for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

The X-marginal of distribution (6), which is related to the optimization gap, was first utilized by Nitanda et al. (2022) to establish convergence of MLD. Note that if $\hat{\mu}^X(x) \propto \exp(-\frac{\delta F}{\delta \rho}(\mu^X, x))$ satisfies LSI for any $\mu \in \mathcal{P}_2(\mathbb{R}^{2d})$ with constant $\tau > 0$, then Assumption 2.5 is satisfied with the choice $\mathscr{C}_{\mathsf{LSI}} = \min\{1/2, \tau\}$. We refer our readers to Chen et al. (2022; 2023); Suzuki et al. (2023) for the verification of Assumptions 2.3 and 2.5 in a variety of settings. Suzuki et al. (2023) consider a weaker smoothness assumption than Assumption 2.4 where they use W_2 distance in place of W_1 distance. They verify smoothness in W_2 distance for three examples including training mean-field neural networks, MMD minimization and KSD minimization, whereas Chen et al. (2022) verify smoothness in W_1 distance only for the example of training mean-field neural networks. In this paper, we verify $\mathscr{L}$-smoothness in W_1 distance (Assumption 2.4) for the other two examples (see Section C.1). Beyond Assumptions 2.3-2.5, we introduce four additional assumptions that are sufficient for our spacetime discretization analysis.

Assumption 2.6 (Bounded Gradient). For any $\rho \in \mathcal{P}_2(\mathbb{R}^d)$, the intrinsic derivative of F satisfies (where $\mathscr{L} > 0$)

$$\|D_\rho F(\rho, x)\| \leq \mathscr{L}(1 + \|x\|). \tag{9}$$

Notably, Suzuki et al. (2023) assume that F can be decomposed as $F(\rho) = U(\rho) + \mathbb{E}_{x\sim\rho}[r(x)]$ where $\|D_\rho U(\rho, x)\| \leq R$ for any $\rho \in \mathcal{P}(\mathbb{R}^d)$, $x \in \mathbb{R}^d$, and where $r(x)$ is a differentiable function satisfying $\|\nabla r(x) - \nabla r(y)\| \leq \lambda_2\|x - y\|$ with $\nabla r(0) = 0$ in order to establish the convergence of their spacetime discretization of MLD. Thus, their assumption that $\|D_\rho F(\rho, x)\| \leq \|D_\rho U(\rho, x)\| + \|\nabla r(x)\| \leq R + \lambda_2\|x\|$ implies Assumption 2.6 holds with the choice $\mathscr{L} \geq \max\{R, \lambda_2\}$. The next three assumptions are needed for bounding the second moment of the iterates $(x_t, v_t)_{t\geq 0}$ and $(x_t^i, v_t^i)_{t\geq 0}$ along MULD and N-ULD, which is crucial for the establishment of our discrete-time convergence.

Assumption 2.7. For all $\mu \in \mathcal{P}_2(\mathbb{R}^{2d})$, the distribution (6) given F satisfies $\mathbb{E}_{\hat{\mu}}\|\cdot\|^2 \lesssim d$.

Assumption 2.8. The functional F and the initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^{2d})$ satisfy $F(\mu_0^X) \lesssim \mathscr{L}d$.

Assumption 2.9. The functional F and the initial distribution $\mu_0^N \in \mathcal{P}_2(\mathbb{R}^{2Nd})$ satisfy $\mathbb{E}_{\mathbf{x}_0\sim(\mu_0^X)^N} F(\mu_{\mathbf{x}_0}) \lesssim \mathscr{L}d$, where μ_0^N is the N-tensor product of μ_0 and $\mu_{\mathbf{x}_0} = \frac{1}{N}\sum_{i=1}^N \delta_{x_0^i}$ with $x_0^i \sim \mu_0^X$.

While Assumptions 2.7-2.9 are sufficient, they may not be necessary for the iterates to be bounded. Nevertheless, we argue these assumptions are not too restrictive by verifying them for three examples introduced above including training mean-field neural networks, MMD minimization and KSD minimization in Section 4.

2.4. Related Work

Techniques for establishing the continuous-time convergence of the mean-field underdamped systems and their space-discretization (N-particle systems) are centered around *coupling* and *hypocoercivity*. The latter one is also known as functional approaches (Villani, 2009). The coupling approach generally constructs a joint probability of the mean-field and N-particle systems to make the analytic comparison between them. Based on coupling approaches, Guillin et al. (2022); Bolley et al. (2010); Bou-Rabee & Schuh (2023) show convergence of the underdamped dynamics with mean-field interaction and its space-discretization. Duong & Tugaut (2018); Kazeykina et al. (2020) study the ergodicity of the MULD without a quantitative rate. Under the setting of small mean-field dependence, Kazeykina et al. (2020) show exponential contraction using coupling techniques in Eberle et al. (2019a;b). The functional approach (hypocoercivity) generally constructs appropriate Lyapunov functionals and studies how their values change along the dynamics. Based on hypocoercivity, Monmarché (2017); Guillin et al. (2021); Guillin & Monmarché (2021); Bayraktar et al. (2022) establish the exponential convergence of the mean-field underdamped systems and its propagation of chaos by constructing a suitable Lyapunov functional. Nevertheless, most of the works above only consider specific settings of MULD such as singular interactions and two-body interactions, which restricts the application to real-world problems. Setting $\gamma = 1$, Chen et al. (2023) establish the exponential convergence of MULD and N-ULD using the hypocoercivity technique in Villani (2009). Under Assumptions 2.3-2.5, they derive the convergence without restricting the size of interactions, which subsumes many settings above. Notably, the techniques of our Theorems 3.1 and 3.2 are adopted from Chen et al. (2023) based on hypocoercivity where we consider other choices of γ to improve the decaying rate of MULD and N-ULD established in Chen et al. (2023).

3. N-particle Underdamped Langevin Algorithm

Our first step is to establish the global convergence of the *mean-field underdamped Langevin algorithm* (MULA),

$$\begin{aligned} \mathrm{d}x_t &= v_t \mathrm{d}t, \\ \mathrm{d}v_t &= -\gamma v_t \mathrm{d}t - D_\rho F(\mu^X_{kh}, x_{kh})\mathrm{d}t + \sqrt{2\gamma}\mathrm{dB}_t, \end{aligned} \tag{MULA}$$

for stepsize h, $t \in [kh, (k+1)h]$ and $k = 1, ..., K$. Note that MULA is the EI time-discretization of the MULD, where each step will now require integrating from $t = kh$ to $t = (k+1)h$ for stepsize h. MULA is intractable to implement in most instances given we do not often have access to μ^X_{kh}. This prompts us to consider the particle approximation which uses $\mu_{\mathbf{x}_{kh}} = \frac{1}{N}\sum_{i=1}^N \delta_{x^i_{kh}}$ to approximate μ^X_{kh} where $(x^i_k)_{i=1}^N$ are iid samples from μ^X_k:

$$\begin{aligned} \mathrm{d}x^i_t &= v^i_t \mathrm{d}t, \\ \mathrm{d}v^i_t &= -\gamma v^i_t \mathrm{d}t - D_\rho F(\mu_{\mathbf{x}_{kh}}, x^i_{kh})\mathrm{d}t + \sqrt{2\gamma}\mathrm{dB}^i_t, \end{aligned} \tag{11}$$

for stepsize h, $t \in [kh, (k+1)h]$, $i = 1, ..., N$, $k \in \mathbb{N}$ and $\mu_{\mathbf{x}_{kh}} = \frac{1}{N}\sum_{i=1}^N \delta_{x^i_{kh}}$. Integrating the particle system (11) from $t = kh$ to $t = (k+1)h$ for stepsize h and $i = 1, ..., N$, we obtain our proposed Algorithm 1 which we refer to as the *N-particle underdamped Langevin algorithm* (N-ULA).

Algorithm 1 (N-ULA)

```
Require: $F$ satisfies Assumptions 2.3-2.7 and 2.9
Initialize $\mathbf{x}_0 = (x^1_0, ..., x^N_0)$, $\mathbf{v}_0 = (v^1_0, ..., v^N_0)$, $h, \gamma$
  Specify $\varphi_0, \varphi_1, \varphi_2, \Sigma_{11}, \Sigma_{12}, \Sigma_{22}$ using (33) and (34).
for $k = 0, ..., K-1$ do
    for $i = 1, ..., N$ do
        $\begin{bmatrix} (\mathrm{B}^i_k)^x \\ (\mathrm{B}^i_k)^v \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_{11} I_d & \Sigma_{12} I_d \\ \Sigma_{12} I_d & \Sigma_{22} I_d \end{bmatrix}\right)$
        $x^i_{k+1} \leftarrow x^i_k + \varphi_0\, v^i_k - \varphi_1\, D_\rho F(\mu_{\mathbf{x}_k}, x^i_k) + (\mathrm{B}^i_k)^x$
        $v^i_{k+1} \leftarrow \varphi_2\, v^i_k - \varphi_0\, D_\rho F(\mu_{\mathbf{x}_k}, x^i_k) + (\mathrm{B}^i_k)^v$
    end for
end for
output $(x^1_K, ..., x^N_K)$
```

The update parameters of Algorithm 1, φ_0, φ_1, φ_2 and Σ_{11}, Σ_{12}, Σ_{22}, are functions of γ and stepsize h. Thus, we need to specify the value of γ and h to compute the update parameters and initialize $(\mathbf{x}_0, \mathbf{v}_0) \sim \mu^N_0 \in \mathcal{P}_2(\mathbb{R}^{2Nd})$ before running the algorithm.

3.1. Convergence Analysis

We begin by leveraging entropic hypocoercivity and Theorems 2.1 and 2.2 from Chen et al. (2023) to analyze the continuous-time dynamics MULD and N-ULD. Let

$$S = \begin{pmatrix} 1/\mathscr{L} & 1/\sqrt{\mathscr{L}} \\ 1/\sqrt{\mathscr{L}} & 2 \end{pmatrix} \otimes I_d. \tag{12}$$

We construct the Lyapunov functional similar to Chen et al. (2023), but with a different choice of S. Theorem 3.1 is established by showing the following functional is decaying along the trajectory of MULD.

$$\begin{aligned} \mathcal{E}(\mu) &:= \mathcal{F}(\mu) + \mathsf{FI}_S(\mu\|\hat{\mu}), \text{ where} \\ \mathcal{F}(\mu) &:= F(\mu^X) + \int \frac{1}{2}\|v\|^2 \mu(\mathrm{d}x\mathrm{d}v) + \mathrm{Ent}(\mu). \end{aligned} \tag{13}$$

Our second Theorem 3.2 establishes the convergence of N-ULD. Denote $\mathbf{x} = (x^1, ..., x^N)$, $\mathbf{v} = (v^1, ..., v^N)$, $\mu^N = \mathrm{Law}(\mathbf{x}, \mathbf{v})$, and μ^N_* as the limiting distribution of N-ULD satisfying $\mu^N_*(\mathbf{x}, \mathbf{v}) \propto \exp\left(-NF(\mu_\mathbf{x}) - \frac{1}{2}\|\mathbf{v}\|^2\right)$ (see the derivation of limiting distribution in Appendix A.4). Denote $\nabla_i := (\nabla_{x^i}, \nabla_{v^i})^\mathsf{T}$. We obtain our guarantee by showing the functional is decaying along the trajectory of N-ULD:

$$\begin{aligned} &\mathcal{E}^N(\mu^N) := \mathcal{F}^N(\mu^N) + \mathsf{FI}^N_S(\mu^N\|\mu^N_*), \text{ where} \\ &\mathsf{FI}^N_S(\mu^N\|\mu^N_*) := \sum_{i=1}^N \mathbb{E}_{\mu^N}\left\|S^{1/2}\nabla_i \log\frac{\mu^N}{\mu^N_*}\right\|^2, \text{ and} \\ &\mathcal{F}^N(\mu^N) := \int NF(\mu_\mathbf{x}) + \frac{1}{2}\|\mathbf{v}\|^2 \mu^N(\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}) + \mathrm{Ent}(\mu^N). \end{aligned} \tag{14}$$

Theorem 3.1 (Convergence of MULD)**.** *If Assumptions 2.3-2.5 hold, μ_0 has finite second moment, finite entropy and finite Fisher information, then the law μ_t of the MULD with $\gamma = \sqrt{\mathscr{L}}$ and $\mathcal{E}$ defined in (13) satisfy,*

$$\mathcal{F}(\mu_t) - \mathcal{F}(\mu_*) \le (\mathcal{E}(\mu_0) - \mathcal{E}(\mu_*))\exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{3\sqrt{\mathscr{L}}}t\right).$$

Theorem 3.2 (Convergence of N-ULD)**.** *If Assumptions 2.3-2.5 hold, μ^N_0 has finite second moment, finite entropy, finite Fisher information, and $N \ge (\mathscr{L}/\mathscr{C}_{\mathsf{LSI}})(32 + 24\mathscr{L}/\mathscr{C}_{\mathsf{LSI}})$, then the joint law μ^N_t of N-ULD with $\gamma = \sqrt{\mathscr{L}}$ and $\mathcal{E}^N$ defined in (14) satisfy*

$$\frac{1}{N}\mathcal{F}^N(\mu^N_t) - \mathcal{F}(\mu_*) \le \frac{\mathcal{E}^N_0}{N}\exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{6\sqrt{\mathscr{L}}}t\right) + \frac{\mathcal{B}}{N},$$

where $\mathcal{B} := \frac{60\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}} + \frac{36\mathscr{L}^2 d}{\mathscr{C}^2_{\mathsf{LSI}}}$, $\mathcal{E}^N_0 := \mathcal{E}^N(\mu^N_0) - N\mathcal{E}(\mu_*)$.

Note that $\mathcal{E}^N_0 = \mathcal{F}^N(\mu^N_0) - N\mathcal{F}(\mu_*) + \mathsf{FI}^N_S(\mu^N_0\|\mu^N_*) \ge 0$ by Lemma B.4. The decaying rate given in Theorem 3.1 resembles the decaying rate of ULD in Zhang et al. (2023) with similar choices of γ and S. Theorem 3.2 implies the non-uniform-in-N convergence of N-ULD, which incorporates a bias term involving N due to the particle approximation. Our proof technique is more refined but parallel to that of Chen et al. (2023) where our faster convergence and smaller bias is achieved by choosing $\gamma = \sqrt{\mathscr{L}}$ instead of $\gamma = 1$ (see Table 1).

Our main results analyze the convergence of the discrete-time processes MULA and N-ULA as well as their mixing time guarantees to generate an ϵ-approximate solution in TV distance with the specific choice of initialization, damping coefficient γ, and stepsize h.

Theorem 3.3 (Convergence of MULA)**.** *In addition to the assumptions specified in Theorems 3.1, let Assumptions 2.6-2.8 hold. Denote* $\bar{\mu}_K := Law(x_K, v_K)$ *and* $\kappa := \mathscr{L}/\mathscr{C}_{\mathsf{LSI}}$. *Then in order to ensure* $\|\bar{\mu}_K - \mu_*\|_{\mathsf{TV}} \leq \epsilon$ *for MULA, it suffices to choose* $\gamma = \sqrt{\mathscr{L}}$, $\bar{\mu}_0 = \mathcal{N}(0, I_{2d})$, *and*

$$h = \widetilde{\Theta}\left(\frac{\mathscr{C}_{\mathsf{LSI}}\epsilon}{\mathscr{L}^{3/2}d^{1/2}}\right), \; K = \widetilde{\Theta}\left(\frac{\kappa^2 d^{1/2}}{\epsilon}\right).$$

Theorem 3.4 (Convergence of N-ULA)**.** *In addition to the assumptions specified in Theorem 3.2, let Assumptions 2.6, 2.7 and 2.9 hold. Denote* $\bar{\mu}^i_K := Law(x^i_K, v^i_K)$ *for* $i = 1, ..., N$ *and* $\kappa := \mathscr{L}/\mathscr{C}_{\mathsf{LSI}}$. *Then in order to ensure* $\frac{1}{N}\sum_{i=1}^N \|\bar{\mu}^i_K - \mu_*\|_{\mathsf{TV}} \leq \epsilon$ *for N-ULA, it suffices to choose* $\gamma = \sqrt{\mathscr{L}}$, $\bar{\mu}^N_0 = \mathcal{N}(0, I_{2Nd})$,

$$h = \widetilde{\Theta}\left(\frac{\mathscr{C}_{\mathsf{LSI}}\epsilon}{\mathscr{L}^{3/2}d^{1/2}}\right), \; K = \widetilde{\Theta}\left(\frac{\kappa^2 d^{1/2}}{\epsilon}\right),$$

and the number of particles $N = \Theta\left(\kappa^2 d/\epsilon^2\right)$.

3.2. Proof Sketches

For the continuous-time results, we outline the proof of Theorem 3.1 (and analogously Theorem 3.2) in this section to provide intuition for how choosing $\gamma = \sqrt{\mathscr{L}}$ can improve the decaying rate of MULD. Inheriting the analysis of Theorem 2.1 in Chen et al. (2023) and Lemma 32 in Villani (2009), we show that for a general γ, the Lyapunov functional (13) with $S = [s_{ij}] \otimes I_d \in \mathbb{R}^{2d\times 2d}$ along MULD satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\mu_t) \leq -Y_t^{\mathsf{T}}\mathcal{K}Y_t, \tag{15}$$

where $s_{11} = c$, $s_{12} = s_{21} = b$, $s_{22} = a$; $Y_t \in \mathbb{R}^4$ is defined by (54) in Appendix D.1; $\mathcal{K}$ is an upper triangle matrix with diagonal elements $(\gamma + 2\gamma a - 4\mathscr{L}b,\, 2\gamma a,\, 2b,\, 2\gamma c)$. To ensure $S \succ 0$ and $Y_t^{\mathsf{T}}\mathcal{K}Y_t > 0$, the criteria of choosing positive constants γ, a, b, c should be $ac > b^2$ and $\mathcal{K} \succ 0$. Our choices of $\gamma = \sqrt{\mathscr{L}}$ as well as a, b, c specified in (12) satisfy the criterion and upper bound the right hand side of (15) differently from Chen et al. (2023) with $\gamma = 1$, which leads to an improved decaying rate by utilizing Grönwall's inequality.

For discretization errors, we outline the proof of Theorem 3.3 (and analogously Theorem 3.4) in this section. Let $(\mu_t)_{t\geq 0}$ and $(\bar{\mu}_{t/h})_{t\geq 0}$ represent the law of MULD and MULA initialized at μ_0. Let $\mathbf{Q}_{Kh}$ and $\mathbf{P}_{Kh}$ denote probability measures of MULD and MULA on the space of paths $C([0, Kh], \mathbb{R}^{2d})$. Invoking Girsanov's theorem (Girsanov, 1960; Kutoyants, 2004; Le Gall, 2016) and Assumption 2.4, we can upper bound the pathwise KL divergence between MULD and MULA for stepsize h and $k = 1, ..., K$ under Assumptions 2.4 and 2.6:

$$\mathsf{KL}(\mathbf{Q}_{Kh}\|\mathbf{P}_{Kh}) \lesssim \frac{\mathscr{L}^4 h^5}{\gamma}\sum_{k=0}^{K-1}\mathbb{E}_{\mathbf{Q}_{Kh}}\|x_{kh}\|^2 + \frac{\mathscr{L}^4 h^5 K}{\gamma} + \frac{\mathscr{L}^2 h^3}{\gamma}\sum_{k=0}^{K-1}\mathbb{E}_{\mathbf{Q}_{Kh}}\|v_{kh}\|^2 + \mathscr{L}^2 h^4 K d \tag{16a}$$

The derivation of (16a) is similar to that of Zhang et al. (2023); they show a discretization error of EI-ULMC measured in the q-th order Rényi divergence ($q \in [1, 2)$), which has KL divergence as a special case ($q = 1$). Their smoothness assumption on the potential function V is $(\mathscr{L}, s)$-weak smoothness, which recovers $\mathscr{L}$-smoothness when $s = 1$. We use techniques similar to Zhang et al. (2023) to bound the discretization error. Their Lemma 26 can be generalized to our Lemma B.7 in the mean-field setting, which describes an intermediate process of deriving (16a). Uniformly upper bounding the right-hand side of (16a) requires obtaining uniform bounds for $\mathbb{E}_{\mathbf{Q}_{Kh}}\|x_{kh}\|^2$ and $\mathbb{E}_{\mathbf{Q}_{Kh}}\|v_{kh}\|^2$; If we were to rely on existing techniques (Zhang et al., 2023), we would need a χ^2-convergence guarantee of MULD. Given χ^2-convergence is not established for MULD by previous works, we develop different techniques to uniformly upper bound the iterates of MULD and N-ULD. Denote $z_t = (x_t, v_t)$. More specifically, we have for $t \in [0, T]$

$$\mathbb{E}_{\mathbf{Q}_T}\|z_t\|^2 = W_2^2(\mu_t, \delta_0) \lesssim \underbrace{W_2^2(\mu_t, \mu_*)}_{\mathsf{I}} + \underbrace{W_2^2(\mu_*, \delta_0)}_{\mathsf{II}},$$

where δ_0 is Dirac measure on $0 \in \mathbb{R}^{2d}$, and II is the second moment of μ_*. Under Assumption 2.5, μ_* satisfies LSI which allows us to upper bound I using Talagrand's inequality: $\mathsf{I} \lesssim \mathsf{KL}(\mu_t\|\mu_*)/\mathscr{C}_{\mathsf{LSI}}$. Under Assumptions 2.3 and 2.4, Lemma 4.2 in Chen et al. (2023) establishes the following relation between KL divergence and energy gap: $\mathsf{KL}(\mu_t\|\mu_*) \leq \mathcal{F}(\mu_t) - \mathcal{F}(\mu_*)$. Along with this upper bound of KL divergence, we obtain

$$\mathsf{I} \lesssim \frac{\mathsf{KL}(\mu_t\|\mu_*)}{\mathscr{C}_{\mathsf{LSI}}} \leq \frac{\mathcal{F}(\mu_t) - \mathcal{F}(\mu_*)}{\mathscr{C}_{\mathsf{LSI}}} \leq \frac{\mathcal{F}(\mu_0)}{\mathscr{C}_{\mathsf{LSI}}},$$

where the last inequality follows from the fact that $\mathcal{F}(\mu_t)$ is decreasing along MULD (Kazeykina et al., 2020; Chen et al., 2023) and our assumption that $\mathcal{F}(\mu_*) \geq 0$. Therefore, under Assumptions 2.7 and 2.8, our Lemma B.8 establishes the upper bound of $\mathbb{E}_{\mathbf{Q}_T}\|z_t\|^2$ in terms of $\mathscr{L}$, $\mathscr{C}_{\mathsf{LSI}}$ and d, which implies the uniform upper bound of $\mathsf{KL}(\mu_T\|\bar{\mu}_K)$ by data processing inequality. Applying Pinsker's inequality, we can convert the discretization error bound in KL divergence to that in TV distance. The remaining details of this proof is deferred to Appendix E.

Discretization	Method	# of particles	Mixing time
Time-discretizations	MLA (Nitanda et al., 2022)	*	$\widetilde{\Theta}\left(\kappa^2\mathscr{L}d/\epsilon^2\right)$
	EI-ULMC (Zhang et al., 2023)	*	$\widetilde{\Theta}\left(\kappa^{3/2}d^{1/2}/\epsilon\right)$
	MULA **(Ours)**	*	$\widetilde{\Theta}\left(\kappa^2 d^{1/2}/\epsilon\right)$
Space-discretizations	N-ULD (Chen et al. (2023))	$\Theta\left(\kappa^2\mathscr{L}d/\epsilon^2\right)$	$\widetilde{\Theta}\left(\kappa\right)$
	N-ULD **(Ours)**	$\Theta\left(\kappa^2 d/\epsilon^2\right)$	$\widetilde{\Theta}\left(\kappa/\mathscr{L}^{1/2}\right)$
Spacetime discretizations	N-LA (Suzuki et al., 2023)	$\Theta\left(\kappa\mathscr{L}^3 d/\epsilon^2\right)$	$\widetilde{\Theta}\left(\kappa^2\mathscr{L}d/\epsilon^2\right)$
	N-ULA **(Ours)**	$\Theta\left(\kappa^2 d/\epsilon^2\right)$	$\widetilde{\Theta}\left(\kappa^2 d^{1/2}/\epsilon\right)$

Table 1. Comparison of algorithms in terms of the mixing time and number of particles to achieve ϵ-approximate solutions in TV distance. $\kappa := \mathscr{L}/\mathscr{C}_{\mathsf{LSI}}$. * represents that we do not need particle approximation for this method.

3.3. Discussion of Mixing Time Results

We summarize the convergence results of MULA, N-ULA and several existing methods including EI-ULMC, the EM-discretization of MLD (referred to as MLA (Nitanda et al., 2022)), and its finite-particle system (referred to as N-LA (Suzuki et al., 2023)) in Table 1. For the mixing time to generate an ϵ-approximate solution in TV distance, our proposed MULA and N-ULA achieve better dependence on $\mathscr{L}$, d and ϵ than MLA and N-LA, and keep the same dependence on $\mathscr{C}_{\mathsf{LSI}}$ as MLA and N-LA. For the number of particles, we improve the dependence on $\mathscr{L}$ for N-ULD ($\gamma = \sqrt{\mathscr{L}}$) when compared with N-ULD ($\gamma = 1$) in Chen et al. (2023) (and for N-ULA when compared with N-LA). Particularly, our dependence on the smoothness constant in the number of particle guarantee of N-ULA is $\Theta(\mathscr{L}^2)$ whereas the counterpart of N-LA is $\Theta(\mathscr{L}^4)$. However, our dependence on the LSI constant in the number of particle guarantee of N-ULA is $\Theta(\mathscr{C}_{\mathsf{LSI}}^{-2})$ whereas the counterpart of N-LA is $\Theta(\mathscr{C}_{\mathsf{LSI}}^{-1})$.

4. Applications of Algorithm 1

In this section, we will show how Algorithm 1 can be applied to several applications by verifying Assumptions 2.3-2.9 hold for these examples. We present these results in full details in Appendix C.

4.1. Training Mean-field Neural Networks

Consider a two-layer mean-field neural network (with infinite depth), which can be parameterized as $h(\rho; a) := \mathbb{E}_{x\sim\rho}[h(x; a)]$, where $h(x; a)$ represents a single neuron with trainable parameter x and input a (e.g. $h(x; a) = \sigma(x^\mathsf{T} a)$ for activation function σ); ρ is the probability distribution of the parameter x. Given dataset $(a_i, b_i)_{i=1}^n$ and loss function ℓ, we choose F in objective (2) to be

$$F(\mu^X) = \frac{1}{n}\sum_{i=1}^n \ell(h(\mu^X; a_i), b_i) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\mu^X}\|x\|^2, \quad (17)$$

The objectives (17) satisfy Assumptions 2.3-2.6 for specific common choices ℓ and h described in several works (Nitanda et al., 2022; Chen et al., 2022; 2023; Suzuki et al., 2023). If there exists $\mathscr{L} > 0$ such that the activation function satisfies $|h(x; a)| \leq \sqrt{\mathscr{L}}$ (also proposed in Suzuki et al. (2023)) and the convex loss function ℓ is quadratic or satisfies $|\partial_1 \ell| \leq \sqrt{\mathscr{L}}$ (also proposed in Nitanda et al. (2022)), F satisfies Assumption 2.7 with $\lambda' \leq (2\pi)^3\exp(-8\mathscr{L})$. Finally, if in addition we assume ℓ is $\sqrt{\mathscr{L}}$-Lipschitz and choose $\lambda' \leq (2\pi)^3\exp(-8\mathscr{L})$, $\mu_0 = \mathcal{N}(0, I_{2d})$ and $\mu_0^N = \mathcal{N}(0, I_{2Nd})$, Assumptions 2.8 and 2.9 will be satisfied.

4.2. Density Estimation via MMD Minimization

The maximum mean discrepancy between two probability measures ρ and π is defined as $\mathcal{M}(\rho\|\pi) = \iint[k(x,x) - 2k(x,y) + k(y,y)]\mathrm{d}\rho(x)\mathrm{d}\pi(y)$, where k is a positive definite kernel. Similar to Example 2 in Suzuki et al. (2023), we consider the non-parametric density estimation using the Gaussian mixture model, which can be parameterized as $p(\rho; z) := \mathbb{E}_{x\sim\rho}[p(x; z)]$, where $p(x; z)$ is the Gaussian density function of z with mean x and a user-specified variance σ^2. Given a set of samples $\{z_i\}_{i=1}^n$ from the target distribution p^*, our goal is to fit p^* by minimizing the empirical version of $\mathcal{M}(p(\rho; z)\|p^*)$, defined as

$$\begin{aligned}\hat{\mathcal{M}}(\rho) = &\iiint p(x; z)p(x'; z')k(z, z')\mathrm{d}z\mathrm{d}z'\mathrm{d}\rho(x)\mathrm{d}\rho(x') \\ &- 2\int\left(\frac{1}{n}\sum_{i=1}^n\int p(x; z)k(z, z_i)\mathrm{d}z\right)\mathrm{d}\rho(x).\end{aligned}$$

We choose F in objective (2) to be

$$F(\mu^X) = \hat{\mathcal{M}}(\mu^X) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\mu^X}\|x\|^2, \quad (18)$$

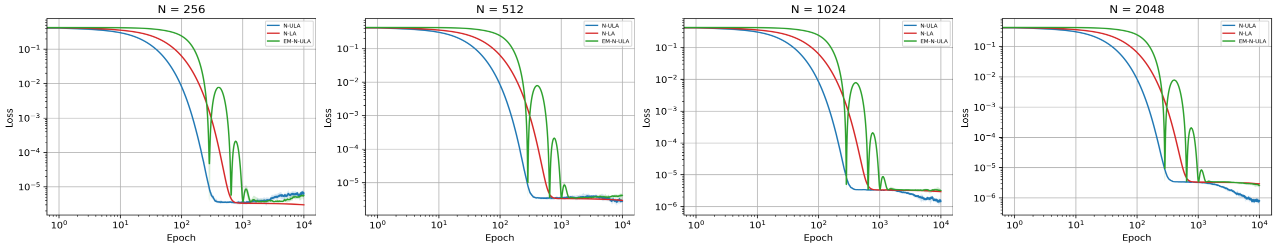

Figure 1. Evaluation on N-ULA, N-LA and EM-N-ULA with different number of particles N where x-axis represents the training epochs and y-axis represents the value of $\frac{1}{2n}\sum_{i=1}^{n}(\frac{1}{N}\sum_{s=1}^{N}h(x^s;a_i)-f(a_i))^2$. As the number of particles grows, our method outperforms both N-LA and EM-N-ULA, however if the number of particles is small, the momentum-based methods suffer from unstable convergence (the loss will slightly go up after convergence) as is shown in the first figure.

where $\lambda' > 0$. Suzuki et al. (2023) show that objective (18) satisfies Assumptions 2.3, 2.5 and 2.6 by choosing a smooth and light-tailed kernel k, such as Gaussian radial basis function (RBF) kernel defined as $k(z,z') := \exp(-\|z-z'\|^2/2\sigma'^2)$ for $\sigma' > 0$. We also verify that objective (18) also satisfies our Assumption 2.4 with the same choice of kernel. With Gaussian RBF kernel k ($\sigma' = \sigma$), we provide verification in Appendix C that objective (18) satisfies Assumptions 2.7-2.9 when $\lambda' \le 3\pi/25$, $\mu_0 = \mathcal{N}(0, I_{2d})$ and $\mu_0^N = \mathcal{N}(0, I_{2Nd})$.

4.3. Kernel Stein Discrepancy Minimization

Kernel Stein discrepancy (KSD) minimization is a method for sampling from a target distribution ρ_* if we have the access to the score function $s_{\rho_*}(x) = \nabla \log \rho_*(x)$ (Chwialkowski et al., 2016; Liu et al., 2016). For a positive definite kernel k, the Stein kernel is defined as

$$\begin{aligned} u_{\rho_*}(x,x') &= s_{\rho_*}^{\mathsf{T}}(x)k(x,x')s_{\rho_*}(x') + s_{\rho_*}^{\mathsf{T}}(x)\nabla_{x'}k(x,x') \\ &\quad + \nabla_x^{\mathsf{T}}k(x,x')s_{\rho_*}(x') + \mathsf{tr}(\nabla_{x,x'}k(x,x')). \end{aligned}$$

The KSD between ρ and ρ_* is defined as $\mathsf{KSD}(\rho) = \iint u_{\rho_*}(x,x')\mathrm{d}\rho(x)\mathrm{d}\rho(x')$. We choose F in (2) to be

$$F(\mu^X) = \mathsf{KSD}(\mu^X) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\mu^X}\|x\|^2, \qquad (19)$$

where $\lambda' > 0$. Suzuki et al. (2023) show that objective (19) satisfies Assumptions 2.3, 2.5 and 2.6 by choosing light-tailed kernel and assume the score function satisfies

$$\max_{k=1,2,3}\{\|\nabla^{\otimes k}\log\rho_*(x)\|_{\mathsf{op}}\} \le \mathscr{L}(1+\|x\|). \qquad (20)$$

Choosing the same kernel as in Suzuki et al. (2023), we verify in Appendix C that (19) also satisfies Assumption 2.4 and satisfies our Assumptions 2.7-2.9 with $\lambda' \le \min\{(2\pi)^3\exp(-4\mathscr{L}), \mathscr{L}, d\}$, $\mu_0 = \mathcal{N}(0, I_{2d})$ and $\mu_0^N = \mathcal{N}(0, I_{2Nd})$.

5. Experiments

We evaluate our method N-ULA in training a two-layer mean-field neural network to approximate a Gaussian function $f(z) = \exp(-\|z-m\|^2/2d)$ for $z \in \mathbb{R}^d$ and unknown $m \in \mathbb{R}^d$. Consider objective (17) given quadratic loss ℓ and n randomly generated data samples $(a_i, f(a_i))_{i=1}^n$:

$$F(\rho) = \frac{1}{2n}\sum_{i=1}^{n}(h(\rho;a_i)-f(a_i))^2 + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}\|x\|^2,$$

where $h(\rho;a) = \mathbb{E}_{x\sim\rho}[\mathsf{tanh}(x^{\mathsf{T}}a)]$. In this task, we compare our N-ULA with N-LA (the spacetime discretization of MLD) and EM-N-ULA (the Euler-Maruyama time discretization of N-ULD) for the number of particles $N \in \{2^8, 2^9, 2^{10}, 2^{11}\}$. We average our results over 5 random seeds from $\{0,1,2,3,4\}$ and present the convergence curves of three methods with their error bars in Figure 1. Figure 1 demonstrate the superiority of N-ULA over N-LA and EM-N-ULA in terms of the convergence speed. Codes of our experiments are available at https://github.com/QiangFu09/NULA. More details of the experimental settings and discussion are postponed to Section F.

6. Discussion

To summarize, this paper (1) improves the convergence guarantees in Chen et al. (2023) with a refined Lyapunov analysis (Theorems 3.1 and 3.2); (2) discretizes the MULD and N-ULD with a scheme which results in smaller bias than the EM scheme; and (3) presents a novel discretization analysis of MULD and N-ULD. We also verify that these methods work when the objective is W_1 smooth. We provide experimental evidence in Appendix F that MULA and N-ULA outperforms MLA and N-LA across a variety of experiments, consistent with our theoretical findings. We now note several directions for future potential developments. First, we only choose $\lambda = 1$ in our bound, as considering a general

λ is even more complicated in the analysis of MULD in Chen et al. (2023). However, it is interesting to incorporate a general λ for generalization in machine learning, which we leave for the future work. Second, it is unclear what the optimal choice of damping coefficient γ is for MULD and N-ULD. Understanding whether the optimal choice has been found is helpful to improving our established mixing time. What's more, we obtain convergence rates for the MULA and N-ULA in TV distance, which are not consistent with the convergence rates of MULD, N-ULD, MLA and N-LA in energy gap (e.g. $\mathcal{F}(\mu_t) - \mathcal{F}(\mu_*)$). We hope to establish our results in the energy gap or KL divergence in the future. Finally, our techniques on uniformly bounding the iterates of MULD and N-ULD combined with Assumptions 2.7-2.9 generates an additional $\mathscr{C}_{\mathsf{LSI}}$ after applying Talagrand's inequality, which leads to non-improvement of $\mathscr{C}_{\mathsf{LSI}}$ for MULA and N-ULA. We hope to explore whether it is possible to weaken those assumptions and refine the analysis of uniformly bounding the iterates to improve the dependence of $\mathscr{C}_{\mathsf{LSI}}$ in the mixing time and number of particles of MULA and N-ULA.

Acknowledgment

We would like to thank Andre Wibisono for valuable discussion and suggestions on this paper, thank Songbo Wang for correction on our Assumption 2.4, thank Zhenjie Ren for sharing their codes and thank our anonymous reviewers for their constructive comments.

Impact Statement

This paper presents theoretical work whose goal is to advance the field of Machine Learning.

References

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32, 2019.

Bayraktar, E., Feng, Q., and Li, W. Exponential entropy dissipation for weakly self-consistent vlasov-fokker-planck equations. *arXiv preprint arXiv:2204.12049*, 2022.

Bolley, F., Guillin, A., and Malrieu, F. Trend to equilibrium and particle approximation for a weakly selfconsistent vlasov-fokker-planck equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(5):867–884, 2010.

Bou-Rabee, N. and Schuh, K. Convergence of unadjusted hamiltonian monte carlo for mean-field models. *Electronic Journal of Probability*, 28:1–40, 2023.

Cao, Y., Lu, J., and Wang, L. On explicit l 2-convergence rate estimate for underdamped langevin dynamics. *Archive for Rational Mechanics and Analysis*, 247(5):90, 2023.

Chen, F., Ren, Z., and Wang, S. Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv preprint arXiv:2212.03050*, 2022.

Chen, F., Lin, Y., Ren, Z., and Wang, S. Uniform-in-time propagation of chaos for kinetic mean field langevin dynamics. *arXiv preprint arXiv:2307.02168*, 2023.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on learning theory*, pp. 300–323. PMLR, 2018.

Chizat, L. Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 2022.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *International conference on machine learning*, pp. 2606–2615. PMLR, 2016.

Claisse, J., Conforti, G., Ren, Z., and Wang, S. Mean field optimization problem regularized by fisher information. *arXiv preprint arXiv:2302.05938*, 2023.

Duong, M. H. and Tugaut, J. The vlasov-fokker-planck equation in non-convex landscapes: convergence to equilibrium. 2018.

Eberle, A., Guillin, A., and Zimmer, R. Couplings and quantitative contraction rates for langevin dynamics. 2019a.

Eberle, A., Guillin, A., and Zimmer, R. Quantitative harris-type theorems for diffusions and mckean–vlasov processes. *Transactions of the American Mathematical Society*, 371(10):7135–7173, 2019b.

Foster, J., Lyons, T., and Oberhauser, H. The shifted ode method for underdamped langevin mcmc. *arXiv preprint arXiv:2101.03446*, 2021.

Foster, J., Reis, G. d., and Strange, C. High order splitting methods for sdes satisfying a commutativity condition. *arXiv preprint arXiv:2210.17543*, 2022.

Fu, Q., Xu, D., and Wilson, A. C. Accelerated stochastic optimization methods under quasar-convexity. In *International Conference on Machine Learning*, pp. 10431–10460. PMLR, 2023.

Girsanov, I. V. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability & Its Applications*, 5(3): 285–301, 1960.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Guillin, A. and Monmarché, P. Uniform long-time and propagation of chaos estimates for mean field kinetic particles in non-convex landscapes. *Journal of Statistical Physics*, 185:1–20, 2021.

Guillin, A., Liu, W., Wu, L., and Zhang, C. The kinetic fokker-planck equation with mean field interaction. *Journal de Mathématiques Pures et Appliquées*, 150:1–23, 2021.

Guillin, A., Le Bris, P., and Monmarché, P. Convergence rates for the vlasov-fokker-planck equation and uniform in time propagation of chaos in non convex cases. *Electronic Journal of Probability*, 27:1–44, 2022.

He, Y., Balasubramanian, K., and Erdogdu, M. A. On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method. *Advances in Neural Information Processing Systems*, 33:7366–7376, 2020.

Hinder, O., Sidford, A., and Sohoni, N. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, pp. 1894–1938. PMLR, 2020.

Hörmander, L. Hypoelliptic second order differential equations. 1967.

Hu, K., Ren, Z., Siska, D., and Szpruch, L. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019.

Johnston, T., Lytras, I., and Sabanis, S. Kinetic langevin mcmc sampling without gradient lipschitz continuity–the strongly convex case. *arXiv preprint arXiv:2301.08039*, 2023.

Kazeykina, A., Ren, Z., Tan, X., and Yang, J. Ergodicity of the underdamped mean-field langevin dynamics. *arXiv preprint arXiv:2007.14660*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kloeden, P. E., Platen, E., Gelbrich, M., and Romisch, W. Numerical solution of stochastic differential equations. *SIAM Review*, 37(2):272–274, 1995.

Kolmogoroff, A. Zufallige bewegungen (zur theorie der brownschen bewegung). *The Annals of Mathematics*, 35 (1):116, 1934.

Kutoyants, Y. A. *Statistical inference for ergodic diffusion processes*. Springer Science & Business Media, 2004.

Laborde, M. and Oberman, A. A lyapunov analysis for accelerated gradient methods: From deterministic to stochastic case. In *International Conference on Artificial Intelligence and Statistics*, pp. 602–612. PMLR, 2020.

Le Gall, J.-F. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

Leimkuhler, B., Paulin, D., and Whalley, P. A. Contraction and convergence rates for discretized kinetic langevin dynamics. *arXiv preprint arXiv:2302.10684*, 2023.

Li, X., Wu, Y., Mackey, L., and Erdogdu, M. A. Stochastic runge-kutta accelerates langevin monte carlo and beyond. *Advances in neural information processing systems*, 32, 2019.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR, 2016.

Ma, Y.-A., Chatterji, N. S., Cheng, X., Flammarion, N., Bartlett, P. L., and Jordan, M. I. Is there an analog of nesterov acceleration for gradient-based mcmc? 2021.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Monmarché, P. Long-time behaviour and propagation of chaos for mean field kinetic particles. *Stochastic Processes and their Applications*, 127(6):1721–1737, 2017.

Monmarché, P. High-dimensional mcmc with a standard splitting scheme for the underdamped langevin diffusion. *Electronic Journal of Statistics*, 15(2):4117–4166, 2021.

Nesterov, Y. E. A method of solving a convex programming problem with convergence rate o\bigl(k^2\bigr). In *Doklady Akademii Nauk*, volume 269, pp. 543–547. Russian Academy of Sciences, 1983.

Nitanda, A., Wu, D., and Suzuki, T. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 9741–9757. PMLR, 2022.

Platen, E. and Bruti-Liberati, N. *Numerical solution of stochastic differential equations with jumps in finance*, volume 64. Springer Science & Business Media, 2010.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Shen, R. and Lee, Y. T. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Srinivasan, V. and Wilson, A. C. Sufficient conditions for non-asymptotic convergence of riemannian optimization methods. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Suzuki, T., Wu, D., and Nitanda, A. Convergence of mean-field langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. *arXiv preprint arXiv:2306.07221*, 2023.

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Villani, C. Limites hydrodynamiques de l'équation de boltzmann. *Séminaire Bourbaki*, 2000:365–405, 2001.

Villani, C. *Hypocoercivity*, volume 202. American Mathematical Society, 2009.

Wilson, A. C., Recht, B., and Jordan, M. I. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

Zhang, M., Chewi, S., Li, M. B., Balasubramanian, K., and Erdogdu, M. A. Improved discretization analysis for underdamped langevin monte carlo. *arXiv preprint arXiv:2302.08049*, 2023.

A. Supplementary Background

A.1. Mean-field Langevin Dynamics

The law $(\rho_t)_{t\geq 0}$ of MLD solves the following non-linear Fokker-Planck equation:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t D_\rho F(\rho_t, \cdot)) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\hat{\rho}_t}\right), \tag{21}$$

where $\hat{\rho}_t(x) \propto \exp\left(-\frac{\delta F}{\delta \rho}(\rho_t, x)\right)$. Let $E(\rho) := F(\rho) + \mathrm{Ent}(\rho)$. The optimality condition of the EMO problem is

$$\frac{\delta E}{\delta \rho} = \frac{\delta F}{\delta \rho} + \log \rho + c = 0, \tag{22}$$

where c is a constant. Given the condition (22), the solution of EMO problem ρ_* satisfies $\rho_*(x) = \hat{\rho}_*(x) \propto \exp\left(-\frac{\delta F}{\delta \rho}(\rho_*, x)\right)$, which solves $\nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\hat{\rho}_t}\right) = 0$. Thus we conclude that MLD converges to the minimizer of EMO objective.

A.2. N-particle Langevin Dynamics

The space-discretization of MLD is referred to as the *N-particle Langevin dynamics*,

$$\mathrm{d}x_t^i = -D_\rho F(\rho_{\mathbf{x}_t}, x_t^i)\mathrm{d}t + \sqrt{2}\mathrm{dB}_t, \tag{N-LD}$$

where $\rho_{\mathbf{x}_t} = \frac{1}{N}\sum_{i=1}^N \delta_{x_t^i}$. Let ρ_t^i denotes the law of x_t^i and ρ_t^N denotes the joint law of $\mathbf{x}_t := (x_t^1, ..., x_t^N)$. The joint law $(\rho_t^N)_{t\geq 0}$ of N-LD solves the following linear Fokker-Planck equation:

$$\frac{\partial \rho_t^N}{\partial t} = \sum_{i=1}^N \nabla_i \cdot \left(\rho_t^N D_\rho F(\rho_{\mathbf{x}_t}, x_t^i)\right) + \Delta_i \rho_t^N = \sum_{i=1}^N \nabla_i \cdot \left(\rho_t^N \nabla_i \log \frac{\rho_t^N}{\rho_*^N}\right), \tag{23}$$

where $\nabla_i := \nabla_{x^i}$, $\Delta_i := \Delta_{x^i}$ and $\rho_*^N(\mathbf{x}) \propto \exp(-NF(\rho_\mathbf{x}))$. Define the *N-particle free energy*:

$$E^N(\rho^N) = N \int F(\rho_\mathbf{x})\rho^N(\mathrm{d}x) + \mathrm{Ent}(\rho^N). \tag{24}$$

The optimality condition of minimizing the N-particle free energy (24) over $\mathcal{P}_2(\mathbb{R}^{Nd})$ is

$$\frac{\delta E^N}{\delta \rho^N} = NF(\rho_\mathbf{x}) + \log \rho^N + c = 0, \tag{25}$$

where c is a constant. Given the optimality condition (25), the minimizer of (24) satisfies $\rho_*^N(x) \propto \exp(-NF(\rho_\mathbf{x}))$, which is exactly the limiting distribution of N-LD according to (23). Thus we conclude that N-LD converges to the minimizer of (24).

A.3. Mean-field Underdamped Langevin Dynamics

The law $(\mu_t)_{t\geq 0}$ of MULD solves the following non-linear Fokker-Planck equation:

$$\begin{aligned} \frac{\partial \mu_t}{\partial t} &= \gamma \Delta_v \mu_t + \gamma \nabla_v \cdot (\mu_t v_t) - v \cdot \nabla_x \mu_t + D_\rho F(\mu_t^x, x_t) \cdot \nabla_v \mu_t \\ &= \nabla \cdot \left(\mu_t J_\gamma \nabla \log \frac{\mu_t}{\hat{\mu}_t}\right), \end{aligned} \tag{26}$$

where $J_\gamma = \begin{pmatrix} 0 & 1 \\ -1 & \gamma \end{pmatrix}$, $\nabla := (\nabla_x, \nabla_v)^\mathsf{T}$ and $\hat{\mu}_t(x, v) \propto \exp\left(-\frac{\delta F}{\delta \rho}(\mu_t^X, x) - \frac{1}{2}\|v\|^2\right)$. The optimality condition of the augmented EMO problem is

$$\frac{\delta \mathcal{F}}{\delta \mu} = \frac{\delta F}{\delta \mu} + \log \mu + \frac{1}{2}\|v\|^2 + c = 0, \tag{27}$$

where $\mathcal{F}$ is defined in (13) and c is a constant. Note that $\frac{\delta F(\mu^X)}{\delta \mu} = \frac{\delta F(\mu^X)}{\delta \rho}$. Given the optimality condition (27), the solution of the augmented EMO problem satisfies $\mu_*(x,v) = \hat{\mu}_*(x,v) \propto \exp\left(-\frac{\delta F}{\delta \rho}(\mu_*^X, x) - \frac{1}{2}\|v\|^2\right)$, which solves $\nabla \cdot \left(\mu_t J_\gamma \nabla \log \frac{\mu_t}{\hat{\mu}_t}\right) = 0$. Thus we conclude that MULD converges to the minimizer of the augmented EMO objective.

A.4. N-particle Underdamped Langevin Dynamics

The law $(\mu_t^N)_{t\geq 0}$ of N-ULD solves the following linear Fokker-Planck equation:

$$\begin{aligned}\frac{\partial \mu_t^N}{\partial t} &= \sum_{i=1}^N \left(\gamma \Delta_{v^i} \mu_t^N + \gamma \nabla_{v^i} \cdot (\mu_t^N v_t^i) - v_t^i \cdot \nabla_{x^i} \mu_t^N + D_\rho F(\mu_{\mathbf{x}_t}, x_t^i) \cdot \nabla_{v^i} \mu_t^N\right) \\ &= \sum_{i=1}^N \nabla_i \cdot \left(\mu_t^N J_\gamma \nabla_i \log \frac{\mu_t^N}{\hat{\mu}_*^N}\right),\end{aligned} \tag{28}$$

where $J_\gamma = \begin{pmatrix} 0 & 1 \\ -1 & \gamma \end{pmatrix}$, $\nabla_i := (\nabla_{x^i}, \nabla_{v^i})^\mathsf{T}$ and $\hat{\mu}_*^N(x,v) \propto \exp\left(-NF(\mu_\mathbf{x}) - \frac{1}{2}\|v\|^2\right)$. Define the *N-particle free energy*:

$$\mathcal{F}^N(\mu^N) = \int NF(\mu_\mathbf{x}) + \frac{1}{2}\|\mathbf{v}\|^2 \mu^N(\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}) + \mathrm{Ent}(\mu^N). \tag{29}$$

The optimality condition of minimizing the N-particle free energy (29) over $\mathcal{P}_2(\mathbb{R}^{2Nd})$ is

$$\frac{\delta \mathcal{F}^N}{\delta \mu^N} = NF(\mu_\mathbf{x}) + \frac{1}{2}\|\mathbf{v}\|^2 + \log \mu^N + c = 0, \tag{30}$$

where c is a constant. Given the optimality condition (30), the minimizer of (29) satisfies $\mu_*^N(x) \propto \exp(-NF(\mu_\mathbf{x}) - \frac{1}{2}\|\mathbf{v}\|^2)$, which is exactly the limiting distribution of N-ULD according to (28). Thus we conclude that N-ULD converges to the minimizer of (29).

B. Helpful Lemmas

Lemma B.1. *The solution (x_t, v_t) to the discrete-time process (MULA) for $t \in [kh, (k+1)h]$ is*

$$\begin{aligned} x_t &= x_{kh} + \frac{1 - e^{-\gamma(t-kh)}}{\gamma} v_{kh} - \frac{\gamma h - (1 - e^{-\gamma(t-kh)})}{\gamma^2} D_\rho F(\mu_{kh}^X, x_{kh}) + \mathrm{B}_{kh}^x, \\ v_t &= e^{-\gamma(t-kh)} v_{kh} - \frac{1 - e^{-\gamma(t-kh)}}{\gamma} D_\rho F(\mu_{kh}^X, x_{kh}) + \mathrm{B}_{kh}^v, \end{aligned} \tag{31}$$

where $(\mathrm{B}_{kh}^x, \mathrm{B}_{kh}^v) \in \mathbb{R}^{2d}$ is independent of k and has the joint distribution

$$\begin{bmatrix} \mathrm{B}_{kh}^x \\ \mathrm{B}_{kh}^v \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \frac{2}{\gamma}\left(h - \frac{2(1-e^{-\gamma(t-kh)})}{\gamma} + \frac{1-e^{-2\gamma(t-kh)}}{2\gamma}\right) & \frac{1}{\gamma}\left(1 - 2e^{-\gamma(t-kh)} + e^{-2\gamma(t-kh)}\right) \\ * & 1 - e^{-2\gamma(t-kh)} \end{bmatrix} \otimes I_d\right)$$

The solution (x_t^i, v_t^i) to the discrete-time process (11) for $i = 1, ..., N$ and $t \in [kh, (k+1)h]$ is

$$\begin{aligned} x_t^i &= x_{kh}^i + \frac{1 - e^{-\gamma(t-kh)}}{\gamma} v_{kh}^i - \frac{\gamma h - (1 - e^{-\gamma(t-kh)})}{\gamma^2} D_\rho F(\mu_{\mathbf{x}_{kh}}, x_{kh}^i) + (\mathrm{B}_{kh}^i)^x, \\ v_t^i &= e^{-\gamma(t-kh)} v_{kh}^i - \frac{1 - e^{-\gamma(t-kh)}}{\gamma} D_\rho F(\mu_{\mathbf{x}_{kh}}, x_{kh}^i) + (\mathrm{B}_{kh}^i)^v. \end{aligned} \tag{32}$$

where $((\mathrm{B}_{kh}^i)^x, (\mathrm{B}_{kh}^i)^v) \in \mathbb{R}^{2d}$ is independent of i, k and has the joint distribution

$$\begin{bmatrix} (\mathrm{B}_{kh}^i)^x \\ (\mathrm{B}_{kh}^i)^v \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \frac{2}{\gamma}\left(h - \frac{2(1-e^{-\gamma(t-kh)})}{\gamma} + \frac{1-e^{-2\gamma(t-kh)}}{2\gamma}\right) & \frac{1}{\gamma}\left(1 - 2e^{-\gamma(t-kh)} + e^{-2\gamma(t-kh)}\right) \\ * & 1 - e^{-2\gamma(t-kh)} \end{bmatrix} \otimes I_d\right)$$

Proof. The proof technique is similar to the proof of Lemmas 10 and 11 proposed in Cheng et al. (2018). □

Choosing $t = (k+1)h$ for (32) generates the update parameters of Algorithm 1:

$$\varphi_0 = \frac{1-e^{-\gamma h}}{\gamma},\ \varphi_1 = \frac{\gamma h - (1-e^{-\gamma h})}{\gamma^2},\ \varphi_2 = e^{-\gamma h}; \tag{33}$$

$$\Sigma_{11} = \frac{2}{\gamma}\left(h - \frac{2(1-e^{-\gamma h})}{\gamma} + \frac{1-e^{-2\gamma h}}{2\gamma}\right),\ \Sigma_{12} = \frac{1}{\gamma}\left(1-2e^{-\gamma h}+e^{-2\gamma h}\right),\ \Sigma_{22} = 1-e^{-2\gamma h}. \tag{34}$$

Lemma B.2. *Suppose $D_\rho F : \mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}^d$ admits a continuous first variation $\delta D_\rho F : \mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}^d$. Then, $D_\rho F$ is $\mathscr{L}$-Lipschitz with respect to W_1 distance satisfying*

$$\|D_\rho F(\rho_1, x) - D_\rho F(\rho_2, x)\| \le \mathscr{L} W_1(\rho_1, \rho_2) \tag{35}$$

with $\mathscr{L} := \sup_{\rho'\in\mathcal{P}_2(\mathbb{R}^d)} \sup_{x,x'\in\mathbb{R}^d} \left\|D_\rho^2 F(\rho', x, x')\right\|_{\mathsf{op}}$

Proof. By the definition of functional derivative, we have

$$\|D_\rho F(\rho_1,x) - D_\rho F(\rho_2,x)\| \le \int_0^1 \left\|\int \frac{\delta}{\delta\rho} D_\rho F((1-t)\rho_1 + t\rho_2, x, x')(\rho_1-\rho_2)\mathrm{d}x'\right\| \mathrm{d}t \tag{36}$$

By Kantorovich duality and the definition of $\mathscr{L}$, which is the Liptschiz constant of $\frac{\delta}{\delta\rho}D_\rho F(\cdot, x)$, we obtain

$$\left\|\int \frac{\delta}{\delta\rho} D_\rho F((1-t)\rho_1 + t\rho_2, x, x')(\rho_1-\rho_2)\mathrm{d}x'\right\| \le \mathscr{L} W_1(\rho_1,\rho_2).$$

Combining with (36), we complete the proof. □

Lemma B.3 (Mean-field Entropy Sandwich, Chen et al. 2023, Lemma 4.2)**.** *Assume F satisfies Assumptions 2.3-2.5. Then for every $\mu \in \mathcal{P}_2(\mathbb{R}^{2d})$ we have*

$$\mathsf{KL}(\mu\|\mu_*) \le \mathcal{F}(\mu) - \mathcal{F}(\mu_*) \le \mathsf{KL}(\mu\|\hat{\mu}) \le \left(1 + \frac{\mathscr{L}}{\mathscr{C}_{\mathsf{LSI}}} + \frac{\mathscr{L}^2}{2\mathscr{C}_{\mathsf{LSI}}^2}\right)\mathsf{KL}(\mu\|\mu_*). \tag{37}$$

Lemma B.4 (Particle System's Entropy Inequality, Chen et al. 2023, Lemma 4.2)**.** *Assume that F satisfies Assumption 2.3 and there exists a measure $\mu_* \in \mathcal{P}(\mathbb{R}^{2d})$ that admits the structure $\mu_*(x,v) \propto \exp\left(-\frac{\delta F}{\delta\rho}(\mu_*^X, x) - \frac{1}{2}\|v\|^2\right)$. Then for all $\mu^N \in \mathcal{P}(\mathbb{R}^{2dN})$, we have*

$$\mathsf{KL}(\mu^N\|\mu_*^{\otimes N}) \le \mathcal{F}^N(\mu^N) - N\mathcal{F}(\mu_*). \tag{38}$$

Lemma B.5 (Information Inequality)**.** *Let $X_1, \ldots, X_N$ be measurable spaces, μ be a probability on the product space $X = X_1 \times \ldots \times X_N$ with $\mu = \mu^1 \otimes \ldots \otimes \mu^N$ and $\nu = \nu^1 \otimes \ldots \otimes \nu^N$ is a σ-finite measure. Then*

$$\sum_{i=1}^N \mathsf{KL}(\mu^i\|\nu^i) \le \mathsf{KL}(\mu\|\nu). \tag{39}$$

Lemma B.6 (Matrix Grönwall's Inequality, Zhang et al. 2023)**.** *Let $x : \mathbb{R}_+ \to \mathbb{R}^d$, and $c \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$, where A has non-negative entries. Suppose that the following inequality is satisfied componentwise:*

$$x(t) \le c + \int_0^t Ax(s)\mathrm{d}s, \quad \textit{for all } t \ge 0.$$

Then the following inequality holds where $I_d \in \mathbb{R}^{d\times d}$ is the d-dimensional identity matrix:

$$x(t) \le \left(AA^\dagger e^{At} - AA^\dagger + I_d\right)c.$$

Lemma B.7. *Let $(x_t, v_t)_{t\geq 0}$ and $(x_t^i, v_t^i)_{t\geq 0}$ respectively denote the iterates of the* MULD *and* N-ULD*. Assume that $h \lesssim \mathscr{L}^{-1/2} \wedge \gamma^{-1}$. Under Assumption 2.4 and Assumption 2.6, for $t \in [kh, (k+1)h]$, we have*

$$\sup_{t\in[kh,(k+1)h]} \|x_t - x_{kh}\| \leq 2\mathscr{L}h^2\|x_{kh}\| + 4h\|v_{kh}\| + 2\mathscr{L}h^2 + 2\sqrt{2\gamma}h \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t - \mathrm{B}_{kh}\|$$

$$\sup_{t\in[kh,(k+1)h]} \|x_t^i - x_{kh}^i\| \leq 2\mathscr{L}h^2\|x_{kh}^i\| + 4h\|v_{kh}^i\| + 2\mathscr{L}h^2 + 2\sqrt{2\gamma}h \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t^i - \mathrm{B}_{kh}^i\|$$

for $i = 1, ..., N$.

Proof. We only prove the first relation, and the proof of the second relation is similar.

$$\begin{aligned}
\|x_t - x_{kh}\| &= \left\|\int_{kh}^t v_\tau \mathrm{d}\tau\right\| \leq h\|v_{kh}\| + \left\|\int_{kh}^t v_\tau - v_{kh}\mathrm{d}\tau\right\| \\
&\leq h\|v_{kh}\| + \left\|\int_{kh}^t \int_0^\tau \gamma v_{\tau'}\mathrm{d}\tau'\mathrm{d}\tau\right\| + \left\|\int_{kh}^t\int_{kh}^\tau D_\rho F(\mu_{\tau'}^X, x_{\tau'})\mathrm{d}\tau'\mathrm{d}\tau\right\| + \left\|\int_{kh}^t\int_{kh}^\tau \sqrt{2\gamma}\mathrm{d}\mathrm{B}_{\tau'}\mathrm{d}\tau\right\| \\
&\leq h\|v_{kh}\| + \gamma h\left(h\|v_{kh}\| + \int_{kh}^t \|v_\tau - v_{kh}\|\mathrm{d}\tau\right) + \left\|\int_{kh}^t\int_{kh}^\tau D_\rho F(\mu_{\tau'}^X, x_{\tau'})\mathrm{d}\tau'\mathrm{d}\tau\right\| \\
&\quad + \left\|\int_{kh}^t\int_{kh}^\tau \sqrt{2\gamma}\mathrm{d}\mathrm{B}_{\tau'}\mathrm{d}\tau\right\| \\
&\leq h\|v_{kh}\| + \gamma h\left(h\|v_{kh}\| + \int_{kh}^t \|v_\tau - v_{kh}\|\mathrm{d}\tau\right) + \mathscr{L}h\int_{kh}^t \|x_\tau - x_{kh}\|\,\mathrm{d}\tau + \mathscr{L}h^2\|x_{kh}\| \\
&\quad + \mathscr{L}h^2 + \sqrt{2\gamma}h \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t - \mathrm{B}_{kh}\|
\end{aligned}$$

where the last inequality follows from Assumptions 2.4 and 2.6. Likewise for V:

$$\begin{aligned}
\|v_t - v_{kh}\| &= \left\|\int_{kh}^t \gamma v_\tau \mathrm{d}\tau\right\| + \left\|\int_{kh}^t D_\rho F(\mu_\tau^X, x_\tau)\mathrm{d}\tau\right\| + \left\|\int_{kh}^t \sqrt{2\gamma}\mathrm{d}\mathrm{B}_t\right\| \\
&\leq \gamma\left(h\|v_{kh}\| + \int_{kh}^t \|v_\tau - v_{kh}\|\mathrm{d}\tau\right) + \left\|\int_{kh}^t D_\rho F(\mu_\tau^X, x_\tau)\mathrm{d}\tau\right\| + \sqrt{2\gamma} \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t - \mathrm{B}_{kh}\| \\
&\leq \gamma\left(h\|v_{kh}\| + \int_{kh}^t \|v_\tau - v_{kh}\|\mathrm{d}\tau\right) + \mathscr{L}\int_{kh}^t \|x_\tau - x_{kh}\|\mathrm{d}\tau + \mathscr{L}h + \mathscr{L}h\|x_{kh}\| \\
&\quad + \sqrt{2\gamma} \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t - \mathrm{B}_{kh}\|
\end{aligned}$$

where the last inequality follows from Assumptions 2.4 and 2.6. Before applying matrix form of Grönwall's inequality, let $c = c_1 + c_2$ with $c_2 = \begin{bmatrix} h\|v_{kh}\| \\ 0 \end{bmatrix}$,

$$A = \begin{bmatrix} \mathscr{L}h & \gamma h \\ \mathscr{L} & \gamma \end{bmatrix},\ c_1 = \begin{bmatrix} \mathscr{L}h^2\|x_{kh}\| + \gamma h^2\|v_{kh}\| + \mathscr{L}h^2 + \sqrt{2\gamma}h \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t - \mathrm{B}_{kh}\| \\ \mathscr{L}h\|x_{kh}\| + \gamma h\|v_{kh}\| + \mathscr{L}h + \sqrt{2\gamma} \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t - \mathrm{B}_{kh}\| \end{bmatrix}.$$

c_1 lies in the image space of A, and $\exp(A_t)c_1$ also lies in the image space of A. For the first component:

$$\begin{aligned}
\sup_{t\in[kh,(k+1)h]} \|x_t - x_{kh}\| &\leq h\exp\left((\mathscr{L}h + \gamma)h\right)\left(\mathscr{L}h\|x_{kh}\| + \gamma h\|v_{kh}\| + \mathscr{L}h + \sqrt{2\gamma} \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t - \mathrm{B}_{kh}\|\right) \\
&\quad + \frac{\mathscr{L}h\exp((\mathscr{L}h + \gamma)h) + \gamma}{\mathscr{L}h + \gamma} h\|v_{kh}\| \\
&\leq 2h\left(\mathscr{L}h\|x_{kh}\| + 2\|v_{kh}\| + \mathscr{L}h + \sqrt{2\gamma} \sup_{t\in[kh,(k+1)h]} \|\mathrm{B}_t - \mathrm{B}_{kh}\|\right)
\end{aligned}$$

where the second inequality comes from choosing $h \lesssim \frac{1}{\mathscr{L}^{1/2}} \wedge \frac{1}{\gamma}$.

$$((AA^{\dagger}(\exp(Ah) - I) + I)c_2)_{(1)} = \frac{\mathscr{L}h\exp((\mathscr{L}h + \gamma)h) + \gamma}{\mathscr{L}h + \gamma}h\|v_{kh}\| \le 2h\|v_{kh}\|$$

Combining relations above and Lemma B.6 completes the proof. □

Lemma B.8. *Let $(x_t, v_t)_{t\ge 0}$ denote the iterates of the MULD with $(x_0, v_0) \sim \mu_0 = \mathcal{N}(0, I_{2d})$. Under Assumption 2.7 and Assumption 2.8, we have $\mathbb{E}\|(x_t, v_t)\|^2 \lesssim \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}}$.*

Proof.

$$\begin{aligned}
\mathbb{E}\|(x_t, v_t)\|^2 = W_2^2(\mu_t, \delta_0) &\le 2W_2^2(\mu_t, \mu_*) + 2W_2^2(\mu_*, \delta_0)\\
&\le \frac{2}{\mathscr{C}_{\mathsf{LSI}}}\mathsf{KL}(\mu_t\|\mu_*) + 2\mathbf{m}_2^2\\
&\le \frac{2}{\mathscr{C}_{\mathsf{LSI}}}(\mathcal{F}(\mu_t) - \mathcal{F}(\mu_*)) + 2\mathbf{m}_2^2\\
&\le \frac{2}{\mathscr{C}_{\mathsf{LSI}}}(\mathcal{F}(\mu_0) - \mathcal{F}(\mu_*)) + 2\mathbf{m}_2^2\\
&\le \frac{2}{\mathscr{C}_{\mathsf{LSI}}}\mathcal{F}(\mu_0) + 2\mathbf{m}_2^2
\end{aligned}$$

The second inequality follows from Talagrand's inequality which can be implied by Assumption 2.5.[1] The third inequality follows from Lemma B.3. The fourth inequality follows that $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\mu_t) < 0$ along the MULD (Proof of Theorem 2.1 in Chen et al. (2023)) and the last inequality follows from the assumption that $\mathcal{F}(\mu_*) \ge 0$. By the definition of $\mathcal{F}(\mu)$, we have $\mathcal{F}(\mu_0) = F(\mu_0^X) + \int \frac{1}{2}\|v\|^2\mu_0(\mathrm{d}x\mathrm{d}v) + \mathrm{Ent}(\mu_0)$. Since $(x_0, v_0) \sim \mathcal{N}(0, I_{2d})$, we have $\int \frac{1}{2}\|v\|^2\mu_0(\mathrm{d}x\mathrm{d}v) \lesssim d$ and

$$|\mathrm{Ent}(\mu_0)| = \left|\int \mu_0 \log \mu_0\right| = \frac{d}{2}\log(2\pi) + \frac{1}{2}\mathbb{E}_{\mu_0}\|\cdot\|^2 \lesssim d.$$

By Assumption 2.8, we have $F(\mu_0^X) \lesssim \mathscr{L}d$. By Assumption 2.7, we have $\mathbf{m}_2^2 \lesssim d$. Thus we have

$$\mathbb{E}\|(x_t, v_t)\|^2 \le \frac{2}{\mathscr{C}_{\mathsf{LSI}}}\mathcal{F}(\mu_0) + 2\mathbf{m}_2^2 \lesssim \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}} + d$$

□

Lemma B.9. *Let (x_t^i, v_t^i) denote the iterates of the N-ULD with $(x_0^i, v_0^i) \sim \mu_0^i = \mathcal{N}(0, I_{2d})$ for $i = 1, ..., N$ and $t \ge 0$. Under Assumption 2.7 and Assumption 2.9, we have $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\|(x_t^i, v_t^i)\|^2 \lesssim \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}}$.*

Proof.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \mathbb{E}\|(x_t^i, v_t^i)\|^2 = \frac{1}{N}\sum_{i=1}^N W_2^2(\mu_t^i, \delta_0) &\le \frac{2}{N}\sum_{i=1}^N W_2^2(\mu_t^i, \mu_*) + 2W_2^2(\mu_*, \delta_0)\\
&\le \frac{2}{\mathscr{C}_{\mathsf{LSI}}}\frac{1}{N}\sum_{i=1}^N \mathsf{KL}(\mu_t^i\|\mu_*) + 2\mathbf{m}_2^2\\
&\le \frac{2}{\mathscr{C}_{\mathsf{LSI}}}\frac{1}{N}\mathsf{KL}(\mu_t^N\|\mu_*^{\otimes N}) + 2\mathbf{m}_2^2\\
&\le \frac{2}{\mathscr{C}_{\mathsf{LSI}}}\left(\frac{1}{N}\mathcal{F}^N(\mu_t^N) - \mathcal{F}(\mu_*)\right) + 2\mathbf{m}_2^2\\
&\le \frac{2}{N\mathscr{C}_{\mathsf{LSI}}}\mathcal{F}^N(\mu_0^N) + 2\mathbf{m}_2^2
\end{aligned}$$

[1] Assumption 2.5 states that $\hat{\mu}$ satisfies the LSI. Note that μ_* also has the form of the $\hat{\mu}_*$ and thus satisfies LSI.

The second inequality follows from Talagrand's inequality which can be implied by Assumption 2.5. The third inequality follows from Lemma B.5. The fourth inequality follows from Lemma B.4 and the last inequality follows that $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}^N(\mu_t^N) < 0$ along the N-ULD (Proof of Theorem 2.2 in Chen et al. (2023)) and $\mathcal{F}(\mu_*) \geq 0$. By the definition of $\mathcal{F}^N(\mu^N)$, we have $\mathcal{F}^N(\mu_0^N) = \int (NF(\mu_{\mathbf{x}}) + \frac{1}{2}\|\mathbf{v}\|^2)\mu_0^N(\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}) + \mathrm{Ent}(\mu_0^N)$. Similar to the proof of Lemma B.8, since $(\mathbf{x}, \mathbf{v}) \sim \mathcal{N}(0, I_{2Nd})$, we have $\int \frac{1}{2}\|\mathbf{v}\|^2\mu_0^N(\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}) \lesssim Nd$ and $|\mathrm{Ent}(\mu_0^N)| \lesssim Nd$. By Assumption 2.9 and Assumption 2.7, we also have $\int NF(\mu_{\mathbf{x}})\mu_0^N(\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}) \lesssim N\mathscr{L}d$ and $\mathbf{m}_2^2 \lesssim d$. Thus we have

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \mathbb{E}\|(x_t^i, v_t^i)\|^2 &\leq \frac{2}{N\mathscr{C}_{\mathsf{LSI}}}\mathcal{F}^N(\mu_0^N) + 2\mathbf{m}_2^2 \\
&= \frac{2}{N\mathscr{C}_{\mathsf{LSI}}}\left(\int (NF(\mu_{\mathbf{x}}) + \frac{1}{2}\|\mathbf{v}\|^2)\mu_0^N(\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}) + \mathrm{Ent}(\mu_0^N)\right) + 2\mathbf{m}_2^2 \\
&\lesssim \frac{1}{N\mathscr{C}_{\mathsf{LSI}}}(N\mathscr{L}d + Nd) + d \\
&\lesssim \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}} + d
\end{aligned}
$$

□

Lemma B.10 (Girsanov's Theorem, (Zhang et al. (2023), Theorem 19))**.** *Consider stochastic processes $(x_t)_{t\geq 0}$, $(b_t^{\boldsymbol{P}})_{t\geq 0}$, $(b_t^{\boldsymbol{Q}})_{t\geq 0}$ adapted to the same filtration, and $\sigma \in \mathbb{R}^{d\times d}$ any constant matrix (possibly degenerate). Let $\boldsymbol{P}_T$ and $\boldsymbol{Q}$ be probability measures on the path space $C([0, T]; \mathbb{R}^d)$ such that $(x_t)_{t\geq 0}$ follows*

$$
\begin{aligned}
\mathrm{d}x_t &= b_t^{\boldsymbol{P}}\mathrm{d}t + \sigma\mathrm{d}\mathrm{B}_t^{\boldsymbol{P}} \quad \textit{under } \boldsymbol{P}_T, \\
\mathrm{d}x_t &= b_t^{\boldsymbol{Q}}\mathrm{d}t + \sigma\mathrm{d}\mathrm{B}_t^{\boldsymbol{Q}} \quad \textit{under } \boldsymbol{Q}_T,
\end{aligned}
$$

where $\mathrm{B}^{\boldsymbol{P}}$ and $\mathrm{B}^{\boldsymbol{Q}}$ are $\boldsymbol{P}_T$-Brownian motion and $\boldsymbol{Q}_T$-Brownian motion. Suppose there exists a process $(y_t)_{t\geq 0}$ such that

$$
\sigma y_t = b_t^{\boldsymbol{P}} - b_t^{\boldsymbol{Q}},
$$

and

$$
\mathbb{E}_{\mathbf{Q}_T}\exp\left(\frac{1}{2}\int_0^T \|y_t\|^2\,\mathrm{d}t\right) < \infty.
$$

If we define $\sigma^\dagger$ as the Moore-Penrose pseudo-inverse of σ, then we have

$$
\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{Q}_T} = \exp\left(\int_0^T \langle \sigma_t^\dagger(b_t^{\boldsymbol{P}_T} - b_t^{\boldsymbol{Q}_T}), \mathrm{d}\mathrm{B}_t^{\boldsymbol{Q}_T}\rangle - \frac{1}{2}\int_0^T \|\sigma_t^\dagger(b_t^{\boldsymbol{P}_T} - b_t^{\boldsymbol{Q}_T})\|^2\mathrm{d}t\right)
$$

Besides, $(\tilde{\mathrm{B}}_t)_{t\in[0,T]}$ defined by $\mathrm{d}\tilde{\mathrm{B}}_t := \mathrm{d}\mathrm{B}_t + \sigma_t^\dagger(b_t^Y - b_t^X)$ is a $\mathbf{P}_T$-Brownian motion.

C. Verification of Assumptions

C.1. Verification of Assumption 2.4

Smoothness in W_1 distance has been verified for training mean-field neural networks in Chen et al. (2022). Thus we only verify smoothness in W_1 distance for examples of density estimation via MMD minimization and KSD minimization. Lemma B.2 provides sufficient conditions for smoothness in W_1 distance. In particular, we have

$$
\|D_\rho F(\rho_1, x_1) - D_\rho F(\rho_2, x_2)\| \leq \|D_\rho F(\rho_1, x_1) - D_\rho F(\rho_2, x_1)\| + \|D_\rho F(\rho_2, x_1) - D_\rho F(\rho_2, x_2)\| \tag{40}
$$

Suzuki et al. (2023) verify that $\|D_\rho F(\rho_2, x_1) - D_\rho F(\rho_2, x_2)\| \leq \mathscr{L}\|x_1 - x_2\|$ for three examples mentioned above. Thus it suffices to verify (35) for the last two examples.

MMD Minimization We now prove that objective (18) satisfies Assumption 2.4 with Gaussian RBF kernel. We choose σ' in Gaussian RBF kernel k to be σ for brevity. We reformulate (18) as

$$F(\rho) = \hat{\mathcal{M}}(\rho) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}\|x\|^2. \tag{41}$$

According to the definition of $\hat{\mathcal{M}}$ in Section 4, the intrinsic derivative of F is

$$\begin{aligned} D_\rho F(\rho,x) &= D_\rho\hat{\mathcal{M}}(\rho,x) + \frac{\lambda'}{2}\|x\|^2 \\ &= 2\iiint \nabla_x p(x;z)p(x';z')k(z,z')\mathrm{d}z\mathrm{d}z'\mathrm{d}\rho(x') - \frac{2}{n}\sum_{i=1}^{n}\int \nabla_x p(x;z)k(z,z_i)\mathrm{d}z + \frac{\lambda'}{2}\|x\|^2 \end{aligned}$$

We only need to prove $D_\mu\hat{\mathcal{M}}(\mu,x)$ is smooth. The second-order intrinsic derivative $D_\rho\hat{\mathcal{M}}(\rho,x)$ is

$$\begin{aligned} D_\rho^2\hat{\mathcal{M}}(\rho,x,x') &= 2\iint \nabla_x p(x;z)\otimes\nabla_{x'}p(x';z')k(z,z')\mathrm{d}z\mathrm{d}z' \\ &= \frac{2}{(2\pi\sigma^2)^d\sigma^4}\iint (x-z)\otimes(x'-z')\exp\left(-\frac{\|x-z\|^2+\|x'-z'\|^2+\|z-z'\|^2}{2\sigma^2}\right)\mathrm{d}z\mathrm{d}z' \end{aligned}$$

From the relation $x\cdot\exp(-x^2/2\sigma^2)\le\sigma$ for $x\ge 0$, we have

$$\begin{aligned} \left\|D_\rho^2\hat{\mathcal{M}}(\rho,x,x')\right\| &\le \frac{1}{(2\pi\sigma^2)^d\sigma^4}\iint \|x-z\|\|x'-z'\|\exp\left(-\frac{\|x-z\|^2+\|x'-z'\|^2+\|z-z'\|^2}{2\sigma^2}\right)\mathrm{d}z\mathrm{d}z' \\ &\le \frac{1}{(2\pi\sigma^2)^d\sigma^2}\iint \exp\left(-\frac{\|z-z'\|^2}{2\sigma^2}\right)\mathrm{d}z\mathrm{d}z' = \frac{1}{(2\pi\sigma^2)^{d/2}\sigma^2} \end{aligned}$$

According to Lemma B.2 and (40), F defined in (41) satisfies Assumption 2.4.

KSD Minimization We now prove that objective (19) satisfies Assumption 2.4 with kernel

$$k(x,x') = \exp\left(-\frac{\|x\|^2}{2\sigma_1^2} - \frac{\|x'\|^2}{2\sigma_1^2} - \frac{\|x-x'\|^2}{2\sigma_2^2}\right). \tag{42}$$

We also assume the score function of μ_* satisfies (20). Under this assumption on score function and with this choice of kernel, Suzuki et al. (2023) show in their Appendix A that the Stein kernel u_{ρ_*} satisfies $\sup_{x,x'\in\mathbb{R}^d}\max\{|u_{\rho_*}|, \|\nabla_x u_{\rho_*}\|, \|\nabla_x\nabla_{x'}u_{\rho_*}\|_{\mathsf{op}}\}\le\mathscr{L}$. We reformulate (19) as

$$F(\rho) = \mathsf{KSD}(\rho) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}\|x\|^2. \tag{43}$$

Similarly, we only need to verify that KSD is smooth with respect to W_1 distance. The intrinsic derivative of KSD is

$$D_\rho\mathsf{KSD}(\rho,x) = \int \nabla_x u_{\rho_*}(x,x')\mathrm{d}\rho(x').$$

The second-order intrinsic derivative of $D_\rho\mathsf{KSD}(\rho,x)$ is

$$D_\rho^2\mathsf{KSD}(\rho,x,x') = \nabla_x\nabla_{x'}u_{\rho_*}(x,x')$$

The following relation implies Assumption 2.4 by Lemma B.2.

$$\|D_\rho^2\mathsf{KSD}(\rho,x,x')\| = \|\nabla_x\nabla_{x'}u_{\rho_*}(x,x')\| \le \mathscr{L}$$

C.2. Verification of Assumption 2.7

Training Mean-field Neural Networks Denote $\hat{\mu}(x,v) = \hat{\mu}^X(x) \otimes \mathcal{N}(0, I_d)$ where $\hat{\mu}^X(x) \propto \exp\left(-\frac{\delta F}{\delta \rho}(\mu^X, x)\right)$. Since the second moment of $\mathcal{N}(0, I_d)$ is $O(d)$, it suffices to ensure $\mathbb{E}_{x\sim\hat{\mu}^X}\|x\|^2 = O(d)$. We reformulate objective (17) as:

$$F(\rho) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\rho; a_i), b_i) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}[\|x\|^2]. \tag{44}$$

- We will prove that Assumption 2.7 holds if $|h(x;a)| \le \sqrt{\mathscr{L}}$ (such activation functions include tanh and $\mathsf{sigmoid}$) and $|\partial_1 \ell| \le \sqrt{\mathscr{L}}$ (such loss functions include logistic loss, Huber loss and log-cosh loss) or ℓ is quadratic. The functional derivative of F is
$$\frac{\delta F}{\delta \rho}(\mu^X, x) = \frac{1}{n}\sum_{i=1}^{n}\left[\partial_1 \ell(h(\mu^X; a_i), b_i) h(x; a_i)\right] + \frac{\lambda'}{2}\|x\|^2$$
Consider the case where $|\partial_1 \ell| \le \sqrt{\mathscr{L}}$. Since $|h(x;a)| \le \sqrt{\mathscr{L}}$, we have $|\partial_1 \ell(h(\mu^X; a_i), b_i) h(x; a_i)| \le \mathscr{L}$. Let $Z = \int \exp\left(-\frac{\delta F}{\delta \rho}(\mu^X, x)\right) \mathrm{d}x$, and we have
$$\mathbb{E}_{\hat{\mu}^X}\|\cdot\|^2 = \frac{1}{Z}\int \|x\|^2 \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\left[\partial_1 \ell(h(\mu^X; a_i), b_i) h(x; a_i)\right] - \frac{\lambda'}{2}\|x\|^2\right) \mathrm{d}x \triangleq \frac{Z'}{Z} \tag{45}$$
Now we bound Z' and Z respectively.
$$Z' \le \int \|x\|^2 \exp\left(\mathscr{L} - \frac{\lambda'}{2}\|x\|^2\right) \mathrm{d}x \lesssim \frac{\exp(\mathscr{L}) d}{\lambda'},$$
$$Z \ge \int \exp\left(-\mathscr{L} - \frac{\lambda'}{2}\|x\|^2\right) \mathrm{d}x = \exp(-\mathscr{L})\left(\frac{2\pi}{\lambda'}\right)^{d/2}$$
Choose $\lambda' \le (2\pi)^3 \exp(-4\mathscr{L})$ which implies $\lambda' \le \frac{(2\pi)^{\frac{d}{d-2}}}{\exp\left(\frac{4\mathscr{L}}{d-2}\right)}$, and we have $\mathbb{E}_{\hat{\mu}^X}\|\cdot\|^2 = \frac{Z'}{Z} \lesssim \frac{\exp(2\mathscr{L})}{\lambda'\left(\frac{2\pi}{\lambda'}\right)^{d/2}} d \le d$. Consider the case where ℓ is quadratic. $|h(\mu^X; a_i)| = |\int h(x; a_i)\mu^X(\mathrm{d}x)| \le \int |h(x; a_i)|\mu(\mathrm{d}x) \le \sqrt{\mathscr{L}}$, thus we have $|\partial_1 \ell(h(\mu^X; a_i), b_i) h(x; a_i)| = |(h(\mu^X; a_i) - b_i) h(x; a_i)| \le \mathscr{L} + |b_i|\sqrt{\mathscr{L}}$. We can scale the label to ensure $\max_{i=1}^n |b_i| \le \sqrt{\mathscr{L}}$, and we obtain $|\partial_1 \ell(h(\mu^X; a_i), b_i) h(x; a_i)| \le 2\mathscr{L}$. The remaining proof keeps the same with $\lambda' \le (2\pi)^3 \exp(-8\mathscr{L})$.

- We will prove that Assumption 2.7 holds if $|h(x;a)| \le \sqrt{\mathscr{L}}(1 + \|x\|)$ (such activation functions include ReLU, GeLU, Softplus, SiLU) and $|\partial_1 \ell| \le \sqrt{\mathscr{L}}$. Under these conditions, we have $|\partial_1 \ell(h(\mu^X; a_i), b_i) h(x; a_i)| \le \mathscr{L}(1 + \|x\|)$. Then, based on (45), we obtain
$$\begin{aligned} Z' &\le \int \|x\|^2 \exp\left(\mathscr{L}(1 + \|x\|) - \frac{\lambda'}{2}\|x\|^2\right) \mathrm{d}x \le \exp(\mathscr{L}) \int \|x\|^2 \exp\left(\frac{3\mathscr{L}^2}{2\lambda'} - \frac{\lambda'}{3}\|x\|^2\right) \mathrm{d}x \\ &\lesssim \exp\left(\mathscr{L} + \frac{3\mathscr{L}^2}{2\lambda'}\right)\frac{d}{\lambda'}. \end{aligned}$$
We also have
$$\begin{aligned} Z &\ge \int \exp\left(-\mathscr{L}(1 + \|x\|) - \frac{\lambda'}{2}\|x\|^2\right) \mathrm{d}x \ge \exp(\mathscr{L}) \int \exp\left(-\frac{\mathscr{L}^2}{\lambda'} - \frac{3\lambda'}{4}\|x\|^2\right) \mathrm{d}x \\ &= \exp\left(\mathscr{L} - \frac{\mathscr{L}^2}{\lambda'}\right)\left(\frac{4\pi}{3\lambda'}\right)^{d/2} \end{aligned}$$
Combining the upper bound of Z' and the lower bound of Z, if $d \ge \frac{5\mathscr{L}^2}{\lambda'}\left(\log \frac{4\pi}{3}\right)^{-1}$, we obtain
$$\mathbb{E}_{\hat{\mu}^X}\|\cdot\|^2 = \frac{Z'}{Z} \lesssim \exp\left(\frac{5\mathscr{L}^2}{2\lambda'}\right)\frac{d}{\lambda'}\left(\frac{3\lambda'}{4\pi}\right)^{d/2} \le \exp\left(\frac{5\mathscr{L}^2}{2\lambda'}\right)\left(\frac{3}{4\pi}\right)^{d/2} d \le d.$$
Note that $d \ge \frac{5\mathscr{L}^2}{\lambda'}\left(\log \frac{4\pi}{3}\right)^{-1}$ is possible for large-scale problems.

MMD Minimization We now prove that objective (18) satisfies Assumption 2.7 with Gaussian RBF kernel. We choose σ' in Gaussian RBF kernel k to be σ for brevity. We reformulate (18) as

$$F(\rho) = \hat{\mathcal{M}}(\rho) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}\|x\|^2. \tag{46}$$

According to the definition of $\hat{\mathcal{M}}(\rho)$ in Section 4, the functional derivative of $\hat{\mathcal{M}}(\rho)$ is

$$\frac{\delta\hat{\mathcal{M}}}{\delta\rho}(\rho,x) = \underbrace{2\iiint p(x;z)p(x';z')k(z,z')\mathrm{d}z\mathrm{d}z'\mathrm{d}\rho(x')}_{\mathsf{P}} - \underbrace{\frac{2}{n}\sum_{i=1}^{n}\int p(x;z)k(z,z_i)\mathrm{d}z}_{\mathsf{Q}} \tag{47}$$

Next we bound each part of $\frac{\delta\hat{\mathcal{M}}}{\delta\rho}(\rho,x)$. For P, we have

$$\begin{aligned}
\frac{1}{2}\mathsf{P} &= \frac{1}{(2\pi\sigma^2)^d}\iiint \exp\left(-\frac{\|x-z\|^2}{2\sigma^2} - \frac{\|x'-z'\|^2}{2\sigma^2} - \frac{\|z-z'\|^2}{2\sigma^2}\right)\mathrm{d}z\mathrm{d}z'\mathrm{d}\rho(x') \\
&= \frac{(\pi\sigma^2)^{\frac{d}{2}}}{(2\pi\sigma^2)^d}\iint \exp\left(-\frac{\|x-x'\|^2}{6\sigma^2} - \frac{3\|z'-\frac{2}{3}x'-\frac{1}{3}x\|^2}{4\sigma^2}\right)\mathrm{d}z'\mathrm{d}\rho(x') \\
&= \left(\frac{1}{\sqrt{3}}\right)^d \int \exp\left(-\frac{\|x-x'\|^2}{6\sigma^2}\right)\mathrm{d}\rho(x') \le \left(\frac{1}{\sqrt{3}}\right)^d
\end{aligned}$$

where the last inequality follows from the relation $\exp\left(-\frac{\|x-x'\|^2}{6\sigma^2}\right) \le 1$. For Q, we have

$$\begin{aligned}
\frac{1}{2}\mathsf{Q} &= \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\frac{1}{n}\sum_{i=1}^{n}\int \exp\left(-\frac{\|x-z\|^2}{2\sigma^2} - \frac{\|z-z_i\|^2}{2\sigma^2}\right)\mathrm{d}z \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{\|x\|^2+\|z_i\|^2}{2\sigma^2} + \frac{\|z_i+x\|^2}{4\sigma^2}\right)\int \exp\left(-\frac{\|z-\frac{1}{2}z_i-\frac{1}{2}x\|^2}{\sigma^2}\right)\mathrm{d}z \\
&= \left(\frac{1}{\sqrt{2}}\right)^d \frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{\|x\|^2+\|z_i\|^2}{2\sigma^2} + \frac{\|z_i+x\|^2}{4\sigma^2}\right) \le \left(\frac{1}{\sqrt{2}}\right)^d
\end{aligned}$$

where the last inequality follows from the relation $\|z_i+x\|^2 \le 2\|z_i\|^2 + 2\|x\|^2$. Note that $\mathsf{P} \ge 0$ and $\mathsf{Q} \ge 0$. Combining the bound of P and Q, we obtain the bound of $\frac{\delta\hat{\mathcal{M}}}{\delta\rho}(\rho,x)$ as follows:

$$-\sqrt{2} \le -2\left(\frac{1}{\sqrt{2}}\right)^d \le \frac{\delta\hat{\mathcal{M}}(\mu)}{\delta\mu}(x) = \mathsf{P}-\mathsf{Q} \le 2\left(\frac{1}{\sqrt{3}}\right)^d \le \sqrt{3} \tag{48}$$

Let $\hat{\mu}^X(x) = \exp\left(-\frac{\delta F}{\delta\rho}(\mu^X,x)\right)/Z$ where $Z = \int \exp\left(-\frac{\delta F}{\delta\rho}(\mu^X,x)\right)\mathrm{d}x$, and we have

$$\mathbb{E}_{\hat{\mu}^X}\|\cdot\|^2 = \frac{1}{Z}\int \|x\|^2\exp\left(-\frac{\delta\hat{\mathcal{M}}}{\delta\rho}(\mu^X,x) - \frac{\lambda'}{2}\|x\|^2\right)\mathrm{d}x \triangleq \frac{Z'}{Z} \tag{49}$$

Now we bound Z' and Z respectively.

$$\begin{aligned}
Z' &\le \int \|x\|^2\exp\left(\sqrt{2} - \frac{\lambda'}{2}\|x\|^2\right)\mathrm{d}x \lesssim \frac{\exp(\sqrt{2})d}{\lambda'}, \\
Z &\ge \int \exp\left(-\sqrt{3} - \frac{\lambda'}{2}\|x\|^2\right)\mathrm{d}x = \exp(-\sqrt{3})\left(\frac{2\pi}{\lambda'}\right)^{d/2}
\end{aligned}$$

Thus in order to ensure $\mathbb{E}_{\hat{\mu}^X}\|\cdot\|^2 = \frac{Z'}{Z} \lesssim \frac{\exp(\sqrt{2}+\sqrt{3}){\lambda'}^{\frac{d-2}{2}}}{(2\pi)^{\frac{d}{2}}}d \le d$, it suffices to choose $\lambda' \le 3\pi/25$.

KSD Minimization Assume the score function s_{ρ_*} satisfies (20) and choose the kernel k to be (42), and the Stein kernel u_{ρ_*} satisfies $\sup_{x,x'\in\mathbb{R}^d}\max\{|u_{\rho_*}|, \|\nabla_x u_{\rho_*}\|, \|\nabla_x^2 u_{\rho_*}\|_{\mathsf{op}}\} \le \mathscr{L}$ (Suzuki et al., 2023). We now prove the following objective

$$F(\rho) = \mathsf{KSD}(\rho) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}\|x\|^2 \tag{50}$$

satisfies Assumption 2.7, with KSD defined by $\mathsf{KSD}(\rho) = \iint u_{\rho_*}(x,x')\mathrm{d}\rho(x)\mathrm{d}\rho(x')$. The functional derivative of KSD is

$$\frac{\delta\mathsf{KSD}}{\delta\rho}(\rho,x) = \int u_{\rho_*}(x,x')\mathrm{d}\rho(x').$$

The functional derivative is bounded as

$$\left|\frac{\delta\mathsf{KSD}}{\delta\rho}(\rho,x)\right| \le \int |u_{\rho_*}(x,x')|\mathrm{d}\rho(x') \le \mathscr{L}.$$

Let $\hat{\mu}^X(x) = \exp\left(-\frac{\delta F}{\delta\rho}(\mu^X,x)\right)/Z$ where $Z = \int\exp\left(-\frac{\delta F}{\delta\rho}(\mu^X,x)\right)\mathrm{d}x$, and we have

$$\mathbb{E}_{\hat{\mu}^X}\|\cdot\|^2 = \frac{1}{Z}\int\|x\|^2\exp\left(-\frac{\delta\mathsf{KSD}}{\delta\rho}(\mu^X,x) - \frac{\lambda'}{2}\|x\|^2\right)\mathrm{d}x \triangleq \frac{Z'}{Z} \tag{51}$$

Now we bound Z' and Z respectively.

$$Z' \le \int\|x\|^2\exp\left(\mathscr{L} - \frac{\lambda'}{2}\|x\|^2\right)\mathrm{d}x \lesssim \frac{\exp(\mathscr{L})d}{\lambda'},$$
$$Z \ge \int\exp\left(-\mathscr{L} - \frac{\lambda'}{2}\|x\|^2\right)\mathrm{d}x = \exp(-\mathscr{L})\left(\frac{2\pi}{\lambda'}\right)^{d/2}$$

Thus we have $\mathbb{E}_{\hat{\mu}^X}\|\cdot\|^2 = \frac{Z'}{Z} \lesssim \frac{\exp(2\mathscr{L})d\lambda'^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}}} \le d$ for $\lambda' \le (2\pi)^3\exp(-4\mathscr{L})$.

C.3. Verification of Assumption 2.8

Training Mean-field Neural Networks Reformulate the objective (17) with $\mu_0 = \mathcal{N}(0, I_d)$:

$$F(\rho) = \frac{1}{n}\sum_{i=1}^n \ell(h(\rho;a_i),b_i) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}[\|x\|^2].$$

- If l is $\sqrt{\mathscr{L}}$-Lipschitz, we have $|\ell(h(\rho;a),b)| \le \sqrt{\mathscr{L}}|h(\rho;a)-b|$. If $|h(x;a)| \le \sqrt{\mathscr{L}}$, we have $|h(\rho;a)| \le \sqrt{\mathscr{L}}$. Since $\mu_0 = \mathcal{N}(0,I_{2d})$, $\mathbb{E}_{x\sim\mu_0^X}[\|x\|^2] \lesssim d$. With $\lambda' \le \min\{\mathscr{L}, d\}$, we have $F(\mu_0^X) \lesssim \sqrt{\mathscr{L}}(\sqrt{\mathscr{L}} + \max_{i=1}^n|b_i|) + d$. We can normalize the data samples to ensure $\max_{i=1}^n|b_i| \lesssim d \wedge \sqrt{\mathscr{L}}$. Thus $F(\mu_0^X) \lesssim \mathscr{L} + d$.
- If $|h(x;a)| \le \sqrt{\mathscr{L}}(1+\|x\|)$, we have $|h(\mu_0^X;a)| \le \sqrt{\mathscr{L}}\int(1+\|x\|)\mu_0^X(\mathrm{d}x) \lesssim \sqrt{\mathscr{L}}d^{1/2}$. If ℓ is $\sqrt{\mathscr{L}}$-Lipschitz, we have $|l(h(\mu_0^X;a_i),b_i)| \le \sqrt{\mathscr{L}}|h(\mu_0^X;a_i)-b_i| \lesssim \mathscr{L}d^{1/2} + \sqrt{\mathscr{L}}\max_{i=1}^n|b_i|$. We can normalize the data samples to ensure $\max_{i=1}^n|b_i| \lesssim d\wedge\sqrt{\mathscr{L}}$. Thus we have $F(\mu_0^X) \lesssim \mathscr{L}d + d$.

MMD Minimization Reformulate the objective (18) with Gaussian RBF kernel ($\sigma' = \sigma$) and $\mu_0 = \mathcal{N}(0, I_d)$:

$$F(\rho) = \hat{\mathcal{M}}(\rho) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}\|x\|^2, \tag{52}$$

where

$$\begin{aligned}\hat{\mathcal{M}}(\rho) &= \iiint p(x;z)p(x';z')k(z,z')\mathrm{d}z\mathrm{d}z'\mathrm{d}(\rho\times\rho)(x,x') - 2\int\left(\frac{1}{n}\sum_{i=1}^n\int p(x;z)k(z,z_i)\mathrm{d}z\right)\mathrm{d}\rho(x)\\ &= \frac{1}{3^{d/2}}\int\exp\left(-\frac{\|x-x'\|^2}{6\sigma^2}\right)\mathrm{d}(\rho\times\rho)(x,x') - \frac{2}{2^{d/2}}\frac{1}{n}\sum_{i=1}^n\int\exp\left(-\frac{\|x-z_i\|^2}{4\sigma^2}\right)\mathrm{d}\rho(x)\\ &\le \frac{1}{3^{d/2}}\int\exp\left(-\frac{\|x-x'\|^2}{6\sigma^2}\right)\mathrm{d}(\rho\times\rho)(x,x') \le \frac{1}{3^{d/2}} \le \mathscr{L}\end{aligned}$$

Thus $F(\mu_0^X) = \hat{\mathcal{M}}(\mu_0^X) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\mu_0^X}\|x\|^2 \lesssim \mathscr{L} + d$, which satisfies Assumption 2.8.

KSD Minimization Consider the same objective in (19) with $\mu_0 = \mathcal{N}(0, I_d)$:

$$F(\rho) = \mathsf{KSD}(\rho) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}\|x\|^2.$$

If we choose kernel $k(x,x') = \exp\left(-\frac{\|x\|^2}{2\sigma_1^2} - \frac{\|x'\|^2}{2\sigma_1^2} - \frac{\|x-x'\|^2}{2\sigma_2^2}\right)$ and assume the score function of ρ_* satisfies $\max\{\|\nabla \log\rho_*(x)\|, \|\nabla^{\otimes 2}\log\rho_*(x)\|_{\mathsf{op}}, \|\nabla^{\otimes 3}\log\rho_*(x)\|_{\mathsf{op}}\} \leq \mathscr{L}(1+\|x\|)$, then the Stein kernel u_{ρ_*} satisfies $\sup_{x,x'\in\mathbb{R}^d}\max\{|u_{\rho_*}|, \|\nabla_x u_{\rho_*}\|, \|\nabla_x^2 u_{\rho_*}\|_{\mathsf{op}}\} \leq \mathscr{L}$ according to the statement of Appendix A in Suzuki et al. (2023). We have

$$\begin{aligned}
F(\mu_0^X) &= \mathsf{KSD}(\mu_0^X) + \frac{\lambda'}{2}\mathbb{E}_{x\sim\mu_0^X}\|x\|^2 \\
&= \iint u_{\rho_*}(x,x')\mathrm{d}\mu^X(x)\mathrm{d}\mu^X(x') + \frac{\lambda'}{2}\mathbb{E}_{x\sim\mu_0^X}\|x\|^2 \\
&\lesssim \mathscr{L} + d,
\end{aligned}$$

which satisfies Assumption 2.8.

C.4. Verification of Assumption 2.9

Training Mean-field Neural Networks Similar to examples of training mean-field neural networks above, we initialize $\mu_0^N = \mathcal{N}(0, I_{2Nd})$.

$$\mathbb{E}_{\mathbf{x}\sim\mu^N}F(\mu_{\mathbf{x}}) := \mathbb{E}_{\mathbf{x}\sim\mu^N}\frac{1}{n}\sum_{i=1}^{n}\left[\ell\left(\frac{1}{N}\sum_{s=1}^{N}h(x^s;a_i), b_i\right)\right] + \frac{\lambda'}{2}\mathbb{E}_{\mathbf{x}\sim\mu^N}\frac{1}{N}\sum_{s=1}^{N}\left[\|x^s\|^2\right],$$

where $\mathbf{x} = (x^1, ..., x^N)$, $x^i \sim \mu^i$ for $i = 1, ..., N$ and $\mu^N = \otimes_{i=1}^N \mu^i = \mathrm{Law}(x^1, ..., x^N)$.

- If $|h(x;a)| \leq \sqrt{\mathscr{L}}$ and ℓ is $\sqrt{\mathscr{L}}$-Lipschitz, and $\mathbb{E}_{\mathbf{x}_0\sim\mu_0^N}\frac{1}{n}\sum_{i=1}^n\left[\ell\left(\frac{1}{N}\sum_{i=1}^N h(x_0^i;a_i), b_i\right)\right] \lesssim \sqrt{\mathscr{L}}(\sqrt{\mathscr{L}} + \max_{i=1}^n |b_i|)$ and thus $\mathbb{E}_{\mu_0^N}F(\mu_{\mathbf{x}_0}) \lesssim \mathscr{L} + \sqrt{\mathscr{L}}\max_{i=1}^n|b_i| + d$. We can normalize the data samples to ensure $\max_{i=1}^n|b_i| \lesssim d \wedge \sqrt{\mathscr{L}}$. Thus we have $\mathbb{E}_{\mu_0^N}F(\mu_{\mathbf{x}_0}) = O(\mathscr{L} + d)$.
- If $|h(x;a)| \leq \sqrt{\mathscr{L}}(1 + \|x\|)$ and ℓ is $\sqrt{\mathscr{L}}$-Lipschitz, $\mathbb{E}_{\mathbf{x}_0\sim\mu_0^N}\frac{1}{n}\sum_{i=1}^n\left[\ell\left(\frac{1}{N}\sum_{s=1}^N h(x_0^s;a_i), b_i\right)\right] \leq \sqrt{\mathscr{L}}\left(\sqrt{\mathscr{L}}\frac{1}{N}\sum_{s=1}^N(1 + \mathbb{E}_{\mathbf{x}_0\sim\mu_0^N}\|x_0^s\|) + \max_{i=1}^n|b_i|\right) \lesssim \mathscr{L}d^{1/2} + \sqrt{\mathscr{L}}\max_{i=1}^n|b_i|$. We can normalize the data samples to ensure $\max_{i=1}^n|b_i| \lesssim d \wedge \sqrt{\mathscr{L}}$. Thus we have $\mathbb{E}_{\mu_0^N}F(\mu_{\mathbf{x}_0}) = O(\mathscr{L}d + d)$

MMD Minimization Now we verify Assumption 2.9 for the example of density estimation. We consider the N-particle approximation of the objective (52) with the initialization $\mu_0^N = \mathcal{N}(0, I_{Nd})$.

$$\begin{aligned}
&\mathbb{E}_{\mu^N}\hat{\mathcal{M}}(\mu_{\mathbf{x},\mathbf{y}}) \\
&\quad := \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mu^N}\left[\frac{1}{N^2}\sum_{s=1}^N\sum_{t=1}^N\iint p(x^s;z)p(y^t;z')k(z,z')\mathrm{d}z\mathrm{d}z' - \frac{2}{nN}\sum_{i=1}^n\sum_{s=1}^N\int p(x^s;z)k(z,z_i)\mathrm{d}z\right] \\
&\quad \leq \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mu^N}\left[\frac{1}{N^2}\sum_{s=1}^N\sum_{t=1}^N\iint p(x^s;z)p(y^t;z')k(z,z')\mathrm{d}z\mathrm{d}z'\right] \\
&\quad = \left(\frac{1}{\sqrt{3}}\right)^d\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mu^N}\left[\frac{1}{N^2}\sum_{s=1}^N\sum_{t=1}^N\exp\left(-\frac{\|x^s - y^t\|^2}{6\sigma^2}\right)\right] \leq \left(\frac{1}{\sqrt{3}}\right)^d \leq \mathscr{L}
\end{aligned}$$

where $\mathbf{x} = (x^1, ..., x^N)$ and $\mathbf{y} = (y^1, ..., y^N)$. Thus we can upper bound $\mathbb{E}_{\mathbf{x}_0,\mathbf{y}_0\sim\mu_0^N}F(\mu_{\mathbf{x}_0,\mathbf{y}_0})$ as follows:

$$\mathbb{E}_{\mathbf{x}_0,\mathbf{y}_0\sim\mu_0^N}F(\mu_{\mathbf{x}_0,\mathbf{y}_0}) = \mathbb{E}_{\mathbf{x}_0,\mathbf{y}_0\sim\mu_0^N}\hat{\mathcal{M}}(\mu_{\mathbf{x},\mathbf{y}}) + \frac{\lambda'}{2}\mathbb{E}_{\mathbf{x}_0\sim\mu_0^N}\frac{1}{N}\sum_{s=1}^N\left[\|x_0^s\|^2\right] \lesssim \mathscr{L} + d$$

which satisfies Assumption 2.9.

KSD Minimization Similar to the verification of Assumption 2.8 above, we have the following relation for $\mu_0^N = \mathcal{N}(0, I_{Nd})$ under the same assumptions on the score function and kernel:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_0\sim\mu_0}F(\mu_{\mathbf{x}_0}) &= \mathsf{KSD}(\mu_{\mathbf{x}_0}) + \frac{\lambda'}{2}\mathbb{E}_{\mathbf{x}_0\sim\mu_0^N}\frac{1}{N}\sum_{s=1}^{N}\left[\|x_0^s\|^2\right] \\ &= \mathbb{E}_{\mathbf{x}_0\sim\mu_0}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}u_{\mu_*}(x_0^i, x_0^j) + \frac{\lambda'}{2}\mathbb{E}_{\mathbf{x}_0\sim\mu_0^N}\frac{1}{N}\sum_{s=1}^{N}\left[\|x_0^s\|^2\right] \lesssim \mathscr{L} + d,\end{aligned}$$

which satisfies Assumption 2.9.

D. Continuous-time Results

In this section, we give the explicit rate of Theorem 2.1 and Theorem 2.2 proposed by Chen et al. (2023) with a specific choice of parameters and then provide the detailed proof of Theorem 3.1 and Theorem 3.2 by reparameterizing γ.

D.1. Proof of Theorem 3.1

Our proof is directly adapted from Theorem 2.1 in Chen et al. (2023) using hypocoercivity in Villani (2009). Chen et al. (2023) prove the Lyapunov functional

$$\mathcal{E}(\mu_t) = \mathcal{F}(\mu_t) + \mathsf{FI}_S(\mu_t\|\hat{\mu}_t) \tag{53}$$

is decaying along the MULD with $S = \begin{pmatrix} c & b \\ b & a \end{pmatrix} \otimes I_d$ and $\gamma = 1$. Let $A_t = \nabla_v$, $B_t = v\cdot\nabla_x - D_\rho F(\mu_t^x, x)\cdot\nabla_v$, $C_t = [A_t, B_t] = A_tB_t - B_tA_t = \nabla_x$ and

$$Y_t = (\|A_tu_t\|, \|A_t^2u_t\|, \|C_tu_t\|, \|C_tA_tu_t\|)^\mathsf{T} \tag{54}$$

where $u_t = \log\frac{\mu_t}{\hat{\mu}_t}$ and $\|\cdot\| := \|\cdot\|_{L^2(\mu_t)}$. More specifically, Chen et al. (2023) prove that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\mu_t) \le -Y_t^\mathsf{T}\mathcal{K}Y_t, \tag{55}$$

where

$$\mathcal{K} = \begin{pmatrix} 1+2a-4\mathscr{L}b & -2b & -2a-2\mathscr{L}c & 0 \\ 0 & 2a & -2\mathscr{L}c & -4b \\ 0 & 0 & 2b & 0 \\ 0 & 0 & 0 & 2c \end{pmatrix}.$$

The choice of a, b, c should satisfies $ac > b^2$ and $K \succ 0$. If we choose $a = c = 2\mathscr{L}$ and $b = 1$, the smallest eigenvalue of $\mathcal{K}$ is $\lambda_{\mathsf{min}}(\mathcal{K}) = 1$, and thus we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{E}(\mu_t) - \mathcal{E}(\mu_*)) &\le -(\|A_tu_t\|^2 + \|A_t^2u_t\|^2 + \|C_tu_t\|^2 + \|C_tA_tu_t\|^2) \\ &\le -(\|A_tu_t\|^2 + \|C_tu_t\|^2) = -\frac{1}{2}\mathsf{FI}(\mu_t\|\hat{\mu}_t) - \frac{1}{2}\mathsf{FI}(\mu_t\|\hat{\mu}_t) \\ &\le -\mathscr{C}_{\mathsf{LSI}}\mathsf{KL}(\mu_t\|\hat{\mu}_t) - \frac{1}{2\lambda_{\mathsf{max}}(S)}\mathsf{FI}_S(\mu_t\|\hat{\mu}_t) \\ &\le -\mathscr{C}_{\mathsf{LSI}}(\mathcal{F}(\mu_t) - \mathcal{F}(\mu_*)) - \frac{1}{4\mathscr{L}+2}\mathsf{FI}_S(\mu_t\|\hat{\mu}_t) \\ &\le -\frac{\mathscr{C}_{\mathsf{LSI}}}{6\mathscr{L}}(\mathcal{E}(\mu_t) - \mathcal{E}(\mu_*))\end{aligned}$$

Applying Grönwall's inequality, we obtain

$$\mathcal{F}(\mu_t) - \mathcal{F}(\mu_*) \le \mathcal{E}(\mu_t) - \mathcal{E}(\mu_*) \le (\mathcal{E}(\mu_0) - \mathcal{E}(\mu_*))\exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{6\mathscr{L}}t\right). \tag{56}$$

Note that the proof in Chen et al. (2023) also considers the approximation technique to remove some restrictive assumptions they make, which we omit in our proof. Now we consider a more general γ in the proof above. Analogous to the proof of Lemma 32 in Villani (2009), if we incorporate a general γ, the diagonal elements of upper triangular matrix $\mathcal{K}$ will become $(\gamma + 2\gamma a - 4\mathscr{L} b,\ 2\gamma a,\ 2b,\ 2\gamma c)$. If we choose $\gamma = \sqrt{\mathscr{L}}$, $b = 1/\sqrt{\mathscr{L}}$, $a = 2$ and $c = 1/\mathscr{L}$, the smallest eigenvalue of K will become $\lambda_{\mathsf{min}}(\mathcal{K}) = 2/\sqrt{\mathscr{L}}$. Similar to the previous proof, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{E}(\mu_t) - \mathcal{E}(\mu_*)) &\le -\frac{2}{\sqrt{\mathscr{L}}}(\|A_t u_t\|^2 + \|A_t^2 u_t\|^2 + \|C_t u_t\|^2 + \|C_t A_t u_t\|^2) \\
&\le -\frac{2}{\sqrt{\mathscr{L}}}(\|A_t u_t\|^2 + \|C_t u_t\|^2) = -\frac{1}{\sqrt{\mathscr{L}}}\mathsf{FI}(\mu_t\|\hat{\mu}_t) - \frac{1}{\sqrt{\mathscr{L}}}\mathsf{FI}(\mu_t\|\hat{\mu}_t) \\
&\le -\frac{2\mathscr{C}_{\mathsf{LSI}}}{\sqrt{\mathscr{L}}}\mathsf{KL}(\mu_t\|\hat{\mu}_t) - \frac{1}{\lambda_{\mathsf{max}}(S)\sqrt{\mathscr{L}}}\mathsf{FI}_S(\mu_t\|\hat{\mu}_t) \\
&\le -\frac{2\mathscr{C}_{\mathsf{LSI}}}{\sqrt{\mathscr{L}}}(\mathcal{F}(\mu_t) - \mathcal{F}(\mu_*)) - \frac{1}{3\sqrt{\mathscr{L}}}\mathsf{FI}_S(\mu_t\|\hat{\mu}_t) \\
&\le -\frac{\mathscr{C}_{\mathsf{LSI}}}{3\sqrt{\mathscr{L}}}(\mathcal{E}(\mu_t) - \mathcal{E}(\mu_*))
\end{aligned}$$

where the fourth inequality follows from $\lambda_{\mathsf{max}}(S) = \frac{\frac{1}{\mathscr{L}}+2+\sqrt{\frac{1}{\mathscr{L}^2}+4}}{2} \le \frac{1}{\mathscr{L}} + 2 \le 3$. Applying Grönwall's inequality, we obtain

$$\mathcal{F}(\mu_t) - \mathcal{F}(\mu_*) \le \mathcal{E}(\mu_t) - \mathcal{E}(\mu_*) \le (\mathcal{E}(\mu_0) - \mathcal{E}(\mu_*)) \exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{3\sqrt{\mathscr{L}}}t\right), \tag{57}$$

which completes the proof of Theorem 3.1. Eq. (57) exhibits a faster rate than the rate of Eq. (56).

D.2. Proof of Theorem 3.2

Our proof is directly adapted from Theorem 2.2 in Chen et al. (2023) using hypocoercivity in Villani (2009). Chen et al. (2023) prove that the Lyapunov functional

$$\mathcal{E}^N(\mu_t^N) = \mathcal{F}^N(\mu_t^N) + \mathsf{FI}_S^N(\mu_t^N\|\mu_*^N) \tag{58}$$

is decaying along the N-ULD with $S = \begin{pmatrix} c & b \\ b & a \end{pmatrix} \otimes I_d$ and $\gamma = 1$. Let $u_t^N = \log\frac{\mu_t^N}{\mu_*^N}$, $\|\cdot\| := \|\cdot\|_{L^2(\mu_t^N)}$ and

$$Y_t^N = (\|\nabla_{\mathbf{v}} u_t^N\|, \|\nabla_{\mathbf{v}}^2 u_t^N\|, \|\nabla_{\mathbf{x}} u_t^N\|, \|\nabla_{\mathbf{x}}\nabla_{\mathbf{v}} u_t^N\|)^{\mathsf{T}}.$$

Chen et al. (2023) prove that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^N(\mu_t^N) \le -(Y_t^N)^{\mathsf{T}}\mathcal{K}Y_t^N \tag{59}$$

where

$$\mathcal{K} = \begin{pmatrix} 1 + 2a - 4\mathscr{L}b & -2b & -2a & 0 \\ 0 & 2a & -4\mathscr{L}c & -4b \\ 0 & 0 & 2b & 0 \\ 0 & 0 & 0 & 2c \end{pmatrix}.$$

The choice of a, b, c should satisfies $ac > b^2$ and $K \succ 0$. If we choose $a = c = 2\mathscr{L}$ and $b = 1$, the smallest eigenvalue of K is $\lambda_{\mathsf{min}}(\mathcal{K}) = 1$, and thus we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^N(\mu_t^N) &\le -(\|\nabla_{\mathbf{v}} u_t^N\|^2 + \|\nabla_{\mathbf{v}}^2 u_t^N\|^2 + \|\nabla_{\mathbf{x}} u_t^N\|^2 + \|\nabla_{\mathbf{x}}\nabla_{\mathbf{v}} u_t^N\|^2) \\
&\le -(\|\nabla_{\mathbf{v}} u_t^N\|^2 + \|\nabla_{\mathbf{x}} u_t^N\|^2) = -\mathsf{FI}(\mu_t^N\|\mu_*^N)
\end{aligned} \tag{60}$$

Since μ_*^N does not satisfy the uniform LSI, we can not utilize the same technique to upper bound $-\mathsf{FI}(\mu_t^N\|\mu_*^N)$. Chen et al. (2022) and Chen et al. (2023) obtain the lower bound of the relative Fisher information $\mathsf{FI}(\mu_t^N\|\mu_*^N)$ using other technique to circumvent the uniform LSI of μ_*^N. We will directly provide the conclusion instead of providing many details about that technique in this paper, and we refer our readers to Chen et al. (2022; 2023) for the precise proof. Chen et al. (2023) propose that

$$\begin{aligned}\mathsf{FI}(\mu_t^N\|\mu_*^N) &= \frac{1}{2}\mathsf{FI}(\mu_t^N\|\mu_*^N) + \frac{1}{2}\mathsf{FI}(\mu_t^N\|\mu_*^N)\\ &\geq \frac{1}{2}\left[2(1-\varepsilon)\mathscr{C}_{\mathsf{LSI}} - \frac{\mathscr{L}}{N}\left(16 + 12(\varepsilon^{-1}-1)\frac{\mathscr{L}}{\mathscr{C}_{\mathsf{LSI}}}\right)\right](\mathcal{F}^N(\mu_t^N) - N\mathcal{F}(\mu_*))\\ &\quad + \frac{1}{2}\mathsf{FI}(\mu_t^N\|\mu_*^N) - \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3(\varepsilon^{-1}-1)\mathscr{L})\end{aligned}$$

for $\varepsilon \in (0,1)$. If we choose $\varepsilon = 1/2$ and $N \geq \frac{32\mathscr{L}}{\mathscr{C}_{\mathsf{LSI}}} + \frac{24\mathscr{L}^2}{\mathscr{C}_{\mathsf{LSI}}^2}$, we have

$$\begin{aligned}\mathsf{FI}(\mu_t^N\|\mu_*^N) &\geq \frac{\mathscr{C}_{\mathsf{LSI}}}{4}(\mathcal{F}^N(\mu_t^N) - N\mathcal{F}(\mu_*)) + \frac{1}{2\lambda_{\mathsf{max}}(S)}\mathsf{FI}_S(\mu_t^N\|\mu_*^N) - \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3\mathscr{L})\\ &\geq \frac{\mathscr{C}_{\mathsf{LSI}}}{4}(\mathcal{F}^N(\mu_t^N) - N\mathcal{F}(\mu_*)) + \frac{1}{6\mathscr{L}}\mathsf{FI}_S(\mu_t^N\|\mu_*^N) - \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3\mathscr{L})\\ &\geq \frac{\mathscr{C}_{\mathsf{LSI}}}{24\mathscr{L}}(\mathcal{E}^N(\mu_t^N) - N\mathcal{E}(\mu_*)) - \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3\mathscr{L})\end{aligned}$$

Combining (60) with the lower bound of Fisher information above, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{E}^N(\mu_t^N) - N\mathcal{E}(\mu_*)) \leq -\frac{\mathscr{C}_{\mathsf{LSI}}}{24\mathscr{L}}(\mathcal{E}^N(\mu_t^N) - N\mathcal{E}(\mu_*)) + \frac{\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3\mathscr{L})$$

Applying Grönwall's inequality, we obtain

$$\begin{aligned}\mathcal{F}^N(\mu_t^N) - N\mathcal{F}(\mu_*) &\leq \mathcal{E}^N(\mu_t^N) - N\mathcal{E}(\mu_*)\\ &\leq (\mathcal{E}^N(\mu_0^N) - N\mathcal{E}(\mu_*))\exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{24\mathscr{L}}t\right) + \frac{\mathscr{L}\mathrm{d}t}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3\mathscr{L})\exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{24\mathscr{L}}t\right)\\ &\leq (\mathcal{E}^N(\mu_0^N) - N\mathcal{E}(\mu_*))\exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{24\mathscr{L}}t\right) + \frac{120\mathscr{L}^2 d}{\mathscr{C}_{\mathsf{LSI}}} + \frac{72\mathscr{L}^3 d}{\mathscr{C}_{\mathsf{LSI}}^2}\end{aligned} \tag{61}$$

where the last inequality follows from $\exp(-x) \leq (1+x)^{-1}$ for $x > -1$. Now we consider a more general γ in the proof above. Analogous to the proof of Lemma 32 in Villani (2009), if we incorporate γ, the diagonal elements of upper triangular matrix $\mathcal{K}$ will become $(\gamma + 2\gamma a - 4\mathscr{L}b,\, 2\gamma a,\, 2b,\, 2\gamma c)$. If we choose $\gamma = \sqrt{\mathscr{L}}$, $b = 1/\sqrt{\mathscr{L}}$, $a = 2$ and $c = 1/\mathscr{L}$, the smallest eigenvalue of K will become $\lambda_{\mathsf{min}}(\mathcal{K}) = 2/\sqrt{\mathscr{L}}$. Similar to the previous proof, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{E}^N(\mu_t^N) - N\mathcal{E}(\mu_*)) &\leq -\frac{2}{\sqrt{\mathscr{L}}}(\|\nabla_{\mathbf{v}} u_t^N\|^2 + \|\nabla_{\mathbf{v}}^2 u_t^N\|^2 + \|\nabla_{\mathbf{x}} u_t^N\|^2 + \|\nabla_{\mathbf{x}}\nabla_{\mathbf{v}} u_t^N\|^2)\\ &\leq -\frac{2}{\sqrt{\mathscr{L}}}(\|\nabla_{\mathbf{v}} u_t^N\|^2 + \|\nabla_{\mathbf{x}} u_t^N\|^2) = -\frac{2}{\sqrt{\mathscr{L}}}\mathsf{FI}(\mu_t^N\|\mu_*^N)\\ &\leq -\frac{\mathscr{C}_{\mathsf{LSI}}}{2\sqrt{\mathscr{L}}}(\mathcal{F}^N(\mu_t^N) - N\mathcal{F}(\mu_*)) - \frac{1}{\lambda_{\mathsf{max}}(S)\sqrt{\mathscr{L}}}\mathsf{FI}_S(\mu_t^N\|\mu_*^N) + \frac{2\sqrt{\mathscr{L}}d}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3\mathscr{L})\\ &\leq -\frac{\mathscr{C}_{\mathsf{LSI}}}{2\sqrt{\mathscr{L}}}(\mathcal{F}^N(\mu_t^N) - N\mathcal{F}(\mu_*)) - \frac{1}{3\sqrt{\mathscr{L}}}\mathsf{FI}_S(\mu_t^N\|\mu_*^N) + \frac{2\sqrt{\mathscr{L}}d}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3\mathscr{L})\\ &\leq -\frac{\mathscr{C}_{\mathsf{LSI}}}{6\sqrt{\mathscr{L}}}(\mathcal{E}^N(\mu_t^N) - N\mathcal{E}(\mu_*)) + \frac{2\sqrt{\mathscr{L}}d}{\mathscr{C}_{\mathsf{LSI}}}(5\mathscr{C}_{\mathsf{LSI}} + 3\mathscr{L})\end{aligned}$$

Applying Grönwall's inequality, we obtain

$$\mathcal{F}^N(\mu_t^N) - N\mathcal{F}(\mu_*) \leq \mathcal{E}^N(\mu_t^N) - N\mathcal{E}(\mu_*) \leq \mathcal{E}_0^N \exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{6\sqrt{\mathscr{L}}}t\right) + \frac{60\mathscr{L}d}{\mathscr{C}_{\mathsf{LSI}}} + \frac{36\mathscr{L}^2 d}{\mathscr{C}_{\mathsf{LSI}}^2} \tag{62}$$

where $\mathcal{E}_0^N := \mathcal{E}^N(\mu_0^N) - N\mathcal{E}(\mu_*)$. This completes the proof of Theorem 3.2. The convergence rate exhibited in Eq. (62) is faster and incurs a smaller bias than the rate exhibited in Eq. (61).

E. Discretization Analysis

In this section, we provide the proof of Theorem 3.3 and Theorem 3.4 establishing the global convergence of the discrete-time-space processes. Our discretization analysis is unified for the MULA and N-ULA.

E.1. Proof of Theorem 3.3

Suppose $\mathbf{Q}_{Nh}$ is the joint law of the MULD for $t \in [0, Nh]$ and $\mathbf{P}_{Nh}$ is the joint law of the MULA for $t \in [kh, (k+1)h]$ and $k = 0, 1, ..., K-1$. Applying Girsanov's theorem (Lemma B.10), we have

$$
\begin{aligned}
\mathsf{KL}(\mathbf{Q}_{Kh}\|\mathbf{P}_{Kh}) &= \mathbb{E}_{\mathbf{Q}_{Kh}} \log \frac{\mathrm{d}\mathbf{Q}_{Kh}}{\mathrm{d}\mathbf{P}_{Kh}} \\
&= \mathbb{E}_{\mathbf{Q}_{Kh}} \sum_{k=0}^{K-1} \left(-\frac{1}{\sqrt{2\gamma}} \int_{kh}^{(k+1)h} \left\langle \begin{pmatrix} 0 \\ D_\rho F(\mu_t^X, x_t) - D_\rho F(\mu_{kh}^X, x_{kh}) \end{pmatrix}, \mathrm{dB}_t \right\rangle \right. \\
&\quad \left. + \frac{1}{4\gamma} \int_{kh}^{(k+1)h} \left\| D_\rho F(\mu_t^X, x_t) - D_\rho F(\mu_{kh}^X, x_{kh}) \right\|^2 \mathrm{d}t \right) \\
&= \frac{1}{4\gamma} \sum_{k=0}^{K-1} \int_{kh}^{(k+1)h} \mathbb{E}_{\mathbf{Q}_{Kh}} \left\| D_\mu F(\mu_t^X, x_t) - D_\mu F(\mu_{kh}^X, x_{kh}) \right\|^2 \mathrm{d}t
\end{aligned}
$$

And we obtain

$$
\begin{aligned}
\mathsf{KL}(\mathbf{Q}_{Kh}\|\mathbf{P}_{Kh}) &= \frac{1}{4\gamma} \sum_{k=0}^{K-1} \int_{kh}^{(k+1)h} \mathbb{E}_{\mathbf{Q}_{Kh}} \left\| D_\rho F(\mu_t^X, x_t) - D_\rho F(\mu_{kh}^X, x_{kh}) \right\|^2 \mathrm{d}t \\
&\leq \frac{\mathscr{L}^2}{2\gamma} \sum_{k=0}^{K-1} \int_{kh}^{(k+1)h} \mathbb{E}_{\mathbf{Q}_{Kh}} \|x_t - x_{kh}\|^2 + W_1^2(\mu_t^X, \mu_{kh}^X) \mathrm{d}t \\
&\leq \frac{\mathscr{L}^2}{2\gamma} \sum_{k=0}^{K-1} \int_{kh}^{(k+1)h} \mathbb{E}_{\mathbf{Q}_{Kh}} \|x_t - x_{kh}\|^2 + \mathbb{E}_{\mathbf{Q}_{Kh}} \|x_t - x_{kh}\|^2 \mathrm{d}t \\
&= \frac{\mathscr{L}^2}{\gamma} \sum_{k=0}^{K-1} \int_{kh}^{(k+1)h} \mathbb{E}_{\mathbf{Q}_{Kh}} \|x_t - x_{kh}\|^2 \mathrm{d}t
\end{aligned}
$$

where the first inequality follows from Assumption 2.4 and the last inequality follows from Lemma B.7 and the inequality $\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 \leq \frac{1}{n}\sum_{i=1}^n x_i^2$:

$$
\mathbb{E}_{\mathbf{Q}_{Kh}} \|x_t - x_{kh}\|^2 \leq 16\mathscr{L}^2 h^4 \mathbb{E}_{\mathbf{Q}_{Kh}} \|x_{kh}\|^2 + 64h^2 \mathbb{E}_{\mathbf{Q}_{Kh}} \|v_{kh}\|^2 + 16\mathscr{L}^2 h^4 + 32\gamma h^3 d
$$

Combined with Lemma B.8 and $\gamma = \sqrt{\mathscr{L}}$, the discretization error is upper bounded as follows:

$$
\begin{aligned}
\mathsf{KL}(\mathbf{Q}_{Kh}\|\mathbf{p}_{Kh}) &\leq \frac{16\mathscr{L}^4 h^5 K}{\gamma} \max_{0 \leq k \leq K} \mathbb{E}_{\mathbf{Q}_{Kh}} \|x_{kh}\|^2 + \frac{64\mathscr{L}^2 h^3 K}{\gamma} \max_{0 \leq k \leq K} \mathbb{E}_{\mathbf{Q}_{Kh}} \|v_{kh}\|^2 \\
&\quad + \frac{16\mathscr{L}^4 h^5 K}{\gamma} + 32\mathscr{L}^2 h^4 K d \\
&\lesssim \frac{\mathscr{L}^{9/2} h^5 K d}{\mathscr{C}_{\mathsf{LSI}}} + \frac{\mathscr{L}^{5/2} h^3 K d}{\mathscr{C}_{\mathsf{LSI}}} + \mathscr{L}^{7/2} h^5 K + \mathscr{L}^2 h^4 K d \\
&= \frac{\mathscr{L}^{9/2} h^4 T d}{\mathscr{C}_{\mathsf{LSI}}} + \frac{\mathscr{L}^{5/2} h^2 T d}{\mathscr{C}_{\mathsf{LSI}}} + \mathscr{L}^{7/2} h^4 T + \mathscr{L}^2 h^3 T d
\end{aligned}
$$

where $T = Kh$. By Lemma B.3 and Theorem 3.1, we obtain

$$
\mathsf{KL}(\mu_t\|\mu_*) \leq \mathcal{F}(\mu_t) - \mathcal{F}(\mu_*) \leq (\mathcal{E}(\mu_0) - \mathcal{E}(\mu_*)) \exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{3\sqrt{\mathscr{L}}} t\right) \tag{63}
$$

Combining with (63), we upper bound the TV distance between $\bar{\mu}_K$, the probability measure of MULA at Kh and μ_*, the limiting distribution of MULD as follows:

$$
\begin{aligned}
\|\bar{\mu}_K - \mu_*\|_{\mathsf{TV}} &\le \|\bar{\mu}_K - \mu_{Kh}\|_{\mathsf{TV}} + \|\mu_{Kh} - \mu_*\|_{\mathsf{TV}} \\
&= \|\mu_{Kh} - \bar{\mu}_K\|_{\mathsf{TV}} + \|\mu_{Kh} - \mu_*\|_{\mathsf{TV}} \\
&\lesssim \sqrt{\mathsf{KL}(\mu_{Kh}\|\bar{\mu}_K)} + \sqrt{\mathsf{KL}(\mu_{Kh}\|\mu_*)} \\
&\lesssim \sqrt{\mathsf{KL}(\mathbf{Q}_{Kh}\|\mathbf{p}_{Kh})} + \sqrt{\mathsf{KL}(\mu_{Kh}\|\mu_*)} \\
&\lesssim \frac{\mathscr{L}^{9/4}h^2T^{1/2}d^{1/2}}{\mathscr{C}_{\mathsf{LSI}}^{1/2}} + \frac{\mathscr{L}^{5/4}hT^{1/2}d^{1/2}}{\mathscr{C}_{\mathsf{LSI}}^{1/2}} + \mathscr{L}^{7/4}h^2T^{1/2} + \mathscr{L}h^{3/2}T^{1/2}d^{1/2} \\
&\quad + (\mathcal{E}(\mu_0) - \mathcal{E}(\mu_*))^{1/2}\exp\left(-\mathscr{C}_{\mathsf{LSI}}T/6\sqrt{\mathscr{L}}\right)
\end{aligned}
$$

where the first inequality follows from the triangle inequality of TV distance; the second inequality follows from Pinsker's inequality, and the fourth inequality follows from the data processing inequality. In order to ensure $\|\mu_{Kh} - \mu_*\|_{\mathsf{TV}} \le \frac{1}{2}\epsilon$, it suffices to choose $T = Kh = \widetilde{\Theta}\left(\sqrt{\mathscr{L}}/\mathscr{C}_{\mathsf{LSI}}\right)$. In order to ensure $\|\bar{\mu}_K - \mu_{Kh}\|_{\mathsf{TV}} \le \frac{1}{2}\epsilon$, it suffices to choose the stepsize

$$
h = \Theta\left(\frac{\mathscr{C}_{\mathsf{LSI}}^{1/2}\epsilon}{\mathscr{L}^{5/4}T^{1/2}d^{1/2}}\right) = \widetilde{\Theta}\left(\frac{\mathscr{C}_{\mathsf{LSI}}\epsilon}{\mathscr{L}^{3/2}d^{1/2}}\right), \tag{64}
$$

and the mixing time

$$
K = \frac{T}{h} = \widetilde{\Theta}\left(\frac{\mathscr{L}^2 d^{1/2}}{\mathscr{C}_{\mathsf{LSI}}^2\epsilon}\right). \tag{65}
$$

The choice of $T,\ h,\ K$ above ensures $\|\bar{\mu}_K - \mu_*\|_{\mathsf{TV}} \le \epsilon$.

E.2. Proof of Theorem 3.4

Suppose Q^i_{Nh} is the joint law of the N-ULD for the i-th particle and $t \in [0, Kh]$; P^i_{Nh} is the joint law of the N-ULA for the i-th particle. Applying Girsanov's theorem (Lemma B.10), we have

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathsf{KL}(\mathbf{Q}^i_{Kh}\|\mathbf{P}^i_{Kh}) &= \frac{1}{4\gamma}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{Q}^i_{Kh}}\left\|D_\rho F(\mu_{\mathbf{x}_t}, x^i_t) - D_\rho F(\mu_{\mathbf{x}_{kh}}, x^i_{kh})\right\|^2 \mathrm{d}t \\
&\le \frac{\mathscr{L}^2}{2\gamma}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|x^i_t - x^i_{kh}\|^2 + W_1^2(\mu_{\mathbf{x}_t}, \mu_{\mathbf{x}_{kh}})\mathrm{d}t \\
&\le \frac{\mathscr{L}^2}{\gamma}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|x^i_t - x^i_{kh}\|^2\mathrm{d}t \\
&\le \frac{16\mathscr{L}^4h^5}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|x^i_{kh}\|^2 + \frac{64\mathscr{L}^2h^3}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|v_{kh}\|^2 \\
&\quad + \frac{16\mathscr{L}^4h^5K}{\gamma} + 32\mathscr{L}^2h^4Kd
\end{aligned}
$$

where the first inequality follows from Assumption 2.4 and the last inequality follows from Lemma B.7 and the inequality $\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2 \le \frac{1}{n}\sum_{i=1}^{n}x_i^2$:

$$
\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|x^i_t - x^i_{kh}\|^2 \le 16\mathscr{L}^2h^4\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|x^i_{kh}\|^2 + 64h^2\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|v^i_{kh}\|^2 + 16\mathscr{L}^2h^4 + 32\gamma h^3 d
$$

for $t \in [kh, (k+1)h]$ and $k = 0, 1, ..., K-1$. Combining Lemma B.9 and $\gamma = \sqrt{\mathscr{L}}$, the discretization error is upper bounded as follows:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathsf{KL}(\mathbf{Q}^i_{Kh}\|\mathbf{P}^i_{Kh}) &\le \frac{16\mathscr{L}^4h^5K}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\max_{0\le k\le K}\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|x^i_{kh}\|^2 + \frac{64\mathscr{L}^2h^3K}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\max_{0\le k\le K}\mathbb{E}_{\mathbf{Q}^i_{Kh}}\|v_{kh}\|^2 \\
&\quad + \frac{16\mathscr{L}^4h^5K}{\gamma} + 32\mathscr{L}^2h^4Kd \\
&\lesssim \frac{\mathscr{L}^{9/2}h^5Kd}{\mathscr{C}_{\mathsf{LSI}}} + \frac{\mathscr{L}^{5/2}h^3Kd}{\mathscr{C}_{\mathsf{LSI}}} + \mathscr{L}^{7/2}h^5K + \mathscr{L}^2h^4Kd \\
&= \frac{\mathscr{L}^{9/2}h^4Td}{\mathscr{C}_{\mathsf{LSI}}} + \frac{\mathscr{L}^{5/2}h^2Td}{\mathscr{C}_{\mathsf{LSI}}} + \mathscr{L}^{7/2}h^4T + \mathscr{L}^2h^3Td
\end{aligned}$$

where $T = Kh$. By Lemma B.4 and Theorem 3.2, we obtain

$$\frac{1}{N}\mathsf{KL}(\mu^N_T\|\mu_*^{\otimes N}) \le \frac{1}{N}\mathcal{F}^N(\mu^N_T) - \mathcal{F}(\mu_*) \le \frac{\mathcal{E}^N_0}{N}\exp\left(-\frac{\mathscr{C}_{\mathsf{LSI}}}{6\sqrt{\mathscr{L}}}T\right) + \frac{60\mathscr{L}d}{N\mathscr{C}_{\mathsf{LSI}}} + \frac{36\mathscr{L}^2d}{N\mathscr{C}_{\mathsf{LSI}}^2}, \tag{66}$$

where $\mathcal{E}^N_0 := \mathcal{E}^N(\mu^N_0) - N\mathcal{E}(\mu_*)$. Combining with (66), we upper bound the averaged TV distance between $\bar{\mu}^i_K$ and μ_* over N particles as follows:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\|\bar{\mu}^i_K - \mu_*\|_{\mathsf{TV}} &\le \frac{1}{N}\sum_{i=1}^{N}\|\bar{\mu}^i_K - \mu^i_{Kh}\|_{\mathsf{TV}} + \frac{1}{N}\sum_{i=1}^{N}\|\mu^i_{Kh} - \mu_*\|_{\mathsf{TV}} \\
&= \frac{1}{N}\sum_{i=1}^{N}\|\mu^i_{Kh} - \bar{\mu}^i_K\|_{\mathsf{TV}} + \frac{1}{N}\sum_{i=1}^{N}\|\mu^i_{Kh} - \mu_*\|_{\mathsf{TV}} \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathsf{KL}(\mu^i_{Kh}\|\bar{\mu}^i_K)} + \frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathsf{KL}(\mu^i_{Kh}\|\mu_*)}
\end{aligned}$$

where the first inequality follows from the triangle inequality of TV distance; the second inequality follows from Pinsker's inequality. Then we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\|\bar{\mu}^i_K - \mu_*\|_{\mathsf{TV}} &\lesssim \frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathsf{KL}(\mu^i_{Kh}\|\bar{\mu}^i_K)} + \frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathsf{KL}(\mu^i_{Kh}\|\mu_*)} \\
&\lesssim \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathsf{KL}(\mu^i_{Kh}\|\bar{\mu}^i_K)} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathsf{KL}(\mu^i_{Kh}\|\mu_*)} \\
&\le \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathsf{KL}(\mathbf{Q}^i_{Kh}\|\mathbf{P}^i_{Kh})} + \sqrt{\frac{1}{N}\mathsf{KL}(\mu^N_{Kh}\|\mu_*^{\otimes N})} \\
&\le \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathsf{KL}(\mathbf{Q}^i_{Kh}\|\mathbf{P}^i_{Kh})} + \sqrt{\frac{1}{N}\mathcal{F}^N(\mu^N_{Kh}) - \mathcal{F}(\mu_*)} \\
&\lesssim \frac{\mathscr{L}^{9/4}h^2T^{1/2}d^{1/2}}{\mathscr{C}_{\mathsf{LSI}}^{1/2}} + \frac{\mathscr{L}^{5/4}hT^{1/2}d^{1/2}}{\mathscr{C}_{\mathsf{LSI}}^{1/2}} + \mathscr{L}^{7/4}h^2T^{1/2} + \mathscr{L}h^{3/2}T^{1/2}d^{1/2} \\
&\quad + \left(\frac{1}{N}\mathcal{E}^N(\mu^N_0) - \mathcal{E}(\mu_*)\right)^{1/2}\exp\left(-\mathscr{C}_{\mathsf{LSI}}T/12\sqrt{\mathscr{L}}\right) + \frac{\mathscr{L}^{1/2}d^{1/2}}{N^{1/2}\mathscr{C}_{\mathsf{LSI}}^{1/2}} + \frac{\mathscr{L}d^{1/2}}{N^{1/2}\mathscr{C}_{\mathsf{LSI}}}
\end{aligned}$$

where the second inequality follows from Jensen's inequality; the third inequality follows from data processing inequality and the information inequality (Lemma B.5) and the fourth inequality follows from Lemma B.4. In order to ensure

$\frac{1}{N}\sum_{i=1}^N \|\mu^i_{Kh} - \mu_*\|_{\mathsf{TV}} \le \frac{1}{2}\epsilon$, it suffices to choose $T = Kh = \widetilde{\Theta}\left(\sqrt{\mathscr{L}}/\mathscr{C}_{\mathsf{LSI}}\right)$. In order to ensure $\frac{1}{N}\sum_{i=1}^N \|\bar{\mu}^i_K - \mu^i_{Kh}\|_{\mathsf{TV}} \le \frac{1}{2}\epsilon$, it suffices to choose the stepsize

$$h = \Theta\left(\frac{\mathscr{C}_{\mathsf{LSI}}^{1/2}\epsilon}{\mathscr{L}^{5/4}T^{1/2}d^{1/2}}\right) = \widetilde{\Theta}\left(\frac{\mathscr{C}_{\mathsf{LSI}}\epsilon}{\mathscr{L}^{3/2}d^{1/2}}\right), \tag{67}$$

the mixing time

$$K = \frac{T}{h} = \widetilde{\Theta}\left(\frac{\mathscr{L}^2 d^{1/2}}{\mathscr{C}_{\mathsf{LSI}}^2\epsilon}\right), \tag{68}$$

and the number of particles

$$N = \Theta\left(\frac{\mathscr{L}^2 d}{\mathscr{C}_{\mathsf{LSI}}^2\epsilon^2}\right). \tag{69}$$

The choice of $T,\ h,\ K,\ N$ above ensures $\frac{1}{N}\sum_{i=1}^N \|\bar{\mu}^i_K - \mu_*\|_{\mathsf{TV}} \le \epsilon$.

F. Numerical Experiments

We verify our theoretical findings by providing empirical support in this section. Our experiment is to approximate a Gaussian function $f(z) = \exp(-\|z-m\|^2/2d)$ for $z \in \mathbb{R}^d$ and unknown $m \in \mathbb{R}^d$ by a mean-field two-layer neural network with tanh activation. Consider the empirical risk minimization problem (17) with quadratic loss function l, $d = 10^3$, $\lambda' = 10^{-4}$ and n randomly generated data samples from $f(z)$ ($n = 100$), described by

$$F(\rho) = \frac{1}{2n}\sum_{i=1}^n (h(\rho; a_i) - f(a_i))^2 + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}[\|x\|^2].$$

F satisfy Assumptions 2.3-2.9 with the choice of $l,\ h$, and thus we apply Algorithm 1 for minimizing the objective above. Note that the number of neurons in the first hidden layer is equivalent to the number of particles in N-ULA, and we choose $N \in \{256, 512, 1024, 2048\}$. The intrinsic derivative of $F(\mu)$ for the j-th particle in our method is given by

$$D_\rho F(\mu_{\mathbf{x}}, x^j) = \frac{1}{n}\sum_{i=1}^n (\frac{1}{N}\sum_{s=1}^N h(x^s; a_i) - f(a_i))\nabla h(x^j; a_i) + \lambda' x^j.$$

Note that $\frac{1}{N}\sum_{s=1}^N h(x^s; a)$ is in fact a two-layer neural network with N neurons. Instead of fine-tuning γ and stepsize h in N-ULA, we directly fine-tune the value of φ_0, φ_1 and φ_2 in Algorithm 1 by grid search. For simplifying the computation, we approximate $(\mathrm{B}^i_k)^x$ and $(\mathrm{B}^i_k)^v$ by $\eta\xi^x_k$ and $\eta\xi^v_k$ where ξ^x_k and ξ^v_k are independent standard Gaussian, and then we fine-tune the scaling scalar η. We compare our method (N-ULA) to N-LA with stepsize h_1 and scaling scalar λ_1 given by,

$$x^j_{k+1} = x^j_k - h_1 D_\rho F(\mu_{\mathbf{x}_k}, x^j_k) + \sqrt{2\lambda_1 h_1}\xi^i_k \tag{N-LA}$$

for $i = 1, ..., N$, $k = 1, ..., K$ and $\xi^i_k \sim \mathcal{N}(0, I_d)$, and EM-N-ULA (the EM discretization of the N-ULD with stepsize h_2 and scaling scalar λ_2) whose update is given by

$$\begin{aligned} x^j_{k+1} &= x^j_k + h_2 v^j_k \\ v^j_{k+1} &= (1-\gamma h_2)v^j_k - h_2 D_\rho F(\mu_{\mathbf{x}_k}, x^j_k) + \sqrt{2\lambda_2 h_2}\xi^i_k \end{aligned} \tag{EM-N-ULA}$$

for $i = 1, ..., N$, $k = 1, ..., K$ and $\xi^i_k \sim \mathcal{N}(0, I_d)$ in the same task. We choose $K = 10^4$ and also fine-tune h_1, λ_1 and h_2, λ_2 to make fair comparison. We postpone our choice of hyperparameters to the Appendix F.1. For each algorithm in our experiment, we initialize $x^j_0 \sim \mathcal{N}(0, 10^{-2}I_d)$ and $v^j_0 \sim \mathcal{N}(0, 10^{-2}I_d)$ for $j = 1, ..., N$, average 5 runs over random seeds in $\{0, 1, 2, 3, 4\}$ and generate the error bars by filling between the largest and the smallest value per iteration. Fig. 2 illustrates the effectiveness of N-ULA. For each N, N-ULA enjoys faster convergence than N-LA and EM-N-ULA. Notably, there is an interesting phenomenon in our experiments. For $N = 256$, both N-ULA and EM-N-ULA suffer from convergence instability, which means that the loss will escape the stable convergence regime and slightly go up after many training epochs. However, N-ULA outperforms N-LA and EM-N-ULA without convergence instability for $N = 512,\ 1024,\ 2048$, and the loss of N-ULA even goes on decreasing when the losses of N-LA and EM-N-ULA keep stable for $N = 1024,\ 2048$. This phenomenon matches our theory that we do not reduce the number of particles for N-ULA when compared with N-LA (see Table 1). These observations suggest that our method performs better in the high particle-approximation regime. Fig. 3 demonstrates this finding more transparently. The second row of Fig. 2 also suggests that EM discretization incurs a larger bias than LPM.

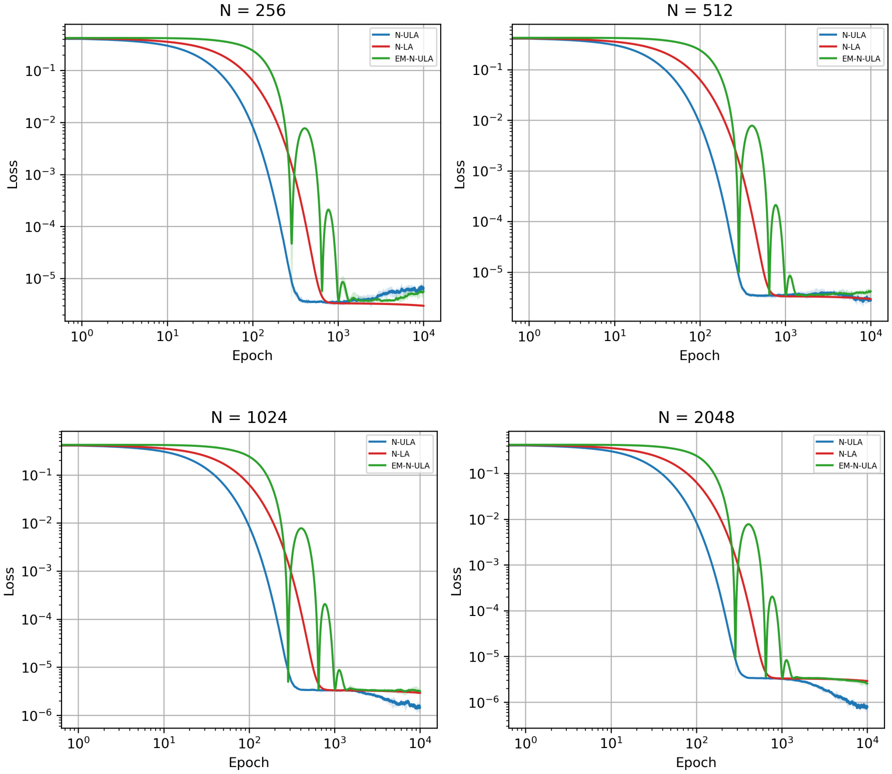

Figure 2. Evaluation on N-ULA, N-LA and EM-N-ULA with different number of particles N where x-axis represents the training epochs and y-axis represents the value of $\frac{1}{2n}\sum_{i=1}^{n}(\frac{1}{N}\sum_{s=1}^{N}h(x^s;a_i)-f(a_i))^2$. Our method often enjoys better performance in the high particle-approximation regime which is consistent with our theoretical findings.

F.1. Experimental Settings

We give the actual updates of the methods involved in our experiment and provide the precise value of parameters in Table 2. The update of the N-ULA is given by

$$\begin{aligned} x_{k+1}^j &= x_k^j + \varphi_0\, v_k^j - \varphi_1\, D_\rho F(\mu_{\mathbf{x}_k}, x_k^j) + \eta \xi_k^x, \\ v_{k+1}^j &= \varphi_2\, v_k^j - \varphi_3\, D_\rho F(\mu_{\mathbf{x}_k}, x_k^j) + \eta \xi_k^v. \end{aligned}$$

for $j = 1, ..., N$. The update of EM-N-ULA is given by

$$\begin{aligned} x_{k+1}^j &= x_k^j + h_2\, v_k^j, \\ v_{k+1}^j &= (1-h_3)v_k^j - h_2\, D_\rho F(\mu_{\mathbf{x}_k}, x_k^j) + \sqrt{2\lambda_2 h_2}\xi_k. \end{aligned}$$

for $j = 1, ..., N$. The update of the N-LA is given by

$$x_{k+1}^j = x_k^j - h_1\, D_\rho F(\mu_{\mathbf{x}_k}, x_k^j) + \sqrt{2\lambda_1 h_1}\xi_k.$$

for $j = 1, ..., N$.

Parameters	φ_0	φ_1	φ_2	φ_3	η	h_1	h_2	h_3	λ_1	λ_2
Value	10^{-4}	0.02	0.99	0.02	10^{-3}	10^{-2}	10^{-2}	10^{-2}	10^{-4}	10^{-4}

Table 2. Choice of hyperparameters.

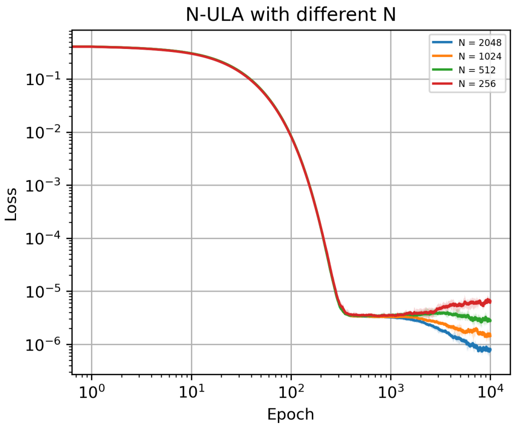

Figure 3. N-ULA with different number of particles

G. Methods for Comparisons

In this section, we review the convergence result of MLA in Nitanda et al. (2022) and N-LA in Suzuki et al. (2023), which consider problem (1) in more specific settings. Nitanda et al. (2022) suppose $F(\rho) = \mathbb{E}_{(a,b)\sim\mathcal{D}}\left[\ell(h(\rho;a),b)\right] + \frac{\lambda'}{2}\mathbb{E}_{x\sim\rho}\|x\|^2$ whereas Suzuki et al. (2023) suppose $F(\rho) = U(\rho) + \lambda'\mathbb{E}_{x\sim\rho}[r(x)]$. While our convergence results are established in TV distance, we consider more general settings compared with the previous two. Since the problem setting in Nitanda et al. (2022) is only for training neural networks, we perform convergence analysis of the MLA in Suzuki et al. (2023)'s setting to make a comparison with our results. Define the free energy

$$E(\rho) = F(\rho) + \text{Ent}(\rho), \tag{70}$$

where $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Let $\bar{\rho}_k$ denotes the law of k-th iterate of the MLA and ρ_* denotes the minimizer of (70), and Nitanda et al. (2022) obtain the following results in Theorem 2:

$$E(\bar{\rho}_k) - E(\rho_*) \le \exp(-\mathscr{C}_{\mathsf{LSI}}hk)(E(\bar{\rho}_0) - E(\rho_*)) + \frac{\delta_h}{2\mathscr{C}_{\mathsf{LSI}}}, \tag{71}$$

where $\delta_{hk} := \mathbb{E}\|D_\rho F(\bar{\rho}_{k+1}, x_{k+1}) - D_\rho F(\bar{\rho}_k, x_k)\|^2$ and $\mathbb{E}$ is taken under the joint law of $\bar{\rho}_{k+1}$ and $\bar{\rho}_k$. Now we bound δ_{hk} uniformly in k with a different method from the one in Nitanda et al. (2022). We do not need to specify F to be the objective of training nerual networks. Since F is $\mathscr{L}$-smooth[2] and satisfies Assumption 2.6, we obtain

$$\begin{aligned}
\mathbb{E}\|D_\rho F(\bar{\rho}_{k+1}, x_{k+1}) - D_\rho F(\bar{\rho}_k, x_k)\|^2 &\le 2\mathscr{L}^2\mathbb{E}(\|x_{k+1} - x_k\|^2 + W_2^2(\bar{\rho}_{k+1}, \bar{\rho}_k)) \\
&\le 4\mathscr{L}^2\mathbb{E}\|x_{k+1} - x_k\|^2 \\
&= 4\mathscr{L}^2\mathbb{E}\| - hD_\rho F(\bar{\rho}_k, x_k) + \sqrt{2h}\xi\|^2 \\
&\le 4\mathscr{L}^2h^2\mathbb{E}\|D_\rho F(\bar{\rho}_k, x_k)\|^2 + 8\mathscr{L}^2hd \\
&\le 8\mathscr{L}^4h^2(1 + \mathbb{E}\|x_k\|^2) + 8\mathscr{L}^2hd
\end{aligned}$$

We refer to Lemma 1 in Suzuki et al. (2023) to uniformly bound $\mathbb{E}\|x_k\|^2$. Before applying Lemma 1, we translate some constants in Suzuki et al. (2023) into our constants systems. Suzuki et al. (2023) assumes that $\|D_\rho U(\rho, x)\| \le R$, $\lambda_1 I_d \preceq \nabla^2 r(x) \preceq \lambda_2 I_d$. We let $R = \mathscr{L}$ and $\lambda_2 = \mathscr{L}$ (since this specification matches our Assumption 2.6). We prove Lemma 1 proposed by Suzuki et al. (2023) in the mean-field setting without particle approximation. But we also assume the decomposition $F(\rho) = U(\rho) + \mathbb{E}_{x\sim\rho}[r(x)]$ with $\|D_\rho U(\rho, x)\| \le \mathscr{L}$ and $\lambda_1 I_d \preceq \nabla^2 r \preceq \mathscr{L} I_d$. Given the update of

[2] We inherit the weaker smoothness assumption in Suzuki et al. (2023) with respect to W_2 distance.

the MLA, if $h \leq \frac{\lambda_1}{2\mathscr{L}^2}$, we have

$$\begin{aligned}\mathbb{E}\|x_{k+1}\|^2 &= \mathbb{E}\|x_k\|^2 + h^2\mathbb{E}\|D_\rho F(\rho_k, x_k))\|^2 + 2hd - 2h\mathbb{E}\left\langle x_k, D_\rho U(\rho_k, x_k) + \nabla r(x_k)\right\rangle \\ &\leq \mathbb{E}\|x_k\|^2 + \mathscr{L}^2h^2(1+\mathbb{E}\|x_k\|^2) + 2hd + 2h\mathscr{L}\mathbb{E}\|x_k\| - 2h\lambda_1\mathbb{E}\|x_k\|^2 \\ &\leq (1-\lambda_1 h)\mathbb{E}\|x_k\|^2 + \mathscr{L}^2h^2 + 2hd + \frac{2\mathscr{L}^2 h}{\lambda_1}\end{aligned}$$

Recursively, we obtain

$$\mathbb{E}\|x_k\|^2 \leq (1-\lambda_1 h)^k\mathbb{E}\|x_0\|^2 + \frac{\mathscr{L}^2 h + 2d}{\lambda_1} + \frac{2\mathscr{L}^2}{\lambda_1^2} \leq \mathbb{E}\|x_0\|^2 + \frac{\mathscr{L}^2 h + 2d}{\lambda_1} + \frac{2\mathscr{L}^2}{\lambda_1^2}. \tag{72}$$

If $x_0 \sim \mathcal{N}(0, I_d)$, $\mathbb{E}\|x_0\|^2 \lesssim d$. Thus (72) implies $\mathbb{E}\|x_k\|^2 \lesssim \mathscr{L}^2 d$. Plugging into the inequality above, we obtain

$$\mathbb{E}\|D_\rho F(\bar\rho_{k+1}, x_{k+1}) - D_\rho F(\bar\rho_k, x_k)\|^2 \lesssim \mathscr{L}^6 h^2 d + \mathscr{L}^2 hd. \tag{73}$$

Applying Lemma B.3 and pinsker's inequality, we obtain

$$\begin{aligned}\|\bar\rho_K - \rho_*\|_{\mathsf{TV}} \lesssim \sqrt{\mathsf{KL}(\bar\rho_K\|\rho_*)} &\leq \sqrt{E(\bar\rho_K) - E(\rho_*)} \\ &\lesssim \exp(-\mathscr{C}_{\mathsf{LSI}} hK/2)(E(\bar\rho_0) - E(\rho_*))^{1/2} + \frac{\mathscr{L}^3 hd^{1/2}}{\mathscr{C}_{\mathsf{LSI}}^{1/2}} + \frac{\mathscr{L}h^{1/2}d^{1/2}}{\mathscr{C}_{\mathsf{LSI}}^{1/2}}\end{aligned}$$

In order to ensure $\|\bar\rho_K - \rho_*\|_{\mathsf{TV}} \leq \epsilon$, it suffices to choose

$$h = \Theta\left(\frac{\mathscr{C}_{\mathsf{LSI}}\epsilon^2}{\mathscr{L}^3 d}\right), \quad K = \widetilde{\Theta}\left(\frac{\mathscr{L}^3 d}{\mathscr{C}_{\mathsf{LSI}}^2\epsilon^2}\right). \tag{74}$$

Now we translate the convergence results in Suzuki et al. (2023). Define the free energy of the particle system:

$$E^N(\mu^N) = N\mathbb{E}_{\mathbf{x}\sim\mu^N}F(\mu_{\mathbf{x}}) + \mathrm{Ent}(\mu^N), \tag{75}$$

where $\mu_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N \delta_{x^i}$. Similar to the analysis above, Theorem 2 in Suzuki et al. (2023) implies the TV-convergence of the N-LA, given by

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \|\bar\rho_K^i - \rho_*\|_{\mathsf{TV}} \lesssim \sqrt{\frac{1}{N}\sum_{i=1}^N \mathsf{KL}(\bar\rho_K^i\|\rho_*)} &\leq \sqrt{\frac{1}{N}E^N(\rho_K^N) - E(\rho_*)} \\ &\lesssim \exp\left(-\mathscr{C}_{\mathsf{LSI}} hK/4\right) + h^{1/2}K^{1/2}(\mathscr{L}^3 hd^{1/2} + \mathscr{L}h^{1/2}d^{1/2}) \\ &\quad + h^{1/2}K^{1/2}\frac{\mathscr{L}^2 d^{1/2}}{N^{1/2}}\end{aligned}$$

In order to ensure $\|\bar\rho_K - \rho_*\|_{\mathsf{TV}} \leq \epsilon$, it suffices to choose

$$h = \Theta\left(\frac{\mathscr{C}_{\mathsf{LSI}}\epsilon^2}{\mathscr{L}^3 d}\right), \quad K = \widetilde{\Theta}\left(\frac{\mathscr{L}^3 d}{\mathscr{C}_{\mathsf{LSI}}^2\epsilon^2}\right), \quad N = \Theta\left(\frac{\mathscr{L}^4 d}{\mathscr{C}_{\mathsf{LSI}}\epsilon^2}\right). \tag{76}$$