Proceedings of Machine Learning Research vol 291:1–31, 2025 38th Annual Conference on Learning Theory

Beyond propagation of chaos: A stochastic algorithm for mean field optimization

Chandan Tankala CHANDANT@UOREGON.EDU
University of Oregon
Dheeraj M. Nagaraj DHEERAJNAGARAJ@GOOGLE.COM
Google DeepMind
Anant Raj ANANTRAJ@IISC.AC.IN
Indian Institute of Science

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Gradient flow in the 2-Wasserstein space is widely used to optimize functionals over probability distributions and is typically implemented using an interacting particle system with n particles. Analyzing these algorithms requires showing (a) that the finite-particle system converges and/or (b) that the resultant empirical distribution of the particles closely approximates the optimal distribution (i.e., propagation of chaos). However, establishing efficient sufficient conditions can be challenging, as the finite particle system may produce heavily dependent random variables.

In this work, we study the virtual particle stochastic approximation, originally introduced for Stein Variational Gradient Descent Das and Nagaraj (2023). This method can be viewed as a form of stochastic gradient descent in the Wasserstein space and can be implemented efficiently. In popular settings, we demonstrate that our algorithm's output converges to the optimal distribution under conditions similar to those for the infinite particle limit, and it produces i.i.d. samples without the need to explicitly establish propagation of chaos bounds.

Keywords: List of keywords

1. Introduction

Optimizing a functional $\mathcal{E}()$ over the space of all probability distributions over $\mathbb{R}^d$ with finite second moments ($\mathcal{P}_2(\mathbb{R}^d)$) has gained immense interest in the recent years with applications in machine learning and Bayesian inference. A notable example is the training and analysis of neural networks in the infinite-width regime. While analyzing neural network training is challenging due to inherent non-linearity, the infinite-width limit—known as the mean-field limit—facilitates a more tractable analysis. In this regime, the optimization problem reduces to optimizing over the distribution of neuron weights to achieve accurate label prediction Chizat and Bach (2018); Mei et al. (2018); Nitanda et al. (2022); Suzuki et al. (2023); Nitanda (2024). As other important examples, the task of sampling can be re-cast as the optimization of the Kullback-Leibler (KL) divergence to the target distribution Vempala and Wibisono (2019); Durmus et al. (2019), and variational inference involves constrained optimization over the space of distributions to fit the given data Yao and Yang (2022); Lacker (2023); Liu and Wang (2016); Lambert et al. (2022); Yan et al. (2024). A prototypical example of such a functional is given by $\mathcal{E}(\mu) = \int V\, d\mu + \int \log \mu\, d\mu$. More broadly, this work examines the following optimization objective. Given an energy functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, and regularization strength $\sigma > 0$, we consider functionals of the form $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defined as:

$$\mathcal{E}(\mu) = \mathcal{F}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu)\,, \tag{1}$$

where $\mathcal{H}(\mu)$ is the negative entropy defined as follows:

$$\mathcal{H}(\mu) = \begin{cases} \int \mu(x) \log \mu(x) dx & \text{if } \mu \ll \mathsf{Leb} \text{ and } d\mu(x) = \mu(x)dx \\ \infty & \text{otherwise}. \end{cases}$$

A common approach to optimization over $\mathcal{P}_2(\mathbb{R}^d)$ is gradient flow with respect to the Wasserstein metric. The well-known Langevin dynamics was shown to be the gradient flow of the Kullback-Leibler (KL) divergence to the target distribution in the seminal work of Jordan et al. (1998). This framework can be extended to a broader class of functionals, including interaction energy and entropy McCann (1997); Ambrosio et al. (2008). While Langevin dynamics can be implemented algorithmically through time discretization of Itô stochastic differential equations (SDE), the more general case—leading to McKean-Vlasov type SDEs Carmona and Delarue (2015), where the drift function depends on the distribution of the variable—is not straightforward. The popular computational approximation in this context is the particle approximation where n instances (or particles) of the SDE are implemented computationally, with the distribution of the variables replaced by their empirical distribution. Theoretically, one way to show that this approximation optimizes the objective functional is to show that (a) the particle approximation converges rapidly to its nd dimensional stationary distribution and (b) a sample from the n particle stationary distribution gives a representative sample from the optimal distribution (called propagation of chaos). Propagation of chaos type of results show that whenever we pick any k out of the n particles, then the resulting k particle distribution is close to the k fold product distribution of the target with an error of $O(\mathsf{poly}(k/n))$ in metrics such as KL divergence.

To show the convergence of n particle approximation, prior works such as Chen et al. (2022); Chewi et al. (2024); Wang (2024) consider the finite-particle stationary distribution and establish Logarithmic Sobolev inequalities (LSI) for the nd dimensional system with a constant independent of n. This can be used to establish computational complexity of standard sampling algorithms such as LMC, ULMC, and MALA, allowing us to obtain guarantees for a large class of sampling algorithms instead of specialized analysis for each. However, this can be technically involved and yield pessimistic estimates for the LSI which are worse than those established in the mean-field case (compare to (Suzuki et al., 2024, Theorem 1)). Additionally, the complexity of sampling can be polynomial in the dimension nd, requiring more iterations with more particles (see Vempala and Wibisono (2019)). Obtaining guarantees for the LSI constant $C_{\mathsf{LSI},n}$ for the n-particle stationary distribution can be very hard and $C_{\mathsf{LSI},n}$ could be much worse than the LSI constant C_{LSI} for the mean field stationary distribution π. For instance, in Wang (2024)[Theorem 1], $C_{\mathsf{LSI},n} = O(dC_{\mathsf{LSI}}^3)$ (along with additional assumptions), and in Chewi et al. (2024)[Theorem 2], $C_{\mathsf{LSI},n} = O(e^d)$, while Chewi et al. (2024)[Equation 2.2], C_{LSI} does not depend on d.

Our Contribution We consider stochastic approximations of the mean field SDEs, which can be implemented exactly, when an unbiased estimator for the Wasserstein gradients is available (Algorithm 1). The virtual particle stochastic approximation method, first proposed by Das and Nagaraj (2023) for Stein Variational Gradient Descent, is the theoretical foundation of our analysis. We expand its applicability to sampling from the stationary distributions of McKean-Vlasov SDEs. This extension presents novel analytical challenges, particularly in quantifying the approximation of the non-linear stationary probability measure induced by the McKean-Vlasov SDE and the Brownian diffusion process, which presents almost everywhere non differentiable paths. The algorithm , which outputs n particles for T iterations/time steps, has the following properties:

1. **Computational complexity** $O(nT + T^2)$, unlike the $O(n^2T)$ complexity of standard particle methods. The **output particles are i.i.d.** from a distribution close to the minimizer of $\mathcal{E}$, and does not require us to establish propagation of chaos separately.

2. Our general result in Theorem 7 shows that the number of time steps T required to sample from an ϵ optimal distribution is polynomial in the problem parameters, depends on the **isoperimetry constant of only the mean field dynamics** and is independent of n. As noted above, the isoperimetry bounds for the n particle system, $C_{\mathsf{LSI},n}$, can be much worse than that of the mean field optimal distribution C_{LSI}

3. We illustrate our result in two important scenarios: pairwise interaction energy in the weak interaction regime (Equation (3)) and the mean field neural network with square loss (Equation (6)), we establish these results under standard assumptions.

We note that prior works can deal with a much larger class of functionals Chen et al. (2022); Wang (2024); Nitanda et al. (2022), without access to unbiased estimators for the gradients.

1.1. Prior Work

Stochastic Approximation and Sampling Algorithms: Stochatic Gradient Langevin Dynamics Welling and Teh (2011) was introduced as a stochastic, computationally viable variant of Langevin Monte Carlo (LMC) and has been extensively studied in the literature Raginsky et al. (2017); Kinoshita and Suzuki (2022); Das et al. (2023). This has been extended to other sampling algorithms and interacting particle systems Huang et al. (2024); Jin et al. (2020). Recently, virtual particle stochastic approximation Das and Nagaraj (2023) was introduced in the context of Stein Variational Gradient Descent (SVGD) Liu and Wang (2016), where the algorithm directly produced an unbiased estimate of the flow in the space of probability distributions, giving the first provably fast finite particle variant of SVGD. Note that such bounds have been obtained for the traditional SVGD algorithm since then Balasubramanian et al. (2024). Stochastic approximations have also been utilized to obtain speedup of sampling algorithm with randomized mid-point based time discretization Kandasamy and Nagaraj (2024); Yu et al. (2023); Shen and Lee (2019).

Propagation of Chaos: The propagation of chaos problem for McKean-Vlasov SDEs was originally studied by Sznitman (1991), which established convergence rates in the Wasserstein metric via coupling arguments. These bounds were first made uniform in time by Malrieu (2001, 2003) in the quadratic Wasserstein and relative entropy metrics. In the case of pairwise interaction energy (Equation (3)), these works obtain a bound on the error of order $O(k/n)$ in the squared quadratic Wasserstein distance and assume strong convexity for the *external potential* and convexity for the *interaction potential*. A uniform in time propagation of chaos was recently shown by Chen et al. (2022) by assuming convexity of the mean-field functional, as opposed to imposing convexity conditions on the interaction potential. The error in the squared quadratic Wasserstein error bound was improved to $O((k/n)^2)$ in Lacker and Le Flem (2023) by assuming a uniform-in-n log-Sobolev inequality for the stationary distribution of the n-particle system and using the recursive BBGKY proof technique. Kook et al. (2024) build on this proof technique and also obtain an error bound of order $O((k/n)^2)$ in the squared quadratic Wasserstein idstance and KL-divergence under a slightly different assumption on a ratio involving log-Sobolev inequality, smoothness and diffusion constants, which is referred to as the "weak interaction" condition. Recently, Bou-Rabee and Schuh (2023) presented a non-linear

Hamiltonian Monte Carlo algorithm and prove its rate of convergence in L^1-Wasserstein distance, without using the propagation of chaos arguments.

Mean Field Optimization: Mean-field analysis of neural networks emerged as a theoretical framework for understanding the optimization dynamics of wide neural networks. Early foundational work by Nitanda and Suzuki (2017); Chizat and Bach (2018); Mei et al. (2018) established that gradient flow on infinite width two-layer neural networks converges to the global minimum under appropriate conditions, demonstrating that we can successfully study neural networks in the infinite-dimensional space of parameter distributions by exploiting convexity. The connection with the mean-field Langevin dynamics arises with the addition of Gaussian noise to the gradient, corresponding to the entropy-regularized term in the objective function. Nitanda et al. (2022); Chizat (2022) were among the first to establish exponential convergence rates under certain LSIs, which are verifiable in regularized risk minimization problems using two-layer neural networks. Subsequently, Suzuki et al. (2023, 2024) study uniform in time propagation of chaos result in the context of mean-field neural networks where the main ingredient is the proximal Gibbs distribution, which also satisifies a LSI for convex losses with smooth and bounded activation functions.

Wasserstein Gradient Flows: These describe the evolution of probability measures over $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the Wasserstein-2 metric. Given an energy functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, the gradient flow $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ is formally the solution to the evolution equation

$$\frac{d}{dt}\mu_t = -\text{grad}_W \mathcal{F}(\mu_t),$$

where $\text{grad}_W \mathcal{F}(\mu)$ is the Wasserstein gradient of $\mathcal{F}$. More rigorously, the gradient flow $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ is a curve in $\mathcal{P}_2(\mathbb{R}^d)$ which satisfies the following *continuity equation* in the sense of distributions:

$$\frac{\partial \mu_t(x)}{\partial t} + \nabla_x \cdot (\mu_t(x) v_t(x)) = 0\,,$$

where the 'velocity field' $v_t(x) : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d = -\nabla_{\mathcal{W}}\mathcal{F}(\mu_t)(x)$ and $\nabla_{\mathcal{W}}\mathcal{F}(x, \mu_t)$ is often also called the Wasserstein gradient. Under certain regularity conditions, which hold in all the cases considered in this work, $\nabla_{\mathcal{W}}\mathcal{F}(\mu_t)(x) = \nabla_x \frac{\delta \mathcal{F}}{\delta \mu}(x, \mu_t)$. Here $\frac{\delta \mathcal{F}}{\delta \mu}$ denotes the first variation (Eulerian derivative) of $\mathcal{F}$. Important cases include the heat equation (when $\mathcal{F}$ is the entropy functional) and the Fokker-Planck equation (when $\mathcal{F}$ includes an external potential term). These flows provide a geometric perspective on evolution equations in probability spaces with applications in statistical physics and PDEs. For a rigorous treatment, see Ambrosio et al. (2008); Villani (2021).

1.2. Notation

For any measure, ρ over $\mathbb{R}^d$ and functions $f, g : \mathbb{R}^d \to \mathbb{R}^d$. Let $\langle f, g \rangle_{L_2(\rho)} := \int \rho(dx) \langle f(x), g(x) \rangle$ and $L_2^2(\rho; f)^2 := \int \rho(dx) \|f(x)\|^2$ whenever f, g are square integrable with respect to ρ. For a vector field $f : \mathbb{R}^d \to \mathbb{R}^d$, its divergence is given by $\nabla \cdot f = \sum_{i=1}^d \frac{\partial f}{\partial x_i}$, and for a function $f : \mathbb{R} \to \mathbb{R}$, the Laplacian is defined as $\Delta f := \sum_{i=1}^d \frac{\partial^2 f}{\partial x_i^2}$. Let $\mathcal{P}_2(\mathbb{R}^d), \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ with finite second moment, and those that are absolutely continuous with respect to the Lebesgue measure. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we let $\mathsf{Var}(\mu)$ denote the trace of its covariance. The Wasserstein distance $\mathcal{W}_2(\mu, \nu)$ between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as:

$$\mathcal{W}_2^2(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|X - Y\|^2 d\gamma(x, y)\,,$$

where $\Gamma(\mu,\nu)$ is the set of all joint distributions over $\mathbb{R}^d \times \mathbb{R}^d$ such that the marginal distribution of X is μ and of Y is ν. The Fisher Divergence of a probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with respect to $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as: $\mathsf{FD}(\mu||\nu) := \int_{\mathbb{R}^d} \mu(x)\left\|\nabla \log \frac{\mu(x)}{\nu(x)}\right\|^2 dx$. The first variation of a functional $\mathcal{F}$ at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is denoted by $\delta_\mu \mathcal{F}(\mu)(x)$ or just $\delta\mathcal{F}(x,\mu)$, where $x \in \mathbb{R}^d$, and is defined as the quantity which satisfies the equality:

$$\left.\frac{d\mathcal{F}(\mu+\varepsilon(\mu'-\mu))}{d\varepsilon}\right|_{\varepsilon=0} = \int \delta_\mu \mathcal{F}(\mu)(x)(\mu'-\mu)(x)dx\,.$$

If the probability measures $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d)$ have densities p, q respectively, then the total-variation distance between them is defined as: $\|\mu-\nu\|_{\mathsf{TV}} := \frac{1}{2}\int_{\mathbb{R}^d} |p(x)-q(x)|dx$.

2. Preliminaries and Problem Setup

In this work, we consider the energy functional given in equation (1) as $\mathcal{E}(\mu) = \mathcal{F}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu)$, where $\mathcal{H}(\mu)$ is the negative entropy. Under mild conditions on $\mathcal{F}$ Ambrosio et al. (2008), given $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, the gradient flow of this functional with respect to the Wasserstein metric is the curve $\mu : [0,T] \to \mathcal{P}_2(\mathbb{R}^d)$ given by the following non-linear Fokker-Planck Equation:

$$\frac{\partial \mu_t}{\partial t} - \nabla \cdot (\nabla_{\mathcal{W}} \mathcal{F}(x,\mu_t)\mu_t) = \frac{\sigma^2}{2}\Delta\mu_t\,. \tag{2}$$

Here $\nabla_{\mathcal{W}}\mathcal{F}(\cdot,\cdot) : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^d$ is the Wasserstein gradient. In our applications, we can show that $\nabla_{\mathcal{W}}\mathcal{F} = \nabla\delta\mathcal{F}$, where $\delta\mathcal{F}$ is the first variation of $\mathcal{F}$. This is considered non-linear since the drift depends on μ_t. The SDE corresponding to this is called the McKean-Vlasov equation given by:

$$dX_t = -\nabla_{\mathcal{W}}\mathcal{F}(X_t,\mu_t) + \sigma dB_t\,; \quad \mu_t := \mathsf{Law}(X_t)\,; \quad X_0 \sim \mu_0\,. \tag{MKV}$$

(MKV) can be seen as a sampling algorithm whenever it converges to a stationary distribution which is same as the global minimizer of $\mathcal{E}(\mu)$. The drift in (MKV) being dependent on μ_t makes it hard to approximate the dynamics algorithmically. Therefore particle approximation is used: let $X_0^{(1)},\dots,X_0^{(n)} \overset{\text{i.i.d.}}{\sim} \mu_0$ be i.i.d. $\hat{\mu}_k$ be the empirical distribution of $(X_k^{(i)})_{i\in[n]}$. Let η be the timestep size and $(Z_k^{(i)})_{k,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\mathbf{I})$. The particle approximation is then given by:

$$X_{k+1}^{(i)} = X_k^{(i)} - \eta\nabla_{\mathcal{W}}\mathcal{F}(X_k^{(i)},\hat{\mu}_k) + \sigma\sqrt{\eta}Z_k^{(i)}\,; \quad \forall\, i \in [n]\,. \tag{pMKV}$$

Optimization, Dynamics and Stochastic Approximation: We motivate our setting via an analogy with gradient descent (GD) over $\mathbb{R}^d$, when the objective function is $F(x) = \mathbb{E}_{\theta\sim P} f(x,\theta)$ and only samples $\theta_1,\dots,\theta_N \overset{\text{i.i.d.}}{\sim} P$ are available. One approach is to consider the minimization of empirical risk $\hat{F}(x) := \frac{1}{N}\sum_{i=1}^N f(x,\theta_i)$ via GD: $x_{t+1} = x_t - \alpha\nabla\hat{F}(x_t)$. However this can be a) computationally expensive and b) need not converge to $\arg\min_x F(x)$. This can lead to sub-optimal convergence of GD Amir et al. (2021). GD's stochastic approximation, stochastic gradient descent (SGD), updates $x_{t+1} = x_t - \alpha\nabla f(x_t; z_t)$ for $t = 1,\dots,N$. Since $\mathbb{E}[f(x_t;z_t)|x_t] = \nabla F(x_t)$, SGD with one pass over the data minimizes F (not $\hat{F}$) Polyak and Juditsky (1992); Chen et al. (2020); Robbins and Monro (1951); Kushner and Yin (2003); Godichon-Baggioni et al. (2021) with near optimal convergence in problems of interest, making it popular in machine learning.

The stochastic approximation algorithm we introduce below directly approximates the population level dynamics in (MKV) directly whenever we have an estimator $\hat{G}(x,y,\xi) : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ such that $\mathbb{E}_{Y\sim\mu,\xi\sim\nu}\hat{G}(x,Y,\xi) = -\nabla_{\mathcal{W}}\mathcal{F}(x,\mu)$ for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and for a given $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. We refer to Section 2.2 for a rich class of functionals $\mathcal{E}$ with such an estimator.

2.1. The Virtual Particle Stochastic Approximation

Given a McKean-Vlasov type SDE of the form $dX_t = G(X_t, \mu_t)dt + \sigma dB_t$, where $\mu_t = \mathsf{Law}(X_t)$, $G : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^d$ is the drift and B_t is the standard Brownian motion. The Euler-Maruyama discretization of this is given by: $X_{k+1} = X_k + \eta G(X_k, \mu_k) + \sigma\sqrt{\eta}Z_k$. Here $(Z_k)_k$ are i.i.d. $\mathcal{N}(0, \mathbf{I})$ random variables, η is the timestep size, and $\mu_k = \mathsf{Law}(X_k)$. However, this is not tractable since we do not know the distribution μ_k. Thus, we introduce the tractable approximation, called the virtual particle stochastic approximation (Algorithm 1), a generalization of VP-SVGD Das and Nagaraj (2023). Suppose there exists an estimator $\hat{G} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ such that whenever $Y, \xi \sim \mu \times \nu$ then $\mathbb{E}[\hat{G}(x, Y, \xi)] = G(x, \mu)$ for every $x \in \mathbb{R}^d$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and ν is a fixed, known distribution. Let the initial distribution be μ_0. Fix number of timesteps T, number of desired samples n and timestep size η.

Algorithm 1: Virtual Particle Stochastic Approximation

Data: Time steps T, number of samples n, timestep size η. Initial Distribution μ_0, distribution ν, estimator $\hat{G}$
Result: $X_T^{(1)}, \dots, X_T^{(n)}$
$X_0^{(1)}, \dots, X_0^{(n)}, Y_0^{(0)}, \dots, Y_0^{(T)} \overset{\text{i.i.d.}}{\sim} \mu_0$
$\xi_1, \dots, \xi_T \overset{\text{i.i.d.}}{\sim} \nu$
$k \leftarrow 0$
while $k \leq T$ **do**
  for $i = 1, .., n$ **do**

$$X_{k+1}^{(i)} = X_k^{(i)} + \eta\hat{G}(X_k^{(i)}, Y_k^{(k)}, \xi_k) + \sigma\sqrt{\eta}Z_k^{(i)} \quad Z_k^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$$

  end
  for $j = k+1, ...T,$ **do**

$$Y_{k+1}^{(j)} = Y_k^{(j)} + \eta\hat{G}(Y_k^{(j)}, Y_k^{(k)}, \xi_k) + \sigma\sqrt{\eta}W_k^{(j)} \quad W_k^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$$

  end
end

At timestep k, the particle $Y_k^{(k)}$ is used to estimate μ_k for all the "real" particles $X_k^{(i)}$, and the remaining "virtual" $Y_k^{(j)}$ are then discarded. Let $\mathcal{R}_k$ be the sigma algebra of $Y_0^{(0)}, \dots, Y_k^{(k)}, \xi_0, \dots, \xi_k$. Then the following can be easily demonstrated:

Lemma 1 *Conditioned on $\mathcal{R}_{k-1}$, $X_k^{(i)}$,$Y_k^{(j)}$ for $i = 1, \dots, k$ and $j = k, \dots, T$ are i.i.d.*

Let $\mu_k|\mathcal{R}_{k-1}$ be the law of $X_k^{(1)}$ conditioned on $\mathcal{R}_{k-1}$. It is a random probability measure which is measurable with respect to $\mathcal{R}_{k-1}$.

Witness Path: In algorithm 1, we call the diagonal trajectory $Y_k^{(k)}$ for $0 \leq k \leq T$ as the 'witness path'. Notice that given a 'witness path, we can obtain a sample from $\mu_T|\mathcal{R}_{T-1}$ in T steps, without fixing n beforehand. Therefore, storing the witness path yields an approximate sampling algorithm from the global minima π of $\mathcal{E}$ whenever the Markov process Equation (MKV) converges to π.

Computational Complexity: Algorithm 1 produces n i.i.d. samples from $\mu_T|\mathcal{R}_{T-1}$, which is shown to converge to π in Theorem 7. The computational complexity is $O(nT+T^2)$, avoiding the $O(n^2T)$ complexity incurred in the straight forward particle approximation (pMKV). Additionally Theorems 7, 10 and 13 show that we can pick T independent of the n to ensure ϵ error.

2.2. Examples

Example 1: Pairwise Interaction Energy Let $V, W : \mathbb{R}^d \to \mathbb{R}$ and $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$. In this case, V is commonly referred to as the external potential and W as the interaction potential. Let W be even (i.e, $W(x) = W(-x)$). The functional $\mathcal{E}$ in this case is defined as:

$$\mathcal{E}(\mu) := \int V(x)d\mu(x) + \tfrac{1}{2}\int\int W(x-y)d\mu(x)d\mu(y) + \tfrac{\sigma^2}{2}\mathcal{H}(\mu)\,. \tag{3}$$

Here the Wasserstein gradient flow gives the following McKean-Vlasov dynamics $dX_t = -\nabla V(X_t)dt - (\nabla W * \mu_t)(X_t) + \sqrt{\sigma}dB_t$ where $\mu_t = \mathsf{Law}(X_t)$. From Ambrosio and Savaré (2007)[Proposition 4.13], for the potential energy functional $\mathcal{V}(\mu) := \int V(x)\mu(dx)$, the Wasserstein gradient is given by $\nabla_{\mathcal{W}}\mathcal{V}(\mu) = \nabla V$. Next, from Ambrosio and Savaré (2007)[Theorem 4.19], for the interaction energy functional $\mathscr{W}(\mu) := \frac{1}{2}\int\int W(x-y)\mu(dx)\mu(dy)$, the Wasserstein gradient is given by $\nabla_{\mathcal{W}}\mathscr{W}(\mu) = (\nabla W) * \mu$ if $(\nabla W) * \mu \in L^2(\mu; \mathbb{R}^d)$. Finally, from Ambrosio and Savaré (2007)[Theorem 4.16], for the entropy functional $\mathcal{H}(\mu) := \int \log \mu d\mu$, the Wasserstein gradient is given by $\nabla_{\mathcal{W}}\mathcal{H}(\mu) = \nabla \log \mu$ if $\nabla \log \mu \in L^2(\mu;\, \mathbb{R}^d)$. Thus,

$$\nabla_{\mathcal{W}}\mathcal{E}(\mu) = \nabla V + (\nabla W) * \mu + \tfrac{\sigma^2}{2}\nabla \log \mu\,. \tag{4}$$

The unique minimizer of the functional $\mathcal{E}$ satisfies a fixed point equation and is given in Kook et al. (2024)[Equation 1.1] as:

$$\pi(x) \propto \exp\left(-\tfrac{2}{\sigma^2}V(x) - \tfrac{2}{\sigma^2}W * \pi(x)\right)\,. \tag{5}$$

Example 2: Mean Field Neural Network Let $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$. Consider the activation function $h(x,z) : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$. We consider the two layer mean-field neural network to be $f(\mu; z) := \int h(x,z)d\mu(x)$. Given data $(Z,W) \sim P$, we consider the square loss:

$$\mathcal{E}(\mu) = \mathbb{E}_{(Z,W)}(f(\mu;Z) - W)^2 + \frac{\lambda}{2}\int \|x\|^2 d\mu(x) + \frac{\sigma^2}{2}\mathcal{H}(\mu)\,. \tag{6}$$

If we have samples $(z_1, w_1), \ldots, (z_n, w_n)$, then we take P to be the empirical distribution. From Nitanda et al. (2022)[Section 2.3], the first variation of the functional $\mathcal{E}$ defined above is:

$$\delta\mathcal{E}(x;\mu) = \mathbb{E}_{(Z,W)}\left[2(f(\mu;Z) - W)h(x,Z)] + \frac{\lambda}{2}\|x\|^2\right] + \tfrac{\sigma^2}{2}(\log\mu + 1)\,,$$

and since $\nabla_{\mathcal{W}}\mathcal{E} = \nabla\delta\mathcal{E}$, we have:

$$\nabla_{\mathcal{W}}\mathcal{E}(\mu) = 2\mathbb{E}_{(Z,W)}\left[(f(\mu;Z) - W)\nabla_x h(x,Z)] + \lambda x\right] + \tfrac{\sigma^2}{2}\nabla\log\mu\,. \tag{7}$$

The unique minimizer of the functional $\mathcal{E}$ is given in Nitanda et al. (2022)[Equation 15] as the solution of the fixed point equation:

$$\pi(x) \propto \exp\left(-\tfrac{2}{\sigma^2}\delta\mathcal{F}(x,\pi)\right)\,. \tag{8}$$

3. Results

In this section, we first establish a convergence theory for Algorithm 1. We begin with a key descent lemma (Lemma 6) that bounds the evolution of the energy functional by decomposing the error into discretization, stochastic, and linearization terms. Using this, we prove our main result (Theorem 7) showing that Algorithm 1 produces i.i.d. samples converging to the minimizer of the energy functional, with rates independent of the number of particles n. We then demonstrate how this framework applies to two important cases: pairwise interaction energy (Section 3.2) and mean-field neural networks (Section 3.3). In both cases, we verify the assumptions of Theorem 7 and establish explicit convergence rates while avoiding the need for separate propagation of chaos bounds.

3.1. General Convergence:

For some functional $\bar{\mathcal{F}} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, let π be the unique minimizer of the functional $\bar{\mathcal{F}} + \frac{\sigma^2}{2}\mathcal{H}$. Define $\bar{\mathcal{E}}(\mu) := \bar{\mathcal{F}}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu) - \bar{\mathcal{F}}(\pi) - \frac{\sigma^2}{2}\mathcal{H}(\pi)$ (not necessarily $\mathcal{E}$). The functional $\bar{\mathcal{E}}$ is introduced, rather than simply analyzing $\mathcal{E}$ since $\bar{\mathcal{E}}$ can have better contraction properties. This is indeed the case with pairwise interaction energy where the KL functional to the minimizer π has a contraction whenever π satisifes an LSI.

We consider Algorithm 1, with $\hat{G}$ such that whenever $(Y, \xi) \sim \mu \times \nu$, we have: $\mathbb{E}[\hat{G}(x, Y, \xi)] = -\nabla_{\mathcal{W}}\mathcal{F}(x, \mu)$ for every $x \in \mathbb{R}^d$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. We do not assume $\mathcal{F} \neq \bar{\mathcal{F}}$, which is important for the case of the interaction energy. We track the evolution of $\mathbb{E}\bar{\mathcal{E}}(\mu_k|\mathcal{R}_{k-1})$ along the discrete time trajectory $\mu_k|\mathcal{R}_{k-1}$ via continuous interpolations. We then specialize to the case of pairwise interaction energy (Equation (3)) in Section 3.2 and mean field neural networks (Equation (6)) in Section 3.3, allowing us to obtain convergence bounds of Algorithm 1 for these cases.

To simplify notation, consider X_0, Y i.i.d. from a distribution $\rho_0 \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ and $\xi \sim \nu$ independent of X_0, Y. For $t \in [0, \eta]$, we consider the random variable $X_t := X_0 + tu(X_0, Y, \xi) + \sigma B_t$, where $u : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a velocity field and B_t is the standard $\mathbb{R}^d$ Brownian motion independent of everything else. Let $\rho_t(\cdot|Y, \xi) := \mathsf{Law}(X_t|Y, \xi)$. Assume that $\mathbb{E}_{Y,\xi}[u(x, Y, \xi)] = -\nabla_{\mathcal{W}}\mathcal{F}(x, \rho_0)$ for every $x \in \mathbb{R}^d$ and that $u(x, Y, \xi)$ has a finite second moment when $x \sim \rho_t(\cdot|Y, \xi)$ almost surely Y, ξ. Here ρ_0, X_0, Y, ξ corresponds to $\mu_k|\mathcal{R}_{k-1}, X_k^{(1)}, Y_k^{(k)}, \xi_k$ respectively. The velocity field $u(x, Y, \xi)$ corresponds to $\hat{G}(x, Y, \xi)$. We use the notation u and $\hat{G}$ interchangeably. Under this correspondence, it is clear that $X_\eta|Y, \xi$ has the same distribution as $\mu_{k+1}|\mathcal{R}_k$. Following the proof of (Vempala and Wibisono, 2019, Lemma 3), we conclude that ρ_t satisfies the Fokker-Planck equation:

$$\begin{aligned}\frac{\partial \rho_t(x|Y, \xi)}{\partial t} &= -\nabla_x.(\rho_t(x|Y, \xi)\mathbb{E}[u(X_0, Y, \xi)|X_t = x, Y, \xi]) + \frac{\sigma^2}{2}\Delta_x \rho_t(x|Y, \xi) \\ &= -\nabla_x.(\rho_t(x|Y, \xi)v_t(x, Y, \xi)), \end{aligned} \tag{9}$$

where $v_t(x, Y, \xi) = \mathbb{E}[u(X_0, Y, \xi)|X_t = x, Y, \xi] - \frac{\sigma^2}{2}\nabla \log \rho_t(x|Y, \xi), \forall x \in \mathbb{R}^d, t \in [0, \eta]$. Taking $\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x, \rho_t(\cdot|Y, \xi)) = \nabla_{\mathcal{W}}\bar{\mathcal{F}}(x, \rho_t(\cdot|Y, \xi)) + \frac{\sigma^2}{2}\nabla \log \rho_t(x|Y, \xi)$, we have the following evolution of $\bar{\mathcal{E}}$ along the trajectory $\rho_t(\cdot|y, \xi)$.

Lemma 2 *$\rho_0 \times \nu$ almost surely (y, ξ), suppose that the energy functional $\bar{\mathcal{E}}(\rho_t(\cdot|y, \xi))$ satisfies:*

$$\nabla_{\mathcal{W}}\bar{\mathcal{E}}(, \rho_t(\cdot|y, \xi)) \in L^2(\rho_t(\cdot|y, \xi)); \quad \frac{d\bar{\mathcal{E}}(\rho_t(\cdot|y, \xi))}{dt} = \left\langle \nabla_{\mathcal{W}}\bar{\mathcal{E}}(\cdot, \rho_t(\cdot|y, \xi)), v_t(\cdot, y, \xi)\right\rangle_{L_2(\rho_t(\cdot|y,\xi))}$$

Then,

$$\frac{d\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))}{dt} = -\int d\rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 + \mathsf{DE}_1(t) + \mathsf{DE}_2(t) + \mathsf{SE}(t) + \mathsf{LE}(t) \tag{10}$$

1. $\mathsf{DE}_1(t) := \left\langle \nabla_{\mathcal{W}}\bar{\mathcal{E}}(\cdot,\rho_t(\cdot|Y,\xi)), \mathbb{E}[u(X_0,Y,\xi)|X_t=\cdot,Y,\xi] - u(\cdot,Y,\xi)\right\rangle_{L_2(\rho_t(\cdot|Y,\xi))}$
2. $\mathsf{DE}_2(t) := \left\langle \nabla_{\mathcal{W}}\bar{\mathcal{E}}(\cdot,\rho_t(\cdot|Y,\xi)), \nabla_{\mathcal{W}}\mathcal{F}(\cdot,\rho_t(\cdot|Y,\xi)) - \nabla_{\mathcal{W}}\mathcal{F}(\cdot,\rho_0)\right\rangle_{L_2(\rho_t(\cdot|Y,\xi))}$
3. $\mathsf{SE}(t) := \left\langle \nabla_{\mathcal{W}}\bar{\mathcal{E}}(\cdot,\rho_t(\cdot|Y,\xi)), u(\cdot,Y,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(\cdot,\rho_0)\right\rangle_{L_2(\rho_t(\cdot|Y,\xi))}$
4. $\mathsf{LE}(t) := \left\langle \nabla_{\mathcal{W}}\bar{\mathcal{E}}(\cdot,\rho_t(\cdot|Y,\xi)), \nabla_{\mathcal{W}}\bar{\mathcal{F}}(\cdot,\rho_t(\cdot|Y,\xi)) - \nabla_{\mathcal{W}}\mathcal{F}(\cdot,\rho_t(\cdot|Y,\xi)\right\rangle_{L_2(\rho_t(\cdot|Y,\xi))}$

Remark 3 *Here,* $\mathsf{DE}_1(t), \mathsf{DE}_2(t), \mathsf{SE}(t), \mathsf{LE}(t)$ *are all functions of* (Y,ξ). *The **discretization error*** DE_1 *arises since* X_0 *is used instead of* X_t *in* $u(X_0,Y,\xi)$ *and* DE_2 *arises since* $-\mathbb{E}[u(x,Y,\xi)] = \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0) \neq \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_t(\cdot|Y,\xi))$. SE *is the **stochastic error**, since we are using an estimator* $u(\cdot,Y,\xi)$ *of* $\nabla_{\mathcal{W}}\mathcal{F}(\cdot,\rho_0)$. LE *is the **linearization error** since we consider the evolution of* $\bar{\mathcal{E}}$ *where as the gradient descent is for* $\mathcal{E} \neq \bar{\mathcal{E}}$. *This is important for linearizing the non-linear Fokker-Planck equation, in the case of pairwise interaction energy in Section 3.2.*

The Lemma below, proved in Section B.1, bounds each of the error terms in expectation.

Lemma 4 *In the setting of Lemma 2, let* π *be the unique global minimizer of* $\bar{\mathcal{E}}$. *Let* $(Y^*,\xi) \sim \pi \times \nu$. *Define* $(\sigma^*)^2 := \mathbb{E}_{x\sim\pi^*}\mathsf{Var}(u(x,Y^*,\xi))$, $(G_\pi)^2 := \mathbb{E}_{x\sim\pi}\|\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)\|^2$ *and* $J_t := \sqrt{\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t,\rho_t(\cdot|Y,\xi))\|^2}$

1. *Suppose the function* $x \to u(x,y,\xi)$ *and* $y \to u(x,y,\xi)$ *are* L_u*-Lipschitz. Then,*

$$\mathbb{E}[\mathsf{DE}_1(t)] \leq L_u J_t \cdot \sqrt{2t^2\left(2L_u^2\mathcal{W}_2^2(\rho_0,\pi) + (\sigma^*)^2 + (G_\pi)^2\right) + \sigma^2 td}\,. \tag{11}$$

2. *Suppose* $\|\nabla_{\mathcal{W}}\mathcal{F}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\nu)\| \leq L_{\mathcal{F}}\mathcal{W}_2(\mu,\nu)$ *Then,*

$$\mathbb{E}[\mathsf{DE}_2(t)] \leq L_{\mathcal{F}} J_t \cdot \sqrt{2t^2\left(2L_u^2\mathcal{W}_2^2(\rho_0,\pi) + (\sigma^*)^2 + (G_\pi)^2\right) + \sigma^2 td}\,. \tag{12}$$

3. *Assume that* $x \to u(x,y,\xi)$, $y \to u(x,y,\xi)$ *and* $x \to \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)$ *are continuously differentiable and* L_u *Lipschitz. Define* $\Theta(x,y,\xi) := u(x,y,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)$. *Assume* $\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x_1,\mu) - \nabla_{\mathcal{W}}\bar{\mathcal{F}}(x_2,\nu)\| \leq L_{\bar{u}}\|x_1-x_2\| + L_{\bar{\mathcal{F}}}\mathcal{W}_2(\mu,\nu)$

$$\begin{aligned}\mathbb{E}[\mathsf{SE}(t)] &\leq \sigma\sqrt{2t}L_u d\sqrt{\mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x,Y,\xi)]} \\ &\quad + 2\sqrt{\mathbb{E}[\mathcal{W}_2^2(\rho_t,\rho_t(\cdot|Y,\xi))]}\left[(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\sqrt{\mathbb{E}\left\|\Theta(X_t,Y,\xi)\right\|^2} + L_u\sqrt{\mathbb{E}\left\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_t,\rho_t)\right\|^2}\right]\end{aligned} \tag{13}$$

4. *Suppose* $\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)\| \leq L_l\mathcal{W}_2(\pi,\mu)$. *Then,*

$$\mathbb{E}[\mathsf{LE}(t)] \leq L_l J_t \cdot \sqrt{\mathbb{E}\left[\mathcal{W}_2^2(\pi,\rho_t(\cdot|Y,\xi))\right]}\,. \tag{14}$$

The Assumption 1 concerns the regularity of the functional $\bar{\mathcal{E}}$ and its stochastic gradients, whereas Assumption 2 gives growth conditions and functional inequalities $\bar{\mathcal{E}}$. Assumption 3 bounds the fluctuations of the stochastic gradient. As we show in specific examples of pairwise interacting systems (Section 3.2) and mean field neural networks (Section 3.3), these are implied by standard assumptions in the literature, and allow us to establish state-of-the-art convergence bounds.

Assumption 1 (Lipschitz continuity) *For some $L_u, L_{\bar{u}}, L_{\bar{\mathcal{F}}}, L_{\mathcal{F}} > 0$, the function $x \to u(x,y,\xi)$ and $y \to u(x,y,\xi)$ are L_u-Lipschitz. For every $x, y \in \mathbb{R}^d$, $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:*

$$(i)\ \ \|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu) - \nabla_{\mathcal{W}}\bar{\mathcal{F}}(y,\nu)\| \le L_{\bar{\mathcal{F}}}\mathcal{W}_2(\mu,\nu) + L_{\bar{u}}\|x-y\|$$
$$(ii)\ \ \|\nabla_{\mathcal{W}}\mathcal{F}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\nu)\| \le L_{\mathcal{F}}\mathcal{W}_2(\mu,\nu) + L_u\|x-y\|$$
$$(iii)\ \ \|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)\| \le L_l\mathcal{W}_2(\pi,\mu)$$

Assumption 2 *Let π be the minimizer of the functional $\bar{\mathcal{F}} + \frac{\sigma^2}{2}\mathcal{H}$. For some $C_{\bar{\mathcal{E}}}, C_{\mathsf{LSI}}, C_{\mathrm{KL}} > 0$, the functional $\bar{\mathcal{E}}$ satisfies the:*

$$(i)\ \ \|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\mu)\|^2_{L^2(\mu)} \ge C_{\bar{\mathcal{E}}}\bar{\mathcal{E}}(\mu) \quad \textit{for all } \mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d) \qquad \text{(Polyak-Łojasiewicz inequality)}$$
$$(ii)\ \ \mathrm{KL}(\mu\,||\,\pi) \le \frac{C_{\mathsf{LSI}}}{2}\mathsf{FD}(\mu\,||\,\pi) \quad \textit{for all } \mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d) \qquad \text{(Log-Sobolev inequality)}$$
$$(iii)\ \ \mathrm{KL}(\mu||\pi) \le C_{\mathrm{KL}}\bar{\mathcal{E}}(\mu) \quad \textit{for all } \mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d) \qquad \text{(KL-Growth)}$$

Remark 5 *We believe we can consider the more natural condition of $\mathcal{W}_2^2$-Growth- $\mathcal{W}_2^2(\mu,\pi) \le C_{\mathcal{W}}\bar{\mathcal{E}}(\mu)$-instead of KL-Growth by a straightforward modification of our proofs.*

Assumption 3 *If $Y, \xi \sim \rho_0 \times \nu$, then from some $C^{\mathsf{Var}}, C_\nu^{\mathsf{Var}} > 0$, and for all $x \in \mathbb{R}^d$:*

$$\mathbb{E}\|u(x,Y,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(x;\rho_0)\|^2 \le C^{\mathsf{Var}}\mathsf{Var}(\rho_0) + C_\nu^{\mathsf{Var}}\,.$$

We now demonstrate the following descent lemma, proved in Section B.2.

Lemma 6 (Descent lemma) *Suppose that Assumptions 1, 2 and 3 are satisfied.*

$$(\sigma^*)^2 := \mathbb{E}_{x\sim\pi}\mathsf{Var}(u(x,Y^*,\xi))\,,\ (G_\pi)^2 := \mathbb{E}_{x\sim\pi}\|\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)\|^2\,,$$
$$(G_{\mathsf{mod}})^2 := \mathbb{E}_{x\sim\pi}\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\pi)\|^2\,.$$

Assume that the following inequalities are satisfied for some small enough $c_0 > 0$:

1. $C_{\bar{\mathcal{E}}} \ge 8L_l^2 C_{\mathrm{KL}} C_{\mathsf{LSI}}$
2. $\eta < c_0 \min\left(\frac{1}{L_u}, \frac{1}{C_{\bar{\mathcal{E}}}}, \frac{C_{\bar{\mathcal{E}}}}{C_{\mathsf{LSI}}C_{\mathrm{KL}}L_u^2(L_{\bar{u}}+L_{\bar{\mathcal{F}}})}, \frac{1}{L_u(L_{\bar{u}}+L_{\bar{\mathcal{F}}})}\sqrt{\frac{C_{\bar{\mathcal{E}}}}{C_{\mathsf{LSI}}C_{\mathrm{KL}}}}, \frac{C_{\bar{\mathcal{E}}}}{C^{\mathsf{Var}}C_{\mathrm{KL}}C_{\mathsf{LSI}}(L_{\bar{u}}+L_{\bar{\mathcal{F}}})}\right)$

Then, for some universal constant $C > 0$:

$$\mathbb{E}\bar{\mathcal{E}}(\rho_\eta) \le e^{-\frac{\eta C_{\bar{\mathcal{E}}}}{8}}\mathbb{E}\bar{\mathcal{E}}(\rho_0) + C\left[\gamma_3\eta^3 + \gamma_2\eta^2 + \gamma_1\eta^{\frac{3}{2}}\right]\,,$$

where $\gamma_3 := (L_u^2 + L_{\mathcal{F}}^2)((\sigma^*)^2 + (G_\pi)^2) + \frac{L_u^2 G_{\mathsf{mod}}^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{C_{\bar{\mathcal{E}}}}$, $\gamma_2 := (L_u^2 + L_{\mathcal{F}}^2)\sigma^2 d + L_u G_{\mathsf{mod}}\sqrt{C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}} + (L_u + L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}) + \frac{\sigma^2 L_u^2 d^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{C_{\bar{\mathcal{E}}}}$ *and* $\gamma_1 := \sigma d L_u\sqrt{C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}}\,.$

Unrolling the recursion established in Lemma 6 allows us to prove our main result:

Theorem 7 *Consider Algorithm 1. Let π be the unique minimizer of $\bar{\mathcal{E}}$. Suppose Assumptions 1, 2, 3 hold with $G_\pi = u$. With $\gamma_1, \gamma_2, \gamma_3$ as defined in Lemma 6, the following holds:*

1. *Conditioned on $\mathcal{R}_{T-1}$, $X_T^{(1)}, \dots, X_T^{(n)} \overset{i.i.d.}{\sim} \mu_T | \mathcal{R}_{T-1}$.*
2. $\mathbb{E}\bar{\mathcal{E}}(\mu_T|\mathcal{R}_{T-1}) \leq e^{-\frac{\eta C_{\bar{\mathcal{E}}} T}{8}} \bar{\mathcal{E}}(\mu_0) + C\left[\frac{\gamma_3 \eta^2}{C_{\bar{\mathcal{E}}}} + \frac{\gamma_2 \eta}{C_{\bar{\mathcal{E}}}} + \frac{\gamma_1 \sqrt{\eta}}{C_{\bar{\mathcal{E}}}}\right]$.

Remark 8 *Note that in Algorithm 1, if we use a batch size $B > 1$, then we would modify the gradient estimator as*

$$\hat{G}_B(x, Y_1, Y_2, \dots, Y_B, \xi) := \frac{1}{B} \sum_{i=1}^{B} \hat{G}(x, Y_i, \xi)\,,$$

where $\{Y_i\}_{i=1}^B$ are i.i.d. Therefore

$$\mathsf{Var}(\hat{G}_B(x, Y_1, Y_2, \dots, Y_B, \xi)) = \frac{1}{B} \mathsf{Var}(\hat{G}(x, Y_1, \xi))\,.$$

The parameter B affects $\mathbb{E}[\mathsf{SE}(t)]$ via $\mathsf{Var}(\hat{G}(x, Y, \xi))$. From Lemma 4, we obtain that the first term (which is the leading order term) in (13) changes to

$$\frac{\sigma\sqrt{2t} L_u d}{\sqrt{B}} \sqrt{\mathbb{E}_{x \sim \rho_0}[\mathsf{Var}(u(x, Y_1, \xi)]}\,. \tag{15}$$

3.2. Pairwise Interaction Energy Functional

Let $V, W : \mathbb{R}^d \to \mathbb{R}$ and $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$. Let W be even (i.e, $W(x) = W(-x)$). Recall that the definition of the functional $\mathcal{F}$, its Wasserstein gradient, and its unique minimizer are in (3), (4), (5) respectively. We call $f : \mathbb{R}^d \to \mathbb{R}$ to be L-smooth if $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for every $x, y \in \mathbb{R}^d$.

Assumption 4 (Smoothness) *V is L_V smooth and W is L_W smooth.*

Assumption 5 (LSI) *π satisfies LSI with constant C_{LSI}, i.e., for all $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$:*

$$\mathrm{KL}(\mu \,||\, \pi) \leq \frac{C_{\mathsf{LSI}}}{2} \mathsf{FD}(\mu \,||\, \pi)\,.$$

The assumption $L_W \leq \frac{\sigma^2}{\sqrt{24 C_{\mathsf{LSI}}}}$ is called " weak interaction" in Kook et al. (2024). Our assumption below is less restrictive in terms of multiplicative constants.

Assumption 6 (Weak Interaction) $L_W \leq \frac{\sigma^2}{4 C_{\mathsf{LSI}}}$.

We define the velocity field $\hat{G}(x, y) := -\nabla V(x) - \nabla W(x - y)\,, \forall\, x \in \mathbb{R}^d$ and $\bar{\mathcal{E}}(\mu) = \frac{\sigma^2}{2} \mathrm{KL}(\mu || \pi)$ (which corresponds to picking $\bar{\mathcal{F}}(\mu) = \int V(x) d\mu(x) + \int W(x - y) d\pi(y) d\mu(x)$), where π is the minimizer of Equation (3). The following Lemma establishes the general Assumptions required for Theorem 7 using the Assumptions 4 and 5. We refer to Section C.1 for the proof.

Lemma 9 *Under Assumptions 4 and 5, the general Assumption 1 is satisfied with* $L_u = L_{\bar{u}} = L_V + L_W$, $L_{\mathcal{F}} = L_W$, $L_{\bar{\mathcal{F}}} = 0$ *and* $L_l = L_W$. *The Assumption 2 is satisfied with* $C_{\mathsf{LSI}} = C_{\mathsf{LSI}}$, $C_{\bar{\mathcal{E}}} = \frac{\sigma^2}{C_{\mathsf{LSI}}}$ *and* $C_{\mathrm{KL}} = \frac{2}{\sigma^2}$. *The Assumption 3 is satisfied with* $C^{\mathsf{Var}} = 2L_W^2$ *and* $C_\nu^{\mathsf{Var}} = 0$.

Theorem 10, proved in Section C.2, instantiates Theorem 7 to the Pairwise Interaction Energy.

Theorem 10 *Consider the Pairwise Interaction Energy in Equation (3) under Assumptions 4, 5 and 6. There exist universal constants* $c_0, C > 0$ *such that Algorithm 1 with* $\hat{G}$ *as above with* $\eta < c_0 \min\left(\frac{C_{\mathsf{LSI}}}{\sigma^2}, \frac{\sigma^4}{C_{\mathsf{LSI}}^2 (L_V + L_W)^3}\right)$ *satisfies:*

$$\begin{aligned}\mathbb{E}\left[\mathrm{KL}(\mu_T | \mathcal{R}_{T-1} || \pi)\right] \le e^{\left(-\frac{T\sigma^2\eta}{8C_{\mathsf{LSI}}}\right)} \mathrm{KL}(\mu_0 || \pi) + C\frac{\sqrt{\eta} d^{3/2} (C_{\mathsf{LSI}})^{1/2} \sigma (L_V + L_W)}{4} \\ + C\eta(L_V + L_W)^2 d^2 C_{\mathsf{LSI}} \,.\end{aligned} \tag{16}$$

Remark 11 *Given* $\epsilon/3 \in (0, \mathrm{KL}(\mu_0||\pi) \wedge 1)$, *as per Theorem 10, we can achieve* $\mathbb{E}\left[\mathrm{KL}(\mu_T | \mathcal{R}_{T-1}) - \mathrm{KL}(\pi)\right] \le \epsilon$ *by picking:*

1. $\eta = \frac{8C_{\mathsf{LSI}}}{\sigma^2 T} \log\left(\frac{3\,\mathrm{KL}(\mu_0||\pi)}{\epsilon}\right)$
2. $T \gtrsim \max\left(\frac{C_{\mathsf{LSI}}^2 d^3 (L_V + L_W)^2}{\epsilon^2}, \frac{C_{\mathsf{LSI}}^2 d^2 (L_V + L_W)^2}{\sigma^2 \epsilon}, \frac{C_{\mathsf{LSI}}^3 (L_V + L_W)^3}{\sigma^6}\right) \log\left(\frac{3\,\mathrm{KL}(\mu_0||\pi)}{\epsilon}\right)$

Comparison with prior work: We refer to the general discussion on establishing isoperimetry for n particle systems in Section 1. In the specific case of pairwise interaction, prior work Kook et al. (2024) considers the standard particle algorithm for this problem and shows that the n-particle stationary distribution with $n = \frac{\sqrt{d}}{\epsilon}$, which ensures that the law of the first particle Y_1 is such that $\mathsf{KL}(\mathsf{Law}(Y_1)||\pi) \le \epsilon^2$. To obtain sampling guarantees for the n particle system, they further assume the potentials V, W are decomposed as $V = V_0 + V_1, W = W_0 + W_1$, where V_0, W_0 are uniformly convex and $\mathsf{osc}(V_1) := \sup V_1 - \inf V_1 < \infty$, $\mathsf{osc}(W_1) := \sup W_1 - \inf W_1 < \infty$. Note that V_1, W_1 can be interpreted as perturbations of V_0, W_0, which are uniformly convex, and the assumption $\mathsf{osc}(V_1), \mathsf{osc}(W_1) < \infty$ means the perturbations V_1, W_1 are bounded. The isoperimetric constant $C_{\mathsf{LSI},n}$ is then bounded with an exponential dependence on $\mathsf{osc}(V_1) + \mathsf{osc}(W_1)$. Our work does not require these additional assumptions. However, the standard particle based method as above can use any standard sampling algorithm to sample from the n-particle stationary distribution (not restricted to LMC style Euler-Maruyama discretization as in this work).

3.3. Mean Field Neural Network

In (6), we assume P to be empirical distribution of m data samples $(z_1, w_1), \ldots, (z_m, w_m)$. Then, the functional in (6) simplifies to:

$$\mathcal{F}(\mu) = \frac{1}{m}\sum_{i=1}^m \left(\int h(z_i, x) d\mu(x) - w_i\right)^2 + \frac{\lambda}{2}\int \|x\|^2 d\mu(x)\,, \tag{17}$$

The Wasserstein gradient of this functional follows from (7) to be:

$$\nabla_{\mathcal{W}} \mathcal{F}(x; \mu) = \frac{2}{m}\sum_{i=1}^m \left(\int h(z_i, y) d\mu(y) - w_i\right) \nabla_x h(z_i, x) + \lambda x\,.$$

The unique minimizer of the functional $\mathcal{E}$ is given in (8). We set $\bar{\mathcal{E}}(\mu) = \mathcal{E}(\mu) - \mathcal{E}(\pi)$. We consider the proximal Gibbs distribution corresponding to μ, which is given by:

$$\pi_\mu(x) \propto \exp\left(-\frac{2\delta\mathcal{F}(x,\mu)}{\sigma^2}\right) ,$$

where $\delta\mathcal{F}$ is the first-variation of the functional $\mathcal{F}$. We let $\nu = \mathsf{Unif}([m])$. We denote the random variable ξ, used in the definition of the Wasserstein gradient estimator $\hat{G}$, by I and choose $\hat{G}$ to be:

$$\hat{G}(x,y,i) := -(h(z_i,y) - w_i)\nabla_x h(z_i,x) - \lambda x \,,\, \forall\, x,y \in \mathbb{R}^d \,, i \in [m]\,.$$

Assumption 7 (Boundedness) *For every $x, z \in \mathbb{R}^d$:*

$$\|h(z,x)\| \leq B; \quad \|z\| \leq R; \quad |w| \leq R; \quad \|\nabla_x h(z,x)\| \leq M\|z\|\,.$$

Assumption 8 (Lipschitz continuity) *The function $x \to \nabla h(z,x)$ is $L\|z\|$-Lipschitz.*

Excluding special cases, the boundedness assumption on $h, \nabla_x h$ and Lipschitz assumption on $\nabla_x h$ are necessary to satisfy the general assumptions in prior works Wang (2024); Nitanda (2024) for the square loss. For instance, we refer to (Nitanda, 2024, Assumption 1).

Assumption 9 (LSI) *For any $q \in \mathcal{P}_2(\mathbb{R}^d)$, the proximal Gibbs distribution π_q satisfies the LSI with constant C_{LSI}. π also satisfies LSI with the same constant.*

The following lemma is proved in Section D.1

Lemma 12 *Under Assumptions 7, 8 and 9, the general Assumption 1 is satisfied with $L_u = L_{\bar{u}} = (B+R)LR + \lambda + M^2R^2$, $L_{\mathcal{F}} = L_{\bar{\mathcal{F}}} = M^2R^2$, and $L_l = 0$. The general Assumption 2 is satisfied with $C_{\mathsf{LSI}} = C_{\mathsf{LSI}}$, $C_{\bar{\mathcal{E}}} = \frac{\sigma^2}{C_{\mathsf{LSI}}}$ and $C_{\mathrm{KL}} = \frac{2}{\sigma^2}$. The Assumption 3 is satisfied with $C^{\mathsf{Var}} = 0$ and $C^{\mathsf{Var}}_\nu = 4M^2R^2(B+R)^2$.*

Theorem 13 *Consider the case of the Mean Field Neural Network with square loss in Equation (6) under Assumptions 8, 7 and 9. We consider Algorithm 1 with $\hat{G}$ as defined above and $\eta < c_0 \min\left(\frac{C_{\mathsf{LSI}}}{\sigma^2}, \frac{\sigma^4}{C_{\mathsf{LSI}}^2 L_u^3}\right)$ for some $c_0 > 0$ small enough and $L_u = (B+R)LR + \lambda + M^2R^2$. Then for some universal constant $C > 0$:*

$$\begin{aligned}&\mathbb{E}\mathcal{E}(\mu_T|\mathcal{R}_{T-1}) - \mathcal{E}(\pi)\\ &\leq e^{-\frac{T\eta\sigma^2}{8C_{\mathsf{LSI}}}}(\mathcal{E}(\mu_0) - \mathcal{E}(\pi)) + C\frac{C_{\mathsf{LSI}}}{\sigma^2}\left[\eta(\sigma^2L_u^2d + L_uM^2R^2(B+R)^2) + \sqrt{\eta}\sigma dL_uMR(B+R)\right]\,.\end{aligned}$$

Remark 14 *Given $\epsilon/3 \in (0, \mathcal{E}(\mu_0) - \mathcal{E}(\pi))$, as per Theorem 13, we can achieve $\mathbb{E}\mathcal{E}(\mu_T|\mathcal{R}_{T-1}) - \mathcal{E}(\pi) \leq \epsilon$ by picking*

1. $\eta = \frac{8C_{\mathsf{LSI}}}{\sigma^2 T}\log\left(\frac{3(\mathcal{E}(\mu_0)-\mathcal{E}(\pi))}{\epsilon}\right)$
2. $T \gtrsim \max\left(\frac{C_{\mathsf{LSI}}^3 d^2 L_u^2 M^2R^2(B+R)^2}{\sigma^4\epsilon^2}, \frac{C_{\mathsf{LSI}}^2(\sigma^2L_u^2d + L_uM^2R^2(B+R)^2)}{\sigma^4\epsilon}, \frac{L_u^3C_{\mathsf{LSI}}^3}{\sigma^6}\right)\log\left(\frac{3(\mathcal{E}(\mu_0)-\mathcal{E}(\pi))}{\epsilon}\right)$

4. Conclusion

We introduce a novel convergence analysis of virtual particle stochastic approximation, achieving state-of-the-art convergence when unbiased estimators for the Wasserstein gradient are accessible. Future research directions include extending this analysis to scenarios involving biased-but-consistent estimators, leveraging techniques to convert them into unbiased estimators with heavy tails Blanchet and Glynn (2015). Furthermore, investigating gradient flows within the Wasserstein-Fischer-Rao Geometry Chizat et al. (2018b,a), which has demonstrated significant potential in statistics and inference, presents a promising avenue. Finally, exploring the algorithm's behavior under weaker conditions, such as the Poincaré inequality, is another interesting area for future study.

Acknowledgments

We thank the reviewers for many helpful comments. Anant Raj was supported by a grant from Ittiam Systems Private Limited through the Ittiam Equitable AI Lab.

References

Luigi Ambrosio and Giuseppe Savaré. Gradient flows of probability measures. In *Handbook of differential equations: evolutionary equations*, volume 3, pages 1–136. Elsevier, 2007.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

Idan Amir, Tomer Koren, and Roi Livni. Sgd generalizes better than gd (and regularization doesn't help). In *Conference on Learning Theory*, pages 63–92. PMLR, 2021.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Krishnakumar Balasubramanian, Sayan Banerjee, and Promit Ghosal. Improved finite-particle convergence rates for stein variational gradient descent. *arXiv preprint arXiv:2409.08469*, 2024.

Jose H Blanchet and Peter W Glynn. Unbiased monte carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE, 2015.

Nawaf Bou-Rabee and Katharina Schuh. Nonlinear hamiltonian monte carlo & its particle approximation. *arXiv preprint arXiv:2308.11491*, 2023.

René Carmona and François Delarue. Forward–backward stochastic differential equations and controlled mckean–vlasov dynamics. *The Annals of Probability*, pages 2647–2700, 2015.

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

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234, 2020.

Sinho Chewi, Atsushi Nitanda, and Matthew S Zhang. Uniform-in-n log-sobolev inequality for the mean-field langevin dynamics with convex energy. *arXiv preprint arXiv:2409.10440*, 2024.

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

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18:1–44, 2018a.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018b.

Aniket Das and Dheeraj Nagaraj. Provably fast finite particle variants of svgd via virtual particle stochastic approximation. *Advances in Neural Information Processing Systems*, 36:49748–49760, 2023.

Aniket Das, Dheeraj M Nagaraj, and Anant Raj. Utilising the clt structure in stochastic gradient based sampling: Improved analysis and faster algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4072–4129. PMLR, 2023.

John Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3(1):2325–5870, 2007.

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

Antoine Godichon-Baggioni, Nicklas Werge, and Olivier Wintenberger. Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *arXiv preprint arXiv:2109.07117*, 2021.

Xunpeng Huang, Difan Zou, Yi-An Ma, Hanze Dong, and Tong Zhang. Faster sampling via stochastic gradient proximal sampler. *arXiv preprint arXiv:2405.16734*, 2024.

Shi Jin, Lei Li, and Jian-Guo Liu. Random batch methods (rbm) for interacting particle systems. *Journal of Computational Physics*, 400, 2020.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Saravanan Kandasamy and Dheeraj Nagaraj. The poisson midpoint method for langevin dynamics: Provably efficient discretization for diffusion models. *arXiv preprint arXiv:2405.17068*, 2024.

Yuri Kinoshita and Taiji Suzuki. Improved convergence rate of stochastic gradient langevin dynamics with variance reduction and its application to optimization. *Advances in Neural Information Processing Systems*, 35:19022–19034, 2022.

Yunbum Kook, Matthew S Zhang, Sinho Chewi, Murat A Erdogdu, and Mufan Bill Li. Sampling from the mean-field stationary distribution. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 3099–3136. PMLR, 2024.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Daniel Lacker. Independent projections of diffusions: gradient flows for variational inference and optimal mean field approximations. *arXiv preprint arXiv:2309.13332*, 2023.

Daniel Lacker and Luc Le Flem. Sharp uniform-in-time propagation of chaos. *Probability Theory and Related Fields*, 187(1-2):443–480, 2023.

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

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

Florent Malrieu. Convergence to equilibrium for granular media equations and their euler schemes. *The Annals of Applied Probability*, 13(2):540–560, 2003.

Florient Malrieu. Logarithmic sobolev inequalities for some nonlinear pde's. *Stochastic processes and their applications*, 95(1):109–132, 2001.

Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1): 153–179, 1997.

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

Atsushi Nitanda. Improved particle approximation error for mean field neural networks. *arXiv preprint arXiv:2405.15767*, 2024.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

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

Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Taiji Suzuki, Atsushi Nitanda, and Denny Wu. Uniform-in-time propagation of chaos for the mean-field gradient langevin dynamics. In *The Eleventh International Conference on Learning Representations*, 2023.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Mean-field langevin dynamics: Time-space discretization, stochastic gradient, and variance reduction. *Advances in Neural Information Processing Systems*, 36, 2024.

Alain-Sol Sznitman. Topics in propagation of chaos. *Ecole d'été de probabilités de Saint-Flour XIX—1989*, 1464:165–251, 1991.

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

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Songbo Wang. Uniform log-sobolev inequalities for mean field particles with flat-convex energy. *arXiv preprint arXiv:2408.03283*, 2024.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Yuling Yan, Kaizheng Wang, and Philippe Rigollet. Learning gaussian mixtures using the wasserstein–fisher–rao gradient flow. *The Annals of Statistics*, 52(4):1774–1795, 2024.

Rentian Yao and Yun Yang. Mean field variational inference via wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*, 2022.

Lu Yu, Avetik Karagulyan, and Arnak Dalalyan. Langevin monte carlo for strongly log-concave distributions: Randomized midpoint revisited. *arXiv preprint arXiv:2306.08494*, 2023.

Appendix A. Useful Technical Results

Recall that the random variable X_t is defined as $X_t := X_0 + tu(X_0, Y, \xi) + \sigma B_t$, where $u : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is the velocity field. Also, note that $\rho_t := \mathsf{Law}(X_t)$ and $\rho_t(\cdot|Y,\xi) := \mathsf{Law}(X_t|Y,\xi)$. Let $\mathsf{Var}(\mu)$ denote the trace of the covariance matrix of the probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathsf{Var}(u(x,Y,\xi))$ the variance of the velocity field u with respect to the random variables Y, ξ, for every fixed $x \in \mathbb{R}^d$. Next, $\mathcal{W}_p(\mu,\nu)$ denotes the p-Wasserstein distance between the probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\nabla_{\mathcal{W}}\mathcal{F}$ the Wasserstein gradient of the functional $\mathcal{F}$. Recall that, for a functional $\bar{\mathcal{F}}$, we define $\bar{\mathcal{E}}(\mu) := \bar{\mathcal{F}}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu) - \bar{\mathcal{F}}(\pi) - \frac{\sigma^2}{2}\mathcal{H}(\pi)$, where π is the minimizer of the functional $\bar{\mathcal{F}} + \frac{\sigma^2}{2}\mathcal{H}$. The functional $\bar{\mathcal{E}}$ need not be the same as the functional $\mathcal{E}$. Let $C_{\bar{\mathcal{E}}}, C_{\mathsf{LSI}}, C_{\mathsf{KL}}$ be the constants corresponding to (Polyak-Łojasiewicz inequality), (Log-Sobolev inequality), (KL-Growth) in Assumption 2, respectively and $L_{\bar{u}}, L_{\bar{\mathcal{F}}}$ the constants defined in Assumption 1. Furthermore, $\|\cdot\|$ denotes the L^2 norm of a function.

The following lemma bounds the expected Wasserstein distance between $\rho_t, \rho_t(|Y,\xi)$.

Lemma 15

$$\mathbb{E}\mathcal{W}_2^2(\rho_t, \rho_t(\cdot|Y,\xi)) \leq 2t^2\mathbb{E}_{x\sim\rho_0}\mathsf{Var}(u(x,Y,\xi))$$

$$\mathbb{E}\mathcal{W}_2^2(\rho_0, \rho_t(\cdot|Y,\xi)) \leq \sigma^2 td + t^2\mathbb{E}\|u(X_0,Y,\xi)\|^2$$

Proof Notice that $\rho_t(\cdot) = \mathbb{E}\rho_t(\cdot|Y,\xi)$. Since Wasserstein distance is convex in each of its coordinates and $x \to x^2$ is increasing and convex over $\mathbb{R}^+$, we have:

$$\mathbb{E}\mathcal{W}_2^2(\rho_t, \rho_t(\cdot|Y,\xi)) \leq \mathbb{E}\mathcal{W}_2^2(\rho_t(\cdot|Y',\xi'), \rho_t(\cdot|Y,\xi))\,,$$

where (Y',ξ') is an independent copy of (Y,ξ). We now couple $\rho_t(\cdot|Y',\xi')$ and $\rho_t(\cdot|Y,\xi)$ for a given Y,ξ as follows:

$$X_t' = X_0 + tu(X_0,Y',\xi') + \sigma B_t\,; \qquad X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t\,.$$

Therefore,

$$\mathbb{E}\mathcal{W}_2^2(\rho_t, \rho_t(\cdot|Y,\xi)) \leq t^2\mathbb{E}\|u(X_0,Y',\xi') - u(X_0,Y,\xi)\|^2 = 2t^2\mathbb{E}_{x\sim\rho_0}\mathsf{Var}(u(x,Y,\xi)) \tag{18}$$

Finally, we couple ρ_0 and $\rho_t(\cdot|Y,\xi)$ by $X_0 \sim \rho_0$ and $X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t$. This implies, $\mathbb{E}\mathcal{W}_2^2(\rho_0, \rho_t(\cdot|Y,\xi)) \leq \sigma^2 td + t^2\mathbb{E}\|u(X_0,Y,\xi)\|^2$. ∎

Lemma 16 (Lemma 11 in Vempala and Wibisono (2019)) *Let ν be a probability measure over $\mathbb{R}^d$ with density $\nu(x) \propto e^{-F(x)}$ where F is L-smooth. Then,*

$$\mathbb{E}_{x\sim\nu}\|\nabla F(x)\|^2 \leq dL$$

Lemma 17 (Otto-Villani Theorem Otto and Villani (2000)) *Let π satisfy LSI with constant C_{LSI}. Then π satisfies Talangrand's inequality T_p for any $p \in [1,2]$, i.e., for all $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$:*

$$\mathcal{W}_p^2(\mu,\pi) \leq 2C_{\mathsf{LSI}}\,\mathrm{KL}(\mu,\pi)\,.$$

Lemma 18 *For any $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, we have:*

1. $$\mathsf{Var}(\mu) \leq 2\mathcal{W}_2^2(\mu,\pi) + 2\mathsf{Var}(\pi)$$

2. *Suppose Assumption 1 holds. Let $X \sim \mu$ and $X^* \sim \pi$. Then:*

$$\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X,\mu)\|^2 \leq 3(L_{\bar{u}}^2 + L_{\bar{\mathcal{F}}}^2)\mathcal{W}_2^2(\mu,\pi) + 3\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X^*,\pi)\|^2$$

Proof Let X, X' be i.i.d. from μ. Let Y, Y' be i.i.d. from π such that Y, X are optimally coupled and Y', X' are optimally coupled. Now consider:

$$\begin{aligned}\mathsf{Var}(\mu) &= \frac{1}{2}\mathbb{E}\|X - Y + Y - Y' + Y' - X'\|^2 \\ &\leq \mathbb{E}\|X - Y + Y' - X'\|^2 + \mathbb{E}\|Y - Y'\|^2 \\ &= \mathbb{E}\|X - Y + Y' - X'\|^2 + 2\mathsf{Var}(\pi) \\ &= 2\mathbb{E}\|X - Y\|^2 - 2\|\mathbb{E}X - \mathbb{E}Y\|^2 + 2\mathsf{Var}(\pi) \\ &\leq 2\mathcal{W}_2^2(\mu, \pi) + 2\mathsf{Var}(\pi)\,,\end{aligned}$$

where in the penultimate step we have used the fact that $X - Y, X' - Y'$ are i.i.d. Next, let $X^* \sim \pi$ be optimally coupled to $X \sim \mu$. By Assumption 1 and triangle inequality, we have:

$$\begin{aligned}\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X, \mu)\| &= \|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X, \mu) - \nabla_{\mathcal{W}}\bar{\mathcal{F}}(X^*, \pi) + \nabla_{\mathcal{W}}\bar{\mathcal{F}}(X^*, \pi)\| \\ &\leq L_{\bar{\mathcal{F}}}\mathcal{W}_2(\mu, \pi) + L_{\bar{u}}\|X - X^*\| + \|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X^*, \pi)\|\,.\end{aligned}$$

By squaring, applying $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$, and taking expectation proves the second statement of this lemma. ■

Lemma 19 $V(u, \rho_0) := \mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x, Y, \xi))]$. *Under Assumptions 2 and 3, we have:*

$$V(u, \rho_0) \lesssim C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}} \bar{\mathcal{E}}(\rho_0) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}} \tag{19}$$

$$\mathsf{Var}(\rho_0) \lesssim C_{\mathsf{LSI}} C_{\mathrm{KL}} \bar{\mathcal{E}}(\rho_0) + \mathsf{Var}(\pi) \tag{20}$$

Proof By Assumption 3, we have:

$$\begin{aligned}V(u, \rho_0) &\leq C^{\mathsf{Var}}\mathsf{Var}(\rho_0) + C_\nu^{\mathsf{Var}} \\ &\lesssim C^{\mathsf{Var}}\mathcal{W}_2^2(\rho_0, \pi) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}} && \text{(By Lemma 18)} \\ &\lesssim C^{\mathsf{Var}} C_{\mathsf{LSI}}\,\mathrm{KL}(\rho_0 || \pi) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}} && \text{(By Lemma 17)} \\ &\lesssim C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}} && \text{(By Assumption 2-(KL-Growth))}\end{aligned}$$

Applying a similar reasoning to the bounds in Lemma 18, we conclude the bound on $\mathsf{Var}(\rho_0)$ ■

Lemma 20 *Consider a probability measure π over $\mathbb{R}^d$, with density $\pi(x) \propto \exp(-F(x))$, which satisfies the LSI with constant C_{LSI}. Then,* $\mathsf{Var}(\pi) \leq dC_{\mathsf{LSI}}$.

Proof By Bakry et al. (2013)[Definition 4.2.1], the probability measure π satisfies the Poincaré inequality with constant $C_{\mathsf{PI}} > 0$ if

$$\mathsf{Var}_\pi(f) \leq C_{\mathsf{PI}}\mathbb{E}_\pi[\|\nabla f\|^2]\,,$$

for all $f : \mathbb{R}^d \to \mathbb{R}$ such that f is continuously differentiable and ∇f is square integrable with respect to π, . For any $i \in \{1, 2, \ldots, d\}$ and $x = (x_1, x_2, \ldots, x_d)$, let $f(x) = x_i$. Then $\|\nabla f(x)\| = 1$, for all $x \in \mathbb{R}^d$. Thus, by Poincaré inequality, $\mathsf{Var}_\pi(f(X)) = \mathsf{Var}(X_i) \leq C_{\mathsf{PI}}$. Now, $\mathsf{Var}(\pi) = \sum_{i=1}^d \mathsf{Var}(X_i) \leq dC_{\mathsf{PI}}$. Finally, since π satisfying the LSI implies that it also satisfies the Poincaré inequality with the same constant C_{LSI}, we conclude that $\mathsf{Var}(\pi) \leq dC_{\mathsf{LSI}}$. ■

Lemma 21 *Consider a probability measure π over $\mathbb{R}^d$ with density $\pi(x) \propto \exp(-F(x))$ satisfies the Logarithmic Sobolev Inequality with constant C_{LSI}. Assume that F is L-smooth (i.e, ∇F is L-Lipschitz) and ∇F is square integrable with respect to π. Then,*

$$C_{\mathsf{LSI}} \geq \frac{1}{L}$$

Proof Let $x_\pi \in \mathbb{R}^d$ be the mean of π (a fixed vector). Without loss of generality, we assume $\pi(x) = e^{-F(x)}$. This requires adding a constant to $F(x)$ in the original definition, which does not change ∇F. Using integration by parts, we have:

$$\begin{aligned} d &= \int \langle x - x_\pi, \nabla F(x) \rangle e^{-F(x)} dx \\ &\leq \sqrt{\mathsf{Var}(\pi)} \sqrt{\mathbb{E}_\pi \|\nabla F\|^2} && \text{(Cauchy-Schwarz Inequality)} \\ &\leq \sqrt{d C_{\mathsf{LSI}}} \sqrt{dL} && \text{(Lemma 20 and Lemma 16)} \end{aligned}$$

The claim follows from the above equation. ■

Appendix B. Proof of Technical Lemmas

B.1. Proof of Lemma 4

Proof

1. First, we bound $\mathbb{E}_{Y,\xi}[\mathsf{DE}_1(t)]$. Moving the expectation out of the inner product, applying Cauchy-Schwarz inequality, using the assumption that the function $(x, y) \to u(x, y, \xi)$ is L_u-Lipschitz, and the fact that $X_t = X_0 + tu(X_0, Y, \xi) + \sigma B_t$, we get:

$$\begin{aligned} \mathbb{E}_{Y,\xi}[\mathsf{DE}_1(t)] &= \mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{E}}(X_t, \rho_t(\cdot|Y, \xi)), u(X_0, Y, \xi) - u(X_t, Y, \xi) \rangle \\ &\leq \mathbb{E}\left[\|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(X_t, \rho_t(\cdot|Y, \xi))\| \cdot \|u(X_0, Y, \xi) - u(X_t, Y, \xi)\| \right] \\ &\leq L_u \mathbb{E}\left[\|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(X_t, \rho_t(\cdot|Y, \xi))\| \cdot \|X_t - X_0\| \right] \\ &= L_u \mathbb{E}\left[\|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(X_t, \rho_t(\cdot|Y, \xi))\| \cdot \|tu(X_0, Y, \xi) + \sigma B_t\| \right] \\ &\leq L_u \sqrt{\mathbb{E}\|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(X_t, \rho_t(\cdot|Y, \xi))\|^2} \sqrt{\mathbb{E}\|tu(X_0, Y, \xi) + \sigma B_t\|^2} \\ &= L_u \sqrt{\mathbb{E}\|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(X_t, \rho_t(\cdot|Y, \xi))\|^2} \sqrt{t^2 \mathbb{E}\|u(X_0, Y, \xi)\|^2 + \sigma^2 t d}\,. \end{aligned} \tag{21}$$

   Next, we bound $\mathbb{E}\|u(X_0, Y, \xi)\|^2$. Let $(X^*, Y^*) \sim \pi \times \pi$ be optimally coupled to $(X_0, Y) \sim \rho_0 \times \rho_0$ in the 2-Wasserstein distance and independent of ξ. Thus, by the triangle inequality, the inequality $(a + b)^2 \leq 2a^2 + 2b^2$, by L_u-Lipschitz continuity of $(x, y) \to u(x, y, \xi)$, and the fact that $\mathcal{W}_2^2(\mu \times \mu, \nu \times \nu) \leq 2\mathcal{W}_2^2(\mu, \nu)$, for any probability measures μ, ν, we have:

$$\begin{aligned} \mathbb{E}\|u(X_0, Y, \xi)\|^2 &\leq 2\mathbb{E}\|u(X_0, Y, \xi) - u(X^*, Y^*, \xi)\|^2 + 2\mathbb{E}\|u(X^*, Y^*, \xi)\|^2 \\ &\leq 2L_u^2 \mathbb{E}[\|X_0 - X^*\|^2 + \|Y - Y^*\|^2] + 2\mathbb{E}\|u(X^*, Y^*, \xi)\|^2 \\ &= 2L_u^2 \mathcal{W}_2^2(\rho_0 \times \rho_0, \pi \times \pi) + 2\mathbb{E}\|u(X^*, Y^*, \xi)\|^2 \\ &\leq 4L_u^2 \mathcal{W}_2^2(\rho_0, \pi) + 2\mathbb{E}\|u(X^*, Y^*, \xi)\|^2\,. \end{aligned} \tag{22}$$

Next, using $\mathbb{E}[u(x,Y^*,\xi)] = -\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)$, for every x, we obtain:

$$\begin{aligned}\mathbb{E}\|u(X^*,Y^*,\xi)\|^2 &= \mathbb{E}\|u(X^*,Y^*,\xi)+\nabla_{\mathcal{W}}\mathcal{F}(X^*,\pi)\|^2 + \mathbb{E}\|\nabla_{\mathcal{W}}\mathcal{F}(X^*,\pi)\|^2 \\ &= (\sigma^*)^2 + (G_\pi)^2 \,. \end{aligned} \tag{23}$$

Now by using the bounds (22) and (23) in (21) proves (11).

2. Next, we bound $\mathbb{E}_{Y,\xi}[\mathsf{DE}_2(t)]$. Using the assumption $\|\nabla_{\mathcal{W}}\mathcal{F}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\nu)\| \leq L_{\mathcal{F}}\mathcal{W}_2(\mu,\nu)$, applying the Cauchy-Schwarz inequality and Jensen's inequality, we get:

$$\begin{aligned}\mathbb{E}_{Y,\xi}[\mathsf{DE}_2(t)] &\leq L_{\mathcal{F}}\mathbb{E}\left[\mathcal{W}_2(\rho_t(\cdot|Y,\xi),\rho_0)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t,\rho_t(\cdot|Y,\xi))\|\right] \\ &\leq L_{\mathcal{F}}\sqrt{\mathbb{E}[\mathcal{W}_2^2(\rho_t(\cdot|Y,\xi),\rho_0)]}\cdot\sqrt{\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t,\rho_t(\cdot|Y,\xi))\|^2}\,. \end{aligned} \tag{24}$$

Next, we bound $\mathbb{E}[\mathcal{W}_2^2(\rho_t(\cdot|Y,\xi),\rho_0)]$. Since $X_t|Y,\xi \sim \rho_t(\cdot|Y,\xi)$, $X_0 \sim \rho_0$ and $X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t$. Therefore, by the definition of the Wasserstein distance, we have:

$$\begin{aligned}\mathbb{E}[\mathcal{W}_2^2(\rho_t(\cdot|Y,\xi),\rho_0)] &\leq \mathbb{E}[\mathbb{E}[\|X_t - X_0\|^2|Y,\xi]] \\ &= \mathbb{E}\|tu(X_0,Y,\xi)+\sigma B_t\|^2 \\ &= t^2\mathbb{E}_{(X,Y,\xi)\sim\rho_0\times\rho_0\times\nu}\|u(X,Y,\xi)\|^2 + \sigma^2 td \\ &\leq 4t^2L_u^2\mathcal{W}_2^2(\rho_0,\pi) + 2t^2((\sigma^*)^2+(G_\pi)^2) + \sigma^2 td\,, \end{aligned} \tag{25}$$

where the last inequality follows from (22) and (23). By plugging the bound in (25) into (24) proves (12).

3. Next, we bound $\mathbb{E}_{Y,\xi}[\mathsf{SE}(t)]$. Define $\Theta(x,y,\xi) := u(x,y,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)$. Note that since $\mathbb{E}[u(x,Y,\xi)] = -\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)$, for every x, we have:

$$\begin{aligned}\mathbb{E}[\nabla_x\cdot\Theta(x,Y,\xi)] &= \mathbb{E}[\nabla_x\cdot u(x,Y,\xi) + \nabla_x\cdot\nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)] \\ &= \nabla_x\cdot\mathbb{E}[u(x,Y,\xi)+\nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)] = 0\,. \end{aligned} \tag{26}$$

Here, we have exchanged the integral and derivative using the dominated convergence theorem along with Lipchitz continuity of $u, \nabla_{\mathcal{W}}\mathcal{F}$. Using $\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|y,\xi)) = \nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t(\cdot|y,\xi)) + \frac{\sigma^2}{2}\nabla\log\rho_t(x|y,\xi)$ and integrating by parts, we get:

$$\begin{aligned}\mathbb{E}_{Y,\xi}[\mathsf{SE}(t)] &= \mathbb{E}\left\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(\cdot,\rho_t(\cdot|Y,\xi)),\Theta(\cdot,Y,\xi)\right\rangle_{L_2(\rho_t(\cdot|Y,\xi))} \\ &\quad+\frac{\sigma^2}{2}\mathbb{E}\int\rho_t(x|Y,\xi)\langle\nabla_x\log\rho_t(x|Y,\xi)\,,\,\Theta(x,Y,\xi)\rangle dx \\ &= \mathbb{E}\left\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(\cdot,\rho_t(\cdot|Y,\xi)),\Theta(\cdot,Y,\xi)\right\rangle_{L_2(\rho_t(\cdot|Y,\xi))} \\ &\quad+\frac{\sigma^2}{2}\mathbb{E}\int\langle\nabla_x\rho_t(x|y,\xi)\,,\,\Theta(x,y,\xi)\rangle dx, \qquad (\text{since } \nabla\log p_t = \tfrac{\nabla p_t}{p_t}) \\ &= \mathbb{E}\left\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(\cdot,\rho_t(\cdot|Y,\xi)),\Theta(\cdot,Y,\xi)\right\rangle_{L_2(\rho_t(\cdot|Y,\xi))} \\ &\quad-\frac{\sigma^2}{2}\mathbb{E}\int(\nabla_x\cdot\Theta(x,Y,\xi))\rho_t(x|Y,\xi)dx, \qquad (\text{integration by parts})\,. \end{aligned} \tag{27}$$

Now we bound the term $\mathbb{E}\int(\nabla_x \cdot \Theta(x,Y,\xi))\rho_t(x|Y,\xi)dx$. From (26) it follows that

$$\mathbb{E}\int(\nabla_x \cdot \Theta(x,Y,\xi))\rho_t(x|y,\xi)dx = \mathbb{E}\int(\nabla_x \cdot \Theta(x,Y,\xi))(\rho_t(x|Y,\xi) - \rho_t(x))dx\,.$$

Since the functions $x \to u(x,y,\xi)$ and $x \to \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)$ are continuously differentiable and L_u-Lipschitz continuous, $\Theta(x,y,\xi)$ is $2L_u$-Lipschitz continuous and continuously differentiable. Thus $|\nabla_x \cdot \Theta(x,y,\xi)| \leq 2dL_u$ uniformly. By noting $\frac{1}{2}\int|\rho_t(x|y,\xi) - \rho_t(x)|dx$ is the total-variation distance between $\rho_t(\cdot|y,\xi)$ and ρ_t, and applying Pinsker's inequality, we get:

$$\begin{aligned}
\left|\mathbb{E}\int(\nabla \cdot \Theta(x,Y,\xi))(\rho_t(x|Y,\xi) - \rho_t(x))dx\right| &\leq \mathbb{E}\int|\nabla \cdot \Theta(x,Y,\xi)| \cdot |\rho_t(x|Y,\xi) - \rho_t(x)|dx \\
&\leq 4dL_u\mathbb{E}[\mathsf{TV}(\rho_t, \rho_t(\cdot|Y,\xi))] \\
&\leq 4dL_u\mathbb{E}\sqrt{\frac{1}{2}\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t)} \\
&= \sqrt{8}dL_u\mathbb{E}\sqrt{\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t)}\,. \qquad (28)
\end{aligned}$$

Since $\rho_t = \mathbb{E}[\rho_t(\cdot \mid Y,\xi)]$, we consider Y', ξ' to be an i.i.d. copy of (Y,ξ) respectively and to be independent of X_0. Since the KL divergence functional is convex jointly in its arguments, we have:

$$\begin{aligned}
\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t) &= \mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\mathbb{E}[\rho_t(\cdot \mid Y',\xi')]) \\
&\leq \mathbb{E}\left[\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t(\cdot \mid Y',\xi'))|Y,\xi\right]\,. \qquad (29)
\end{aligned}$$

Further conditioning on X_0 and taking an expectation yields $\rho_t(\cdot|Y,\xi) = \mathbb{E}[\rho_t(\cdot|X_0,Y,\xi)|Y,\xi]$. Another application of the joint convexity of the KL divergence functional in the above inequality gives

$$\mathbb{E}\left[\mathrm{KL}(\rho_t(\cdot|Y,\xi)\,||\,\rho_t(\cdot|Y',\xi'))|Y,\xi\right] \leq \mathbb{E}\left[\mathrm{KL}(\rho_t(\cdot \mid X_0,Y,\xi)\,||\,\rho_t(\cdot \mid X_0,Y',\xi'))|Y,\xi\right]\,. \qquad (30)$$

Since $X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t$, we have $\rho_t(\cdot|X_0,Y,\xi) = \mathcal{N}(\mu_1, \sigma^2 tI)$, $\rho_t(\cdot|X_0,Y',\xi') = \mathcal{N}(\mu_2, \sigma^2 tI)$, where $\mu_1 := X_0 + tu(X_0,Y,\xi)$, $\mu_2 := X_0 + tu(X_0,Y',\xi')$. Using the formula for KL-divergence between multivariate normal distributions in Duchi (2007) yields:

$$\begin{aligned}
\mathrm{KL}(\rho_t(\cdot \mid X_0,Y,\xi)\,||\,\rho_t(\cdot \mid X_0,Y',\xi')) &= \frac{\|\mu_1 - \mu_2\|^2}{2\sigma^2 t} \\
&= \frac{t}{2\sigma^2}\|u(X_0,Y,\xi) - u(X_0,Y',\xi')\|^2\,. \qquad (31)
\end{aligned}$$

From the bounds in (29), (30), (31), and by applying Jensen's inequality to the outer expectation, we obtain

$$\begin{aligned}
\mathbb{E}\sqrt{\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t)} &\leq \mathbb{E}\sqrt{\frac{t}{2\sigma^2}\mathbb{E}\left[\|u(X_0,Y,\xi) - u(X_0,Y',\xi')\|^2\middle|Y,\xi\right]} \\
&\leq \frac{\sqrt{t}}{\sigma}\sqrt{\frac{1}{2}\mathbb{E}\|u(X_0,Y,\xi) - u(X_0,Y',\xi')\|^2} \\
&\leq \frac{\sqrt{t}}{\sigma}\sqrt{\mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x,Y,\xi))]}\,. \qquad (32)
\end{aligned}$$

Then by plugging the bound in (32) into (28), we get:

$$\left|\mathbb{E}\int(\nabla_x\cdot\Theta(x,Y,\xi))(\rho_t(x|Y,\xi)-\rho_t(x))dx\right| \le \frac{\sqrt{8}\sqrt{t}L_u d}{\sigma}\sqrt{\mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x,Y,\xi))]}\,. \quad (33)$$

Next, we bound the first term in (27). First, note that from (26), we have:

$$\mathbb{E}\int\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\,,\,\Theta(x,Y,\xi)\rangle\rho_t(x)dx = 0\,.$$

Using this fact, we obtain:

$$\begin{aligned}
&\mathbb{E}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_t,\rho_t(\cdot|Y,\xi))\,,\,\Theta(X_t,Y,\xi)\rangle_{L_2(\rho_t(\cdot|Y,\xi))}\\
&=\mathbb{E}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_t,\rho_t(\cdot|Y,\xi))\,,\,\Theta(X_t,Y,\xi)\rangle_{L_2(\rho_t(\cdot|Y,\xi))}-\mathbb{E}\int\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\,,\,\Theta(x,Y,\xi)\rangle\rho_t(x)dx\\
&=\mathbb{E}\int\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t(\cdot|Y,\xi))\,,\,\Theta(x,Y,\xi)\rangle\rho_t(x|Y,\xi)dx\\
&\quad-\mathbb{E}\int\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\,,\,\Theta(x,Y,\xi)\rangle\rho_t(x)dx\,.
\end{aligned}$$

For a given Y,ξ, let $Z_1\sim\rho_t(\cdot|Y,\xi)$ and $Z_2\sim\rho_t$ be optimally coupled in the 2-Wasserstein distance. By the Cauchy-Schwarz inequality we conclude that almost surely Y,ξ:

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t(\cdot|y,\xi))\,,\,\Theta(x,Y,\xi)\rangle-\mathbb{E}_{x\sim\rho_t}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\,,\,\Theta(x,Y,\xi)\rangle\\
&\quad=\mathbb{E}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(Z_1,\rho_t(\cdot|y,\xi))\,,\,\Theta(Z_1,Y,\xi)\rangle-\mathbb{E}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(Z_2,\rho_t)\,,\,\Theta(Z_2,Y,\xi)\rangle\\
&\quad=\mathbb{E}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(Z_1,\rho_t(\cdot|Y,\xi))-\nabla_{\mathcal{W}}\bar{\mathcal{F}}(Z_2,\rho_t)\,,\,\Theta(Z_1,Y,\xi)\rangle\\
&\quad+\mathbb{E}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(Z_2,\rho_t)\,,\,\Theta(Z_1,y,\xi)-\Theta(Z_2,y,\xi)\rangle\\
&\quad\le\mathbb{E}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(Z_1,\rho_t(\cdot|Y,\xi))-\nabla_{\mathcal{W}}\bar{\mathcal{F}}(Z_2,\rho_t)\big\|\cdot\big\|\Theta(Z_1,Y,\xi)\big\|\\
&\quad+\mathbb{E}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(Z_2,\rho_t)\big\|\cdot\big\|\Theta(Z_1,Y,\xi)-\Theta(Z_2,Y,\xi)\big\|\,. \qquad (34)
\end{aligned}$$

Next, note that by the definition of $\Theta(x,y,\xi)$, the assumption that $(x,y)\to u(x,y,\xi)$ and $x\to\nabla_{\mathcal{W}}\mathcal{F}(x,\mu)$ are L_u-Lipschitz, and the assumption $\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu)-\nabla_{\mathcal{W}}\bar{\mathcal{F}}(y,\nu)\|\le L_{\bar{u}}\|x-y\|+L_{\bar{\mathcal{F}}}\mathcal{W}_2(\mu,\nu)$, we obtain:

$$\big\|\Theta(X_1,Y,\xi)-\Theta(X_2,Y,\xi)\big\|\le 2L_u\|X_1-X_2\|\,,$$

$$\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_1,\rho_t(\cdot|Y,\xi))-\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_2,\rho_t)\big\|\le L_{\bar{u}}\|X_1-X_2\|+L_{\bar{\mathcal{F}}}\mathcal{W}_2(\rho_t,\rho_t(\cdot|Y,\xi))\,.$$

Using the above bounds in (34), applying the Cauchy-Schwarz inequality, and the Jensen's inequality yields the following almost surely Y,ξ:

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t(\cdot|Y,\xi))\,,\,\Theta(x,Y,\xi)\rangle-\mathbb{E}_{x\sim\rho_t}\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\,,\,\Theta(x,Y,\xi)\rangle\\
&\quad\le\sqrt{2}(L_{\bar{u}}+L_{\bar{\mathcal{F}}})\mathcal{W}_2(\rho_t,\rho_t(\cdot|Y,\xi))\sqrt{\mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\big\|\Theta(x,Y,\xi)\big\|^2}\\
&\quad+2L_u\mathcal{W}_2(\rho_t,\rho_t(\cdot|Y,\xi))\sqrt{\mathbb{E}_{x\sim\rho_t}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\big\|^2}\,.
\end{aligned}$$

Hence, by taking expectation with respect to Y, ξ and another application of the Cauchy-Schwarz inequality, we obtain:

$$\mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(X_t, \rho_t(\cdot|Y,\xi))\,,\, \Theta(X_t, Y, \xi)\rangle_{L_2(\rho_t(\cdot|Y,\xi))} \\ \leq 2\sqrt{\mathbb{E}[\mathcal{W}_2^2(\rho_t, \rho_t(\cdot|Y,\xi))]}\left[(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\sqrt{\mathbb{E}\big\|\Theta(X_t,Y,\xi)\big\|^2} + L_u\sqrt{\mathbb{E}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_t,\rho_t)\big\|^2}\right]. \tag{35}$$

Finally, by multiplying the bound in (33) by $\sigma^2/2$, adding the resultant to (35), and using (27) we prove (13).

4. Finally, we bound $\mathbb{E}_{y,\xi}[\mathsf{LE}(t)]$. By applying the Cauchy-Schwartz inequality twice and the assumption $\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)\| \leq L_l \mathcal{W}_2(\pi,\mu)$, we have almost surely Y, ξ:

$$\begin{aligned}\mathsf{LE}(t) &= \mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\langle \nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi)), \nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t(\cdot|Y,\xi)) - \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_t(\cdot|Y,\xi)\rangle \\ &\leq \mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\left[\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\| \cdot \|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t(\cdot|Y,\xi)) - \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_t(\cdot|Y,\xi)\|\right] \\ &\leq L_l \mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\left[\mathcal{W}_2(\pi,\rho_t(\cdot|Y,\xi))\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|\right] \\ &\leq L_l\sqrt{\mathcal{W}_2^2(\pi,\rho_t(\cdot|Y,\xi))}\sqrt{\mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2}\,. \end{aligned} \tag{36}$$

Next, by taking an expectation with respect to $(Y,\xi) \sim \rho_0 \times \nu$ and applying the Jensen's inequality, we obtain (14).

■

B.2. Proof of Lemma 6

Before delving into the proof of Lemma 6, we note the following bound on the stochastic error.

Lemma 22 (Stochastic error bound) *Define $V(u,\rho_0) := \mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x,Y,\xi))]$ and $G_{\mathsf{mod}}^2 := \mathbb{E}_{X^*\sim\pi}\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X^*,\pi)\|^2$. Let Assumptions 1, 2 and 3 hold. Assume $L_u t \leq 1$. Then for arbitrary $\beta > 0$, we have:*

$$\begin{aligned}\mathbb{E}[\mathsf{SE}(t)] \lesssim{}& (\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\sqrt{(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}})} + t(L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}) \\ &+ \frac{(\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{\beta} + (tC^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}} + \beta)\bar{\mathcal{E}}(\rho_0) \\ &+ tL_u^2(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))\,.\end{aligned}$$

Proof We define the following to reflect the result of Lemma 4:

$$\begin{aligned}&\Theta(x,y,\xi) := u(x,y,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0) \\ &T_4 := 2\sqrt{\mathbb{E}[\mathcal{W}_2^2(\rho_t,\rho_t(\cdot|Y,\xi))]}\left[(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\sqrt{\mathbb{E}\big\|\Theta(X_t,Y,\xi)\big\|^2} + L_u\sqrt{\mathbb{E}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_t,\rho_t)\big\|^2}\right] \\ &T_5 := \sigma\sqrt{2t}L_u d\sqrt{\mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x,Y,\xi)]}\,.\end{aligned}$$

From Lemma 4, we have $\mathbb{E}[\mathsf{SE}(t)] \leq T_4 + T_5$. We now bound each term of $T_4(t)$. By Lemma 15, we have:

$$\sqrt{\mathbb{E}\mathcal{W}_2^2(\rho_t, \rho_t(\cdot|Y,\xi))} \leq \sqrt{2t^2 V(u,\rho_0)}\,. \tag{37}$$

For the second term, given Y,ξ, let $Z_1 \sim \rho_t(\cdot|Y,\xi)$ and $Z_2 \sim \rho_t$ be optimally coupled in the 2-Wasserstein distance. By Assumption 1, we obtain:

$$\begin{aligned}\big\|\Theta(Z_1,Y,\xi)\big\| &\leq \big\|\Theta(Z_1,Y,\xi) - \Theta(Z_2,y,\xi)\big\| + \big\|\Theta(Z_2,Y,\xi)\big\| \\ &\leq 2L_u\|Z_1 - Z_2\| + \big\|\Theta(Z_2,Y,\xi)\big\|\,.\end{aligned}$$

Next, by squaring, applying the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, taking expectation, applying Lemma 15, we get almost surely Y,ξ:

$$\mathbb{E}[\big\|\Theta(Z_1,Y,\xi)\big\|^2|Y,\xi] \leq 8L_u^2\mathcal{W}_2^2(\rho_0,\rho_t(\cdot|Y,\xi)) + 2\mathbb{E}[\big\|\Theta(Z_2,Y,\xi)\big\|^2|Y,\xi]\,.$$

Applying Lemma 15, Lemma 19 and Assumption 3, after noting that $X_t|Y,\xi \overset{d}{=} Z_1$ we obtain:

$$\begin{aligned}\mathbb{E}\big\|\Theta(X_t,Y,\xi)\big\|^2 &\leq 16L_u^2t^2V(u,\rho_0) + 2\mathbb{E}[\mathbb{E}[\big\|\Theta(Z_2,Y,\xi)\big\|^2|Y,\xi]] \\ &\leq 16L_u^2t^2V(u,\rho_0) + 2C^{\mathsf{Var}}\mathsf{Var}(\rho_0) + 2C_\nu^{\mathsf{Var}} \qquad \text{(By Assumption 3)} \\ &\lesssim C^{\mathsf{Var}}C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}\,. \end{aligned} \tag{38}$$

Here we have used the assumption that $L_ut \leq 1$. The last step above follows by an application of Lemma 19. Now, additionally consider Assumptions 2 and 3. We apply Lemma 18, to conclude:

$$\begin{aligned}\mathbb{E}_{x\sim\rho_t}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\big\|^2 &\leq 3(L_{\bar{\mathcal{F}}}^2 + L_{\bar{u}}^2)\mathcal{W}_2^2(\rho_t,\pi) + 3\mathbb{E}_{x\sim\pi}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\pi)\big\|^2\,. \\ &\leq 3(L_{\bar{\mathcal{F}}}^2 + L_{\bar{u}}^2)\mathbb{E}\mathcal{W}_2^2(\rho_t(\cdot|Y,\xi),\pi) + 3\mathbb{E}_{x\sim\pi}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\pi)\big\|^2\,.\end{aligned}$$

The last step follows from the usual convexity of $\mathcal{W}_2^2$. By applying the Talagrand's T_2-inequality implied by Lemma 17 and Assumption 2-(Log-Sobolev inequality) along with Assumption 2-(KL-Growth), we get:

$$\mathbb{E}_{x\sim\rho_t}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\big\|^2 \lesssim (L_{\bar{\mathcal{F}}}^2 + L_{\bar{u}}^2)C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathbb{E}_{x\sim\pi}\big\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\pi)\big\|^2\,. \tag{39}$$

Let $\zeta(\rho_0) := \sqrt{C^{\mathsf{Var}}C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}}$. We now use equations (39), (38) and (37), along with Lemma 19 to bound $V(u,\rho_0)$ and hence $\mathbb{E}[\mathsf{SE}(t)]$ as:

$$\begin{aligned}\mathbb{E}[\mathsf{SE}(t)] &\lesssim (\sigma L_u d\sqrt{t} + tL_uG_{\mathsf{mod}})\zeta(\rho_0) + t(L_{\bar{u}} + L_{\bar{F}})\zeta^2(\rho_0) \\ &\quad + tL_u(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\zeta(\rho_0)\sqrt{C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))}\,.\end{aligned}$$

Applying AM-GM inequality on $tL_u(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\zeta(\rho_0)\sqrt{C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))}$, we conclude:

$$\begin{aligned}\mathbb{E}[\mathsf{SE}(t)] &\lesssim (\sigma L_u d\sqrt{t} + tL_uG_{\mathsf{mod}})\zeta(\rho_0) + t(L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}) \\ &\quad + tL_u^2(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + tC^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0)\,.\end{aligned}$$

Now, using the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for every $x, y \geq 0$ on $\zeta(\rho_0)$, we conclude:

$$(\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\zeta(\rho_0) \leq (\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\sqrt{(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}})} \\ + (\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\sqrt{C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}} \bar{\mathcal{E}}(\rho_0)}\,.$$

Plugging this into the bound for $\mathbb{E}\mathsf{SE}(t)$ above and applying the AM-GM inequality on $(\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\sqrt{C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}} \bar{\mathcal{E}}(\rho_0)}$, we conclude the result. ■

Proof [Proof of Lemma 6] Let $t \in [0, \eta]$. First, we consider the error terms $\mathsf{DE}_1, \mathsf{DE}_2, \mathsf{SE}, \mathsf{LE}$ in (10) and obtain:

$$\begin{aligned} \frac{d}{dt}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) &= -\int dx \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 + \mathsf{DE}_1(t) + \mathsf{DE}_2(t) + \mathsf{SE}(t) + \mathsf{LE}(t) \\ &= -\frac{1}{2}\int dx \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 + \mathsf{res}(t)\,, \end{aligned} \tag{40}$$

where $\mathsf{res}(t) := \mathsf{DE}_1(t) + \mathsf{DE}_2(t) + \mathsf{SE}(t) + \mathsf{LE}(t) - \frac{1}{2}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx$. We begin by bounding the linearization error $\mathsf{LE}(t)$ using the assumptions stated in this lemma. From (36), the observation that π satisfies Talagrand's T_2-inequality by Lemma 17, Assumption 2-(KL-Growth), and the inequality $ab \leq \frac{a^2}{4} + b^2$ for all $a, b \in \mathbb{R}$, we get:

$$\begin{aligned} \mathsf{LE}(t) &\leq L_l\sqrt{2C_{\mathsf{LSI}}}\sqrt{\mathrm{KL}(\pi,\rho_t(\cdot|Y,\xi))}\sqrt{\mathbb{E}_{\rho_t(\cdot|Y,\xi)}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2} \\ &\leq L_l\sqrt{2C_{\mathrm{KL}}C_{\mathsf{LSI}}}\sqrt{\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))}\sqrt{\mathbb{E}_{\rho_t(\cdot|Y,\xi)}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2} \\ &\leq 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \frac{1}{4}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx \\ &\leq 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \frac{1}{4}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx\,. \end{aligned}$$

By plugging the above inequality into (40) and applying Assumption 2-(Polyak-Łojasiewicz inequality) with $\mu = \rho_t(\cdot|Y,\xi)$, we get:

$$\begin{aligned} \frac{d}{dt}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) &\leq -\frac{1}{2}\int dx \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 + 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_1(t) \\ &\leq -\frac{C_{\bar{\mathcal{E}}}}{2}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_1(t) \\ &\leq -C'\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_1(t)\,, \end{aligned} \tag{41}$$

where

$$\mathsf{res}_1(t) := \mathsf{DE}_1(t) + \mathsf{DE}_2(t) + \mathsf{SE}(t) - \frac{1}{4}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx\,,\ C' := \frac{C_{\bar{\mathcal{E}}}}{2} - 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\,.$$

From Equation (41) we get:

$$\frac{d}{dt}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) \leq -C'\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{SE}(t) + \mathsf{res}_2(t)\,, \tag{42}$$

where

$$\mathsf{res}_2(t) := \mathsf{DE}_1(t) + \mathsf{DE}_2(t) - \frac{1}{4}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx\,.$$

The next step is to bound $\mathbb{E}[\mathsf{SE}(t)]$. To simplify notation, we define:

$$\begin{aligned}A_1^{\mathsf{SE}}(t,\beta) &:= (\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\sqrt{(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}})} + t(L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}) \\ &\quad + \frac{(\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{\beta} + (tC^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}} + \beta)\bar{\mathcal{E}}(\rho_0) \\ A_2^{\mathsf{SE}}(t) &:= tL_u^2(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}\,.\end{aligned}$$

By Lemma 22, we conclude that for arbitrary $\beta > 0$, we have:

$$\mathbb{E}[\mathsf{SE}(t)] \lesssim A_1^{\mathsf{SE}}(t,\beta) + A_2^{\mathsf{SE}}(t)\mathbb{E}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))\,, \tag{43}$$

Using this notation in Equation (42), we obtain:

$$\begin{aligned}\frac{d}{dt}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) &\le (-C' + A_2^{\mathsf{SE}}(t))\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{SE}(t) - A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_2(t) \\ &\le -C''(\eta)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{SE}(t) - A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_2(t)\,,\end{aligned} \tag{44}$$

where $C''(\eta) := \frac{C_{\bar{\mathcal{E}}}}{2} - 2L_l^2 C_{\mathrm{KL}} C_{\mathsf{LSI}} - A_2^{\mathsf{SE}}(\eta)$. Note that by our assumptions on η and the parameter L_l in the statement of this lemma, for a small enough c_0, we must have $C''(\eta) \ge \frac{C_{\bar{\mathcal{E}}}}{4}$.

Next, by multiplying both sides of (44) by $e^{C''(\eta)t}$, we get:

$$\frac{d}{dt}\left[e^{C''(\eta)t}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))\right] \le e^{C''(\eta)t}\left[\mathsf{SE}(t) - A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_2(t)\right]\,.$$

Thus, by integrating on both sides and taking expectation (along with Fubini's theorem), we obtain:

$$\mathbb{E}\bar{\mathcal{E}}(\rho_\eta(\cdot|Y,\xi)) \le e^{-C''(\eta)\eta}\bar{\mathcal{E}}(\rho_0) + e^{-C''(\eta)\eta}\int_0^\eta \mathbb{E}\left[\mathsf{SE}(t) - A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_2(t)\right]e^{C''(\eta)t}dt\,. \tag{45}$$

To simplify notation in the subsequent steps of the proof, we define:

$$\begin{aligned}T_1(t) &:= \sqrt{\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t,\rho_t(\cdot|Y,\xi))\|^2} \\ A(t) &:= L_u\sqrt{2t^2\left(2L_u^2\mathcal{W}_2^2(\rho_0,\pi) + (\sigma^*)^2 + (G_\pi)^2\right) + \sigma^2 td} \\ B(t) &:= L_{\mathcal{F}}\sqrt{2t^2\left(2L_u^2\mathcal{W}_2^2(\rho_0,\pi) + (\sigma^*)^2 + (G_\pi)^2\right) + \sigma^2 td}\,.\end{aligned}$$

Next, by using the inequality $ab \le \frac{a^2}{8} + 2b^2$ for all $a, b \in \mathbb{R}$, the bounds for $\mathsf{DE}_1(t) + \mathsf{DE}_2(t)$ using Lemma 4, we get:

$$
\begin{aligned}
e^{-C''(\eta)\eta}\int_0^\eta \mathbb{E}\left[\mathsf{res}_2(t)\right]e^{C''(\eta)t}dt &= e^{-C''(\eta)\eta}\int_0^\eta e^{C''(\eta)t}\left[\mathbb{E}[\mathsf{DE}_1(t)+\mathsf{DE}_2(t)]-\tfrac{1}{4}\mathbb{E}T_1(t)^2\right]dt \\
&\le e^{-C''(\eta)\eta}\int_0^\eta e^{C''(\eta)t}\left[\mathbb{E}T_1(t)(A(t)+B(t))-\frac{1}{4}\mathbb{E}T_1(t)^2\right]dt \\
&\le 2e^{-C''(\eta)\eta}\int_0^\eta e^{C''(\eta)t}\mathbb{E}(A(t)^2+B(t)^2)dt \\
&\le 2\int_0^\eta \left[\mathbb{E}A(t)^2+\mathbb{E}B(t)^2\right]dt\,. \qquad (46)
\end{aligned}
$$

Next, we bound $\int_0^\eta \left[A(t)^2+B(t)^2\right]dt$. To further simplify notation, we define:

$$
T_2 := 2\left(2L_u^2\mathcal{W}_2^2(\rho_0,\pi)+(\sigma^*)^2+(G_\pi)^2\right)\,,\; T_3 := \sigma^2 d\,.
$$

Therefore, integration yields:

$$
\begin{aligned}
e^{-C''(\eta)\eta}\int_0^\eta \mathbb{E}\left[\mathsf{res}_2(t)\right]e^{C''(\eta)t}dt &\le 2\int_0^\eta \left[A(t)^2+B(t)^2\right]dt \\
&\le 2(L_u^2+L_{\mathcal{F}}^2)\int_0^\eta (T_2t^2+T_3t)dt \\
&= 2(L_u^2+L_{\mathcal{F}}^2)(\tfrac{T_2\eta^3}{3}+\tfrac{T_3\eta^2}{2}) \\
&\lesssim (L_u^2+L_{\mathcal{F}}^2)\left[\eta^3\left[L_u^2\mathcal{W}_2^2(\rho_0,\pi)+(\sigma^*)^2+(G_\pi)^2\right]+\sigma^2d\eta^2\right]\,.
\end{aligned}
$$

Now, we apply Assumptions 2-(KL-Growth) and Assumption 2-(Log-Sobolev inequality) along with Lemma 17 to upper bound $\mathcal{W}_2^2(\rho_0,\pi)$ in the equation above to conclude:

$$
e^{-C''(\eta)\eta}\int_0^\eta \mathbb{E}\left[\mathsf{res}_2(t)\right]e^{C''(\eta)t}dt \lesssim (L_u^2+L_{\mathcal{F}}^2)\left[\eta^3\left[L_u^2C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0)+(\sigma^*)^2+(G_\pi)^2\right]+\sigma^2d\eta^2\right]\,. \qquad (47)
$$

Now we consider:

$$
\begin{aligned}
&e^{-C''(\eta)\eta}\int_0^\eta e^{C''(\eta)t}\mathbb{E}[\mathsf{SE}(t)-A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))]dt \\
&\lesssim e^{-C''(\eta)\eta}\int_0^\eta e^{C''(\eta)t}A_1^{\mathsf{SE}}(t,\beta)dt \qquad \text{(From Equation (43))} \\
&\le \eta A_1^{\mathsf{SE}}(\eta,\beta)\,,
\end{aligned}
$$

where we recall

$$
\begin{aligned}
A_1^{\mathsf{SE}}(\eta,\beta) &= (\sigma L_u d\sqrt{\eta}+\eta L_u G_{\mathsf{mod}})\sqrt{(C^{\mathsf{Var}}\mathsf{Var}\pi+C_\nu^{\mathsf{Var}})}+\eta(L_{\bar{u}}+L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}\pi+C_\nu^{\mathsf{Var}}) \\
&\quad+\frac{(\sigma L_u d\sqrt{\eta}+\eta L_u G_{\mathsf{mod}})^2C^{\mathsf{Var}}C_{\mathsf{LSI}}C_{\mathrm{KL}}}{\beta}+(\eta C^{\mathsf{Var}}(L_{\bar{u}}+L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}+\beta)\bar{\mathcal{E}}(\rho_0)\,. \qquad (48)
\end{aligned}
$$

First, note that since for every $x \geq 0$, we have $1 - x \leq e^{-x} \leq 1 - x + \frac{x^2}{2}$ and as demonstrated above $C''(\eta) \geq \frac{C_{\bar{\mathcal{E}}}}{4}$. With $\beta = c_0 C_{\bar{\mathcal{E}}}$ for some small enough constant c_0 and letting η small enough as noted in the statement of this lemma, we conclude:

$$e^{-C''(\eta)\eta} + (L_u^2 + L_{\mathcal{F}}^2)\eta^3 L_u^2 C_{\mathsf{LSI}} C_{\mathrm{KL}} + (\eta^2 C^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}}) C_{\mathsf{LSI}} C_{\mathrm{KL}} + \eta\beta) \leq e^{-\frac{\eta C_{\bar{\mathcal{E}}}}{8}} .$$

We now plug (47) and (48) along with the above bound into (45) to conclude:

$$\mathbb{E}\bar{\mathcal{E}}(\rho_\eta(\cdot|Y,\xi)) \leq e^{-\frac{\eta C_{\bar{\mathcal{E}}}}{8}} \bar{\mathcal{E}}(\rho_0) + C\gamma_3\eta^3 + C\gamma_2\eta^2 + C\gamma_1\eta^{\frac{3}{2}} .$$

■

Appendix C. Proof for Pairwise Interaction Potential

C.1. Proof of Lemma 9

Proof Note that $\hat{G}$ satisfies the Lipschitz continuity property of Assumption 1 with $L_u = L_V + L_W$ because for every $x_1, x_2, y_1, y_2 \in \mathbb{R}^d$:

$$\|\hat{G}(x_1, y_1) - \hat{G}(x_2, y_2)\| \leq L_V\|x_1 - x_2\| + L_W(\|x_1 - x_2\| + \|y_1 - y_2\|) .$$

Next, the functional $\mathcal{F}$ satisfies the Lipschitz continuity property of Assumption 1, with $L_u = L_V + L_W$ and $L_{\mathcal{F}} = L_W$, because, if $Z_1 \sim \mu$ and $Z_2 \sim \nu$ are optimally coupled in the 1-Wasserstein distance, then by definition of $\nabla_{\mathcal{W}}\mathcal{F}$ in (4), for every $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\begin{aligned}
&\|\nabla_{\mathcal{W}}\mathcal{F}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(y,\nu)\| \\
&\leq \|\nabla V(x) - \nabla V(y)\| + \|\int \nabla W(x-z)\mu(dz) - \int \nabla W(y-z)\nu(dz)\| \\
&\leq L_V\|x-y\| \\
&+ \|\int \nabla W(x-z)\mu(dz) - \int \nabla W(y-z)\mu(dz) + \int \nabla W(y-z)\mu(dz) - \int \nabla W(y-z)\nu(dz)\| \\
&= (L_V + L_W)\|x-y\| + \mathbb{E}_{(Z_1,Z_2)\sim\mu\times\nu}\|\nabla W(y-Z_1) - \nabla W(y-Z_2)\| \\
&= (L_V + L_W)\|x-y\| + L_W\mathcal{W}_1(\mu,\nu) \\
&\leq (L_V + L_W)\|x-y\| + L_W\mathcal{W}_2(\mu,\nu) .
\end{aligned}$$

Similarly, we consider $\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu) = \nabla V(x) + \nabla W * \pi(x)$ and conclude $L_{\bar{u}} = L_V + L_W$, $L_{\bar{\mathcal{F}}} = 0$ and $L_l = L_W$. Clearly, Assumption 2-(Log-Sobolev inequality) is satisfied with constant C_{LSI} since it is the same as Assumption 5. Notice that $\bar{\mathcal{E}}(\mu) := \frac{\sigma^2}{2}\mathrm{KL}(\mu||\pi)$, and that $\nabla_{\mathcal{W}}\mathrm{KL}(\mu||\pi) = \mathsf{FD}(\mu||\pi)$, we conclude that Assumption 5 implies Assumption 2-(Polyak-Łojasiewicz inequality) with $C_{\bar{\mathcal{E}}} = \frac{\sigma^2}{C_{\mathsf{LSI}}}$. The choice of $\bar{\mathcal{E}}$ also implies Assumption 2-(KL-Growth) with $C_{\mathrm{KL}} = \frac{2}{\sigma^2}$.

Now consider Assumption 3. Note that by Jensen's inequality:

$$\begin{aligned}
\mathbb{E}_{Y\sim\rho_0}\|u(x,Y) + \nabla_{\mathcal{W}}\mathcal{F}(x;\rho_0)\|^2 &= \mathbb{E}_{Y\sim\rho_0}\|\nabla W(x-Y) - \nabla W * \rho_0(x)\|^2 \\
&\leq \mathbb{E}_{Y\sim\rho_0}\int \|\nabla W(x-Y) - \nabla W(x-z)\|^2\rho_0(dz) \\
&= L_W^2\mathbb{E}_{(Y,Z)\sim\rho_0\times\rho_0}\|Y-Z\|^2 \\
&= 2L_W^2\mathsf{Var}(\rho_0) .
\end{aligned}$$

Hence, Assumption 3 is satisfied with $C^{\mathsf{Var}} = 2L_W^2$ and $C_\nu^{\mathsf{Var}} = 0$. ■

C.2. Proof of Theorem 10

Proof Using the notation in Theorem 7, we conclude that $\bar{\mathcal{F}}(x,\pi) = \mathcal{F}(x,\pi)$ for every $x \in \mathbb{R}^d$ and hence $G_\pi = G_{\mathsf{mod}}$. Since $\pi \propto \exp(-\frac{2\delta\mathcal{F}(x,\pi)}{\sigma^2})$ and $\nabla_{\mathcal{W}}\mathcal{F}(x,\pi) = \nabla_x\delta\mathcal{F}(x,\pi)$, we apply Lemma 16 to conclude that $(G_\pi)^2 := \mathbb{E}_{x\sim\pi}\|\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)\|^2 \leq \frac{d\sigma^2(L_V+L_W)}{2}$. Applying Lemma 20, we conclude $\mathsf{Var}(\pi) \leq C_{\mathsf{LSI}}d$. Using Assumption 3 instantiated to our case, we conclude $(\sigma^*)^2 \leq 2L_W^2 C_{\mathsf{LSI}}d$. Under the assumption on η, all the requisite assumptions for Theorem 7 are satisfied (after simplifying with the fact that $C_{\mathsf{LSI}} \geq \frac{\sigma^2}{2(L_V+L_W)}$ from Lemma 21). Thus, we note that $\gamma_1, \gamma_2, \gamma_3$ in Theorem 7 can be instantiated as

$$\gamma_1 \lesssim \sigma d^{\frac{3}{2}}\sqrt{C_{\mathsf{LSI}}}L_W(L_V+L_W)\,;\; \gamma_2 \lesssim (L_V+L_W)^2\sigma^2 d^2\,;\; \gamma_3 \lesssim (L_V+L_W)^3 d\sigma^2\,.$$

Here, we have Assumption 6 that $L_W \leq \frac{\sigma^2}{4C_{\mathsf{LSI}}}$ to simplify the expressions. Invoking Theorem 7, using the fact that $\eta(L_V+L_W) < c_0$ by assumption in the statement of this theorem, and $C_{\mathsf{LSI}} \geq \frac{\sigma^2}{2(L_V+L_W)}$ from Lemma 21, we conclude the result. ■

Appendix D. Proof for Mean Field Neural Networks

D.1. Proof of Lemma 12

Proof Note that $u = \hat{G}$ satisfies the Lipschitz continuity property of $x \to u(x,y,\xi)$ and $y \to u(x,y,\xi)$ in Assumption 1 with $L_u = (B+R)(LR+N)+\lambda+M^2R^2$ because for every $x_1, x_2 \in \mathbb{R}^d$:

$$\begin{aligned}
\|u(x_1,y,i) - u(x_2,y,i)\| &\leq (h(z_i,Y)-w_i)\|\nabla_x h(z_i,x_1) - \nabla_x h(z_i,x_2)\| + \lambda\|x_1-x_2\| \\
&\leq (h(z_i,y)-w_i)(L\|z_i\|)\|x_1-x_2\| + \lambda\|x_1-x_2\| \\
&\leq ((B+R)LR+\lambda)\|x_1-x_2\|\,,
\end{aligned}$$

where the last two inequalities follow from Assumptions 7 and 8. Similarly considering $\|u(x,y_1,i) - u(x,y_2,i)\|$, we conclude the result. Similarly, we can take $L_{\mathcal{F}} = M^2R^2$ because, for every $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\begin{aligned}
&\|\nabla_{\mathcal{W}}\mathcal{F}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\nu)\| \\
&= \left\|\frac{1}{m}\sum_{i=1}^m \left[\left(\int h(z_i,w)d\mu(w) - \int h(z_i,w)d\mu(v)\right)\nabla_x h(z_i,x)\right]\right\| \\
&\leq \frac{1}{m}\sum_{i=1}^m \|\nabla_x h(z_i,x)\|M\|z_i\|\mathcal{W}_1(\mu,\nu) \\
&\leq M^2R^2\mathcal{W}_2(\mu,\nu)\,.
\end{aligned}$$

Next, since the proximal Gibbs measure satisfies the LSI by Assumption 9, the LSI condition in Assumption 2 is satisfied with $C_{\mathsf{LSI}} = C_{\mathsf{LSI}}$. By (Nitanda et al., 2022, Proposition 1), we conclude that $C_{\mathrm{KL}} = \frac{2}{\sigma^2}$. Again, by (Nitanda et al., 2022, Proposition 1), we have: $\nabla_{\mathcal{W}} \bar{\mathcal{E}}(x, \mu) = \frac{\sigma^2}{2} \nabla_x \log(\frac{\mu(x)}{\pi_\mu(x)})$. Therefore, have:

$$
\begin{aligned}
\int \|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(x, \mu)\|^2 d\mu(x) &= \frac{\sigma^4}{4} \mathsf{FD}(\mu || \pi_\mu) \\
&\geq \frac{\sigma^4}{2C_{\mathsf{LSI}}} \mathrm{KL}(\mu || \pi_\mu) && \text{(By Assumption 9)} \\
&\geq \frac{\sigma^2}{C_{\mathsf{LSI}}} \bar{\mathcal{E}}(\mu)\,. && \text{(By (Nitanda et al., 2022, Proposition 1))}
\end{aligned}
$$

Therefore, we conclude that Assumption 2-(Polyak-Łojasiewicz inequality) holds with $C_{\bar{\mathcal{E}}} = \frac{\sigma^2}{C_{\mathsf{LSI}}}$. Let $Y, I \sim \rho_0 \times \nu$ and let Y', I' be an independent copy of Y, I. By Assumptions 7, 8 we have $\|u(x, y, i) + \nabla_{\mathcal{W}} \mathcal{F}(x, \rho_0)\| \leq 2(B + R)MR$ almost surely. Thus, the mean-field neural network case satisfies Assumption 3 with $C^{\mathsf{Var}} = 0$ and $C_\nu^{\mathsf{Var}} = 4(B + R)^2 M^2 R^2$.

■

D.2. Proof of Theorem 13

Proof Under the parameter correspondence established in Lemma 12, with our choice of η, we conclude that the conditions for Theorem 7 are satisfied (once considered with the fact that $C_{\mathsf{LSI}} \geq \frac{\sigma^2}{2L_u}$ from Lemma 21). Since $\pi \propto \exp(-\frac{2\delta\mathcal{F}(x,\pi)}{\sigma^2})$ and $\nabla_{\mathcal{W}} \mathcal{F}(x, \pi) = \nabla_x \delta\mathcal{F}(x, \pi)$, we apply Lemma 16 to conclude that $(G_\pi)^2 = G_{\mathsf{mod}}^2 := \mathbb{E}_{x \sim \pi} \|\nabla_{\mathcal{W}} \mathcal{F}(x, \pi)\|^2 \leq \frac{d\sigma^2 L_u}{2}$. Using Assumption 3 instantiated to our case, we conclude $(\sigma^*)^2 \leq 4M^2 R^2 (B + R)^2$.

Instantiating the quantities γ_1, γ_2 and γ_3 found in Theorem 7 to the case of mean field neural networks, we have:

$$
\gamma_1 \lesssim \sigma d L_u M R (B + R)\,,\ \gamma_2 \lesssim \sigma^2 L_u^2 d + L_u M^2 R^2 (B + R)^2\,,\ \gamma_3 \lesssim \sigma^2 L_u^3 d + L_u^2 M^2 R^2 (B + R)^2\,.
$$

We then apply Theorem 7 and simplify using the fact that $\eta L_u \leq c_0$, and $C_{\mathsf{LSI}} \geq \frac{\sigma^2}{2L_u}$ from Lemma 21 to conclude the result.

■