UNIFORM-IN-TIME PROPAGATION OF CHAOS FOR THE MEAN-FIELD GRADIENT LANGEVIN DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

The mean-field Langevin dynamics is characterized by a stochastic differential equation that arises from (noisy) gradient descent on an infinite-width two-layer neural network, which can be viewed as an interacting particle system. In this work, we establish a quantitative weak propagation of chaos result for the system, with a finite-particle discretization error of $\mathcal{O}(1/N)$ *uniformly over time*, where N is the width of the neural network. This allows us to directly transfer the optimization guarantee for infinite-width networks to practical finite-width models without excessive overparameterization. On the technical side, our analysis differs from most existing studies on similar mean field dynamics in that we do not require the interaction between particles to be sufficiently weak to obtain a uniform propagation of chaos, because such assumptions may not be satisfied in neural network optimization. Instead, we make use of a logarithmic Sobolev-type condition which can be verified in appropriate regularized risk minimization settings.

1 INTRODUCTION

Mean-field neural networks. We consider the optimization of a two-layer neural network in the *mean-field regime* (Mei et al., 2018), which is represented as an average over N neurons:

$$f_{\boldsymbol{X}}(z) = \frac{1}{N}\sum_{i=1}^{N} h_z(x_i),$$

where given the input $z \in \mathbb{R}^d$, each neuron computes a nonlinear transformation based on trainable parameters $x \in \mathbb{R}^p$; for example, we may set $h_z(x) = \tanh(r\sigma(w^\top z))$ for $x = (r, w) \in \mathbb{R} \times \mathbb{R}^d$.

Importantly, the mean-field parameterization allows for the parameters to move away from initialization during gradient descent and hence learn informative features (Yang and Hu, 2020) even when the network width is large ($N \to \infty$), in contrast to the *Neural Tangent Kernel* (NTK) parameterization (Jacot et al., 2018) (corresponding to a $1/\sqrt{N}$ prefactor), which freezes the model at initialization under overparameterization. This *feature learning* ability enables mean-field neural networks to outperform the NTK counterpart (or linear estimators in general) in learning a wide range of target functions (Ghorbani et al., 2019; Li et al., 2020; Abbe et al., 2022; Ba et al., 2022).

Optimization guarantees of gradient-based methods for mean-field neural networks are typically obtained by lifting the finite-width network to the infinite-dimensional space of parameter distributions and then exploiting the convexity of the objective function. Using this viewpoint, convergence of gradient flow on infinite-width networks to the global optimal solution can be shown under appropriate conditions (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020). However, most existing results are *qualitative* in nature, in that they do not characterize the rate of convergence and the finite-particle discretization error.

Mean-field Langevin dynamics. An often-studied optimization method for mean-field neural networks is the noisy particle gradient descent (NPGD) algorithm (Mei et al., 2018; Hu et al., 2019; Chen et al., 2020b), where Gaussian noise is injected to the gradient to encourage "exploration" and enable global optimality to be shown under milder conditions than the noiseless case. The large particle and vanishing step size limit is termed the *mean-field Langevin dynamics* (Hu et al., 2019), which globally minimizes an entropy-regularized convex functional in the space of measures.

Recently, Nitanda et al. (2022); Chizat (2022) established exponential convergence for the mean-field Langevin dynamics under certain logarithmic Sobolev inequalities which can be easily verified in regularized risk minimization problems using two-layer neural networks (1). This represents a significant step towards a *quantitative* optimization analysis of neural networks in the presence of feature learning, yet the limitation is also clear: these results are obtained from directly analyzing the large particle limit (i.e., the McKean-Vlasov stochastic differential equation), and cannot be easily transferred to practical finite-width networks. In fact, applying the quantitative bounds from Mei et al. (2018; 2019) leads to discretization error bounds that blow up exponentially in time, rendering the guarantee vacuous beyond the very early stages of gradient descent learning. Therefore, it is important to derive a finite-particle discretization error bound that holds *uniformly over time*, that is, the error remains stable even when t is large.

1.1 Our contributions

In this paper, we establish finite-particle guarantees for the mean-field Langevin dynamics via a *propagation of chaos* calculation (Sznitman, 1991) which controls the weak error between the empirical distribution of the interacting particle system and the corresponding infinite particle limit along the optimization trajectory. This allows us to bound the difference in the function value between the finite-width neural network optimized by NPGD and the infinite-width counterpart. In particular, starting from the initialized particles $\boldsymbol{X}_0 \overset{\text{i.i.d}}{\sim} \mu_0$, if we denote the finite-particle model at time t of optimization as $f_{\boldsymbol{X}_t}$, and its corresponding infinite-particle limit (starting from μ_0) as f_{μ_t}, then our main result is the following.

Theorem (informal)**.** *Under suitable regularity conditions,* $\mathbb{E}\big[\,(f_{\boldsymbol{X}_t} - f_{\mu_t})^2\,\big] = \mathcal{O}\,(1/N)$ *for any* $t \geq 0$.

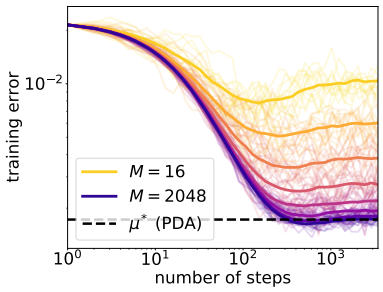

Figure 1: Training error of two-layer NNs optimized by NPGD. Solid curve: mean over 100 runs. Translucent curve: individual runs. Dashed black line: global optimum approximated by the PDA algorithm (Nitanda et al., 2020).

We make the following remarks on the main theorem:

- To our knowledge, we provide the first rigorous *uniform-in-time* propagation of chaos result in the context of mean-field neural networks. This is in contrast to prior works in which the discretization error typically increases as optimization proceeds (cf. Mei et al. (2018, Theorem 3)). The theorem implies that as the width N becomes larger, the deviation of the network output (and hence the training and test error) from the infinite-width model diminishes rapidly, as shown in Figure 1.
- Our analysis assumes a modified Log-Sobolev condition which is satisfied in regularized risk minimization problems using neural network when the convex regularizer on parameters has super-quadratic tail. Noticeably, we do not impose any constraint on the strength of regularization and interaction; this differs from many existing results where uniform propagation of chaos is only achieved under weak interaction or large noise (Eberle et al., 2019; Delarue and Tse, 2021).

Due to the space constraint, we defer discussions on additional related works to Appendix A.

2 Preliminaries

In this section, we formulate the problem setting and introduce some useful notations for the following sections. We optimize a two-layer neural network by minimizing the empirical or expected risk in a supervised learning setting, where the input is included in a set $\mathcal{Z} \subset \mathbb{R}^d$ and the output is in a bounded set $\mathcal{Y} \subset \mathbb{R}$. As defined in the Introduction, $h_{(\cdot)}(x) : z \in \mathcal{Z} \mapsto h_z(x) \in \mathcal{Y}$ represents one neuron (particle) with parameters $x \in \mathbb{R}^p$, and the mean-field neural network is written as the average over M neurons: $f_{\boldsymbol{X}}(z) = \frac{1}{N}\sum_{i=1}^N h_z(x_i)$, where $\boldsymbol{X} = (x_i)_{i=1}^N \subset \mathbb{R}^d$ denotes the collection of parameters and $z \in \mathcal{Z}$. The continuous limit of the neural network is obtained by taking $N \to \infty$, and in analogy to the law of large numbers, $f_{\boldsymbol{X}}$ converges to the following integral form:

$$f_\mu(z) = \int h_z(x)\mathrm{d}\mu(x),$$

where μ is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ representing the weight of each parameter. Let $\mathcal{P}$ be the set of probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, and $\mathcal{P}_p$ be those with a finite p-th moment $(p \geq 1)$. As typical for the mean-field analysis, we aim to optimize the density function μ so that the neural network f_μ accurately predicts the output $y \in \mathcal{Y}$ from the input $z \in \mathcal{Z}$.

In the following, we take (regularized) empirical risk minimization as a concrete example, and we note that the exact same analysis applies to the minimization of expected risk. Let $\ell(z, y) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a convex loss function, such as the squared loss $\ell(z, y) = (z - y)^2/2$ for regression, or the logistic loss $\ell(z, y) = \log(1 + \exp(-yz))$ for classification. For each (z_i, y_i) in the given training data $(z_i, y_i)_{i=1}^n \subset \mathcal{Z} \times \mathcal{Y}$, we use the notation $h_i(x)$ and $\ell_i(f)$ to indicate $h_{z_i}(x)$ and $\ell(f(z_i), y_i)$ respectively. Our goal is to find an *approximate* minimizer of the following objective over $\mathcal{P}$:

$$F(\mu) := \frac{1}{n} \sum_{i=1}^{n} \ell_i(f_\mu) + \lambda_1 \int r(x) \mathrm{d}\mu(x), \tag{1}$$

where $\lambda_1 > 0$ is the regularization strength and $r(\cdot)$ is a convex regularizer. More specifically, we will analyze the mean-field Langevin dynamics which solves an *entropy-regularized* version of (1). It is worth noting that this entropy-regularized objective can also be globally optimized by the recently proposed particle gradient-type methods in Nitanda et al. (2020); Oko et al. (2022), for which finite-width convergence rates have been provided. However, those methods employ an intricate double-loop structure which does not mirror the commonly-used gradient descent algorithm. Therefore, an important question to be addressed is whether the vanilla noisy gradient descent algorithm on a finite-width neural network also enjoys similar convergence guarantee – this is precisely the motivation of the current paper.

3 Mean-field gradient Langevin dynamics

Derivation of the continuous dynamics. The basic idea of the mean-field Langevin dynamics is to optimize the aforementioned objective via *Wasserstein gradient flow* over a set of measures $\mathcal{P}$. To define the gradient with respect to the measure, we introduce the *first-variation* $\frac{\delta G}{\delta \mu}$ of a functional $G : \mathcal{P}_2 \to \mathbb{R}$ at $\mu \in \mathcal{P}_q$ (for a given $q \geq 1$) as a continuous functional $\mathcal{P}_q \times \mathbb{R}^d \to \mathbb{R}$ that satisfies

$$\lim_{\epsilon \to 0} \frac{G(\epsilon\nu + (1-\epsilon)\mu)}{\epsilon} = \int \frac{\delta G}{\delta \mu}(\mu)(x) \mathrm{d}(\nu - \mu),$$

for any $\nu \in \mathcal{P}_q$. If there exists such a functional $\frac{\delta G}{\delta \mu}$, we say G admits a first-variation at μ, or simply G is differentiable at μ. To avoid the ambiguity of $\frac{\delta G}{\delta \mu}$ up to constant shift, we follow the convention of imposing $\int \frac{\delta G}{\delta \mu}(\mu) \mathrm{d}\mu = 0$. In our setting, the first-variation of the objective F is given by

$$\frac{\delta F}{\delta \mu}(\mu)(x) = \frac{1}{n} \sum_{j=1}^{n} \ell_j'(f_\mu) h_j(x) + \lambda_1 r(x).$$

In our analysis, we need to trace the behavior of $F(\mu_t)$ along a trajectory of measure $(\mu_t)_t$in $\mathcal{P}_2$ following a *continuity equation*:

$$\partial_t \mu_t = \nabla \cdot (\mu_t v_t),$$

where $v_t : \mathbb{R}^d \to \mathbb{R}^d$ is a vector field included in $L^2(\mu_t)$, and the time-derivative and the divergence operator are defined in a *weak sense*, that is, for any continuously differentiable function ϕ with a compact support, $\int \phi \mathrm{d}\mu_t - \int \phi \mathrm{d}\mu_s = -\int_s^t \int \nabla\phi \cdot v_\tau \mathrm{d}\mu_\tau \mathrm{d}\tau$. Then, the time-derivative of $G(\mu_t)$ along the trajectory can be written as

$$\partial_t G(\mu_t) = \int v_t \cdot \nabla \tfrac{\delta}{\delta \mu} G(\mu_t) \mathrm{d}\mu_t. \tag{2}$$

We refer readers to Villani (2009); Ambrosio et al. (2005); Bakry et al. (2014) for more details. In this sense, $\nabla \frac{\delta}{\delta \mu} F(\mu_t)$ can be seen as a *gradient direction* in the measure space (endowed with a Wasserstein metric). The mean-field Langevin dynamics approximately minimizes the objective F based on the Wasserstein gradient flow. Specifically, define the nonlinear drift term:

$$b(x, \mu) = \nabla \frac{\delta F}{\delta \mu}(\mu)(x) = \frac{1}{n} \sum_{j=1}^{n} \ell_j'(f_\mu) \nabla h_j(x) + \lambda_1 \nabla r(x),$$

the *mean-field Langevin dynamics* is given by the following stochastic differential equation:

$$\mathrm{d}X_t = -b(X_t, \mu_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t, \tag{3a}$$

$$\mu_t = \mathrm{Law}(X_t), \tag{3b}$$

for $X_0 \sim \mu_0$, where $\mathrm{Law}(X)$ denotes the distribution (probability law) of the random variable X, and $(W_t)_{t\geq 0}$ is the d-dimensional standard Brownian motion. The existence and uniqueness of the solution are ensured by Theorem 3.3 of Huang et al. (2021) (see also Corollary 3 for more details). It is known that the law $m(t, \mu_0) = \mu_t$ in the mean-field Langevin dynamics satisfies the following *nonlinear* Fokker-Planck equation:

$$\partial_t m(t, \mu_0) = \lambda \Delta m(t, \mu_0) + \nabla \cdot [m(t, \mu_0) b(\cdot, m(t, \mu_0))],$$

with $m(0, \mu_0) = \mu_0$ (this is again defined in a weak sense, that is, $\int \phi \mathrm{d}(m(t, \mu_0) - m(s, \mu_0)) = \int_s^t \int (\lambda \Delta \phi - b(\cdot, m(\tau, \mu_0))^\top \nabla \phi) \mathrm{d}m(\tau, \mu_0)\mathrm{d}\tau$ for smooth test function f with compact support). This dynamics is an example of distribution depending SDEs originating from the study of interacting particle systems which dates back to 1950s (Kahn and Harris, 1951; Kac, 1956; 1959; McKean, 1966; 1967). A fundamental characterization of the mean-field Langevin dynamics is that it is a Wasserstein gradient flow that minimizes the following objective (Mei et al., 2018; Hu et al., 2019):

$$\mathcal{L}(\mu) = F(\mu) + \lambda \mathrm{Ent}(\mu), \tag{4}$$

where $\mathrm{Ent}(\mu) = -\int \log(\mathrm{d}\mu(z)/\mathrm{d}z)\mathrm{d}\mu(z)$ is the negative entropy of μ. Indeed, it is known that $\nabla \frac{\delta \mathcal{L}(\mu)}{\delta \mu} = \nabla \frac{\delta F}{\delta \mu} + \lambda \nabla \log(\mu) = \lambda \nabla \log(\mu) + b(\cdot, \mu_t)$ (e.g, Theorem 4.16 of Ambrosio et al. (2005)) and the continuity equation corresponding to μ_t can be rewritten as $\partial_t \mu_t = \nabla \cdot [(\lambda \nabla \log(\mu_t) + b(\cdot, \mu_t))\mu_t] = \nabla \cdot (\nabla \frac{\delta \mathcal{L}(\mu)}{\delta \mu} \mu_t)$, which, in combination with the identity (2), yields that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) = -\int \left\| \nabla \frac{\delta \mathcal{L}(\mu)}{\delta \mu} \right\|^2 \mathrm{d}\mu_t.$$

We therefore see that μ_t decreases $\mathcal{L}(\mu_t)$ unless $\frac{\delta \mathcal{L}(\mu)}{\delta \mu} \neq 0$, which is the crucial property that guarantees the convergence of μ_t to the global optimal solution (see Lemma 1). Remarkably, this can be seen as a nonlinear extension of the usual gradient Langevin dynamics, where $F(\mu)$ is a linear functional in the form of $F(\mu) = \int L(x)\mathrm{d}\mu$ with some objective function L.

It is easy to see that the objective $\mathcal{L}$ can be reformulated as the following objective (up to constant) that employs the KL divergence from a distribution characterized by the regularization term r:

$$\mathcal{L}(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\mu) + \lambda \mathrm{KL}(\mu, \nu_r), \tag{5}$$

where ν_r is a distribution with density proportional to $\exp(-\lambda_1 r/\lambda)$ and $\mathrm{KL}(\cdot, \cdot)$ is the KL divergence (relative entropy) defined as $\mathrm{KL}(\mu, \nu) := \int \log(\mathrm{d}\mu/\mathrm{d}\nu)\mathrm{d}\mu$.

Particle discretization. One of the main difficulties to simulate the mean-field Langevin dynamics is that we cannot access the exact information of μ_t in the practical setting. Instead, we approximate this infinite-dimensional objective by a finite set of particles, which yields the following SDE:

$$\mathrm{d}\hat{X}_t^i = -b(\hat{X}_t^i, \mu_t^N)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t^i, \quad \mu_t^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{X}_t^i},$$

and $\hat{X}_0^i \overset{\text{i.i.d.}}{\sim} \mu_0$. Our goal is to analyze the finite-particle approximation error due to replacing μ_t with μ_t^N. The main mechanism of this approximation is the *propagation of chaos* (Sznitman, 1991), which roughly refers to the phenomenon that as the number of particles $N \to \infty$, the interaction between particles vanishes and $\mu_t^N \to \mu_t$. However, it is far from trivial to establish a *quantitative* estimate of such approximation that provides meaningful guarantee for the finite-particle system. In the following sections, we will show that under appropriate conditions, the finite-particle error can be (weakly) controlled uniformly over time, which allows us to prove optimization guarantees for finite-width neural networks that are not excessively overparameterized.

4 CONVERGENCE ANALYSIS

In this section we present our main theoretical result – the quantitative propagation of chaos. First, we introduce the main assumption in our analysis.

Assumption 1. *We assume $h_i, \ell_i, r \in \mathcal{C}^\infty$ and satisfy the following conditions:*

1. ***Convexity of loss:*** *ℓ_i is a convex function.*
2. ***Boundedness and smoothness:*** *There exists $B > 0$ such that $\|h_i\|_\infty \leq B$, $\|\nabla h_i\|_\infty \leq B$, $\|\nabla\nabla^\top h_i\|_\infty \leq B$, $\max\{|\ell_j(f_\mu)|, |\ell_j'(f_\mu)|, |\ell_j''(f_\mu)|\} \leq B$ uniformly over $\mathcal{P}$.*
3. ***Regularity of r:*** *The regularization term r is a convex function satisfying $c_r\|x\|^{2+\delta} \leq r(x) \leq C_r(1+\|x\|^{2+\delta})$, $\nabla r(x)\cdot x \geq c_r\|x\|^{2+\delta}$ and $0 \preceq \nabla\nabla^\top r(x) \preceq C_r(1+\|x\|^\delta)I$ for constants $0 < \delta$ and $c_r, C_r > 0$.*

We make the following remarks on the main assumption.

- The loss convexity is a standard assumption to ensure that the objective $\mathcal{L}$ is convex with respect to μ (note that this does not imply convexity with respect to the parameters $(x_i)_{i=1}^N$ of the network).
- The second assumption is satisfied for standard two-layer models under the following conditions: (i) $\|z\| \leq C$; (ii) smooth loss function, such as the squared loss and logistic loss. For example, $h_z(x) = \tanh(r)\sigma(w^\top z)$ with smooth activation function σ and $x = (r, w)$, or $h_z(x) = \sigma(w^\top z)$ with smooth and bounded activation and $x = w$.
- The constraint on $r(\cdot)$ requires the regularization term to have a super-quadratic tail, which is satisfied, for example, by $r(x) = \|x\|^4$. While this does not cover the standard weight decay, we note that regularizers with stronger tail growth have been employed in the theoretical analysis of neural networks (Chen et al., 2020a; Allen-Zhu and Li, 2022). The purpose of this assumption is to ensure good isoperimetry of μ_t along the trajectory.
- The infinite differentiability condition is imposed only for the simplicity of the analysis.

Proximal Gibbs distribution. An important quantity in the convergence analysis is the *proximal Gibbs distribution*: for $\mu \in \mathcal{P}$, we define the proximal density function as

$$p_\mu(x) = \frac{1}{Z(\mu)}\exp\left(-\frac{1}{\lambda}\frac{\delta F(\mu)}{\delta\mu}(x)\right),$$

where $Z(\mu)$ is a normalization constant. One may check that this corresponds to the minimizer of the *linearized potential*: $\min_{\nu\in\mathcal{P}}\int\frac{\delta F(\mu)}{\delta\mu}\mathrm{d}\nu + \lambda\mathrm{Ent}(\nu)$. For μ_t, we denote by $\tilde{\mu}_t$ its proximal Gibbs measure, that is, the probability measure with the density p_{μ_t}. Then, we have the following characterization of the minimizer of $\mathcal{L}$.

Proposition 1. *Under Assumption 1, the functional $\mathcal{L}$ has a unique minimizer in $\mathcal{P}$ that is absolutely continuous with respect to the Lebesgue measure. Moreover, $\mu^* \in \mathcal{P}_{2+\delta}$ is the optimal solution if and only if μ^* is absolutely continuous and its density function satisfies $\frac{\mathrm{d}\mu^*}{\mathrm{d}x} = p_{\mu^*}$.*

This proposition can be shown in the same manner as Proposition 2.5 of Hu et al. (2019). We remark that although this prior result assumed r to have at most quadratic growth, its proof does not require such growth condition but requires only the integrability of ν_r and $\nu_r(x)\log(\nu_r(x))$.

Many convergence properties of the mean-field Langevin dynamics can be characterized by properties of μ^* and $\tilde{\mu}_t$. We first introduce the *logarithmic Sobolev inequality* (LSI) which will be very useful in the subsequent analysis.

Definition 1 (Logarithmic Sobolev inequality)**.** *Let $p(\theta)$ be a smooth probability density function on $\mathbb{R}^d$. $p(\theta)$ (or its corresponding probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}^d))$) satisfies the LSI with constant $\alpha' > 0$ if and only if, for any smooth function $\phi : \mathbb{R}^p \to \mathbb{R}$ with $\mathbb{E}_p[\|\phi\|^2] < \infty$, it holds that*

$$\mathbb{E}_p[\phi^2\log(\phi^2)] - \mathbb{E}_p[\phi^2]\log(\mathbb{E}_p[\phi^2]) \leq \frac{2}{\alpha'}\mathbb{E}_p[\|\nabla\phi\|_2^2].$$

We can verify that in our setting, the proximal Gibbs measure satisfies the LSI condition.

Proposition 2. *Under Assumption 1, p_μ satisfies the log-Sobolev inequality with a constant α that depends on $d, c_r, B, \lambda, \delta$. If additionally $\nabla\nabla^\top r \succeq I$, then the LSI holds with $\alpha = \frac{2\lambda_1}{\lambda}\exp\left(-\frac{4B}{\lambda}\right)$.*

The proof is given in Corollary 5 in the Appendix. Note what similar characterization was obtained in Nitanda et al. (2022); Chizat (2022) under quadratic regularizer $r(x) = \|x\|^2$, via the standard Bakry–Emery and Holley–Stroock arguments (Bakry and Émery, 1985; Holley and Stroock, 1987) (see also Corollary 5.7.2 and 5.1.7 of Bakry et al. (2014)). However, our Assumption 1 does not entail $\nabla\nabla^\top r \succeq I$ and thus our proof follows a different strategy. This LSI condition is crucial to the geometric ergodicity of the mean-field Langevin dynamics described in Theorem 1 below.

4.1 Uniform propagation of chaos for finite-width neural networks

To present our *weak* propagation of chaos result, we first introduce a test function in a form of

$$\mathcal{U}(t, \mu) := \Phi(m(t, \mu)),$$

where $\Phi : \mathcal{P} \to \mathbb{R}$ is assumed to be sufficiently smooth, that is, Φ is twice differentiable with respect to μ and x, and the derivatives are bounded as $\sup_{x_1,\dots,x_k \in \mathbb{R}^d} |\partial_{x_1}^{j_1} \dots \partial_{x_k}^{j_k} \frac{\delta^k \Phi(\mu)}{\delta \mu^k}(x_1, \dots, x_k)| < C$ for $k = 0, 1, 2$ and $j_i = 0, 1, 2$ uniformly over all $\mu \in \mathcal{P}$ with some constant C (see Delarue and Tse (2021) for a related discussion).

Example 1. *Under Assumption 1, we allow for the following test functions.*

(i) ***Neural network function value****:* $\Phi(\mu) = \int h_z(x) \mathrm{d}\mu(x)$ *with a fixed* $z \in \mathcal{Z}$.

(ii) ***Training and test error****:* $\Phi(\mu) = \mathbb{E}_{(Z,Y)\sim P}[\ell(f_\mu(Z), Y)]$ *where* P *is a distribution on* $\mathcal{Z} \times \mathcal{Y}$. *For a smooth loss* ℓ, Φ *satisfies the smoothness condition. If* $P = \frac{1}{n}\sum_{i=1}^n \delta_{(z_i,y_i)}$, *then* Φ *is the training loss, and if* P *is the test distribution, it is the test loss.*

To control the difference between the finite-particle system at time t and the optimal μ^*, we will bound $\mathbb{E}[\Phi(\mu_t^N)] - \Phi(\mu^*)$. For that purpose, we utilize the following decomposition of the error:

$$\mathbb{E}[\Phi(\mu_t^N)] - \Phi(\mu^*) = \mathbb{E}[\mathcal{U}(t, \mu_0^N) - \Phi(\mu^*)] + \mathbb{E}[\mathcal{U}(0, \mu_t^N) - \mathcal{U}(t, \mu_0^N)] =: (I) + (II).$$

4.1.1 Bounding the first term (I): $\mathbb{E}[\mathcal{U}(t, \mu_0^N)] - \Phi(\mu^*)$

Let $W_p(\mu, \nu)$ denote the p-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p$. We first observe that $m(t, \mu_0^N)$ converges to μ^* in an exponential order (geometric ergodicity) in the following sense.

Lemma 1. *Under Assumption 1, for* $\mu_t = m(t, \mu_0^N)$, *it holds that* $\mathcal{L}(\mu_t) < \infty$ *for any* $t > 0$, *and*

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \le \exp(-2\alpha\lambda(t - \tau_0))\lambda\psi_2(\tau_0) W_2(\mu_0, \mu^*)^2,$$
$$\mathrm{KL}(\mu_t, \mu^*) \le \exp(-2\alpha\lambda(t - \tau_0))\psi_2(\tau_0) W_2(\mu_0, \mu^*)^2,$$

for any $t > \tau_0$ *where* $\tau_0 > 0$ *is arbitrary,* $\psi_2(t) = \frac{1}{2\lambda}\left(\frac{B^2}{1-e^{-B^2 t}} + \frac{tB^2 \exp(4B^2 t)}{2}\right)$, *and* α *is the LSI constant of* $\tilde{\mu}_t$ *shown in Proposition 2.*

The proof can be found in Corollary 4 in the Appendix. This is an extension of the "entropy sandwich" argument in Nitanda et al. (2022); Chizat (2022), in which the right hand side of the bound is given by $\mathrm{KL}(\mu_0, \mu^*)$ instead of the Wasserstein distance. However, in our setting, $\mu_0 = \mu_0^N$ is a discrete distribution and thus the KL divergence from μ^* is not finite. To resolve this issue, our analysis shows an upper bound of the KL divergence at $t > \tau_0$ via the Wasserstein distance $W_2(\mu_0, \mu^*)$ (see Corollary 3). Then, we obtain the following theorem about the convergence of term (I).

Theorem 1 (Geometric ergodicity)**.** *Under Assumption 1, the term (I) converges as*

$$\mathbb{E}[\mathcal{U}(t, \mu_0^N)] - \Phi(\mu^*) \le C\sqrt{2\alpha^{-1}\psi_2(\tau_0)} \exp(-\alpha\lambda(t - \tau_0))\mathbb{E}[W_2(\mu_0^N, \mu^*)],$$

for any $t > \tau_0$ *where* $\tau_0 > 0$ *is an arbitrary positive real number.*

Proof. By Otto-Villani's theorem (Otto and Villani, 2000), the LSI condition implies Talagrand's inequality: $W_2(\mu, \mu^*) \le \sqrt{\frac{2}{\alpha}\mathrm{KL}(\mu, \mu^*)}$. Also, the smoothness of Φ implies that $\Phi(\mu) - \Phi(\mu^*) \le CW_2(\mu, \mu^*)$. The assertion is obtained by combining Lemma 1 and Talagrand's inequality. □

4.1.2 BOUNDING THE SECOND TERM (II): $\mathbb{E}[\mathcal{U}(0,\mu_t^N)-\mathcal{U}(t,\mu_0^N)]$

Bounding the second term (II) is much more involved. We utilize the following evaluation adapted from Delarue and Tse (2021) (see Arnaudon and Del Moral (2020) for similar argument):

$$\mathbb{E}[\mathcal{U}(0,\mu_t^N)-\mathcal{U}(t,\mu_0^N)]=\frac{1}{N}\sum_{i=1}^{d}\int_0^t\mathbb{E}\left[\int\left(\partial_{(x_1)_i}\partial_{(x_2)_i}\frac{\delta^2\mathcal{U}}{\delta\mu^2}(t-s,\mu_s^N)(x,x)\right)\mu_s^N(\mathrm{d}x)\right]\mathrm{d}s.$$

Intuitively, the integrand on the right hand side approximately represents $\mathbb{E}[\mathcal{U}(0,\mu_{s+\epsilon}^N)-\mathcal{U}(t,\mu_s^N)]$ for small ϵ, that is, how a small time difference between the finite particle and continuous limit propagates to the terminal time t. Here, for a linear operator q acting on a function $f:\mathbb{R}^d\to\mathbb{R}$, we write $f(q):=q(f)$. Then it holds (Delarue and Tse (2021); see also Appendix C.1 and C.2) that

$$\begin{aligned}\partial_{(x_1)_i}\partial_{(x_2)_i}\frac{\delta^2\mathcal{U}}{\delta\mu^2}(t,\mu_0)(x_1,x_2)=&\frac{\delta^2\Phi}{\delta\mu^2}(\mu_t)(d_i^{(1)}(t;\mu_0,\xi,x_1),d_i^{(1)}(t;\mu_0,\xi,x_2))\\&+\frac{\delta\Phi}{\delta\mu}(\mu_t)(d_{i,i}^{(2)}(t;\mu_0,x_1,x_2)),\end{aligned}$$

where $d_i^{(1)}$ and $d_{i,j}^{(2)}$ are linear operators defined by

$$\begin{aligned}d_i^{(1)}(t;\mu_0,\xi,x_1)(\phi)&=\partial_{(x_1)_i}\tfrac{\delta}{\delta\mu}(m(t;\cdot)(\phi))|_{\mu_0}(x_1),\\d_{i,j}^{(2)}(t;\mu_0,x_1,x_2)(\phi)&=\partial_{(x_1)_i}\partial_{(x_2)_j}\tfrac{\delta^2}{\delta\mu^2}(m(t;\cdot)(\phi))|_{\mu_0}(x_1,x_2),\end{aligned}$$

for a smooth test function $\phi:\mathbb{R}^d\to\mathbb{R}$, where $m(t,\mu)(\phi)=\int\phi(x)\mathrm{d}m(t,\mu)(x)$ (we will also use the same notation for a general measure μ). Given a linear operator q, we introduce a differential operator L_μ as follows,

$$L_\mu q=\lambda\Delta q+\nabla\cdot(b(\cdot,\mu)q)+\nabla\cdot\left(\mu\tfrac{\delta b}{\delta\mu}(\cdot,\mu)(q)\right),$$

which is defined in a weak sense, i.e., $(L_\mu q)(\phi)=q(L_\mu^*\phi)=q(\lambda\Delta\phi-b(\cdot,\mu)\cdot\nabla\phi-\int\frac{\delta b}{\delta\mu}(y,\mu)(\cdot)\cdot\nabla\phi(y)\mu(\mathrm{d}y))$ for a test function $\phi:\mathbb{R}^d\to\mathbb{R}$ with appropriate regularity condition. Following Delarue and Tse (2021), we know that these operators obey the following dynamics.

Proposition 3. *$d_i^{(1)}$ and $d_{i,j}^{(2)}$ follows the following differential equation: for $t>0$,*

$$\begin{cases}\partial_t d_i^{(1)}(t;\mu,x)=L_{m(t,\mu)}d_i^{(1)}(t;\mu,x),\\d_i^{(1)}(0;\mu,x)=(D'_x)_i,\end{cases}$$

where $(D'_x)_i$ is defined by $(D'_x)_i(\phi)=\partial_{x_i}\phi(x)$, and

$$\begin{cases}\partial_t d_{i,j}^{(2)}(t;\mu,x_1,x_2)=L_{m(t,\mu)}d_{i,j}^{(2)}(t;\mu,x_1,x_2)\\\qquad+\nabla\cdot\left(d_j^{(1)}(t;\mu,x_2)\frac{\delta b}{\delta\mu}(\cdot,m(t,\mu))(d_i^{(1)}(t;\mu,x_1))\right)\\\qquad+\nabla\cdot\left(d_i^{(1)}(t;\mu,x_1)\frac{\delta b}{\delta\mu}(\cdot,m(t,\mu))(d_j^{(1)}(t;\mu,x_2))\right)\\\qquad+\nabla\cdot\left(m(t,\mu),\frac{\delta^2 b}{\delta\mu^2}(\cdot,m(t,\mu))(d_i^{(1)}(t;\mu,\delta_{x_1}),d_j^{(1)}(t;\mu,\delta_{x_2}))\right),\\d_{i,j}^{(2)}(0;\mu,x_1,x_2)=0.\end{cases}$$

The above equations should be interpreted in a weak sense,i.e., when $\partial_t q_t-L_{m(t;\mu)}-r_t=0$ means that $q_t(\phi(t,\cdot))-q_s(\phi(s,\cdot))=\int_s^t q_\tau(\partial_\tau\phi(\tau,\cdot))\mathrm{d}\tau+\int_s^t q_\tau(L_{m(\tau;\mu)}^*\phi(\tau,\cdot))\mathrm{d}\tau+\int_s^t r_\tau(\phi(\tau,\cdot))\mathrm{d}\tau$ for a smooth test function $\phi:[0,\infty)\times\mathbb{R}^d\to\mathbb{R}$.

The well-posedness of this differential equation is justified in Delarue and Tse (2021); Tse (2021) for mean-field dynamics on the d-dimensional torus. Although our dynamics is defined on $\mathbb{R}^d$ and the regularization term r has unbounded gradient, the arguments there can be applied because r is convex and does not depend on the distribution. Here we choose to skip the technical details.

Our strategy is to show boundedness of $\int_0^t\mathbb{E}\left[\int\partial_{(x_1)_i}\partial_{(x_2)_i}\frac{\delta^2\mathcal{U}}{\delta\mu^2}(t-s,\mu_s^N)(x,x)\mu_s^N(\mathrm{d}x)\right]\mathrm{d}s$ by proving the exponential convergence of $\partial_{(x_1)_i}\partial_{(x_2)_i}\frac{\delta^2\mathcal{U}}{\delta\mu^2}(t-s,\mu_s^N)(x,x)$. However, this requires

a local evaluation around $m(t-s;\mu_s^N)$, for which we cannot exploit "global" properties such as the log-Sobolev condition of the optimal solution μ^*. In particular, we need to show that $m(t-s;\mu_s^N)$ also satisfies the LSI condition, which is technically demanding. To overcome this difficulty, we need to impose a stronger-than-quadratic tail growth condition on the regularizer, i.e., $r(x)=\Omega(\|x\|^{2+\delta})$ (see third point in Assumption 1). Indeed, under this assumption, we can show that $\tilde{\mu}_\tau$ for $\mu_\tau = m(\tau+s;\mu_s^N)$ satisfies the *super logarithmic Sobolev inequality*.

Definition 2 (super logarithmic Sobolev inequality (super LSI))**.** *We say that a probability measure μ satisfies super log-Sobolev inequality if there exists a monotonically non-increasing function $\beta:(0,\infty)\to\mathbb{R}$ such that for any ϕ satisfying $\mathbb{E}_\mu[\phi^2]=1$ and $\mathbb{E}_\mu[\|\nabla\phi\|^2]<\infty$, it holds that*

$$\mu(\phi^2\log\phi^2)\le r\int\|\nabla\phi\|^2\mathrm{d}\mu+\beta(r)\quad(\forall r>0).$$

It is known that super LSI implies LSI if there exists $r>0$ such that $\beta(r)=0$.

Lemma 2. *For any $\mu\in\mathcal{P}$, the probability measure $\tilde{\mu}$ corresponding to the proximal Gibbs density p_μ satisfies the super log-Sobolev inequality with $\beta(r)=C'-\frac{4+2\delta}{\delta}\log(r/2)$ with a constant $C'>0$. Furthermore, it satisfies the log-Sobolev inequality with the LSI-constant $\tilde{\alpha}=\exp(-\frac{\delta C'}{4+2\delta})$.*

See Corollary 5 in the Appendix for the proof. An important consequence of this lemma is that μ_t and $\tilde{\mu}_t$ have a bounded density ratio when t is sufficiently large, i.e., $t>T_0$ for some constant T_0. This implies that many properties of $\tilde{\mu}_t$ is also inherited by μ_t. Crucially, the bound on density ratio is strong enough for the LSI condition to be transferred from $\tilde{\mu}_t$ to μ_t, which allows us to establish the exponential convergence of $d_i^{(1)}$ and $d_{i,j}^{(2)}$.

Corollary 1. *Under Assumption 1, there exists some $T_0>0$ depending on $d,B,\delta,\alpha,\lambda,c_r,C_r$ such that for all $t\ge T_0$, we have*

$$\frac{1}{\sqrt{2}}\le\left|\frac{\mathrm{d}\mu_t}{\mathrm{d}\tilde{\mu}_t}(x)\right|\le\sqrt{2}\ \ (\forall x\in\mathbb{R}^d),\qquad\left\|\frac{\mathrm{d}\mu_t}{\mathrm{d}\tilde{\mu}_t}-1\right\|_\infty\le C'\exp(-\alpha\lambda(t-T_0)),$$

where C' is some positive constant. Moreover, for $t\ge T_0$, μ_t satisfies $(\alpha/2)$-LSI.

Equipped with the LSI condition on μ_t, we can now control term (II) in the error decomposition, the proof of which can be found in Appendix C.4.

Theorem 2 (Uniform-in-time Bound for (II))**.** *Suppose that the support of μ_0 is bounded. Then for any $0<s<t$ and $1\le i\le d$, it holds that*

$$\mathbb{E}\left[\int\left(\partial_{(x_1)_i}\partial_{(x_2)_i}\frac{\delta^2\mathcal{U}}{\delta\mu^2}(t-s,\mu_s^N)(x,x)\right)\mu_s^N(\mathrm{d}x)\right]=\mathcal{O}(\exp(-\lambda\alpha(t-s-T_0)/2)),$$

with a constant $T_0>0$. As a consequence, we have that

$$\mathbb{E}[\mathcal{U}(0,\mu_t^N)-\mathcal{U}(t,\mu_0^N)]=\mathcal{O}\left(N^{-1}\right).$$

4.1.3 Putting things together

By combining Theorems 1 and 2, we arrive at the following characterization on the difference between the finite-width neural network and the optimal (infinite-width) solution μ^*.

Corollary 2. *Under Assumption 1, if the initial distribution μ_0 has bounded support, then we have*

$$\mathbb{E}[\mathcal{U}(0,\mu_t^N)]-\Phi(\mu^*)\le\mathcal{O}(N^{-1})+C\sqrt{2\alpha^{-1}\psi_2(\tau_0)}\exp(-\alpha\lambda(t-\tau_0)),$$

for constant $C>0$ and any $\tau_0>0$.

As an important consequence of this general convergence guarantee, we know that the finite particle dynamics also exhibits a similar decay of the training and test losses compared to the continuous limit up to a $\mathcal{O}(1/N)$ discretization error; this can be straightforwardly checked by considering a special setting where $\Phi(\mu)=\mathbb{E}_{(Z,Y)\sim P}[\ell(f_\mu(Z),Y)]$ (see Example 1).

Remark 1. *Our current convergence result holds in expectation, but one may also be interested in proving a high probability statement. For that purpose, as discussed in Arnaudon and Del Moral (2020), we may apply a martingale concentration inequality (e.g., see Lemma 3.2 of Nishiyama (1997)) to obtain a guarantee in the form of $\mathcal{U}(0,\mu_t^N)-\mathcal{U}(t,\mu_0^N)-\mathbb{E}[\mathcal{U}(0,\mu_t^N)-\mathcal{U}(t,\mu_0^N)]=c\sqrt{\log(\epsilon^{-1})/N}$ with probability $1-\epsilon$.*

5 NUMERICAL EXPERIMENTS

We provide empirical support for our propagation of chaos result in a synthetic student-teacher setting. We consider the empirical risk minimization problem, where the training labels are generated by a teacher model which is a Gaussian function defined as $f^*(z) = \exp\left(-\frac{\|z-a\|^2}{2d}\right)$. We set $n = 2000, d = 20$. The loss is chosen to be the squared error, and for the regularization term we set $r(x) = \|x\|^2$ or $r(x) = \|x\|^4$, and the regularization strength $\lambda_1 = \lambda = 10^{-2}$. The student model is a two-layer neural network with tanh activation, and the width N is taken to be $\{16, 32, 64, 128, 256, 512, 1024, 2048\}$. We optimize the student model using NPGD with step size $\eta = 10^{-2}$. The global optimal solution μ^* is approximated via the particle dual averaging (PDA) algorithm (Nitanda et al., 2020). We set the model width $N = 2048$ and number of outer loop steps $T = 250$. We scale the number of inner loop steps T_t with t, and the step size η_t with $1/\sqrt{t}$, where t is the outer loop iteration.

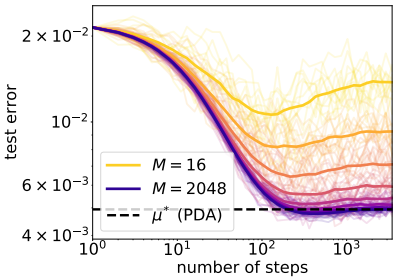

Figure 2: Test error of NNs optimized by NPGD ($r(x) = \|x\|^2$). Solid curve: mean over 100 runs. Translucent curve: individual runs.

For the figures, we report the training or test error without the regularization terms. In Figure 1 in the Introduction, we plot the training error for $r(x) = \|x\|^4$; whereas in Figure 2, we plot the test error for $r(x) = \|x\|^2$. Observe that even though our current theoretical analysis do not cover the latter setting, the empirical trends are almost identical: as the width N increases, the performance of the finite-width network improves and approaches that of the (approximate) optimal solution μ^*.

6 CONCLUSION

In this paper, we established a uniform-in-time quantitative propagation of chaos for the mean-field Langevin dynamics in the context of neural network optimization. In contrast to most existing works, our analysis gives a quantitative discretization error and does not blow up through time, and we do not impose the commonly-assumed weak interaction condition. This is achieved by utilizing the super LSI condition that is satisfied by a regularization term with super-quadratic tail. This condition enables us to establish the log-Sobolev inequality of the intermediate solution μ_t, which gives an exponential convergence of the error propagation.

Limitations and future directions. Our current analysis requires the super-quadratic tail of the regularization term, which does not cover the commonly-used ℓ_2 regularization (weight decay). An interesting research problem is to relax such assumption to cover more general regularization terms. Indeed, our numerical experiments show that the quadratic regularization also converges similarly (Fig. 2). Another limitation is that our result is restricted to the continuous-time dynamics. An extension to discrete-time dynamics is also an important future direction.

REFERENCES

E. Abbe, E. B. Adsera, and T. Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

S. Akiyama and T. Suzuki. On learnability via gradient method for two-layer relu neural networks in teacher-student setting. In *International Conference on Machine Learning*, pages 152–162. PMLR, 2021.

Z. Allen-Zhu and Y. Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE, 2022.

L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2005.

M. Arnaudon and P. Del Moral. A second order analysis of Mckean–Vlasov semigroups. *The Annals of Applied Probability*, 30(6):2613–2664, 2020.

J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, 2022.

D. Bakry and M. Émery. Diffusions hypercontractives. In J. Azéma and M. Yor, editors, *Séminaire de Probabilités XIX 1983/84*, pages 177–206, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. ISBN 978-3-540-39397-9.

D. Bakry, I. Gentil, M. Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

M. Chen, Y. Bai, J. D. Lee, T. Zhao, H. Wang, C. Xiong, and R. Socher. Towards understanding hierarchical learning: Benefits of neural representations. *Advances in Neural Information Processing Systems*, 33:22134–22145, 2020a.

Z. Chen, Y. Cao, Q. Gu, and T. Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. *arXiv preprint arXiv:2002.04026*, 2020b.

Z. Chen, G. M. Rotskoff, J. Bruna, and E. Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. *arXiv preprint arXiv:2008.09623*, 2020c.

L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

L. Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=BDqzLH1gEm.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 3040–3050, 2018.

V. De Bortoli, A. Durmus, X. Fontaine, and U. Simsekli. Quantitative propagation of chaos for sgd in wide neural networks. *Advances in Neural Information Processing Systems*, 33:278–288, 2020.

F. Delarue and A. Tse. Uniform in time weak propagation of chaos on the torus. *arXiv preprint arXiv:2104.14973*, 2021.

M. G. Delgadino, R. S. Gvalani, G. A. Pavliotis, and S. A. Smith. Phase transitions, logarithmic sobolev inequalities, and uniform-in-time propagation of chaos for weakly interacting diffusions. *arXiv preprint arXiv:2112.06304*, 2021.

A. Durmus, A. Eberle, A. Guillin, and R. Zimmer. An elementary approach to uniform in time propagation of chaos. *Proceedings of the American Mathematical Society*, 148(12):5387–5398, 2020.

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

H. Föllmer. Time reversal on wiener space. In *Stochastic processes—mathematics and physics*, pages 119–129. Springer, 1986.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9111–9121, 2019.

A. Guillin, W. Liu, L. Wu, and C. Zhang. Uniform poincaré and logarithmic sobolev inequalities for mean field particle systems. *The Annals of Applied Probability*, 32(3):1590–1614, 2022.

R. Holley and D. Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of statistical physics*, 46(5-6):1159–1194, 1987.

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

X. Huang, P. Ren, and F.-Y. Wang. Distribution dependent stochastic differential equations. *Frontiers of Mathematics in China*, 16(2):257–301, 2021.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589, 2018.

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

M. Kac. Foundations of kinetic theory. In *Proceedings of The third Berkeley symposium on mathematical statistics and probability*, volume 3, pages 171–197, 1956.

M. Kac. *Probability and related topics in physical sciences*, volume 1. American Mathematical Soc., 1959.

H. Kahn and T. E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series*, 12:27–30, 1951.

Y.-H. Kim. A note on the moment inequalities for stochastic integral. *Math. Inequal. Appl*, 16(4): 1023–1029, 2013.

D. Lacker. Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions. *arXiv preprint arXiv:2105.02983*, 2021.

Y. Li, T. Ma, and H. R. Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Proceedings of Conference on Learning Theory 33*, pages 2613–2682, 2020.

H. P. McKean. A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911, 1966.

H. P. McKean. Propagation of chaos for a class of non-linear parabolic equations. *Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7, Catholic Univ., 1967)*, pages 41–57, 1967.

S. Mei, A. Montanari, and P.-M. 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.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019.

S. Mischler. An introduction to evolution PDEs, Chapter 0: On the Gronwall lemma, 2019. URL https://www.ceremade.dauphine.fr/~mischler/Enseignements/M2evol2018/chap0.pdf.

Y. Nishiyama. Some central limit theorems for ℓ^∞-valued semimartingales and their applications. *Probability theory and related fields*, 108(4):459–494, 1997.

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

A. Nitanda, D. Wu, and T. Suzuki. Particle dual averaging: Optimization of mean field neural networks with global convergence rate analysis, 2020.

A. Nitanda, D. Wu, and T. Suzuki. Convex analysis of the mean field langevin dynamics. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9741–9757. PMLR, 28–30 Mar 2022.

K. Oko, T. Suzuki, A. Nitanda, and D. Wu. Particle stochastic dual coordinate ascent: Exponential convergent algorithm for mean field neural network optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PQQp7AJwz3.

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

P. Ren and F.-Y. Wang. Bismut formula for lions derivative of distribution dependent sdes and applications. *Journal of Differential Equations*, 267(8):4745–4777, 2019.

P. Ren and F.-Y. Wang. Exponential convergence in entropy and Wasserstein for Mckean–Vlasov sdes. *Nonlinear Analysis*, 206:112259, 2021.

G. M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

G. M. Rotskoff, S. Jelassi, J. Bruna, and E. Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *Proceedings of International Conference on Machine Learning 36*, pages 9689–9698, 2019.

S. Salem. A gradient flow approach of uniform in time propagation of chaos for particles in double a well confinement. *arXiv preprint arXiv:1810.08946*, 2018.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

A.-S. Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

A. Tse. Higher order regularity of nonlinear fokker-planck pdes with respect to the measure component. *Journal de Mathématiques Pures et Appliquées*, 150:134–180, 2021.

C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

F. Wang. *Functional inequalities, Markov semigroups and spectral theory*. The Science Series of the Contemporary Elite Youth. Elsevier, 2005.

F.-Y. Wang. Integration by parts formula and shift harnack inequality for stochastic equations. *The Annals of Probability*, 42(3):994–1019, 2014.

F.-Y. Wang. Distribution dependent sdes for landau type equations. *Stochastic Processes and their Applications*, 128(2):595–621, 2018.

C. Wei, J. D. Lee, Q. Liu, and T. Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9712–9724, 2019.

G. Yang and E. J. Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

Appendix

A Additional Related Works

Optimization of mean-field neural networks. Mean-field analysis of two-layer neural networks describes the optimization dynamics as a partial differential equation (PDE) of the parameter distribution, from which global convergence xB be shown (Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020). However, quantitative convergence results usually requires additional assumptions on the learning problem (Chizat, 2019; Akiyama and Suzuki, 2021), or modification of the dynamics (Rotskoff et al., 2019; Wei et al., 2019). For the entropy-regularized objective (4), efficient optimization algorithms have been proposed in Nitanda et al. (2020); Oko et al. (2022); noticeably, the quantitative convergence rate guarantees FOR these particle-based methods remain valid in both finite-width and discrete-time settings. On the other hand, if we restrict ourselves the standard gradient descent-based methods, then quantitative propagation of chaos has also been shown in De Bortoli et al. (2020); Chen et al. (2020c). In particular, De Bortoli et al. (2020) studied the impact of learning rate in the stochastic gradient descent update, but did not provide a convergence rate or uniform control of the discretization error over time (due to the lack of regularization). Chen et al. (2020c) showed that the long-time fluctuation due to finite-width can be controlled assuming that the mean-field dynamics converges at a specific rate, but such condition is very challenging to establish in their setting. In contrast, our result provides a uniform-in-time bound on the finite-width discretization error under conditions that can be verified for regularized risk minimization problems.

Interacting particle systems and propagation of chaos. The propagation of chaos has been analyzed mainly in the McKean-Vlasov equations: that has a bias having form of $b(x,\mu) = \nabla V(x) - \nabla \int W(\cdot, y)\mathrm{d}\mu(y)|_x$. Generally, neural network optimization is not included in this class, but techniques to analyze such the equation can be applied to the neural network setting. There are several analyses on the discretization error with a bounded time horizon (see Lacker (2021) and references therein, for example). That is, for a fixed time horizon T, it has been shown that $\sup_{t\in[0,T]} |\Psi(\mu_t^N) - \Psi(\mu_t)| \leq C_T/N$. However, the constant C_T depends on the time horizon T and usually has exponential dependency. To obtain uniform in time analyses, there are roughly two approaches: (1) uniform log-Sobolev (or Poincare) inequality approach and (2) local Taylor expansion approach. (1) the first approach directly derives the LSI constant of the the N-particle dynamics ($X = (X_t^i)_{i=1}^N$) and show that the constant can be bounded from below *uniformly* over all N. Guillin et al. (2022) established a device to derive the uniform LSI constant (see also the references therein). This is useful because it can deal with a non-convex potential, but it still requires a small interaction assumption. Ren and Wang (2021); Delgadino et al. (2021) showed geometric ergodicity based on a similar evaluation. Salem (2018) showed a uniform *WJ-inequality* instead of the uniform log-Sobolev inequality, but it also requires a weak interaction condition. (2) the second approach is what we have employed. Arnaudon and Del Moral (2020); Delarue and Tse (2021) developed a frame-work in this direction. Durmus et al. (2020) developed another technique using a sophisticated coupling argument. All of them still require a weak interaction condition to show the uniform-in-time evaluation, and thus cannot be applied to the neural network setting.

B Basic properties of the solution

B.1 Boundedness and uniqueness of the solution

Proposition 4 (Theorem 2.1 and Corollary 4.3 of Wang (2018), adapted)**.** *Suppose that there exit* $K_1, K_2, K_3 \in C([0,\infty);(0,\infty))$ *such that*

- $-2\langle b(x,\mu) - b(y,\nu), x-y\rangle \leq K_1(t)\|x-y\|^2 + K_2(t)W_2(\mu,\nu)\|x-y\|$ *for any* $x, y \in \mathbb{R}^d$, $t \geq 0$ *and* $\mu, \nu \in \mathcal{P}_2$,
- $\|b(0,\mu)\|^2 \leq K_3(t)\{1 + \mu(\|\cdot\|^2)\}$ *for any* $\mu \in \mathcal{P}_2$ *and* $t \geq 0$.

Then, the mean-filed Langevin dynamics (3) *has a unique strong solution.*

Moreover, for the two different solutions $X_t^{(1)}$ and $X_t^{(2)}$ with different initial distributions $\mu_0^{(1)}, \mu_0^{(2)} \in \mathcal{P}_2$, the corresponding laws $\mu_t^{(k)} = \mathrm{Law}(X_t^{(k)})$ $(k = 1, 2)$ are equivalent for $t > 0$ and satisfy the following contraction property:

$$W_2(\mu_t^{(1)}, \mu_t^{(2)})^2 \leq \psi_1(t) W_2(\mu_0^{(1)}, \mu_0^{(2)})^2, \quad \mathrm{KL}(\mu_t^{(1)}, \mu_t^{(2)}) \leq \psi_2(t) W_2(\mu_0^{(1)}, \mu_0^{(2)})^2, \tag{6}$$

for $t > 0$, where $\psi_j : (0, \infty) \to [0, \infty)$ depends only on K_1, K_2, K_3, λ and ϕ_1 is an increasing function.

Here, it possibly holds that $\lim_{t \to 0} \psi_2(t) = \infty$. Indeed, it can be written as

$$\psi_2(t) = \frac{1}{2\lambda} \left(\frac{K_1(t)}{1 - e^{-K_1(t)t}} + \frac{t K_2(t) \exp(2t(K_1(t) + K_2(t)))}{2} \right).$$

Corollary 3. *Under Assumption 1, the mean-filed Langevin dynamics (3) has a unique strong solution. Moreover, the two distributions $\mu_t^{(1)}$ and $\mu_t^{(2)}$ corresponding to different initial distributions in $\mathcal{P}_2$ are equivalent and satisfy the contraction property (6).*

In particular, μ_t is equivalent to μ^* and hence is equivalent to the Lebesgue measure. Therefore, μ_t has a density that is positive for all $x \in \mathbb{R}^d$, and if $\mu_0 \in \mathcal{P}_2$, then the density of μ_t satisfies $(t, x) \mapsto \frac{\mathrm{d}\mu_t}{\mathrm{d}x}(x) \in C^{1,\infty}((0, \infty) \times \mathbb{R}^d, \mathbb{R})$.

Proof. Let $H(x, \mu) = \frac{1}{n} \sum_{j=1}^n \ell_j'(f_\mu) h_j(x)$. We just need to check the two conditions in Proposition 4. The first condition can be checked as follows: by noticing the convexity of r, we have

$$\begin{aligned}
&- 2\langle b(x, \mu) - b(y, \nu), x - y \rangle \\
&= -2\langle \nabla H(x, \mu) - \nabla H(y, \nu), x - y \rangle - 2\lambda_1 \langle \nabla r(x) - \nabla r(y), x - y \rangle \\
&= -2\langle \nabla H(x, \mu) - \nabla H(y, \mu), x - y \rangle - 2\langle \nabla H(y, \mu) - \nabla H(y, \nu), x - y \rangle \\
&\quad - 2\lambda_1 \langle \nabla r(x) - \nabla r(y), x - y \rangle \\
&\leq 2B^2 \|x - y\|^2 + 2B^2 \|f_\mu - f_\nu\|_\infty \|x - y\| \quad (\because \text{Assumption 1 and convexity of } r) \\
&\leq 2B^2 \|x - y\|^2 + 2B^3 W_2(\mu, \nu) \|x - y\|,
\end{aligned}$$

which yields the first condition. Next, the second condition can be guaranteed as

$$\|b(0, \mu)\| \leq \|\nabla H(0, \mu)\| + \lambda_1 \|\nabla r(0)\| \leq B^2 + \lambda C_r.$$

Therefore, applying Proposition 4, we obtain the first assertion.

As for the secone assertion, we first note that μ^* is an invariant measure of the mean field Langevin dynamics. Hence, if $\mu_0 = \mu^*$, then $\mu_t = \mu^*$ for any $t > 0$. Moreover, remind that $p_{\mu^*} = \mu^*$ by Proposition 1. Combining these relations, we have $\mu^* = p_{\mu^*} = \mu_t$ $(t > 0)$ when the initial distribution is $\mu_0 = \mu_t$. Therefore, μ_t with a general initial distribution is equivalent to μ^*. $(t, x) \mapsto \frac{\mathrm{d}\mu_t}{\mathrm{d}x}(x) \in C^{1,\infty}((0, \infty) \times \mathbb{R}^d, \mathbb{R})$ follows from Theorem 5.1 of Jordan et al. (1998). □

In the analysis, it is important to ensure the boundedness of the moments of X_t. Indeed, we have the following estimate.

Lemma 3. *Under Assumption 1, for any $p \geq 2$, $\mathbb{E}[\|X_0\|^p] < \infty$ implies*

$$\mathbb{E}\left[\sup_{0 \in [0, T]} \|X_t\|^p \right] < \infty$$

for any $T > 0$.

This is a direct consequence from Theorem 2.1 of Wang (2018). Therefore, we have that $\mu_t \in \mathcal{P}_p$ as long as $\mu_0 \in \mathcal{P}_p$. Indeed, we are interested in a situation where $\mu_0 = \mu_s^N = \frac{1}{N} \sum_{i=1}^N \hat{X}_s^i$, and hence $\mu_t \in \mathcal{P}_p$ for any $p \geq 2$ because a discrete measure has a finite moment for any $p \geq 2$.

If $p = 2$ or $p = 2 + \delta$, we have more strict uniform bound as follows.

Lemma 4. *Under Assumption 1, we have the following uniform boundedness of the moments:*

$$\sup_{t>0} \mathbb{E}[\|X_t\|^2] \le \max\left\{ \frac{1}{\lambda_1 c(1+\delta/2)} \left[\frac{B^4}{\lambda_1 c(1+\delta/2)} + \lambda_1 c\delta + 2\lambda d \right], \mathbb{E}[\|X_0\|^2] \right\},$$

$$\begin{aligned}
&\sup_{t>0} \mathbb{E}[r(X_t)] \\
&\le \max\left\{ \tfrac{C_r(2+\delta)}{\lambda_1 c_r^2(1+\delta)} \left[\frac{1}{2+\delta} \frac{(B^2 C_r)^{2+\delta}}{(\lambda_1 c_r/2)^{1+\delta}} + \frac{(1+2\delta)\lambda_1 c_r}{2+\delta} + \lambda C_r + \frac{2}{2+\delta} (\lambda C_r)^{(2+\delta)/2} \left(\tfrac{\delta}{\lambda c_r(1+\delta)} \right)^{\delta/2} \right], \right. \\
&\qquad \left. \mathbb{E}[r(X_0)] \right\}.
\end{aligned}$$

The same bounds also hold with respect to $\mathbb{E}[\|\hat{X}_t^i\|^2]$ *and* $\mathbb{E}[r(\hat{X}_t^i)]$.

Proof. Let $H_t(x) = \frac{1}{n} \sum_{j=1}^n \ell_j'(f_{\mu_t}) h_j(x)$. By the formula of the infinitesimal generator, we have that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[\|X_t\|^2] = \mathbb{E}[-2X_t^\top b(X_t, \mu_t)] + 2\lambda d.$$

By the Young's inequality, the right hand side can be bounded as

$$\begin{aligned}
&-2\mathbb{E}\left[X_t^\top \left(\nabla H_t(X_t) + \lambda_1 \nabla r(X_t) \right) \right] + 2\lambda d \\
&\le 2B^2 \mathbb{E}[\|X_t\|] - 2\lambda_1 c \mathbb{E}[\|X_t\|^{2+\delta}] + 2\lambda d \\
&\le \frac{4B^4}{4\lambda_1 c(1+\delta/2)} + \frac{2\lambda_1 c(1+\delta/2)\mathbb{E}[\|X_t\|^2]}{2} + 2\lambda_1 c \left\{ \frac{\delta}{2} - (1+\tfrac{\delta}{2})\mathbb{E}[\|X_t\|^2] \right\} + 2\lambda d \\
&\le \frac{B^4}{\lambda_1 c(1+\delta/2)} + \lambda_1 c\delta + 2\lambda d - \lambda_1 c(1+\tfrac{\delta}{2})\mathbb{E}[\|X_t\|^2].
\end{aligned}$$

Hence, we obtain that

$$\mathbb{E}[\|X_t\|^2] \le \frac{1}{\lambda_1 c(1+\delta/2)} \left[\frac{B^4}{\lambda_1 c(1+\delta/2)} + \lambda_1 c\delta + 2\lambda d \right] \vee \mathbb{E}[\|X_0\|^2].$$

In the same vein, we can show the bound for $r(X_t)$ as follows. First, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[r(X_t)] = \mathbb{E}[-\nabla^\top r(X_t) b(X_t, \mu_t)] + \lambda \mathbb{E}[\mathrm{Tr}[\nabla \nabla^\top r(X_t)]].$$

By the Young's inequality, the right hand side can be bounded as

$$\begin{aligned}
&-\mathbb{E}\left[\nabla^\top r(X_t) \left(\nabla H_t(X_t) + \lambda_1 \nabla r(X_t) \right) \right] + \lambda C_r (1 + \mathbb{E}[\|X_t\|^\delta]) \\
&\le B^2 \mathbb{E}[\|\nabla r(X_t)\|] - \lambda_1 \mathbb{E}[\|\nabla r(X_t)\|^2] + \lambda C_r (1 + \mathbb{E}[\|X_t\|^\delta]) \\
&\le B^2 C_r \mathbb{E}[\|X_t\|^{1+\delta}] - \lambda_1 c_r \mathbb{E}[\|X_t\|^{2(1+\delta)}] + \lambda C_r (1 + \mathbb{E}[\|X_t\|^\delta]) \\
&\le \frac{1}{2+\delta} \frac{(B^2 C_r)^{2+\delta}}{(\lambda_1 c_r/2)^{1+\delta}} + \frac{\lambda_1 c_r (1+\delta)}{2(2+\delta)} \mathbb{E}[\|X_t\|^{2+\delta}] - \lambda_1 c_r \left\{ \frac{2(1+\delta)}{2+\delta} \mathbb{E}[\|X_t\|^{2+\delta}] - \frac{\delta}{2+\delta} \right\} \\
&\quad + \lambda C_r + \frac{2}{2+\delta} (\lambda C_r)^{(2+\delta)/2} \left(\frac{\delta}{\lambda c_r (1+\delta)} \right)^{\delta/2} + \frac{\lambda_1 c_r (1+\delta)}{2(2+\delta)} \mathbb{E}[\|X_t\|^{2+\delta}] \\
&\le \frac{1}{2+\delta} \frac{(B^2 C_r)^{2+\delta}}{(\lambda_1 c_r/2)^{1+\delta}} + \frac{\delta \lambda_1 c_r}{2+\delta} + \lambda C_r + \frac{2}{2+\delta} (\lambda C_r)^{(2+\delta)/2} \left(\frac{\delta}{\lambda c_r (1+\delta)} \right)^{\delta/2} \\
&\quad - \frac{\lambda_1 c_r (1+\delta)}{2+\delta} \underbrace{\mathbb{E}[\|X_t\|^{2+\delta}]}_{\ge \mathbb{E}[r(X_t)]/C_r - 1}.
\end{aligned}$$

This gives the second bound.

The same argument can be applied to $\hat{X}_t^i$, which concludes the assertion. □

Lemma 5. *Under Assumption 1, for $\mu_0 \in \mathcal{P}_p$ with $p \geq 2$, it holds that*

$$\nabla \log(\mu_t)(x) = -\frac{1}{t\sqrt{2\lambda}}\mathbb{E}\left[\int_0^t (I + s\nabla b(\cdot, \mu_s)|_{X_s}) \cdot \mathrm{d}W_s | X_t = x\right],$$

and, if $p/\delta \geq 1$, it also holds that

$$\mathbb{E}[\|\nabla \log(\mu_t)(X_t)\|^{p/\delta}] < \infty,$$

for any $t > 0$.

Proof. First, note that $\mu_0 \in \mathcal{P}_p$ ensures differentiability of μ_t. The characterization of $\nabla \log(\mu_t)$ is given by the integration by parts formula investigated by Föllmer (1986) (see also Lemma 6.2 of Hu et al. (2019), Wang (2014) and Theorem 5.1 of Wang (2018)).

As for the moment bound, first we note that

$$\|\nabla b(\cdot, \mu)|_x\| = O(1 + \|x\|^\delta),$$

by the assumption. Then, by the Jensen's inequality and the moment inequality of stochastic integral (Kim, 2013), we have that, for $q = p/\delta$,

$$\begin{aligned}\mathbb{E}[\|\nabla \log(\mu_t)\|^q] &\lesssim \left(\frac{1}{t\sqrt{2\lambda}}\right)^q C_{t,q}\mathbb{E}\left[\int_0^t (1 + s\|\nabla b(\cdot, \mu_s)|_{X_s}\|)^q \mathrm{d}s\right] \\ &\lesssim \left(\frac{1}{t\sqrt{2\lambda}}\right)^q C_{t,q}\mathbb{E}\left[\int_0^t (1 + s^q(1 + \|X_s\|^\delta))^q \mathrm{d}s\right] < \infty \quad (\because \text{Lemma 3}),\end{aligned}$$

where $C_{t,q} = \left(\frac{q(q-1)}{2}\right)^{q/2} t^{\frac{q-2}{2}}$. □

According to Lemma 3 and the remark following the lemma, we have $\mu_t \in \mathcal{P}_p$ for any $p \geq 2$ in our situation where μ_0 is a discrete measure like $\mu_0 = \mu_s^N$. Hence, we may assume $\mathbb{E}[\|\nabla \log(\mu_t)(X_t)\|^p] < \infty$ for any $p \geq 2$. In particular, the Fisher divergence $I(\mu_t||\mu^*)$ is well-defined for any $t > 0$ (but not well defined for $t = 0$).

B.2 Geometric ergodicity

For $\mu, \nu \in \mathcal{P}$ where ν is absolutely continuous with respect to μ and thus can be written as $\mathrm{d}\nu = f\mathrm{d}\mu$, the *Fisher* divergence of ν with respect to μ is defined as

$$I(\nu||\mu) = 4\int \|\nabla\sqrt{f}\|^2 \mathrm{d}\mu = \int \|\nabla \log(f)\|^2 \mathrm{d}\nu.$$

Proposition 5 (Geometric ergodicity of mean-field Langevin dynamics (Nitanda et al., 2022; Chizat, 2022))**.**

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \exp(-2\alpha\lambda t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu^*)).$$
$$\lambda \mathrm{KL}(\mu||\mu^*) \leq \mathcal{L}(\mu) - \mathcal{L}(\mu^*) \leq \lambda \mathrm{KL}(\mu||p_\mu).$$

Although Nitanda et al. (2022); Chizat (2022) assumed $r(x) = \Theta(\|x\|^2)$, we can adapt the same argument also to our situation. Indeed, the quadraticity of the regularization term is used to ensure the well-posedness of the solution, but we have it by Corollary 3 in our setting, which yields the assertion of the proposition.

Combining Corollary 3 and Propositions 5, we obtain the following corollary.

Corollary 4. *Under Assumption 1, for any $\mu_0 \in \mathcal{P}$ with $W_2(\mu_0, \mu^*) < \infty$, it holds that $\mathcal{L}(\mu_t) < \infty$ and*

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \exp(-2\alpha\lambda(t - \tau_0))\lambda\psi_2(\tau_0)W_2(\mu_0, \mu^*)^2,$$
$$\mathrm{KL}(\mu_t, \mu^*) \leq \exp(-2\alpha\lambda(t - \tau_0))\psi_2(\tau_0)W_2(\mu_0, \mu^*)^2,$$

for any $t > \tau_0$ where $\tau_0 > 0$ can be arbitrary.

Proof. We know that $\mu^* = p_{\mu^*} = \mu_t$ $(t > 0)$ when the initial distribution is $\mu_0 = \mu_t$ from the proof of Corollary 3. Plugging into this relation to Corollary 3 and Propositions 5 gives the assertion. □

In the convergence analysis, we need to assume the LSI (or PI) condition on μ_t instead of $\tilde{\mu}_t$. This is not generally ensured. However, we can have it if the semi-group satisfies the *super log-Sobolev inequality*. Indeed, super log-Sobolev inequality ensures the *ultra-contractivity* yielding an L^∞-convergence of the density ratio between μ_t and $\tilde{\mu}_t$. This is remarkably useful to transfer the LSI property of $\tilde{\mu}_t$ to μ_t.

Definition 3 (super log-Sobolev inequality (super LSI))**.** *We say that a probability measure μ satisfies super log-Sobolev inequality if there exits a monotonically non-increasing function $\beta : (0, \infty) \to \mathbb{R}$ such that*

$$\mu(f^2 \log f^2) \le r \int \nabla f \cdot \nabla f \mathrm{d}\mu + \beta(r) \quad (\forall r > 0,\ \forall f \in D(\mathcal{E}),\ \mu(f^2) = 1).$$

Proposition 6. *If a probability measure μ is given by $\mu = \exp(-V)$ where $V(x) = \lambda_1 r(x) + H(x)$ with a convex function $r : \mathbb{R}^d \to \mathbb{R}$ satisfying $\lambda_1 \nabla r(x) \cdot x \ge c\|x\|^{2+\delta}$ for $\delta > 0$ and $h : \mathbb{R}^d \to \mathbb{R}$ satisfying $\|\nabla H\|_\infty \le C < \infty$, then μ satisfies the super log-Sobolev inequality with $\beta(r) = C' - \frac{4+2\delta}{\delta}\log(r/2)$ where $C' > 0$ is a constant depending on d, c, C, δ. In particular, it satisfies the log-Sobolev inequality with a constant $\alpha' > 0$ such that $\beta(2/\alpha') = 0$.*

Proof. This can be proven by adapting Corollary 5.7.5 of Wang (2005). Let P_t be the semigroup that corresponds to the generator $L^*\phi = \Delta\phi - \nabla V \cdot \nabla\phi$. Then, we have that

$$\begin{aligned} L^*\|x\|^2 = d - (\lambda_1 \nabla r + \nabla H) \cdot (2x) &\le d - 2c\|x\|^{2+\delta} + 2\|\nabla H\|_\infty \|x\| \\ &\le d - 2c\|x\|^{2+\delta} + c\|x\|^{2+\delta} + \frac{\|\nabla H\|_\infty^{\frac{2+\delta}{\delta}}}{c} \quad (\because \text{Young's inequality}) \\ &= d + \frac{C^{\frac{2+\delta}{\delta}}}{c} - c\|x\|^{2+\delta}. \end{aligned}$$

Corollary 5.7.5 of Wang (2005) yields that

$$\|P_t\|_{L^2(\mu)\to L^\infty(\mu)} \le \exp[c' t^{-(1+\delta/2)/(\delta/2)}] = \exp[c' t^{-(2+\delta)/\delta}],$$

for a constant $c' > 0$ depending on d, c, C. Indeed, (5.7.9) of Wang (2005) holds for $c \leftarrow d + \frac{C^{\frac{2+\delta}{\delta}}}{c}$ and $\gamma(r) \leftarrow cr^{1+\delta/2}$ in their notations, which yields the bound.

Then, Theorem 5.1.7 of Wang (2005) states that the super log-Sobolev inequality holds for $\beta(r) = 2\log\|P_{r/2}\|_{L^2(\mu)\to L^\infty(\mu)} \le 2\log(c') - 2\frac{2+\delta}{\delta}\log(r/2)$. By resetting $C' \leftarrow 2\log(c')$, we obtain the assertion. □

Corollary 5. *$\tilde{\mu}_t$ satisfies the super log-Sobolev inequality with $\beta(r) = C' - \frac{4+2\delta}{\delta}\log(r/2)$. In particular, it satisfies the log-Sobolev inequality with the LSI-constant $\tilde{\alpha} = \exp(-\frac{\delta C'}{4+2\delta})$.*

Let $P_{s,t}$ be the subgroup associated with μ_t, i.e., $\mu_s P_{s,t} f = \mu_t f$.

Theorem 3. *There exists $t_0 \in (0, 1]$ such that*

$$\|P_{s,s+t_0}\|_{L^2(\tilde{\mu}_s)\to L^\infty(\tilde{\mu}_{s+t_0})} \le \exp\left[\int_2^\infty \frac{\beta(1/p)}{p^2}\mathrm{d}p + Ct_0\right]$$

where $C > 0$ is a constant. In particular, there exists $C_0 > 0$ such that $\|P_{s,s+t_0}\|_{L^2(\tilde{\mu}_s)\to L^\infty(\tilde{\mu}_{s+t_0})} \le C_0 < \infty$ uniformly over $s > 0$.

Proof. We notice that

$$\tilde{\mu}_t(f^p \log(f)) \le -r\tilde{\mu}_t(f^{p-1}L^*_{\mu_t} f) + \beta\left(\frac{4(p-1)r}{p}\right)p^{-1}\tilde{\mu}_t(f^p) + \tilde{\mu}_t(f^p)\log(\|f\|_{L^p(\tilde{\mu}_t)}). \quad (7)$$

for all $f \in \mathcal{D}(L^*_{\mu_t})$ such that $f \geq 0$ (which will be noted by $f \in \mathcal{D}_+$), $2 < p < \infty$, $r > 0$ (The invariant measure corresponding to $L^*_{\mu_t}$ is $\tilde{\mu}_t$, i.e., $L_{\mu_t}\tilde{\mu}_t = 0$).

Let $t_0 := \int_2^\infty \frac{\gamma(p)}{p}\mathrm{d}p \leq 1$ for $\gamma(p) = \frac{1}{2p}$. We define $p(\tau)$ and $N(\tau)$ as functions on $[s,t)$ such that

$$p'(\tau) = \frac{p(\tau)}{\gamma \circ p(\tau)}, \quad p(s) = 2,$$
$$N'(\tau) = \frac{p'(\tau)\beta\left(\frac{4(p(\tau)-1)\gamma\circ p(\tau)}{p(\tau)}\right)}{p(\tau)^2}, \quad N(0) = 0.$$

Then, for $f \in \mathcal{D}_+$, one has

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\|P_{s,\tau}\|_{L^{p(\tau)}(\tilde{\mu}_\tau)} =& \frac{\tilde{\mu}_\tau(\frac{\mathrm{d}}{\mathrm{d}\tau}(P_{s,\tau}f)^{p(\tau)})}{p(\tau)\|P_{s,\tau}f\|^{p(\tau)-1}_{p(\tau)}} - \frac{p'(\tau)}{p(\tau)}\|P_{s,\tau}f\|_{L^{p(\tau)}(\tilde{\mu}_\tau)}\log(\|P_{s,\tau}f\|_{L^{p(\tau)}(\tilde{\mu}_\tau)}) \\ &+ \frac{1}{p(\tau)}\|P_{s,\tau}f\|^{1-p(\tau)}_{L^{p(\tau)}(\tilde{\mu}_\tau)}\int |P_{s,\tau}f|^{p(\tau)}\frac{\mathrm{d}}{\mathrm{d}\tau}\tilde{\mu}_\tau \mathrm{d}x. \end{aligned} \tag{8}$$

The last term of the right hand side can be bounded as

$$\begin{aligned}&\frac{1}{p(\tau)}\|P_{s,\tau}f\|^{1-p(\tau)}_{L^{p(\tau)}(\tilde{\mu}_\tau)}\int |P_{s,\tau}f|^{p(\tau)}\frac{\mathrm{d}}{\mathrm{d}\tau}\tilde{\mu}_\tau \mathrm{d}x \\ \leq& \frac{1}{p(\tau)}\|P_{s,\tau}f\|^{1-p(\tau)}_{L^{p(\tau)}(\tilde{\mu}_\tau)}\int |P_{s,\tau}f|^{p(\tau)}|\frac{1}{n}\sum_{j=1}^n \ell_j''(f_{\mu_\tau})h_j\mu_t(L^*_{\mu_t}f)|\tilde{\mu}_\tau \mathrm{d}x \\ \leq& C\frac{1}{p(\tau)}\|P_{s,\tau}f\|^{1-p(\tau)}_{L^{p(\tau)}(\tilde{\mu}_\tau)}\int |P_{s,\tau}f|^{p(\tau)}\tilde{\mu}_\tau \mathrm{d}x = C\frac{\|P_{s,\tau}f\|_{L^{p(\tau)}(\tilde{\mu}_\tau)}}{p(\tau)}. \end{aligned} \tag{9}$$

Combining these inequalities (8) and (9) with the super log-Sobolev inequality (7) yields that

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\tau}(e^{-N(\tau)}\|P_{s,\tau}f\|_{L^{p(\tau)}(\tilde{\mu}_\tau)}) \\ &= \frac{p'(\tau)e^{-N(s)}}{p(\tau)\|P_{s,\tau}f\|^{p(\tau)-1}_{L^{p(\tau)}(\tilde{\mu}_\tau)}}\Big\{\tilde{\mu}_\tau((P_{s,\tau}f)^{p(\tau)}\log P_{s,\tau}f) + \frac{p(\tau)}{p'(\tau)}\tilde{\mu}_\tau((P_{s,\tau}f)^{p(\tau)-1}L^*_{\mu_t}P_{s,\tau}f) \\ &\quad - \frac{N'(\tau)p(\tau)}{p'(\tau)}\tilde{\mu}_\tau((P_{s,\tau}f)^{p(\tau)}) - \tilde{\mu}_\tau((P_{s,\tau}f)^{p(\tau)})\log\|P_{s,\tau}f\|_{L^{p(\tau)}(\tilde{\mu}_\tau)}\Big\} \\ &\quad + e^{-N(\tau)}\frac{1}{p(\tau)}\|P_{s,\tau}f\|^{1-p(\tau)}_{L^{p(\tau)}(\tilde{\mu}_\tau)}\int |P_{s,\tau}f|^{p(\tau)}\frac{\mathrm{d}}{\mathrm{d}\tau}\tilde{\mu}_\tau \mathrm{d}x \\ &\leq C\frac{1}{p(\tau)}e^{-N(\tau)}\|P_{s,\tau}f\|_{L^{p(\tau)}(\tilde{\mu}_\tau)}. \end{aligned}$$

Hence, we obtain that

$$\|P_{s,\tau}f\|_{L^{p(\tau)}(\tilde{\mu}_\tau)} \leq e^{N(\tau)+C\int_s^\tau p(\tau')^{-1}\mathrm{d}\tau'}\|f\|_{L^2(\tilde{\mu}_s)} \leq e^{N(\tau)+C(\tau-s)}\|f\|_{L^2(\tilde{\mu}_s)}.$$

Here, notice that

$$\int_2^{p(s+t_0)}\frac{\gamma(\tau)}{\tau}\mathrm{d}\tau = \int_s^{s+t_0}\frac{p'(\tau)\gamma(p(\tau))}{p(\tau)}\mathrm{d}\tau = \int_s^{s+t_0} 1\mathrm{d}\tau = t_0.$$

But, since we also have $t_0 = \int_2^\infty \frac{\gamma(\tau)}{\tau}\mathrm{d}\tau$ by its definition, we must have $\lim_{\tau\to s+t_0} p(\tau) = \infty$ and hence

$$\lim_{\tau\to s+t_0} N(\tau) = \int_2^\infty \frac{\beta\left(\frac{4(p-1)\gamma(p)}{p}\right)}{p^2}\mathrm{d}p \leq \int_2^\infty \frac{\beta\left(2\gamma(p)\right)}{p^2}\mathrm{d}p.$$

This yields the first assertion. Applying Corollary 5, we can see that $\int_2^\infty \frac{\beta(2\gamma(p))}{p^2}\mathrm{d}p = \int_2^\infty \frac{\beta(1/p)}{p^2}\mathrm{d}p \lesssim \int_2^\infty \frac{1+\log(p)}{p^2}\mathrm{d}p =: C_0 < \infty$. This yields the second assertion. □

Let $p(t,y)$ be the density function with respect to $\tilde{\mu}_t$ that corresponds to $P_{0,t}$, i.e., $p(t,y) = \frac{\mathrm{d}\mu_t}{\mathrm{d}\tilde{\mu}_t}(y)$. Then, we have that

$$\left|\int (p(t+2t_0,y)-1)f(y)\mathrm{d}\tilde{\mu}_{t+2t_0}(\mathrm{d}y)\right| \leq |P_{t+t_0,t+2t_0}\cdot P_{0,t+t_0}(f-\tilde{\mu}_{t+t_0}f)| + |(\tilde{\mu}_{t+t_0}-\tilde{\mu}_{t+2t_0})f|$$
$$\leq c_0\|P_{0,t+t_0}f-\tilde{\mu}_{t+t_0}f\|_{L^2(\tilde{\mu}_{t+t_0})} + |(\tilde{\mu}_{t+t_0}-\tilde{\mu}_{t+2t_0})f|$$
$$\leq C_0\exp(-\alpha\lambda t)\|P_{0,t_0}f-\tilde{\mu}_{t_0}f\|_{L^2(\tilde{\mu}_{t_0})} + C\exp(-\alpha\lambda t)\|f\|_{L^2(\tilde{\mu}_\infty)} \quad (\because \text{Poincare inequality})$$
$$\leq \exp(-\alpha\lambda t)[C_0^2\|f\|_{L^2(\tilde{\mu}_0)} + C\|f\|_{L^2(\tilde{\mu}_\infty)}] \leq \exp(-\alpha\lambda t)(C_0^2+C)C\|f\|_{L^2(\tilde{\mu}_{t+2t_0})},$$

where we used that the density ratio between $\tilde{\mu}_\tau$ and $\tilde{\mu}_{\tau'}$ are bounded from above and below for any τ,τ' because of the boundedness of ℓ_i and h. Therefore, we have that

$$\|p(t+2t_0,y)-1\|_\infty \leq C'\exp(-2\alpha\lambda t)$$

for a constant C'. This means that μ_t and $\tilde{\mu}_t$ are mutually absolutely continuous for sufficiently large $t>0$ and the Radon-Nikodým density satisfies

$$\left\|\frac{\mathrm{d}\mu_t}{\mathrm{d}\tilde{\mu}_t}-1\right\|_\infty \leq C'\exp(-\alpha\lambda(t-2t_0)).$$

Importantly, this observation indicates that μ_t satisfies the $(\alpha/2)$-LSI for sufficiently large t such as $C'\exp(-\alpha\lambda(t-2t_0)) \leq \min\{\sqrt{2}-1, 1-1/\sqrt{2}\}$. In the following, we let such t be T_0.

Corollary 6. *Under Assumption 1, there exits T_0 depending on $d,B,\delta,\alpha,\lambda,c_r,C_r$ such that μ_t satisfies the $(\alpha/2)$-LSI condition for $t\geq T_0$. Moreover,*

$$\frac{1}{\sqrt{2}} \leq \left|\frac{\mathrm{d}\mu_t}{\mathrm{d}\tilde{\mu}_t}(x)\right| \leq \sqrt{2}, \quad \left\|\frac{\mathrm{d}\mu_t}{\mathrm{d}\tilde{\mu}_t}-1\right\|_\infty \leq C'\exp(-\alpha\lambda(t-T_0)),$$

for all $t\geq T_0$, where C' is a constant.

C Computation of Error Terms

C.1 First order differentiation

Let $H_t(x) = \frac{1}{n}\sum_{j=1}^n \ell_j'(f_{\mu_t})h_j(x)$. Here, we evaluate the first order derivative $d_i^{(1)}(t;\mu,x)$. For that purpose, we define an operator $d^{(1)}(t;\mu,\xi)(=d^{(1)}(t;\mu,\xi,x))$ where $\xi\in\mathbb{R}^d, x\in\mathbb{R}^d, \mu\in\mathcal{P}$ (x is omitted if there is no confusion) as

$$\begin{aligned} d^{(1)}(t;\mu,\xi,x)(\phi) &= \xi^\top\nabla\frac{\partial}{\partial\epsilon}m(t,\epsilon\delta_x+(1-\epsilon)\mu)(\phi)|_{\epsilon=0} \\ &= \xi^\top\nabla\frac{\delta}{\delta\mu}m(t,\mu)(\phi)(x). \end{aligned}$$

We can check that $d_i^{(1)}(t;\mu,x) = d^{(1)}(t;\mu,e_i,x)$ where e_i is the indicator vector that has 1 in its i-th coordinate and 0 in other coordinates. In our setting, we are interested in a setting $\mu_0 = \mu_s^N = \frac{1}{N}\sum_{i=1}^N \hat{X}_s^i$ which is a discrete measure with a support consisting of N points. In that case, $d^{(1)}(t;\mu_0,\xi,x)$ can be reformulated as a gradient flow as

$$\mathbb{E}_{\mu_0}\left[\tilde{\xi}(X_0)^\top\nabla\frac{\delta}{\delta\mu}m(t,\mu_0)(\phi)\right] = \frac{1}{N}\sum_{i=1}^N \tilde{\xi}(\hat{X}_0^i)^\top\nabla\frac{\delta}{\delta\mu}m(t,\mu_0)(\phi).$$

where $\tilde{\xi}(X_0) = N\xi$ when $X_0 = x$ and $\tilde{\xi}(X_0) = 0$ otherwise. Clearly, $\tilde{\xi}\in L^2(\mu_0)$. Accordingly, we define the following SDE:

$$\begin{aligned} &\mathrm{d}X_t^\epsilon = -b(X_t^\epsilon,\mu_t^\epsilon)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t, \\ &\mu_t^\epsilon = \mathrm{Law}(X_t^\epsilon), \quad X_0^\epsilon = X_0 + \epsilon\tilde{\xi}(X_0). \end{aligned}$$

Then, if we define $\tilde{v}_t^\xi := \lim_{\epsilon\to 0}\frac{X_t^\epsilon - X_t}{\epsilon}$, it holds that

$$d^{(1)}(t;\mu_0,\xi,x)(\phi) = \mathbb{E}[\tilde{v}_t^\xi\cdot\nabla\phi(X_t)].$$

We can see that the infinitesimal displacement $\tilde{v}_t^\xi$ follows the following differential equation:

$$\mathrm{d}\tilde{v}_t^\xi = -\left[\tilde{v}_t^\xi \cdot \nabla b(x,\mu_t)|_{x=X_t} + \frac{\delta}{\delta\mu} b(X_t,\mu_t)(\tilde{q}_t)\right]\mathrm{d}t, \tag{10}$$

and then we can check that $\tilde{v}_t^\epsilon$ has a bounded moment because

$$\begin{aligned}\frac{1}{p}\frac{\mathrm{d}\|\tilde{v}_t^\xi\|^p}{\mathrm{d}t} &= -\|\tilde{v}_t^\xi\|^{p-2}\left[\tilde{v}_t^\xi \cdot \nabla b(x,\mu_t)|_{x=X_t} \cdot \tilde{v}_t^\xi + \frac{\delta}{\delta\mu} b(X_t,\mu_t)(\tilde{q}_t)\cdot \tilde{v}_t^\xi\right] \\ &= -\|\tilde{v}_t^\xi\|^{p-2}\left[(\tilde{v}_t^\xi)^\top \nabla\nabla^\top H_t(X_t)\tilde{v}_t^\xi + \lambda(\tilde{v}_t^\xi)^\top\nabla\nabla^\top r(X_t)\tilde{v}_t^\xi + \frac{\delta}{\delta\mu} b(X_t,\mu_t)(\tilde{q}_t)\cdot\tilde{v}_t^\xi\right] \\ &\le B^2\|\tilde{v}_t^\xi\|^p + B^2\|\tilde{v}_t^\xi\|^{p-1}\mathbb{E}[\|\tilde{v}_t^\xi\|],\end{aligned}$$

where we used Assumption 1 and convexity of r in the last inequality. We notice that, when $X_0 \neq x$, $\tilde{v}_0^\xi = 0$ and then $\tilde{v}_t^\xi = 0$ in that situation. This also yields $\mathbb{E}[\|\tilde{v}_t^\xi\|] = \mathbb{E}[\|\tilde{v}_t^\xi\||X_0 = x]/N$. Therefore, when $X_0 = x$, we have that

$$\|\tilde{v}_t^\xi\|^p \le \mathcal{O}(\exp(B^2(t+1)))\|\tilde{v}_0^\xi\|^p = \mathcal{O}(\exp(B^2(t+1)))N^p\|\xi\|^p \tag{11}$$

for $p \ge 1$ and $t > 0$.

By Corollary 3, $m(t, \epsilon\delta_x - (1-\epsilon)\mu_0)$ has a smooth density for $t > 0$, which we denote by $\mu_t^{(\epsilon)}$. Corollary 3 also asserts that $\mu_t^{(\epsilon)}(x) > 0$ and equivalent to μ^* for any $t > 0$. Let

$$q_{t,x}^{(1)}(x') = \frac{1}{\mu_t(x')}\xi^\top\nabla\frac{\partial}{\partial\epsilon}\mu_t^{(\epsilon)}(x')|_{\epsilon=0}.$$

Here after, we use a notation of

$$q_{t,x}^{(1)}(\phi) := \mathbb{E}_{\mu_t}[q_{t,x}^{(1)}\phi].$$

We also a similar notation below for related quantity.

Lemma 6 (Bismut formula)**.** *Suppose Assumption 1 holds. For a bounded measurable function $f : \mathbb{R}^d \to \mathbb{R}$, $\mu_0 \in \mathcal{P}_2$ and $t > 0$, then $m(t, (I+\epsilon\xi)_\sharp\mu_0)(f)$ is differentiable with respect to ϵ at $t = 0$, and*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\epsilon} m(t,(I+\epsilon\xi)_\sharp\mu_0)(f)|_{\epsilon=0} &= \mathbb{E}[\nabla f(X_t)\cdot v_t(X_t)] \\ &= \mathbb{E}\left[f(X_t)\int_0^t \zeta_s^\xi \cdot \mathrm{d}W_s\right]\end{aligned}$$

where

$$\zeta_s^\xi = \frac{(\sqrt{2\lambda})^{-1}}{t}\left(\tilde{v}_s^\xi + s\frac{\delta}{\delta\mu}b(X_s,\mu_s)(\tilde{q}_s)\right).$$

In particular, $q_{t,x}^{(1)}(x') = \mathbb{E}[\int_0^t \zeta_s^\xi \cdot \mathrm{d}W_s | X_t = x']$ and

$$\mathbb{E}[(q_{t,x}^{(1)}(X_t))^2|X_0 = x] < K(t)N^2,$$

where $K(t)$ is a constant depending on t, and $\mathbb{E}[(q_{t,x}^{(1)}(X_t))^2|X_0 = x'] = 0$ for $x' \neq x$.

Proof. The first assertion is obtained by the Bismut formula with respect to the Lions derivative by the initial distribution μ_0 (Ren and Wang, 2019). Indeed, Theorem 2.1 of Ren and Wang (2019) yields the assertion by setting $g_s = s/t$ in their notation. We note that they assumed $b(x,\mu)$ and their derivatives are bounded, however, their argument can be applied to our setting because $r(x)$ is a convex function and forces X_t contracts to origin instead of make it diverge.

As for the second assertion, first we note that $\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathbb{E}[f(X_t)] = \int f(x)\frac{\partial_\epsilon\mu_t(x)}{\mu_t(x)}\mathrm{d}\mu_t = \int f(x)\partial_\epsilon\log\mu_t(x)\mathrm{d}\mu_t$ for all f (note that $\mu_t(x) > 0$ for all $x \in \mathbb{R}^d$ by Corollary 3). This indicates that $\partial_\epsilon\log(\mu_t)(x) = \mathbb{E}[\int_0^t \zeta_t^\xi \cdot \mathrm{d}W_s | X_t = x]$ almost surely. The moment bound can be ensured by noticing that

$$\mathbb{E}[q_{t,x}^{(1)}(X_t)^2|X_0 = x] = \mathbb{E}\left[\mathbb{E}\left[\int_0^t \zeta_s^\xi\cdot\mathrm{d}W_s\middle|X_t\right]^2|X_0 = x\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left(\int_0^t \zeta_t^\xi \cdot \mathrm{d}W_s\right)^2 |X_0 = x\right] \\
&\leq C_{t,2}\int_0^t \mathbb{E}[\|\zeta_s^\xi\|^2|X_0 = x]\mathrm{d}s \\
&\lesssim \frac{C_{t,2}}{t^2}\int_0^t (1+s)^2\mathbb{E}[\|\tilde{v}_s^\xi\|^2|X_0 = x]\mathrm{d}s \\
&\lesssim \frac{C_{t,2}}{t^2}(1+t)^3N^2\|\xi\|^2 = O\left(\frac{C_{t,2}(1+t)^3}{t^2}N^2\right),
\end{aligned}
$$

where the second inequality is by the moment inequality of stochastic integral (Kim, 2013) ($C_{t,q} = (\frac{q(q-1)}{2})^{q/2}t^{\frac{q-2}{2}}$) and the last inequality is due to Eq. (11) and Assumption 1. When $X_0 \neq x$, we notice that $\|\tilde{v}_s^\xi\| = 0$ yielding the final remark. □

Let

$$
\hat{L}_t\phi = \lambda\Delta\phi - b_t^\top\nabla\phi
$$

for $\phi : \mathbb{R}^d \to \mathbb{R}$. Then, the derivative of $q_{t,x}^{(1)}$ with respect to t can be evaluated as

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\mu_t}[q_{t,x}^{(1)}f] &= \frac{\partial}{\partial t}\frac{\partial}{\partial\epsilon}\mathbb{E}[f(X_t^\epsilon)]|_{\epsilon=0} \\
&= \frac{\partial}{\partial\epsilon}\frac{\partial}{\partial t}\mathbb{E}[f(X_t^\epsilon)]|_{\epsilon=0} \\
&= \frac{\partial}{\partial\epsilon}\mathbb{E}[(\lambda\Delta - b_t^\top\nabla)f(X_t^\epsilon)]|_{\epsilon=0} \\
&= \int q_{t,x}^{(1)}(\lambda\Delta - b_t^\top\nabla)f - \frac{\delta b_t}{\delta\mu}(q_{t,x}^{(1)})\nabla f\mathrm{d}\mu_t.
\end{aligned}
$$

See Tse (2021) for higher order regularity of the nonlinear PDEs induced by the derivative with respect to the initial distribution in the torus setting. Therefore, by applying Theorem 4, we obtain the following convergence bound.

Lemma 7. *Under Assumption 1, it holds that*

$$
\begin{aligned}
\mathbb{E}[(q_{t,x}^{(1)}(X_t))^2] &\leq \mathcal{O}(\Lambda_{\mu_0}\exp(-\lambda\alpha(t-T_0)/2)\mathbb{E}[(q_{0,x}^{(1)}(X_0))^2]) \\
&= \mathcal{O}(\Lambda_{\mu_0}\exp(-\lambda\alpha(t-T_0)/2)N\|\xi\|^2).
\end{aligned}
$$

Proof. We apply Theorem 4. We notice that the condition of Theorem 4 holds for $\epsilon_t = C\exp(-\alpha\lambda t)W_2(\mu_0,\mu^*)$ by Corollary 4 and $\delta_t = 0$. Moreover, by Corollary 1, we can assume $\alpha_t = \alpha/2$ for $t \geq T_0$ and $\alpha_t = 0$ for $0 \leq t < T_0$. Next, we check the integrability of $\sqrt{I(\mu_t||\tilde{\mu}_t)}$ with respect to t. According to Nitanda et al. (2022); Chizat (2022), we know that

$$
\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*)) \leq -\lambda^2 I(\mu_t||\tilde{\mu}_t) \leq 0,
$$

which indicates that

$$
\int_{t'}^t I(\mu_t||\tilde{\mu}_t)\mathrm{d}t \leq \frac{1}{\lambda^2}(\mathcal{L}(\mu_{t'}) - \mathcal{L}(\mu^*) - (\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*))).
$$

Hence,

$$
\int_{t'}^t \sqrt{I(\mu_t||\tilde{\mu}_t)}\mathrm{d}t \leq \sqrt{\frac{(t-t')}{\lambda^2}(\mathcal{L}(\mu_{t'}) - \mathcal{L}(\mu_t))} \leq \mathcal{O}\left(\frac{1}{\lambda}\sqrt{t-t'}\exp(-\lambda\alpha(t'-\tau_0))\sqrt{\mathcal{L}(\mu_{\tau_0}) - \mathcal{L}(\mu^*)}\right).
$$

By taking $t = k+1+\tau_0$ and $t' = k+\tau_0$ for $k = 0, 1, \ldots$ and taking summation over k, we obtain

$$
\int_{\tau_0}^\infty \sqrt{I(\mu_t||\tilde{\mu}_t)}\mathrm{d}t \lesssim \mathcal{O}\left(\sqrt{\mathcal{L}(\mu_{\tau_0}) - \mathcal{L}(\mu^*)}\right) = \mathcal{O}(1+\sqrt{\psi_2(\tau_0)}\mathcal{W}_2(\mu_0,\mu^*)) = \mathcal{O}_p(1), \quad (12)
$$

where we omitted the dependence on λ, α in the order symbol. Therefore, Theorem 4 yields that

$$
\mathbb{E}[(q_{t,x}^{(1)}(X_t))^2] \leq \mathcal{O}(\Lambda_{\mu_0}\exp(-\lambda\alpha(t-T_0)/2)\mathbb{E}[(q_{t,x}^{(1)}(X_0))^2]),
$$

where $\Lambda_{\mu_0} = \exp(\mathcal{O}(\mathcal{W}_2(\mu_0,\mu^*)))$. This yields the assertion. □

We finally remark that, when $t > t_0$, for any $\phi \in C_b(\mathbb{R}^d)$, Eq. (11) yields that

$$\begin{aligned}
d^{(1)}(t;\mu_0,\xi,x)(\phi) &= \mathbb{E}_{\mu_t}[q^{(1)}_{t,x}\phi] = \frac{1}{N}\mathbb{E}[q^{(1)}_{t,x}(X_t)\phi(X_t)|X_0 = x] \\
&\leq \frac{1}{N}\sqrt{\mathbb{E}[(q^{(1)}_{t,x}(X_t))^2|X_0 = x]\|\phi\|^2_\infty} \\
&\leq \frac{1}{\sqrt{N}}\sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}[(q^{(1)}_{t,x}(X_t))^2|X_0 = \hat{X}^i_s]\|\phi\|^2_\infty} \\
&\quad (\because \mathbb{E}[(q^{(1)}_{t,x}(X_t))^2|X_0 = x'] = 0 \text{ for } x \neq x') \\
&= \frac{1}{\sqrt{N}}\sqrt{\mathbb{E}[(q^{(1)}_{t,x}(X_t))^2]\|\phi\|^2_\infty} \\
&\leq \mathcal{O}\left(\Lambda_{\mu_0}\exp(-\lambda\alpha(t-T_0)/2)\|\xi\|\|\phi\|_\infty\right).
\end{aligned}$$

Moreover, it holds that

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N d^{(1)}(t;\mu_0,\xi,\hat{X}^i_s)(\phi) &= \frac{1}{N}\sum_{i=1}^N \frac{1}{N}\mathbb{E}[q^{(1)}_{t,\hat{X}^i_s}(X_t)\phi(X_t)|X_0 = \hat{X}^i_s] \\
&\leq \frac{1}{N}\sum_{i=1}^N \frac{1}{N}\sqrt{\mathbb{E}[q^{(1)}_{t,\hat{X}^i_s}(X_t)^2|X_0 = \hat{X}^i_0]\mathbb{E}[\phi(X_t)^2|X_0 = \hat{X}^i_s]} \\
&\leq \frac{1}{N}\sqrt{\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}[q^{(1)}_{t,\hat{X}^i_s}(X_t)^2|X_0 = \hat{X}^i_s]\right)\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\phi(X_t)^2|X_0 = \hat{X}^i_s]\right)} \\
&= \frac{1}{N}\sqrt{\left(\sum_{i=1}^N \mathbb{E}[q^{(1)}_{t,\hat{X}^i_s}(X_t)^2]\right)\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\phi(X_t)^2|X_0 = \hat{X}^i_s]\right)} \\
&= \mathcal{O}\left(\Lambda_{\mu_0}\exp(-\lambda\alpha(t-T_0)/2)\|\xi\|\sqrt{\mathbb{E}[\phi^2(X_t)]}\right). \qquad (13)
\end{aligned}$$

In the following, we write

$$q^{(1)}_t := \frac{1}{N}\sum_{i=1}^N q^{(1)}_{t,\hat{X}^i_s}.$$

C.2 Second order differentiation

Here, we evaluate the second order derivatives. Let $x_1, x_2 \in \mathbb{R}^d$ fixed. For $\epsilon = (\epsilon_1, \epsilon_2)$ with $\epsilon_k \geq 0$ and $\xi^{[1]}, \xi^{[2]} \in \mathbb{R}^d$, we note that

$$\begin{aligned}
&\xi^{[1]\top}\nabla_{x_1}\nabla^\top_{x_2}\frac{\delta^2}{\delta\mu^2}\mathcal{U}(t,\mu_0)(x_1,x_2)\xi^{[2]} \\
=&\xi^{[1]\top}\nabla_{x_1}\nabla^\top_{x_2}\frac{\partial^2}{\partial\epsilon_1\partial\epsilon_2}\Phi(m(t,\epsilon_1\delta_{x_1}+\epsilon_2\delta_{x_2}-(1-\epsilon_1-\epsilon_2)\mu_0))|_{\epsilon=(0,0)}\xi^{[2]} \\
=&\xi^{[2]\top}\nabla_{x_2}\frac{\partial}{\partial\epsilon_2}\frac{\delta}{\delta m}\Phi(m(t,\epsilon_2\delta_{x_2}+(1-\epsilon_2)\mu_0)(d^{(1)}(t;\epsilon_2\delta_{x_2}+(1-\epsilon_2)\mu_0,\xi^{[1]})) \\
=&\frac{\delta^2}{\delta m^2}\Phi(m(t,\mu_0))(d^{(1)}(t;\mu_0,\xi^{[1]}),d^{(1)}(t;\mu_0,\xi^{[2]})) + \frac{\delta}{\delta m}\Phi(m(t,\mu_0))(d^{(2)}(t;\mu_0,\xi^{[1]},\xi^{[2]},x_1,x_2)),
\end{aligned}$$

where

$$\begin{aligned}
&d^{(2)}(t;\mu_0,\xi^{[1]},\xi^{[2]},x_1,x_2)(\phi) \\
&= \xi^{[1]}\nabla_{x_1}\nabla^\top_{x_2}\frac{\partial^2}{\partial\epsilon_1\partial\epsilon_2}m(t,\epsilon_1\delta_{x_1}+\epsilon_2\delta_{x_2}-(1-\epsilon_1-\epsilon_2)\mu_0))(\phi)|_{\epsilon=(0,0)}\xi^{[2]}.
\end{aligned}$$

By Corollary 3, $m(t, \epsilon_1\delta_{x_1} + \epsilon_2\delta_{x_2} - (1-\epsilon_1-\epsilon_2)\mu_0)$ has a smooth density for $t > 0$, which is denoted by $\mu_t^{(\epsilon_1,\epsilon_2)}$. Corollary 3 also asserts that $\mu_t^{(\epsilon_1,\epsilon_2)}(x) > 0$ and equivalent to μ^* for any $t > 0$. If we write

$$q^{(2)}_{t,(x_1,x_2)}(x) = \frac{1}{\mu_t(x)}\xi^{[1]\top}\nabla\nabla^\top \frac{\partial^2}{\partial\epsilon_1\partial\epsilon_2}\mu_t^{(\epsilon_1,\epsilon_2)}(x)\xi^{[2]},$$

then we can notice that $d^{(2)}(t;\mu_0,\xi^{[1]},\xi^{[2]},x_1,x_2)(\phi) = \mathbb{E}_{\mu_t}[q^{(2)}_{t,(x_1,x_2)}\phi]$. Here, we denote by $d^{(1)}_{t,[k]}(\phi) = d^{(1)}(t;\mu_0,\xi^{[k]},x_k)(\phi) = \xi^\top\nabla_x\frac{\partial}{\partial\epsilon}m(t,\epsilon\delta_{x_k} - (1-\epsilon)\mu_0))(\phi)|_{\epsilon=0}$. Then, by taking the derivative of $\frac{\mathrm{d}}{\mathrm{d}t}d^{(1)}(t;\cdot)$ with respect to ϵ_2, then $q_t^{(2)}$ (and $d^{(2)}$) follows the following dynamics:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}d^{(2)}(t;\mu_0,\xi^{(1)},\xi^{(2)},x_1,x_2)(\phi) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\mu_t}[q^{(2)}_{t,(x_1,x_2)}\hat{L}_t(\phi)]\\
=&\mathbb{E}_{\mu_t}[q^{(2)}_{t,(x_1,x_2)}\hat{L}_t(\phi)] - \int\frac{\delta b_t}{\delta\mu}(d^{(2)}(t;\mu_0,\xi^{(1)},\xi^{(2)},x_1,x_2))\cdot\nabla\phi\mathrm{d}\mu_t\\
&- d^{(1)}_{t,[1]}\left(\frac{\delta b_t}{\delta\mu}(d^{(1)}_{t,[2]})\cdot\nabla\phi\right) - d^{(1)}_{t,[2]}\left(\frac{\delta b_t}{\delta\mu}(d^{(1)}_{t,[1]})\cdot\nabla\phi\right) - \mathbb{E}_{\mu_t}\left[\frac{\delta^2 b_t}{\delta^2\mu}(d^{(1)}_{t,[1]},d^{(1)}_{t,[2]})\cdot\nabla\phi\right],
\end{aligned}$$

and

$$q^{(2)}_{0,(x_1,x_2)}(\cdot) = 0.$$

See also Delarue and Tse (2021); Arnaudon and Del Moral (2020) for related calculations and their technical details.

Now, let

$$q_t^{(2)} = \frac{1}{N}\sum_{i=1}^N q_{t,(\hat{X}_s^i,\hat{X}_s^i)},$$

and consider a situation where $x_1 = x_2 \in \{\hat{X}_0^i\}_{i=1}^N$. We write $q^{(1)}_{t,[k],i}$ to indicate $q^{(1)}_{t,\hat{X}_t^i}$ for $d^{(1)}(t;\mu_0,\xi^{[k]},\hat{X}_0^i)$ when $x_k = \hat{X}_0^i$. Then, we have that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\mu_t}[q_t^{(2)}\hat{L}_t(\phi)]\\
=&\mathbb{E}_{\mu_t}[q_t^{(2)}\hat{L}_t(\phi)] - \int\frac{\delta b_t}{\delta\mu}(q_t^{(2)})\cdot\nabla\phi\mathrm{d}\mu_t\\
&- \frac{1}{N}\sum_{i=1}^N\mathbb{E}_{\mu_t}\left[q^{(1)}_{t,[1],i}\frac{\delta b_t}{\delta\mu}(q^{(1)}_{t,[2],i})\cdot\nabla\phi\right] - \frac{1}{N}\sum_{i=1}^N\mathbb{E}_{\mu_t}\left[q^{(1)}_{t,[2],i}\frac{\delta b_t}{\delta\mu}(q^{(1)}_{t,[1],i})\cdot\nabla\phi\right]\\
&- \mathbb{E}_{\mu_t}\left[\frac{\delta^2 b_t}{\delta^2\mu}(q^{(1)}_{t,[1],i},q^{(1)}_{t,[2],i})\cdot\nabla\phi\right].
\end{aligned}$$

Then, by Eq. (13)

$$\begin{aligned}
&-\frac{1}{N}\sum_{i=1}^N\mathbb{E}_{\mu_t}\left[q^{(1)}_{t,[1],i}\frac{\delta b_t}{\delta\mu}(q^{(1)}_{t,[2],i})\cdot\nabla\phi\right] \le \mathcal{O}\left(\Lambda_{\mu_0}\exp(-\lambda\alpha(t-T_0)/2)\|\xi^{[1]}\|\sqrt{\mathbb{E}_{\mu_t}[\|\nabla\phi\|^2]}\right)\\
&-\frac{1}{N}\sum_{i=1}^N\mathbb{E}_{\mu_t}\left[q^{(1)}_{t,[2],i}\frac{\delta b_t}{\delta\mu}(q^{(1)}_{t,[1],i})\cdot\nabla\phi\right] \le \mathcal{O}\left(\Lambda_{\mu_0}\exp(-\lambda\alpha(t-T_0)/2)\|\xi^{[2]}\|\sqrt{\mathbb{E}_{\mu_t}[\|\nabla\phi\|^2]}\right).
\end{aligned}$$

In the same vein, we also have

$$-\mathbb{E}_{\mu_t}\left[\frac{\delta^2 b_t}{\delta^2\mu}(q^{(1)}_{t,[1],i},q^{(1)}_{t,[2],i})\cdot\nabla\phi\right] \le \mathcal{O}\left(\Lambda_{\mu_0}^2\exp(-\lambda\alpha(t-T_0))\|\xi^{[1]}\|\|\xi^{[2]}\|\sqrt{\mathbb{E}_{\mu_t}[\|\nabla\phi\|^2]}\right).$$

Then applying Theorem 4, we obtain the following lemma.

Lemma 8. *Under Assumption 1, it holds that*

$$\mathbb{E}[(q_t^{(2)}(X_t))^2] \le \mathcal{O}(\Lambda_{\mu_0}\exp(-\lambda\alpha(t-T_0)/2)).$$

where $\Lambda_{\mu_0} = \exp(\mathcal{O}(W_2(\mu_0,\mu^*)))$.

Proof. Due to the argument above, the assumption in Theorem 4 guaranteed for $\delta_t = \exp(-\lambda t/2)$. The other conditions are also satisfied as in the proof of Lemma 7. Hence, for arbitrary small positive time $\tau_0' > 0$, we have that

$$\mathbb{E}[(q^{(2)}_{\tau_0'}(X_{\tau_0'}))^2] = \mathcal{O}(1 + \mathbb{E}[(q^{(2)}_0(X_0))^2]) = \mathcal{O}(1),$$

because of Theorem 4-(i) and $q^{(2)}_0 = 0$.

On the other hand, Corollary 4 yields that $\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \lambda\psi_2(\tau_0)W_2(\mu_0, \mu^*)^2$. Combining this with the argument in Eq. (12), we see that

$$\int_{\tau_0'}^{\infty} \sqrt{I(\mu_t, \tilde{\mu}_t)} \leq \mathcal{O}(1 + \sqrt{\psi_2(\tau_0')}\mathcal{W}_2(\mu_0, \mu^*)).$$

Then, by the same argument as Lemma 7 and Theorem 4-(ii), we obtain the assertion. □

C.3 General convergence guarantee

We can see that $q_t = q_t^{(1)}$ and $q_t = q_t^{(2)}$ satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\mu_t}[q_t\phi] = \mathbb{E}[q_t\hat{L}_t(\phi)] - \int \frac{\delta b_t}{\delta\mu}(q_t) \cdot \nabla\phi \mathrm{d}\mu_t + r_t(\phi), \tag{14}$$

where $|r_t(\phi)| \leq C\exp(-c_0\lambda\alpha t)\sqrt{\mathbb{E}_{\mu_t}[\|\nabla\phi\|^2]}$ and by abuse of notation we write $q_t(\phi) := \mathbb{E}_{\mu_t}[q_t\phi]$.

We define

$$D_1(t) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_{\mu_t}[q_t h_j]^2 \ell''_{j,t}, \quad D_2(t) := \int q_t^2 \mathrm{d}\mu_t,$$

where $\ell''_{j,t} = \ell''_j(f_{\mu_t})$. Then, we have the following bound on D_1 and D_2.

Theorem 4. *Assume Assumption 1 and suppose that $|\frac{\mathrm{d}}{\mathrm{d}t}\ell''_{j,t}| \leq \epsilon_t$, q_t satisfies Eq. (14) with $r_t(q_t)$ satisfying $|r_t(q_t)| \leq C\sqrt{\delta_t\mathbb{E}_{\mu_t}[\|\nabla q_t\|^2]}$ with a sequence $1/2 \geq \delta_t \geq 0$ and $C \geq 0$, and μ_t satisfies α_t-PI for $\alpha_t \geq 0$ ($\alpha_t = 0$ is also allowed in case that μ_t does not satisfy the LSI), then the following bounds hold:*

(i)

$$D_2(t) \leq \exp\left[t\left(\frac{B^4}{\lambda} + \frac{C^2}{2\lambda}\right)\right]\left(D_2(0) + \frac{C^2}{2\left(\frac{B^4}{\lambda} + \frac{C^2}{2\lambda}\right)}\right).$$

(ii)

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}(D_1(t) + \lambda D_2(t)) \\ &\leq -2(1-\delta_t)\lambda\alpha_t(D_1(t) + \lambda D_2(t)) + B^2\left(2B\lambda\sqrt{I(\mu_t||\tilde{\mu}_t)} + \epsilon_t + 2B^2\frac{\delta_t}{1-\delta_t}\right)D_2(t) + \frac{C^2}{2}\delta_t.\end{aligned}$$

In particular, it holds that, for $0 \leq \tau \leq t$,

$$D_1(t) + \lambda D_2(t) \leq \int_{\tau}^{t} \frac{C^2}{2}\delta_s e^{A_t - A_s}\mathrm{d}s + \exp(A_t - A_\tau)(D_1(\tau) + \lambda D_2(\tau)),$$

where

$$A_s = \int_0^s -2(1-\delta_s)\lambda\alpha_s + C_1\left(\lambda\sqrt{I(\mu_s||\tilde{\mu}_s)} + \epsilon_s + \frac{\delta_s}{1-\delta_s}\right)\mathrm{d}s,$$

and $C_1 = \max\{2B^3, 1, 2B^4\}/\lambda$.

Proof. By substituting $\phi \leftarrow q_t$, it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int (q_t)^2 \mathrm{d}\mu_t = 2\mathbb{E}_{\mu_t}[q_t \hat{L}_t(q_t)] - \int q_t^2 \frac{\partial}{\partial t}\mathrm{d}\mu_t - 2\int \frac{\delta b_t}{\delta \mu}(q_t)\cdot \nabla q_t \mathrm{d}\mu_t + 2r_t(q_t).$$

Here, the first two terms in the right hand side can be evaluated as

$$\begin{aligned}
&2\mathbb{E}_{\mu_t}[q_t \hat{L}_t(q_t)] - \int (q_t)^2 \frac{\partial}{\partial t}\mathrm{d}\mu_t \\
=&2\mathbb{E}_{\mu_t}[q_t(\lambda\Delta - b_t^\top \nabla)(q_t)] - \int (\lambda\Delta - b_t^\top \nabla)(q_t)^2 \mathrm{d}\mu_t \\
=&2\mathbb{E}_{\mu_t}[q_t(\lambda\Delta - b_t^\top \nabla)(q_t)] - 2\mathbb{E}_{\mu_t}[\lambda(\Delta q_t)q_t + \lambda\|\nabla q_t\|^2 - b_t^\top(\nabla q_t)q_t] \\
=&-2\lambda\mathbb{E}_{\mu_t}[\|\nabla q_t\|^2].
\end{aligned}$$

(1) By the assumption, the Cauchy-Schwarz inequality, and the arithmetic-geometric mean inequality, we can see that

$$\begin{aligned}
&-2\int \frac{\delta b_t}{\delta \mu}(q_t)\cdot \nabla q_t \mathrm{d}\mu_t + 2r_t(q_t) \\
\leq& \frac{B^4}{\lambda}\mathbb{E}_{\mu_t}[q_t^2] + \lambda\mathbb{E}_{\mu_t}[\|\nabla q_t\|^2] + \frac{C^2}{\lambda}\delta_t + \lambda\mathbb{E}_{\mu_t}[\|\nabla q_t\|^2] \quad = 2\lambda\mathbb{E}_{\mu_t}[\|\nabla q_t\|^2] + \frac{B^4}{\lambda}\mathbb{E}_{\mu_t}[q_t^2] + \frac{C^2}{\lambda}\delta_t.
\end{aligned}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}D_2(t) \leq \frac{B^4}{\lambda}D_2(t) + \frac{C^2}{\lambda}\delta_t.$$

This yields that

$$\begin{aligned}
D_2(t) &\leq \exp\left(t\frac{B^4}{\lambda} + \frac{C^2}{\lambda}\int_0^t \delta_s \mathrm{d}s\right) D_2(0) + \frac{C^2}{2}\int_0^t \delta_s \exp\left((t-s)\frac{B^4}{\lambda} + \frac{C^2}{\lambda}\int_s^t \delta_\tau \mathrm{d}\tau\right)\mathrm{d}s \\
&\leq \exp\left[t\left(\frac{B^4}{\lambda} + \frac{C^2}{2\lambda}\right)\right] D_2(0) + \frac{C^2}{2\left(\frac{B^4}{\lambda} + \frac{C^2}{2\lambda}\right)}\exp\left[t\left(\frac{B^4}{\lambda} + \frac{C^2}{2\lambda}\right)\right].
\end{aligned}$$

This gives the first inequality.

(2) Next, we evaluate the time differentiation of D_1 as

$$\begin{aligned}
&n\frac{\mathrm{d}}{\mathrm{d}t}D_1(t) \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\sum_{j=1}^n (\mathbb{E}_{\mu_t}[q_t h_j])^2 \ell''_{j,t} \\
&= \sum_{j=1}^n \mathbb{E}_{\mu_t}[q_t h_j]\left\{2\ell''_{j,t}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\mu_t}[q_t h_j] + \mathbb{E}_{m_t}[q_t h_j]\frac{\mathrm{d}}{\mathrm{d}t}\ell''_{j,t}\right\} \\
&= 2\sum_{j=1}^n \mathbb{E}_{m_t}[q_t h_j]\ell''_{j,t}\mathbb{E}_{\mu_t}\left[q_t(\lambda\Delta h_j - b_t^\top \nabla h_j) + \frac{\delta b_t^\top}{\delta\mu}(q_t)\nabla h_j\right] + B^2 n\epsilon_t D_2(t) \\
&= -2\lambda\sum_{j=1}^n \mathbb{E}_{m_t}[q_t h_j]\ell''_{j,t}\int \nabla q_t \cdot \nabla h_j \mathrm{d}\mu_t \\
&\quad + \lambda\sum_{j=1}^n \mathbb{E}_{\mu_t}[q_t h_j]\ell''_{j,t}\int q_t(b_t^\top + \lambda\nabla\log(\mu_t))\cdot \nabla h_j \mathrm{d}\mu_t \\
&\quad - \frac{2}{n}\|(\ell''_{i,t}\mathbb{E}_{\mu_t}[q_t h_j])_{j=1}^n\|_{Q_t}^2 + B^2 n\epsilon_t D_2(t) \qquad \left(\text{where } (Q_t)_{i,j} := \int \nabla h_i \cdot \nabla h_j \mathrm{d}\mu_t\right).
\end{aligned}$$

Here, we notice that

$$\sum_{j=1}^n \mathbb{E}_{\mu_t}[q_t h_j]\ell''_{j,t}\int q_t(b_t + \lambda\nabla\log(\mu_t))\cdot \nabla h_j \mathrm{d}\mu_t$$

$$
\begin{aligned}
&\leq \lambda \sum_{j=1}^{n} \sqrt{\int q_t^2 \mathrm{d}\mu_t} \sqrt{\int \|\nabla \log(\tilde{\mu}_t) - \nabla \log(\mu_t)\|^2 \mathrm{d}\mu_t} \sqrt{\int q_t^2 \mathrm{d}\mu_t} \ell''_{j,t} \|\nabla h_j\|_\infty \|h_j\|_\infty \\
&\leq \lambda n B^3 \sqrt{I(\mu_t||\tilde{\mu}_t)} \mathbb{E}_{\mu_t}[q_t^2] = \lambda n B^3 \sqrt{I(\mu_t||\tilde{\mu}_t)} D_2(t).
\end{aligned}
$$

We also notice that

$$
\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{\mu_t}[q_t h_j] \ell''_{j,t} \int \nabla q_t \cdot \nabla h_j \mathrm{d}x = \int \nabla q_t \cdot \frac{\delta b_t}{\delta \mu}(q_t) \mathrm{d}x.
$$

Therefore,

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}(D_1(t) + \lambda D_2(t)) \\
\leq& -\frac{2}{n^2} \|(\ell''_{i,t} \mathbb{E}_{\mu_t}[q_t h_j])_{j=1}^n\|_{Q_t}^2 - 4\lambda \int \nabla q_t \cdot \frac{\delta b_t}{\delta \mu}(q_t) \mathrm{d}\mu_t - 2\lambda^2 \int \|\nabla q_t\|^2 \mathrm{d}\mu_t \\
&+ B^3 \lambda \sqrt{I(\mu_t||\tilde{\mu}_t)} D_2(t) + B^2 \epsilon_t D_2(t) + 2\lambda r_t(q_t) \\
\leq& -\frac{2}{n^2} \|(\ell''_{i,t} \mathbb{E}_{\mu_t}[q_t h_j])_{j=1}^n\|_{Q_t}^2 - 4\lambda \int \nabla q_t \cdot \frac{\delta b_t}{\delta \mu}(q_t) \mathrm{d}\mu_t - 2\lambda^2 \int \|\nabla q_t\|^2 \mathrm{d}\mu_t \\
&+ 2B^3 \lambda \sqrt{I(\mu_t||\tilde{\mu}_t)} D_2(t) + B^2 \epsilon_t D_2(t) + 2\delta_t \lambda^2 \int \|\nabla q_t\|^2 \mathrm{d}\mu_t + \frac{C^2}{2} \delta_t \\
=& -2 \int \left\| \frac{1}{\sqrt{1-\delta_t}} \frac{\delta b_t}{\delta \mu}(q_t) + \sqrt{1-\delta_t} \lambda \nabla q_t \right\|^2 \mathrm{d}\mu_t + 2\left(\frac{1}{1-\delta_t} - 1\right) \int \left\| \frac{\delta b_t}{\delta \mu}(q_t) \right\|^2 \mathrm{d}\mu_t \\
&+ B^2 (2B\lambda \sqrt{I(\mu_t||\tilde{\mu}_t)} + \epsilon_t) D_2(t) + \frac{C^2}{2} \delta_t \\
\leq& -2 \int \left\| \frac{1}{\sqrt{1-\delta_t}} \frac{\delta b_t}{\delta \mu}(q_t) + \sqrt{1-\delta_t} \lambda \nabla q_t \right\|^2 \mathrm{d}\mu_t \\
&+ B^2 \left(2B\lambda \sqrt{I(\mu_t||\tilde{\mu}_t)} + \epsilon_t + 2B^2 \frac{\delta_t}{1-\delta_t} \right) D_2(t) + \frac{C^2}{2} \delta_t.
\end{aligned}
$$

When μ_t satisfies α_t-PI (which is implied by α_t-LSI), it holds that

$$
\alpha_t \int \left[\theta_1 \frac{1}{n} \sum_{j=1}^{n} \ell''_{j,t} h_j \mathbb{E}_{\mu_t}[q_t h_j] + \theta_2 \lambda q_t \right]^2 \mathrm{d}\mu_t \leq \int \left\| \theta_1 \frac{\delta b_t}{\delta \mu}(q_t) + \theta_2 \lambda \nabla q_t \right\|^2 \mathrm{d}\mu_t
$$

for any $\theta_1, \theta_2 > 0$, which gives that

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}(D_1(t) + \lambda D_2(t)) \\
\leq& -2\alpha_t \int \left[\frac{1}{\sqrt{1-\delta_t}} \frac{1}{n} \sum_{j=1}^{n} \ell''_{j,t} h_j \mathbb{E}_{\mu_t}[q_t h_j] + \sqrt{1-\delta_t} \lambda q_t \right]^2 \mathrm{d}\mu_t \\
&+ B^2 \left(2B\lambda \sqrt{I(\mu_t||\tilde{\mu}_t)} + \epsilon_t + 2B^2 \frac{\delta_t}{1-\delta_t} \right) D_2(t) + \frac{C^2}{2} \delta_t \\
\leq& -2\alpha_t \frac{1}{1-\delta_t} \frac{1}{n^2} \sum_{i,j=1}^{n} \ell''_{i,t} \ell''_{j,t} \mathbb{E}_{\mu_t}[q_t h_i] \mathbb{E}_{\mu_t}[q_t h_j] \int h_i h_j \mathrm{d}\mu_t \\
&- 4\lambda \alpha_t \frac{1}{n} \sum_{j=1}^{n} \ell''_{j,t} \mathbb{E}_{\mu_t}[q_t h_j] \mathbb{E}_{\mu_t} E[q_t h_j] - (1-\delta_t) 2\alpha_t \lambda^2 \int q_t^2 \mathrm{d}\mu_t \\
&+ B^2 \left(2B\lambda \sqrt{I(\mu_t||\tilde{\mu}_t)} + \epsilon_t + 2B^2 \frac{\delta_t}{1-\delta_t} \right) D_2(t) + \frac{C^2}{2} \delta_t \\
\leq& -2(1-\delta_t) \lambda \alpha_t (D_1(t) + \lambda D_2(t))
\end{aligned}
$$

$$+ B^2 \left(2B\lambda \sqrt{I(\mu_t || \tilde{\mu}_t)} + \epsilon_t + 2B^2 \frac{\delta_t}{1-\delta_t} \right) D_2(t) + \frac{C^2}{2} \delta_t.$$

Then, by noticing $\lambda D_2(t) \leq D_1(t) + \lambda D_2(t)$ and the Grönwall's inequality (Mischler, 2019), we arrive at

$$D_1(t) + \lambda D_2(t) \leq \int_0^t \frac{C^2}{2} \delta_s e^{A_t - A_s} \mathrm{d}s + \exp(A_t)(D_1(0) + \lambda D_2(0)),$$

where

$$A_s = \int_0^s -2(1-\delta_s)\lambda\alpha_s + C_1 \left(\lambda \sqrt{I(\mu_s || \tilde{\mu}_s)} + \epsilon_s + \frac{\delta_s}{1-\delta_s} \right) \mathrm{d}s,$$

for a constant $C_1 = \max\{2B^3, 1, 2B^4\}/\lambda$. □

C.4 Combining all bounds together

Remember that it holds that

$$\mathbb{E}[\mathcal{U}(0, \mu_t^N) - \mathcal{U}(t, \mu_0^N)] = \frac{1}{N} \sum_{i=1}^d \int_0^t \mathbb{E} \left[\int \left(\partial_{(x_1)_i} \partial_{(x_2)_i} \frac{\delta^2 \mathcal{U}}{\delta \mu^2} (t-s, \mu_s^N)(x,x) \right) \mu_s^N(\mathrm{d}x) \right] \mathrm{d}s.$$

Hence, by applying Lemma 7 and 8 with $\mu_0 = \mu_s^N, \xi = \xi^{[1]} = \xi^{[2]} = e_i$, we obtain

$$\begin{aligned}
&\int \left(\partial_{(x_1)_i} \partial_{(x_2)_i} \frac{\delta^2 \mathcal{U}}{\delta \mu^2} (t-s, \mu_s^N)(x,x) \right) \mu_s^N(\mathrm{d}x) \\
&= \int \frac{\delta^2 \Phi}{\delta \mu^2} (m(t-s, \mu_s^N))(d^{(1)}(t-s; \mu_s^N, \xi, x), d^{(1)}(t-s; \mu_s^N, \xi, x)) \\
&\quad + \frac{\delta \Phi}{\delta \mu} (m(t-s, \mu_s^N))(d^{(2)}(t-s; \mu_s^N, \xi, \xi, x, x)) \mu_s^N(\mathrm{d}x) \\
&= \mathcal{O}(\Lambda_{\mu_s^N} \exp(-\lambda\alpha(t-s-T_0)/2)),
\end{aligned}$$

where $\Lambda_{\mu_s^N} = \exp(\mathcal{O}(W_2(\mu_s^N, \mu^*)))$. Therefore, we just need to evaluate $\mathbb{E}[\Lambda_{\mu_s^N}]$. Suppose that C is a constant such that $\Lambda_{\mu_s^N} \leq \exp(CW_2(\mu_s^N, \mu^*))$. Then, we can easily verify that $\Lambda_{\mu_s^N} \leq \exp\left(C \sqrt{\mathbb{E}_{\mu_s^N}[\|X\|^2] + \mathbb{E}_{\mu^*}[\|X\|^2]} \right)$. Since $\mu^*(x) \lesssim \exp(-\lambda c_r \|x\|^{2+\delta})$, $\mathbb{E}_{\mu^*}[\|X\|^2] = \mathcal{O}(1)$ and thus we need to evaluate $\mathbb{E}\left[\exp\left(C \sqrt{\mathbb{E}_{\mu_s^N}[\|X\|^2]} \right) \right]$ This can be upper bounded by $\mathbb{E}\left[\exp\left(\frac{1}{2} + \frac{1}{2} \mathbb{E}_{\mu_s^N}[\|X\|^2] \right) \right] = \mathbb{E}\left[\exp\left(\frac{1}{2} + \frac{1}{2N} \sum_{i=1}^N \|\hat{X}_s^i\|^2 \right) \right]$. Let $Q_s := \frac{1}{N} \sum_{i=1}^N \|\hat{X}_s^i\|^2$. Then, by the Ito's formula, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E} \left[\exp\left(\frac{1}{2N} \sum_{i=1}^N \|\hat{X}_t^i\|^2 \right) \right] \\
&= \mathbb{E} \left[\exp\left(Q_t/2\right) \left(-\frac{1}{N} \sum_{i=1}^N b(\hat{X}_t^i, \mu_t^N)^\top \hat{X}_t^i + \frac{\lambda}{2N} \sum_{i=1}^N (d + \|\hat{X}_t^i\|^2) \right) \right] \\
&\leq \mathbb{E} \left[\exp\left(Q_t/2\right) \frac{1}{N} \sum_{i=1}^n \left(B^2 \|\hat{X}_t^i\| - (2+\delta)\lambda c_r \|\hat{X}_t^i\|^{2+\delta} + \frac{\lambda}{2}(d + \|\hat{X}_t^i\|^2) \right) \right].
\end{aligned}$$

Then, by the Young's inequality, there exists a constant C' depending on $B, \lambda, \delta, c_r, d$ such that

$$B^2 \|\hat{X}_t^i\| - (2+\delta)\lambda c_r \|\hat{X}_t^i\|^{2+\delta} + \frac{\lambda}{2}(d + \|\hat{X}_t^i\|^2) \leq C' - \frac{(2+\delta)\lambda c_r}{2} \|\hat{X}_t^i\|^{2+\delta}.$$

Moreover, by Jensen's inequality, the term related to $\|\hat{X}_t^i\|^{2+\delta}$ can be further bounded by

$$-\frac{1}{N} \sum_{i=1}^n \|\hat{X}_t^i\|^{2+\delta} \leq -\left(\frac{1}{N} \sum_{i=1}^n \|\hat{X}_t^i\|^2 \right)^{(2+\delta)/2} = Q_t^{(2+\delta)/2}.$$

In summary, we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\exp\left(Q_t/2\right)\right] \leq \mathbb{E}[\exp(Q_t/2)(C' - \epsilon Q_t^{(2+\delta)/2})]$$

with another constant $\epsilon = \frac{(2+\delta)\lambda c_r}{2}$. Here, by noticing that

$$C' - \epsilon Q_t^{(2+\delta)/2} \leq 2C'\mathbf{1}[Q_t \leq (2C'/\epsilon)^{2/(2+\delta)}] - C',$$

it holds that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\exp\left(Q_t/2\right)\right] &\leq \mathbb{E}[\exp(Q_t/2)(2C'\mathbf{1}[Q_t \leq (2C'/\epsilon)^{2/(2+\delta)}] - C')] \\ &\leq \mathbb{E}\left\{2C'\exp[\tfrac{1}{2}(2C'/\epsilon)^{2/(2+\delta)}] - C'\exp(Q_t/2)\right\}.\end{aligned}$$

This means that

$$\mathbb{E}[\exp(Q_t/2)] \leq \max\{2\exp[\tfrac{1}{2}(2C'/\epsilon)^{2/(2+\delta)}], \mathbb{E}[\exp(Q_0/2)]\} = \mathcal{O}(1).$$

We can further show an exponential decay of the term related to $\mathbb{E}[\exp(Q_0/2)]$, but this is sufficient for our purpose.

Finally, we arrive at

$$\begin{aligned}\mathbb{E}[\mathcal{U}(0,\mu_t^N) - \mathcal{U}(t,\mu_0^N)] &= \frac{1}{N}\mathcal{O}\left(\mathbb{E}[\Lambda_{\mu_s^N}]\int_0^t \exp(-\lambda\alpha(t-s-T_0)/2)\mathrm{d}s\right) \\ &\leq \frac{1}{N}\mathcal{O}\left(\exp(\lambda\alpha T_0/2)\frac{1}{\lambda\alpha}(1+\mathbb{E}[\Lambda_{\mu_0^N}])\right).\end{aligned}$$

By selecting the initial distribution so that $\mathbb{E}[\Lambda_{\mu_0^N}] = \mathcal{O}(1)$, the right hand side is $\mathcal{O}(1)$. This is satisfied if the support of μ_0 is bounded, e.g., $\mu_0^N = \frac{1}{N}\sum_{i=1}^N \delta_0$.