Parallel Computation Algorithms and Convergence Guarantees for Mean-Field Langevin Dynamics

Yoshihito Okamoto
The University of Tokyo
yo7al7fu717@g.ecc.u-tokyo.ac.jp

Huanjian Zhou
The University of Tokyo
zhou-huanjian185@g.ecc.u-tokyo.ac.jp

Taiji Suzuki
The University of Tokyo
RIKEN Center for Advanced Intelligence Project
taiji@mist.i.u-tokyo.ac.jp

Abstract

Entropy-regularized optimization problems arise in various areas of machine learning, including mean-field neural networks, variational inference, and reinforcement learning. Mean-field Langevin dynamics (MFLD) is a dynamics capable of sampling from the optimal solution of such problems under appropriate conditions, and has been the subject of extensive recent research. As machine learning models grow in scale and parameter dimension, computational efficiency becomes increasingly important, especially in reducing the adaptive complexity of computation of MFLD, that is, the number of sequential rounds they require. In this work, we initiate the study of the adaptive complexity of MFLD and provide the first theoretical evidence that MFLD can indeed be accelerated through parallelism. In particular, we provide polylogarithmic convergence guarantees for both infinite-particle and the finite-particle setting under the log-Sobolev inequality.

1 Introduction

The entropy-regularized convex minimization problem is a class of optimization problems over probability measures that arises in various settings in machine learning. In particular, in the training of two-layer mean-field neural networks (MFNNs), the optimization of the parameter distribution of each neuron in the mean-field limit of infinitely many neurons reduces to a problem of this form [NS17, MMN18, CB18, RV22, SS20]. A standard approach to solving this problem is *noisy gradient descent*, which adds Gaussian noise to the gradient of the loss functional [MMN18, HRŠS21]. In the continuous-time limit, noisy gradient descent induces a dynamics described by a McKean–Vlasov SDE, known as the *(infinite-particle) mean-field Langevin dynamics* (MFLD):

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\,\mathrm{d}t + \sqrt{2\lambda}\,\mathrm{d}B_t, \quad \mu_t = \mathrm{Law}(X_t).$$

Here, $\mathcal{P}_2(\mathbb{R}^d)$ denotes the set of probability measures on $\mathbb{R}^d$ with finite second moment, $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is a convex functional, and $\frac{\delta F}{\delta \mu} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ denotes its first variation. MFLD is the Wasserstein gradient flow of $\mathcal{L}$, and it has been shown that under the uniform *log-Sobolev inequality* (LSI), μ_t converges linearly to the minimizer of $\mathcal{L}$ [NWS22, Chi22]. However, since exactly tracking μ_t in the drift term of MFLD requires following infinitely many particles, implementing MFLD directly on a computer is infeasible. A natural remedy is to approximate μ_t by the empirical distribution $\rho_{\mathbf{X}_t} = \frac{1}{P}\sum_{p=1}^{P} \delta_{X_t^{(p)}}$, leading to the *finite-particle MFLD*:

$$\mathrm{d}X_t^{(p)} = -\nabla \frac{\delta F(\rho_{\mathbf{X}_t})}{\delta \mu}(X_t^{(p)})\,\mathrm{d}t + \sqrt{2\lambda}\,\mathrm{d}B_t^{(p)}.$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The error incurred by this approximation has been studied within the framework of *propagation of chaos* (PoC) [Szn91], and in the context of MFLD, *uniform-in-time* PoC for the finite-particle dynamics has been established in [CRW25, SNW23, NLK$^+$25]. In particular, [NLK$^+$25] refines the defective LSI implicitly introduced in [CRW25] and removes the dependence of the particle approximation error on the LSI constant, thereby establishing sharper convergence guarantees for the loss functional of finite-particle MFLD in both the continuous-time and discrete-time settings.

However, the finite- and infinite-particle MFLD algorithms based on the Euler–Maruyama discretization, for which convergence guarantees have been established above, are inherently sequential, and their iteration depth has already become a central practical bottleneck in several large-scale settings. For instance, in reinforcement learning, the actor network in mean-field Langevin actor-critic methods [YOYS24] is optimized via MFLD over a parameter space of dimension $d = D + 2$, where D is the dimension of the state-action space, making the sequential iteration cost prohibitive in high-dimensional control tasks. Another important example is LoRA-based fine-tuning of large language models. In the mean-field reformulation of LoRA [NLK$^+$25], the parameter dimension of each particle is determined by the hidden dimensions of the original layer, which can reach thousands or more in modern language models, and the sequential iteration cost of MFLD scales accordingly.

For gradient Langevin dynamics (GLD), the special case of MFLD in which F is linear, various parallel algorithms have been developed to improve the adaptive complexity, *i.e.*, the number of sequential rounds required. Under log-concavity, several parallel algorithms have been proposed that achieve an exponential improvement in adaptive complexity, with guarantees in the sense of Wasserstein distance or Kullback–Leibler (KL) divergence [SL19, YD25, ACV24, ZS25b]. Furthermore, [ZS25a] proposed a parallel algorithm building on [ZS25b] and established improved convergence guarantees in the sense of the relative Fisher information (FI) over sequential algorithms, without assuming log-concavity or the LSI for the target distribution.

In contrast, no parallel algorithm has been studied for MFLD to date. The inefficiency of sequential methods, together with this gap between GLD and MFLD, naturally leads to the following question:

Can parallel algorithms for MFLD achieve a computational advantage over sequential ones?

1.1 Our Contributions

In this paper, we initiate the study of parallel algorithms for MFLD in both the infinite- and finite-particle settings. Specifically, we construct parallel algorithms for MFLD by extending the framework of [ZS25a] from GLD to MFLD, and establish the *first* convergence guarantees showing improved adaptive complexity under various assumptions and convergence criteria. Our specific contributions are as follows.

Infinite-particle MFLD. We propose a parallel algorithm for infinite-particle MFLD by extending the algorithm of [ZS25a] for GLD to MFLD. We establish three convergence guarantees for this algorithm. First, under strong LSI on the stationary distribution, we prove convergence of the output distribution in KL divergence. Next, under the uniform LSI, we establish convergence in the loss functional and show that the algorithm improves adaptive complexity over sequential algorithms. Finally, under only a Lipschitz continuity condition on F, we prove a time-averaged FI bound toward the proximal Gibbs distribution along the interpolated trajectory generated by the algorithm. We summarize the comparison between existing results and our results in Table 1.

Finite-particle MFLD. For the finite-particle setting, we propose two parallel algorithms with the same update rule but different output rules tailored to the convergence criterion. First, without assuming any LSI condition, we prove FI convergence to the finite-particle stationary distribution. Next, under assumptions ensuring the defective LSI, we establish convergence in the sense of the loss functional and show that the corresponding algorithm improves adaptive complexity over sequential algorithms. We summarize the comparison between existing results and our results in Table 2.

1.2 Relation to Prior Work

Prior work on MFLD. As discussed above, we propose the first parallel algorithms for MFLD and establish their convergence guarantees. Furthermore, in the context of MFLD convergence analysis,

Table 1: Comparison of adaptive complexity required to achieve ε^2-accuracy for infinite-particle MFLD. We assume $\mathsf{KL}(\mu^0\|\pi^0) = O(d)$, $\mathsf{KL}(\mu^0\|\pi^*) = O(d)$, $\mathsf{FI}(\mu^0\|\pi^0) = O(d)$, and $\mathcal{L}(\mu^0) - \mathcal{L}(\pi^*) = O(d)$ (see Lemma 3.1). Uniform LSI and strong LSI refer to Assumptions 2.5 and 2.4, respectively. PGD denotes the proximal Gibbs distribution. For averaged FI w.r.t. PGD, we report the number of adaptive rounds required to make this quantity at most ε^2.

Reference	Parallel computation	Metric	LSI	Adaptive complexity	Queries per round
[NWS22]	$\times$	Loss	Uniform LSI	$O(\frac{d}{\varepsilon^2}\log(\frac{d}{\varepsilon^2}))$	1
Theorem 3.2	✓	KL	Strong LSI	$O(\log(\frac{d}{\varepsilon^2}))$	$O(\frac{d}{\varepsilon^2}\log(\frac{d}{\varepsilon^2}))$
Theorem 3.4	✓	Loss	Uniform LSI	$O(\log(\frac{d}{\varepsilon^2}))$	$O(\frac{d}{\varepsilon^2}\log(\frac{d}{\varepsilon^2}))$
Theorem 3.7	✓	Averaged FI w.r.t. PGD	None	$O(\frac{d}{\varepsilon^2} + \log(\frac{d}{\varepsilon^2}))$	$O(\frac{d}{\varepsilon^2}\log(\frac{d}{\varepsilon^2}))$

Table 2: Comparison of adaptive complexity required to achieve ε^2-accuracy for finite-particle MFLD. We assume $\frac{1}{P}\mathcal{L}(\mu^0) - \mathcal{L}(\pi^*) = O(d)$, $\mathsf{FI}(\mu^0\|\mu_*^{(P)}) = O(dP)$, and $\mathbb{E}[\|\mathbf{x}_{\mathrm{init}}\|^2] = O(dP)$ (see Lemma 4.1), where $\mu_*^{(P)}$ denotes the stationary distribution of finite-particle MFLD. Defective LSI refers to Lemma 2.7.

Reference	Parallel computation	Metric	LSI	Adaptive complexity	Queries per round
[NLK$^+$25]	$\times$	Loss	Defective LSI	$O(\frac{d}{\varepsilon^4}\log(\frac{d}{\varepsilon^2}))$	1
Theorem 4.2	✓	FI w.r.t. $\mu_*^{(P)}$	None	$O(\frac{d}{\varepsilon^4} + \log(\frac{d}{\varepsilon^2}))$	$O(\frac{d}{\varepsilon^6}\log(\frac{d}{\varepsilon^2}))$
Theorem 4.3	✓	Loss	Defective LSI	$O(\log(\frac{d}{\varepsilon^2}))$	$O(\frac{d}{\varepsilon^4}\log(\frac{d}{\varepsilon^2}))$

Theorem 3.2 provides a new theoretical framework by introducing a strong LSI condition imposed only on the stationary distribution and establishing convergence guarantees under this condition.

Prior work on parallel algorithms for GLD. Our algorithms and proofs build on the frameworks of [ZS25a, ZS25b]. However, the distribution dependence of the drift in MFLD introduces several technical differences in the analysis. In particular, while the proof of Theorem 3.2 broadly follows the KL-based analysis of these works, our other convergence results require analyzing the time evolution of the loss functional, since the KL-based approach leads to error terms that are difficult to control in the MFLD setting. Although analyzing the loss functional is standard in MFLD, we establish a new evolution inequality for the loss functional analogous to the KL-based inequality used in the parallel computation framework of [ZS25a]. Furthermore, the distribution dependence of the drift also necessitates considering the finite-particle setting. In practical computation, the evolving distribution must be approximated by the empirical measure of finitely many particles. This setting requires quantifying additional finite-particle approximation errors, for which the defective LSI of [NLK$^+$25] provides one tool.

2 Preliminaries

2.1 Infinite-Particle MFLD

For a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, a map $\frac{\delta F}{\delta \mu} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ satisfying $\frac{\mathrm{d}F(\mu+\varepsilon(\mu'-\mu))}{\mathrm{d}\varepsilon}|_{\varepsilon=0} = \int \frac{\delta F(\mu)}{\delta \mu}(x)\mathrm{d}(\mu'-\mu)(x)$ for all $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$ is called the first variation of F. It is unique up to an additive constant. A functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is said to be *linearly convex* if for all $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$, $F(\mu') \geq F(\mu) + \int \frac{\delta F(\mu)}{\delta \mu}(x)\mathrm{d}(\mu'-\mu)(x)$. Under these definitions, we consider the *entropy-regularized (convex) minimization problem*

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

where $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is convex, $\lambda > 0$, and $\mathrm{Ent}(\mu) = \int \log(\frac{\mathrm{d}\mu}{\mathrm{d}x}(x))\mathrm{d}\mu(x)$ denotes the entropy of $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. As a dynamics for solving this optimization problem, we consider the mean-field Langevin dynamics (MFLD), described by the following stochastic differential equation:

$$\mathrm{d}x_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(x_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}B_t \tag{2}$$

where $\mu_t = \mathrm{Law}(x_t)$ denotes the distribution of x_t and B_t is a d-dimensional Brownian motion. It is known that this constitutes the Wasserstein gradient flow of $\mathcal{L}$, and that the distribution of x_t converges to the optimal solution of (1) under appropriate conditions [NWS22, Chi22]. For simplicity, we set $\lambda = 1$ hereafter, which corresponds to absorbing the temperature parameter into F. Moreover, from the Fokker–Planck equation, the stationary distribution π^* satisfies $\pi^* \propto \exp(-\frac{\delta F(\pi^*)}{\delta \mu})$. Throughout the paper, unless otherwise stated, we assume that the stationary distribution exists and is unique.

For the purpose of analysis, we introduce the following Lipschitz condition.

Assumption 2.1 (Smoothness). There exist constants $L_1, L_2 \geq 0$ such that for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and all $x, y \in \mathbb{R}^d$, $\|\nabla \frac{\delta F(\mu)}{\delta \mu}(x) - \nabla \frac{\delta F(\nu)}{\delta \mu}(y)\| \leq L_1 W_2(\mu, \nu) + L_2\|x - y\|$ holds, where $\|\cdot\|$ denotes the Euclidean norm and $W_2(\mu, \nu) = (\min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y))^{1/2}$ is the 2-Wasserstein distance between μ and ν. We also define $L = L_1 \vee L_2$.

Although this assumption is restrictive, it is commonly adopted in theoretical analyses of MFLD, and the class of functionals F satisfying it includes examples of practical interest. For instance, it holds for suitably parameterized two-layer neural networks combining bounded and smooth activation functions, such as sigmoid or tanh, with squared or logistic loss; see Proposition 5.1 of [Chi22].

2.2 Finite-Particle MFLD

Let P denote the number of particles, and let boldface denote P-tuples of $\mathbb{R}^d$-valued vectors such as $\mathbf{x} = (x^{(1)}, \ldots, x^{(P)})$, with $\|\mathbf{x}\|$ denoting the Euclidean norm on $(\mathbb{R}^d)^P$. We use the shorthand $\nabla \frac{\delta F(\mu)}{\delta \mu}(\mathbf{x}) = (\nabla \frac{\delta F(\mu)}{\delta \mu}(x^{(1)}), \ldots, \nabla \frac{\delta F(\mu)}{\delta \mu}(x^{(P)}))$, and define $\rho_{\mathbf{x}} = \frac{1}{P}\sum_{p=1}^{P} \delta_{x^{(p)}}$.

In practice, it is generally difficult to directly employ MFLD (Eq. 2), which requires exact knowledge of the distribution of infinitely many particles. Therefore, one considers the following SDE, in which the distribution μ_t is approximated by the empirical distribution $\rho_{\mathbf{X}_t}$ of finitely many particles:

$$\mathrm{d}\mathbf{X}_t = -\nabla \frac{\delta F(\rho_{\mathbf{X}_t})}{\delta \mu}(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \tag{3}$$

where $\mu_t = \mathrm{Law}(\mathbf{X}_t)$ denotes the distribution of $\mathbf{X}_t$ and $B_t = (B_t^{(1)}, \ldots, B_t^{(P)})$ is a dP-dimensional Brownian motion. We refer to this as finite-particle MFLD. It is known that this constitutes the Wasserstein gradient flow of the objective functional $\mathcal{L}^{(P)}$ in the following optimization problem:

$$\min_{\mu^{(P)} \in \mathcal{P}_2((\mathbb{R}^d)^P)} \{\mathcal{L}^{(P)}(\mu^{(P)}) = P\mathbb{E}_{\mathbf{X} \sim \mu^{(P)}}[F(\rho_{\mathbf{X}})] + \mathrm{Ent}(\mu^{(P)})\}. \tag{4}$$

Our original goal is to minimize the loss functional $\mathcal{L}$ in the infinite-particle setting, whose minimum value is $\mathcal{L}(\pi^*)$ under appropriate regularity conditions. Accordingly, the question of interest is whether the per-particle contribution $\frac{1}{P}\mathcal{L}^{(P)}$ of the finite-particle loss functional $\mathcal{L}^{(P)}$ approaches $\mathcal{L}(\pi^*)$ as the algorithm progresses. Under appropriate conditions, $\frac{1}{P}\mathcal{L}^{(P)}(\mu_t)$ approaches $\mathcal{L}(\pi^*)$ for large t and P [CRW25, SNW23, NLK+25]. Moreover, from the Fokker–Planck equation, the stationary distribution of Eq. (3) is given by $\mu_*^{(P)}(\mathbf{x}) \propto \exp(-PF(\rho_{\mathbf{x}}))$.

2.3 Concepts for Computational Complexity Analysis

Parallel gradient oracle. We define a first-order gradient oracle for the first variation $\frac{\delta F}{\delta \mu} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ as an oracle that, given a query (μ, x), returns $\nabla \frac{\delta F(\mu)}{\delta \mu}(x)$. Similarly, for $k = \mathrm{poly}(d)$, we define a parallel gradient oracle as an oracle that, given a query $\{(\mu_1, x_1), \ldots, (\mu_k, x_k)\}$, returns $\{\nabla \frac{\delta F(\mu_1)}{\delta \mu}(x_1), \ldots, \nabla \frac{\delta F(\mu_k)}{\delta \mu}(x_k)\}$.

ε^2-accurate sample. In the context of sampling, we say that the output is an ε^2-accurate sample with respect to the relative Fisher information if the distribution μ of the output satisfies $\mathsf{FI}(\mu\|\pi) =$

$\mathbb{E}_\mu\big[\, \big\|\nabla \log \frac{\mu}{\pi}\big\|^2 \,\big] \le \varepsilon^2$ for the target distribution π. Analogously, μ is said to be an ε^2-accurate sample with respect to the KL divergence if $\mathsf{KL}(\mu\|\pi) := \mathbb{E}_\mu\big[\log \frac{\mu}{\pi}\big] \le \varepsilon^2$. Furthermore, we say that μ is an ε^2-accurate sample with respect to the loss functional if $\mathcal{L}(\mu) - \mathcal{L}(\pi) \le \varepsilon^2$ in the infinite-particle case, and $\frac{1}{P}\mathcal{L}^{(P)}(\mu) - \mathcal{L}(\pi) \le \varepsilon^2$ in the finite-particle case with P particles.

Adaptive complexity. We define the adaptive complexity of an algorithm as the minimum number of rounds of queries to the oracle required for the output to be an ε^2-accurate sample. We treat all parameters except d and ε as constants. For a formal definition, see [DG19, ZS25a].

2.4 Assumptions for Convergence Guarantees

In this section, we present the concepts and propositions used for establishing convergence guarantees. In the convergence analysis of MFLD, the following proximal Gibbs distribution associated with a probability measure μ plays a central role.

Definition 2.2 (Proximal Gibbs distribution). For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the proximal Gibbs distribution $\pi[\mu]$ is defined as $\pi[\mu](x) = \exp\big(-\frac{\delta F(\mu)}{\delta \mu}(x)\big)/Z(\mu)$, where $Z(\mu)$ is the normalizing constant.

In particular, in the case of GLD, $\pi[\mu] = \pi^* \propto \exp(-V)$ for all μ. Hereafter, unless otherwise stated, we assume that for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the first variation $\frac{\delta F(\mu)}{\delta \mu}(x)$ of F exists and is smooth, and that $|Z(\mu)| < \infty$ and $\pi[\mu] \in \mathcal{P}_2(\mathbb{R}^d)$ hold. Estimates using the proximal Gibbs distribution include Entropy sandwich [NWS22], among others. Moreover, the following log-Sobolev inequality is an important tool used in the convergence guarantees of GLD and MFLD.

Definition 2.3 (Log-Sobolev inequality (LSI)). A distribution $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is said to satisfy the *log-Sobolev inequality* if there exists $\alpha > 0$ such that for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mu \ll \nu$, $\mathsf{KL}(\mu\|\nu) \le \frac{1}{2\alpha}\mathsf{FI}(\mu\|\nu)$ holds. In this case, we write that ν satisfies $\mathrm{LSI}(\alpha)$.

We also introduce the following assumption, used in Section 3.2.

Assumption 2.4 (Strong LSI on π^*). There exists $\alpha > \frac{6}{\sqrt{5}}L_1$ such that π^* satisfies $\mathrm{LSI}(\alpha)$.

While the LSI does not hold in general, sufficient conditions are known, such as the Bakry–Émery criterion [BÉ85] and the Holley–Stroock bounded perturbation lemma [HS87]. For convergence guarantees based on the loss functional, the following assumption is commonly made, which requires that the LSI holds uniformly over the input μ for the proximal Gibbs distribution.

Assumption 2.5 (Uniform LSI). There exists $\alpha > 0$ such that $\forall \mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\pi[\mu]$ satisfies $\mathrm{LSI}(\alpha)$.

In particular, in the case of GLD, since the proximal Gibbs distribution is constant, this reduces to the necessary and sufficient condition that the stationary distribution $\pi^* \propto \exp(-V)$ satisfies the LSI.

We next introduce assumptions that serve as sufficient conditions for the defective LSI to hold.

Assumption 2.6 (Assumptions 3.2–3.4 in [NLK$^+$25])**.** The following conditions **(A1)–(A3)** hold.

- **(A1)** There exists $\alpha > 0$ such that for all $\mathbf{x} \in (\mathbb{R}^d)^P$ and all $i \in \{1, \dots, P\}$, the conditional Gibbs distribution $\nu_{i|-i}(\cdot|\mathbf{x}^{-i}) \propto \exp(-PF(\rho_{x\cup\mathbf{x}^{-i}}))$ satisfies $\mathrm{LSI}(\alpha)$, where $\mathbf{x}^{-i} = (x^{(1)}, \dots, x^{(i-1)}, x^{(i+1)}, \dots, x^{(P)})$ and $\rho_{x\cup\mathbf{x}^{-i}} = \frac{1}{P}\sum_{j\neq i}\delta_{x^{(j)}} + \frac{1}{P}\delta_x$.
- **(A2)** In the functional $\mu \mapsto F(\mu) = F_0(\mu) + \mathbb{E}_{X\sim\mu}[r(X)]$, F_0 is differentiable and linearly convex.
- **(A3)** There exists a constant $B > 0$ such that for all $\mathbf{x} \in (\mathbb{R}^d)^P$, $x \in \mathbb{R}^d$, and $i \in \{1, 2, \dots, P\}$, $B_{F_0}(\rho_{x\cup\mathbf{x}^{-i}}, \rho_{\mathbf{x}}) \le \frac{B}{P^2}$ holds, where B_{F_0} denotes the Bregman divergence of F_0, defined for $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$ as $B_{F_0}(\mu, \mu') = F_0(\mu) - F_0(\mu') - \langle \frac{\delta F_0(\mu')}{\delta \mu}, \mu - \mu' \rangle$.

These assumptions are satisfied, for example, by the two-layer mean-field neural network with a bounded activation function and L_2-regularization considered in Example 3.5 of [NLK$^+$25], where the Holley–Stroock bounded perturbation lemma ensures condition **(A1)**. Under this assumption, the following inequality, known as the defective LSI, holds.

Lemma 2.7 (Defective LSI, [NLK$^+$25]). Under Assumption 2.6, for all $\mu \in \mathcal{P}_2((\mathbb{R}^d)^{(P)})$,

$$\frac{1}{P}\mathcal{L}^{(P)}(\mu) - \mathcal{L}(\pi^*) \le \frac{B}{P} + \frac{1}{2\alpha P}\mathsf{FI}(\mu\|\mu_*^{(P)}).$$

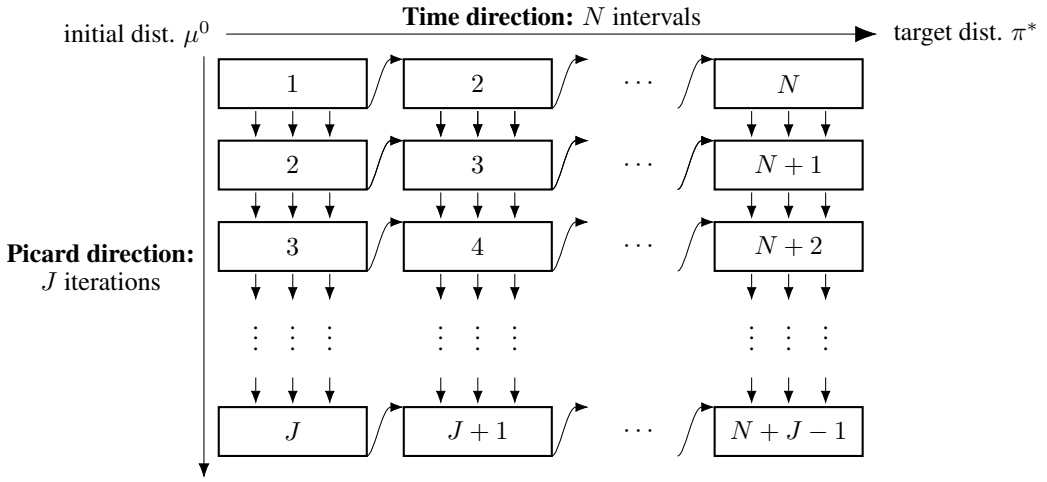


Figure 1: Wave-like parallel Picard computation for MFLD, adapted from Figure 1 of [ZS25b]. Each rectangle represents one time-interval/Picard-iteration block (n, j), where $n = 0, \ldots, N-1$ and $j = 1, \ldots, J$. The number in a rectangle denotes the oracle-query round $k = n + j$; blocks with the same number can be updated in parallel. The total number of rounds of oracle queries is $N + J - 1$.

3 Parallel Picard Method for Infinite-Particle MFLD

In this section, we present the parallel algorithm for infinite-particle MFLD (Algorithm 1) in Section 3.1, and show it achieves $O(\log d)$ adaptive complexity under either the LSI condition for the target distribution or the LSI condition for proximal Gibbs distribution uniformly in Section 3.2. Finally, without assuming any LSI condition, we prove a time-averaged FI bound along the interpolated trajectory generated by Algorithm 1 in Section 3.3. The detailed proofs of the results in this section are deferred to Appendix B.

3.1 Parallel Algorithm for Infinite-Particle MFLD

Following [ZS25a], for the interval $[t_n, t_{n+1}]$, we consider the following Picard method approximation of MFLD:

$$x_{t_{n+1}} - x_{t_n} = -\int_{t_n}^{t_{n+1}} \nabla \frac{\delta F(\mu_s)}{\delta \mu}(x_s)\mathrm{d}s + \sqrt{2}(B_{t_{n+1}} - B_{t_n}) \tag{5}$$

$$\approx -\sum_{i=0}^{M-1} w_i \nabla \frac{\delta F(\mu_{t_n+\tau_{n,i}})}{\delta \mu}(x_{t_n+\tau_{n,i}}) + \sqrt{2}(B_{t_{n+1}} - B_{t_n}) \tag{6}$$

where $t_n = t_n + \tau_{n,0} \le t_n + \tau_{n,1} \le \cdots \le t_n + \tau_{n,M} = t_{n+1}$ defines a discretization of the time interval $[t_n, t_{n+1}]$ and $w_0, \ldots, w_{M-1} \in \mathbb{R}_{\ge 0}$ are weights. Based on this approximation, we perform the following update in Lines 9–18:

$$x^j_{n,m} = x^j_{n,0} - \frac{h}{M}\sum_{m'=0}^{m-1} \nabla \frac{\delta F(\mu^{j-1}_{n,m'})}{\delta \mu}(x^{j-1}_{n,m'}) + \sqrt{2}(B_{nh+\frac{mh}{M}} - B_{nh}). \tag{7}$$

Here, h denotes the step size, $j = 1, \ldots, J$ denotes the Picard iteration index, $n = 0, \ldots, N-1$ corresponds to the index of time interval, and $m = 0, \ldots, M$ corresponds to $\tau_{n,m} = \frac{mh}{M}$. Furthermore, $\mu^j_{n,m}$ denotes the distribution of $x^j_{n,m}$. That is, within each interval, the values in the current interval are computed using the values from the previous Picard iteration within the same interval. At the boundary of each interval, we set $x^j_{n,0} = x^j_{n-1,M}$ using the value at the right endpoint of the preceding interval from the same iteration (Line 11 and Line 16). By employing this update scheme, the computation can be carried out in a wave-like fashion, grouping blocks that share the same value of $j + n$. This enables parallel computation of the blocks within the same group. Figure 1 illustrates this computation schedule.

Algorithm 1 Parallel Picard method for infinite-particle MFLD

Input: $x_{\mathrm{init}} \sim \mu^0$, access to a gradient oracle for $\nabla \frac{\delta F(\cdot)}{\delta \mu}$, number of iterations J, number of intervals $N(\leq J)$, interval width h, number of subdivision points per interval M.

for $n = 0, \dots, N-1, m = 0, \dots, M-1$ **do**
  $\xi_{n,m} \sim \mathcal{N}(0, \frac{h}{M} I)$
for $n = 0, \dots, N-1, m = 0, \dots, M-1$ **do**
  $B_{nh + \frac{(m+1)h}{M}} = \sum_{n'=0}^{n-1} \sum_{m'=0}^{M-1} \xi_{n',m'} + \sum_{m'=0}^{m} \xi_{n,m'}$
for $n = 0, \dots, N-1, m = 0, \dots, M-1$ **do**
  $x^0_{n,m} = x_{\mathrm{init}}$
for $j = 1, \dots, J$ **do**
  $x^j_{0,0} = x_{\mathrm{init}}$
for $k = 1, \dots, N$ **do**
  for $j = 1, \dots, \min\{k, J\}$ (in parallel) **do**
    Set $n = k - j$, and if $n \neq 0$, set $x^j_{n,0} = x^j_{n-1,M}$.
    for $m = 1, \dots, M$ (in parallel) **do**
      $x^j_{n,m} = x^j_{n,0} - \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\mu^{j-1}_{n,m'})}{\delta \mu}(x^{j-1}_{n,m'}) + \sqrt{2}(B_{nh + \frac{mh}{M}} - B_{nh})$
for $k = N+1, \dots, N+J-1$ **do**
  for $n = \max\{0, k-J\}, \dots, N-1$ (in parallel) **do**
    Set $j = k - n$, and if $n \neq 0$, set $x^j_{n,0} = x^j_{n-1,M}$.
    for $m = 1, \dots, M$ (in parallel) **do**
      $x^j_{n,m} = x^j_{n,0} - \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\mu^{j-1}_{n,m'})}{\delta \mu}(x^{j-1}_{n,m'}) + \sqrt{2}(B_{nh + \frac{mh}{M}} - B_{nh})$
return $x^J_{N,0} = x^J_{N-1,M}$

3.2 Analysis of Algorithm 1: $O(\log d)$ Adaptive Complexity

In this section, we provide concrete convergence guarantees for Algorithm 1. In particular, we show convergence to the stationary distribution under an LSI condition for the target distribution (Theorem 3.2), and convergence in the loss under an LSI condition that holds uniformly over all proximal Gibbs distributions (Theorem 3.4).

Before stating the convergence theorem, we first specify the initialization condition required for our analysis. By using Lemma 17 in [CGLL23] and the entropy sandwich [NWS22], the following lemma can be established.

Lemma 3.1 (**Initialization for infinite-particle case).** Suppose the smooth condition (Assumption 2.1) holds and let $\mu^0 = \mathcal{N}(0, L_2^{-1} I_d)$. Assume $\|\nabla \frac{\delta F(\mu^0)}{\delta \mu}(0)\|^2 = O(d)$, and that there exists $\Delta > 0$ satisfying $\frac{\delta F(\mu^0)}{\delta \mu}(0) - \inf_{x \in \mathbb{R}^d} \frac{\delta F(\mu^0)}{\delta \mu}(x) \leq \Delta$. Furthermore, assume that $\mathfrak{m} = \int \|\cdot\| \mathrm{d}\pi^0 < \infty$ for $\pi^0 = \pi[\mu^0]$. Then, the following initialization condition holds,

$$\mathsf{KL}(\mu^0 \| \pi^0) = O(d), \quad \mathsf{KL}(\mu^0 \| \pi^*) = O(d), \quad \mathsf{FI}(\mu^0 \| \pi^0) = O(d), \quad \text{and} \quad \mathcal{L}(\mu^0) - \mathcal{L}(\pi^*) = O(d).$$

Convergence to the stationary distribution in KL. Under Assumption 2.4, the output distribution of Algorithm 1 converges linearly to π^* in KL, up to a discretization error (the second term) and a parallelization error (the third term).

Theorem 3.2. Suppose assumptions 2.1 and 2.4 hold, $Lh \leq \frac{1}{20}$, and $J \geq N$. Then, Algorithm 1 performs $N + J$ iterations and processes at most MJ queries per iteration. Moreover, for $\beta = \frac{3}{4} - \frac{9L_1^2}{2\alpha^2} - 27L^2 \frac{h^2}{M^2}$, $\pi^0 = \pi[\mu^0]$, and $L = L_1 \vee L_2$, the output distribution $\rho = \mu^J_{N,0}$ satisfies

$$\mathsf{KL}(\rho \| \pi^*) \lesssim \exp(-2\alpha\beta N h) \mathsf{KL}(\mu^0 \| \pi^*) + \frac{Ld}{M} + (\mathsf{FI}(\mu^0 \| \pi^0) + Ld) 0.5^{J-N}.$$

A notable aspect is that this result requires only the LSI on π^*, rather than the uniform LSI commonly assumed in the convergence theory of MFLD [NWS22, Chi22]. Instead, a stronger condition on α relative to L_1 is imposed to ensure $\beta > 0$.

Corollary 3.3. With initialization condition (Lemma 3.1), by taking

$h = O(\frac{1}{L}), \quad N = O(L \log(\frac{\mathsf{KL}(\mu^0 \| \pi^*)}{\varepsilon^2})), \quad J - N = O(\log(\frac{\mathsf{FI}(\mu^0 \| \pi^0) + Ld}{\varepsilon^2})), \quad \text{and} \quad M = O(\frac{Ld}{\varepsilon^2})$

the algorithm runs $O(\log(\frac{d}{\varepsilon^2}))$ iterations with $O(\frac{d}{\varepsilon^2} \log(\frac{d}{\varepsilon^2}))$ queries per iteration and returns a sample satisfying $\mathsf{KL}(\rho \| \pi^*) \le \varepsilon^2$.

Convergence in the loss functional Under the uniform LSI (Assumption 2.5), the following linear convergence of the loss functional can be established.

Theorem 3.4. Suppose $Lh \le \frac{1}{20}$ and Assumptions 2.1 and 2.5 hold. Then, Algorithm 1 performs $N + J$ iterations and processes at most MJ queries per iteration. Moreover, for $\Delta_0 = \mathcal{L}(\mu^0) - \mathcal{L}(\pi^*)$ and $L = L_1 \vee L_2$, the output distribution $\rho = \mu^J_{N,0}$ satisfies

$$\mathcal{L}(\rho) - \mathcal{L}(\pi^*) \lesssim \exp(-\alpha N h)\Delta_0 + \frac{Ld}{M} + (\mathsf{FI}(\mu^0 \| \pi^0) + Ld) 0.5^{J-N}.$$

Compared with the convergence rate for the sequential setting shown in Theorem 2 of [NWS22], Theorem 3.4 introduces a new parallelization error term, the third term on the right-hand side, which decreases exponentially in $J - N$.

Proof Sketch of Theorem 3.4. Following [NWS22, NLK+25], we employ the one-step interpolation argument applied to $\partial_t \mathcal{L}(\mu_t)$. By an argument analogous to Lemma 12 in [BCE+22], for $t \in [nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}]$, we obtain

$$\partial_t \mathcal{L}(\mu_t) \le -\frac{3}{4}\mathsf{FI}(\mu_t \| \pi_t) + \mathbb{E}\left[\left\| \nabla \frac{\delta F(\mu_t)}{\delta \mu}(x_t) - \nabla \frac{\delta F(\mu^{J-1}_{n,m})}{\delta \mu}(x^{J-1}_{n,m}) \right\|^2\right].$$

The second term is bounded using the Lipschitz continuity of $\nabla \frac{\delta F(\mu)}{\delta \mu}$ (Assumption 2.1) and the Picard iteration error $\mathcal{E} = \max_{n,m} \mathbb{E}[\|x^J_{n,m} - x^{J-1}_{n,m}\|^2]$, which decays exponentially in $J - N$ by an argument analogous to [ZS25a]. Combining the uniform LSI (Assumption 2.5) with the entropy sandwich [NWS22] to bound $\mathsf{FI}(\mu_t \| \pi_t)$ from below in terms of $\mathcal{L}(\mu_t) - \mathcal{L}(\pi^*)$, and applying Gronwall's inequality, yields the theorem.

Corollary 3.5. With initialization condition (Lemma 3.1), by taking

$$h = O(\tfrac{1}{L}), \quad N = O(L \log(\tfrac{\Delta_0}{\varepsilon^2})), \quad J - N = O(\log(\tfrac{\mathsf{FI}(\mu^0 \| \pi^0) + Ld}{\varepsilon^2})), \quad \text{and} \quad M = O(\tfrac{Ld}{\varepsilon^2}),$$

the algorithm runs $O(\log(\frac{d}{\varepsilon^2}))$ iterations with $O(\frac{d}{\varepsilon^2} \log(\frac{d}{\varepsilon^2}))$ queries per iteration and returns a sample satisfying $\mathcal{L}(\rho) - \mathcal{L}(\pi^*) \le \varepsilon^2$.

Remark 3.6 (Comparison with [NWS22]). In the infinite-particle setting, Theorem 3.4 shows that the adaptive complexity is reduced from $O(\frac{d}{\varepsilon^2} \log(\frac{d}{\varepsilon^2}))$ for the sequential algorithm to $O(\log(\frac{d}{\varepsilon^2}))$ by Algorithm 1 (see Table 1).

3.3 Parallel Methods for Targets Beyond LSI

The preceding guarantees impose either an LSI on π^* or a uniform LSI on the proximal Gibbs distributions. As in standard GLD, where time-averaged Fisher-divergence bounds hold without log-concavity or LSI [BCE+22, ZS25a], we prove an analogous bound for MFLD relative to the proximal Gibbs distributions. This guarantee does not directly control the final output distribution of Algorithm 1, but it quantifies convergence of the intermediate distributions without assuming LSI.

Convergence of intermediate distributions to the proximal Gibbs distribution Let μ_t denote the interpolated distribution generated during the execution of Algorithm 1. Then the following time-averaged Fisher divergence bound holds.

Theorem 3.7. Suppose assumptions 2.1 hold, $Lh \le \frac{1}{20}$, and $J \ge N$. Then, Algorithm 1 performs $N + J$ iterations and processes at most MJ queries per iteration. Moreover, the following holds:

$$\frac{1}{Nh} \int_0^{Nh} \mathsf{FI}(\mu_t \| \pi_t) \mathrm{d}t \lesssim \frac{\mathsf{KL}(\mu^0 \| \pi^0)}{Nh} + \frac{Ld}{M} + (\mathsf{FI}(\mu^0 \| \pi^0) + Ld) 0.5^{J-N}$$

This can be shown by evaluating the time evolution of the loss functional and applying the entropy sandwich [NWS22]. Since $\pi[\pi^*] = \pi^*$ and $\mathsf{FI}(\mu \| \pi[\mu]) = 0$ implies $\mu = \pi^*$ by uniqueness,

$\mathsf{FI}(\mu\|\pi[\mu])$ can be regarded as a measure of convergence to π^*. Theorem 3.7 provides an average convergence guarantee with respect to this measure.

4 Parallel Picard Method for Finite-Particle MFLD

In this section, we present the parallel algorithm for finite-particle MFLD (Algorithm 2) in Section 4.1. In Section 4.2, we first establish a convergence guarantee without assuming the LSI and discuss its fundamental limitation, and then show that Algorithm 2′ achieves $O(\log d)$ adaptive complexity under the defective LSI. All proofs in this section are deferred to Appendix C.

4.1 Parallel Algorithm for Finite-Particle MFLD

Corresponding to the finite-particle MFLD, we consider an algorithm (Algorithm 2) based on the following update, analogous to Algorithm 1:

$$\mathbf{x}_{n,m}^j = \mathbf{x}_{n,0}^j - \frac{h}{M}\sum_{m'=0}^{m-1} \nabla \frac{\delta F(\rho_{n,m'}^{j-1})}{\delta \mu}(\mathbf{x}_{n,m'}^{j-1}) + \sqrt{2}(B_{nh+\frac{mh}{M}} - B_{nh}).$$

Here, $\rho_{n,m}^j = \rho_{\mathbf{x}_{n,m}^j}$. The boundary points within each interval $[nh, (n+1)h]$ are updated as $\mathbf{x}_{n,0}^j = \mathbf{x}_{n-1,M}^j$, in the same manner as the infinite-particle case. Moreover, parallelization is performed not only over the Picard iteration index j (or equivalently, the interval index n) and the subdivision index m, but also over the particle index p. Furthermore, by employing random sampling of the time from $[0, Nh]$, the distribution of the output is set to the time-averaged distribution of the interpolation process (see the next section): $\overline{\mu}_{Nh} = \frac{1}{Nh}\int_0^{Nh} \mu_t \mathrm{d}t$. We also define Algorithm 2′ as the variant of Algorithm 2 that outputs $\mathbf{x}_{N,0}^J$ instead of sampling a random time from $[0, Nh]$. As the remaining details are analogous to Algorithm 1, they are deferred to Appendix C.

4.2 Analysis of Algorithm 2: Convergence without LSI

In this section, we establish a convergence guarantee for Algorithm 2 without assuming any form of LSI, and discuss its fundamental limitation.

Before stating the convergence theorem, we also state the initialization condition required for our analysis. By using Lemma 17 in [CGLL23] and the defective entropy sandwich [NLK+25], the following lemma can be established.

Lemma 4.1 (Initialization for finite-particle case). Suppose Assumptions 2.1 and 2.6-(A3) hold, and let $\mu^0 = \mathcal{N}(0, (2L)^{-1} I_{dP})$. Assume that $\|\nabla \frac{\delta F(\rho_{\mathbf{0}})}{\delta \mu}(\mathbf{0})\|^2 = O(dP)$ holds, and that there exists $\Delta > 0$ such that for all $\mathbf{x} \in (\mathbb{R}^d)^P$ and all i, $\frac{\delta F(\rho_{\mathbf{0}\cup\mathbf{x}^{-i}})}{\delta \mu}(0) - \inf_{x\in\mathbb{R}^d} \frac{\delta F(\rho_{x\cup\mathbf{x}^{-i}})}{\delta \mu}(x) \le \Delta$. Furthermore, assume that for all i, $\mathfrak{m}_i = \int \|\cdot\| \mathrm{d}\nu_{i|-i}(\cdot|\mathbf{x}^{-i}) < \infty$ holds for $\nu_{i|-i}(\cdot|\mathbf{x}^{-i})$. Then, the following initialization condition holds,

$$\tfrac{1}{P}\mathcal{L}(\mu^0) - \mathcal{L}(\pi^*) = O(d), \quad \mathsf{FI}(\mu^0\|\mu_*^{(P)}) = O(dP), \quad \text{and} \quad \mathbb{E}[\|\mathbf{x}_{\mathrm{init}}\|^2] = O(dP).$$

By exploiting the convexity of the Fisher divergence in its first argument and following a similar argument to Theorem 3.7, the following result for Algorithm 2 is obtained.

Theorem 4.2. Suppose Assumption 2.1 holds, $Lh \le \frac{1}{20}$, and $J \ge N$. Then, Algorithm 2 performs $N + J$ iterations and processes at most PMJ queries per iteration, and for $\Delta_0^{(P)} = \frac{1}{P}\mathcal{L}^{(P)}(\mu^0) - \mathcal{L}(\pi^*)$, its output distribution $\rho = \overline{\mu}_{Nh} = \frac{1}{Nh}\int_0^{Nh} \mu_t \mathrm{d}t$ satisfies

$$\mathsf{FI}(\rho\|\mu_*^{(P)}) \lesssim \frac{P\Delta_0^{(P)}}{Nh} + \frac{LdP}{M} + (\mathsf{FI}(\mu^0\|\mu_*^{(P)}) + LdP)0.5^{J-N}.$$

Theorem 4.2 illustrates a fundamental limitation of parallel computation for finite-particle MFLD. Although the time-averaged distribution $\overline{\mu}_{Nh}$ converges to the finite-particle stationary distribution $\mu_*^{(P)}$ in the sense of FI, this does not imply convergence to the target distribution π^* in general. That is, without additional assumptions, convergence to the target distribution is not guaranteed.

4.3 Analysis of Algorithm 2′: $O(\log d)$ Adaptive Complexity under Defective LSI

Since finite-particle MFLD is inherently an approximate dynamics, the output distribution does not converge to the target distribution π^* in general. In this section, we show that, as in the sequential setting, convergence to the target distribution can be established under conditions ensuring the defective LSI. By an argument analogous to that of Theorem 3.4, with the defective LSI (Lemma 2.7), we obtain the following.

Theorem 4.3. Suppose Assumptions 2.1 and 2.6 hold, $Lh \leq \frac{1}{20}$, and $J \geq N$. Then, Algorithm 2′ performs $N + J$ iterations and processes at most PMJ queries per iteration, and for $\Delta_0^{(P)} = \frac{1}{P}\mathcal{L}^{(P)}(\mu^0) - \mathcal{L}(\pi^*)$, the output distribution $\rho = \mu_{N,0}^J$ satisfies

$$\frac{1}{P}\mathcal{L}^{(P)}(\rho) - \mathcal{L}(\pi^*) \lesssim \exp(-\alpha N h)\Delta_0^{(P)} + \frac{1}{P} + \frac{Ld}{M} + \frac{1}{P}(\mathsf{FI}(\mu^0 \| \mu_*^{(P)}) + LdP)0.5^{J-N}.$$

Compared with the convergence rate for the sequential setting shown in Theorem 3.8 of [NLK$^+$25], Theorem 4.3 introduces a new parallelization error term, the fourth term on the right-hand side, which decreases exponentially in $J - N$.

Corollary 4.4. With initialization condition (Lemma 4.1), by taking

$P = O(\frac{1}{\varepsilon^2}), h = O(\frac{1}{L}), N = O(L\log(\frac{\Delta_0^{(P)}}{\varepsilon^2})), J - N = O(\log(\mathsf{FI}(\mu^0 \| \mu_*^{(P)}) + \frac{Ld}{\varepsilon^2})), M = O(\frac{Ld}{\varepsilon^2})$, the algorithm runs $O(\log(\frac{d}{\varepsilon^2}))$ iterations with $O(\frac{d}{\varepsilon^4}\log(\frac{d}{\varepsilon^2}))$ queries per iteration and returns a sample satisfying $\frac{1}{P}\mathcal{L}^{(P)}(\rho) - \mathcal{L}(\pi^*) \leq \varepsilon^2$.

Remark 4.5 (**Comparison with [NLK$^+$25]).** In the finite-particle setting, Theorem 4.3 shows that the adaptive complexity is reduced from $O(\frac{d}{\varepsilon^4}\log(\frac{d}{\varepsilon^2}))$ for the sequential algorithm to $O(\log(\frac{d}{\varepsilon^2}))$ by Algorithm 2′ (see Table 2).

5 Conclusion and Discussion

In this work, based on the algorithm of [ZS25a], we proposed the first parallel algorithms for MFLD in both the infinite-particle and finite-particle settings, and established convergence guarantees under various conditions including the LSI of the stationary distribution, uniform LSI, defective LSI, and without any LSI assumption. Moreover, we demonstrated that the proposed algorithms achieve a speedup over sequential algorithms due to parallelization.

We discuss limitations of this work and directions for future research. First, in the finite-particle setting, there is potential to refine the PoC framework further, which could weaken the defective LSI assumption required in Theorem 4.3 and improve the convergence rate. Second, the guarantees in the setting without any LSI assumption remain limited. While this work bounds the time-averaged FI, [ZS25a] establishes stronger guarantees for GLD by exploiting the invariance of the proximal Gibbs distribution and convexity, obtaining bounds on the FI between the time-averaged distribution and its proximal Gibbs distribution. Whether analogous results hold for MFLD remains an open question. Finally, numerical verification has not been conducted. In particular, verifying whether parallelization can achieve speedup in the LoRA-based LLM fine-tuning setting of [NLK$^+$25] is an important direction for future work.

Acknowledgments

YO was partially supported by JST CREST (JPMJCR2115). TS was partially supported by JSPS KAKENHI (24K02905, 25H01107). This work was partially supported by JST ERATO (JPMJER2601). This research is supported by the National Research Foundation, Singapore, Infocomm Media Development Authority under its Trust Tech Funding Initiative, and the Ministry of Digital Development and Information under the AI Visiting Professorship Programme (award number AIVP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Infocomm Media Development Authority, and the Ministry of Digital Development and Information.

References

[ACV24] Nima Anari, Sinho Chewi, and Thuy-Duong Vuong. Fast parallel sampling under isoperimetry. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 161–185. PMLR, 2024.

[BCE+22] Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling:first-order stationarity guarantees for Langevin Monte Carlo. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2896–2923. PMLR, 2022.

[BÉ85] D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206, 1985.

[CB18] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[CEL+22] Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of Langevin Monte Carlo from Poincare to log-Sobolev. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1–2. PMLR, 2022.

[CGLL23] Sinho Chewi, Patrik Gerber, Holden Lee, and Chen Lu. Fisher information lower bounds for sampling. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 375–410. PMLR, 2023.

[Chi22] L'enaic Chizat. Mean-field Langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

[CRW25] Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field Langevin dynamics. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 61(4):2357 – 2404, 2025.

[DG19] Jelena Diakonikolas and Cristóbal Guzmán. Lower bounds for parallel and randomized convex optimization. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1132–1157. PMLR, 2019.

[GL13] Nicola Gigli and M. Ledoux. From log Sobolev to Talagrand: A quick proof. *Discrete and continuous dynamical systems*, 33(5):1927–1935, 2013.

[HRŠS21] Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(4):2043 – 2065, 2021.

[HS87] Richard Holley and Daniel W. Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[NLK+25] Atsushi Nitanda, Anzelle Lee, Damian Tan Xing Kai, Mizuki Sakaguchi, and Taiji Suzuki. Propagation of chaos for mean-field Langevin dynamics and its application to model ensemble. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 46586–46610. PMLR, 2025.

[NS17] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles, 2017.

[NWS22] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field Langevin dynamics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9741–9757. PMLR, 2022.

[OV00] 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.

[RV22] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, volume 32, pages 2100–2111. Curran Associates, Inc., 2019.

[SNW23] Taiji Suzuki, Atsushi Nitanda, and Denny Wu. Uniform-in-time propagation of chaos for the mean-field gradient langevin dynamics. In *The Eleventh International Conference on Learning Representations*, 2023.

[SS20] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[Szn91] Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

[YD25] Lu Yu and Arnak Dalalyan. Parallelized midpoint randomization for Langevin Monte Carlo. *Stochastic Processes and their Applications*, 190:104764, 2025.

[YOYS24] Kakei Yamamoto, Kazusato Oko, Zhuoran Yang, and Taiji Suzuki. Mean field Langevin actor-critic: Faster convergence and global optimality beyond lazy learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55706–55738. PMLR, 2024.

[ZS25a] Huanjian Zhou and Masashi Sugiyama. The adaptive complexity of minimizing relative Fisher information. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[ZS25b] Huanjian Zhou and Masashi Sugiyama. Parallel simulation for log-concave sampling and score-based diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 79192–79225. PMLR, 2025.

A Useful Facts

In this section, we list useful lemmas employed in the analysis. First, the following estimate concerning the proximal Gibbs distribution is known.

Lemma A.1 (Entropy sandwich, [NWS22, Chi22])**.** For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathsf{KL}(\mu\|\pi^*) \le \mathcal{L}(\mu) - \mathcal{L}(\pi^*) \le \mathsf{KL}(\mu\|\pi[\mu])$$

holds.

Moreover, the following two sufficient conditions for the LSI are known.

Lemma A.2 (Bakry–Émery criterion, [BÉ85])**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-strongly convex, and let the probability measure p satisfy $p(x) \propto \exp(-f(x))$. Then p satisfies $\mathrm{LSI}(\alpha)$.

Lemma A.3 (Holley–Stroock bounded perturbation lemma, [HS87])**.** Let the probability measure q satisfy $\mathrm{LSI}(\alpha')$, and let $h : \mathbb{R}^d \to \mathbb{R}$ satisfy $\|h\|_\infty \le B$ for some $B > 0$. Then the probability measure $p(x) \propto q(x)\exp(h(x))$ satisfies the LSI with constant $\alpha = \alpha' \exp(-4B)$.

The Bakry–Émery criterion shows, in particular for GLD, that if the potential function V is strongly convex, then the stationary distribution $\pi^*(x) \propto \exp(-V(x))$ satisfies the LSI. The Holley–Stroock bounded perturbation lemma states that a probability distribution obtained by applying a bounded perturbation to a distribution satisfying the LSI also satisfies the LSI.

We introduce the following Talagrand inequality.

Definition A.4 (Talagrand inequality)**.** A distribution $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is said to satisfy the Talagrand inequality if there exists $\beta > 0$ such that for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mu \ll \nu$,

$$W_2^2(\mu,\nu) \le \frac{2}{\beta}\mathsf{KL}(\mu\|\nu).$$

In this case, we write that ν satisfies $\mathrm{T}(\beta)$.

This does not hold in general, but when ν satisfies the log-Sobolev inequality, the Talagrand inequality is automatically guaranteed to hold.

Lemma A.5 ([OV00, GL13])**.** If $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ satisfies $\mathrm{LSI}(\alpha)$, then ν satisfies $\mathrm{T}(\alpha)$. Moreover, for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mu \ll \nu$,

$$W_2^2(\mu,\nu) \le \frac{1}{\alpha^2}\mathsf{FI}(\mu\|\nu)$$

holds.

B Missing Proofs in Section 3

B.1 Proof of Theorem 3.2

In this section, we show that under Assumption 2.4, the distribution $\mu_{N,0}^J$ of the output $x_{N,0}^J$ of Algorithm 1 converges to the stationary distribution π^*. We use the following estimate of the Fisher divergence for Lipschitz continuous gradients.

Lemma B.1 (Lemma 20 in [CEL$^+$22])**.** Suppose ∇V is L-Lipschitz continuous. Then for any probability distribution μ,

$$\mathbb{E}_\mu\left[\|\nabla V\|^2\right] \le \mathsf{FI}(\mu\|\pi) + 2dL$$

holds, where $\pi(x) \propto \exp(-V(x))$.

Corollary B.2. Under Assumption 2.1, for any probability distributions μ, ν,

$$\mathbb{E}_\mu\left[\left\|\nabla\frac{\delta F(\nu)}{\delta\mu}\right\|^2\right] \le \mathsf{FI}(\mu\|\pi[\nu]) + 2dL_2$$

holds.

Proof. By Assumption 2.1, $x \mapsto \frac{\delta F(\nu)}{\delta\mu}(x)$ is L_2-Lipschitz continuous. Hence, setting $V(x) = \frac{\delta F(\nu)}{\delta\mu}(x)$, since $\pi[\nu](x) \propto \exp(-V(x))$, the result follows by applying Lemma B.1. □

We first evaluate the error arising from increasing iterations at the same time point.

Lemma B.3. Under $Lh \le \frac{1}{20}$ and Assumption 2.1, the following holds:

1. $\mathcal{E}_n^j \le 2\mathcal{E}_{n-1}^j + 8L^2h^2\mathcal{E}_n^{j-1} \quad (j = 2, \dots, J, n = 1, \dots, N-1)$,
2. $\mathcal{E}_n^1 \le 2\mathcal{E}_{n-1}^1 + 2h^2\mathsf{FI}(\mu^0\|\pi^0) + 5dh \quad (j = 1, n = 1, \dots, N-1)$,
3. $\mathcal{E}_0^j \le 4L^2h^2\mathcal{E}_0^{j-1} \quad (j = 2, \dots, J, n = 0)$,
4. $\mathcal{E}_0^1 \le 2h^2\mathsf{FI}(\mu^0\|\pi^0) + 5dh \quad (j = 1, n = 0)$,

where $\mathcal{E}_n^j = \max_{m=0,1,\dots,M} \mathbb{E}\left[\left\|x_{n,m}^j - x_{n,m}^{j-1}\right\|^2\right]$ for $j = 1, \dots, J,\ n = 0, \dots, N-1$.

Proof. 1. For $j = 2, 3, \dots, J,\ n = 1, 2, \dots, N-1,\ m = 0, 1, \dots, M$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|x_{n,m}^j - x_{n,m}^{j-1}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\left(x_{n,0}^j - \frac{h}{M}\sum_{m'=0}^{m-1}\nabla\frac{\delta F(\mu_{n,m'}^{j-1})}{\delta\mu}(x_{n,m'}^{j-1}) + \sqrt{2}(B_{nh+\frac{mh}{M}} - B_{nh})\right)\right.\right. \\
&\qquad \left.\left. - \left(x_{n,0}^{j-1} - \frac{h}{M}\sum_{m'=0}^{m-1}\nabla\frac{\delta F(\mu_{n,m'}^{j-2})}{\delta\mu}(x_{n,m'}^{j-2}) + \sqrt{2}(B_{nh+\frac{mh}{M}} - B_{nh})\right)\right\|^2\right] \\
&\le 2\mathbb{E}\left[\left\|x_{n,0}^j - x_{n,0}^{j-1}\right\|^2\right] \\
&\qquad + 2\frac{h^2}{M^2}\mathbb{E}\left[\left\|\sum_{m'=0}^{m-1}\left(\nabla\frac{\delta F(\mu_{n,m'}^{j-1})}{\delta\mu}(x_{n,m'}^{j-1}) - \nabla\frac{\delta F(\mu_{n,m'}^{j-2})}{\delta\mu}(x_{n,m'}^{j-2})\right)\right\|^2\right] \\
&\le 2\mathbb{E}\left[\left\|x_{n,0}^j - x_{n,0}^{j-1}\right\|^2\right] \\
&\qquad + 2h^2\max_{m'=0,\dots,M}\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu_{n,m'}^{j-1})}{\delta\mu}(x_{n,m'}^{j-1}) - \nabla\frac{\delta F(\mu_{n,m'}^{j-2})}{\delta\mu}(x_{n,m'}^{j-2})\right\|^2\right] \\
&\le 2\mathbb{E}\left[\left\|x_{n,0}^j - x_{n,0}^{j-1}\right\|^2\right] + 4h^2\max_{m'=0,\dots,M}(L_1^2W_2^2(\mu_{n,m'}^{j-1}, \mu_{n,m'}^{j-2}) + L_2^2\mathbb{E}\left[\left\|x_{n,m'}^{j-1} - x_{n,m'}^{j-2}\right\|\right])
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\mathbb{E}\left[\left\|x_{n,0}^{j}-x_{n,0}^{j-1}\right\|^2\right]+4(L_1^2+L_2^2)h^2\max_{m'=0,\ldots,M}\mathbb{E}\left[\left\|x_{n,m'}^{j-1}-x_{n,m'}^{j-2}\right\|\right]\\
&=2\mathbb{E}\left[\left\|x_{n-1,M}^{j}-x_{n-1,M}^{j-1}\right\|^2\right]+4(L_1^2+L_2^2)h^2\max_{m'=0,\ldots,M}\mathbb{E}\left[\left\|x_{n,m'}^{j-1}-x_{n,m'}^{j-2}\right\|\right]\\
&\leq 2\mathcal{E}_{n-1}^{j}+8L^2h^2\mathcal{E}_{n}^{j-1},
\end{aligned}
$$

which gives the claim.

2. For $n=1,\ldots,N-1,\ m=0,\ldots,M$, using Lemma B.1,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|x_{n,m}^{1}-x_{n,m}^{0}\right\|^2\right]\\
&\leq 2\mathbb{E}\left[\left\|x_{n,0}^{1}-x_{n,0}^{0}\right\|^2\right]+2h^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu^0)}{\delta\mu}(x_{\mathrm{init}})\right\|^2\right]+4dh\\
&\leq 2\mathbb{E}\left[\left\|x_{n,0}^{1}-x_{n,0}^{0}\right\|^2\right]+2h^2\left(\mathsf{FI}(\mu^0\|\pi^0)+2dL_2\right)+4dh\\
&\leq 2\mathcal{E}_{n-1}^{1}+2h^2\mathsf{FI}(\mu^0\|\pi^0)+5dh.
\end{aligned}
$$

3. For $j=2,3,\ldots,J,\ m=0,1,\ldots,M$, using $x_{0,0}^{j}=x_{0,0}^{j-1}=x_{\mathrm{init}}$,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|x_{0,m}^{j}-x_{0,m}^{j-1}\right\|^2\right]\\
&=\frac{h^2}{M^2}\mathbb{E}\left[\left\|\sum_{m'=0}^{m-1}\left(\nabla\frac{\delta F(\mu_{0,m'}^{j-1})}{\delta\mu}(x_{0,m'}^{j-1})-\nabla\frac{\delta F(\mu_{0,m'}^{j-2})}{\delta\mu}(x_{0,m'}^{j-2})\right)\right\|^2\right]\\
&\leq h^2\max_{m'=0,1,\ldots,M}\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu_{0,m'}^{j-1})}{\delta\mu}(x_{0,m'}^{j-1})-\nabla\frac{\delta F(\mu_{0,m'}^{j-2})}{\delta\mu}(x_{0,m'}^{j-1})\right\|^2\right]\\
&\leq 2h^2\max_{m'=0,1,\ldots,M}(L_1^2W_2^2(\mu_{0,m'}^{j-1},\mu_{0,m'}^{j-2})+L_2^2\mathbb{E}\left[\left\|x_{0,m'}^{j-1}-x_{0,m'}^{j-2}\right\|^2\right])\\
&\leq 2h^2(L_1^2+L_2^2)\max_{m'=0,1,\ldots,M}\mathbb{E}\left[\left\|x_{0,m'}^{j-1}-x_{0,m'}^{j-2}\right\|^2\right]\\
&\leq 4L^2h^2\mathcal{E}_{0}^{j-1},
\end{aligned}
$$

which gives the claim.

4. For $m=0,\ldots,M-1$,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|x_{0,m}^{1}-x_{0,m}^{0}\right\|^2\right]\\
&\leq 2h^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu^0)}{\delta\mu}(x_{\mathrm{init}})\right\|^2\right]+4\mathbb{E}\left[\left\|B_{\frac{mh}{M}}\right\|^2\right]\\
&\leq 2h^2(\mathsf{FI}(\mu^0\|\pi^0)+2dL_2)+4dh\\
&\leq 2h^2\mathsf{FI}(\mu^0\|\pi^0)+5dh.
\end{aligned}
$$

This completes the proof. □

Using this lemma, we can establish the following uniform estimate on the error when the number of iterations is increased. In particular, as can be seen from the coefficient 0.5^{J-N}, the parallelization error decreases exponentially with the degree of parallelism.

Lemma B.4. Under Assumption 2.1 and $Lh\leq\frac{1}{20}$,

$$
\mathcal{E}\leq 200\left(2h^2\mathsf{FI}(\mu^0\|\pi^0)+5dh\right)0.5^{J-N},
$$

where $\mathcal{E}=\max_{n=0,\ldots,N-1}\mathcal{E}_n^J$.

Proof. From Lemma B.3, in particular,

1. $\mathcal{E}_n^j \le 2\mathcal{E}_{n-1}^j + 0.02h^2\mathcal{E}_n^{j-1} \quad (j = 2, \dots, J, n = 1, \dots, N-1)$,
2. $\mathcal{E}_n^1 \le 2\mathcal{E}_{n-1}^1 + 2h^2\mathsf{FI}(\mu^0\|\pi^0) + 5dh \quad (j = 1, n = 1, \dots, N-1)$,
3. $\mathcal{E}_0^j \le 0.01\mathcal{E}_0^{j-1} \quad (j = 2, \dots, J, n = 0)$,
4. $\mathcal{E}_0^1 \le 2h^2\mathsf{FI}(\mu^0\|\pi^0) + 5dh \quad (j = 1, n = 0)$.

We follow the same argument as Lemma A.5 in [ZS25a].
Setting $A = 2h^2\mathsf{FI}(\mu^0\|\pi^0) + 5dh$, for $n \ge 1$ and $j = 1$,

$$\mathcal{E}_n^1 \le 2\mathcal{E}_{n-1}^1 + A$$

implies

$$\mathcal{E}_n^1 + A \le 2(\mathcal{E}_{n-1}^1 + A) \le 2^n(\mathcal{E}_0^1 + A),$$

and by the estimate in 4.,

$$\mathcal{E}_n^1 \le 2^n\mathcal{E}_0^1 + (2^n - 1)A \le 2^{n+1}A.$$

For $n = 0$ and $j \le 2$, by repeatedly applying 3. and then using the estimate in 4.,

$$\mathcal{E}_0^j \le 0.01^{j-1}A.$$

Using these together with the estimate in 1., $\binom{m}{n} \le \left(\frac{em}{n}\right)^n$, and $\left(1 + \frac{m}{n}\right)^n \le e^m$, for $j \ge n$,

$$\begin{aligned}
\mathcal{E}_n^j &\le \sum_{a=2}^{n} 0.02^{j-1}2^{n-a}\binom{n-a+j-2}{j-2}\mathcal{E}_a^1 + \sum_{b=2}^{j} 0.02^{j-b}2^n\binom{n+j-b}{j-b}\mathcal{E}_0^b \\
&\le A\left(\sum_{a=2}^{n} 0.02^{j-1}2^{n+1}\binom{n-a+j-2}{j-2} + \sum_{b=2}^{j}\binom{n+j-b}{j-b}0.02^{j-b}2^n 0.01^{b-1}\right) \\
&\le 0.02^{j-1}2^{n+1}A\left(\sum_{a=2}^{n}\binom{n-a+j-2}{j-2} + \sum_{b=2}^{j}\binom{n+j-b}{j-b}\right) \\
&\le 0.02^{j-1}2^{n+1}A\left(\sum_{a=2}^{n} e^{j-2}\left(\frac{n-a+j-2}{j-2}\right)^{j-2} + \sum_{b=2}^{j} e^{j-b}\left(\frac{n+j-b}{j-b}\right)^{j-b}\right) \\
&\le 0.02^{j-1}2^{n+1}A\left(\sum_{a=2}^{n} e^{j-2}\left(\frac{n-a+j-2}{j-2}\right)^{j-2} + \sum_{i=0}^{j-2} e^{j-2}\left(\frac{n+i}{i}\right)^{i}\right) \\
&\le 0.02^{j-1}2^{n+1}A\left(\sum_{a=2}^{n} e^{j-2}e^{n-2} + \sum_{i=0}^{j-2} e^{j-2}e^{n-2}\right) \\
&\le 0.02^{j-1}2^{n+1}e^{2j}(2j)A \\
&\le 200 \cdot 0.5^j 2^n A,
\end{aligned}$$

where the last inequality uses $\left(\frac{e}{5}\right)^{2j} \cdot j \le 1$, i.e., $0.02^j e^{2j} j \le 0.5^j$. This yields the claim. □

Next, we evaluate the distribution of the interpolation x_t of $x_{n,m}^J$. We use the following lemma concerning the one-step interpolation.

Lemma B.5 (Lemma 12 in [BCE+22])**.** Let $(B_t)_{t\ge 0}$ be a standard Brownian motion independent of (x_0, g_0). For

$$x_t = x_0 - tg_0 + \sqrt{2}B_t \quad (t \ge 0),$$

we have

$$\partial_t\mathsf{KL}(\mu_t\|\pi) \le -\frac{3}{4}\mathsf{FI}(\mu_t\|\pi) + \mathbb{E}\left[\|\nabla V(x) - g_0\|^2\right],$$

where $\pi(x) \propto \exp(-V(x))$.

Proof. Let $\mathcal{F}_0$ be the σ-algebra generated by (x_0, g_0). The Fokker–Planck equation for the conditional distribution $\mu_{t|\mathcal{F}_0}(x, \omega)$ of x_t given $\mathcal{F}_0$, using Itô's formula, independence of B_t, and the martingale

property of the stochastic integral, is

$$\partial_t \mu_{t|\mathcal{F}_0}(x,\omega) = \Delta \mu_{t|\mathcal{F}_0}(x,\omega) + \nabla \cdot (\mu_{t|\mathcal{F}_0}(x,\omega) g_0).$$

Taking expectations on both sides,

$$\partial_t \mu_t(x) = \Delta \mu_t(x) + \nabla \cdot (\mathbb{E}\left[\mu_{t|\mathcal{F}_0}(x,\omega) g_0\right]).$$

For the second term, Lebesgue-a.e.,

$$\begin{aligned}
\mathbb{E}\left[\mu_{t|\mathcal{F}_0}(x,\omega) g_0\right] &= \int \mu_{t|\mathcal{F}_0}(x,\omega) g_0(\omega) \mathrm{dP}(\omega) \\
&= \mu_t(x) \int g_0(\omega) \mu_{\mathcal{F}_0|t}(\mathrm{d}\omega|x) \\
&= \mu_t(x) \mathbb{E}[g_0 | x_t = x]
\end{aligned}$$

holds, so that

$$\partial_t \mu_t(x) = \Delta \mu_t(x) + \nabla \cdot (\mu_t(x) \mathbb{E}\left[g_0 | x_t = x\right]).$$

From this, the time derivative of the KL divergence can be computed as

$$\begin{aligned}
\partial_t \mathsf{KL}(\mu_t \| \pi) &= \partial_t \int \left(\log \frac{\mu_t}{\pi}\right) \mu_t \mathrm{d}x \\
&= \int \left(\partial_t \mu_t + \left(\log \frac{\mu_t}{\pi}\right) \partial_t \mu_t\right) \mathrm{d}x \\
&= \int \left(\log \frac{\mu_t}{\pi}\right) \partial_t \mu_t \mathrm{d}x \\
&= \int \left(\log \frac{\mu_t}{\pi}\right) \nabla \cdot (\mu_t \left(\nabla \log \mu_t + \mathbb{E}\left[g_0 | x_t = x\right]\right)) \mathrm{d}x \\
&= -\int \left\langle \nabla \log \frac{\mu_t}{\pi}, \nabla \log \mu_t + \mathbb{E}\left[g_0 | x_t = x\right] \right\rangle \mathrm{d}\mu_t(x) \\
&= -\mathsf{FI}(\mu_t \| \pi) + \int \left\langle \nabla \log \frac{\mu_t}{\pi}, \nabla V - \mathbb{E}\left[g_0 | x_t = x\right] \right\rangle \\
&\leq -\frac{3}{4} \mathsf{FI}(\mu_t \| \pi) + \mathbb{E}\left[\left\| \nabla V(x_t) - \mathbb{E}\left[g_0 | x_t\right] \right\|^2\right] \\
&\leq -\frac{3}{4} \mathsf{FI}(\mu_t \| \pi) + \mathbb{E}\left[\left\| \nabla V(x_t) - g_0 \right\|^2\right],
\end{aligned}$$

where Jensen's inequality and Young's inequality were used. This completes the proof. □

Using this, we can bound the time evolution of the KL divergence as follows.

Lemma B.6. Under $Lh \leq \frac{1}{6\sqrt{6}}$ and Assumptions 2.1 and 2.4,

$$\begin{aligned}
\partial_t \mathsf{KL}(\mu_t \| \pi^*) \leq -2\alpha\beta \mathsf{KL}(\mu_t \| \pi^*) &+ 54 L^3 d \left(t - nh - \frac{mh}{M}\right)^2 \\
&+ 27 L^2 d \left(t - nh - \frac{mh}{M}\right) + \frac{27}{2} L^2 \mathbb{E}\left[\left\| x^J_{n,m} - x^{J-1}_{n,m} \right\|^2\right]
\end{aligned}$$

holds, where $\beta = \frac{3}{4} - \frac{9L_1^2}{2\alpha^2} - 27L^2 \frac{h^2}{M^2}$.

Proof. By Lemma B.5, for $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$,

$$\partial_t \mathsf{KL}(\mu_t \| \pi^*) \leq -\frac{3}{4} \mathsf{FI}(\mu_t \| \pi^*) + \mathbb{E}\left[\left\| \nabla \frac{\delta F(\pi^*)}{\delta \mu}(x_t) - \nabla \frac{\delta F(\mu^{J-1}_{n,m})}{\delta \mu}(x^{J-1}_{n,m}) \right\|^2\right].$$

Here,

$$\begin{aligned}
&\mathbb{E}\left[\left\| x_t - x^{J-1}_{n,m} \right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\| x_t - x^J_{n,m} \right\|^2\right] + 2\mathbb{E}\left[\left\| x^J_{n,m} - x^{J-1}_{n,m} \right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= 2\mathbb{E}\left[\left\|-\left(\left(t-nh-\frac{mh}{M}\right)\nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})+\sqrt{2}(B_t-B_{nh+\frac{mh}{M}})\right)\right\|^2\right]\\
&\quad+2\mathbb{E}\left[\left\|x_{n,m}^J-x_{n,m}^{J-1}\right\|^2\right]\\
&\le 2\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right]+4\mathbb{E}\left[\left\|B_t-B_{nh+\frac{mh}{M}}\right\|^2\right]\\
&\quad+2\mathbb{E}\left[\left\|x_{n,m}^J-x_{n,m}^{J-1}\right\|^2\right]\\
&\le 4\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\pi^*)}{\delta\mu}(x_t)-\nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right]\\
&\quad+4\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\pi^*)}{\delta\mu}(x_t)\right\|^2\right]+4\mathbb{E}\left[\left\|B_t-B_{nh+\frac{mh}{M}}\right\|^2\right]\\
&\quad+2\mathbb{E}\left[\left\|x_{n,m}^J-x_{n,m}^{J-1}\right\|^2\right],
\end{aligned}
$$

where we used the independence of $x_{n,m}^{J-1}$ and $B_t-B_{nh+\frac{mh}{M}}$. By this and Lemma B.1,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\nabla\frac{\delta F(\pi^*)}{\delta\mu}(x_t)-\nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right]\\
&\le 2L_1^2W_2^2(\mu_{n,m}^{J-1},\pi^*)+2L_2^2\mathbb{E}\left[\left\|x_t-x_{n,m}^{J-1}\right\|^2\right]\\
&\le 4L_1^2(W_2^2(\mu_{n,m}^{J-1},\mu_t)+W_2^2(\mu_t,\pi^*))+2L_2^2\mathbb{E}\left[\left\|x_t-x_{n,m}^{J-1}\right\|^2\right]\\
&\le \frac{4L_1^2}{\alpha^2}\mathsf{FI}(\mu_t\|\pi^*)+(4L_1^2+2L_2^2)\mathbb{E}\left[\left\|x_t-x_{n,m}^{J-1}\right\|^2\right]\\
&\le \frac{4L_1^2}{\alpha^2}\mathsf{FI}(\mu_t\|\pi^*)+6L^2\mathbb{E}\left[\left\|x_t-x_{n,m}^{J-1}\right\|^2\right]\\
&\le \frac{4L_1^2}{\alpha^2}\mathsf{FI}(\mu_t\|\pi^*)+24L^2\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\pi^*)}{\delta\mu}(x_t)-\nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right]\\
&\quad+24L^2\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\pi^*)}{\delta\mu}(x_t)\right\|^2\right]+24L^2\mathbb{E}\left[\left\|B_t-B_{nh+\frac{mh}{M}}\right\|^2\right]\\
&\quad+12L^2\mathbb{E}\left[\left\|x_{n,m}^J-x_{n,m}^{J-1}\right\|^2\right]\\
&\le \frac{4L_1^2}{\alpha^2}\mathsf{FI}(\mu_t\|\pi^*)+24L^2\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\pi^*)}{\delta\mu}(x_t)-\nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right]\\
&\quad+24L^2\left(t-nh-\frac{mh}{M}\right)^2(\mathsf{FI}(\mu_t\|\pi^*)+2dL_2)+24L^2d\left(t-nh-\frac{mh}{M}\right)\\
&\quad+12L^2\mathbb{E}\left[\left\|x_{n,m}^J-x_{n,m}^{J-1}\right\|^2\right]\\
&\le \left(\frac{4L_1^2}{\alpha^2}+24L^2\left(t-nh-\frac{mh}{M}\right)^2\right)\mathsf{FI}(\mu_t\|\pi^*)\\
&\quad+24L^2\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\pi^*)}{\delta\mu}(x_t)-\nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right]
\end{aligned}
$$

$$+ 48L^3 d\left(t - nh - \frac{mh}{M}\right)^2 + 24L^2 d\left(t - nh - \frac{mh}{M}\right)$$
$$+ 12L^2 \mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right].$$

Since $t - nh - \frac{mh}{M} \leq \frac{h}{M} \leq h$,

$$\left(1 - 24L^2h^2\right)\mathbb{E}\left[\left\|\nabla \frac{\delta F(\pi^*)}{\delta \mu}(x_t) - \nabla \frac{\delta F(\mu_{n,m}^{J-1})}{\delta \mu}(x_{n,m}^{J-1})\right\|^2\right]$$
$$\leq \left(\frac{4L_1^2}{\alpha^2} + 24L^2\frac{h^2}{M^2}\right)\mathsf{FI}(\mu_t\|\pi^*) + 48L^3 d\left(t - nh - \frac{mh}{M}\right)^2$$
$$+ 24L^2 d\left(t - nh - \frac{mh}{M}\right) + 12L^2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right].$$

Choosing $\frac{8}{9} \leq 1 - 24L^2h^2$, i.e., $Lh \leq \frac{1}{6\sqrt{6}}$,

$$\mathbb{E}\left[\left\|\nabla \frac{\delta F(\pi^*)}{\delta \mu}(x_t) - \nabla \frac{\delta F(\mu_{n,m}^{J-1})}{\delta \mu}(x_{n,m}^{J-1})\right\|^2\right]$$
$$\leq \left(\frac{9L_1^2}{2\alpha^2} + 27L^2\frac{h^2}{M^2}\right)\mathsf{FI}(\mu_t\|\pi^*) + 54L^3 d\left(t - nh - \frac{mh}{M}\right)^2$$
$$+ 27L^2 d\left(t - nh - \frac{mh}{M}\right) + \frac{27}{2}L^2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right].$$

Therefore,

$$\partial_t \mathsf{KL}(\mu_t\|\pi^*)$$
$$\leq -\frac{3}{4}\mathsf{FI}(\mu_t\|\pi^*) + \left(\frac{9L_1^2}{2\alpha^2} + 27L^2\frac{h^2}{M^2}\right)\mathsf{FI}(\mu_t\|\pi^*) + 54L^3 d\left(t - nh - \frac{mh}{M}\right)^2$$
$$+ 27L^2 d\left(t - nh - \frac{mh}{M}\right) + \frac{27}{2}L^2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right]$$
$$= -\beta\mathsf{FI}(\mu_t\|\pi^*) + 54L^3 d\left(t - nh - \frac{mh}{M}\right)^2$$
$$+ 27L^2 d\left(t - nh - \frac{mh}{M}\right) + \frac{27}{2}L^2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right]$$
$$\leq -2\alpha\beta\mathsf{KL}(\mu_t\|\pi^*) + 54L^3 d\left(t - nh - \frac{mh}{M}\right)^2$$
$$+ 27L^2 d\left(t - nh - \frac{mh}{M}\right) + \frac{27}{2}L^2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right],$$

which gives the claim. □

From these results, we can establish the following convergence.

Theorem B.7. Under Assumptions 2.1 and 2.4 with $Lh \leq \frac{1}{20}$,

$$\mathsf{KL}(\mu_{n,m}^J\|\pi^*) \leq \exp\left(-2\alpha\beta\left(nh + \frac{mh}{M}\right)\right)\mathsf{KL}(\mu_{0,0}^J\|\pi^*) + \delta,$$

where $\delta = \frac{1}{\alpha\beta}\left(27L^3 d\frac{h^2}{M^2} + \frac{27}{2}L^2 d\frac{h}{M} + \frac{27}{4}L^2\mathcal{E}\right)$.

Proof. Using the result of Lemma B.6, in particular,

$$\partial_t \mathsf{KL}(\mu_t\|\pi^*) \leq -2\alpha\beta\mathsf{KL}(\mu_t\|\pi^*) + 54L^3 d\left(t - nh - \frac{mh}{M}\right)^2$$

$$
\begin{aligned}
&\quad + 27L^2 d\left(t - nh - \frac{mh}{M}\right) + \frac{27}{2}L^2\mathcal{E} \\
&\leq -2\alpha\beta \mathsf{KL}(\mu_t\|\pi^*) + 54L^3 d\frac{h^2}{M^2} + 27L^2 d\frac{h}{M} + \frac{27}{2}L^2\mathcal{E} \\
&= -2\alpha\beta\left(\mathsf{KL}(\mu_t\|\pi^*) - \frac{c}{2\alpha\beta}\right),
\end{aligned}
$$

so that

$$
\partial_t\left(\mathsf{KL}(\mu_t\|\pi^*) - \frac{c}{2\alpha\beta}\right) \leq -2\alpha\beta\left(\mathsf{KL}(\mu_t\|\pi^*) - \frac{c}{2\alpha\beta}\right),
$$

where $c = 54L^3 d\frac{h^2}{M^2} + 27L^2 d\frac{h}{M} + \frac{27}{2}L^2\mathcal{E}$. By Grönwalls lemma, for $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$,

$$
\mathsf{KL}(\mu_t\|\pi^*) - \frac{c}{2\alpha\beta} \leq \exp\left(-2\alpha\beta\left(t - nh - \frac{mh}{M}\right)\right)\left(\mathsf{KL}\left(\mu^J_{n,m}\|\pi^*\right) - \frac{c}{2\alpha\beta}\right).
$$

In particular, setting $t = nh + \frac{mh}{M}$,

$$
\mathsf{KL}\left(\mu^J_{n,m+1}\|\pi^*\right) - \frac{c}{2\alpha\beta} \leq \exp\left(-2\alpha\beta\frac{h}{M}\right)\left(\mathsf{KL}\left(\mu^J_{n,m}\|\pi^*\right) - \frac{c}{2\alpha\beta}\right).
$$

Applying this repeatedly,

$$
\begin{aligned}
\mathsf{KL}\left(\mu^J_{n,m}\|\pi^*\right) - \frac{c}{2\alpha\beta} &\leq \exp\left(-2\alpha\beta\left(nh + \frac{mh}{M}\right)\right)\left(\mathsf{KL}(\mu^0\|\pi^*) - \frac{c}{2\alpha\beta}\right) \\
&\leq \exp\left(-2\alpha\beta\left(nh + \frac{mh}{M}\right)\right)\mathsf{KL}(\mu^0\|\pi^*),
\end{aligned}
$$

from which

$$
\mathsf{KL}\left(\mu^J_{n,m}\|\pi^*\right) \leq \exp\left(-2\alpha\beta\left(nh + \frac{mh}{M}\right)\right)\mathsf{KL}(\mu^0\|\pi^*) + \frac{c}{2\alpha\beta}.
$$

Setting $\delta = \frac{c}{2\alpha\beta}$ completes the proof. □

Theorem 3.2 follows from Theorem B.7 and Lemma B.3.

B.2 Proof of Theorem 3.7

First, corresponding to Lemma B.5 from the preceding section, the following estimate on the time evolution of the loss functional along the one-step interpolation holds.

Lemma B.8. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion independent of (x_0, g_0), and let

$$
x_t = x_0 - tg_0 + \sqrt{2}B_t \quad (t \geq 0).
$$

Then for the distribution μ_t of x_t and its proximal Gibbs distribution π_t,

$$
\partial_t\mathcal{L}(\mu_t) \leq -\frac{3}{4}\mathsf{FI}(\mu_t\|\pi_t) + \mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu_t)}{\delta\mu}(x_t) - g_0\right\|^2\right]
$$

holds.

Proof. As in the proof of Lemma B.5,

$$
\partial_t\mu_t(x) = \nabla\cdot(\mu_t(x)(\nabla\log\mu_t(x) + \mathbb{E}\left[g_0|x_t = x\right]))
$$

holds. Moreover,

$$
\frac{\delta\mathcal{L}(\mu_t)}{\delta\mu} = \frac{\delta F(\mu_t)}{\delta\mu} + \log\mu_t + \text{Const.} = \log\frac{\mu_t}{\pi_t},
$$

where we used the freedom of an additive constant in the first variation. By these facts together with Young's inequality and Jensen's inequality,

$$
\begin{aligned}
\partial_t\mathcal{L}(\mu_t) &= \int\frac{\delta\mathcal{L}(\mu_t)}{\delta\mu}(x)\frac{\partial\mu_t}{\partial t}(x)\mathrm{d}x \\
&= \int\left(\log\frac{\mu_t}{\pi_t}\right)\nabla\cdot\left\{\mu_t\left(\nabla\log\mu_t + \mathbb{E}\left[g_0|x_t = x\right]\right)\right\}\mathrm{d}x
\end{aligned}
$$

$$
\begin{aligned}
&= -\int \left\langle \nabla \log \frac{\mu_t}{\pi_t}, \nabla \log \mu_t + \mathbb{E}\left[g_0 | x_t = x\right] \right\rangle \mathrm{d}\mu_t(x) \\
&= -\int \left\langle \nabla \log \frac{\mu_t}{\pi_t}, \nabla \log \frac{\mu_t}{\pi_t} + \nabla \log \pi_t + \mathbb{E}\left[g_0 | x_t = x\right] \right\rangle \mathrm{d}\mu_t(x) \\
&= -\mathsf{FI}(\mu_t \| \pi_t) + \int \left\langle -\nabla \log \frac{\mu_t}{\pi_t}, \nabla \log \pi_t + \mathbb{E}\left[g_0 | x_t = x\right] \right\rangle \mathrm{d}\mu_t(x) \\
&\leq -\mathsf{FI}(\mu_t \| \pi_t) + \frac{1}{4} \int \left\| \nabla \log \frac{\mu_t}{\pi_t} \right\|^2 \mathrm{d}\mu_t + \int \left\| \nabla \log \pi_t + \mathbb{E}\left[g_0 | x_t = x\right] \right\|^2 \mathrm{d}\mu_t(x) \\
&= -\frac{3}{4}\mathsf{FI}(\mu_t \| \pi_t) + \mathbb{E}\left[\left\| \nabla \log \pi_t(x_t) + \mathbb{E}\left[g_0 | x_t\right] \right\|^2 \right] \\
&\leq -\frac{3}{4}\mathsf{FI}(\mu_t \| \pi_t) + \mathbb{E}\left[\left\| \nabla \log \pi_t(x_t) + g_0 \right\|^2 \right] \\
&= -\frac{3}{4}\mathsf{FI}(\mu_t \| \pi_t) + \mathbb{E}\left[\left\| \nabla \frac{\delta F(\mu_t)}{\delta \mu}(x_t) - g_0 \right\|^2 \right],
\end{aligned}
$$

which completes the proof. □

Using this, the following estimate on the time evolution along the one-step interpolation of Algorithm 1 can be established.

Lemma B.9. Under Assumption 2.1 and $Lh \leq \frac{1}{6\sqrt{6}}$, for $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$,

$$
\begin{aligned}
\partial_t \mathcal{L}(\mu_t) \leq &-\left(\frac{3}{4} - 36L^2\frac{h^2}{M^2}\right)\mathsf{FI}(\mu_t \| \pi_t) \\
&+ 36L^3 d\left(t - nh - \frac{mh}{M}\right)^2 + 18L^2 d\left(t - nh - \frac{mh}{M}\right) \\
&+ \left(108L^4\left(t - nh - \frac{mh}{M}\right)^2 + 9L^2\right)\mathbb{E}\left[\left\| x_{n,m}^J - x_{n,m}^{J-1} \right\|^2\right]
\end{aligned}
$$

holds.

Proof. By Lemma B.8, for $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$,

$$
\partial_t \mathcal{L}(\mu_t) \leq -\frac{3}{4}\mathsf{FI}(\mu_t \| \pi_t) + \mathbb{E}\left[\left\| \nabla \frac{\delta F(\mu_t)}{\delta \mu}(x_t) - \nabla \frac{\delta F(\mu_{n,m}^{J-1})}{\delta \mu}(x_{n,m}^{J-1}) \right\|^2 \right].
$$

Moreover,

$$
\begin{aligned}
\mathsf{FI}(\mu_t \| \pi_t) &= \int \left\| \nabla \log \mu_t - \nabla \log \pi_t \right\|^2 \mathrm{d}\mu_t \\
&\leq 2\int \left\| \nabla \log \mu_t - \nabla \log \pi_{n,m}^J \right\|^2 \mathrm{d}\mu_t + 2\int \left\| \nabla \log \pi_{n,m}^J - \nabla \log \pi_t \right\|^2 \mathrm{d}\mu_t \\
&= 2\mathsf{FI}(\mu_t \| \pi_{n,m}^J) + 2\int \left\| \nabla \frac{\delta F(\mu_{n,m}^J)}{\delta \mu} - \nabla \frac{\delta F(\mu_t)}{\delta \mu} \right\|^2 \mathrm{d}\mu_t \\
&\leq 2\mathsf{FI}(\mu_t \| \pi_{n,m}^J) + 2L_1^2 W_2^2(\mu_{n,m}^J, \mu_t) \\
&\leq 2\mathsf{FI}(\mu_t \| \pi_{n,m}^J) + 2L^2\mathbb{E}\left[\left\| x_t - x_{n,m}^J \right\|^2\right],
\end{aligned}
$$

and by Lemma B.1,

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \nabla \frac{\delta F(\mu_{n,m}^J)}{\delta \mu}(x_t) \right\|^2 \right] \\
&\leq \mathsf{FI}(\mu_t \| \pi_{n,m}^J) + 2dL_2
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2\mathsf{FI}(\mu_t\|\pi_t) + 2L_1^2\mathbb{E}\left[\left\|x_t - x_{n,m}^J\right\|^2\right] + 2dL_2 \\
&\leq 2\mathsf{FI}(\mu_t\|\pi_t) + 4L^2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right] + 4L^2\mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right] + 2dL.
\end{aligned}
$$

Also,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu_{n,m}^J)}{\delta\mu}(x_t) - \nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right] \\
&\leq 2L_1^2 W_2(\mu_{n,m}^J, \mu_{n,m}^{J-1}) + 2L_2^2\mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right] \\
&\leq 2L^2\left(\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right] + \mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right]\right).
\end{aligned}
$$

Using these estimates together with the same argument as in the proof of Lemma B.6,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right] \\
&\leq 4\left(t - nh - \frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu_{n,m}^J)}{\delta\mu}(x_t) - \nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right] \\
&\quad + 4\left(t - nh - \frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu_{n,m}^J)}{\delta\mu}(x_t)\right\|^2\right] + 4\mathbb{E}\left[\left\|B_t - B_{nh+\frac{mh}{M}}\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right] \\
&\leq 8L^2\left(t - nh - \frac{mh}{M}\right)^2\left(\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right] + \mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right]\right) \\
&\quad + 4\left(t - nh - \frac{mh}{M}\right)^2\left(2\mathsf{FI}(\mu_t\|\pi_t) + 4L^2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right]\right. \\
&\quad \left. + 4L^2\mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right] + 2dL\right) + 4d\left(t - nh - \frac{mh}{M}\right) + 2\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right] \\
&\leq 24L^2\left(t - nh - \frac{mh}{M}\right)^2\mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right] \\
&\quad + 8\left(t - nh - \frac{mh}{M}\right)^2\left(\mathsf{FI}(\mu_t\|\pi_t) + dL\right) \\
&\quad + 4d\left(t - nh - \frac{mh}{M}\right) + \left(24L^2\left(t - nh - \frac{mh}{M}\right)^2 + 2\right)\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&(1 - 24L^2h^2)\mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right] \\
&\leq 8\frac{h^2}{M^2}\mathsf{FI}(\mu_t\|\pi_t) + 8Ld\left(t - nh - \frac{mh}{M}\right)^2 \\
&\quad + 4d\left(t - nh - \frac{mh}{M}\right) + \left(24L^2\left(t - nh - \frac{mh}{M}\right)^2 + 2\right)\mathbb{E}\left[\left\|x_{n,m}^J - x_{n,m}^{J-1}\right\|^2\right],
\end{aligned}
$$

and for $1 - 24L^2h^2 \geq \frac{8}{9}$, i.e., $Lh \leq \frac{1}{6\sqrt{6}}$,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|x_t - x_{n,m}^{J-1}\right\|^2\right] \\
&\leq 9\frac{h^2}{M^2}\mathsf{FI}(\mu_t\|\pi_t) + 9Ld\left(t - nh - \frac{mh}{M}\right)^2
\end{aligned}
$$

$$+\frac{9}{2}d\left(t-nh-\frac{mh}{M}\right)+\left(27L^2\left(t-nh-\frac{mh}{M}\right)^2+\frac{9}{4}\right)\mathbb{E}\left[\left\|x_{n,m}^J-x_{n,m}^{J-1}\right\|^2\right].$$

Hence, using the initial inequality,

$$\begin{aligned}\partial_t\mathcal{L}(\mu_t)&\le-\frac{3}{4}\mathsf{FI}(\mu_t\|\pi_t)+\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mu_t)}{\delta\mu}(x_t)-\nabla\frac{\delta F(\mu_{n,m}^{J-1})}{\delta\mu}(x_{n,m}^{J-1})\right\|^2\right]\\&\le-\frac{3}{4}\mathsf{FI}(\mu_t\|\pi_t)+4L^2\mathbb{E}\left[\left\|x_t-x_{n,m}^{J-1}\right\|^2\right]\\&\le-\left(\frac{3}{4}-36L^2\frac{h^2}{M^2}\right)\mathsf{FI}(\mu_t\|\pi_t)\\&\quad+36L^3d\left(t-nh-\frac{mh}{M}\right)^2+18L^2d\left(t-nh-\frac{mh}{M}\right)\\&\quad+\left(108L^4\left(t-nh-\frac{mh}{M}\right)^2+9L^2\right)\mathbb{E}\left[\left\|x_{n,m}^J-x_{n,m}^{J-1}\right\|^2\right],\end{aligned}$$

which completes the proof. □

From this, the following convergence in the sense of the time-averaged Fisher divergence can be established.

Theorem B.10. Under Assumption 2.1 and $Lh\le\frac{1}{20}$,

$$\frac{1}{Nh}\int_0^{Nh}\mathsf{FI}(\mu_t\|\pi_t)\mathrm{d}t\le\frac{2\mathsf{KL}(\mu^0\|\pi^0)}{Nh}+20L^2d\frac{h}{M}+20L^2\mathcal{E},$$

where $\mathcal{E}=\max_{n=0,\dots,N-1}\mathcal{E}_n^J$.

Proof. Integrating both sides of Lemma B.9,

$$\begin{aligned}\mathcal{L}(\mu_{n,m+1}^J)-\mathcal{L}(\mu_{n,m}^J)&\le-\left(\frac{3}{4}-36L^2\frac{h^2}{M^2}\right)\int_{nh+\frac{mh}{M}}^{nh+\frac{(m+1)h}{M}}\mathsf{FI}(\mu_t\|\pi_t)\mathrm{d}t\\&\quad+12L^3d\frac{h^3}{M^3}+9L^2d\frac{h^2}{M^2}\\&\quad+\left(36L^4\frac{h^3}{M^3}+9L^2\frac{h}{M}\right)\mathbb{E}\left[\left\|x_{n,m}^J-x_{n,m}^{J-1}\right\|^2\right]\\&\le-\frac{1}{2}\int_{nh+\frac{mh}{M}}^{nh+\frac{(m+1)h}{M}}\mathsf{FI}(\mu_t\|\pi_t)\mathrm{d}t+10L^2d\frac{h^2}{M^2}+10L^2\frac{h}{M}\mathcal{E}.\end{aligned}$$

Summing and rearranging,

$$\frac{2}{Nh}(\mathcal{L}(\mu_{N-1,M}^J)-\mathcal{L}(\mu_{0,0}^J))\le-\frac{1}{Nh}\int_0^{Nh}\mathsf{FI}(\mu_t\|\pi_t)\mathrm{d}t+20L^2d\frac{h}{M}+20L^2\mathcal{E},$$

so that

$$\frac{1}{Nh}\int_0^{Nh}\mathsf{FI}(\mu_t\|\pi_t)\mathrm{d}t\le\frac{2}{Nh}(\mathcal{L}(\mu^0)-\mathcal{L}(\mu_{N-1,M}^J))+20L^2d\frac{h}{M}+20L^2\mathcal{E}.$$

By the entropy sandwich (Lemma A.1),

$$\begin{aligned}\mathcal{L}(\mu^0)-\mathcal{L}(\mu_{N-1,M}^J)&=(\mathcal{L}(\mu^0)-\mathcal{L}(\pi^*))-(\mathcal{L}(\mu_{N-1,M}^J)-\mathcal{L}(\pi^*))\\&\le\mathsf{KL}(\mu^0\|\pi^0)-\mathsf{KL}(\mu_{N-1,M}^J\|\pi^*)\\&\le\mathsf{KL}(\mu^0\|\pi^0),\end{aligned}$$

and therefore

$$\frac{1}{Nh}\int_0^{Nh}\mathsf{FI}(\mu_t\|\pi_t)\mathrm{d}t\le\frac{2\mathsf{KL}(\mu^0\|\pi^0)}{Nh}+20L^2d\frac{h}{M}+20L^2\mathcal{E},$$

which completes the proof. □

Theorem 3.7 follows from Theorem B.10 and Lemma B.3.

B.3 Proof of Theorem 3.4

Theorem B.11. Under $Lh \le \frac{1}{20}$ and Assumptions 2.1 and 2.5,

$$\mathcal{L}(\mu^J_{n,m}) - \mathcal{L}(\pi^*) \le \exp\left(-\alpha\left(nh + \frac{mh}{M}\right)\right)(\mathcal{L}(\mu^0) - \mathcal{L}(\pi^*)) + \delta,$$

where $\delta = \frac{1}{\alpha}\left(36L^3 d\frac{h^2}{M^2} + 18L^2 d\frac{h}{M} + 10L^2\mathcal{E}\right)$.

Proof. By Lemma B.9, Assumption 2.5, and the entropy sandwich (Lemma A.1), for any $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$,

$$\begin{aligned}
&\partial_t \mathcal{L}(\mu_t) \\
&\le -\frac{1}{2}\mathsf{FI}(\mu_t\|\pi_t) + 36L^3 d\left(t - nh - \frac{mh}{M}\right)^2 + 18L^2 d\left(t - nh - \frac{mh}{M}\right) + 10L^2\mathcal{E} \\
&\le -\alpha\mathsf{KL}(\mu_t\|\pi_t) + 36L^3 d\left(t - nh - \frac{mh}{M}\right)^2 + 18L^2 d\left(t - nh - \frac{mh}{M}\right) + 10L^2\mathcal{E} \\
&\le -\alpha(\mathcal{L}(\mu_t) - \mathcal{L}(\pi^*)) + 36L^3 d\left(t - nh - \frac{mh}{M}\right)^2 + 18L^2 d\left(t - nh - \frac{mh}{M}\right) + 10L^2\mathcal{E} \\
&\le -\alpha(\mathcal{L}(\mu_t) - \mathcal{L}(\pi^*)) + 36L^3 d\frac{h^2}{M^2} + 18L^2 d\frac{h}{M} + 10L^2\mathcal{E} \\
&= -\alpha\left(\mathcal{L}(\mu_t) - \frac{c}{\alpha}\right),
\end{aligned}$$

so that

$$\partial_t\left(\mathcal{L}(\mu_t) - \frac{c}{\alpha}\right) \le -\alpha\left(\mathcal{L}(\mu_t) - \frac{c}{\alpha}\right),$$

where $c = \alpha\mathcal{L}(\pi^*) + 36L^3 d\frac{h^2}{M^2} + 18L^2 d\frac{h}{M} + 10L^2\mathcal{E}$. By GrönwalLs lemma, for $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$,

$$\mathcal{L}(\mu_t) - \frac{c}{\alpha} \le \exp\left(-\alpha\left(t - nh - \frac{mh}{M}\right)\right)\left(\mathcal{L}(\mu^J_{n,m}) - \frac{c}{\alpha}\right),$$

and in particular, setting $t = nh + \frac{(m+1)h}{M}$,

$$\mathcal{L}(\mu^J_{n,m+1}) - \frac{c}{\alpha} \le \exp\left(-\alpha\frac{h}{M}\right)\left(\mathcal{L}(\mu^J_{n,m}) - \frac{c}{\alpha}\right).$$

Applying this repeatedly,

$$\begin{aligned}
\mathcal{L}(\mu^J_{n,m}) - \frac{c}{\alpha} &\le \exp\left(-\alpha\left(nh + \frac{mh}{M}\right)\right)\left(\mathcal{L}(\mu^0) - \frac{c}{\alpha}\right) \\
&\le \exp\left(-\alpha\left(nh + \frac{mh}{M}\right)\right)\left(\mathcal{L}(\mu^0) - \mathcal{L}(\pi^*)\right).
\end{aligned}$$

Setting $\delta = \frac{1}{\alpha}\left(36L^3 d\frac{h^2}{M^2} + 18L^2 d\frac{h}{M} + 10L^2\mathcal{E}\right)$,

$$\mathcal{L}(\mu^J_{n,m}) - \mathcal{L}(\pi^*) \le \exp\left(-\alpha\left(nh + \frac{mh}{M}\right)\right)\left(\mathcal{L}(\mu^0) - \mathcal{L}(\pi^*)\right) + \delta$$

holds. □

Theorem 3.4 follows from Theorem B.11 and Lemma B.3.

C Missing Proofs in Section 4

C.1 Details of Algorithm 2 and 2′

Algorithm 2 Parallel Picard method for finite-particle MFLD

Input: $\mathbf{x}_{\mathrm{init}} \sim \mu^0$, access to a gradient oracle for $\nabla \frac{\delta F(\cdot)}{\delta \mu}$, number of iterations J, number of intervals $N(\leq J)$, interval width h, number of subdivision points per interval M.

for $n = 0, \dots, N-1, p = 1, \dots, P, m = 0, \dots, M-1$ **do**
  $\xi_{n,m}^{(p)} \sim \mathcal{N}\left(0, \frac{h}{M} I\right)$
for $n = 0, \dots, N-1, p = 1, \dots, P, m = 0, \dots, M-1$ **do**
  $B_{nh+\frac{(m+1)h}{M}}^{(p)} = \sum_{n'=0}^{n-1} \sum_{m'=0}^{M-1} \xi_{n',m'}^{(p)} + \sum_{m'=0}^{m} \xi_{n,m'}^{(p)}$
for $n = 0, \dots, N-1, p = 1, \dots, P, m = 0, \dots, M-1$ **do**
  $x_{n,m}^{0,(p)} = x_{\mathrm{init}}^{(p)}$
for $j = 1, \dots, J, p = 1, \dots, P$ **do**
  $x_{0,0}^{j,(p)} = x_{\mathrm{init}}^{(p)}$
for $k = 1, \dots, N$ **do**
  for $j = 1, \dots, \min\{k, J\}$ (in parallel) **do**
    Set $n = k - j$, and if $n \neq 0$, set $\mathbf{x}_{n,0}^j = \mathbf{x}_{n-1,M}^j$.
    for $p = 1, \dots, P$ (in parallel), $m = 1, \dots, M$ (in parallel) **do**
      $x_{n,m}^{j,(p)} = x_{n,0}^{j,(p)} - \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\rho_{n,m'}^{j-1})}{\delta \mu}(x_{n,m'}^{j-1,(p)}) + \sqrt{2}(B_{nh+\frac{mh}{M}}^{(p)} - B_{nh}^{(p)})$
for $k = N+1, \dots, N+J-1$ **do**
  for $n = \max\{0, k-J\}, \dots, N-1$ (in parallel) **do**
    Set $j = k - n$, and if $n \neq 0$, set $\mathbf{x}_{n,0}^j = \mathbf{x}_{n-1,M}^j$.
    for $p = 1, \dots, P$ (in parallel), $m = 1, \dots, M$ (in parallel) **do**
      $x_{n,m}^{j,(p)} = x_{n,0}^{j,(p)} - \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\rho_{n,m'}^{j-1})}{\delta \mu}(x_{n,m'}^{j-1,(p)}) + \sqrt{2}(B_{nh+\frac{mh}{M}}^{(p)} - B_{nh}^{(p)})$
Sample $t \sim \mathrm{Uniform}([0, Nh])$ and set $k = \left\lfloor \frac{t}{h/M} \right\rfloor$.
Set $n_k = \left\lfloor \frac{k}{M} \right\rfloor$ and $m_k = k - n_k M$.
$\mathbf{x}_t = \mathbf{x}_{n_k,m_k}^J - \left(t - \frac{kh}{M}\right) \nabla \frac{\delta F\left(\rho_{n_k,m_k}^{J-1}\right)}{\delta \mu} \left(\mathbf{x}_{n_k,m_k}^{J-1}\right) + \sqrt{2}(B_t - B_{\frac{kh}{M}})$
return $\mathbf{x}_t$

Algorithm 2′ Parallel Picard method for finite-particle MFLD

Input: $\mathbf{x}_{\mathrm{init}} \sim \mu^0$, access to a gradient oracle for $\nabla \frac{\delta F(\cdot)}{\delta \mu}$, number of iterations J, number of intervals $N(\leq J)$, interval width h, number of subdivision points per interval M.

for $n = 0, \dots, N-1, p = 1, \dots, P$, $m = 0, \dots, M-1$ **do**
  $\xi_{n,m}^{(p)} \sim \mathcal{N}\left(0, \frac{h}{M} I\right)$
for $n = 0, \dots, N-1, p = 1, \dots, P$, $m = 0, \dots, M-1$ **do**
  $B_{nh+\frac{(m+1)h}{M}}^{(p)} = \sum_{n'=0}^{n-1} \sum_{m'=0}^{M-1} \xi_{n',m'}^{(p)} + \sum_{m'=0}^{m} \xi_{n,m'}^{(p)}$
for $n = 0, \dots, N-1, p = 1, \dots, P$, $m = 0, \dots, M-1$ **do**
  $x_{n,m}^{0,(p)} = x_{\mathrm{init}}^{(p)}$
for $j = 1, \dots, J, p = 1, \dots, P$ **do**
  $x_{0,0}^{j,(p)} = x_{\mathrm{init}}^{(p)}$
for $k = 1, \dots, N$ **do**
  for $j = 1, \dots, \min\{k, J\}$ (in parallel) **do**
    Set $n = k - j$, and if $n \neq 0$, set $\mathbf{x}_{n,0}^j = \mathbf{x}_{n-1,M}^j$.
    for $p = 1, \dots, P$ (in parallel), $m = 1, \dots, M$ (in parallel) **do**
      $x_{n,m}^{j,(p)} = x_{n,0}^{j,(p)} - \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\rho_{n,m'}^{j-1})}{\delta \mu}(x_{n,m'}^{j-1,(p)}) + \sqrt{2}(B_{nh+\frac{mh}{M}}^{(p)} - B_{nh}^{(p)})$
for $k = N+1, \dots, N+J-1$ **do**
  for $n = \max\{0, k-J\}, \dots, N-1$ (in parallel) **do**
    Set $j = k - n$, and if $n \neq 0$, set $\mathbf{x}_{n,0}^j = \mathbf{x}_{n-1,M}^j$.
    for $p = 1, \dots, P$ (in parallel), $m = 1, \dots, M$ (in parallel) **do**
      $x_{n,m}^{j,(p)} = x_{n,0}^{j,(p)} - \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\rho_{n,m'}^{j-1})}{\delta \mu}(x_{n,m'}^{j-1,(p)}) + \sqrt{2}(B_{nh+\frac{mh}{M}}^{(p)} - B_{nh}^{(p)})$
return $\mathbf{x}_{N,0}^J = \mathbf{x}_{N-1,M}^J$

C.2 Proof of Theorem 4.2

In this section, we show that the output distribution of Algorithm 2 converges to the stationary distribution in the finite-particle setting. First, we provide the following lemma on the Lipschitz continuity of $\frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(\mathbf{x})$ in the following sense.

Lemma C.1. Under Assumption 2.1, $\nabla \frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(\mathbf{x})$ is $2L$-Lipschitz continuous. That is, for any $\mathbf{x}, \mathbf{y} \in (\mathbb{R}^d)^P$,

$$\left\| \nabla \frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(\mathbf{x}) - \nabla \frac{\delta F(\rho_{\mathbf{y}})}{\delta \mu}(\mathbf{y}) \right\| \leq 2L \|\mathbf{x} - \mathbf{y}\|$$

holds.

Proof. Let $Y \sim \mathrm{Uniform}(\{1, \dots, P\})$ be a random variable independent of all other random variables. For any $\mathbf{x} = (x^{(1)}, \dots, x^{(P)}), \mathbf{y} = (y^{(1)}, \dots, y^{(P)}) \in (\mathbb{R}^d)^P$,

$$\begin{aligned}
&\left\| \nabla \frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(\mathbf{x}) - \nabla \frac{\delta F(\rho_{\mathbf{y}})}{\delta \mu}(\mathbf{y}) \right\|^2 \\
&= \sum_{p=1}^{P} \left\| \nabla \frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(x^{(p)}) - \nabla \frac{\delta F(\rho_{\mathbf{y}})}{\delta \mu}(y^{(p)}) \right\|^2 \\
&\leq L^2 \sum_{p=1}^{P} \left(W_2(\rho_{\mathbf{x}}, \rho_{\mathbf{y}}) + \left\| x^{(p)} - y^{(p)} \right\| \right)^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 2L^2 \sum_{p=1}^{P} \left(W_2^2(\rho_{\mathbf{x}}, \rho_{\mathbf{y}}) + \left\| x^{(p)} - y^{(p)} \right\|^2 \right) \\
&\leq 2L^2 \sum_{p=1}^{P} \left(\mathbb{E}\left[\left\| x^{(Y)} - y^{(Y)} \right\|^2 \right] + \left\| x^{(p)} - y^{(p)} \right\|^2 \right) \\
&= 4L^2 \left\| \mathbf{x} - \mathbf{y} \right\|^2 .
\end{aligned}
$$

Therefore, $P\nabla F(\rho_{\mathbf{x}}) = \nabla \frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(\mathbf{x})$ is $2L$-Lipschitz continuous. □

Next, as in the infinite-particle case, we show that the parallelization error converges exponentially with respect to the degree of parallelism.

Lemma C.2. Under $Lh \leq \frac{1}{20}$ and Assumption 2.1, the following holds:

1. $\mathcal{E}_n^j \leq 2\mathcal{E}_{n-1}^j + 8L^2h^2\mathcal{E}_n^{j-1} \quad (j = 2, \ldots, J,\ n = 1, \ldots, N-1)$,
2. $\mathcal{E}_n^1 \leq 2\mathcal{E}_{n-1}^1 + 2h^2 \mathsf{FI}(\mu^0 \| \mu_*^{(P)}) + 5dPh \quad (j = 1,\ n = 1, \ldots, N-1)$,
3. $\mathcal{E}_0^j \leq 4L^2h^2\mathcal{E}_n^{j-1} \quad (j = 2, \ldots, J,\ n = 0)$,
4. $\mathcal{E}_0^1 \leq 2h^2 \mathsf{FI}(\mu^0 \| \mu_*^{(P)}) + 5dPh \quad (j = 1,\ n = 0)$,

where $\mathcal{E}_n^j \coloneqq \max_{m=0,1,\ldots,M} \mathbb{E}\left[\left\| \mathbf{x}_{n,m}^j - \mathbf{x}_{n,m}^{j-1} \right\|^2 \right]$ for $j = 1, \ldots, J,\ n = 0, \ldots, N-1$.

Proof. 1. For $j = 2, \ldots, J$, $n = 1, \ldots, N-1$, $m = 0, \ldots, M$,

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \mathbf{x}_{n,m}^j - \mathbf{x}_{n,m}^{j-1} \right\|^2 \right] \\
&= \mathbb{E}\left[\left\| (\mathbf{x}_{n,0}^j - \mathbf{x}_{n,0}^{j-1}) - \frac{h}{M} \sum_{m'=0}^{m-1} \left(\nabla \frac{\delta F(\rho_{n,m'}^{j-1})}{\delta \mu}(\mathbf{x}_{n,m'}^{j-1}) - \nabla \frac{\delta F(\rho_{n,m'}^{j-2})}{\delta \mu}(\mathbf{x}_{n,m'}^{j-2}) \right) \right\|^2 \right] \\
&\leq 2\mathbb{E}\left[\left\| \mathbf{x}_{n,0}^j - \mathbf{x}_{n,0}^{j-1} \right\|^2 \right] \\
&\qquad + 2\frac{h^2}{M^2}\mathbb{E}\left[\left\| \sum_{m'=0}^{m-1} \left(\nabla \frac{\delta F(\rho_{n,m'}^{j-1})}{\delta \mu}(\mathbf{x}_{n,m'}^{j-1}) - \nabla \frac{\delta F(\rho_{n,m'}^{j-2})}{\delta \mu}(\mathbf{x}_{n,m'}^{j-2}) \right) \right\|^2 \right] \\
&\leq 2\mathbb{E}\left[\left\| \mathbf{x}_{n,0}^j - \mathbf{x}_{n,0}^{j-1} \right\|^2 \right] + 8L^2h^2 \max_{m'=0,1,\ldots,M} \mathbb{E}\left[\left\| \mathbf{x}_{n,m'}^{j-1} - \mathbf{x}_{n,m'}^{j-2} \right\|^2 \right] \\
&= 2\mathcal{E}_{n-1}^j + 8L^2h^2\mathcal{E}_n^{j-1},
\end{aligned}
$$

which gives the claim.

2. For $j = 1$, $n = 1, \ldots, N-1$, $m = 0, \ldots, M$, by Lemma B.1,

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \mathbf{x}_{n,m}^1 - \mathbf{x}_{n,m}^0 \right\|^2 \right] \\
&\leq 2\mathbb{E}\left[\left\| \mathbf{x}_{n,0}^1 - \mathbf{x}_{n,0}^0 \right\|^2 \right] + 2\mathbb{E}\left[\left\| \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\rho_{\mathbf{x}^0})}{\delta \mu}(\mathbf{x}^0) + \sqrt{2}\left(B_{nh+\frac{mh}{M}} - B_{nh} \right) \right\|^2 \right] \\
&= 2\mathbb{E}\left[\left\| \mathbf{x}_{n,0}^1 - \mathbf{x}_{n,0}^0 \right\|^2 \right] + 2\frac{h^2}{M^2}\mathbb{E}\left[\left\| \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\rho_{\mathbf{x}^0})}{\delta \mu}(\mathbf{x}^0) \right\|^2 \right] + 4dP\frac{mh}{M} \\
&\leq 2\mathbb{E}\left[\left\| \mathbf{x}_{n,0}^1 - \mathbf{x}_{n,0}^0 \right\|^2 \right] + 2h^2\mathbb{E}\left[\left\| \frac{h}{M} \sum_{m'=0}^{m-1} \nabla \frac{\delta F(\rho_{\mathbf{x}^0})}{\delta \mu}(\mathbf{x}^0) \right\|^2 \right] + 4dPh \\
&\leq 2\mathbb{E}\left[\left\| \mathbf{x}_{n,0}^1 - \mathbf{x}_{n,0}^0 \right\|^2 \right] + 2h^2\left(\mathsf{FI}(\mu^0 \| \mu_*^{(P)}) + 4dPL \right) + 4dPh \\
&= 2\mathbb{E}\left[\left\| \mathbf{x}_{n-1,M}^1 - \mathbf{x}_{n-1,M}^0 \right\|^2 \right] + 2h^2\mathsf{FI}(\mu^0 \| \mu_*^{(P)}) + 4dPh(2Lh + 1)
\end{aligned}
$$

$$\leq 2\mathcal{E}_{n-1}^1 + 2h^2\mathsf{FI}(\mu^0\|\mu_*^{(P)}) + 5dPh,$$

which gives the claim.

3. For $j = 2, \dots, J$, $n = 0$, $m = 0, \dots, M$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}_{0,m}^j - \mathbf{x}_{0,m}^{j-1}\right\|^2\right] \\
&\leq \frac{h^2}{M^2}\mathbb{E}\left[\left\|\sum_{m'=0}^{m-1}\left(\nabla\frac{\delta F(\rho_{0,m'}^{j-1})}{\delta\mu}(\mathbf{x}_{0,m'}^{j-1}) - \nabla\frac{\delta F(\rho_{0,m'}^{j-2})}{\delta\mu}(\mathbf{x}_{0,m'}^{j-2})\right)\right\|^2\right] \\
&\leq h^2 \max_{m'=0,\dots,M}\mathbb{E}\left[\left\|\nabla\frac{\delta F(\rho_{0,m'}^{j-1})}{\delta\mu}(\mathbf{x}_{0,m'}^{j-1}) - \nabla\frac{\delta F(\rho_{0,m'}^{j-2})}{\delta\mu}(\mathbf{x}_{0,m'}^{j-2})\right\|^2\right] \\
&\leq 4L^2h^2 \max_{m'=0,\dots,M}\mathbb{E}\left[\left\|\mathbf{x}_{0,n'}^{j-1} - \mathbf{x}_{n,0}^{j-2}\right\|^2\right] \\
&\leq 4L^2h^2\mathcal{E}_n^{j-1},
\end{aligned}$$

which gives the claim.

4. For $j = 1$, $n = 0$, $m = 0, \dots, M$, by Lemma B.1,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}_{0,m}^1 - \mathbf{x}_{0,m}^0\right\|^2\right] \\
&= \mathbb{E}\left[\left\|-\frac{h}{M}\sum_{m'=0}^{m-1}\nabla\frac{\delta F(\rho_{\mathbf{x}^0})}{\delta\mu}(\mathbf{x}^0) + \sqrt{2}B_{\frac{mh}{M}}\right\|^2\right] \\
&\leq h^2 \max_{m'=0,\dots,M}\mathbb{E}\left[\left\|\nabla\frac{\delta F(\rho_{\mathbf{x}^0})}{\delta\mu}(\mathbf{x}^0)\right\|^2\right] + 2dP\frac{mh}{M} \\
&\leq h^2(\mathsf{FI}(\mu^0\|\mu_*^{(P)}) + 4dPL) + 2dPh \\
&= h^2\mathsf{FI}(\mu^0\|\mu_*^{(P)}) + 2dPh(2Lh+1) \\
&\leq 2h^2\mathsf{FI}(\mu^0\|\mu_*^{(P)}) + 5dPh,
\end{aligned}$$

which completes the proof. □

Lemma C.3. Under $J \geq N$ and $Lh \leq \frac{1}{20}$,

$$\mathcal{E} \leq 200(2h^2\mathsf{FI}(\mu^0\|\mu_*^{(P)}) + 5dPh)0.5^{J-N},$$

where $\mathcal{E} = \max_{n=0,\dots,N-1}\mathcal{E}_n^J$.

Proof. The result follows by the same argument as Lemma B.4. □

Next, corresponding to Lemma B.8 in the infinite-particle case, we establish the following estimate on the time evolution of the loss functional in the finite-particle case.

Lemma C.4. Let $(B_t)_{t\geq 0} = (B_t^{(1)}, \dots, B_t^{(P)})_{t\geq 0}$ be a standard Brownian motion on $(\mathbb{R}^d)^P$ independent of $(\mathbf{x}_0, \mathbf{g}_0)$, and let

$$\mathbf{x}_t = \mathbf{x}_0 - t\mathbf{g}_0 + \sqrt{2}B_t \quad (t \geq 0).$$

Then, denoting the distribution of $\mathbf{x}_t$ by μ_t,

$$\partial_t\mathcal{L}^{(P)}(\mu_t) \leq -\frac{3}{4}\mathsf{FI}(\mu_t\|\mu_*^{(P)}) + \mathbb{E}\left[\left\|\nabla\log\mu_*^{(P)} + \mathbf{g}_0\right\|\right]$$

holds.

Proof. As in the proof of Lemma B.5,

$$\partial_t\mu_t(\mathbf{x}) = \nabla\cdot(\mu_t(\mathbf{x})(\nabla\log\mu_t(\mathbf{x}) + \mathbb{E}\left[\mathbf{g}_0|\mathbf{x}_t = \mathbf{x}\right])).$$

Since

$$\frac{\delta \mathcal{L}^{(P)}(\mu_t)}{\delta \mu} = PF(\rho_{\mathbf{x}}) + \log \mu_t + \text{Const.} = \log \frac{\mu_t}{\mu_*^{(P)}},$$

we obtain

$$\begin{aligned}
\partial_t \mathcal{L}^{(P)}(\mu_t) &= \int \frac{\delta \mathcal{L}^{(P)}(\mu_t)}{\delta \mu}(\mathbf{x}) \frac{\partial \mu_t}{\partial t}(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \int \left(\log \frac{\mu_t}{\mu_*^{(P)}} \right) \nabla \cdot \left\{ \mu_t \left(\nabla \log \mu_t + \mathbb{E}\left[\mathbf{g}_0 | \mathbf{x}_t = \mathbf{x} \right] \right) \right\} \mathrm{d}\mathbf{x} \\
&= -\int \left\langle \nabla \log \frac{\mu_t}{\mu_*^{(P)}}, \nabla \log \mu_t + \mathbb{E}\left[\mathbf{g}_0 | \mathbf{x}_t = \mathbf{x} \right] \right\rangle \mathrm{d}\mu_t(\mathbf{x}) \\
&= -\int \left\langle \nabla \log \frac{\mu_t}{\mu_*^{(P)}}, \nabla \log \frac{\mu_t}{\mu_*^{(P)}} + \nabla \log \mu_*^{(P)} + \mathbb{E}\left[\mathbf{g}_0 | \mathbf{x}_t = \mathbf{x} \right] \right\rangle \mathrm{d}\mu_t(\mathbf{x}) \\
&= -\mathsf{FI}(\mu_t \| \mu_*^{(P)}) + \int \left\langle -\nabla \log \frac{\mu_t}{\mu_*^{(P)}}, \nabla \log \mu_*^{(P)} + \mathbb{E}\left[\mathbf{g}_0 | \mathbf{x}_t = \mathbf{x} \right] \right\rangle \mathrm{d}\mu_t(\mathbf{x}) \\
&\leq -\mathsf{FI}(\mu_t \| \mu_*^{(P)}) + \frac{1}{4} \int \left\| \nabla \log \frac{\mu_t}{\mu_*^{(P)}} \right\|^2 \mathrm{d}\mu_t \\
&\quad + \int \left\| \nabla \log \mu_*^{(P)} + \mathbb{E}\left[\mathbf{g}_0 | \mathbf{x}_t = \mathbf{x} \right] \right\|^2 \mathrm{d}\mu_t \\
&= -\frac{3}{4} \mathsf{FI}(\mu_t \| \mu_*^{(P)}) + \mathbb{E}\left[\left\| \nabla \log \mu_*^{(P)} + \mathbb{E}\left[\mathbf{g}_0 | \mathbf{x}_t \right] \right\|^2 \right] \\
&\leq -\frac{3}{4} \mathsf{FI}(\mu_t \| \mu_*^{(P)}) + \mathbb{E}\left[\left\| \nabla \log \mu_*^{(P)} + \mathbf{g}_0 \right\|^2 \right],
\end{aligned}$$

which completes the proof. □

Applying this to the one-step interpolation of Algorithm 2, we obtain the following estimate.

Lemma C.5. Under $Lh \leq \frac{1}{12}$ and Assumption 2.1,

$$\begin{aligned}
\partial_t \mathcal{L}^{(P)}(\mu_t) \leq &-\left(\frac{3}{4} - \frac{9h^2}{2M^2} \right) \mathsf{FI}(\mu_t \| \mu_*^{(P)}) + 9dPL^2 \left(t - nh - \frac{mh}{M} \right) \\
&+ 72dPL^3 \left(t - nh - \frac{mh}{M} \right)^2 + 9L^2 \mathbb{E}\left[\left\| \mathbf{x}_{n,m}^J - \mathbf{x}_{n,m}^{J-1} \right\|^2 \right]
\end{aligned}$$

holds.

Proof. By Lemma C.4,

$$\partial_t \mathcal{L}^{(P)}(\mu_t) \leq -\frac{3}{4} \mathsf{FI}(\mu_t \| \mu_*^{(P)}) + \mathbb{E}\left[\left\| \nabla \log \mu_*^{(P)} + \nabla \frac{\delta F(\rho_{n,m}^{J-1})}{\delta \mu}(\mathbf{x}_{n,m}^{J-1}) \right\|^2 \right].$$

Here, by [Chi22], $\nabla \log \mu_*^{(P)} = -P \nabla F(\rho_{\mathbf{x}}) = -\nabla \frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(\mathbf{x})$, so that

$$\begin{aligned}
&\mathbb{E}\left[\left\| \nabla \log \mu_*^{(P)} + \nabla \frac{\delta F(\rho_{n,m}^{J-1})}{\delta \mu}(\mathbf{x}_{n,m}^{J-1}) \right\|^2 \right] \\
&= \mathbb{E}\left[\left\| \nabla \frac{\delta F(\rho_{\mathbf{x}_t})}{\delta \mu}(\mathbf{x}_t) - \nabla \frac{\delta F(\rho_{n,m}^{J-1})}{\delta \mu}(\mathbf{x}_{n,m}^{J-1}) \right\|^2 \right] \\
&\leq 4L^2 \mathbb{E}\left[\left\| \mathbf{x}_t - \mathbf{x}_{n,m}^{J-1} \right\|^2 \right]
\end{aligned}$$

$$
\begin{aligned}
&\leq 8L^2\mathbb{E}\left[\left\|\mathbf{x}_{n,m}^J-\mathbf{x}_{n,m}^{J-1}\right\|^2\right]+8L^2\mathbb{E}\left[\left\|B_t-B_{nh+\frac{mh}{M}}\right\|^2\right]\\
&\quad+8L^2\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\rho_{n,m}^{J-1})}{\delta\mu}(\mathbf{x}_{n,m}^{J-1})\right\|^2\right]\\
&\leq 8L^2\mathbb{E}\left[\left\|\mathbf{x}_{n,m}^J-\mathbf{x}_{n,m}^{J-1}\right\|^2\right]+16L^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mathbf{x}_t)}{\delta\mu}(\mathbf{x}_t)\right\|^2\right]+8L^2\mathbb{E}\left[\left\|B_t-B_{nh+\frac{mh}{M}}\right\|^2\right]\\
&\quad+16L^2\left(t-nh-\frac{mh}{M}\right)^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\rho_{\mathbf{x}_t})}{\delta\mu}(\mathbf{x}_t)-\nabla\frac{\delta F(\rho_{n,m}^{J-1})}{\delta\mu}(\mathbf{x}_{n,m}^{J-1})\right\|^2\right],
\end{aligned}
$$

from which

$$
\begin{aligned}
&\left(1-16L^2\frac{h^2}{M^2}\right)\mathbb{E}\left[\left\|\nabla\frac{\delta F(\rho_{\mathbf{x}_t})}{\delta\mu}(\mathbf{x}_t)-\nabla\frac{\delta F(\rho_{n,m}^{J-1})}{\delta\mu}(\mathbf{x}_{n,m}^{J-1})\right\|^2\right]\\
&\leq 8L^2\mathbb{E}\left[\left\|\mathbf{x}_{n,m}^J-\mathbf{x}_{n,m}^{J-1}\right\|^2\right]+16L^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mathbf{x}_t)}{\delta\mu}(\mathbf{x}_t)\right\|^2\right]+8L^2\mathbb{E}\left[\left\|B_t-B_{nh+\frac{mh}{M}}\right\|^2\right].
\end{aligned}
$$

For $1-16L^2h^2\leq\frac{8}{9}$, i.e., $Lh\leq\frac{1}{12}$, using Lemma B.1,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\nabla\frac{\delta F(\rho_{\mathbf{x}_t})}{\delta\mu}(\mathbf{x}_t)-\nabla\frac{\delta F(\rho_{n,m}^{J-1})}{\delta\mu}(\mathbf{x}_{n,m}^{J-1})\right\|^2\right]\\
&\leq 9L^2\mathbb{E}\left[\left\|\mathbf{x}_{n,m}^J-\mathbf{x}_{n,m}^{J-1}\right\|^2\right]+18L^2\mathbb{E}\left[\left\|\nabla\frac{\delta F(\mathbf{x}_t)}{\delta\mu}(\mathbf{x}_t)\right\|^2\right]+9L^2\mathbb{E}\left[\left\|B_t-B_{nh+\frac{mh}{M}}\right\|^2\right]\\
&\leq 9L^2\mathbb{E}\left[\left\|\mathbf{x}_{n,m}^J-\mathbf{x}_{n,m}^{J-1}\right\|^2\right]+18L^2\left(\mathsf{FI}(\mu_t\|\mu_*^{(P)})+2dP(2L)\right)\\
&\quad+9L^2dP\left(t-nh-\frac{mh}{M}\right).
\end{aligned}
$$

From the initial inequality,

$$
\begin{aligned}
\partial_t\mathcal{L}^{(P)}(\mu_t)\leq&-\left(\frac{3}{4}-18L^2\frac{h^2}{M^2}\right)\mathsf{FI}(\mu_t\|\mu_*^{(P)})+9dPL^2\left(t-nh-\frac{mh}{M}\right)\\
&+72dPL^3\left(t-nh-\frac{mh}{M}\right)^2+9L^2\mathbb{E}\left[\left\|\mathbf{x}_{n,m}^J-\mathbf{x}_{n,m}^{J-1}\right\|^2\right].
\end{aligned}
$$

□

Noting that in the Fisher divergence, the target distribution is fixed, we can show that the output distribution $\overline{\mu}_{Nh}$ of Algorithm 2 converges to $\mu_*^{(P)}$ in the sense of Fisher divergence at the following rate.

Theorem C.6. Under $Lh\leq\frac{1}{20}$ and Assumption 2.1,

$$
\mathsf{FI}(\overline{\mu}_{Nh}\|\mu_*^{(P)})\leq\frac{2\mathcal{L}^{(P)}(\mu^0)}{Nh}+13dPL^2\frac{h}{M}+18L^2\mathcal{E},
$$

where $\overline{\mu}_{Nh}=\frac{1}{Nh}\int_0^{Nh}\mu_t\mathrm{d}t$.

Proof. Integrating both sides of Lemma C.5,

$$
\begin{aligned}
&\mathcal{L}^{(P)}(\mu_{n,m+1}^J)-\mathcal{L}^{(P)}(\mu_{n,m}^J)\\
&\leq-\left(\frac{3}{4}-18\frac{h^2}{M^2}\right)\int_{nh+\frac{mh}{M}}^{nh+\frac{(m+1)h}{M}}\mathsf{FI}(\mu_t\|\mu_*^{(P)})+24dPL^3\frac{h^3}{M^3}+\frac{9}{2}dPL^2\frac{h^2}{M^2}+9L^2\mathcal{E}\frac{h}{M}
\end{aligned}
$$

$$\leq -\frac{1}{2}\int_{nh+\frac{mh}{M}}^{nh+\frac{(m+1)h}{M}} \mathsf{FI}(\mu_t\|\mu_*^{(P)}) + \frac{13}{2}dPL^2\frac{h^2}{M^2} + 9L^2\mathcal{E}\frac{h}{M},$$

from which

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t\|\mu_*^{(P)})\mathrm{d}t \leq \frac{2}{Nh}(\mathcal{L}^{(P)}(\mu^0) - \mathcal{L}^{(P)}(\mu_{N-1,M}^J)) + 13dPL^2\frac{h}{M} + 18L^2\mathcal{E}.$$

Since $\frac{1}{P}\mathcal{L}^{(P)}(\mu) - \mathcal{L}(\pi^*) \geq 0$ for any distribution μ (shown in [CRW25]),

$$\begin{aligned}\mathcal{L}^{(P)}(\mu^0) - \mathcal{L}^{(P)}(\mu_{N-1,M}^J) &= (\mathcal{L}^{(P)}(\mu^0) - P\mathcal{L}(\pi^*)) - (\mathcal{L}^{(P)}(\mu_{N-1,M}^J) - P\mathcal{L}(\pi^*)) \\ &\leq (\mathcal{L}^{(P)}(\mu^0) - P\mathcal{L}(\pi^*)).\end{aligned}$$

Using the convexity of the Fisher divergence in its first argument,

$$\begin{aligned}\mathsf{FI}(\overline{\mu}_{Nh}\|\mu_*^{(P)}) &\leq \frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t\|\mu_*^{(P)})\mathrm{d}t \\ &\leq \frac{2P\Delta_0^{(P)}}{Nh} + 13dPL^2\frac{h}{M} + 18L^2\mathcal{E}.\end{aligned}$$

This completes the proof. □

Theorem 4.2 follows from Theorem C.6 and Lemma B.4.

C.3 Proof of Theorem 4.3

In this section, we show that under assumptions ensuring the defective LSI, the output $\mu_{N,0}^J$ of Algorithm 2′ converges to the true stationary distribution π^*.

Theorem C.7. Under $Lh \leq \frac{1}{20}$ and Assumptions 2.1 and 2.6,

$$\frac{1}{P}\mathcal{L}^{(P)}(\mu_{n,m}^J) - \mathcal{L}(\pi^*) \leq \exp\left(-\alpha\left(nh + \frac{mh}{M}\right)\right)\Delta_0^{(P)} + \delta$$

holds, where $\Delta_0^{(P)} = \frac{1}{P}\mathcal{L}^{(P)}(\mu^0) - \mathcal{L}(\pi^*)$ and $\delta = \frac{B}{P} + 9\frac{L^2\mathcal{E}}{\alpha P} + 15\frac{dL^2}{\alpha}\frac{h}{M}$.

Proof. By Lemma C.5 and Lemma 2.7, for $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$,

$$\begin{aligned}\partial_t\mathcal{L}^{(P)}(\mu_t) &\leq -\left(\frac{3}{4} - \frac{9h^2}{2M^2}\right)\mathsf{FI}(\mu_t\|\mu_*^{(P)}) + 9dPL^2\left(t - nh - \frac{mh}{M}\right) \\ &\quad + 72dPL^3\left(t - nh - \frac{mh}{M}\right)^2 + 9L^2\mathbb{E}\left[\left\|\mathbf{x}_{n,m}^J - \mathbf{x}_{n,m}^{J-1}\right\|^2\right] \\ &\leq -\frac{1}{2}\mathsf{FI}(\mu_t\|\mu_*^{(P)}) + 15dPL^2\left(t - nh - \frac{mh}{M}\right) + 9L^2\mathcal{E} \\ &\leq -\frac{1}{2}\cdot 2\alpha P\left(\frac{1}{P}\mathcal{L}^{(P)}(\mu_t) - \mathcal{L}(\pi^*) - \frac{B}{P}\right) + 15dPL^2\frac{h}{M} + 9L^2\mathcal{E} \\ &= -\alpha\mathcal{L}^{(P)}(\mu_t) + \left(\alpha P\mathcal{L}(\pi^*) + \alpha B + 9L^2\mathcal{E} + 15dPL^2\frac{h}{M}\right) \\ &= -\alpha\left(\mathcal{L}^{(P)}(\mu_t) - \frac{c}{\alpha}\right),\end{aligned}$$

so that

$$\partial_t\left(\mathcal{L}^{(P)}(\mu_t) - \frac{c}{\alpha}\right) \leq -\alpha\left(\mathcal{L}^{(P)}(\mu_t) - \frac{c}{\alpha}\right),$$

where

$$c = \alpha P\mathcal{L}(\pi^*) + \alpha B + 9L^2\mathcal{E} + 15dPL^2\frac{h}{M}.$$

By GrönwalLs lemma, for $t \in \left[nh + \frac{mh}{M}, nh + \frac{(m+1)h}{M}\right]$,

$$\mathcal{L}^{(P)}(\mu_t) - \frac{c}{\alpha} \leq \exp\left(-\alpha\left(t - nh - \frac{h}{M}\right)\right)\left(\mathcal{L}^{(P)}(\mu_{n,m}^J) - \frac{c}{\alpha}\right),$$

and in particular,

$$\mathcal{L}^{(P)}(\mu^J_{n,m+1}) - \frac{c}{\alpha} \leq \exp\left(-\alpha \frac{h}{M}\right)\left(\mathcal{L}^{(P)}(\mu^J_{n,m}) - \frac{c}{\alpha}\right).$$

Applying this repeatedly,

$$\mathcal{L}^{(P)}(\mu^J_{n,m}) - \frac{c}{\alpha} \leq \exp\left(-\alpha\left(nh + \frac{mh}{M}\right)\right)\left(\mathcal{L}^{(P)}(\mu^0) - \frac{c}{\alpha}\right).$$

That is, setting $\delta = \frac{B}{P} + 9\frac{L^2\mathcal{E}}{\alpha P} + 15\frac{dL^2}{\alpha}\frac{h}{M} > 0$,

$$\begin{aligned}\frac{1}{P}\mathcal{L}^{(P)}(\mu^J_{n,m}) - \mathcal{L}(\pi^*) &\leq \exp\left(-\alpha\left(nh + \frac{mh}{M}\right)\right)\left(\frac{1}{P}\mathcal{L}^{(P)}(\mu^0) - \mathcal{L}(\pi^*) - \delta\right) + \delta \\ &\leq \exp\left(-\alpha\left(nh + \frac{mh}{M}\right)\right)\left(\frac{1}{P}\mathcal{L}^{(P)}(\mu^0) - \mathcal{L}(\pi^*)\right) + \delta.\end{aligned}$$

This completes the proof. □

Theorem 4.3 follows from Theorem C.7 and Lemma B.4.

C.4 Proofs of Lemmas on the Initial Distribution

In this section, we prove the lemmas on the initial distribution stated in Section 3.2 and 4.2. We use the following lemma on the KL divergence of the initial distribution.

Lemma C.8 (Lemma 17 in [CGLL23])**.** Let $V : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and satisfy $V(0) - \inf V \leq \Delta$. Assume that for the distribution $\pi \propto \exp(-V)$, $\mathfrak{m} := \int \|\cdot\| \mathrm{d}\pi < \infty$ holds. Then for $\mu^0 = \mathcal{N}(0, L^{-1}I_d)$,

$$\mathsf{KL}(\mu^0\|\pi) \lesssim \Delta + d(1 \vee \log(L\mathfrak{m}^2))$$

holds.

As a corollary of this lemma, we can establish the following.

Corollary C.9. Suppose Assumption 2.1 holds and let $\mu^0 = \mathcal{N}(0, L_2^{-1}I_d)$. Furthermore, assume that $\frac{\delta F(\mu^0)}{\delta \mu}(0) - \inf_{x\in\mathbb{R}^d} \frac{\delta F(\mu^0)}{\delta \mu}(x) \leq \Delta$, and that for the distribution $\pi^0 = \pi[\mu^0]$, $\mathfrak{m} := \int \|\cdot\| \mathrm{d}\pi^0 < \infty$ holds. Then

$$\mathsf{KL}(\mu^0\|\pi^*) \leq \mathcal{L}(\mu^0) - \mathcal{L}(\pi^*) \lesssim \Delta + d(1 \vee \log(L_2\mathfrak{m}^2))$$

holds.

Proof. Setting $V(x) = \frac{\delta F(\mu^0)}{\delta \mu}(x)$ and applying Lemma C.8,

$$\mathsf{KL}(\mu^0\|\pi^0) \lesssim \Delta + d(1 \vee \log(L_2\mathfrak{m}^2)).$$

Combining with the entropy sandwich (Lemma A.1),

$$\mathsf{KL}(\mu^0\|\pi^*) \leq \mathcal{L}(\mu^0) - \mathcal{L}(\pi^*) \leq \mathsf{KL}(\mu^0\|\pi^0) \lesssim \Delta + d(1 \vee \log(L_2\mathfrak{m}^2)),$$

which gives the claim. □

Next, we establish the following lemma on the Fisher divergence of the initial distribution.

Lemma C.10. Suppose Assumption 2.1 holds and let $\mu^0 = \mathcal{N}(0, L_2^{-1}I_d)$. Then

$$\mathsf{FI}(\mu^0\|\pi^0) \leq 6L_2 + 4\left\|\nabla \frac{\delta F(\mu^0)}{\delta \mu}(0)\right\|^2$$

holds. In particular, if $\left\|\nabla \frac{\delta F(\mu^0)}{\delta \mu}(0)\right\|^2 = O(d)$, then $\mathsf{FI}(\mu^0\|\pi^0) = O(L + d)$.

Proof. Noting that $\pi^0(x) \propto \exp\left(-\frac{\delta F(\mu^0)}{\delta \mu}(x)\right)$,

$$\begin{aligned}&\mathsf{FI}(\mu^0\|\pi^0) \\ &= \int \left\|\nabla \log \frac{\mathrm{d}\mu^0}{\mathrm{d}\pi^0}\right\|^2 \mathrm{d}\mu^0\end{aligned}$$

$$
\begin{aligned}
&= \int \left\| -L_2 x + \nabla \frac{\delta F(\mu^0)}{\delta \mu}(x) \right\|^2 \mathrm{d}\mu^0(x) \\
&\leq 2 \int \left\| -L_2 x + \nabla \frac{\delta F(\mu^0)}{\delta \mu}(0) \right\|^2 \mathrm{d}\mu^0(x) + 2 \int \left\| \nabla \frac{\delta F(\mu^0)}{\delta \mu}(x) - \nabla \frac{\delta F(\mu^0)}{\delta \mu}(0) \right\|^2 \mathrm{d}\mu^0(x) \\
&\leq 6L_2^2 \int \|x\|^2 \mathrm{d}\mu^0(x) + 4 \left\| \nabla \frac{\delta F(\mu^0)}{\delta \mu}(0) \right\|^2 \\
&= 6L_2 + 4 \left\| \nabla \frac{\delta F(\mu^0)}{\delta \mu}(0) \right\|^2,
\end{aligned}
$$

which completes the proof. □

Lemma C.11. Suppose Assumption 2.1 holds and let $\mu^0 = \mathcal{N}(0, (2L)^{-1} I_{dP})$. Then

$$
\mathsf{FI}(\mu^0 \| \mu_*^{(P)}) \leq 6L + 4 \left\| \nabla \frac{\delta F(\rho_{\mathbf{0}})}{\delta \mu}(\mathbf{0}) \right\|^2
$$

holds. In particular, if $\left\| \nabla \frac{\delta F(\rho_{\mathbf{0}})}{\delta \mu}(\mathbf{0}) \right\|^2 = O(dP)$, then $\mathsf{FI}(\mu^0 \| \mu_*^{(P)}) = O(L + dP)$.

Proof. Using Lemma C.1, the same argument as in Lemma C.10 gives

$$
\begin{aligned}
&\mathsf{FI}(\mu^0 \| \pi^0) \\
&= \int \left\| \nabla \log \frac{\mathrm{d}\mu^0}{\mathrm{d}\pi^0} \right\|^2 \mathrm{d}\mu^0 \\
&= \int \left\| -L_2 \mathbf{x} + \nabla \frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(\mathbf{x}) \right\|^2 \mathrm{d}\mu^0(\mathbf{x}) \\
&\leq 2 \int \left\| -L_2 \mathbf{x} + \nabla \frac{\delta F(\rho_{\mathbf{0}})}{\delta \mu}(\mathbf{0}) \right\|^2 \mathrm{d}\mu^0(\mathbf{x}) + 2 \int \left\| \nabla \frac{\delta F(\rho_{\mathbf{x}})}{\delta \mu}(\mathbf{x}) - \nabla \frac{\delta F(\rho_{\mathbf{0}})}{\delta \mu}(\mathbf{0}) \right\|^2 \mathrm{d}\mu^0(\mathbf{x}) \\
&\leq (4L_2^2 + 8L^2) \int \|\mathbf{x}\|^2 \mathrm{d}\mu^0(\mathbf{x}) + 4 \left\| \nabla \frac{\delta F(\rho_{\mathbf{0}})}{\delta \mu}(\mathbf{0}) \right\|^2 \\
&\leq 6L + 4 \left\| \nabla \frac{\delta F(\rho_{\mathbf{0}})}{\delta \mu}(\mathbf{0}) \right\|^2.
\end{aligned}
$$

□

In the finite-particle setting, the entropy sandwich cannot be applied directly, but under Assumption 2.6-(A3), the following estimate is known to hold.

Lemma C.12 (Lemma 3.7 in [NLK+25])**.** Under Assumption 2.6-(A3), for any $\mu \in \mathcal{P}_2((\mathbb{R}^d)^P)$,

$$
\frac{1}{P} \mathcal{L}^{(P)}(\mu) - \mathcal{L}(\pi^*) \leq \frac{B}{P} + \frac{1}{P} \sum_{i=1}^{P} \mathbb{E}_{\mathbf{X} \sim \mu} \left[\mathsf{KL}(\mu_{i \| -i}(\cdot | \mathbf{X}^{-i}) \| \nu_{i|-i}(\cdot | \mathbf{X}^{-i})) \right]
$$

holds, where $\mu_{i|-i}(\cdot | \mathbf{x}^{-i})$ denotes the conditional distribution of X^i given $\mathbf{X}^{-i} = \mathbf{x}^{-i}$ for $\mathbf{X} \sim \mu$.

Using this, we can establish the following lemma.

Lemma C.13. Suppose Assumptions 2.1 and 2.6-(A3) hold, and that for all $\mathbf{x} \in (\mathbb{R}^d)^P$ and all i, $\frac{\delta F(\rho_{0 \cup \mathbf{x}^{-i}})}{\delta \mu}(0) - \inf_{x \in \mathbb{R}^d} \frac{\delta F(\rho_{x \cup \mathbf{x}^{-i}})}{\delta \mu}(x) \leq \Delta$. Let $\mu^0 = \mathcal{N}\left(0, (2L)^{-1} I_{dP}\right)$, and assume that for all i, $\mathfrak{m}_i = \int \| \cdot \| \mathrm{d}\nu_{i|-i}(\cdot | \mathbf{X}^{-i}) < \infty$ holds for $\nu_{i|-i}(\cdot | \mathbf{X}^{-i})$, and set $\mathfrak{m} = \max_{i=1,\dots,P} \mathfrak{m}_i$. Then

$$
\frac{1}{P} \mathcal{L}^{(P)}(\mu^0) - \mathcal{L}(\pi^*) \lesssim \frac{B}{P} + \Delta + d \left(1 \vee \log \left(2L\mathfrak{m}^2\right)\right)
$$

holds.

Proof. Viewing with respect to $x^i = x$, by [Chi22], $P\nabla_x F(\rho_{x\cup\mathbf{x}^{-i}}) = \left[\nabla \frac{\delta F(\rho_{x\cup\mathbf{x}^{-i}})}{\delta \mu}\right](x)$, and in particular,

$$\begin{aligned}
&\left\|P\nabla_x F(\rho_{x\cup\mathbf{x}^{-i}}) - P\nabla_x F(\rho_{y\cup\mathbf{x}^{-i}})\right\| \\
&= \left\|\nabla \frac{\delta F(\rho_{x\cup\mathbf{x}^{-i}})}{\delta \mu}(x) - \nabla \frac{\delta F(\rho_{y\cup\mathbf{x}^{-i}})}{\delta \mu}(y)\right\| \\
&\leq L_1 W_2(\rho_{x\cup\mathbf{x}^{-i}}, \rho_{y\cup\mathbf{x}^{-i}}) + L_2 \left\|x - y\right\| \\
&\leq \left(\frac{L_1}{\sqrt{P}} + L_2\right) \left\|x - y\right\| \\
&\leq 2L \left\|x - y\right\|,
\end{aligned}$$

so $P\nabla_x F(\rho_{x\cup\mathbf{x}^{-i}})$ is $2L$-smooth. When $\mu^0 = \mathcal{N}\left(0, (2L)^{-1} I_{dP}\right)$, $\mu_{i|-i}(\cdot|\mathbf{x}^{-i}) = \mathcal{N}\left(0, (2L)^{-1} I_d\right) =: \gamma_d$. Applying Lemma C.8 with $V_i(x) = P\nabla_x F(\rho_{x\cup\mathbf{x}^{-i}})$,

$$\begin{aligned}
\frac{1}{P}\mathcal{L}^{(P)}(\mu^0) - \mathcal{L}(\pi^*) &\leq \frac{B}{P} + \frac{1}{P}\sum_{i=1}^{P} \mathbb{E}_{\mathbf{X}\sim\mu}\left[\mathsf{KL}(\mu_{i\|-i}(\cdot|\mathbf{X}^{-i})\|\nu_{i|-i}(\cdot|\mathbf{X}^{-i}))\right] \\
&\lesssim \frac{B}{P} + \Delta + \frac{d}{P}\sum_{i=1}^{P}\left(1 \vee \log\left(2L\mathfrak{m}_{\mathfrak{i}}{}^2\right)\right) \\
&\lesssim \frac{B}{P} + \Delta + d\left(1 \vee \log\left(2L\mathfrak{m}^2\right)\right).
\end{aligned}$$

This completes the proof. □

Lemma 3.1 follows from Lemmas C.8, C.9, and C.10. Lemma 4.1 follows from Lemmas C.11 and C.13.

D Broader Impacts

This work proposes parallel algorithms for mean-field Langevin dynamics (MFLD), which may contribute to accelerating the training of mean-field neural networks and other related models. As a consequence, the proposed methods could potentially lead to faster training of AI models more broadly. While we do not foresee direct negative societal impacts specific to this theoretical work, indirect impacts may arise if the methods are applied in contexts with potential for misuse, as with any advance in computational efficiency for machine learning.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction accurately reflect the theoretical contributions, with assumptions clearly stated in the contributions and summarized in Tables 1 and 2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are stated with full assumptions and complete proofs. The proofs are provided in the appendix, with proof sketches given in the main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This paper is theoretical and does not include experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: This paper does not include experiments requiring code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [N/A]

Justification: This paper is theoretical and does not include experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: This paper is theoretical and does not include experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [N/A]

Justification: This paper is theoretical and does not include experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks as it is purely theoretical and does not release any models or datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used for writing and editing purposes, and occasionally for clarification during proof development. However, they did not contribute to the core methodology, theoretical results, or proofs presented in this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.